Detecting Mechanical Vibrations in Televisions via Audio Spectrogram Classification

Romulo Fabricio[1], Agemilson Pimentel[1], Ruan Belem[1], Anderson Sousa[2], Laura Martinho[2], Leo Araújo[3], Luan Silva[4] and Osmar Silva[2]
[1]TPV Technology Limited, Manaus-AM, Brazil
[2]Institute and Center for Development and Research in Software Technology - ICTS, AM-Brazil
[3]Federal University of Campina Grande (UFCG), PB-Brazil
[4]Federal University of Maranhão (UFMA), MA-Brazil
Emails: {romulo.fabricio, agemilson.pimentel, ruan.belem}@tpv-tech.com
{anderson.souza, laura.martinho, osmar.silva}@grupoicts.com.br
leo.araujo@ee.ufcg.edu.br
luan.souza@discente.ufma.br

***Abstract*—This paper presents a method for contactless detection of mechanical vibrations in televisions through audio spectrogram classification, utilizing Convolutional Neural Networks. The model was trained on a dataset containing simulated samples and demonstrated high accuracy, with excellent learning curves observed during training. In further evaluation with real samples the model performed well, achieving F1-Score rate of 99,02% in the test partition, confirming its potential for use in preventive maintenance processes and in addressing issues in televisions and other audio-dependent equipment, thereby enhancing the efficiency and quality of service.**

***Index Terms*—audio classification, anomaly detection, mechanical vibration, deep learning, convolutional neural network.**

I. Introdução

O setor de fabricação de televisores vivencia uma transformação significativa com a introdução da Indústria 4.0, que incorpora Inteligência Artificial (IA) e Automação Industrial, promovendo a criação de fábricas inteligentes e digitalizadas [1]. Esta era industrial caracteriza-se pela integração de sistemas físicos cibernéticos que otimizam o processo de manufatura, elevando sua eficiência graças à implementação de novas tecnologias. Apesar das oportunidades oferecidas pela Indústria 4.0, muitas linhas de produção de larga escala permanecem desatualizadas, resultando em ineficiências e falhas operacionais que precisam ser solucionadas para manter a competitividade no mercado globalizado.

Tendo em vista o desenvolvimento contínuo de novas tecnologias, espera-se que a produção de televisores apresente baixa taxa de falhas e defeitos, uma vez que esses dispositivos têm alto custo e demanda. Assim, torna-se necessária a aplicação de procedimentos de testagem para identificar eventuais produtos defeituosos [2]. Na etapa de montagem de televisões, problemas como parafusos soltos, componentes mal fixados ou inadequadamente posicionados e a falta de isolamento acústico podem causar vibrações sonoras audíveis, resultado da ressonância do material da TV com as frequências emitidas por seu autofalante, causando vibrações audíveis prejudiciais ao desempenho do produto. Nesse contexto, é fundamental a busca por novas tecnologias e abordagens para contornar este tipo de problema.

Neste artigo, é proposto um sistema de detecção de defeitos de montagem em televisores que identifica a presença de vibração mecânica em espectrogramas, utilizando técnicas de processamento de áudio e imagem, combinadas com Redes Neurais Convolucionais. A Figura 1 mostra um segmento de áudio e seu respectivo espectrograma.

O sistema será desenvolvido para uma linha de montagem com intenso ruído de fábrica com possibilidade de aplicação em televisores de diferentes modelos e tamanhos, que permitiria uma integração eficaz na linha de montagem. A realização deste tipo de teste não invasivo e automático contribui para manutenções preventivas, correção de problemas, redução de custos de fábrica e aumento da qualidade do produto final.

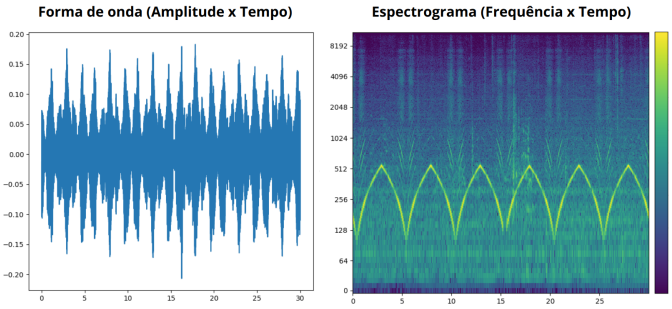

Figura 1. Segmento de áudio e seu respectivo espectrograma.

