ENEMP 2022
XL CONGRESSO BRASILEIRO DE SISTEMAS PARTICULADOS
23 a 26 de outubro de 2022
Uberlândia/MG

CLASSIFICAÇÃO DE CROCÂNCIA DE ALIMENTOS SECOS POR REDES NEURAIS ARTIFICIAIS

RAFAEL Z. LOPES, GUSTAVO C. DACANAL*

Departamento de Engenharia de Alimentos, Faculdade de Ciência Animal e Engenharia de Alimentos,
Universidade de São Paulo, FZEA-USP, 13635-900, Pirassununga, SP, Brasil.
*e-mail: gdacanal@usp.br

RESUMO - A crocância é uma qualidade sensorial obtida pela geração de ruídos sonoros durante a fragmentação de materiais alimentícios. Ela tem um papel importante na decisão de compra do cliente. Conhecer suas características é um fator decisivo para intensificar as sensações causadas pelo alimento crocante. Os poucos estudos que utilizam redes neurais para avaliar a crocância têm seus dados obtidos a partir de um texturômetro. Este trabalho apresenta uma nova abordagem ao classificar três alimentos diferentes a partir de vídeos de mastigação coletados em bases digitais: Frago Frito, Torradas e Batatas Fritas. Os áudios foram convertidos em espectro de potência, via código Python, a fim de obter dois tipos de entrada para as arquiteturas de rede neural: os coeficientes Cepstrais da Frequência Mel e a Transformada Rápida de Fourier. Os modelos desenvolvidos alcançaram 95% de acurácia na identificação do tipo de alimento usado, o que demonstra que a rede conseguiu identificar padrões sonoros diferentes para cada um dos grupos utilizados.

INTRODUÇÃO

O uso de Redes Neurais Artificiais tem aumentado durante as últimas décadas demonstrando resultados viáveis em diversas áreas desde a engenharia à saúde, um exemplo é a análise dos gritos dos bebês para identificar doenças graves (JI et al., 2021). O mercado de análise de alimentos possui pontos ainda não explorados, o som é a característica mais relevante na compra de um produto, porém os mesmos são avaliados através de propriedades mecânicas como a textura. (BUISSON; SILBERZAHN, 2010).

Os alimentos produzem sons durante a mastigação são conhecidos como alimentos crocantes, eles são considerados únicos dentro do espectro da Engenharia de Alimentos. As características dos sons vêm principalmente dos processos de fritura e secagem. Em um desses processos, a água que antes preenchia a estrutura é substituída pelo ar ou pelo óleo, o que resulta em estruturas mais rígidas. Elas são responsáveis pela melhor propagação do som ao serem cisalhadas. Estudos indicam que a diferença entre os sons de cada alimento crocante está na entonação. Ela é subdivida em *crispness* e *crunchness*. A primeira é categorizada como um som mais agudo enquanto a segunda é um som mais grave. (VICKERS, 1984). Frequências menores de 500 Hz são identificadas durante a mastigação, esse som grave ocorre depois de ser filtrado pela mucosa bocal. Já frequências acima de 500 Hz são consideradas mais agudas, elas são captadas durante o primeiro cisalhamento do alimento pela arcada dentária. Essas diferenças empíricas já foram analisadas, isso abre portas para uma rede neural classificar a crocância de diferentes alimentos de forma científica.

Cada alimento crocante tem um som, eles podem variar dependendo do alimento, as batatas fritas têm um tempo médio menor do ruido crocante do que as torradas. Treinar uma rede neural artificial para classificar os

alimentos crocantes é o primeiro passo. Estudos nesta área aproximaram a função de textura usando dados de Força dos texturômetros (TUNICK et al., 2013; KATO et al., 2018, 2019). Eles tiveram desafios em relação ao ruído do equipamento, tanto que uma pequena mudança no som provocava desvios substanciais nos resultados (ANDREANI et al., 2020; DE MORAES et al., 2022). Uma solução proposta foi o desenvolvimento de braço manual de madeira capaz de cisalhar o alimento, e então captar um som mais limpo. O estudo desenvolveu uma função para quantificar a textura do alimento tendo como base a energia da força de cisalhamento e a intensidade sonora (AKIMOTO et al., 2018). A abordagem adotada neste trabalho capturou áudios de vídeos do YouTube dentro da categoria ASMR (Resposta Autônoma do Meridiano) para processá-los em Python. A proposta foi avaliar a crocância apenas usando o próprio som, sem forças ou outras relações envolvidas.

A biblioteca Librosa é um pacote de análise de áudio capaz de isolar parâmetros-chave como os coeficientes Cepstrais da Frequência Mel (MFCCs), que simbolizam a identidade do som. O Keras é uma biblioteca de aprendizagem de máquinas que será usada para validar os padrões identificados no Librosa através da simulação de um cérebro humano que treina aprendendo qual é a crocância de cada alimento (CHOLLET, 2017). Um desafio Kaggle inspirou a primeira arquitetura da rede neural. Eles desenvolveram uma simples rede neural profunda totalmente conectada para classificar os sons alimentares de 20 tipos diferentes de alimentos. (MA et al., 2020) O melhor modelo ganhou usando o MFCC como entrada com uma acurácia de 90%.

Por outro lado, a Rede Neural Convolucional é uma arquitetura mais complexa capaz de lidar com grandes quantidades de dados. Sua aplicação é quase universal, mas mais focada na classificação de imagem e áudio. (ZHOU, 2020). O processo de convolução é uma operação de multiplicação entre os termos de duas matrizes, a original e um filtro, resultando em uma matriz menor. Os dois modelos desenvolvidos neste trabalho receberam a transformada rápida de Fourier (FFT) como a entrada da rede.

