UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

LARISSA VASCONCELLOS DE MORAES

Detecção de Depressão pela Fala Empregando Rede Neurais Profundas

Goiânia
2020

PRPG

UFG

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: **[X] Dissertação** **[ ] Tese**

2. Identificação da Tese ou Dissertação:

Nome completo do autor: Larissa Vasconcellos de Moraes

Título do trabalho: Detecção de Depressão pela Fala Empregando Rede Neurais Profundas

3. Informações de acesso ao documento:

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.

Assinatura do(a) autor(a)[2]

Ciente e de acordo:

Assinatura do(a) orientador(a)[2]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente
- Submissão de artigo em revista científica
- Publicação como capítulo de livro
- Publicação da dissertação/tese em livro

[2] A assinatura deve ser escaneada.

LARISSA VASCONCELLOS DE MORAES

Detecção de Depressão pela Fala Empregando Rede Neurais Profundas

Dissertação apresentada ao Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre no Programa de Pós-Graduação em Ciência da Computação.

Área de concentração: Ciência da Computação.

Orientador: Prof. Dr. Anderson da Silva Soares

Goiânia
2020

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Vasconcellos de Moraes, Larissa
Detecção de Depressão pela Fala Empregando Rede Neurais Profundas [manuscrito] / Larissa Vasconcellos de Moraes. - 2020.
LXII, 62 f.: il.

Orientador: Prof. Dr. Anderson da Silva Soares.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2020.
Bibliografia.

1. Reconhecimento de Emoção de Fala. 2. Reconhecimento de Fala. 3. Rede Neural. 4. Classificação de Depressão. I. Soares, Anderson da Silva , orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **05/2020** da sessão de Defesa de Dissertação de **Larissa Vasconcellos de Moraes**, que confere o título de Mestra em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos dez dias do mês de fevereiro de dois mil e vinte, a partir das quatorze horas, na sala 150 do Instituto de Informática, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Detecção de Depressão pela Fala Empregando Rede Neurais Profundas"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Anderson da Silva Soares (INF/UFG)**,** com a participação dos demais membros da Banca Examinadora: Professor Doutor Rogério Lopes Salvini (INF/UFG), membro titular interno; Professor Doutor Arlindo Rodrigues Galvão Filho (ECEC/PUC-GO), membro titular externo. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido a candidata **aprovada** pelos seus membros. Proclamados os resultados pelo Professor Doutor Anderson da Silva Soares, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos dez dias do mês de fevereiro de dois mil e vinte.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Anderson Da Silva Soares**, **Professor do Magistério Superior**, em 10/02/2020, às 17:08, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rogerio Lopes Salvini**, **Professor do Magistério Superior**, em 10/02/2020, às 17:28, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Arlindo Rodrigues Galvão Filho**, **Usuário Externo**, em 28/02/2020, às 16:41, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1101892** e o código CRC **AC68DAA5**.

Referência: Processo nº 23070.000294/2020-01 SEI nº 1101892

Larissa Vasconcellos de Moraes

Graduou-se em Ciência da Computação na UFMT - Universidade Federal de Mato Grosso. Durante sua graduação, realizou monitorias como bolsista da CAPES. Durante o Mestrado, na UFG - Universidade Federal de Goiás, foi bolsista da CAPES, participou de projetos junto a "Deep Learning Brasil"e Unifesp em um desafio lançado pela RSNA em 2017, conquistou o primeiro lugar, trabalho eleito pelo público, no 2º Workshop de Inteligência Artificial organizado pelo Instituto de Informática da UFG, com o projeto para auxiliar no diagnóstico da depressão por meio da fala com o uso de redes neurais convolucionais.

À memória de minha avó, Isaura da Silva Moraes, maior responsável pela pessoa que me tornei e pela base que criou para eu poder alcançar mais um sonho.

Agradecimentos

Primeiramente agradeço aos meus pais (Renato e Eneida) e irmão (Antônio), por terem me apoiado em todos os momentos não me deixando desistir de concluir essa etapa e me ajudando em todos os momentos, bons e ruins, que passei no decorrer deste caminho. A memória de minha avó (Isaura), que provavelmente estaria me apoiando mesmo não entendendo o que faço.

Quero agradecer também ao meu orientador Prof. Dr. Anderson da Silva Soares, por toda compreensão em todos os momentos que passei no decorrer deste caminho, a paciência, ensinamentos e tranquilidade transmitida e principalmente pelas oportunidades em que confiou a mim, fazendo com que eu crescesse ainda mais. Além de agradecer ao meu chefe, Prof. Dr. Birajara Machado e meus colegas de trabalho, em especial a Prof.ª Dr.ª Luciana Moura e Msc.ª Joselisa Paiva que me apoiaram na conclusão desta etapa.

Agradeço também aos meus primos (Michele e Daniel) que me ajudaram a passar pelo pior momento ao longo deste caminho, além de me apoiarem, se interessarem e entenderem meu projeto. Aos meus colegas que foram meus companheiros no período que passei em Goiânia, em especial ao pessoal da turma de Estrutura de Dados e Análise de Algoritmos. E para finalizar agradeço à alguns amigos da época de faculdade que mesmo distantes aguentaram todas minhas crises e desesperos.

À todos aqueles que não mencionei os nomes, saibam que eu lhes agradeço de coração. Meu muito obrigada a todos que me apoiaram de alguma forma e contribuíram para o fim dessa etapa na minha vida.

A solução de um problema é apenas o início do próximo.
Mark Manson,
*A Sutil Arte de Ligar o F*da-se*.

Resumo

Moraes, Larissa V.. **Detecção de Depressão pela Fala Empregando Rede Neurais Profundas**. Goiânia, 2020. 63p. Dissertação de Mestrado. Instituto de Informática, Universidade Federal de Goiás.

Depressão é um transtorno mental que representa um importante problema para a saúde pública, com aumento de 20% no número de casos na última década. A apresentação dos sintomas depressivos é variada, causando isolamento e prejuízo no trabalho, estudos, sono e alimentação. O diagnóstico precoce continua sendo um dos principais desafios. A literatura do problema apresenta uma prevalência de propostas que utilizam dados de imagem e vídeo, entretanto, avanços recentes de métodos de aprendizado de máquina possibilitam a análise da fala e/ou de textos. Este trabalho propõe o uso de redes neurais profundas para detecção de depressão, a partir da análise da fala do paciente, gravada durante uma entrevista clínica. Para tal, realizou-se o pré-processamento dos áudios, gerando, assim, os espectrogramas, espectrogramas cepstrais de frequência mel e os coeficientes cepstrais de frequência mel. Em seguida, estas medidas foram usadas no treinamento e testes das arquiteturas aqui desenvolvidas. Diferentes combinações de hiper parâmetros de rede e dimensões dos espectrogramas foram analisadas. Os resultados obtidos demonstram menores valores da raiz do erro quadrático médio para aplicação dos coeficientes cepstrais (5,07), em comparação com a literatura (6,50). Apesar de ainda apresentar limitações quanto a um possível uso comercial, foi possível evoluir o estado da arte do problema.

Palavras–chave

Reconhecimento de Emoção de Fala, Reconhecimento de Fala, Rede Neural, Classificação de Depressão.

Abstract

Moraes, Larissa V.. **Detecting Speech Depression Using Deep Neural Networks**. Goiânia, 2020. 63p. MSc. Dissertation. Instituto de Informática, Universidade Federal de Goiás.

Depression is a mental disorder that represents a major public health problem, with a 20% increase in the number of cases in the last decade. The presentation of depressive symptoms is not padronized, causing isolation and impairment in work, studies, sleep and eating. Early diagnosis remains one of the main challenges. Recent advances in machine learning methods make it possible to analyze speech, text, and facial expressions for early diagnosis and detection. This paper proposes the use of deep neural networks to detect depression, based on the patient's speech analysis, recorded during a clinical interview. For this, the pre-processing of the audios was performed, thus generating the spectrograms, mel-frequency cepstral spectrograms and the mel-frequency cepstral coefficients. These measurements were then used in the training and testing of the architectures developed here. Different combinations of network hyperparameters and spectrogram dimensions were analyzed. The results show lower root mean square error values for the application of cepstral coefficients (5.07), compared to the literature (6.50). Therefore, the potential of this method to further assist in detecting depression is envisaged. Future studies are needed to improve and validate this method applied to a sample of national data.

Keywords

Speech Emotion Recognition, Speech Recognition, Neural Network, Depression Classification.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

A depressão é um transtorno mental que afeta cerca de 300 milhões de pessoas em todo o mundo, segundo a Organização Mundial de Saúde (OMS) [61]. As pessoas afetadas por essa doença acabam perdendo o rendimento no trabalho e/ou estudos além da promoção de limitações sociais. Nos piores casos, a depressão pode levar ao suicídio com cerca de 800 mil casos no ano, sendo uma das maiores causas de morte de jovens entre 15 e 29 anos de idade [61].

O diagnóstico da depressão é clínico, feito por médicos especialista. Para obter o diagnóstico é necessário a coleta da história do paciente e a realização de um exame do estado mental [17].

Pessoas depressivas são capazes de transparecer a doença não somente pelas expresariação da fala. Como exemplo, pode-se citar a alteração para um fala mais lenta, monótona, com algumas pausas, ora curtas, ora longas [43]. Essas ocorrências demonstram que a detecção da depressão pode ser realizada a partir da fala. Segundo Hoffman, Gonze e Mendlewicz, [40] o tempo de pausa da fala (*speech pause time, SPT*) está correlacionado com o tempo de reação de pacientes deprimidos e controles, onde os depressivos apresentam um aumento de SPT.

A computação afetiva busca reconhecer emoções humanas ou fazer com que as máquinas expressem emoções [13]. Estudos nessa área mostraram que a fala é a característica que mais exprime atributos notáveis de pessoas deprimidas, tais como: pico em voz baixa, lenta, hesitante, monótona, às vezes gagueira e sussurrante [33, 48]. O uso de dados de som e imagem para tarefas relacionadas a reconhecimento de emoções é explorado no desafio de reconhecimento de emoções audiovisuais (*Audio-Visual Emotion recognition Challenge* – AVEC). Trabalhos apresentados no AVEC mostraram resultados promissores na área de detecção de depressão por meio de ferramentas computacionais, tais como redes neurais.

O estudo de Jan et al. [43] é uma das maiores referências do tema na literatura. Os autores utilizaram algoritmos de aprendizado de máquina clássicos para extrair sinais de vídeo e áudio para representar características da expressão facial e vocal em pacientes com depressão. Os autores utilizaram uma técnica de combinação de modelos com Histograma

do Histórico de Movimentos (MHH), mínimos quadrados parciais e regressão linear. Os resultados obtidos para áudios foram 7,34 de MAE (*Mean Absolute Error*) e 9,09 de RMSE (*Root Mean Square Error*).

Posteriormente, um avanço significativo foi obtido pelos participantes do AVEC 2017, como pelo Yang et al. [82], quando comparados aos resultados dos estudos de Jan et al. [43].

Yang et al. [82] propôs uma estrutura de classificação de depressão multimodal audiovisual composta pelos modelos *Deep Neural Network* (Rede Neural Profunda - DNN) e *Deep Convolutional Neural Network* (Rede Neural Convolucional Profunda - DCNN). Foi adotada uma estratégia de fusão de decisão para melhorar a precisão da estimativa da pontuação do *Patient Health Questionnaire - 8* (PHQ-8), elaborado pela *American Psychological Association*, sendo este um teste utilizado por terapeutas como auxílio no diagnóstico da depressão. Os resultados parciais foram separados por gêneros, onde o masculino obteve um erro absoluto médio (*Mean Absolute Error* - MAE) de 5,107 e um erro quadrárico médio da raiz (*Root Mean Square Error* - RMSE) de 5,590, já o feminino, 4,597 e 5,669 de MAE e RMSE, respectivamente.

O estudo realizado por Ma et al. [54], chamado de DepAudioNet, consiste de um novo modelo de rede que faz uso de uma combinação em série de uma *Convolutional Neural Network* (Rede Neural Convolucional - CNN) com *Long short-term memory* (Unidade de Memória de Curto Prazo - LSTM). O banco de dados utilizado é o DAIC-WOZ, o mesmo escolhido para a realização deste trabalho. Foi realizado um pré-processamento para remover as pausas de longa duração no decorrer dos áudios. Foi feito o uso do Mel-scale para representar o sinal vocal, gerando os espectrogramas que são usados como entrada para a rede. O melhor desempenho alcançado pelo DepAudioNet proposto é de 52% de acerto para depressão e 70% para ausência.

A abordagem proposta para os participantes do AVEC 2017 foi multimodal, onde utiliza-se tipos de dados diferentes, sendo eles vídeos, áudios e textos para realização da classificação de depressão. Entretanto, existe uma carência de estudos na área de classificação de depressão por meio de áudio, quando comparado com abordagens que utilizam imagens e textos. Nos trabalhos de detecção de depressão com abordagem multimodal, o áudio geralmente possui uma acurácia relativamente pior como pode ser visto nos estudos de Yang et al. [83].

Os trabalhos que focam em um único tipo específico, utilizam as imagens referentes a expressões faciais ou conteúdos de textos extraídos a partir da fala. Nesse contexto, a hipótese de trabalho é de que avanços podem ser obtidos no problema de detecção de depressão pela fala, principalmente em abordagens que utilizam redes neurais profundas. Nos últimos anos, diversos *benckmarks* que envolvem reconhecimento de padrões em falas, tiveram relevante aprimoramento de resultados [58, 65, 80].

1.1 Objetivos

Pretende-se com este trabalho realizar o estudo da detecção da presença e ausência de depressão utilizando a fala como principal fonte de dados. Coneguindo com isso mostrar que os humanos demonstram características deprimidas por meio da fala. Esta abordagem apresenta vantagens por ser acessível, de fácil coleta e não invasiva. Por fim, quando comparada as expressões faciais, a fala possui aspectos de difícil disfarce.

1.2 Organização do documento

No decorrer deste trabalho há uma divisão em mais seis capítulos, o seguinte a este dará uma noção geral sobre a depressão. O capitulo três falará sobre o sinal de áudio, dando foco a representação da fala. Já o capítulo quatro, trata-se das técnicas de redes neurais convolucionais aqui aplicadas. Em seguida têm-se o capítulo de métodos e materiais, onde serão especificados os dados aqui utilizados e as técnicas empregadas para a realização deste trabalho. Após especificar as técnicas, no capítulo seguinte é mostrado os resultados e a discussão seguidos pelo capítulo de conclusão. Por fim, as referências bibliográficas que aqui foram utilizadas.

CAPÍTULO 2

Conceitos Básicos sobre Depressão

A depressão é um transtorno afetivo ou do humor envolvendo funções orgânicas, de humor e de pensamentos. É caracterizado, principalmente, por melancolia, ansiedade, baixa autoestima, culpa, fadiga, dificuldade de concentração, distúrbios do sono e do apetite e outros sintomas que podem durar semanas, meses ou anos [63].

Este transtorno, além de ser crônico é também recorrente, onde cerca de 80% dos casos que recebem tratamento para o episódio depressivo terão um segundo episódio ao longo da vida [28].

2.1 Diagnóstico

Ao passar por médicos gerais e serviços de cuidados primários, cerca de 30% a 50% dos casos não são diagnosticados e apenas um terço são tratados [78]. Essa dificuldade ao diagnóstico pode ter relação tanto ao médico quanto ao paciente. O paciente principalmente pelo preconceito e o médico pela falta de treinamento e tempo [28]. Além do fato de que esses médicos não são psiquiatras, os pacientes, em sua maioria, apresentam sintomas somáticos e a minoria, psicológicos, na proporção de dois (sintomas somáticos) para um (sintomas psicológicos) [90]. Os sintomas somáticos incluem a fadiga, queixas gástricas e dores no geral, como cefaleia, epigastralgia, dor lombar e outras de localização imprecisa. Já os sintomas psicológicos envolvem a tristeza, perda de interesse, dificuldades de concentração, problemas de apetite, desânimo e problemas de sono.

