

# FIA 2020/22

XII CONGRESSO/CONGRESO IBEROAMERICANO DE ACÚSTICA  
XXIX ENCONTRO DA SOCIEDADE BRASILEIRA DE ACÚSTICA - SOBRAC

Florianópolis, SC, Brasil

## Classificação de padrões sonoros ocorrendo em simultâneo com operação estacionária utilizando rede neural convolucional

Silva, L. C.<sup>1</sup>; Lenzi, A.<sup>2</sup>; Doi, R. M.<sup>3</sup>; Junior, O. S.<sup>4</sup>

<sup>1</sup> Laboratório de Vibrações e Acústica, Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil, <sup>2</sup> Laboratório de Vibrações e Acústica, Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil, arcanjo.lenzi@ufsc.br, <sup>3</sup> Nidec Global Appliance, Joinville, Brasil, ricardo.m.doi@nidec-ga.com, <sup>4</sup> Nidec Global Appliance, Joinville, Brasil, otavio.santini@nidec-ga.com

### Resumo

Muitos estudos e aplicações de redes neurais estão surgindo, muitas para classificação de sinais sonoros. Esses modelos são treinados de forma empírica, pois as redes neurais são consideradas caixas pretas. Alguns parâmetros são alterados e a validação permite inferir acerca da qualidade dos resultados dos modelos. Este trabalho avalia a influência da relação de energia entre sinais sonoros ocorrendo simultaneamente utilizando a rede neural convolucional. Buscou-se identificar qual é o mínimo de energia sonora necessária para uma correta classificação. Foram realizadas medições de pressão sonora do funcionamento estacionário de uma máquina. Sinais de batida metálica e gotejamento foram adicionados para simular a ocorrência de fenômenos anômalos. Antes da sobreposição, foi aplicada uma redução de energia a esses fenômenos. Isso possibilitou analisar a acurácia do modelo quando a anomalia apresenta uma menor energia sonora em relação à operação estacionária. A rede neural conseguiu identificar e classificar os padrões sonoros ocorrendo em simultâneo com o sinal estacionário gerado pela operação da máquina. Para o modelo categorizar com uma acurácia maior que 90%, a relação de energia entre a operação estacionária e os fenômenos de interesse pode ser até 20 dB global. Caso contrário, a melhor acurácia obtida foi 86%. Portanto, para classificar fenômenos simultâneos com a operação estacionária, os sinais sonoros necessitam ter uma relação de energia sonora de pelo menos 20 dB. Um trabalho futuro é verificar se mais amostras possibilitariam uma maior redução de relação de energia a qual a rede treinada consiga classificar corretamente os fenômenos. Dessa forma, é possível aperfeiçoar os modelos e responder à pergunta: quais as limitações dos modelos de inteligência artificial na classificação de padrões sonoros de ocorrência em simultânea a sinais estacionários como funcionamento de máquinas e ruído?

**Palavras-chave:** Aprendizado profundo, Classificação de padrões, Análise de sinais, Relação sinal ruído.

**PACS:** 43.60.Bf, 43.60.Cg, 43.60.Np.

### Classification of sound patterns occurring simultaneously with stationary operation using convolutional neural network.

#### Abstract

Numerous neural network studies and applications are emerging, with many focusing on the classification of sound signals. As neural networks are regarded black boxes, these models are trained empirically. Some parameters are altered, and the validation enables inferences regarding the quality of the models' outputs. Using a convolutional neural network, this study intends to assess the impact of the energy relationship between simultaneously occurring sound signals. We aimed to determine the minimal sound energy required for accurate classification. The sound pressure of a machine's stationary operation has been measured. The addition of metallic tapping and dripping signals simulates the occurrence of abnormal occurrences. Before the superposition, these events were subjected to an energy reduction. This enabled an analysis of the model's accuracy when the anomaly has a lower sound energy than stationary operation. The neural network was able to identify and classify the sound patterns that occurred simultaneously with the machine's stationary signal. For the model to classify with higher than 90 percent accuracy, the energy ratio between the stationary operation and the phenomenon of interest can be as high as 20 dB. Otherwise, the highest accuracy achieved was 86%. In order to identify events concurrently with stationary operation, the sound signals required a minimum sound energy ratio of 20 dB. Future research will determine if a larger sample size would result in a lower energy ratio at which the trained network can accurately label the phenomena. This allows us to develop the models and answer the question: what are the constraints of artificial intelligence models when identifying sound patterns that occur simultaneously with stationary signals, such as machine activity and noise?

**Keywords:** Deep Learning, Pattern Classification, Signal Analysis, Signal to Noise Ratio.



## 1. INTRODUÇÃO

A identificação de características importantes do comportamento acústico de sistemas dinâmicos possibilita a implementação de uma série de procedimentos, pois gera conhecimento sobre seu funcionamento acústico. Com base nesse entendimento, é possível desenvolver técnicas, aprimorar produtos e serviços ou dotá-los de recursos adicionais, além de orientar o desenvolvimento de ferramentas e a detecção de problemas[1]. Outra aplicação importante é a identificação de falhas ou condições atípicas de operação em máquinas[1, 2].