Este trabalho visa investigar as características únicas do ruido crocante e suas diferenças para cada um dos três alimentos apresentados: Torrada, Batata Chips e Frango Frito. A crocância tem influência direta na percepção das características desejáveis desses produtos, podendo mudar sabores e sensações (SPENCE, 2016). Compreender suas características e diferenças permite otimizar seus processos de fabricação em busca do som mais atraente para o consumido. Contudo, primeiro deve-se explorar se é possível generalizar uma função que descreve o comportamento do ruido crocante ao comparar os sons dos diferentes alimentos nos dois modelos de rede neural propostos.

METODOLOGIA

Aquisição das Amostras de Áudio

Um total de 584 arquivos de áudio foram coletados do YouTube, eles foram a principal fonte das três classes diferentes: Torrada, Batata frita e Frango frito. A categoria de vídeos *ASMR* proveu vídeos limpos, além disso eles foram gravados em quatro países diferentes: Alemanha, Coréia do Sul, Estados Unidos e Japão. Esses vídeos foram cortados em um segundo de duração usando o software editor de vídeo Shotcut© foi o único procedimento necessário. Os arquivos de áudio extraídos tinham uma taxa de amostragem (SR) de 22050 kHz, todos convertidos em áudio. O Shotcut converteu os arquivos de áudio em formato .wav. O uso de segmentos de um segundo seguiu as Diretrizes "Fair use on Youtube". Além disso, não houve exposição dos proprietários das gravações e qualquer uso não autorizado das mesmas.

Python e Wolfram Mathematica foram as duas principais linguagens utilizadas para analisar as propriedades dos arquivos de áudio. Librosa e Tensorflow foram as bibliotecas escolhidas para processar o som e desenvolver os modelos de aprendizagem de máquina. *Librosa.load* gerou uma série temporal de áudio. *Librosa.effects.percussive* decompôs o vetor em componentes percussivos necessários para analisar o som com maior precisão.

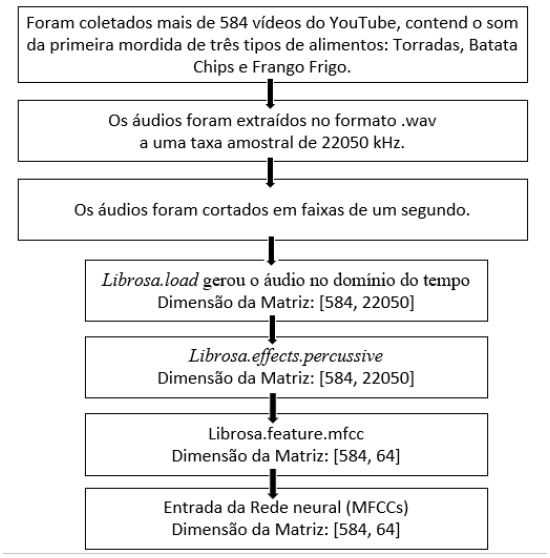


Figura 1: Exemplo de Pré-processamento dos áudios em Python.

Extração da Transformada de Fourier e do *Beat*

Para adquirir os conjuntos de dados de treinamento, as características de áudio precisavam ser extraídas. No domínio do tempo, a característica Beat forneceu uma visualização relevante de como funciona a amplitude sonora do alimento crocante, a batida demonstra o instante de tempo que a mordida termina. *Librosa.beat.beat_track* calcula a batida usando programação dinâmica (ELLIS, 2007). É um método em três etapas que correlaciona a amplitude do som com seu tempo. Para os músicos, o ritmo é a unidade de medida básica de uma melodia. Ela simboliza a velocidade com que a música se desdobra. (LEVITIN, 2007). No entanto, ao extrair esta característica de um som crocante, ela deve ter um novo tipo de interpretação, o qual será explicado mais detalhadamente na próxima seção.

No aprendizado de máquinas, há uma regra ao classificar características dependente do tempo: você tem que padronizar corretamente o tempo entre as amostras. As redes neurais são sensíveis a pequenas variações no conjunto de dados. Dados aleatórios da Internet não podem preencher este requisito, no entanto, é possível utilizá-los no domínio da frequência aplicando a Transformada Discreta de Fourier. O FFT representa um algoritmo que computa a transformada discreta de Fourier de forma mais eficiente. Librosa utiliza o *numpy.fft.fft* para computar a transformada. Este estudo escolheu 100 e 64 como as melhores dimensões das matrizes de entrada da rede neural. A primeira foi usada na abordagem CONV1D e a segunda para computar os MFCCs, respectivamente.

Extração dos MFCCs no Librosa

O conceito dos MFCCs é baseado em dois termos principais: a Frequência Mel e o *Cepstrum*. A Frequência Mel estima o tom de um som, é uma forma relevante de perceber logaritmicamente a frequência. Os humanos discernem melhor as diferenças de tom em baixa frequência do que em alta frequência, mesmo que possam ter a mesma variação linear. A Frequência Mel muda a percepção linear de um espectrograma para a percepção humana, aplicando uma Escala Mel. Isto implica que distâncias iguais na escala têm a mesma distância perceptual, por isso a Frequência Mel comportasse como um logaritmo, ela imita a percepção humana.

Por outro lado, o *Cepstrum* é um espectro de um espectro. Ele proporciona uma separação física natural da relativo à informação proveniente dos formadores do som, ou seja, ele identifica a identidade do som.