Na psiquiatria a depressão é diagnosticada a partir da presença de sintomas que possuem uma certa frequência, duração e intensidade que são descritos por manuais psiquiátricos mundialmente reconhecidos, como por exemplo o Manual diagnóstico e estatístico de transtornos mentais (DSM-IV, 1995), Organização Mundial de Saúde (CID-10, 1996-1997) [69].

Alguns testes, além dos manuais de diagnósticos, também são utilizados como auxílio ao diagnóstico, como o PHQ-8 e o PHQ-9 (*Patient Health Questionnaire* – 8 e 9).

2.2 Questionário da Saúde do Paciente – Módulo Depressão

O Questionário da Saúde do Paciente, são questionários utilizados principalmente por psiquiatras. Aqui serão mencionados o PHQ-8 e o PHQ-9 pelo fato da base de dados aqui utilizada fazer o uso do PHQ-8 como complemento ao diagnóstico.

O PHQ-9 contém nove questões, sendo de rápida aplicação e seria uma vantagem em estudos epidemiológicos, em comparação a outros como o Inventário de Depressão de Beck (*Beck Depression Inventory* – BDI) [73]. O PHQ-9 tem como objetivo identificar corretamente indivíduos em risco de apresentar depressão e a gravidade do transtorno, tendo o diagnóstico final confirmado por um psiquiatra de acordo com suas opiniões e conhecimentos. O PHQ-8 possui oito itens e também é estabelecido como uma medida válida de diagnóstico auxílio na identificação da gravidade da depressão.

O PHQ-8 foi disponibilizado pelos departamentos estaduais de saúde no Inquérito sobre Vigilância de Fatores Comportamentais de 2006 (BRFSS) para poder avaliar a prevalência e impacto da depressão nos Estados Unidos. Há evidências de que o escore do PHQ-8 maior ou igual a 10 representa depressão clinicamente significativa, sendo mais conveniente usá-lo em relação ao PHQ-9 [49]. Ele também foi utilizado pelos psiquiatras para auxílio ao diagnóstico nos dados utilizados neste trabalho.

O PHQ-8 foi criado se baseando no Manual Diagnóstico e Estatístico de Transtornos Mentais (DSM - 5). Nele se pergunta o número de dias nas duas últimas semanas em que o indivíduo teve um sintoma depressivo específico. De 0 a 1 dia foi considerado como "nada", 2 a 6 dias, como "vários dias", 7 a 11 dias, como "mais da metade dos dias"e 12 a 14 dias, como "quase todos os dias". Para cada categoria são atribuídos pontos de zero à três. As pontuações de cada item são somadas produzindo uma pontuação total entre 0 e 24 pontos. Uma pontuação total de 0 a 4 não representa sintomas depressivos significativos. Uma pontuação total de 5 a 9 representa sintomas depressivos leves; 10 a 14, moderado; 15 a 19, moderadamente grave; e 20 a 24, grave [49].

As perguntas presentes neste questionário estão listadas abaixo, onde todas começam com "Quantas vezes nas últimas 2 semanas você se sentiu incomodado por":

1. ter pouco interesse ou prazer em fazer as coisas?
2. se sentir para baixo, deprimido ou sem esperanças?
3. ter problemas para adormecer ou por dormir demais?
4. se sentir cansado ou ter pouca energia?
5. ter pouco apetite ou por comer demais?
6. se sentindo mal consigo mesmo, ou falhando, ou decepcionando a si ou à sua família?
7. ter problemas para se concentrar em coisas, como ler o jornal ou assistir televisão?

8. se mover ou falar tão devagar que outras pessoas poderiam ter notado. Ou o contrário - ser tão inquieto ou inquieta que você se move muito mais do que o habitual?

A partir deste questionário o especialista, psiquiatra, consegue atribuir pontuações de acordo com cada resposta dada. A pontuação final do questionário se dá pela somatória da pontuação referente a cada resposta, quanto maior a frequência de dias dados nas respostas, maiores as chances de o especialista concluir a possibilidade da presença de depressão.

2.3 Depressão na Fala

Alguns trabalhos como o de Brian Helfer et al. [38], mostram que a depressão pode ser expressa pela fala de uma pessoa. As alterações neurofisiológicas associadas com a depressão afetam a coordenação motora e podem interromper a precisão articulatória na fala.

De acordo com Stefan Scherer [16], pessoas que sofrem com depressão possuem uma voz monótona, plana, sem expressão. Além disso essas pessoas arrastam a fala, gaguejam, fazem pausas mais longas, não se esforçam direito para falar como fazem as pessoas que não sofrem de depressão.

Murray Alpert et al. [5] realizaram experimentos em relação a fala automática, como leitura, até liberdade de expressão. Foi notado que os melhores resultados obtidos por Murray Alpert et al. foram em relação a liberdade de expressão, a mesma requer atividade cognitiva (busca de palavras e planejamento do discurso) e motora da fala.

Murray Alpert et al. [5] também concluiram que os pacientes deprimidos apresentam menos prosódia (ênfase e inflexão) em comparação aos indivíduos não deprimidos. E a fluência da fala (pausas na fala) reflete o estado depressivo, enquanto a prosódica reflete o caráter deprimido.

Para melhor entender sobre a fala e suas característica será discutido mais sobre sinal de áudio, voz e fala no capítulo seguinte.

CAPÍTULO 3

Sinal de Áudio

Sinal de áudio é a representação do som no intervalo de tempo e frequência audíveis por seres humanos (20 Hz à 20.000 Hz) [39]. O sinal de áudio é muito abrangente, incluindo não só a fala e a música como todos os outros tipos de sons [55].

Ao falar, os humanos produzem três tipos de sons, a voz ou prosódia, sons fricativos ou trato vocal e fonte de voz ou forma de onda glótica. A prosódia acontece quando ocorre a vibração das cordas vocais que produzem pulsos de ar semi-periodicos nas cavidades vocais. Já o som fricativo se origina da turbulência do ar em passagens estreitas, como lábios e dentes [75]. Já a fonte de voz, diz respeito a velocidade do fluxo de ar através da glote [56].

Uma fala deprimida é caracterizada como uma fala monótona e sem expressões significativas [56]. Estas características podem ser analisadas através da prosódia. Segundo Roddy Cowie e Ellen Douglas Cowie em [15], o domínio prosódico aumentado são relacionados a áreas emotivas. Neste domínio podem ser vistos vários desvios, como da emoção, prejuízos centrais e sensoriais (esquizofrenia e surdez).

A prosódia é o estudo do ritmo e entonação, ou seja, acentuação vocabular e qualidade, relacionada ao nível de legibilidade e naturalidade da fala [16]. As representações das ondas sonoras audíveis e inaudíveis pelo ser humano são mostradas nas representações espectrais.

A partir dessas características da fala citadas, é possível extrair várias outras, dentre elas estão a frequência fundamental (*pitch*), a energia e propriedades relacionadas à duração da prosódia e pausas, a distribuição espectral (espectrogramas), *Mel-Frequency Cepstrum* (MFC), *Mel Frequency Cepstral Coefficients* (MFCC) ou *Log Frequency Power Coefficients* (LFPC), a relação harmônicos/ruído (*Harmonic to Noise Ratio* ou HNR), Nitidez PSY, *Shimmer* e *Zero Crossing Rate* (Taxa de Passagem Zero).

Estudos como os de Zigelboim e Shallom [89], para o reconhecimento da fala, e o trabalho de Bitouk et al. [8], para a classificação de emoções, obtiveram bons resultados utilizando o MFCC. Além destes, o MFCC também foi utilizado em trabalhos para detecção de depressão, como no estudo de Alhanai et al. [4]. O MFCC possui resultados promissores para o reconhecimento de estresse e, segundo [37], são necessárias

melhorias no reconhecimento de emoções. Portanto o MFCC foi uma das características selecionadas para fazer parte deste estudo, com a finalidade de melhor explorá-lo.

Alguns estudos também usaram o espectrograma Mel, tendo bons resultados juntamente com características faciais [88] e com sequências de fonemas [84] para reconhecimentos de emoções. Sendo esta também uma característica selecionada para ser explorada neste estudo.

E como uma terceira abordagem, foi escolhido a representação do áudio no espectro do tempo e da frequência. Um espectrograma ou distribuição espectral na sua forma mais simples, conseguido a partir da Transformada de Fourier, como será especificado mais a seguir.

3.1 Distribuição Espectral da Fala

A análise espectral da fala é o estudo das frequências que compõem a mesma [27]. Essa análise é realizada com o intuito de se obter uma representação visual do sinal da fala. Esta é uma representação da onda sonora no espectro de frequências que variam com o tempo e também é chamada de espectrograma.

O espectro sonoro é dividido em três zonas, faixa de áudio, infrassons e ultrassons [7]. Faixa de áudio, também chamado de sons audíveis, são sons em um intervalo de frequência de 20 Hz e 20.000 Hz que são audíveis por humanos. Sons abaixo de 20 Hz são chamados de infrassons, são inaudíveis por humanos. Assim como os infrassons, os ultrassons também são inaudíveis a ouvidos humanos, são sons com frequência acima de 20.000 Hz.

Os espectrogramas dos sons podem ser utilizados em várias áreas, dentre elas na música [14], em sonares [29] e reconhecimento da fala. Entrando nesse campo de fala observa-se que os espectrogramas também podem ser usados para identificar pessoas [45], identificar emoções [36] e até mesmo para avaliar um sistema que gera a fala a partir do texto [44]. Neste último o espectrograma gerado sintetiza a fala, o mesmo deve ser semelhante com o espectrograma gerado a partir da fala humana. Com isso pode-se ir ajustando o sistema de forma com que o espectrograma gerado seja mais próximo possível do espectrograma obtido a partir da fala humana.

3.1.1 Cálculo do espectrograma utilizando Transformada Rápida de Fourier

Para implementar uma análise espectral de fala é consideravelmente mais eficiente realizar a análise usando o algoritmo de Transformada Rápida de Fourier (FFT - *Fast Fourier Transform*) do que implementar um banco de filtros [62]. Onde um banco de

filtros é um arranjo de filtros passa-faixa, estes permitem a passagem de frequências de um intervalo determinado e rejeitam as que estão fora desse intervalo. O que resultaria na decomposição dos áudios em diversas componentes [21].

Para a realização deste trabalho, os espectrogramas foram gerados utilizando a FFT tendo como base para essa obtenção a biblioteca Numpy [60] do Python.

Segundo descrito em [85] a Transformada de Fourier decompõe uma função em componentes senoidais, porém a fala não se enquadra como onda senoidal, já que uma onda senoidal é uma curva matemática que descreve uma oscilação repetitiva suave, sendo esta uma onda contínua [3]. Como representação de onda senoidal se tem a onda de seno e cosseno mostradas na Figura 3.1, onde no eixo x temos a representação dos ângulos em radiano e em y seus respectivos senos (curva vermelha constante) e cossenos (curva azul pontilhada).

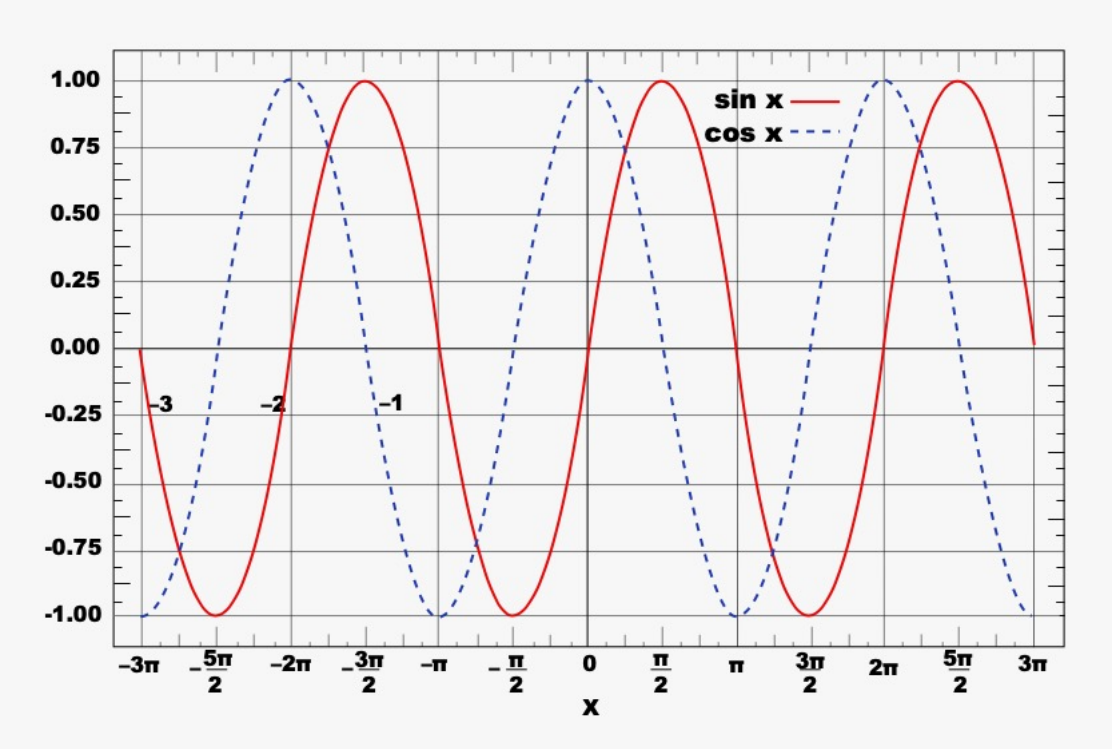


Figura 3.1: *Representação de uma onda sinusoidal*[1].

A fala humana produz sons irregulares, este é um exemplo de onda não senoidal. Estas ondas são consideradas como um conjunto de ondas senoidais de diferentes períodos e frequências [31]. A forma que utilizamos para encontrar esse conjunto de ondas senoidais é aplicando a Transformada Discreta de Fourier (*Discrete Fourier Transform* – DFT). Ela converte um sinal no domínio do tempo com N pontos, tal que $N \in \mathbb{N}$, em

[1] Imagem original em: https://tinyurl.com/wh3u98r.

dois sinais no domínio da frequência com $(\frac{N}{2}+1)$ pontos. Este sinal de saída contêm as amplitudes dos senos e cossenos [10]. A DFT é dada por:

$$F[k] = \sum_{n=0}^{N-1} f[n] e^{-ikn(2\frac{\pi}{N})} , \tag{3-1}$$

onde $0 \leq k \leq (N-1)$; $f[n]$ é a sequência discreta do domínio do tempo que descreve os valores de $f(t)$; e N é o número de amostra da sequência [10].

Porém o número de operações para realizar o cálculo da DFT é de n^2, logo com um custo computacional $O(n^2)$. Já uma Transformada Rápida de Fourier (*Fast Fourier Transform* – FFT) realiza esse mesmo cálculo com $n \log n$ operações, tendo um custo de $O(n \log n)$ [68]. Portanto a FFT é um algoritmo mais eficiente de realizar o cálculo de DFT e o mais usado computacionalmente.

Na Figura 3.2 é mostrado o espectrograma resultante da aplicação da FFT, onde se tem a representação da frequência no eixo y em relação ao tempo no eixo x. A barra de cor lateral faz referência a amplitude da onda sonora medida em decibel (dB).