A proposta tem como foco a melhoria da detecção de defeitos em Tvs, mas também poderá ser expandida para otimizar a produção e garantir a qualidade de outros produtos eletrônicos. O modelo deve ter potencial para ser adaptado a outros tipos de equipamentos com estrutura semelhante que também sejam suscetíveis a problemas de vibração e ruído devido a falhas, como amplificadores e rádios. Em geral, a metodologia propõe uma capacidade de ser aplicada em qualquer aparelho onde a qualidade do som é essencial e a estrutura física pode causar ressonâncias indesejadas.

O artigo está estruturado da seguinte forma: na Seção II, são apresentados os trabalhos relacionados ao contexto da pesquisa. Na Seção III, apresenta-se o sistema proposto. Na seção IV estão os procedimentos experimentais, métricas utilizadas e os resultados obtidos. Por fim, na Seção V, as conclusões do trabalho.

II. Trabalhos Relacionados

No panorama da Indústria 4.0, observa-se um avanço significativo nas metodologias de inspeção de qualidade, essenciais para manter padrões elevados de produção. Pesquisas recentes têm explorado maneiras inovadoras de analisar os processos de áudio utilizando técnicas de visão computacional e estado-da-arte, que não só contribuem significativamente para a detecção e resolução de falhas na linha de produção, como também ampliam a precisão da análise de áudio [3].

Vários autores empregam estratégias na análise de qualidade industrial, como o método proposto por Villalba-Diez et al. [4], que utiliza Redes Neurais Profundas combinadas com câmeras ópticas de alta resolução para detectar defeitos durante a fabricação de cilindros de impressão. O resultado do trabalho evidenciou uma redução significativa nos custos de inspeção de qualidade. Esse tipo de abordagem ilustra a integração da IA na manutenção de processos e destacam seu potencial em refiná-los.

O estudo de Kastelan et al. [5] de 2011 propõe um sistema de verificação e teste em tempo real para avaliação automática da qualidade da imagem em televisores digitais. Esse sistema utiliza uma câmera digital para capturar o conteúdo da tela da TV, garantindo a avaliação da qualidade conforme percebida pelo usuário. Já o trabalho do mesmo autor em 2019 apresenta um conceito inovador de estimulação elétrica de telas sensíveis ao toque, visando automatizar completamente a verificação de dispositivos com essas telas na linha de produção final.

Métodos de ponta na visão computacional aplicados à análise de eventos sonoros têm mostrado grande eficácia no uso de espectrogramas. O conceito de *Eventness*, introduzido por Pham et al. em [6], refere-se à capacidade de detectar eventos de áudio, que é uma analogia ao *Objectness*, que na visão computacional se refere à capacidade de detectar objetos em imagens. Eventos de áudio aparecem como padrões bidimensionais de tempo-frequência em espectrogramas, apresentando texturas e estruturas geométricas específicas. Essa percepção permitiu tratar a detecção de eventos sonoros como um problema de detecção de objetos visuais.

No trabalho de Georges et al. [7] a classificação de cenas acústicas também é explorada, investigada usando espectrogramas para identificar automaticamente o ambiente de onde uma amostra de áudio foi originalmente gravada. Convertendo sinais de áudio em representações espectrais, foram extraídas características usando descritores de textura para examinar a complementariedade dos sinais dos canais de uma gravação de áudio estéreo.

A aplicação de técnicas de visão computacional em espectrogramas ressalta a capacidade de transformar problemas de áudio em análises visuais. Neste trabalho, é proposto um classificador de espectrogramas de áudio no contexto de falhas de montagem de televisores, que permite a interpretação mais precisa e automatizada de padrões complexos, alinhando-se com as inovações na análise de sinais e reforçando a importância de avanços tecnológicos na Indústria 4.0 para aprimorar a eficiência operacional.

III. Metodologia

A abordagem adotada neste artigo propõe um método de análise espectral para identificação de falhas em aparelhos televisores. O sistema foi projetado para funcionar em diferentes modelos e tamanhos, capturando amostras de áudio que posteriormente são transformadas em espectrogramas. As imagens dos gráficos espectrais são processadas usando a arquitetura de CNN *EfficientNet*, selecionada por sua eficiência em termos de velocidade de treinamento e uso de parâmetros, para identificar padrões de vibração que indicam possíveis defeitos de montagem.