Librosa.effects.mfcc calculou os coeficientes necessários para este estudo, além disso, estes valores retornados podem ser escolhidos. Tradicionalmente, o número de MFCCs retornados está entre 12 e 20, nesta pesquisa, o melhor desempenho na Rede Neural foi alcançado utilizando 64 MFCCs. Esta representação significa o número de características necessárias para descrever os padrões crocantes, no total foram utilizados 64 padrões.

Augmentation no Wolfram Mathematica

O *augmentation* foi a última estratégia utilizada no pré-processamento. Ele permitiu um aumento de dez vezes nos dados de entrada usando os mesmos 584 áudios originais. O Wolfram Mathematica centrou os quatro diferentes métodos aplicados neste estudo: *Timeshift*, *Reverb*, *UnformDistribution* e *Noise*.

Timeshift permite deslocar uma parte do áudio no eixo do tempo, que pode ser o áudio inteiro ou pequenas partes do mesmo. Um

exemplo prático permite pegar uma faixa de 0,1s do meio do áudio e substituí-la por uma mesma faixa do início. A facilidade de aplicação desta técnica permitiu que ela fosse reutilizada seis vezes, sendo a única aplicada mais de uma vez. *Reverb* reflete o próprio som. A reverberação, ao contrário do eco, não tem continuidade suficiente para ser distinguível do som original. *UnformDistribution* varia aleatoriamente a amplitude dos áudios de maneira uniforme. Por último, mas não menos importante, *Noise* aplica um ruído ao conjunto de dados.

Modelos de Redes Neurais
Teoria da Convolução Unidimensional

O código foi escrito e processado no Google Colaboratory. Ele foi o meio-termo para provar a hipótese que os sons dos áudios dos três alimentos são diferentes. Apesar de suas limitações no armazenamento de dados, a programação em nuvem traz uma maior praticidade em executar estruturas de aprendizado de máquina complexas como a de rede neural convolucional. A Biblioteca Tensorflow garantiu uma maneira fácil de construir as redes neurais, especialmente estruturas complexas como a CONV1D. Em poucas palavras, a CONV1D é um processo de multiplicação entre um conjunto de números da entrada e a de um filtro de igual tamanho, seguido da adição de cada resultado de multiplicação. Na classificação de áudio, um filtro não é suficiente, portanto, aumentar o número de camadas gradualmente garantiu uma melhor diferenciação entre os dados. No entanto, o uso de filtros cria uma matriz bidimensional cujo comprimento era o vetor FFT, e largura o número de filtros. Este processo reduziu o tamanho das matrizes de saída em relação às camadas de *bypass*.

O primeiro modelo proposto veio como inspiração de um trabalho que classificou as estruturas danificadas utilizando o som de impacto (DORAFSHAN; AZARI, 2020). O modelo proposto pelos autores aplicou a arquitetura CONV1D, cuja entrada foi um espectrograma de imagem. Neste estudo, a mudança proposta é o uso de blocos residuais que inicia o treinamento com o vetor FFT, em vez de uma imagem.

Aprendizado Residual

O processo de Aprendizagem Residual é definido empilhando várias camadas em um bloco e depois somando seu resultado com a primeira camada do bloco. Este procedimento evita a degradação da precisão do treinamento durante a construção de estruturas complexas. (HE et al., 2015). A Figura 1 explica como o bloco residual foi construído enquanto a Figura 2 demonstra a estrutura da rede. Havia 23 camadas no total, além disso, o conjunto não se tornou pequeno devido a aplicação de *padding* nos blocos residuais, o que aplica zeros no começo e no fim o vetor para manter a saída do neurônio igual a sua entrada.

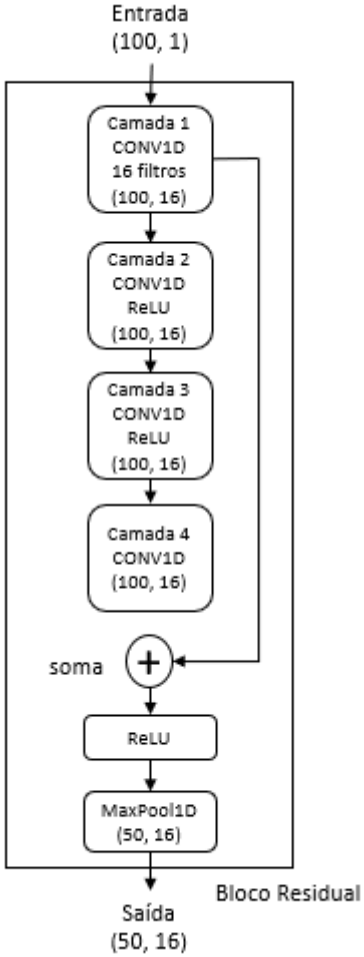


Figura 2: Construindo a Aprendizagem Residual.

Quanto aos hiperparâmetros, o ADAM com uma taxa de aprendizado de 0,001 e a *Categorical Crossentropy* foram as funções otimizadoras e de perda escolhidas para o modelo. O tamanho da batelada escolhido foi 50 e o *dropout* não foi aplicado. Este modelo foi

testado com duas abordagens diferentes. O primeiro teste utilizou um conjunto de dados de 584 FFTs e o último utilizou um conjunto de dados de 5840 FFTs. A última análise utilizou a técnica de *augmentation* para melhorar a capacidade de generalização do modelo

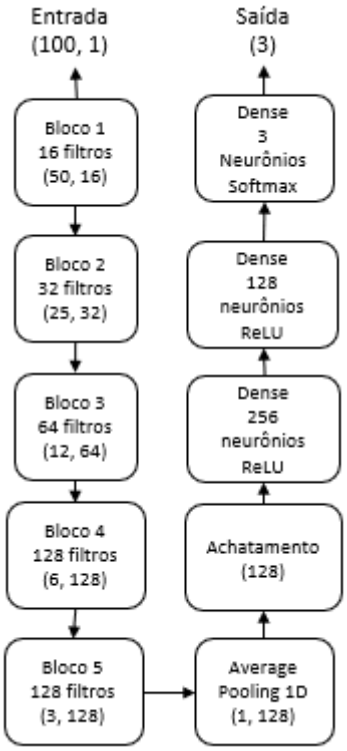


Figura 3: Estrutura da rede neural CONV1D para a entrada FFT.