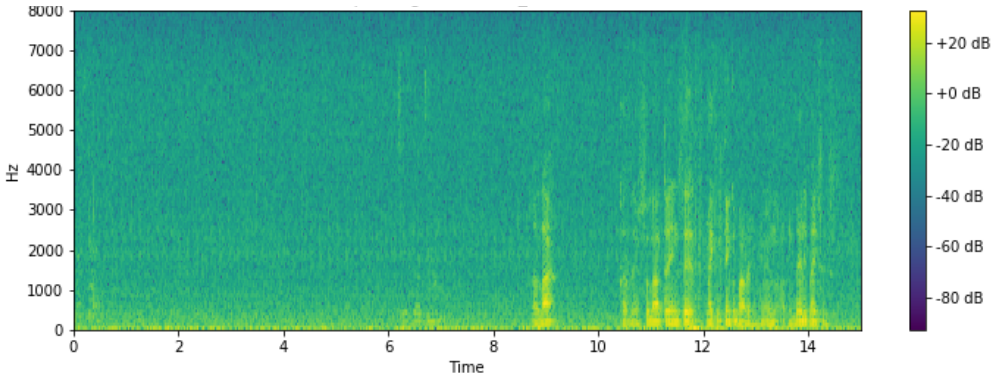


Figura 3.2: *Espectrograma referente ao fragmento 19 do paciente com identificação 302.*

3.2 *Mel-Frequency Cepstrum*

O espectrograma de frequência mel (*Mel-Frequency Cepstrum*), também chamado de *mel-power*, permite verificar frequências que não são observáveis no espectrograma anterior (Figura 3.2). Frequências estas que não são perceptíveis a ouvidos humanos, já que para calcular os bancos de filtros para conseguir esse primeiro espectrograma foram motivados pela natureza do sinal da fala e pela percepção humana dos mesmos [26]. Para se obter um espectrograma *mel-power* é necessário primeiro aplicar um filtro de pré-ênfase. Ele é utilizado para amplificar as altas frequências com a finalidade de conservar informações importantes, obtendo amplitudes mais homogêneas [12].

Após isso é realizado a divisão em pequenos fragmentos do sinal original para não se perder contornos de frequências ao realizar a transformada de Fourier. Para essa divisão é usada uma função de janela, que são utilizadas como forma de aumentar as características espectrais do sinal, como por exemplo a função de *Hamming Window* [74]. Finalmente se aplica a FFT em cada um dos *frames* obtendo um espectro.

Para se obter o espectrograma de frequência mel é aplicado filtros da escala mel [77]. Estes filtros são formados baseados em uma escala psicoacústica de sensibilidade do ouvido. Onde psicoacústica é o estudo da percepção sonora e da forma como os seres humanos percebem os sons [19].

Por fim é aplicado o cálculo logarítmico da energia de saída de cada filtro para então se obter o cepstro [66]. O resultado deste processo pode ser visto na Figura 3.3, onde se relaciona o tempo em x, com a frequência em y, tendo as amplitudes em dBs representadas pelas cores conforme a barra de cores a direita da figura.

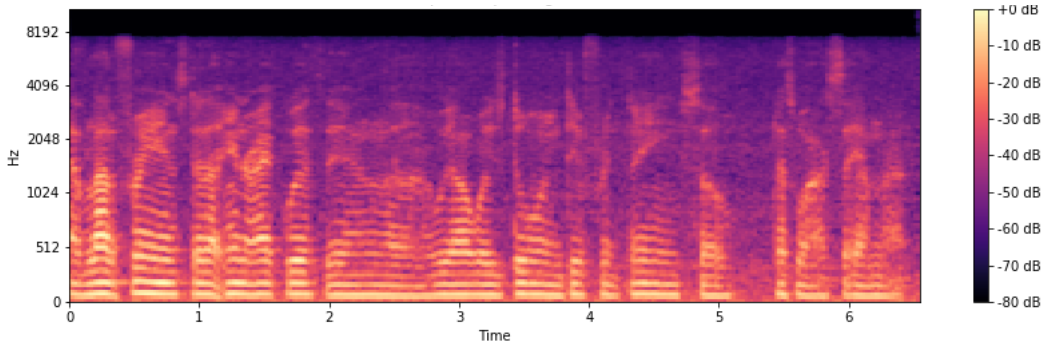


Figura 3.3: *Espectrograma mel-power.*

3.3 *Mel-Frequency Cepstrum Coefficients*

Os MFCC expressam algumas características da fala. Estas são calculadas a partir da sua frequência levando em conta a percepção humana [66], para isso, baseia-se no *pitch* [26]. O *pitch* (altura) relaciona o quão alto ou baixo é o som para nossos ouvidos. Os MFCC são usados principalmente em sistemas de reconhecimento de fala [53] e para sistemas de identificação de gêneros musicais [64].

Para a obtenção do coeficiente cepstral de frequência mel (*Mel-Frequency Cepstrum Coefficients* – MFCC) é necessário primeiro a obtenção do espectrograma *mel-power*. Tendo este é necessário a aplicação da Transformada de Cosseno Discreta (DCT), resultando no espectrograma dos MFCC, Figura 3.4, onde se relaciona o tempo em x, com o próprio coeficiente em y, tendo seus valores representados pelas cores conforme a barra de cores a direita da figura.

Após normalização do MFCC em Figura 3.4, ressalta-se alguns pontos, como pode ser visto na Figura 3.5, onde, também, relaciona-se ao tempo em x, o coeficiente em y, seus valores normalizados representados pelas cores.

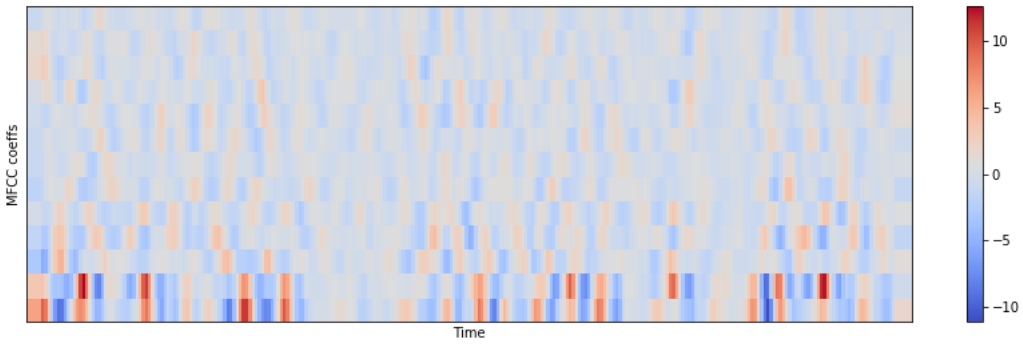


Figura 3.4: *Espectrograma mostrando o MFCC.*

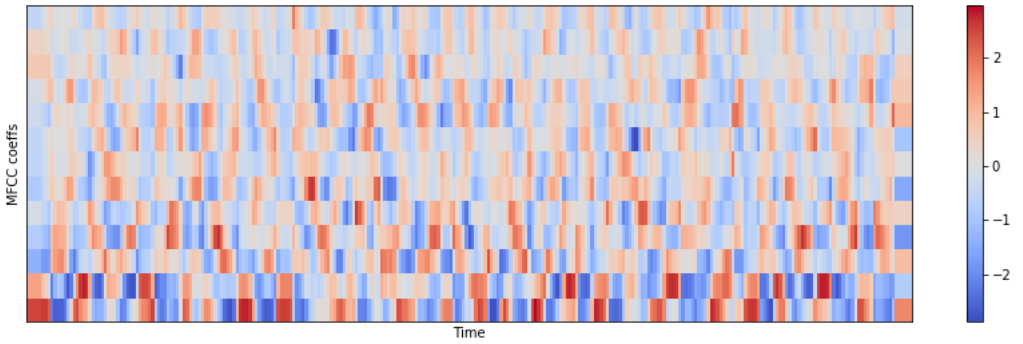


Figura 3.5: *Espectrograma mostrando o MFCC normalizado.*

CAPÍTULO 4

Redes Neurais Convolucionais

A rede neural convolucional (CNN – Convolutional Neural Networks) foi inspirada pelos trabalhos de Hubel e Wiese [41]. Posteriormente o conceito foi computacionalmente desenvolvido por Yann LeCun e Yoshua Bengio em 1995[51].

Em 1998 foi lançada a LeNet, considerada a primeira CNN [52]. Esta rede tinha como problema motivador o reconhecimento de dígitos escritos a mão. A base de dados utilizada para o treino e teste é chamada de MNIST [50], contendo 60 mil imagens para treinamento e 10 mil para teste, todas em escala de cinza e de dimensões 28x28. A LeNet obteve um melhor resultado que as técnicas empregadas na época, como por exemplo o SVM (*Support Vector Machine*) [9].

As CNNs são redes que seguem o padrão *feed-forward*, onde todas as camadas se conectam a camada seguinte, sem haver um caminho inverso, seguindo o caminho da entrada para a saída da rede [71]. Para a realização dos ajustes dos pesos no decorrer do treinamento segue-se a lógica do algoritmo de *back-propagation*. O treinamento por backpropagation ocorre em dois passos. No primeiro é dado uma amostra de entrada para que se obtenha uma saída. Essa saída é comparada com a desejada e é efetuado o cálculo do erro. No segundo passo o erro é propagado da saída para entrada e os pesos e o limiar vão sendo atualizados utilizando a regra delta generalizada [72], esta faz o uso do gradiente descendente. Isso implica em uma gradativa diminuição da somatória dos erros [18].

Redes convolucionais são capazes de extrair características, detectando padrões, além do potencial de classificação [86]. Por essas capacidades de detecção de padrões, as CNNs já tornaram-se "estado da arte"em tarefas de reconhecimento de imagem, segmentação, detecção e recuperação [47].

4.1 ResNet

ResNet ou rede residual (*Residual Networks*) também é uma CNN que foi lançada no desafio ImageNet (desafio de classificação de imagens com mais de mil classes) em 2015 [20] e acabou conquistando o primeiro lugar. Esta arquitetura provou serem mais

fáceis de otimizarem e podem obter precisão com uma profundidade consideravelmente maior do que as arquiteturas anteriores a elas, como a VGG [35]. Esta possui cerca de oito vezes mais camadas do que a rede VGG [2] e obteve melhores desempenhos e resutados.

O que tornou possível esse melhor desempenho foi o desenvolvimento do bloco residual, Figura 4.1.

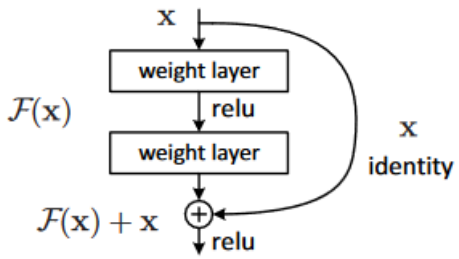


Figura 4.1: *Representação do bloco residual*[1].

O bloco residual, Figura 4.1, tem como objetivo denotar o mapeamento subjacente desejado como $H(x)$, deixando as camadas não lineares empilhadas ajustarem outro mapeamento de $F(x) := H(x) - x$. Onde o mapeamento original é reformulado em $F(x) + x$. Com a ideia de ser mais fácil a otimização do mapeamento residual do que otimização do mapeamento original e não referenciado [35]. A adição da identidade x ao mapeamento $F(x)$ não adiciona parâmetros extras e nem aumenta a complexidade computacional da arquitetura.

Portanto, o resultado inicial apresentado é uma ResNet com 34 camadas, possuindo 16 blocos residuais empilhados, como mostrado na Figura 4.2. Quando se trata de complexidade computacional e número de parâmetros, a ResNet com 34 camadas possui a quantidade equivalente a essa mesma arquitetura de 34 camadas simples, ou não residual.

4.2 UNet

A UNet também é uma CNN que foi lançada em 2016, e, diferente das redes residuais, são mais utilizadas para segmentação e extração de características. Essa arquitetura serviu como ideia base para a criação das arquiteturas aqui desenvolvidas, tendo como principal inspiração as camadas utilizadas por elas para realizar o redimensionamento das imagens para o tamanho da entrada da rede.

As camadas desta rede podem ser agrupadas em duas categorias: o caminho contrativo, também chamado de *encoder*, e o caminho expansivo, chamado de *decoder* [25]. Essa arquitetura tem como propósito garantir que as dimensões de saída da rede

[1] Imagem original em [35]
[2] Imagem original em [35]

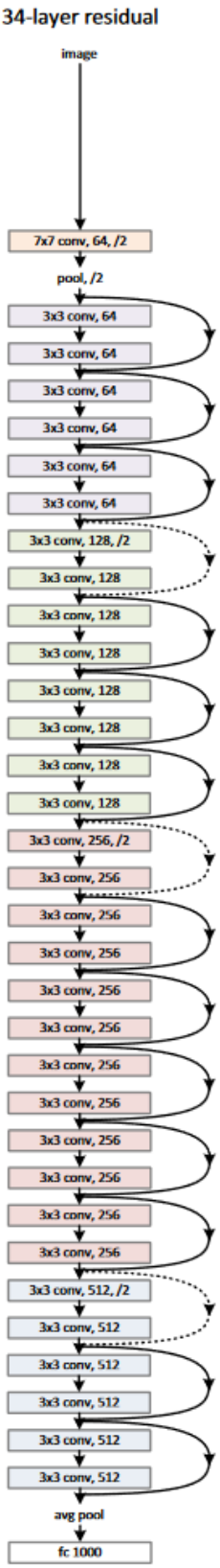


Figura 4.2: *Representação da arquitetura de uma rede residual de 34 camadas, onde cada bloco representa uma camada da rede e as setas mostram o fluxo das saídas*[2].

tenham as mesmas dimensões de entrada, como é mostrado na Figura 4.3. Nesta é possível notar o formato de "U"da arquitetura, onde a "descida"representa o caminho contrativo e a "subida"representa o caminho expansivo.

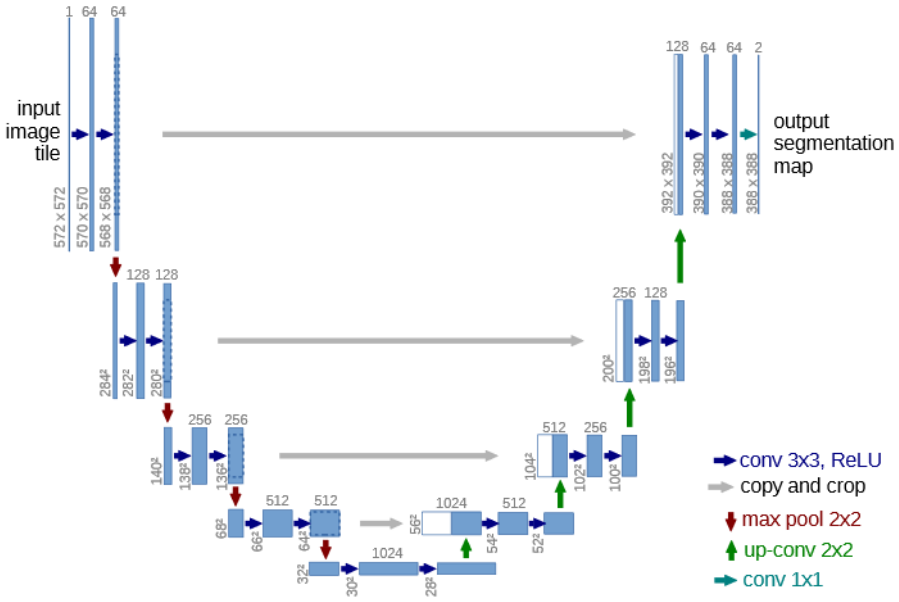


Figura 4.3: *Representação da arquitetura da UNet, onde cada bloco em azul representa um mapa de característica como saída de uma camada e entrada para a próxima. As setas representam as operações realizadas por cada camada da rede* [3].