A seguir, são apresentados alguns conceitos e metodologias utilizados no sistema proposto:

a) ***Sweep****:* Áudio padronizado que é reproduzido nos dispositivos durante os testes. Trata-se de um sinal senoidal cuja frequência cresce linearmente de 100 Hertz a 570 Hertz nos primeiros 2,5 segundos e decresce no mesmo intervalo. Essa faixa de frequência foi definida com base no sinal utilizado no processo existente na fábrica, que foi automatizado para maior eficiência [8].

b) ***Espectrograma****:* Representação visual da energia de um sinal de áudio, expressa como a função de frequência do sinal variando ao longo do tempo. As cores ou níveis de brilho no espectrograma indicam a intensidade das frequências, permitindo uma análise visual detalhada das características temporais e espectrais do sinal de áudio [9].

c) ***Convolutional Neural Network (CNN)****:* Rede Neural Convolucional é uma arquitetura de aprendizado profundo especializada no processamento de dados com estrutura espacial, como imagens, vídeos e sinais de áudio. Ela utiliza camadas convolucionais para extrair características relevantes dos dados e é frequentemente aplicada em tarefas de visão computacional [10].

d) ***Deep Learning (DL)****:* Aprendizado profundo é um método avançado de aprendizado de máquina que utiliza redes neurais com múltiplas camadas para modelar e aprender padrões complexos e hierárquicos em grandes conjuntos de dados [11].

e) ***EfficientNetV2****:* Tipo de rede neural convolucional com velocidade de treinamento mais rápida e melhor eficiência de parâmetros do que modelos anteriores, desenvolvida da combinação de pesquisa e dimensionamento de arquitetura neural com reconhecimento de treinamento para otimizar conjuntamente a velocidade de treinamento [12].

f) ***Grad-CAM****:* Mapa de calor que destaca as regiões da imagem que mais influenciaram a rede neural ao classificar um padrão, visualizando as áreas da imagem que foram mais significativas para a tomada de decisão [13].

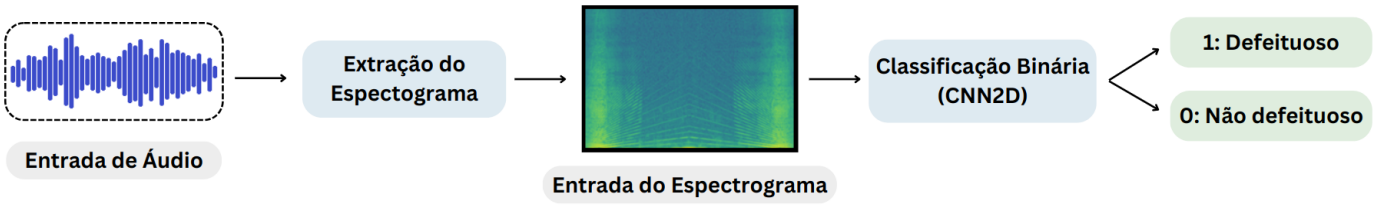

Figura 2. Diagrama de Blocos do Sistema.

A. Sistema Proposto

O processo de detecção de anomalias de áudio começa com a captura de uma amostra de som gerado pelo televisor em operação. O áudio é então transformado em um espectrograma, que visualiza as frequências presentes no sinal ao longo do tempo, aplicando-se uma Transformada de Fourier de Curto Prazo (STFT). A STFT decompõe um sinal em suas componentes de frequência ao longo do tempo, permitindo uma análise simultânea no domínio do tempo e da frequência, como mostra a Equação 1.

$$X(t,\omega) = \int_{-\infty}^{\infty} x(\tau)\, w(\tau - t)\, e^{-j\omega\tau}\, d\tau \tag{1}$$

onde $w(\tau - t)$ é a janela aplicada ao sinal para limitar a análise a um intervalo de tempo específico.