Arquitetura dos MFCCs

No segundo modelo, a rede neural aplicada foi apenas uma rede profunda totalmente conectada com três camadas, isto se deve à natureza dos MFCCs de serem simples. A estrutura da rede teve 512 neurônios na primeira camada, depois 256 neurônios, e três na última. O *Leaky ReLU* proporcionou uma melhor resposta ao modelo enquanto o otimizador ADAM minimizou o overfit com uma taxa de aprendizado de 0,001. A normalização dos lotes tornou a rede neural mais estável quando usada antes da função de ativação. A *Categorical Crossentropy* funcionou bem como uma função de perda neste modelo de classificação. O *dropout* melhorou a generalização, tornando 20% de todos os dados no neurônio em zero de maneira aleatória. O conjunto de treinamento utilizou 467 áudios e 117 no conjunto de dados de validação.

Validação Experimental Complementar

Ensaios experimentais contendo 50 áudios de torradas adquiridos em ensaios de compressão moldaram a segunda validação. Elas foram compradas em um supermercado próximo e mordidos por um voluntário sob uma caixa à prova de som. O áudio foi captado por um microfone modelo bm800 e posteriormente tratado com redução de ruído no Software Audacity. Os modelos gerados anteriormente tiveram que avaliar corretamente os áudios capturados após cinco sessões de treinamento, ou seja, eles tiveram que predizer que os áudios capturados em laboratório eram torradas. Este método compara o modelo que seria preferível utilizar nos próximos estudos, além de simular esse estudo sendo desenvolvido em outros laboratórios.

RESULTADOS E DISCUSSÃO

Análise do pré-processamento

Os projetos de aprendizagem de máquina frequentemente começam com uma hipótese, um desafio ou uma curiosidade: Cada som crocante é diferente? Qual é a implicação de conhecer as características do som para projetos futuros? Isso é o que este estudo responde no final.

A escolha de utilizar três tipos diferentes de alimentos crocantes surgiu com o conceito de que cada alimento tem sua estrutura. Por exemplo, os áudios de frango frito vieram de países diferentes; portanto, as receitas e os ingredientes não são os mesmos, porém a estrutura gerada pela fritura é semelhante. Portanto, eles foram tratados como um grupo a ser classificado.

A base para a escolha das análises de pré-processamento veio de um estudo com áudios de cavalos, cujas, três principais características avaliadas foram os MFCCs, o *Beat* e o Tempogram (ALVES et al., 2021). Os cavalos demonstram um padrão rítmico em seu trote que torna possível a análise temporal. Em contra partida, o som causado pela mordida gera apenas um pulso de energia. Seu pico é

quando ambos os dentes se tocam. A figura 4 demonstra esta relação com um espectrograma filtrado para uma melhor visualização.

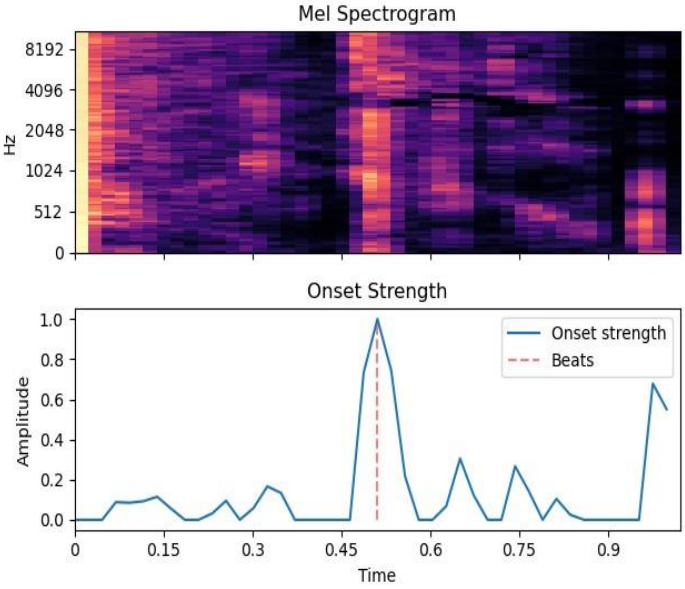


Figura 4: Relação entre a amplitude e a frequência por tempo em uma torrada.

O ritmo se desdobra através de vários pulsos iguais de energia dentro de um intervalo de tempo. O Beat é a representação do ponto médio de um pulso rítmico, no caso das torradas, há apenas um. Apesar de indicar os picos de amplitude, houve variações destes picos, dependendo do vídeo fonte. Não é prático padronizar o tempo de mordida, já que o intervalo de tempo do som variou com o alimento. As batatas fritas tinham um tempo médio de 0,7s para acontecer, torradas 1,2s, e fritas 1,3s. Como havia um número relevante de áudios fora do padrão de um segundo, a análise temporal usando redes neurais tornou-se inviável.