Existem várias técnicas de detecção e classificação de padrões sonoros[3], dentre elas, a inteligência artificial[4]. Essa é uma tecnologia relativamente nova que tende a apresentar bons resultados, entretanto carece de estudos. Como pesquisas sobre a categorização de padrões sonoros simultâneos, em particular, sinais que podem ocorrer durante o funcionamento máquinas. Por exemplo, gotejamento de fluidos, e choques metálicos, os quais indicam anomalia no sistema. Outra análise pouco explorada, é acerca da quantidade de energia necessária para esses fenômenos serem detectados. Isso porque ocorrem durante a operação estacionária e podem apresentar energia sonora extremamente baixa, tornando os fenômenos quase inaudíveis ao ouvido humano e imperceptíveis em inspeções por métodos tradicionais como análise espectral e níveis de ruído.

Há várias arquiteturas de redes neurais<sup>1</sup>, que podem ser escolhidas para gerar modelos de classificação. Dentre eles, este trabalho estudou a arquitetura convolucional, a qual foi desenvolvida para categorizar imagens. Entretanto, essa topologia de rede neural pode ser utilizada em outras aplicações. No caso de sinais de áudio, é possível representá-los por gráficos, tanto os de curva como os de cores, também conhecidos como espectrogramas, sendo é possível aplicar nos dados sobre análise. Foi devido seu histórico de acurácia na classificação de áudios que essa arquitetura foi escolhida.

O banco de dados utilizado, possuía sinais de operação de maquinário com e sem fenômenos de anomalia. Dentre os eventos anômalos escolhidos

estão o gotejamento e o choque metálico no sistema. Além deles, amostras de operação estacionária, sem a influência dos eventos foram obtidas. Esses dados passaram por um pré-processamento para reduzir ruído, e então métodos de extração de características foram aplicados. As informações obtidas foram com base em técnicas de processamento de sinais comumente estudadas em redes treinadas para classificar áudios de fala e música, isso porque não há a definição de qual utilizar em sinal estacionários. Assim, foram estudados o espectro cromático<sup>2</sup>, contraste espectral<sup>3</sup>, espectrograma em escala de mel<sup>4</sup>, coeficientes cepstrais da frequência Mel<sup>5</sup> e a centroide tonal<sup>6</sup>.

## 2. FUNCIONAMENTO REDE NEURAL CONVOLUCIONAL

A CNN<sup>7</sup>[5], é uma categoria de arquitetura de aprendizado profundo, amplamente utilizado na classificação de imagens. Quando aplicada no espectrograma dos sinais sonoros, consegue capturar padrões de modulação de energia ao longo do tempo e da frequência. Para isso, realiza extração e classificação de recursos de imagens através da aplicação filtros, os quais obtêm informações exclusivas dos mesmos dados. Esse processo é efetuado com a operação de convolução, a qual dá o nome a essa arquitetura de rede, sendo possível incluir tantos filtros quantos forem necessários. Assim, cada um deles auxilia o aprendizado de seus os pesos ideais durante o treinamento da rede.

A CNN, como outras redes neurais, pode ser treinada pelo algoritmo de retropropagação, e desenvolvida utilizando o método seque, o qual permite que o modelo seja construído camada por camada, uma após a outra. Seja em 1D, 2D ou 3D, as CNNs têm as mesmas características e usam a mesma abordagem. A principal diferença é a dimensionalidade dos dados de entrada e como o detector de recursos (ou filtro) se move através deles. Esse movimento é controlado pelo passo<sup>8</sup>,

<sup>2</sup>Em Inglês *chroma spectrogram*.

<sup>3</sup>Em Inglês *spectral contrast feature*.

<sup>4</sup>Em Inglês *mel-scaled spectrogram*.

<sup>5</sup>Acrônimo do inglês para *mel-frequency cepstral coefficients*.

<sup>6</sup>Em Inglês *tonal centroid (tonnetz)*.

<sup>7</sup>Acrônimo do inglês para *Convolutional Neural Network*.

<sup>8</sup>Em Inglês *Stride*.

<sup>1</sup>Forma de conexões entre os neurônios artificiais.

que dita a distância na qual o filtro se moverá.

Na arquitetura convolucional, além das camadas de convolução que, como dito, realiza o aprendizado de múltiplos filtros e extrai informações das imagens, são utilizadas outras camadas em conjunto. Dentre elas, *MaxPolling*, *Dropout*, *Flatten*, *Dense*, *Activation*, entre outras[6].