No estágio de pré-processamento, o espectrograma pode ser normalizado, filtrado e ter características extraídas para melhorar a eficiência da classificação. Em seguida, o espectrograma pré-processado é analisado pela CNN, que avalia as características visuais para identificar possíveis defeitos de montagem. O sistema, então, classifica o dispositivo como defeituoso ou não defeituoso, com base na análise do espectrograma. A Figura 2 apresenta o fluxo do sistema proposto.

Na análise espectral de áudio para identificar defeitos em equipamentos, não há um padrão específico no espectrograma que garanta a presença de falhas, devido à complexidade e variação dos sons gerados. Técnicas tradicionais, como a Transformada de Fourier, ajudam a revelar características no domínio da frequência, mas não são suficientes para classificar anomalias com precisão. As CNNs, por sua vez, aprendem diretamente dos dados brutos, identificando padrões complexos sem a necessidade de definições prévias. Treinadas com uma base rotulada, essas redes ajustam seus parâmetros para maximizar a precisão e generalizar para diferentes modelos e condições de funcionamento dos aparelhos.

B. Arquitetura da CNN

A *EfficientNetV2* é uma evolução das arquiteturas de CNNs, desenvolvida para ser altamente eficiente em termos de performance e consumo de recursos computacionais. Ela utiliza uma técnica chamada *compound scaling*, que otimiza simultaneamente a profundidade, largura e resolução da rede para maximizar a precisão com eficiência. A arquitetura básica é composta por blocos MBConv (*Mobile Inverted Bottleneck Convolution*) [14] com camadas SE (*Squeeze-and-Excitation*) [15], que melhoram a recalibração dos canais e a extração de características. Esses blocos são projetados para permitir um fluxo de informações mais eficiente, representado pela Figura 3. O funcionamento da *EfficientNetV2*, desde a entrada de dados até a classificação final, incluindo os aspectos fundamentais do seu treinamento e avaliação são processados da seguinte forma:

a) ***Entrada de Dados****:* A rede *EfficientNetV2* inicia o processo recebendo uma imagem de espectrograma como entrada. Esta imagem $\mathbf{X}$ possui dimensões $W \times H \times C$, onde W é a largura, H a altura, e C o número de canais.

b) ***Pré-processamento****:* O pré-processamento normaliza os dados de entrada e melhora a eficiência do treinamento, realizado conforme consta em Equação 2:

$$\mathbf{X}' = \frac{\mathbf{X} - \text{mean}(\mathbf{X})}{\text{std}(\mathbf{X})} \tag{2}$$

onde *mean* e *std* são calculados por canal. Além disso, o redimensionamento para um tamanho fixo necessário pela rede é geralmente feito usando interpolação bilinear.

c) ***Extração de Recursos (Backbone)****:* O *backbone* é composto por uma série de blocos MBConv, que realizam as convoluções e normalizações necessárias para a extração de características ao longo da rede. Em cada camada, os dados são processados com base nas saídas da camada anterior, conforme Equação 3:

$$\mathbf{Z}_i = f(\mathbf{W}_i * \mathbf{Z}_{i-1} + \mathbf{b}_i) \tag{3}$$

onde $*$ denota a operação de convolução, $\mathbf{W}_i$ e $\mathbf{b}_i$ são os pesos e viés da i-ésima camada, e f é uma função de ativação ReLU [16].

d) ***Extração de Características Globais****:* O *pooling* global [17] é empregado para sintetizar o mapa de características espaciais em um vetor mais compacto e gerenciável, de acordo com Equação 4:

$$\mathbf{y} = \mathit{GlobalPool}(\mathbf{Z}_{\text{final}}) \tag{4}$$

o *pooling* global transforma o mapa de características espaciais $\mathbf{Z}_{\text{final}}$ em um vetor de características $\mathbf{y}$.

*e) **Camada Totalmente Conectada (Classificação)**:* A fase final do processo de classificação envolve uma camada totalmente conectada que computa a probabilidade das classes, de acordo com a Equação 5:

$$\mathbf{p} = softmax(\mathbf{W}_c\mathbf{y} + \mathbf{b}_c) \tag{5}$$

onde $\mathbf{W}_c$ e $\mathbf{b}_c$ são os pesos e viés da camada de classificação.