O domínio da frequência teve uma melhor visualização do comportamento dos áudios ao longo do tempo. Embora a amplitude não fosse clara, o espectrograma indicava uma possível diferença entre os áudios. O brilho de cada áudio se comportava de forma diferente, portanto havia possíveis variações na forma como o ruído crocante se propaga com o tempo. A figura 5 mostra o comportamento deste ruído, as diferenças são sutis, porém detectáveis.

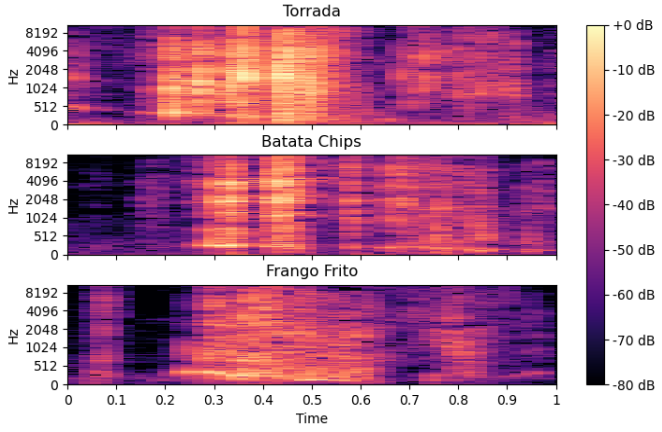


Figura 5: Espectrograma Mel de amostras de torrada, batata chips e frango fritos.

Hipoteticamente, esta diferença nos áudios se deve às características intrínsecas de cada alimento. Espessura e atividade da água foram as propriedades observadas que mais se destacaram durante os experimentos. As batatas fritas têm uma pequena espessura, enquanto as torradas têm uma maior, o que resultou em um som mais enérgico e duradouro. Quanto à atividade da água, uma torrada mergulhada no leite teve seu espectrograma com menor intensidade em comparação com uma torrada seca. As figuras 6 e 7 apresentam uma diminuição na energia resultante para a torrada úmida em comparação com a outra. Talvez seja possível quantificar esta variação em estudos posteriores.

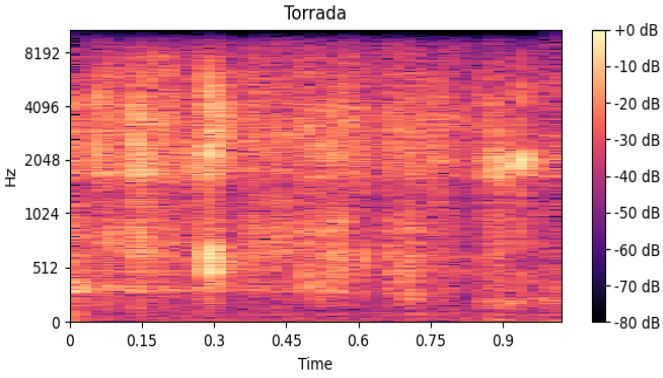


Figura 6: Espectrograma mel não filtrado de uma amostra de torrada.

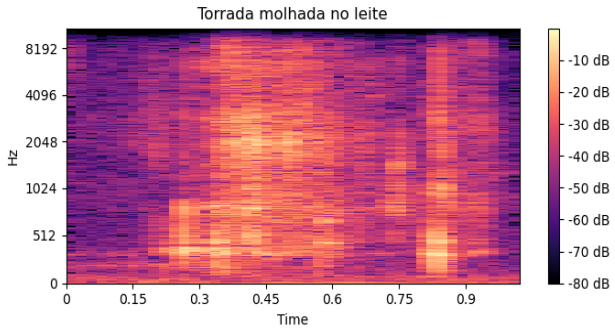


Figura 7: Espectrograma Mel não filtrado de uma amostra de torrada molhada no leite.

O processo de secagem remove a água dos alimentos, o que cria uma estrutura rígida cheia de ar. A energia gerada pelo som depende do ar para se propagar, portanto, quanto maior a energia, melhor a estrutura para fazê-la. Isto provavelmente explica a diferença entre frango frito e a torrada, a parte central do alimento contém uma maior atividade de água em comparação com a crosta. Quando a água está presente na estrutura, ela abafa o som. O domínio da frequência permitiu uma melhor visualização da intensidade em comparação com o domínio do tempo.

Análise do Comportamento da Rede Neural
Arquiteturas anteriormente testadas

Este estudo testou as arquiteturas de redes neurais existentes para encontrar a prova da hipótese anterior. O modelo AlexNet com as entradas de áudio produziu resultados abaixo do esperado com precisão de 42%. (KRIZHEVSKY et al., 2012). A utilização das imagens de espectrograma geradas pelo Librosa como entrada da rede CONV2D também não gerou resultados pertinentes. O conhecido modelo, GoogleNet, não se adaptou o suficiente para alcançar mais de 50% de precisão de validação. Estes resultados mostraram sinais de que havia uma diferença entre os sons, eles eram melhores do que um palpite aleatório. Em Machine Learning, as melhores soluções provêm, muitas vezes, de um redesenho completo dos modelos (SHUKLA et al., 2019).