A camada *Activation* aplica uma função de ativação, que insere um comportamento necessário de não-linearidade. Uma possibilidade é a função Unidade Linear Retificada (ReLU<sup>9</sup>), que retorna 0 para todos os valores negativos, e o próprio valor para valores positivos. Ou seja, produz resultados no intervalo  $[0, \infty]$ . Também é possível utilizar a função Sigmoide, que retorna valores no intervalo entre 0 e 1 ou -1 e 1, sendo muito utilizada em modelos de classificação binária. Outra função comumente utilizada é a *Soft-Max*, que após ser aplicada cada elemento estará no intervalo de 0 a 1, e os elementos somarão 1. Dessa forma, eles podem ser interpretados como uma distribuição de probabilidade. Além disso, diferente da sigmoide, que consegue trabalhar com mais de duas classes, assim sendo utilizada na camada de saída da CNN.

A camada de *Pooling* ou Agrupamento, geralmente usa a função de valor máximo para processamento de imagens, mas também pode ser feito por função de soma ou de média. Ela age reduzindo ou agrupando píxeis de uma determinada região, com a finalidade de diminuir a variância a pequenas alterações, e também de reduzir a quantidade de parâmetros.

A camada *Dropout* define aleatoriamente alguns neurônios para desligar suas conexões durante o treinamento apenas. Isso é realizado por meio da porcentagem de desligamento, e feito para evitar *overfitting* no treinamento. Existe também a camada *Flatten*, redimensiona a saída das camadas convolucionais e *pooling* para se tornar em uma matriz linear, de uma única dimensão. Esta etapa é um preparo para se entrar na camada principal da rede neural totalmente conectada.

Também é possível adicionar camadas totalmente conectada ou *fully connected* nas redes convolucionais, e aplicar *Padding*, que basicamente estende a área de uma imagem, adicionando geralmente

zeros. Essa técnica é adicionada após as camadas convolucionais, pois essas reduzem o tamanho da saída ao aplicar os filtros com o método da convolução.

### 3. EXTRAÇÃO DE RECURSOS

Para a extração, foram escolhidos os métodos de espectrograma cromático, espectrograma em escala de mel, MFCCs[7][8], característica de contraste e Espectral[9], centroide tonal (tonnetz)[10]. O espectro cromático está relacionado ao tom, que pode ser pensado como uma medida de quão alta ou baixa é a frequência de um som. O banco de filtros croma projeta toda a energia contida no som gravado em 12 bins<sup>10</sup>, independentemente da oitava em que está localizada. Ao omitir uma oitava, pode-se criar um mapa de calor representando como a afinação muda temporalmente, que é um aspecto crítico da recuperação de informações musicais.

O centroide tonal é um vetor de recursos de seis dimensões projetado para detectar mudanças harmônicas em fontes musicais. Esta técnica é baseada na treliça harmônica ou Tonnetz, uma representação planar de conexões de alturas nas quais classes de notas com relações harmônicas próximas, como quintas ou terças maiores/menores, têm distâncias euclidianas reduzidas no plano. Alterações harmônicas, como limites de acordes de áudio musical, são descobertas calculando a distância euclidiana entre quadros de análise sucessivos de vetores centroides tonais.

A diferença em decibéis entre picos e vales no espectro de frequência é chamada Contraste Espectral. O quadro de um espectrograma é subdividido em sub-bandas. O contraste de energia é então calculado comparando a energia média no quartil superior (energia de pico) com a energia média no quartil inferior (energia do vale). Valores de contraste maiores que um geralmente correspondem a sinais claros de banda estreita, enquanto valores de contraste menores que um geralmente correspondem a ruídos de banda larga.

A escala Mel é uma transformação logarítmica da frequência de um sinal. Quando essa técnica é utilizada, as frequências tornam-se equidistantes.

<sup>9</sup>Abreviação para *Rectified Linear Unit*.

<sup>10</sup>Um bin é uma linha espectral definida pela Transformada Discreta de Fourier



Essa coerência é fundada na maneira como os humanos ouvem. O espectrograma Mel é criado ao combinar a Transformada de Fourier de Curto Tempo com as frequências da escala Mel[11]. Ou seja, os filtros Mel são multiplicados por cada janela de áudio gerada a partir do cálculo da FFT e sobrepostas. Assim, o número final de janelas do espectrograma é igual ao número de filtros de mel utilizados.

O cepstro, ou espectro de um espectro, pode ser determinado usando o espectrograma de Mel. Isso implica calcular a Transformada Discreta do Cosseno das energias do banco de filtros logarítmicos, que fornece a periodicidade do espectro. O objetivo é demonstrar como as frequências mudam rapidamente.

#### 4. MÉTRICAS DE DESEMPENHO

Após o treinamento, os diferentes métodos de aprendizagem devem ser avaliados para determinar seu desempenho. Para realizar a avaliação, algumas métricas comumente utilizadas em problemas de classificação foram utilizadas. As principais métricas de classificação estudadas neste trabalho são: matriz de confusão, acurácia, *recall*, precisão e *F1-Score*[12].