*f) **Treinamento e Ajuste de Parâmetros**:* Durante o treinamento, o objetivo é ajustar os parâmetros da rede para minimizar a função de perda, tipicamente utilizando métodos de otimização modernos, como pode ser visto na Equação 6:

$$\theta^* = \arg\min_{\theta} L(\mathbf{p}, \mathbf{t}) \tag{6}$$

Minimização da função de perda $L(\mathbf{p}, \mathbf{t})$, pela entropia cruzada, utilizando o algoritmo de otimização Adam. Aqui θ representa todos os parâmetros ajustáveis da rede.

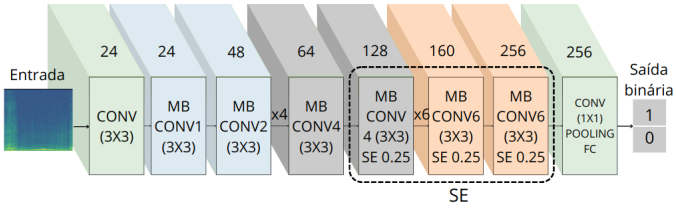

Figura 3. Fluxo da arquitetura da rede.

Após o processamento, o resultado final é um vetor de características que é gerado pela camada de classificação. Para cada classe possível, sendo a classificação binária 1 para defeituoso e 0 ausência de defeito, a *softmax* retorna uma probabilidade. A classe com a maior probabilidade é considerada a predição final do modelo, que identificou padrões de vibração e diferenciou as situações de montagem.

IV. Experimentos

Para desenvolver um sistema robusto para a detecção de falhas em televisores foi realizada uma extensa coleta de dados na linha de produção e em ambiente laboratorial, capturando condições normais e defeituosas. Foram testadas arquiteturas de *Deep Learning* na base de dados, com a *EfficientNetV2* se destacando nas métricas de desempenho. A robustez do modelo foi avaliada sob diferentes condições de ruído, utilizando métricas de desempenho e visualizações para constatar sua eficácia na proposta.

A. Base de dados

A base de dados é composta por porções de *Sweep*. Esse sinal é amplamente utilizado em testes acústicos e de áudio, pois permite avaliar como diferentes frequências são reproduzidas ou percebidas. Durante as coletas, o microfone foi posicionado na parte traseira da TV, à altura dos alto-falantes, para capturar tanto o som gerado pelo sinal *Sweep* quanto o ruído ambiente. As amostras foram capturadas com uma frequência de amostragem de 48 kHz. Como resultado, foram criadas duas bases de dados distintas: VibReal, contendo amostras coletadas em ambientes reais de produção, e VibSim, com dados simulados para complementar as análises.

B. VibReal

A base de dados VibReal consiste em um extenso conjunto de amostras reais, que foram coletadas tanto na linha de produção quanto em ambiente laboratorial, com o objetivo de capturar uma ampla gama de condições operacionais de televisores. Essas coletas foram realizadas em ambientes controlados para garantir a precisão dos dados. As amostras incluem registros de TVs sem defeitos, bem como de TVs com defeitos específicos. Entre as amostras de defeitos coletadas, estão problemas nas tampas traseiras, nos alto-falantes, nas placas internas e em cabos soltos. Ao todo, foram coletadas 7.032 amostras.

C. VibSim

A base de dados VibSim foi criada para aumentar a quantidade de dados disponíveis, simulando diferentes condições em relação ao ambiente de produção. Foi construída a partir de gravações realizadas com diferentes posições do microfone em relação aos televisores e em diferentes volumes. Para aumentar a complexidade e simular ambientes industriais, ruídos de fundo, como sons de fábrica, conversas e ruído branco, foram introduzidos em algumas gravações. Esse procedimento resultou em um total de 22.460 amostras, permitindo uma avaliação mais robusta do modelo em condições simuladas.

Os áudios coletados foram segmentados em porções menores, com uma sobreposição de 20% a 50% entre segmentos vizinhos, o que aumentou significativamente o número de amostras disponíveis. O dataset foi dividido em 42% para treino, 29% para validação e 29% para teste. A Figura 4 apresenta uma amostra do dataset VibSim em um televisor sem falhas no áudio, e a Figura 5 apresenta uma amostra do mesmo equipamento no dataset com defeito simulado de cabos soltos.