Reimaginando imagens como um modelo linear

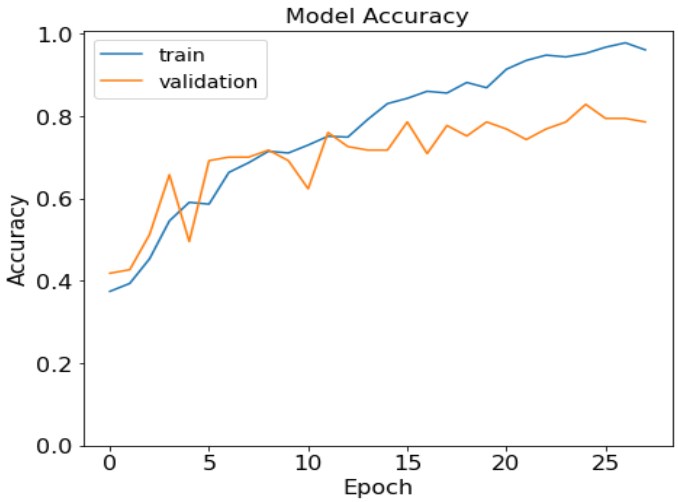
O FFT gerou os dados para moldar os espectros, portanto, substituí-los por esses dados lineares na rede proporcionaria resultados diferentes. Além disso, a rede neural foi modificada para um modelo que utilizava a base Resnet com convoluções unidimensionais. Assumindo uma entrada linear com 100 valores de FFT, o resultado da rede foi 85% de precisão, o que é suficiente para demonstrar que o som crocante de cada alimento é diferente. A entrada de 100 valores foi definida após mais experimentos, mudando-a entre 50, 150, e 200. Os resultados de precisão foram menores para 50, mas permaneceram os mesmos para 150 e 200 entradas. Este estudo continuou com 100 valores de entrada.

O fenômeno de sobreajuste também foi observado no resultado da rede FFT, após um alto número de épocas, um pico de erro de validação apareceu. Este fenômeno é comum no aprendizado de máquinas e significa que a rede neural está memorizando os dados de treinamento. O problema persistiu, independentemente da variação do número de neurônios. A obtenção de uma solução provou ser contraintuitiva, o modelo precisava de um aumento no número de entradas. A técnica de augmentation multiplicou o número de entradas por dez, este método resolveu o problema de ajuste e elevou a precisão do modelo a uma média de 97% após cinco treinamentos distintos do modelo.

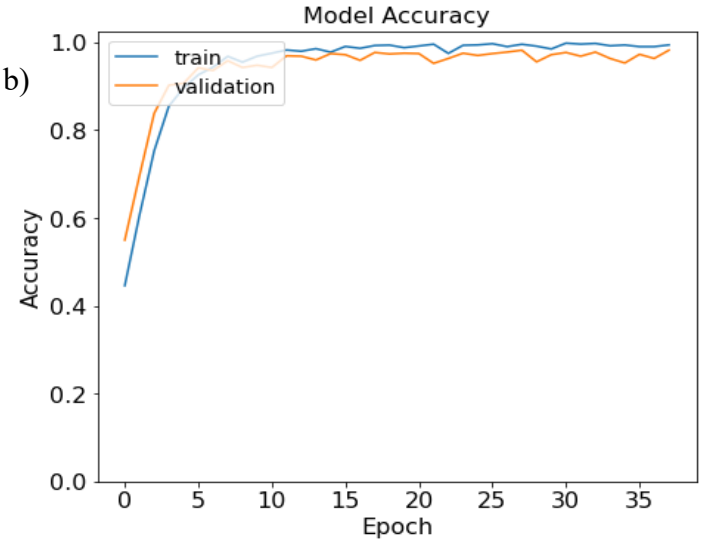
Comparando resultados com o modelo MFCCs

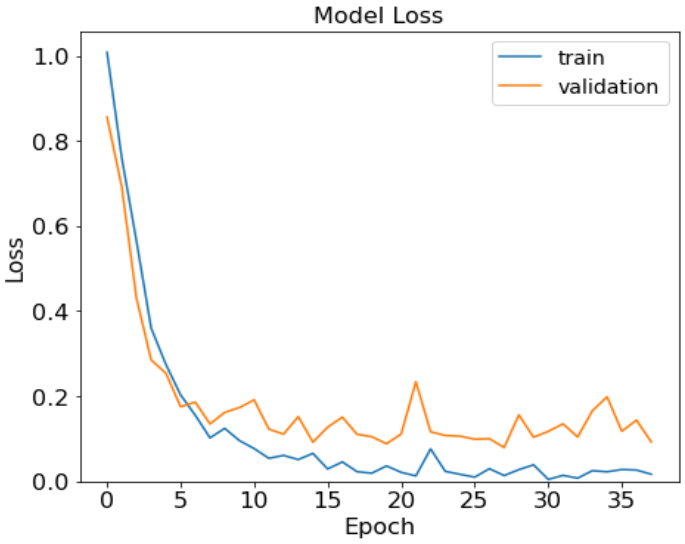
A convolução tem uma relativa complexidade de programação. Houve uma alta demanda de processamento por parte do Google Collaboratory. O desafio era encontrar uma forma que tivesse uma menor exigência e fosse fácil de replicar. O modelo do MFCC atendia às exigências por ser uma rede de três camadas totalmente conectada. O modelo simples atingiu um pico de precisão de 95% para validação, ou seja, teve um desempenho satisfatório. A Figura 8 compara os resultados dos três tipos de treinamento. Todos os resultados chegaram ao mesmo final, uma alta taxa de acerto do modelo sobre o que era cada tipo de alimentos.

a)