A matriz de confusão permite analisar de forma rápida o desempenho de cada sistema. Ela apresenta os resultados de classificação corretas em relação as incorretas. Assim, as possibilidades de resultados de classificação são:

- Verdadeiro Positivo (*True Positive* - TP): o modelo prediz corretamente a classe positiva;
- Verdadeiro Negativo (*True Negative* - TN): o modelo prediz corretamente a classe negativa;
- Falso Positivo (*False Positive* - FP): o modelo prediz incorretamente a classe positiva (erro tipo 1);
- Falso Negativo (*False Negative* - FN): o modelo prediz incorretamente a classe negativa (erro tipo 2).

Também é possível avaliar os modelos a partir das métricas de qualidade, popularmente usadas na

comparação de modelos de classificação. Cada uma delas busca avaliar um aspecto diferente do modelo.

A acurácia é definida pela quantidade de segmentos tanto da classe positiva da negativa classificados corretamente. Essa métrica deve ser tomada em consideração com cuidado caso a base de dados esteja desbalanceada, pois se existirem mais elementos em uma das classes o valor da acurácia pode ser muito alto.

A precisão é a capacidade de um modelo de classificação identificar apenas os pontos de dados relevantes, descrita matematicamente pela Equação 1. O objetivo é responder à pergunta: qual proporção de identificações positivas estava realmente correta? Assim, se um modelo específico não produz falsos positivos, terá uma precisão igual a 1,0.

$$\text{Precisão} = \frac{TP}{TP + FP}. \quad (1)$$

O *recall* é definido como a capacidade de um modelo encontrar todos os casos relevantes dentro de um conjunto de dados, e sua definição é dada pela Equação 2. Essa métrica busca responder à pergunta: qual proporção de positivos reais foi identificada corretamente? Ou seja, um modelo que não produz falsos negativos tem um *recall* igual a 1.

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (2)$$

Constantemente, há a necessidade de encontrar o equilíbrio entre *recall* e precisão. Para isso, utiliza-se uma combinação delas que é conhecida como *F1-Score*. Esse parâmetro é obtido através da média harmônica entre *recall* e precisão, conforme a Equação 3. A *F1-Score* informa o quão preciso é o classificador, ou seja, quantas instâncias classifica corretamente, bem como, o quão robusto é, isto é, se não perde um número significativo de instâncias. O valor mais alto possível de um *F1-Score* é 1,0, indicando precisão e *recall* perfeitos, e o valor mais baixo possível é 0, se a precisão ou a *recall* for zero.

$$F1-Score = 2 \cdot \frac{precisão \cdot recall}{precisão + recall} \quad (3)$$

## 5. NORMA ISO/IEC TR 29119-11

O teste de sistemas tradicionais é mais bem compreendido que os sistemas baseados em IA, os quais apresentam novos desafios. Portanto, é importante definir os critérios de aceitação e como testar sistemas baseados em IA. A ISO/IEC TR 29119-11<sup>11</sup> [13], explica as características específicas para sistemas baseados em Inteligências Artificiais. A norma também procura fornecer diretrizes sobre como testar os modelos sob teste, e esclarece as dificuldades correspondentes de especificar os critérios de aceitação para tais sistemas.

Existem diferentes métricas de desempenho utilizadas para avaliar diferentes modelos de aprendizado de máquina (ML). A ISO/IEC cobriu as métricas de classificação, que, normalmente, são empregadas em duas etapas no fluxo de trabalho de ML. Primeiramente, são aplicadas de modo a definir se há necessidade de ajustar as redes (por exemplo, pela seleção de hiperparâmetros) até que atinjam um nível aceitável de desempenho com seu conjunto de dados de avaliação. Subsequentemente, são usadas para medir a aceitabilidade da performance do modelo final com o conjunto de teste (independente).

Segundo a norma, os critérios de aceitação, (incluindo requisitos funcionais e não funcionais) devem ser documentados e justificados para uso na aplicação. As métricas de desempenho devem ser incluídas para o modelo. No mínimo, as características específicas de IA devem ser consideradas e podem ser usadas como base de uma lista de verificação usada para determinar a integridade dos critérios de aceitação para o sistema baseado em IA.

Os requisitos funcionais definem o comportamento básico do sistema, ou seja, o que fazem ou não devem fazer. Geralmente, definem os comportamentos e incluem cálculos, entrada de dados e processos. Em outras palavras, se os requisitos funcionais não forem atendidos, o sistema não funcionará. Já os requisitos não funcionais espe-

cificam como o sistema deve fazê-lo sem afetar a funcionalidade básica do sistema (por isso o nome, requisitos não funcionais). Mesmo que eles não sejam atendidos, o sistema ainda executará sua finalidade básica.