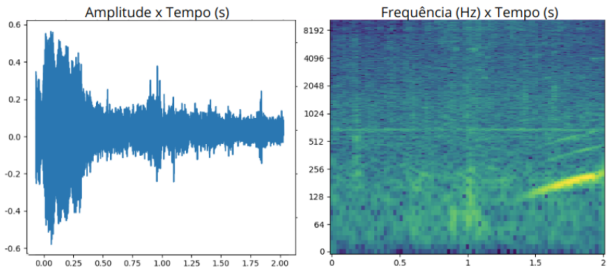

Figura 4. Amostra de aparelho sem defeitos no dataset VibSim.

D. Resultados

Após entender como os dados se comportam, a solução proposta para a detecção de defeitos de vibração em televisores foi desenvolvida utilizando *Deep Learning*, com uma base de dados composta por espectrogramas de áudio com amostras reais (VibReal) e simuladas (VibSim). Foram selecionadas cinco arquiteturas com o intuito de comparar qual delas teria a melhor desenvoltura na detecção de defeitos, sendo estas a

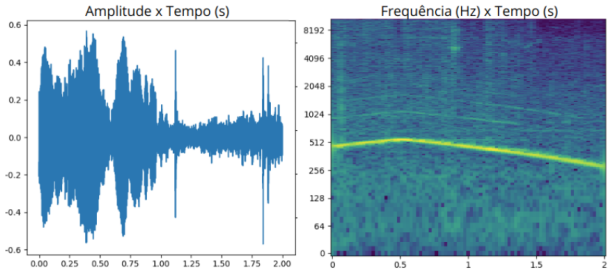

Figura 5. Amostra de aparelho com defeito no dataset VibSim.

ResNet18 [18], ResNet34 [19], MiniXception [20], Xception [21] e *EfficientNetV2* [12].

A *EfficientNetV2* apresentou o melhor desempenho, destacando-se na métrica de *F1-Score*. A escolha do *F1-Score* se deve ao seu balanceamento entre precisão e *recall*, que é fundamental para avaliar a eficácia do modelo na detecção de defeitos. Essa métrica é especialmente importante em contextos que focam em minimizar falsos positivos e falsos negativos. A rede mostrou-se robusta e adequada para a tarefa, conforme evidenciado na Tabela I.

Tabela I
RESULTADOS DE *F1-Score* DAS ARQUITETURAS TESTADAS

Arquitetura	VibSim	VibReal
ResNet18	94,82%	97,20%
ResNet34	95,54%	97,58%
Xception	96,15%	98,00%
miniXception	96,93%	98,35%
EfficientNetV2	98,80%	99,02%

Após o treinamento, a taxa de *F1-Score* da VibSim foi calculada para cada partição da base de dados, dividida em conjunto de treinamento, validação e teste. Esses resultados, apresentados na Tabela II, mostram que o modelo treinado obteve taxas de acerto próximas do ideal em todas as partições de acordo com a quantidade de amostras, demonstrando sua capacidade de generalização para os áudios com vibração simulada. Além disso, os resultados indicam que não há sinais de overfitting, reforçando a robustez do modelo ao lidar com diferentes condições de dados.

Tabela II
TABELA DE RESULTADOS *EfficientNetV2*

Métrica	Treino	Validação	Teste
F1-Score	99,82%	98,75%	99,02%

Os gráficos da Figura 6 mostram o desempenho do modelo ao longo das épocas de treinamento para o dataset VibSim. O gráfico de *F1-Score* reflete uma tendência similar, com valores elevados, mostrando que o modelo equilibra bem precisão e *recall*. O gráfico de perda mostra uma redução significativa na perda, indicando que o modelo está convergindo e minimizando erros, embora oscilações na validação possam sugerir alguma variação nos dados ou sensibilidade a certos exemplos. Esses resultados são bons e indicam que o modelo treina eficientemente, com bom potencial de generalização.