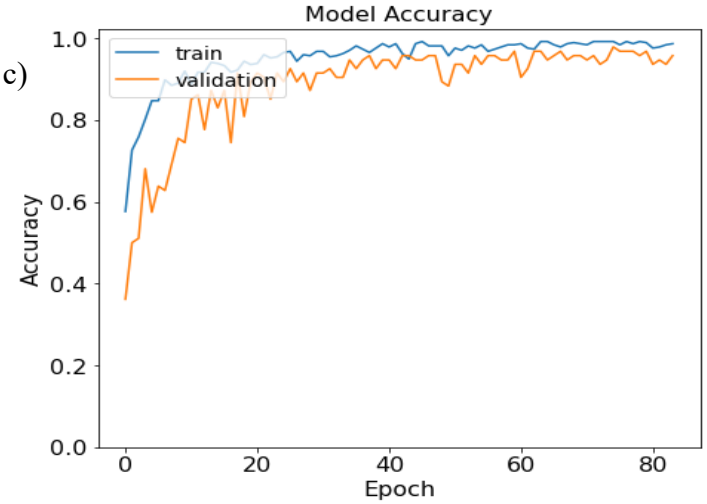


b)





c)



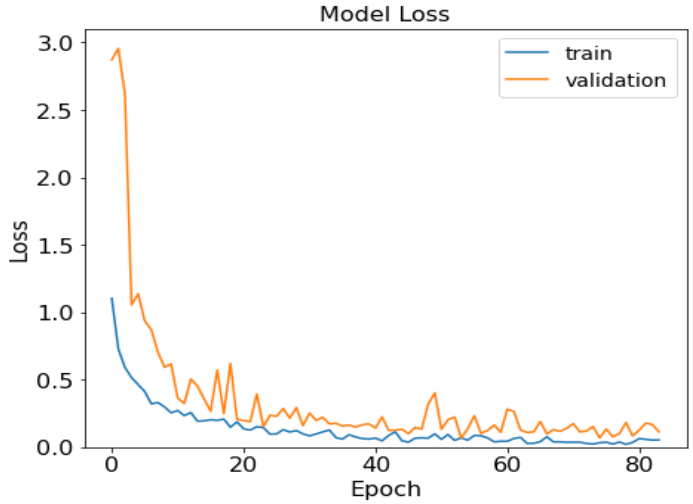


Figura 8: Acurácia (Accuracy) e perda (Loss) dos modelos *FFT* sem augmentation (a), com augmentation (b) e o *MFCCs* (c), respectivamente, pelo número de Épocas (Epochs).

O desempenho de uma rede neural também depende da perda do modelo, ou seja, da soma de todos os erros que o modelo teve na fase de avaliação. A rede neural trabalha para minimizar esta soma de erros, portanto, erros menores representam melhores ajustes. Para comparação, a perda de cada modelo mostrado na Figura 8 foi de 0,40, 0,09, e 0,21, respectivamente. O aumento trouxe esta melhoria de desempenho ao propor um aumento no número de entradas, portanto, a hipótese é que os outros modelos não alcançaram o melhor desempenho por não terem um número maior de dados de entrada.

Por outro lado, seguindo os dados adquiridos em um ambiente controlado, a segunda validação avaliou a generalização do modelo. Os cinquenta áudios tiveram uma origem diferente em relação aos áudios ASMR e, mesmo assim, o modelo adivinhou corretamente este lote de dados laboratoriais. Os modelos FFT divergiam dos resultados ideais, o modelo sem aumento tinha um comportamento estranho. Após cinco sessões de treinamento, a pontuação da avaliação variou entre 42 e 92%, não houve um resultado fixo para todas as situações. Este poderia ser o problema de *overfit*: o modelo não identificava padrões diferentes para a torrada e a fritura. Foi ainda pior para o modelo com *augmentation* com resultados inferiores a 20%. Em contraste, os MFCCs tiveram um desempenho superior a eles, depois de treinar o modelo cinco vezes, ele alcançou 100% de precisão nesta segunda validação. Mesmo assim, os áudios tiveram origens diferentes, dividindo-os em 64 MFCCs provaram ser melhores do que a análise dos dados do espectrograma. Os MFCCs dividiram o som em identidades que o formaram. Por exemplo, o som da fala é formado pelo pulso glótico que contorna o trato vocal. O programa os dividiu em duas identidades: o pulso e os timbres da fala. Em resumo, quaisquer mudanças no espectrograma poderiam trazer resultados divergentes, mas as variações não afetaram os MFCCs.

Seguindo as observações mencionadas, o melhor modelo para uma rede neural é aquele que é mais simples e exige menos esforço manual e computacional. O modelo do MFCC atende muito bem a estes requisitos. Pode-se afirmar que para a classificação de alimentos crocantes, o modelo do MFCC é superior aos outros apresentados. E mais uma coisa, ele alcançou resultados superiores em comparação com o modelo de aumento na segunda validação.

CONCLUSÃO

Este estudo demonstra a natureza simples do ruido crocante. Arquiteturas complexas podem não ser a melhor maneira de analisa-lo, embora sejam possíveis, elas precisam de mais tempo de processamento e poder computacional. Na análise de alimentos, o tempo é um fator decisivo, a velocidade de produzir um resultado generalizável permite impactar mais consumidores.

Os resultados abrem caminhos para pesquisas adicionais sobre as propriedades dos alimentos crocantes usando apenas o próprio ruído de fratura ou da mastigação. Os áudios aleatórios ASMR tiveram diferenças significativas entre si, o que foi validado através dos modelos de rede neural. Ambos os métodos MFCCs e FFT trouxeram a mesma resposta com mais de 95% de acurácia: cada som crocante é único. Portanto, este estudo propõe que não se pode generalizar uma análise sonora ou uma função que descreva propriedades as propriedades que sirva para todos os alimentos crocantes. O processo de otimização do ruído crocante deve ser feito respeitando a estrutura única de cada produto, não podendo ser generalizado para outros tipos.

AGRADECIMENTOS

Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001, e pela FAPESP (Fundação de Pesquisa de São Paulo) sob o subsídio 2021/05317-9. Agradecemos por seu apoio financeiro.