Portanto, para definir as condições para aceitar os modelos, é necessário levantar os requisitos funcionais, e não funcionais com as métricas de desempenho já abordadas. Como os requisitos são, por si só, um estudo para área, este trabalho focou-se nas métricas de desempenho. Dessa forma, foi estabelecido como critério ultrapassar a marca de 95% nos parâmetros estudados (acurácia, precisão, *recall* e *F1-Score*). Portanto, como este documento é o desfecho de uma pesquisa, os dados que não satisfazem esse critério serão apresentados.

É importante salientar, que esse processo deve ser definido antes de iniciar o projeto, e delimitado de acordo com a necessidade do problema. Durante o desenvolvimento dos modelos, o(s) desenvolvedor(es) devem se ater em satisfazer o parâmetro pré determinado.

A norma aqui abordada foi referência para todo o trabalho, visto que descreve não só o funcionamento da matriz de confusão e das métricas citadas. Ela também, expõem definições básicas de ML, assim como métodos avançados de teste e operação dos algoritmos.

## 6. REDE NEURAL DESENVOLVIDA

As primeiras quatro camadas dessa rede neural foram convolucionais com funções de ativação RELU, com, respectivamente, 16, 32, 64 e 128 neurônios. Seus filtros foram definidos usando uma matriz quadrada 3x3 com passo de (1,1), e *padding*, o qual garante que a saída tenha a mesma dimensão que a entrada.

Para reduzir a dimensão dos dados e do filtro, a camada a seguinte foi a *Max Pooling* 2D de tamanho 2x2, *Stride* (2,2) e *padding* zero. Então, para evitar o *overfitting* dos dados, uma camada *Dropout* foi adicionada. O *Dropout* foi definido com uma probabilidade de abandono de 50%, resultando em um alto grau de coadaptação entre as unidades da camada oculta.

Entre as camadas convolucional e densa há uma camada *Flatten* que atua como uma conexão entre

<sup>11</sup> *Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems*



as camadas convolucional e densa. Como resultado, ele conecta todas as camadas mencionadas anteriormente às duas camadas densas seguintes. O primeiro contém 128 neurônios, enquanto o segundo contém 64 neurônios configurados para ativação de ReLu. Por fim, a camada final é densa, com o tamanho de saída igual ao número de classes e a função de ativação *soft-max*. O modelo faz então uma previsão com base na opção provável.

O Quadro 1 ilustra algumas camadas do modelo. É possível realizar uma análise das dimensões de cada camada e dos parâmetros adquiridos durante cada etapa. Mais de 1.600 pontos de aprendizado estão disponíveis para otimização neste caso.

Model: "sequential"

| Layer (type)               | Output Shape        | Param # |
|----------------------------|---------------------|---------|
| conv2d (Conv2D)            | (None, 1, 193, 16)  | 160     |
| conv2d (Conv2D)            | (None, 1, 193, 32)  | 4640    |
| conv2d (Conv2D)            | (None, 1, 193, 64)  | 18496   |
| conv2d (Conv2D)            | (None, 1, 193, 128) | 73856   |
| max_pooling2d (MaxPooling) | (None, 1, 97, 128)  | 0       |
| dropout (Dropout)          | (None, 1, 97, 128)  | 0       |
| flatten (Flatten)          | (None, 12416)       | 0       |
| dense (Dense)              | (None, 128)         | 1589376 |
| dense (Dense)              | (None, 64)          | 8256    |
| dense (Dense)              | (None, 3)           | 195     |

Total params: 1,694,979  
Trainable params: 1,694,979  
Non-trainable params: 0

**Quadro 1:** Resumo da arquitetura da rede neural utilizada.

## 7. RESULTADOS

Os resultados de desempenho dos modelos treinados em sinais de gotejamento e choques metálico serão apresentados com a avaliação do que acontece com a precisão quando a energia desses sinais é reduzida. Foram avaliadas 4 relações de energia sonora, entre elas: sem alteração (0 dB), e com reduções de 6, 12 e 24 dB. Sem mudar o volume dos sinais, a operação estacionária apresenta um nível de 0 dBFS, já os sinais de gotejamento e choques metálicos, respectivamente -3 e -11 dBFS. Quando se descreve que foi reduzida energia sonora dos eventos, esta descrevendo o quanto foi retirado de volume dos sinais de gotejamento e choques metálicos.

Para cada uma dessas variações de amplitude do sinal, os modelos com a arquitetura descrita na seção 6 foram treinados. Isso permitiu a comparação dos resultados e a estimativa da relação de menor energia sonora entre os fenômenos e a operação estacionária. Como resultado, foi viável validar as restrições dos modelos em termos de energia sonora dos eventos.