As principais componentes das CNN são as camadas de entrada, de convolução, de ReLU (*Rectified Linear Units* - Unidades Lineares Retificadas), a de *pool* e um classificador, normalmente se utiliza camadas *fully connected* com a função de *softmax*. Neste trabalho também foram utilizadas camadas de *upsamplig*, convolução transposta, *batch normalization* e *dropout*.

A camada de entrada é onde a rede recebe as imagens, tanto para o treinamento quanto para testes. As imagens são lidas como vetores com três dimensões (h,w,d). Sendo 'h' o número de pixels de altura (*height*), 'w' de largura (*width*) e 'd' a dimensão (*dimension*), ou seja, o número de canal como o RGB (*Red, Green and Blue* - sistemas de cores vermelho, verde e azul).

[3] Imagem original em [70]

4.3 Camada de Convolução

A convolução é uma operação de filtragem realizada nos dados. A função de convolução é representada pelo símbolo "⊛"como na equação [30]:

$$S(t) = (X \circledast W)(t), \tag{4-1}$$

onde $W(t)$ é o *kernel*, também chamado de filtro, que é uma matriz de dimensões definidas pelo usuário que irá convoluir com as imagem de entrada X resultando em uma imagem $S(t)$.

Quando se considera t apenas como valores inteiros obtemos a equação de convolução discreta:

$$(X \circledast W)(t) = \sum_{a=-\infty}^{\infty} X(a)W(t-a). \tag{4-2}$$

Porém quando se trata de aplicações de *machine learning*, como redes neurais convolucionais, normalmente os dados de entrada são matrizes multidimensionais e o *kernel* também é multidimensional. Quando se trata de imagens bidimensionais (I) o *kernel* também é bidimensional (K), para então se realizar a convolução nos dois eixos ao mesmo tempo. Esta é representada pela equação:

$$S(i,j) = (I \circledast K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n), \tag{4-3}$$

onde S representa a matriz de saída da convolução.

Na camada de convolução, a posição do *kernel* é passada por todas as posições da matriz de entrada, com a opção de aplicar *stride* e *padding*. *Padding* é uma operação com o propósito de manter a dimensionalidade da matriz de entrada após a operação convolucional, onde se adicionam linhas e colunas preenchidas com zeros na matriz resultante. Já o *stride* é uma operação que pode reduzir a dimensionalidade da matriz resultante em uma proporção de $\frac{1}{n}$, realizando a deslizamento do *kernel* por n pixels na matriz de entrada, sendo $n \in \mathbb{Z}_+^*$ [87]. Após a operação convolucional ser concluída é gerado um mapa de características que permite a rede reconhecer padrões [30].

4.4 Camada de Convolução Transposta

A camada de convolução transposta é o oposto da camada de convolução, ou seja, transforma algo que tem formato de uma saída de uma camada de convolução para algo que tenha o formato de entrada de uma camada de convolução [24]. Esta pode ser

usada como camada de decodificação de um autoencoder convolucional ou para projetar mapas de características para um espaço de maior dimensão.

Pode-se usar como equivalência para convolução transposta uma convolução direta com *padding* em zero (preenchimento em zero) na imagem de entrada. Porém esse método é menos eficiente do que realizar a transposta da convolução, por gerar um grande número de multiplicações por zero devido o *padding* [24].

Normalmente as convoluções transpostas são utilizadas aumentar a densidade de pixels de uma imagem de baixa densidade de pixels. Isso permite que a rede preencha os detalhes da imagem [59].

Na Figura 4.4 é apresentado na parte superior da figura a entrada dos dados na camada de convolução transposta (com *kernel* de dimensão 4x4 e *stride*, 2) e a parte inferior representa a saída. Quando os valores de dimensões do *kernel* e do *stride* não forem divisíveis, essa camada acaba por gerar alguns artefatos devido a sobreposição, tanto de baixa quanto de alta frequência. Ao mesmo tempo que a mesma camada pode auxiliar a remover outros artefatos.

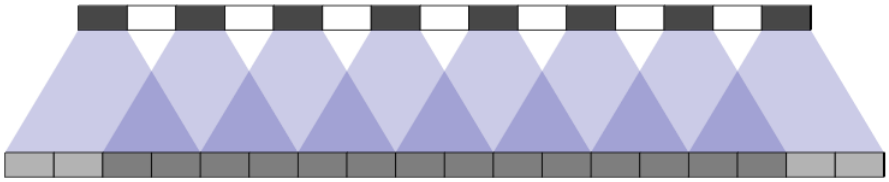


Figura 4.4: *Representação da convolução transposta com stride de 2 e núcleo de dimensões 4 o que faria com que a dimensão da saída fosse o dobro da entrada* [4].

4.5 Função ReLU

A função ReLU é usada como uma função de ativação e é estritamente positiva [57] sendo representado pela equação 4-4.

$$f(x) = max(0,x), \tag{4-4}$$

portanto $f(x)$ é zero caso x seja menor que zero e x, caso x maior que zero. Isto faz com que seja produzido zero em metade de seu domínio. Logo as derivadas através de uma função ReLU permanecem grandes sempre que ela estiver ativa. A segunda derivada da

[4] Imagem retirada de [59]

operação de retificação é 0 em quase toda parte, e a derivada é 1 sempre que ativa. Ou seja, a direção do gradiente é muito mais útil para o aprendizado [30].

Existe outras funções de ativação, como a sigmoide (Equação 4-5) e a tangente hiperbólica (Equação 4-6). Porém no decorrer do treinamento das redes profundas, como a CNN, o gradiente acaba fazendo com que essas funções tenham uma tendência a zerar, devido suas derivadas tenderem a zero o que torna difícil a correção dos pesos nas primeiras camadas da rede no decorrer do *backpropagation*.

$$\sigma(x) = \frac{1}{1+e^{x}} \tag{4-5}$$

$$tanh(x) = 2\sigma(2x) - 1 \tag{4-6}$$

4.6 Camada de *Pool*

A camada de *pooling* recebe cada saída do mapa de características da camada convolucional e prepara um mapa de características de dimensões reduzidas [1]. Para isso pode-se utilizar algumas funções para tornar-se evidentes possíveis padrões nesses mapas de características.

A função mais utilizada é a máxima, também chamada de *max-pooling*, consiste em uma função que seleciona o maior valor de pixel para cada aplicação do filtro no mapa de entrada, dando origem ao novo mapa de características com esses valores. O *Average pooling*, ou *pooling* médio, é uma função que faz a média dos dos valores dos pixels. E por fim o *min-pooling*, ou *pooling* que seleciona o menor valor entre os pixels [30].

4.7 Camada de *Upsampling*

A camada de *upsampling* tem a finalidade de aumentar a dimensão da imagem, sendo ela considerada o oposto da camada de *pooling* [23]. A camada de *upsampling* é bastante utilizada em redes como a UNet e em algumas arquiteturas são usadas como alternativa, por ter uma menor probabilidade de gerar artefatos de sobreposição em relação à camada de convolução transposta.

Uma abordagem alternativa à camada de convolução transposta é o redimensionamento da imagem ultilizando a camada de *upsampling* seguida por uma camada convolucional. Essa abordagem têm funcionado bem para melhorar a resolução de imagens como no trabalho do Dong et al. [22].

Um dos principais filtros de *upsampling* utilizado é o de vizinhos mais próximos (*nearest-neighbor*). O filtro de vizinhos funciona aumentando a dimensão da imagem e

repetindo o valor do pixel nos pixels vizinhos a cada aplicação do filtro na imagem de entrada [34]. A Figura 4.5 representa o funcionamento do filtro de vizinhos de dimensões 2 x 2, onde na esquerda da figura se localiza a imagem de entrada e na direita a imagem resultante.

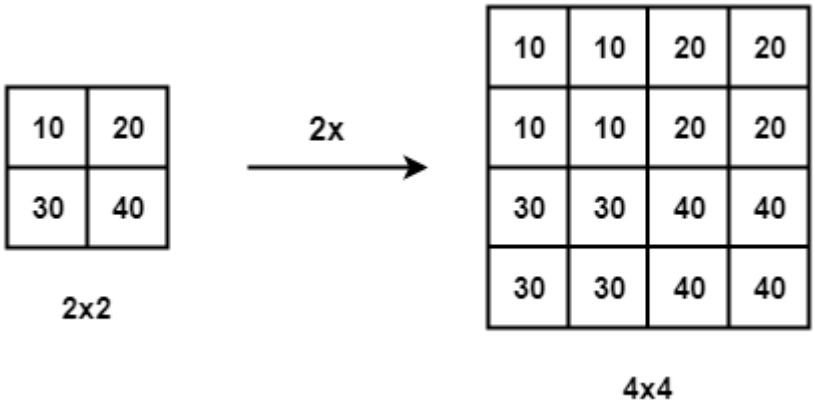


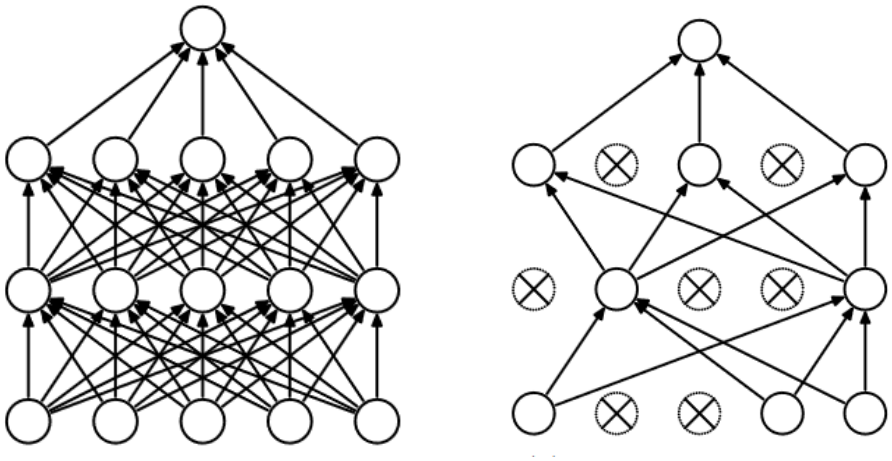
Figura 4.5: *Redimensionamento de 2 x 2 para 4 x 4 utilizando upsampling com filtro de vizinhos mais próximos* [5].

4.8 *Dropout*

O *dropout* é um método poderoso e sem custos para regularização de modelos além de facilitar o *emsemble* (concatenação) eficiente e aproximado de modelos [30]. O ruído é adicionado pela camada de *dropout* nas camadas ocultas, podendo ser visto como uma forma de destruição de conteúdo das entradas de forma adaptativa e inteligente. Esses ruídos adicionados tem como um dos objetivos retardar o *overfitting* (ocorrência em que o modelo decora os dados de treino e deixa de generalizar) [30].

O *dropout* tem como função o desligamento ou adormecimento de alguns neurônios da camada anterior, selecionados de forma aleatória de acordo com uma probalidade escolhida pelo usuário. Na Figura 4.6(a) é representado uma arquitetura de uma rede neural contendo duas camadas ocultas, já a Figura 4.6(b) representa essa mesma arquitetura após aplicação do *dropout* de 40% nas duas primeiras camadas e de 50% na penúltima camada, sendo possível notar que dois neurônios de cada camada foram desativados da Figura 4.6(a) para a Figura 4.6(b).

[5] Imagem original em https://tinyurl.com/ubz8o6t.
[6] Imagem original em [76].



(a) *Rede neural padrão.* (b) *Após aplicação do dropout.*

Figura 4.6: *À esquerda é representado uma rede neural padrão com duas camadas ocultas. À direita um exemplo de rede reduzida produzida pela aplicação do dropout na rede à esquerda* [6].

4.9 *Batch Nomalization*

Um dos grandes problemas para o treinamento de redes neurais profundas é o fato de que as entradas de cada camada oculta sofre com uma grande covariância durante o treinamento. Isso acaba exigindo taxas de aprendizado mais baixas e inicialização cuidadosa dos parâmetros [42].

Como uma forma de amenizar essa questão, foi proposto um método para normalizar essas entradas das camadas, sendo este chamado de *batch normalization* [42]. Este método, assim como o *dropout*, funciona como regularizador e, em alguns casos, elimina a necessidade de fazer o uso do *dropout*, além de permitir o uso de taxas de aprendizado mais altas.

Para realizar a normalização ($\hat{x_i}$) da saída de uma camada de ativação anterior (x_i), é realizado a subtração da média do *batch* (μ_β) e então esse valor é dividido pelo desvio padrão do *batch* (σ^2_β), representado na Equação 4-7.

$$\hat{x_i} = \frac{x_i - \mu_\beta}{\sqrt{\sigma^2_\beta + \varepsilon}} \tag{4-7}$$

4.10 Camada *Fully Connected*

A camada *fully-connected* tem como entrada a saída das camadas anteriores e tem como objetivo chegar a uma decisão de classificação [32]. Onde todos os neurônios da camada anterior são conectados a todos os seus neurônios, como uma *Multilayer*

Perceptron tradicional. Esta camada gera uma saída relacionando o número de neurônios presentes nela com o número de classes presentes no modelo desenvolvido, realizando a classificação. A função mais aplicada para problemas de classificação é a SoftMax [46].

Para realizar essa classificação, a saída das camadas anteriores é concatenada em um único vetor de valores, onde cada um desses valores representam uma probabilidade de que uma determinada característica pertença a uma das classes do problema.

A função softmax, Equação 4-8, também é um tipo de função sigmóide, mas é útil quando tentamos lidar com problemas de classificação. A função SoftMax permite interpretar os valores de saída como probabilidades, já que esta função normaliza as saídas para o intervalo [0, 1] com a soma resultando em 1 [11].

$$\sigma(y)_j = \frac{e^(y_j)}{\sum_{k=1} K e^(y_k)} \tag{4-8}$$

onde $j = 1, ..., K$, sendo j o número do caso testados; K, o númerototal de casos; e y_j, a probabilidade para o caso j.

CAPÍTULO 5

Materiais e Métodos

Neste capítulo é apresentado os dados e recursos utilizados neste trabalho. Além disso, são apresentados os métodos propostos. Também são apresentadas as métricas estatísticas utilizadas para avaliar os modelos gerados.

5.1 Materiais

5.1.1 Dados

Todas as gravações de áudios utilizadas neste estudo foram retiradas do banco de dados DAIC-WOZ [79]. Este foi compilado pelo Instituto de Tecnologias Criativas da *University of Southern California* (USC) e lançado como parte do Desafio e Workshop Emocional de Áudio/Visual de 2016 (AVEC 2016) [79].

Essa base de dados possui 189 áudios, com uma média de 16 minutos. Destes áudios, 139 foram selecionados para o trabalho, onde apenas 42 deles são de pessoas deprimidas e os demais 97 são de pessoas não deprimidas. Dentre os demais áudios, 44 foram desconsiderados por não apresentarem o diagnóstico quando este trabalho se iniciou e os demais, 3, foram eliminados pelo fato de os áudios não estarem legíveis.

Os áudios foram obtidos a partir da gravação das entrevistas realizadas por um agente de computador criado com uma das finalidades de entrevistar pessoas. Durante essa entrevista a Ellie, o agente de computador, realizava perguntas pré-determinadas e interagia com os participantes, sendo controlada por um humano em outra sala. Todos os participantes responderam ao questionário psiquiátrico PHQ-8 antes das entrevistas, com a finalidade de obter um diagnóstico médico para ser usado como referência.