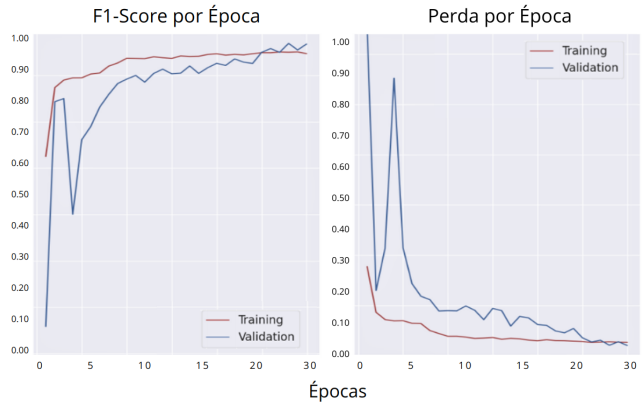

Figura 6. Desempenho do modelo em termos de *F1-Score* e Perda ao longo das épocas de treinamento

As matrizes de confusão fornecem uma visão clara do desempenho do modelo nos conjuntos de teste VibReal e VibSim, destacando a precisão e a taxa de erros. Na Figura 7, a matriz de confusão do teste do VibReal mostra que o modelo atingiu 98,4% de acertos para a classe sem defeito e 99,2% para defeito, mostrando uma taxa de erro muito baixa, com apenas 48 falsos positivos e 34 falsos negativos. No conjunto de teste de VibSim, a matriz de confusão também reflete uma alta acurácia, com 97,9% de acertos para a classe sem defeito e 99,4% para defeito. Apesar do aumento no número absoluto de erros devido ao tamanho maior do conjunto, o modelo minimiza falsos positivos e falsos negativos, sendo eficaz para detectar vibrações, tanto em cenários simulados quanto reais.

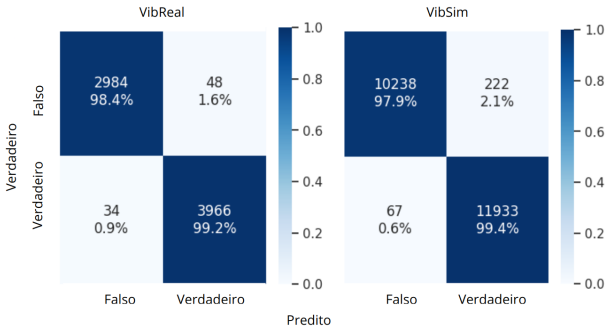

Figura 7. Matriz de confusão dos datasets VibReal e VibSim

A partir da matriz de confusão, é possível calcular e analisar métricas de desempenho importantes, como a Acurácia (ACC), Precisão (*Precision*), *Recall* e *F1-Score*. Essas métricas fornecem uma visão detalhada sobre a performance do modelo em prever corretamente as classes, permitindo avaliar tanto a capacidade do modelo em identificar corretamente as amostras positivas quanto em evitar falsos positivos. A Tabela III evidencia os resultados das métricas nos dois datasets.

Tabela III
MÉTRICAS DE DESEMPENHO PARA VIBREAL E VIBSIM

Dataset	Acurácia	Precisão	*Recall*	*F1-Score*
VibReal	98,80%	98,70%	99,15%	98,92%
VibSim	98,71%	98,17%	99,44%	98,80%

GradCAMs foram extraídos do modelo treinado para identificar defeitos de vibração que permitiram avaliar como a rede estava realizando suas inferências. Ao observar os mapas de calor gerados pelo GradCAM, foi possível visualizar as áreas do espectrograma nas quais a rede concentrou sua atenção ao realizar a classificação, que ajudam a entender os padrões que a rede neural está utilizando para tomar suas decisões, como mostram as Figuras 8, 9 e 10.