No total, foram geradas aproximadamente 4.170 amostras de entrada, com 1.390 áudios para cada classe, que podem ser classificadas em duas categorias (gotejamento e choques metálicos) ou três categorias (gotejamento e choques metálicos e operação estacionária). Esta terceira classe foi incluída para verificar o desempenho dos modelos quando o sinal com maior energia é adicionado como uma categoria a se classificar.

Lembrando que a relação de energia proposta para os fenômenos em relação ao sinal estacionário, foi aplicado de modo a aproximar do comportamento acústico do mundo real desses dispositivos, visto que operam simultaneamente, e, principalmente no caso do gotejamento, ocorre com volume sonoro muito baixo.

A Quadro 1 compara as precisões de quatro modelos, três reduções de potência de som e duas possibilidades de classe.

|          | Redução de energia | 2 classes | 3 classes |
|----------|--------------------|-----------|-----------|
| Acurácia | 0 dB               | 100%      | 100%      |
|          | 6 dB               | 100%      | 99%       |
|          | 12 dB              | 95%       | 99%       |
|          | 24 dB              | 65%       | 97%       |

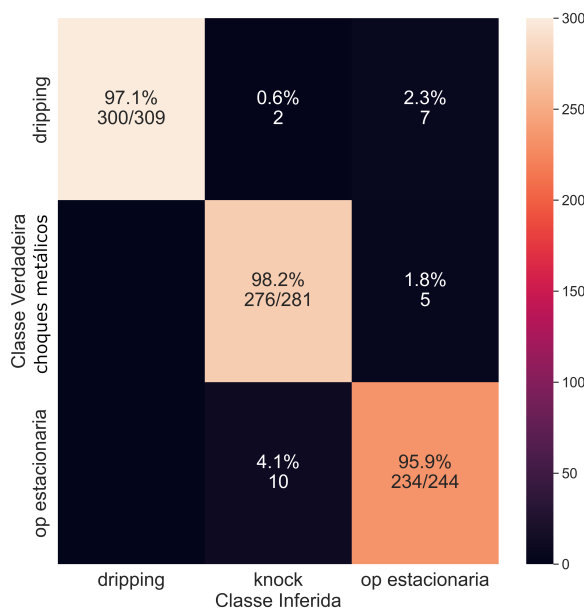
**Figura 1:** Precisão obtida pela redução da energia sonora para duas ou três classes diferentes.

Os resultados obtidos sem redução de energia, 0 dB, com duas ou três classes demonstram a diferença ao introduz uma classe operação estacionária, isso porque as classes gotejamento e choque metálico apresentam ocorrem junco com a operação do maquinário, ou seja, estes sinais possuem a reposta estacionária do sistema. Ao trabalhar com duas classes, a rede apresenta 100% de acurácia independente da redução de energia, neste caso, é necessário obter sinais com relação de amplitude maior entre os eventos e a operação estacionária para testar o comportamento da rede em uma situação mais complexa.

Quando treinados para classificar entre três clas-

ses, a rede apresenta uma acurácia de 100% sem redução de energia. Ao reduzir 6 dB, a diferença de acurácia não é tão significativa, passando a 99%. Reduzindo 12 dB este resultado vai para 99%, e se for removido 24 dB da amplitude do sinal, a acurácia fica em 97%. Ou seja, adicionando a terceira classe, as arquiteturas conseguem inferir com precisão de até 95% em todas as condições estudadas.

Para simplificar a análise dos resultados, será apresentado a seguir os resultados da matriz de confusão, o relatório de classificação, e as curvas de perda e acurácia ao longo das épocas; somente para o modelo treinado para classificar 3 classes com redução de energia de 24 dB, sendo este o resultado mais complexo aplicado ao algoritmo desenvolvido. A Figura 7 mostra a matriz de confusão resultante dessa situação. Nela é possível visualizar os resultados de *recall* para cada um dos áudios estudados.



**Matriz de confusão** A rede CNN 2D foi treinada usando sinais que continham uma redução de 24 dB na intensidade do som.

A matriz de confusão demonstra que os sinais de gotejamento provocaram maior grau de inferência em outras classes, uma explicação para isso é estas amostras possuírem baixo nível de energia, ao ponto do modelo não conseguir identificar seus áudios, confundindo-os com outras classes. Isso é suportado pela ocorrência de falsos negativos, ou classificações errôneas de gotejamentos.

O modelo classificou dez amostras de choques

metálicos como operação estacionário e dois como gotejamento, sendo essa outra indicação de que o desempenho do modelo é influenciado pela baixa energia dos sinais. Tanto o áudios de gotejamento quanto os de choques metálicos foram “confundidos” com operação estacionária, pois o sinal estacionário é comum para os dois outros fenômenos, desta forma, na falta de energia desses eventos sobra somente a operação estacionária como energia sonora. Esses achados são corroborados pelo relatório de classificação apresentado na Figura 2.