As perguntas realizadas nas entrevistas variam de pessoa para pessoa, mas possui alguns padrões. Dentre elas estão perguntas sobre como a pessoa está até se ela já teve depressão ou serviu ao exército, fato esse bem comum nos Estados Unidos.

Os áudios já foram pré-divididos em três conjuntos para a realização do desafio e *workshop* AVEC 2017 [67] (*Audio-Visual Emotion Challenge and workshop* - Desafio e workshop da Emoção Audiovisual). Essa separação foi realizada como um conjunto de

treinamento contendo 107 amostras, um de validação com 35 e um de teste contendo 44. Porém, para a realização deste trabalho não foi utilizada o conjunto de teste, pois o mesmo não tinha a classificação da entrevista quanto a presença ou ausência da depressão até a conclusão dos treinamentos aqui realizados.

5.2 Métodos Utilizados

5.2.1 Pré-processamento dos Dados

Inicialmente os dados de áudios foram editados para retirar das gravações as partes iniciais e finais, onde se tem a comunicação de uma outra pessoa explicando ao participante o funcionamento da entrevista com a Ellie.

Após os cortes, é realizado um processamento dos áudios que consistiram em realizar a fragmentação. Cada áudio foi subdividido em intervalos de 15 segundos para cada fragmento. O total de fragmentos gerados dos 107 áudios de treinamento foram 16.479 e para o conjunto de áudios de validação foram 6.679 para 32 áudios.

Não foram excluídas partes dos áudios em que o entrevistado permanecia em silêncio, por considerarmos o silêncio como uma característica relevante para detecção da presença ou ausência da depressão. Também não foram retirados trechos em que Ellie, o agente de computador, falava, já que todas as entrevistas havia a presença da fala do mesmo além do mesmo não expressar emoções.

Foram realizados os cálculos dos espectrogramas pela transformada rápida de Fourier; dos *mel-power* espectrogramas normalizados; e dos Coeficientes de Frequência Cepstral Mel normalizado (MFCC - *Mel Frequency Cepstral Coefficient*). Cada um dos três tipos de dados formam uma entrada diferente para a rede neural, juntamente com a informação do gênero ao qual cada amostra se refere, já que homens e mulheres demonstram sentimentos de formas diferentes.

5.2.2 Classificação

Para a obtenção dos resultados, foi utilizada uma rede neural convolucional. Os dados de entrada são os dados extraídos dos áudios e as variáveis de gêneros.

O treinamento foi dividido em quatro grupos, assim como representado na Figura 5.1. A diferença entre estes são o tipo de característica extraída do áudio. O primeiro e segundo grupos são utilizados como entradas os espectrogramas com os gêneros, variando as dimensões dos espectrogramas; no terceiro grupo os *mel-power* espectrogramas com os gêneros; e no último grupo os MFCC com os gêneros.

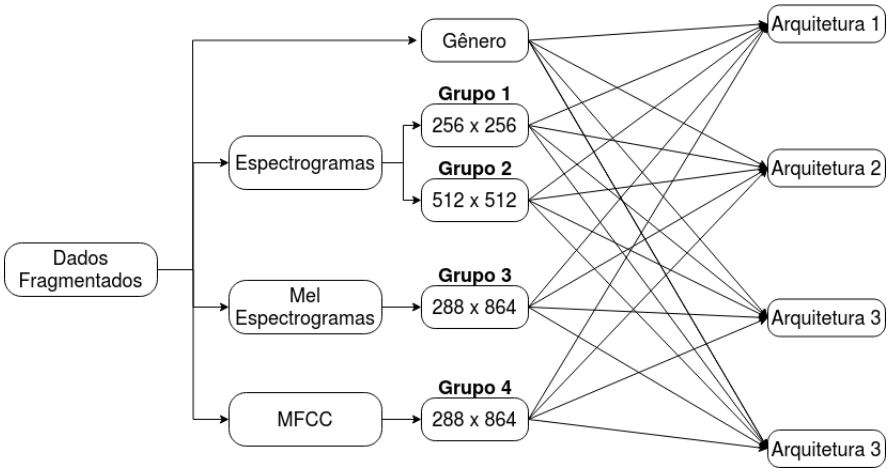


Figura 5.1: *Esquema da abordagem utilizada para o treinamento das arquiteturas aqui desenvolvidas.*

5.2.2.1 As Arquiteturas

As arquiteturas foram treinadas com três dimensões de entradas diferentes. Os espectrogramas calculados com a transformada rápida de Fourier foram utilizados para avaliar a relação das dimensões da entrada para o resultado obtido pela CNN, sendo utilizadas as dimensões de 256x256 e 512x512. Para os espectrogramas *mel-power* e MFCC foi utilizado a dimensão 288x864, como mostrado no esquema da Figura 5.1.

As arquiteturas 1, 2 e 3 aqui desenvolvidas usaram como inspiração o bloco residual da arquitetura ResNet, citada anteriormente em 4.1, e as camadas de *upsampling* e convolução transposta foram obtidas como inspiração da arquitetura UNet, citada em 4.2. Já a arquitetura 4 foi desenvolvida como sendo uma rede convolucional mais simples, não sendo considerada residual. A arquitetura 4 foi gerada com a finalidade de comparar os resultados das demais arquiteturas residuais com esta não residual e analisar quais delas obteriam um melhor desempenho.

A arquitetura 1, representada na Figura 5.2, possui 22 camadas contando as camadas de convolução, *pooling*, *fully-connected* e ainda o *dropout*. Foi utilizado a função de ativação ReLu no decorrer de toda a rede e na última camada, a função Softmax para auxiliar na classificação. Já nas arquiteturas 2 e 3, Figura 5.3 e Figura 5.7 respectivamente, possuem 20 camadas. E a arquitetura 4, Figura 5.8, possui 18 camadas.

Para melhor representar as arquiteturas citadas acima, as mesmas foram divididas em blocos. O bloco de concatenação se faz presente em todos os modelos e está representado na Figura 5.4. Neste bloco é onde se realiza a concatenação da entrada do gênero do paciente após utilizar a camada de *global avarage pooling*, concatenando a saída da camada de convolução e *batch normalization* em um vetor de uma dimensão. Após a concatenação temos camadas de *fully-connected*, *dense*, e a camada de *dropout*. Sendo este o bloco final das arquiteturas aqui desenvolvidas.

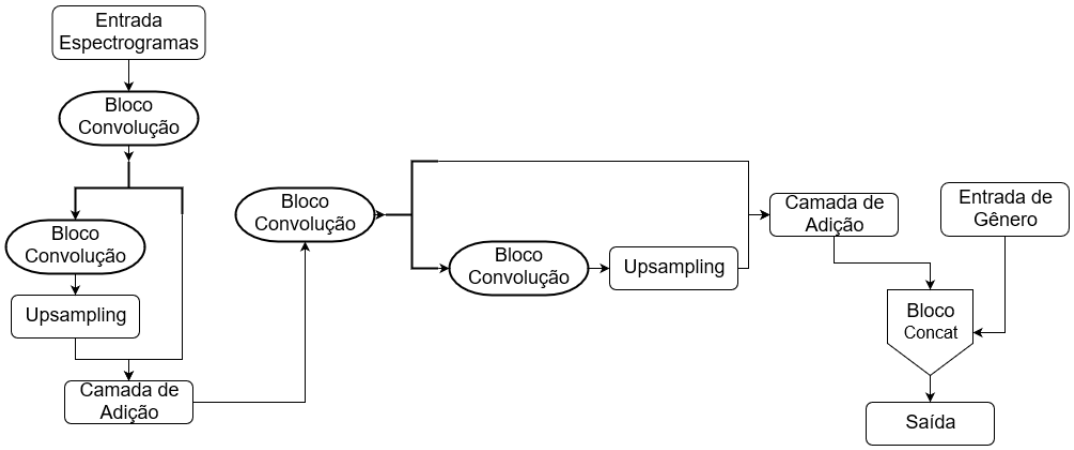


Figura 5.2: *Representação da arquitetura 1, que faz o uso da camada de upsampling. Os blocos Concat e Convolução foram apresentados, respectivamente na Figura 5.4 e Figura 5.5.*

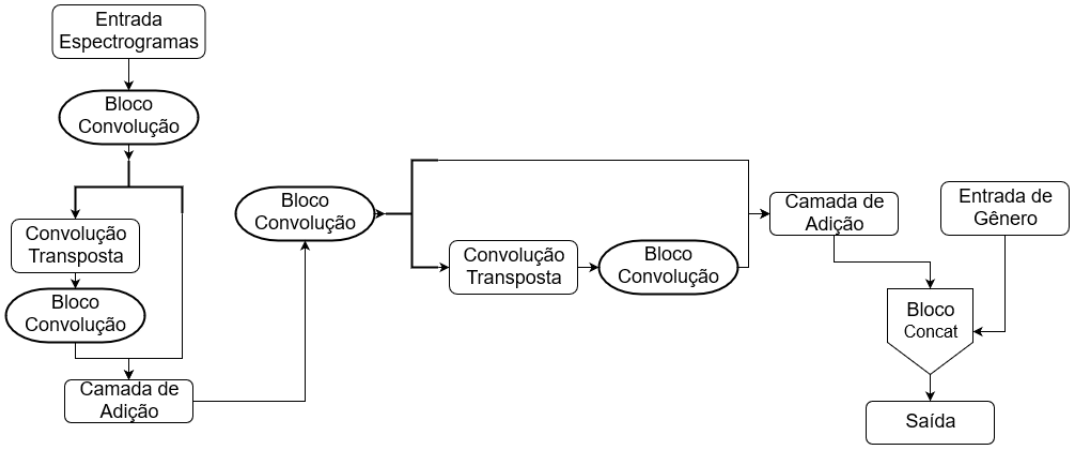


Figura 5.3: *Representação da arquitetura 2, que faz o uso da camada de convolução transposta. Os blocos aqui representados são apresentados na Figura 5.5, Bloco Convolução, e Figura 5.4, Bloco Concat.*

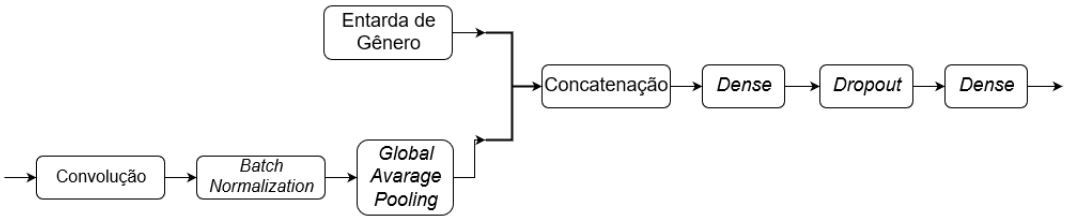


Figura 5.4: *Representação do Bloco Concat presente em todas as arquiteturas aqui desenvolvidas.*

O bloco de convolução é representado na Figura 5.5. Este é composto por uma camada de convolução seguido de uma camada de *batch normalization*, outra de convolução e por fim uma de *max pooling* e outra de *batch normalization*. Este bloco faz-se presente nas arquiteturas, exceto na 3.

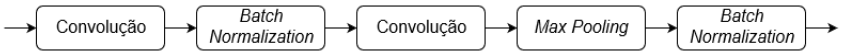


Figura 5.5: *Representação do Bloco Convolução presente nas arquiteturas 1, 2 e 4.*

O último bloco aqui desenvolvido faz-se presente na arquitetura 3 e este é representado na Figura 5.6. Neste bloco percebe-se a presença do bloco de convolução, porém ainda têm-se a realização da adição. Esta é realizada sem fazer o uso de técnicas e camadas para aumentar as dimensões da saída, como utilizado nas arquiteturas 1 e 2. Portanto, nesta arquitetura 3, teve uma reorganização de algumas camadas para tornar-se possível a realização dessa adição. Sendo essa adição responsável para as arquiteturas 1, 2 e 3 serem consideradas residuais.

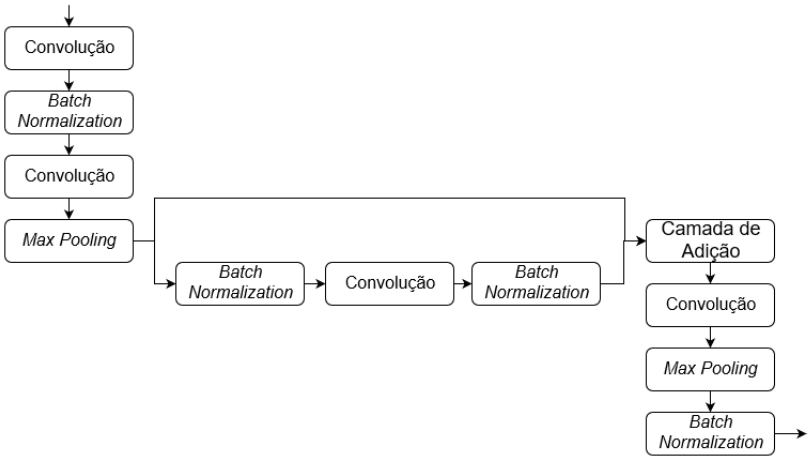


Figura 5.6: *Representação do Bloco Add, presente na arquitetura 3.*

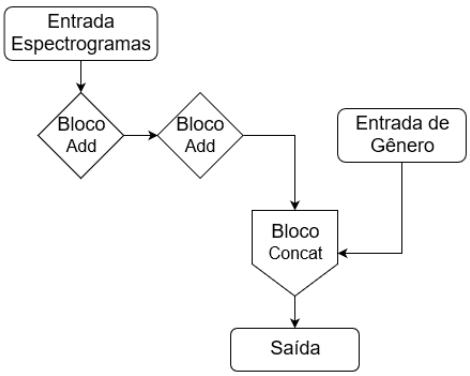


Figura 5.7: *Representação da arquitetura 3, que faz o uso da adição. Os blocos aqui representados são respectivamente o Add, Figura 5.6, e Concat, Figura 5.4.*

As imagens de entradas possuem detalhes pequenos, como alguns valores de frequência por não terem um período constante. Esses detalhes podem se perder ao passarem por tantas camadas de convolução e *pooling*. Esse fato influenciou o uso de

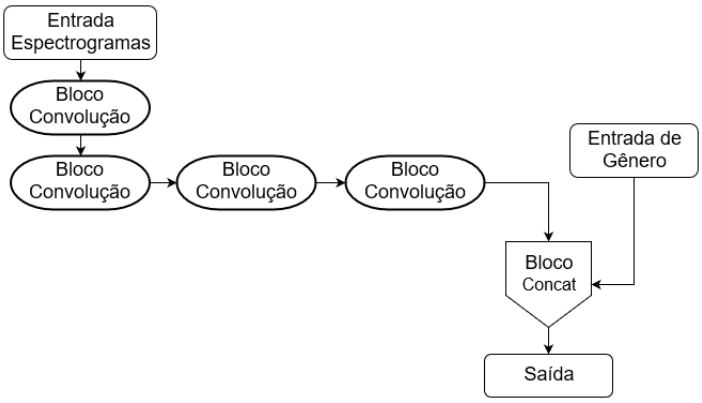


Figura 5.8: *Representação da arquitetura 4, que não realiza a operação de adição. Esta apresenta os blocos Convolução, Figura 5.5, e Concat, Figura 5.4, na sua estrutura.*

uma camada para realizar a adição da saída das camadas iniciais de *pooling* com algumas camadas profundas no decorrer das arquiteturas, variando para cada modelo. Com essas ações, buscou-se recuperar características perdidas que poderiam ser importantes para a classificação da entrada, sendo essas consideras redes residuais. Exceto a arquitetura 4, por não possuir essas camadas de adição.