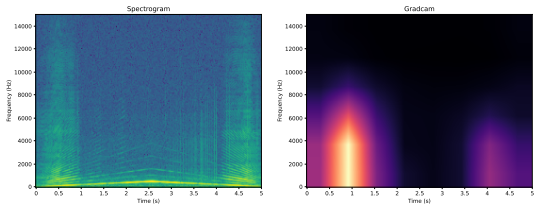

Figura 8. GradCAM de amostra de dispositivo defeituoso

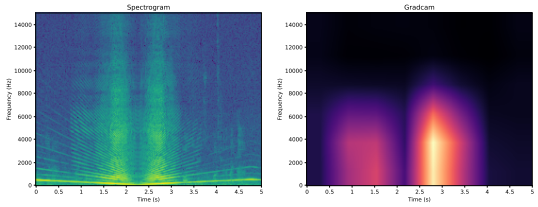

Figura 9. GradCAM de amostra em dispositivo com defeito simulado

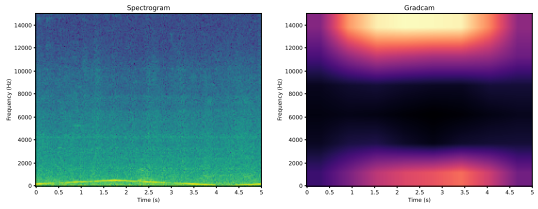

Figura 10. GradCAM de amostra em dispositivo sem defeito

V. CONCLUSÃO

Neste artigo, foi desenvolvida uma abordagem para a detecção automática de defeitos de vibração em televisores, utilizando técnicas de aprendizado profundo que não requerem contato físico com o dispositivo. O sistema foi projetado para operar eficientemente em ambientes com alto nível de ruído e se adapta a diferentes modelos de televisores, demonstrando versatilidade e precisão.

A robustez do sistema foi confirmada por meio de uma série de testes, utilizando tanto datasets reais quanto simulados. O modelo se destacou por sua capacidade de identificar corretamente as vibrações associadas a defeitos de montagem, apresentando resultados consistentes e de alta performance, mesmo em cenários complexos.

Para trabalhos futuros pretende-se expandir o dataset real, aumentando a diversidade e o volume das amostras coletadas, o que permitirá uma validação mais abrangente do modelo. Além disso, poderão ser exploradas novas técnicas de aprendizado profundo de última geração, com o objetivo de aprimorar ainda mais as métricas de desempenho e a capacidade de generalização do sistema.

REFERÊNCIAS

[1] L. Hughes, Y. K. Dwivedi, N. P. Rana, M. D. Williams *et al.*, "Perspectives on the future of manufacturing within the industry 4.0 era," *Production Planning & Control*, vol. 33, pp. 138–158, 2022.

[2] S.-H. Huang and Y.-C. Pan, "Automated visual inspection in the semiconductor industry: A survey," *Computers in industry*, vol. 66, 2015.

[3] R. L. Silva, M. Rudek, A. L. Szejka, and O. C. Junior, "Machine vision systems for industrial quality control inspections," in *Product Lifecycle Management to Support Industry 4.0*, 2018.

[4] J. Villalba-Diez, D. Schmidt, R. Gevers, J. Ordieres-Meré, M. Buchwitz, and W. Wellbrock, "Deep learning for industrial computer vision quality control in the printing industry 4.0," *Sensors*, vol. 19, 2019.

[5] I. Kastelan and M. Katona, "Automated optical inspection system for digital tv sets," *EURASIP Journal on Advances in Signal Processing*, 2011.

[6] P. Pham, J. Li, J. Szurley, and S. Das, "Eventness: Object detection on spectrograms for temporal localization of audio events," in *2018 IEEE ICASSP*, 2018.

[7] G. Z. Felipe, Y. Maldonado, G. d. Costa, and L. G. Helal, "Acoustic scene classification using spectrograms," in *2017 36th International Conference of the Chilean Computer Science Society (SCCC)*, 2017, pp. 1–7.

[8] W. Stefan, G. André, and L. Alexander, "Generalized multiple sweep measurement," *Journal of the Audio Engineering Society*, no. 7767, 2009.

[9] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," *arXiv preprint arXiv:1706.09559*, 2017.

[10] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," *CoRR*, vol. abs/2104.00298, 2021.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[12] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 096–10 106.

[13] R. R. Selvaraju, M. Cogswell, R. Das, Abhishek Vedantam *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE international conference on computer vision*, 2017, pp. 618–626.

[14] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," *CoRR*, 2018.

[15] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2019.

[16] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[17] "tf.keras.layers.globalaveragepooling2d: Large-scale machine learning on heterogeneous systems," 2015.

[18] G. K. Pandey and S. Srivastava, "Resnet-18 comparative analysis of various activation functions for image classification," in *2023 ICICT*, 2023.

[19] L. Gao, X. Zhang, T. Yang, B. Wang, and J. Li, "The application of resnet-34 model integrating transfer learning in the recognition and classification of overseas chinese frescoes," *Electronics*, 2023.

[20] C. Yu, G. Ding, and C. Yan, "Study on mini-xception- based improved lightweight expression detection model," in *International Conference on Control and Intelligent Robotics*, 2022.

[21] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE CVPR*, 2017.