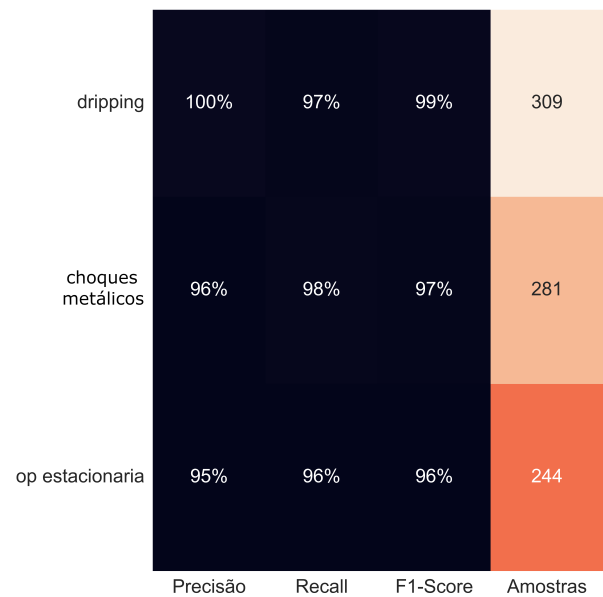


**Figura 2:** Relatório sobre a classificação da rede CNN bidimensional treinada com sinais que reduziram a energia sonora nos fenômenos em 24 dB.

Foram utilizadas 309 amostra de gotejamento, 281 de choques metálicos, e 244 de operação estacionária, logo às classes estão desbalanceadas. Nesse caso, os resultados de precisão, *recall*, e *F1-score* são mais significativos do que os de acurácia, os quais são discutidos a seguir.

Como dito anteriormente, a classe gotejamento teve o maior percentual de falsos negativos. Consequentemente, os sinais desta categoria foram inferidos incorretamente, resultando em um *recall* de 97%. No entanto, a recíproca não ocorreu, indicando ser perfeitamente precisa. A precisão foi de 96 e 95% para as categorias de operação de choques metálicos e estacionária, respectivamente, e 98 e 95% para *recall*. Estes resultados são corroborados ao avaliar os dados de precisão do modelo, que mostra que todos os sinais



que eram de gotejamento foram inferidos corretamente como gotejamento, já as análises de choques metálicos e a operação estacionária tiveram, respectivamente, 96 e 95% de acerto.

Como a precisão e o *recall* são um “cobertor curto”, ou seja, ao tentar diminuir um aumenta o outro, e vice-versa, para encontrar o equilíbrio entre os dois avalia-se então o *F1-Score*, que considera tanto os falsos positivos quanto os falsos negativos. Ao realizar a média da precisão e do *recall* para as 3 classes obtêm-se o seguinte resultado: 97% de precisão, e 97% de *recall*, portanto, eles apresentam valores altos e iguais. Dessa forma, espera-se um *F1-Score* alto e também igual a 97%. Em termos estatísticos, foi possível concluir dois aspectos acerca dos modelos:

- É possível que o modelo não encontre todos os positivos, mas aqueles que o modelo classifica como positivos são muito provavelmente corretos;
- O modelo consegue encontrar todos os casos positivos nos dados, mesmo que também possam identificar erroneamente alguns casos negativos como casos positivos.

A figura 3 mostra a evolução da perda e precisão do modelo CNN ao longo das épocas. Nela a linha azul representa o treinamento, enquanto a linha laranja denota o teste. A tendência deste resultado indica que não ocorreu *overfitting*, mas apresenta várias oscilações ao longo das iterações. É possível avaliar que, a curva de perda tende a zero enquanto a de acurácia se aproxima de um, entretanto elas não chegam a convergir totalmente. Uma forma de alterar esse comportamento é prolongar o período de treinamento, o que melhorará o desempenho do modelo.

## 8. CONSIDERAÇÕES FINAIS

Este trabalho teve por objetivo avaliar modelos treinados com bancos com variações de energia para aferir o comportamento dos algoritmos ao categorizar padrões sonoros de gotejamento, choques metálicos e operação estacionária com diferença de amplitude entre eles. Nessa situação o modelo desenvolvido com a arquitetura CNN nunca apresentou precisão inferior a 95%, valor

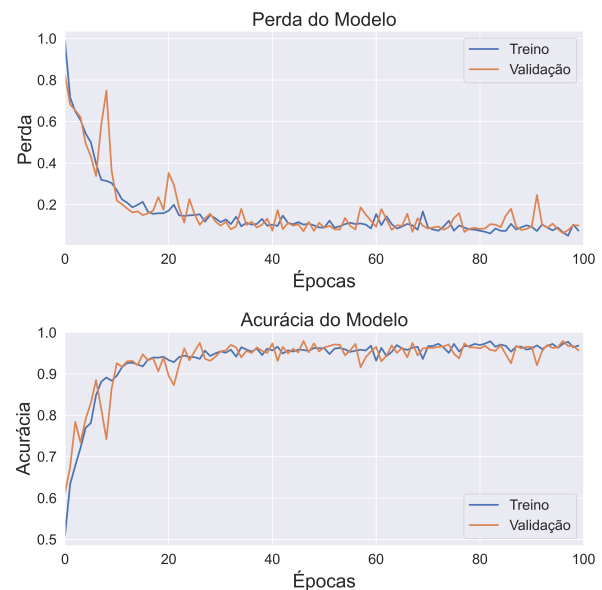


**Figura 3:** Curva de perda e precisão Uma rede CNN bidimensional foi treinada usando sinais que reduzem a energia sonora nos fenômenos em 24 dB.