5.3 Métricas de Avaliação

Foram utilizadas duas métricas de avaliação principais, a raiz do erro quadrático médio (*Root Mean Square Error* - RMSE) e o erro médio absoluto (*Mean Absolute Error* - MAE). Além dessas também serão usados a curva ROC com a área sob a curva (*Area Under Curve* - AUC) e a matriz de confusão.

5.3.1 Raiz do Erro Quadrático Médio (*Root Mean Square Error* - RMSE)

O RMSE é o cálculo da diferença entre os valores preditos por um modelo e os valores reais podendo ser calculado pela equação 5-1.

$$\text{RSME} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - y_{2i})^2}{N}} \tag{5-1}$$

em que y_{2i} é o valor predito referente a amostra i, y_i é o valor real referente a amostra i e N é o número de amostras.

Portanto os valores de RMSE serão usados para comparar o desempenho dos modelos durante os treinamentos, como também para comparar os modelos entre eles.

5.3.2 Erro Médio Absoluto (*Mean Absolute Error* - MAE)

MAE é a média das diferenças absolutas entre a previsão e a observação real, em que todas as diferenças individuais têm peso igual, não considerando a direção do erro. Tendo assim a chance de zerar o erro ao somar erros positivos e negativos. O MAE pode ser calculado pela equação 5-2.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |y_i - y_{2i}|}{N} \tag{5-2}$$

as variáveis nela citada são as mesmas citadas na equação de RMSE, logo y_{2i} é o valor predito, y_i é o valor real e N é o número de amostras.

5.3.3 Curva ROC e Área Sob a Curva ROC

A curva ROC é montada em um gráfico bidimensional usando os valores de sensibilidade e especificidade, Figura 5.9.

A sensibilidade mostra a capacidade da rede de mostrar os casos reais da classe Depressivos. Sendo calculada pela equação: $Sensibilidade = \frac{AP}{TP}$, onde AP refere aos acertos da classe Depressivos e TP ao total de casos da classe.

Já a especificidade é a capacidade da rede de mostrar os casos reais da classe Saudáveis. Sendo calculada pela equação: $Especificidade = \frac{AN}{TN}$, onde AN refere aos acertos da classe Saudáveis e TN ao total de casos da classe.

Porém a curva ROC apresenta no eixo das ordenadas o valor da sensibilidade e no eixo das abscissas o complemento da especificidade (CE). O CE consiste em subtrair de um o valor da especificidade, ou seja, $CE = 1 - Especificidade$.

A curva ROC ideal é representada na Figura 5.9 pela curva 1, onde o modelo analisado estaria classificando corretamente as duas classes. Porém a Figura 5.9 também representa a curva 2 e 3, onde a 2 o modelo está classificando bem, porém há presença de falsos positivos e falsos negativos. Já a curva 3, o modelo não consegue classificar entre positivo e negativo.

A área sob a curva tem a capacidade de discriminar a veracidade da resposta da rede quanto a classificação da depressão. Esta pode ser obtida por interações numéricas, neste caso utilizando o método dos trapézios, calculando a somatória das áreas dos n

[1] Imagem original em: https://tinyurl.com/tfxf9tt.

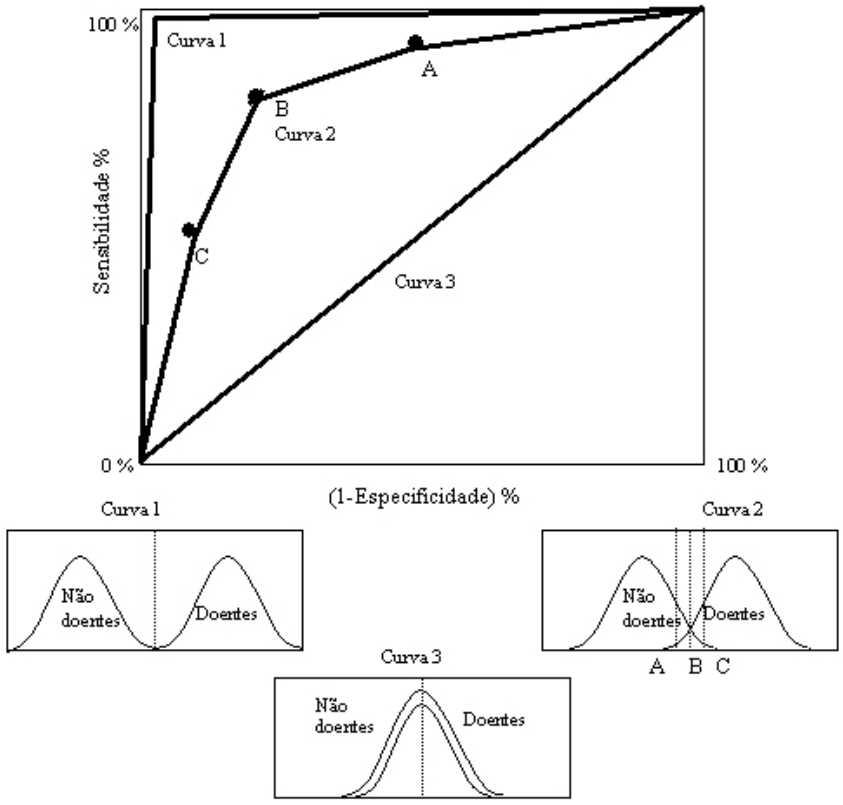


Figura 5.9: *Plano para um gráfico de curva ROC onde a classificação perfeita seria quando a especificidade e a sensibilidade forem igual ao representado pela curva 1, na havendo falsos positivos e falsos negativos.*[1].

trapézios no intervalo sob a curva. Quanto maior essa área, melhor a capacidade da rede de identificação e classificação.

5.3.4 Matriz de Confusão

A matriz de confusão fornece uma visão dos tipos de erros que estão sendo cometidos pelo modelo, além de mostrar o acerto. O tamanho da matriz varia de acordo com o número de classes, neste caso, por serem duas classes (Saudáveis e Depressivos), a matriz será de dimensões 2x2.

Para se obter a matriz são necessários os dados reais e os que foram preditos pelo modelo. Depois disso se organiza os dados corretos na sua classe e as predições erradas são colocadas em qual classe ela foi classificada. Assim podemos ver os falsos positivos (quando a pessoa é saudável e foi classificada como depressiva) e os falsos negativos (quando a pessoa possui depressão e é classificada como saudável).

Na Figura 5.10, onde as linhas representam as classes reais e as colunas as classes preditas pelo modelo, pode-se ver os falsos positivos representados por FP e os falsos

negativos representados por FN. Já TP (verdadeiro positivo) e TN (verdadeiro negativo) representam as classificações corretas realizada pela rede.

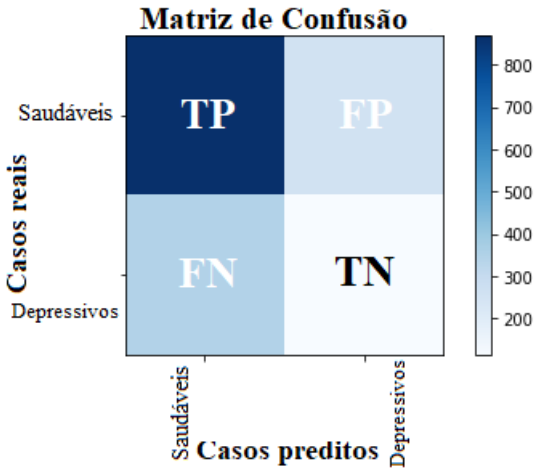


Figura 5.10: *Representação da matriz de confusão.*

CAPÍTULO 6

Resultados

Para avaliação dos resultados utiliza-se três métricas: RMSE, MAE e curva ROC, por fim é utilizada a matriz de confusão para análise entre os resultados dos modelos. A matriz de confusão é uma tabela que mostra as frequências de classificação para cada uma das classes do modelo, mostrando os falsos positivos, verdadeiros positivos, falsos negativos e verdadeiros negativos. A acurácia é a capacidade de um classificador diferenciar corretamente as classes. Para obtê-la é preciso calcular a proporção de verdadeiro positivo e verdadeiro negativo em todos os casos avaliados [6]. Portanto ela não é utilizada como métrica, já que a classe de não depressivos possui um maior número de amostras, logo possui um maior peso no valor final. Portanto a acurácia pode induzir a uma conclusão errada sobre o desempenho do modelo.

Porém, os valores obtidos de AUC ainda não são relevantes e não são utilizados para comparações com outros trabalhos na literatura [4, 81, 43, 83]. Sendo assim, a AUC e a matriz de confusão foram métricas utilizadas com a finalidade de analisar a evolução dos modelos aqui utilizados e o quanto conseguiram aprender, mas sempre buscando a redução dos erros.

O pré-processamento dos áudios seguiu a abordagem citada na Sessão 5.2.1. Portando para as amostras geradas três formas de representação de áudio foram testadas: espectrograma, mel-espectrograma e MFCC. Estas representações foram testadas em quatro arquiteturas.

6.1 Análise das Arquiteturas

A arquitetura 1 se difere dos demais por fazer uso da camada de *Upsampling*, para poder expandir a dimensão da saída de uma camada da rede para ser somada ao resultado de uma camada mais superficial, características de uma rede residual. Já a arquitetura 2, faz o uso da camada de convolução transposta, que aplica uma convolução ao aumentar a dimensão de saída. A arquitetura 3, também realiza a adição, porém as camadas foram rearranjadas de modo a ser possível a adição. Por fim, a arquitetura 4

é um modelo simples, onde não ocorre essa adição, logo é a única arquitetura que não apresenta características de um modelo residual.

A Tabela 6.1 apresenta os resultados obtidos com as arquiteturas aqui desenvolvidas. Onde são realçados de verde os melhores resultados em relação a métrica AUC e azul os menores RMSE, levando em conta cada tipo de entrada.

Analisando em relação aos tipos de entrada a arquitetura 1 obteve um menor erro RMSE (5,300) com os espectrogramas de dimensões 256 x 256, porém obteve uma AUC muito baixa (0,505), tendo um resultado mais equilibrado, em relação ao RMSE e o AUC, utilizando como entrada o MFCC, respectivamente 5,331 e 0,545. Já a arquitetura 2 obteve RMSE semelhantes a arquitetura anterior, porém teve uma AUC com cerca de pelo menos 0,008 de diferença, com exceção da utilização de MFCC em que a AUC foi de 0,003 superior. A arquitetura 3 obteve resultados de RMSE bem abaixo das arquiteturas 1 e 2, independentemente do tipo de entrada utilizada, porém teve um desempenho inferior em relação a AUC. Por fim a arquitetura 4 teve um RMSE melhor em relação aos resultados das arquiteturas 1 e 2 e mostrou RMSE e uma AUC menor do que da arquitetura 3 com os espectrogramas de 256 x 256, porém com os demais tipos de entradas a AUC foi maior do que todos os demais tipos de entradas em relação a arquitetura 3.

Tabela 6.1: *Resultados das Arquiteturas*

Arquiteturas	256 x 256			512 x 512			*mel-power*			MFCC		
	RMSE	MAE	AUC	RMSE	MAE	AUC	RMSE	MAE	AUC	RMSE	MAE	AUC
1	5,300	4,560	0,505	5,367	4,565	0,512	5,638	4,322	0,550	5,331	4,464	0,545
2	5,160	4,754	0,497	5,313	4,611	0,503	5,713	4,561	0,539	5,373	4,574	0,548
3	5,060	4,816	0,499	4,998	4,811	0,491	5,050	4,811	0,501	4,993	4,905	0,503
4	5,021	4,852	0,494	5,063	4,827	0,493	5,057	4,772	0,510	5,056	4,753	0,514

A partir da análise realizada acima, foram selecionadas duas arquiteturas para realização de mais testes. A arquitetura 1 foi a primeira selecionada por ela ter obtido melhores resultados em relação a AUC, mostrando um potencial para evolução. E ao considerar-se o RMSE a arquitetura que obteve um melhor resultado foi a 3, porém a mesma obteve as piores AUC, exceto quando utilizado como entrada o espectrograma de dimensões 256 x 256. Portanto a segunda arquitetura selecionada foi a 4.

Com as arquiteturas selecionados foram gerados outros dois modelos derivados delas. O modelo 1 foi derivado da arquitetura 1 com um núcleo de convolução de dimensões 3 x 3. O modelo 2 também foi derivado da arquitetura 1, porém o núcleo de convolução selecionado foi de dimensões 5 x 5. Já os modelos 3 e 4 foram derivados da arquitetura 4. A diferença entre eles é de que o modelo 3 possui um núcleo de convolução de dimensões 3 x 3 e o modelo 4, 5 x 5. A partir deste iremos analisar não apenas os tipos de entradas dos modelos como também o comportamento das arquiteturas 1 e 4 ao se usar diferentes dimensões dos núcleos de convoluções.

Foram realizados cerca de 30 testes para cada um dos quatro modelos e para cada uma das quatro entradas determinadas, totalizando cerca de 120 teste por modelo e 480

testes resultantes. Os primeiros resultados a serem analisados são os espectrogramas com dimensões de 256 x 256 e 512 x 512.

6.2 Análise dos resultados dos espectrogramas

A Tabela 6.2 mostra os resultados obtidos com os testes realizados com os espectrogramas nos quatro modelos, onde temos as médias dos resultados representada pela letra 'M' e o desvio padrão por 'DP'. Analisando esta tabela, podemos notar que no modelo 1 os espectrogramas de dimensões 256 x 256 obtiveram um menor RSME (5,30), porém os de dimensões 512 x 512 obtiveram uma maior AUC (0,512). Ao se analisar os desvios padrões, o 512 x 512 obteve um melhor resultado de RMSE, por possuir um menor desvio padrão, logo uma menor variância entre os resultados, tanto do RMSE, quanto da AUC. Já os resultados referentes ao modelo 2, os espectrogramas com dimensões 512 x 512 obtiveram um melhor resultado em todas as métricas em relação ao 256 x 256, com pouca diferença de desvio padrão. No modelo 3, o de dimensão 512 x 512 teve um melhor resultado levando em conta a análise da média e do desvio. Já no modelo 4, os espectrogramas de dimensões 512 x 512 obtiveram um melhor resultado, o desvio padrão para o erro RMSE foi muito alto, mas ainda assim teve um melhor resultado, apesar de se aproximarem dos resultados de espectrogramas de 256 x 256.

Tabela 6.2: *Resultados médios utilizando espectrograma.*

Modelos		256 x 256			512 x 512		
		RMSE	MAE	AUC	RMSE	MAE	AUC
1	M	5,30	4,56	0,505	5,36	4,56	0,512
	DP	0,38	0,16	0,021	0,16	0,10	0,012
2	M	6,46	4,73	0,501	6,19	4,54	0,509
	DP	0,29	0,33	0,026	0,32	0,26	0,029
3	M	6,26	4,62	0,494	5,89	4,40	0,511
	DP	0,62	0,39	0,031	0,50	0,16	0,020
4	M	6,74	4,86	0,409	5,84	4,66	0,485
	DP	0,18	0,24	0,035	0,73	0,35	0,028

Ao se comparar as médias e o desvio padrão dos 4 modelos entre si (Tabela 6.2), nota-se que o modelo 1 teve um melhor desempenho em relação aos demais, levando em conta as médias e desvios padrões aqui calculados. Nota-se também que os modelos com o núcleo de convolução com menor dimensões, modelo 1 e 3, tiveram um melhor desempenho ao se comparar com os modelos de maior dimensões do núcleo, 2 e 4. Portanto, os espectrogramas de dimensões 512 x 512 teve melhores resultados do que as dimensões 256 x 256, e em ambas as dimensões os modelos com menores dimensões

dos núcleos de convolução, 1 e 3, mostraram ter melhores resultados sobre os demais modelos, 2 e 4.