REFERÊNCIAS

AKIMOTO, H.; SAKURAI, N.; BLAHOVEC, J. A swing-arm device for the acoustic measurement of food texture. Journal of Texture Studies, n. September 2018, p. 104–113, 2018.

ALVES, A. A. C.; ANDRIETTA, L. T.; LOPES, R. Z.; BUSSIMAN, F. O; E SILVA, F. F.; CARVALHEIRO, R.; BRITO, L. F.; BALIEIRO, J. C. C; ALBUQUERQUE, L. G.; VENTURA, R. V. Integrating Audio Signal Processing and Deep Learning Algorithms for Gait Pattern Classification in Brazilian Gaited Horses. Frontiers in Animal Science, v. 2, 20 ago. 2021. Disponível em: <https://www.frontiersin.org/articles/10.3389/fanim.2021.681557/full>.

ANDREANI, P.; DE MORAES, J. O.; MURTHA, B. H. P.; LINK, J. V.; GIUSTINO, T.; LAURINDO, J. B.; PAUL, S.; CARCIOFI, B. A. M. Spectrum crispness sensory scale correlation with instrumental acoustic high-sampling rate and mechanical analyses. Food Research International, v. 129, p. 108886, mar. 2020. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0963996919307720>.

BUISSON, B.; SILBERZAHN, P. BLUE OCEAN OR FAST-SECOND INNOVATION? A FOUR-BREAKTHROUGH MODEL TO EXPLAIN SUCCESSFUL MARKET DOMINATION. International Journal of Innovation Management, v. 14, n. 03, p. 359–378, 20 jun. 2010. Disponível em: <https://www.worldscientific.com/doi/abs/10.1142/S1363919610002684>.

CHOLLET, F. Deep Learning with Python. 1st. ed. USA: Manning Publications Co., 2017.

DE MORAES, J. O.; MURTHA, B. H. P.; LINK, J. V.; GIUSTINO, T.; LAURINDO, J. B.; PAUL, S.; CARCIOFI, B. A. M.; ANDREANI, P. Mechanical-acoustical measurements to assess the crispness of dehydrated bananas at different water activities. LWT, v. 154, p. 112822, jan. 2022. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0023643821019757>.

DORAFSHAN, S.; AZARI, H. Deep learning models for bridge deck evaluation using impact echo. Construction and Building Materials, v. 263, p. 120109, dez. 2020. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0950061820321140>.

ELLIS, D. P. W. Beat Tracking by Dynamic Programming. Journal of New Music Research, v. 36, n. 1, p. 51–60, mar. 2007. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/09298210701653344>.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep Residual Learning for Image Recognition. 10 dez. 2015. Disponível em: <http://arxiv.org/abs/1512.03385>.

JI, C.; MUDIYANSELAGE, T. B.; GAO, Y.; PAN, Y. A review of infant cry analysis and classification. Eurasip Journal on Audio, Speech, and Music Processing. Springer Science and Business Media Deutschland GmbH, 2021.

KATO, S.; ITO, R.; WADA, N.; KAGAWA, T.; YAMAMOTO, M. Snack Food Texture Estimation by Neural Network. In: 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS), Anais...IEEE, 2018.

KATO, S.; ITO, R.; WADA, N.; KAGAWA, T.; SHIOZAKI, T.; NISHIYAMA, Y. Snack Texture Estimation System Using a Simple Equipment and Neural Network Model. Future Internet, v. 11, n. 3, 2019.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, Red Hook, NY, USA. Anais... Red Hook, NY, USA: Curran Associates Inc., 2012.

LEVITIN, D. J. This Is Your Brain on Music: The Science of a Human Obsession. New York, NY, USA: Plume Books, 2007.

MA, J. S.; GÓMEZ MAUREIRA, M. A.; VAN RIJN, J. N. Eating Sound Dataset for 20 Food Types and Sound Classification Using Convolutional Neural Networks. In: Companion Publication of the 2020 International Conference on Multimodal

Interaction, New York, NY, USA. Anais... New York, NY, USA: ACM, 25 out. 2020. Disponível em: <https://dl.acm.org/doi/10.1145/3395035.3425656>.

SHOTCUT. Versão 22.06. [S.l.]: Meltytech, LLC, 2022. Disponível: <https://shotcut.org/>. Acesso em: 20 Abril 2022

SHUKLA, R; LIPASTI, M.; VAN ESSEN, B.; MOODY, A.; MARUYAMA, N. REMODEL: Rethinking Deep CNN Models to Detect and Count on a NeuroSynaptic System. Frontiers in Neuroscience, v. 13, 22 fev. 2019. Disponível em: <https://www.frontiersin.org/article/10.3389/fnins.2019.00004/full>.

SPENCE, C. Sound: The Forgotten Flavor Sense. [s.l.] Elsevier Ltd, 2016.

TUNICK, M. H.; ONWULATA, C. T.; THOMAS, A. E.; PHILLIPS, J. G.; MUKHOPADHYAY, S.; SHEEN, S.; LIU, C.; LATONA, N. PIMENTEL, M. R.; COOKE, P. H. Critical Evaluation of Crispy and Crunchy Textures: A Review. International Journal of Food Properties, v. 16, n. 5, 2013.

VICKERS, Z. M. CRISPNESS AND CRUNCHINESS - A DIFFERENCE IN PITCH? Journal of Texture Studies, v. 15, n. 2, 1984.

ZHOU, D.-X. Universality of deep convolutional neural networks. Applied and Computational Harmonic Analysis, v. 48, n. 2, 2020.