estipulado com base na referência de dois sigmas de desvio padrão considerando distribuição normal. Para uma redução de 24 dB do volume dos sinais e considerando três sinais de análise, a acurácia foi de 97%, assim como a precisão e o *recall*.

Dos resultados da análise pode-se afirmar que o modelo pode não encontrar todos os dados positivos, mas aqueles que o modelo classifica como positivos estão provavelmente corretos. Entretanto, ele pode identificar erroneamente alguns casos negativos como casos positivos.

A partir dos estudos realizados, as sugestões para trabalhos futuros são:

- Realizar testes com outros padrões sonoros gerados por máquinas. Isso possibilita criar uma forma automática de detecção de padrões ou mau funcionamento do dispositivo.
- Avaliar outros métodos de extração, configurações e arquitetura de redes para aperfeiçoar o desempenho dos modelos. Dentre as possibilidades de otimização estão os hiper-parâmetros do algoritmo, o qual permite realizar estudos em relação ao problema proposto.
- Possibilitar uma estrutura de forma que as rede treinadas continuem aprendendo com

novos dados.

- Treinar modelos de aprendizado não supervisionado para inferir anomalias. Essa categoria de rede não precisa informar o que representa dada amostra. Portanto, é muito utilizada para detectar irregularidade quando não se sabe qual é o problema.

## 9. AGRADECIMENTOS

Agradeço a todos que participaram e ajudaram a execução deste trabalho.

## REFERÊNCIAS

- [1] Yang, Ruo-Yu e Rai, Rahul. Machine auscultation: enabling machine diagnostics using convolutional neural networks and large-scale machine audio data. *Advances in Manufacturing*, 7(2):174–187, 2019.
- [2] Purohit, Harsh; Tanabe, Ryo; Ichige, Kenji; Endo, Takashi; Nikaido, Yuki; Suefusa, Kaori e Kawaguchi, Yohei. Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. *arXiv preprint arXiv:1909.09347*, 2019.
- [3] Alías, Francesc; Socoró, Joan Claudi e Sevillano, Xavier. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, 6(5):143, 2016.
- [4] Purwins, Hendrik; Li, Bo; Virtanen, Tuomas; Schlüter, Jan; Chang, Shuo-Yiin e Sainath, Tara. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219, 2019. doi: [10.1109/JSTSP.2019.2908700](https://doi.org/10.1109/JSTSP.2019.2908700).
- [5] Ciaburro, Giuseppe e Venkateswaran, Balaji. *Neural Networks with R: Smart models using CNN, RNN, deep learning, and artificial intelligence principles*. Packt Publishing Ltd, 2017.
- [6] Sewak, Mohit; Karim, Md Rezaul e Pujari, Pradeep. *Practical convolutional neural networks: implement advanced deep learning models using Python*. Packt Publishing Ltd, 2018.
- [7] Garg, Utkarsh; Agarwal, Sachin; Gupta, Shubham; Dutt, Ravi e Singh, Dinesh. Prediction of emotions from the audio speech signals using mfcc, mel and chroma. Em *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, págs. 87–91. IEEE, 2020.
- [8] Pandey, Sandeep Kumar; Shekhawat, Hanu-mant Singh e Prasanna, SR Mahadeva. Deep learning techniques for speech emotion recognition: A review. Em *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, págs. 1–6. IEEE, 2019.
- [9] Zhang, Jingwen. Music feature extraction and classification algorithm based on deep learning. *Scientific Programming*, 2021, 2021.
- [10] Amlathe, Prakhar. *Standard machine learning techniques in audio beehive monitoring: Classification of audio samples with logistic regression, K-nearest neighbor, random forest and support vector machine*. Tese de doutorado, Utah State University, 2018.
- [11] Fayek, Haytham. Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between. <https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html>, Abril 2016. Acessado em 08 de Agosto 2019.
- [12] Brownlee, J. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery. Disponível em: <https://books.google.com.br/books?id=jaXJDwAAQBAJ>.
- [13] Becker, Nikolas; Junginger, Pauline; Martinez, Lukas; Krupka, Daniel e Beining, Leonie. Ai at work—mitigating safety and discriminatory risk with technical standards. *arXiv preprint arXiv:2108.11844*, 2021.