6.3 *Mel-power*

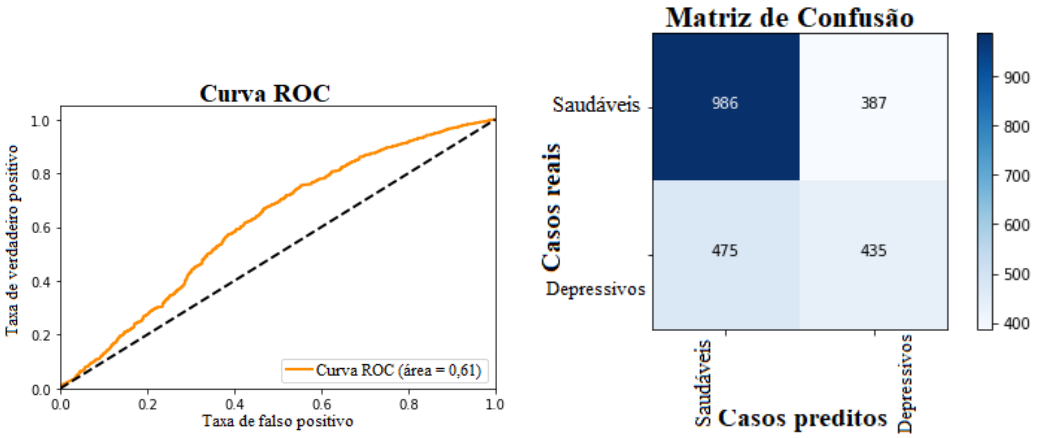
Os resultados obtidos com as entradas *mel-power* são apresentados na Tabela 6.3. Nesta podemos notar que novamente os modelos que possuem a menor dimensão do núcleo de convolução apresentaram melhores resultados. O modelo 1 obteve um RMSE de 5,86 e o modelo 3 de 5,53, enquanto os demais modelos apresentaram um RMSE pouco acima de seis. E ao considerar o RSME como principal métrica, o modelo 3 obteve um melhor resultado, porém ao considerar a AUC, o modelo 1 obteve o melhor desempenho.

Tabela 6.3: *Resultados médios utilizando mel-power.*

Modelos		RMSE	MAE	AUC
1	M	5,86	4,45	0,553
	DP	0,47	0,39	0,030
2	M	6,08	4,29	0,543
	DP	0,29	0,45	0,048
3	M	5,53	4,46	0,507
	DP	0,49	0,22	0,027
4	M	6,04	4,53	0,484
	DP	0,60	0,20	0,026

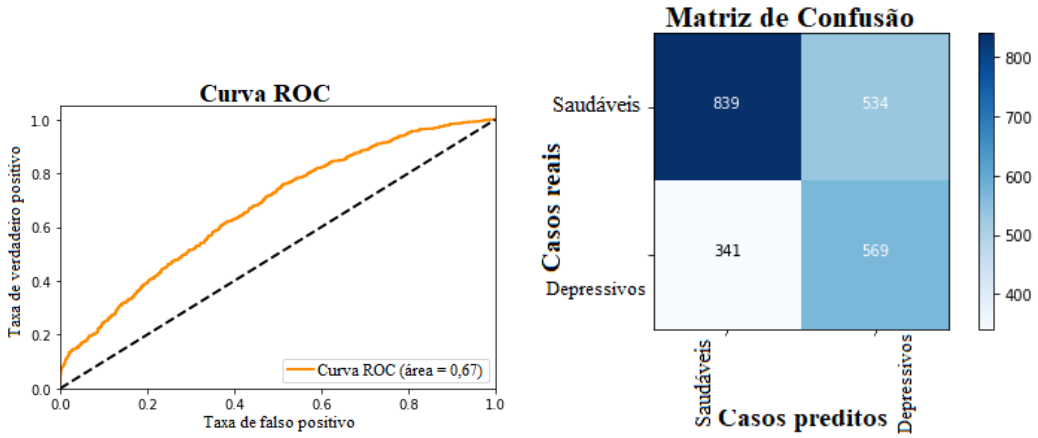
Porém as AUC ainda estão muito baixas onde mostra que os modelos, mesmo com erros mais baixos, ainda apresentam dificuldades de classificação. Isso pode ser visto na Figura 6.1, que representa a curva ROC e a matriz de confusão de um teste do modelo 1 que obteve uma maior AUC dentre os demais testes. Com essas imagens vemos que AUC de 0,614 ainda é um valor baixo, nota-se isso ao confirmar na matriz de confusão, Figura 6.1(b), onde se tem uma quantidade de falsos negativos semelhantes aos verdadeiros negativos. Já para os positivos têm-se uma maior incidência de verdadeiros.

O modelo 2 também obteve um resultado de destaque entre os testes, sendo esses os melhores resultados ao se considerar a AUC e a matriz de confusão como principais métricas. Este pode ser mostrado na curva ROC e na matriz de confusão deste teste que são representados na Figura 6.2. Neste caso, nota-se na matriz de confusão, Figura 6.2(b), uma maior incidência para os verdadeiros positivos e verdadeiros negativos, mas ainda assim possui uma taxa considerável de falsos positivos e negativos.



(a) *Curva ROC* (b) *Matriz de Confusão*

Figura 6.1: *Curva ROC e matriz de confusão referentes ao resultado de um teste realizado com o modelo 1, tendo como entrada mel-power. O RMSE obtido foi de 5,76 e o MAE de 3,89.*



(a) *Curva ROC* (b) *Matriz de Confusão*

Figura 6.2: *Curva ROC e matriz de confusão referentes ao resultado de um teste realizado com o modelo 2, tendo como entrada mel-power. O RMSE obtido foi de 5,82 e o MAE de 4,21.*

6.4 MFCC

Por fim, os resultados utilizando os MFCC como entradas são apresentados na Tabela 6.4. Com esta pode-se observar que o modelo 3 também obteve o menor valor de RMSE, 5,07, além de obter a segunda maior AUC, 0,528. O modelo 2 obteve um alto valor de RMSE, 6,20, porém obteve a maior AUC, 0,537.

Com esses resultados observa-se que novamente uma menor dimensão do núcleo de convolução teve um melhor resultado. Porém, no modelo 1, que faz o uso da camada de *upsamplig*, teve um pior resultado com a menor dimensão do núcleo.

Tabela 6.4: *Resultados médios utilizando MFCC.*

Modelos		RMSE	MAE	AUC
1	M	6,23	4,81	0,524
	DP	0,75	0,66	0,019
2	M	6,20	4,68	0,537
	DP	0,65	0,55	0,053
3	M	5,07	4,60	0,528
	DP	0,30	0,15	0,024
4	M	5,96	4,43	0,524
	DP	0,64	0,32	0,021

6.5 Discussão

A Tabela 6.5 apresenta os resultados obtidos para cada um dos modelos e tipos de entradas colocados lado a lado para melhor análise. Com isso podemos notar que o modelo 1 foi o que obteve o melhor desempenho utilizando as entradas *mel-power* e espectrograma de dimensões 512 x 512 e 256 x 256. Já para o MFCC como entrada, o modelo com melhor desempenho foi o modelo 3.

Com a análise desses resultados podemos concluir que uma maior dimensão do núcleo de convolução não trouxe benefícios para as arquiteturas 1 e 3 aqui desenvolvidas. Já em relação à dimensão do espectrograma, pode-se notar que a 512 x 512 obteve um melhor desempenho, sendo provável uma maior perda de detalhes da imagem de entrada ao passar pela arquitetura os espectrogramas de dimensões 256 x 256.

Tabela 6.5: *Comparação entre os resultados médios dos modelos.*

Testes		Modelo 1			Modelo 2			Modelo 3			Modelo 4		
		RMSE	MAE	AUC	RMSE	MAE	AUC	RMSE	MAE	AUC	RMSE	MAE	AUC
256 x 256	M	5,30	4,56	0,505	6,46	4,73	0,501	6,26	4,62	0,494	6,74	4,86	0,409
	DP	0,38	0,16	0,021	0,29	0,33	0,026	0,62	0,39	0,031	0,18	0,24	0,035
512 x 512	M	5,36	4,56	0,512	6,19	4,54	0,509	5,89	4,40	0,511	5,84	4,66	0,485
	DP	0,16	0,10	0,012	0,32	0,26	0,029	0,50	0,16	0,020	0,73	0,35	0,028
Mel-power	M	5,86	4,45	0,553	6,08	4,29	0,541	5,53	4,46	0,507	6,04	4,53	0,484
	DP	0,47	0,39	0,030	0,29	0,45	0,048	0,49	0,22	0,027	0,60	0,20	0,026
MFCC	M	6,23	4,81	0,524	6,20	4,68	0,537	5,07	4,60	0,528	5,96	4,43	0,524
	DP	0,75	0,66	0,019	0,65	0,55	0,053	0,30	0,15	0,024	0,64	0,32	0,021

Pelos resultados aqui obtidos a arquitetura residual com o uso da camada *upsampling*, modelo 1, obteve os melhores resultados com as entradas de espectrogramas em ambas dimensões analisadas. Porém na arquitetura *mel-power* o modelo 3 teve um RMSE inferior ao modelo 1, sendo respectivamente 5,53 e 5,86, porém o modelo 1 teve uma AUC de 0,553, enquanto o 3, 0,507. Já com o MFCC como entrada o modelo 3 obteve um melhor resultado. Porém as arquiteturas 2 e 3 não foram muito explorados até o momento. Portanto não se pode afirmar que o modelo 1, mesmo obtendo um melhor desempenho até o momento, é o melhor dentre os demais aqui desenvolvidos.

Todos os trabalhos aqui comparados foram realizados utilizando a mesma base de dados aqui utilizada. Além de utilizarem a mesma divisão dos dados entre treino e validação, que foram disponibilizados pela própria base (DAIC-WOZ). No decorrer deste trabalho foram utilizadas abordagens semelhantes e as mesmas métricas para realizar esta comparação com os trabalhos citados na Tabela 6.6, onde os resultados aqui obtidos são apresentados com a soma do desvio padrão entre parênteses (+DP).

Em [4] Alhanai et al. foi utilizado um modelo LSTM baseado em áudio tinha 3 camadas LSTM bidirecionais, cada uma com 128 nós ocultos, modo de mesclagem multiplicativa e taxas de desistência de entrada e de saída de 0,2. Isto na abordagem semelhante a aqui usada em relação ao uso apenas dos áudios, sendo apresentado na Tabela 6.6, onde obtiveram um RMSE de 6,50 e um MAE de 5,13.

Já Williamson et al. [81], utiliza uma abordagem de modelagem estatística que generaliza o uso de distribuições gaussianas para classificação binária no domínio da regressão multivariada. Foi realizada essa técnica para áudio, texto e vídeo, tendo um resultado de RMSE de 6,38 para áudio.

Ao comparar os resultados obtidos aos existentes na literatura, exibidos na Tabela 6.6, pode-se notar um avanço nos resultados levando em consideração o uso da base de dados DAIC-WOZ [79]. O menor resultado obtido pelos trabalhos citados é de 6,38 de RMSE e o melhor aqui obtido foi de 5,37 considerando um desvio padrão a mais.

Tabela 6.6: *Comparação com os resultados reportados na literatura*[1].

Trabalhos	Características	RMSE (+DP)	MAE (+DP)
Baseline AVEC [79]	Áudio	6,74	5,36
Alhanai et al. [4]	Áudio	6,50	5,13
Williamson et al. [81]	Áudio	6,38	5,32
Resultados obtidos neste trabalho			
512 x 512	Áudio	5,36 (5,52)	4,56 (4,66)
Mel-power	Áudio	5,86 (6,33)	4,45 (4,84)
MFCC	Áudio	5,07 (5,37)	4,60 (4,75)

[1] Tabela adaptada de [4].

CAPÍTULO 7

Conclusões

Dentre os resultados obtidos sobre a base de dados utilizada, conclui-se que os espectrogramas com dimensões 512 x 512 obtiveram resultados mais promissores quando comparados aos espectrogramas com dimensões 256 x 256. Deste modo, fornecendo *insights* para a realização de testes com espectrogramas de maiores dimensões.

Já em relação aos espectrogramas Mel e aos MFCC, não foi possível determinar qual obteve um melhor desempenho com os modelos. Portanto, seriam necessários mais testes utilizando os espectrogramas Mel e os MFCC como entrada.

Também se pode concluir que o aumento da dimensão do núcleo de convolução não trouxe melhorias significativas nos resultados para as arquiteturas aqui exploradas. A pesar disto, é possível notar o potencial das arquiteturas aqui desenvolvidas sobre a base de dados utilizada.

Dentre as limitações impostas no desenvolvimento de modelos de aprendizado de máquinas, uma delas é utilizar uma única base de dados (sendo ela do sul da Califórnia), deixando a possibilidade de agregar novas bases de dados e idiomas. Outra limitação enfrentada pelo presente trabalho foi o poder computacional disponível, que mostrou-se ineficiente dada a complexidade dos modelos abordados prejudicando o tempo de treinamento.

Como a base de dados utilizada foi construída em língua inglesa, também fica em aberto a hipótese deste trabalho também poder detectar depressão em língua portuguesa, dado o fato que este trabalho não utiliza o conteúdo falado pelo paciente.

O presente trabalho encontra-se em constante aprimoramento explorando novas técnicas de pré-processamento, explorando mais as arquiteturas aqui desenvolvidas, incluindo a manipulação de seus parâmetros. Encontra-se em desenvolvimento uma base de dados nacional, em São Paulo, em parceria com a rede eCare Life (São Paulo, Brasil) e voluntários filiados à Escola Politécnica da Universidade de São Paulo (São Paulo, Brasil).

Referências Bibliográficas

[1] ACADEMY, D. S. **Deep Learning Book**. http://www.deeplearningbook.com.br/, acessado em: dezembro de 2019, 2019.

[2] AGRAWAL, D. **Very deep convolutional networks for large-scale image recognition**. 2016.

[3] ALEXANDER, C. K.; SADIKU, M. N. **Fundamentos de circuitos elétricos**. AMGH Editora, 2013.

[4] ALHANAI, T.; GHASSEMI, M.; GLASS, J. **Detecting depression with audio/text sequence modeling of interviews**. In: *Proc. Interspeech*, p. 1716–1720, 2018.

[5] ALPERT, M.; POUGET, E. R.; SILVA, R. R. **Reflections of depression in acoustic measures of the patient's speech**. *Journal of affective disorders*, 66(1):59–69, 2001.

[6] BARATLOO, A.; HOSSEINI, M.; NEGIDA, A.; EL ASHAL, G. **Part 1: simple definition and calculation of accuracy, sensitivity and specificity**. 2015.

[7] BISTAFA, S. R. **Acústica aplicada ao controle do ruído**. Editora Blucher, 2018.

[8] BITOUK, D.; VERMA, R.; NENKOVA, A. **Class-level spectral features for emotion recognition**. *Speech communication*, 52(7-8):613–625, 2010.

[9] BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. **A training algorithm for optimal margin classifiers**. In: *Proceedings of the fifth annual workshop on Computational learning theory*, p. 144–152. ACM, 1992.

[10] BRACEWELL, R. N.; BRACEWELL, R. N. **The Fourier transform and its applications**, volume 31999. McGraw-Hill New York, 1986.

[11] BRIDLE, J. S. **Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition**. In: *Neurocomputing*, p. 227–236. Springer, 1990.

[12] BUBLITZ, C. F. **Sistema de reconhecimento de locutor integrado a comunicação e controle dos movimentos de um robô humanoide.** 2015.

[13] COSTA, S. W. S.; SOUZA, A.; PIRES, Y. **Computação afetiva: Uma ferramenta para avaliar aspectos afetivos em aplicações computacionais**. *Anais do Encontro Anual de Tecnologia da Informação e Semana Acadêmica de Tecnologia da Informação*, p. 286–290, 2015.

[14] COSTA, Y. M.; OLIVEIRA, L. S.; KOERICB, A. L.; GOUYON, F. **Music genre recognition using spectrograms**. In: *2011 18th International Conference on Systems, Signals and Image Processing*, p. 1–4. IEEE, 2011.

[15] COWIE, R.; DOUGLAS-COWIE, E. **Automatic statistical analysis of the signal and prosodic signs of emotion in speech**. In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, p. 1989–1992. IEEE, 1996.

[16] CUMMINS, N.; SCHERER, S.; KRAJEWSKI, J.; SCHNIEDER, S.; EPPS, J.; QUATIERI, T. F. **A review of depression and suicide risk assessment using speech analysis**. *Speech Communication*, 71:10–49, 2015.

[17] DA SAÚDE, M. **Depressão: causas, sintomas, tratamentos, diagnóstico e prevenção**. https://saude.gov.br/saude-de-a-z/depressao, acessado em janeiro de 2020.

[18] DA SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais para engenharia e ciências aplicadas curso prático**. *São Paulo: Artliber*, 2010.

[19] DE MOURA, L. F. N.; BISCAINHO, L. W. P.; DA SILVA, E. A.; DE LUCENA, S. C. **Estudo de psicoacústica e suas aplicac oes**. 2006.

[20] DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. **ImageNet: A Large-Scale Hierarchical Image Database**. In: *CVPR09*, 2009.

[21] DINIZ, P. S.; DA SILVA, E. A.; NETTO, S. L. **Processamento Digital de Sinais-: Projeto e Análise de Sistemas**. Bookman Editora, 2014.

[22] DONG, C.; LOY, C. C.; HE, K.; TANG, X. **Image super-resolution using deep convolutional networks**. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[23] DUMITRESCU.; BOIANGIU, C.-A. **A study of image upsampling and downsampling filters**. *Computers*, 8:30, 04 2019.

[24] DUMOULIN, V.; VISIN, F. **A guide to convolution arithmetic for deep learning**, 2016. cite arxiv:1603.07285.

[25] FALK, T.; MAI, D.; BENSCH, R.; ÇIÇEK, Ö.; ABDULKADIR, A.; MARRAKCHI, Y.; BÖHM, A.; DEUBNER, J.; JÄCKEL, Z.; SEIWALD, K.; OTHERS. **U-net: deep learning for cell counting, detection, and morphometry**. *Nature methods*, 16(1):67–70, 2019.

[26] FAYAK, H. **Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between**, 2016.

[27] FERNANDES, J. C. **Acústica e ruídos**. *Bauru: Unesp*, 102, 2002.

[28] FLECK, M. P. D. A.; LAFER, B.; SOUGEY, E. B.; DEL PORTO, J. A.; BRASIL, M. A. A.; JURUENA, M. F. **Diretrizes da associação médica brasileira para o tratamento da depressão (versão integral)**. *Revista brasileira de psiquiatria= Brazilian journal of psychiatry. São Paulo, SP. Vol. 25, n. 2 (jun. 2003), p. 114-122*, 2003.

[29] FRAZER, L. N.; MERCADO, E. **A sonar model for humpback whale song**. *IEEE Journal of Oceanic Engineering*, 25(1):160–182, 2000.

[30] GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; BENGIO, Y. **Deep learning**, volume 1. MIT press Cambridge, 2016.

[31] GUEDES, M. V. **Grandezas periódicas não sinusoidais**, 1992.

[32] HAFEMANN, L. G. **An analysis of deep neural networks for texture classification**. Master's thesis, Universidade Federal do Paraná, Curitiba, 2014.

[33] HAMILTON, M. **A rating scale for depression**. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56, 1960.

[34] HE, K.; SUN, J.; TANG, X. **Guided image filtering**. In: *European conference on computer vision*, p. 1–14. Springer, 2010.

[35] HE, K.; ZHANG, X.; REN, S.; SUN, J. **Deep residual learning for image recognition**. In: *The IEEE conference on computer vision and pattern recognition (CVPR)*, p. 770–778, June 2016.

[36] HE, L.; LECH, M.; MADDAGE, N.; ALLEN, N. **Stress and emotion recognition using log-gabor filter analysis of speech spectrograms**. In: *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, p. 1–6. IEEE, 2009.

[37] HE, L.; LECH, M.; MADDAGE, N. C.; ALLEN, N. B. **Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech**. *Biomedical Signal Processing and Control*, 6(2):139–146, 2011.

[38] HELFER, B. S.; QUATIERI, T. F.; WILLIAMSON, J. R.; MEHTA, D. D.; HORWITZ, R.; YU, B. **Classification of depression state based on articulatory precision.** In: *Interspeech*, p. 2172–2176, 2013.

[39] HODGSON, J. **Understanding records: A field guide to recording practice**. Bloomsbury Publishing, 2010.

[40] HOFFMAN, G. M. A.; GONZE, J. C.; MENDLEWICZ, J. **Speech pause time as a method for the evaluation of psychomotor retardation in depressive illness**. *British Journal of Psychiatry*, 146(5):535–538, 1985.

[41] HUBEL, D. H.; WIESEL, T. N. **Receptive fields and functional architecture of monkey striate cortex**. *The Journal of physiology*, 195(1):215–243, 1968.

[42] IOFFE, S.; SZEGEDY, C. **Batch normalization: Accelerating deep network training by reducing internal covariate shift**. *arXiv preprint arXiv:1502.03167*, 2015.

[43] JAN, A.; MENG, H.; GAUS, Y. F. A.; ZHANG, F.; TURABZADEH, S. **Automatic depression scale prediction using facial expression dynamics and regression**. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, p. 73–80, New York, NY, USA, 2014. ACM.

[44] JIA, Y.; ZHANG, Y.; WEISS, R.; WANG, Q.; SHEN, J.; REN, F.; NGUYEN, P.; PANG, R.; MORENO, I. L.; WU, Y.; OTHERS. **Transfer learning from speaker verification to multispeaker text-to-speech synthesis**. In: *Advances in neural information processing systems*, p. 4480–4490, 2018.

[45] KANNAN, A.; DATTA, A.; SAINATH, T. N.; WEINSTEIN, E.; RAMABHADRAN, B.; WU, Y.; BAPNA, A.; CHEN, Z.; LEE, S. **Large-scale multilingual speech recognition with a streaming end-to-end model**. *arXiv preprint arXiv:1909.05330*, 2019.

[46] KARN, U. **An intuitive explanation of convolutional neural networks**. *The Data Science Blog*, 2016.

[47] KARPATHY, A.; TODERICI, G.; SHETTY, S.; LEUNG, T.; SUKTHANKAR, R.; FEI-FEI, L. **Large-scale video classification with convolutional neural networks**. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, p. 1725–1732, 2014.

[48] KRAEPELIN, E. **Manic depressive insanity and paranoia.** *The Journal of Nervous and Mental Disease*, 53(4):350, 1921.

[49] KROENKE, K.; STRINE, T. W.; SPITZER, R. L.; WILLIAMS, J. B.; BERRY, J. T.; MOKDAD, A. H. **The phq-8 as a measure of current depression in the general population**. *Journal of affective disorders*, 114(1-3):163–173, 2009.

[50] LECUN, Y. **The mnist database of handwritten digits**. *http://yann. lecun. com/exdb/mnist/*, 1998.

[51] LECUN, Y.; BENGIO, Y.; OTHERS. **Convolutional networks for images, speech, and time series**. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[52] LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. **Gradient-based learning applied to document recognition**. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[53] LI, X.; CHEBIYYAM, V.; KIRCHHOFF, K.; AMAZON, A. **Speech audio super-resolution for speech recognition**. *Proc. Interspeech 2019*, p. 3416–3420, 2019.

[54] MA, X.; YANG, H.; CHEN, Q.; HUANG, D.; WANG, Y. **Depaudionet: An efficient deep model for audio based depression classification**. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, p. 35–42, New York, NY, USA, 2016. ACM.

[55] MALLAT, S. **A wavelet tour of signal processing: The sparse way 3th ed**, 2009.

[56] MOORE II, E.; CLEMENTS, M. A.; PEIFER, J. W.; WEISSER, L. **Critical analysis of the impact of glottal features in the classification of clinical depression in speech**. *IEEE transactions on biomedical engineering*, 55(1):96–107, 2007.

[57] NAIR, V.; HINTON, G. E. **Rectified linear units improve restricted boltzmann machines**. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*, p. 807–814, 2010.

[58] NASSIF, A. B.; SHAHIN, I.; ATTILI, I.; AZZEH, M.; SHAALAN, K. **Speech recognition using deep neural networks: A systematic review**. *IEEE Access*, 7:19143–19165, 2019.

[59] ODENA, A.; DUMOULIN, V.; OLAH, C. **Deconvolution and checkerboard artifacts**. *Distill*, 1(10):e3, 2016.

[60] OLIPHANT, T. E. **A guide to NumPy**, volume 1. Trelgol Publishing USA, 2006.

[61] OMS. **Folha Informativa - Depressão**. https://www.paho.org/bra/index.php?option=com_content&view=article&id=5635:folha-informativa-depressao&Itemid=1095, acessado em junho de 2018, 2018.

[62] OPPENHEIM, A. V. **Speech spectrograms using the fast fourier transform**. *IEEE spectrum*, 7(8):57–62, 1970.

[63] PAULINO, C. A.; PREZOTTO, A. O.; CALIXTO, R. F. **Associação entre estresse, depressão e tontura: uma breve revisão**. *Revista Equilíbrio Corporal e Saúde*, 1(1), 2015.

[64] RAJESH, B.; BHALKE, D. **Automatic genre classification of indian tamil and western music using fractional mfcc**. *International Journal of Speech Technology*, 19(3):551–563, 2016.

[65] RAVANELLI, M.; PARCOLLET, T.; BENGIO, Y. **The pytorch-kaldi speech recognition toolkit**. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6465–6469. IEEE, 2019.

[66] RIBEIRO, G.; GOMES, R.; COSTA, S.; COSTA, W. **Análise mel-cepstral na discriminação de patologias laríngeas**. 2014.

[67] RINGEVAL, F.; SCHULLER, B.; VALSTAR, M.; GRATCH, J.; COWIE, R.; PANTIC, M. **Summary for avec 2017: Real-life depression and affect challenge and workshop**. In: *Proceedings of the 25th ACM international conference on Multimedia*, p. 1963–1964. ACM, 2017.

[68] ROCKMORE, D. N. **The fft: an algorithm the whole family can use**. *Computing in Science & Engineering*, 2(1):60–64, 2000.

[69] RODRIGUES, M. **O diagnóstico de depressão**. *Psicologia USP*, 11(1):155–187, 2000.

[70] RONNEBERGER, O.; FISCHER, P.; BROX, T. **U-net: Convolutional networks for biomedical image segmentation**. In: *International Conference on Medical image computing and computer-assisted intervention*, p. 234–241. Springer, 2015.

[71] RUELA, A. S. **Redes neurais feedforward e backpropagation**. *UFOP (http://www. decom. ufop. br/imobilis/wpcontent/uploads/2012/06/03_Feedforward-e-Backpropagation. pdf)(Acedido em 3 05 2014)*, 2012.

[72] RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. **Learning internal representations by error propagation**. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[73] SANTOS, I. S.; TAVARES, B. F.; MUNHOZ, T. N.; ALMEIDA, L. S. P. D.; SILVA, N. T. B. D.; TAMS, B. D.; PATELLA, A. M.; MATIJASEVICH, A. **Sensibilidade e especificidade do patient health questionnaire-9 (phq-9) entre adultos da população geral**. *Cadernos de Saúde Pública*, 29:1533–1543, 2013.

[74] SMITH, J. O. **Spectral Audio Signal Processing**. http://-ccrma.stanford.edu/~jos/sasp/, acessado em: dezembro de 2019. 2011.

[75] SMITH, S. W.; OTHERS. **The scientist and engineer's guide to digital signal processing**. 1997.

[76] SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. **Dropout: a simple way to prevent neural networks from overfitting**. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[77] STEVENS, S.; VOLKMANN, J.; NEWMAN, E. **The mel scale equates the magnitude of perceived differences in pitch at different frequencies**. *Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[78] VALENTINI, W.; LEVAV, I.; KOHN, R.; MIRANDA, C. T.; MELLO, A. D. A. F. D.; MELLO, M. F. D.; RAMOS, C. P. **Treinamento de clínicos para o diagnóstico e tratamento da depressão**. *Revista de Saúde pública*, 38:523–528, 2004.

[79] VALSTAR, M.; GRATCH, J.; SCHULLER, B.; RINGEVAL, F.; LALANNE, D.; TORRES TORRES, M.; SCHERER, S.; STRATOU, G.; COWIE, R.; PANTIC, M. **Avec 2016: Depression, mood, and emotion recognition workshop and challenge**. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, p. 3–10. ACM, 2016.

[80] WATANABE, S.; DELCROIX, M.; METZE, F.; HERSHEY, J. R. **New era for robust speech recognition: exploiting deep learning**. Springer, 2017.

[81] WILLIAMSON, J. R.; GODOY, E.; CHA, M.; SCHWARZENTRUBER, A.; KHORRAMI, P.; GWON, Y.; KUNG, H.-T.; DAGLI, C.; QUATIERI, T. F. **Detecting depression using vocal, facial and semantic communication cues**. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, p. 11–18. ACM, 2016.

[82] YANG, L.; JIANG, D.; HAN, W.; SAHLI, H. **Dcnn and dnn based multi-modal depression recognition**. In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, p. 484–489, Oct 2017.

[83] YANG, L.; JIANG, D.; XIA, X.; PEI, E.; OVENEKE, M. C.; SAHLI, H. **Multimodal measurement of depression using deep learning models**. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, p. 53–59. ACM, 2017.

[84] YENIGALLA, P.; KUMAR, A.; TRIPATHI, S.; SINGH, C.; KAR, S.; VEPA, J. **Speech emotion recognition using spectrogram & phoneme embedding.** In: *Interspeech*, p. 3688–3692, 2018.

[85] ZADEH, L. A. **Theory of filtering**. *Journal of the Society for Industrial and Applied Mathematics*, 1(1):35–51, 1953.

[86] ZEILER, M. D.; FERGUS, R. **Visualizing and understanding convolutional networks**. In: *European conference on computer vision*, p. 818–833. Springer, 2014.

[87] ZHANG, A.; LIPTON, Z. C.; LI, M.; SMOLA, A. J. **Dive into deep learning**. *Unpublished draft. Retrieved*, 3:319, 2019.

[88] ZHANG, S.; ZHANG, S.; HUANG, T.; GAO, W. **Multimodal deep convolutional neural network for audio-visual emotion recognition**. In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, p. 281–284. ACM, 2016.

[89] ZIGELBOIM, G.; SHALLOM, I. D. **A comparison study of cepstral analysis with applications to speech recognition**. In: *2006 International Conference on Information Technology: Research and Education*, p. 30–33. IEEE, 2006.

[90] ZORZETTO FILHO, D. **Sintomas somáticos da depressão: características sociodemográficas e clínicas em pacientes de atenção primária em curitiba, brasil**. 2009.