UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA DE PRODUÇÃO

DETECÇÃO DE DEFEITOS EM MÁQUINAS DE FABRICAÇÃO DE COPOS DE PAPEL UTILIZANDO SINAIS ACÚSTICOS E DE VIBRAÇÃO COM ALGORITMOS DE APRENDIZADO DE MÁQUINA E REDES NEURAIS

PEDRO AUGUSTO MATELLI ANTUNES DE OLIVEIRA

ITAJUBÁ, FEVEREIRO DE 2026

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA DE PRODUÇÃO

PEDRO AUGUSTO MATELLI ANTUNES DE OLIVEIRA

DETECÇÃO DE DEFEITOS EM MÁQUINAS DE FABRICAÇÃO DE COPOS DE PAPEL UTILIZANDO SINAIS ACÚSTICOS E DE VIBRAÇÃO COM ALGORITMOS DE APRENDIZADO DE MÁQUINA E REDES NEURAIS

Dissertação submetida ao programa de Pós-Graduação em Engenharia de Produção como parte dos requisitos para obtenção do Título de Mestre em Ciências em Engenharia de Produção.

Área de Concentração: Engenharia de Produção

Orientador: Prof. Dr. Matheus Brendon Francisco

ITAJUBÁ, FEVEREIRO DE 2026

Resumo

A detecção precoce de defeitos em máquinas industriais é essencial para garantir elevada disponibilidade operacional, reduzir custos de manutenção e evitar perdas de produtividade, especialmente no contexto da Indústria 4.0. Este trabalho propõe e avalia uma abordagem para o diagnóstico de falhas em máquinas de fabricação de copos de papel, integrando sinais acústicos e de vibração como fontes de informação para a detecção de folga na Corrente 1 do sistema de transmissão, componente crítico para o sincronismo do equipamento. Os dados foram coletados em um processo real de produção em uma empresa do setor de embalagens, assegurando representatividade prática das condições operacionais. Foram realizadas 19 coletas experimentais, definidas a partir de um planejamento estruturado de experimentos (*Design of Experiments* – DOE), assegurando a variação sistemática das condições de funcionamento da máquina. Ao todo, foram obtidos aproximadamente 50 minutos de gravações. Os sinais adquiridos foram segmentados em janelas temporais de 5 segundos, com sobreposição de 50%, resultando em uma base de dados composta por 1.453 instâncias. A etapa de modelagem foi estruturada em duas vertentes complementares. Na primeira, adotou-se uma abordagem baseada em engenharia de atributos, na qual foram extraídos descritores estatísticos e espectrais dos sinais de áudio e vibração, utilizados como entrada para modelos estatísticos tradicionais (Regressão Logística e Análise do Discriminante Linear) e algoritmos clássicos de aprendizado de máquina, incluindo *Random Forest, Support Vector Machine*, Perceptron Multicamadas e um modelo de *Ensemble* por votação. Na segunda vertente, foi empregada uma abordagem de aprendizado profundo baseada em Redes Neurais Convolucionais (CNNs), aplicadas a espectrogramas Mel extraídos exclusivamente do sinal de áudio, permitindo o aprendizado automático de representações tempo-frequência relevantes para o diagnóstico de falhas. A avaliação dos modelos foi conduzida por meio de validação interna estratificada e teste externo composto por ensaios completamente inéditos, garantindo uma estimativa rigorosa da capacidade de generalização. Para mitigar efeitos estocásticos, os experimentos foram repetidos ao longo de múltiplas execuções independentes, sendo analisadas métricas médias de desempenho e respectivos intervalos de confiança. Os resultados indicam que a fusão multissensorial de áudio e vibração promoveu ganhos consistentes de desempenho e robustez em relação às abordagens unissensoriais, reduzindo significativamente a discrepância entre validação e teste. Os modelos baseados em atributos apresentaram os melhores desempenhos globais quando utilizadas informações multissensoriais, alcançando elevadas acurácias e maior

estabilidade estatística. A abordagem convolucional, embora restrita ao uso exclusivo do áudio, demonstrou melhor equilíbrio entre desempenho e generalização quando comparada aos modelos baseados apenas em atributos acústicos, evidenciando o potencial do aprendizado automático de representações para aplicações menos intrusivas. De forma geral, os resultados confirmam que a integração de múltiplas fontes sensoriais aumenta a robustez do diagnóstico de falhas, enquanto abordagens baseadas em aprendizado profundo representam uma alternativa promissora para cenários com restrições de instrumentação. A vertente baseada em engenharia de atributos alcançou acurácia média de 94% na melhor configuração multissensorial no teste externo. Já a vertente baseada em redes neurais convolucionais atingiu 92% de acurácia média, demonstrando desempenho competitivo mesmo em uma configuração menos intrusiva. Esses resultados evidenciam o potencial do método para aplicação em sistemas de manutenção preditiva e contribuem de forma quantitativa para o avanço do uso de técnicas de Inteligência Artificial no diagnóstico de falhas em máquinas industriais.

Palavras-chave: Detecção de falhas; Indústria 4.0; Inteligência Artificial; Monitoramento online.

Abstract

Early detection of defects in industrial machinery is essential to ensure high operational availability, reduce maintenance costs, and prevent productivity losses, particularly within the context of Industry 4.0. This study proposes and evaluates an approach for fault diagnosis in paper cup manufacturing machines by integrating acoustic and vibration signals to detect slack in Chain 1 of the transmission system, a critical component for equipment synchronization. Data were collected during a real production process in a packaging company, ensuring practical representativeness of operational conditions. Nineteen experimental acquisitions were conducted based on a structured Design of Experiments (DOE), ensuring systematic variation of machine operating conditions. In total, approximately 50 minutes of recordings were obtained. The acquired signals were segmented into 5-second windows with 50% overlap, resulting in a dataset comprising 1,453 instances. The modeling stage was structured into two complementary approaches. The first was based on manual feature engineering, extracting statistical and spectral descriptors from audio and vibration signals and using them as input to traditional statistical models (Logistic Regression and Linear Discriminant Analysis) and classical machine learning algorithms, including Random Forest, Support Vector Machine, Multilayer Perceptron, and a soft-voting Ensemble model. The second approach employed deep learning through Convolutional Neural Networks (CNNs) applied to Mel spectrograms extracted exclusively from the audio signal, enabling automatic learning of relevant time–frequency representations for fault diagnosis. Model evaluation was performed using stratified internal validation and an external test set composed of completely unseen experimental runs, ensuring a rigorous estimation of generalization capability. To mitigate stochastic effects, experiments were repeated across multiple independent executions, and average performance metrics along with their confidence intervals were analyzed. Results indicate that multisensory integration of audio and vibration signals significantly improved performance and robustness compared to single-sensor approaches, substantially reducing the gap between validation and external testing. The feature-based approach achieved an average external test accuracy of 94% in its best multisensory configuration, while the CNN-based approach reached an average accuracy of 92% using audio alone, demonstrating competitive performance even under less intrusive instrumentation. Overall, the findings confirm that multisource sensory integration enhances diagnostic robustness, while deep learning approaches provide a promising alternative in scenarios with instrumentation constraints. The proposed method shows strong potential for

predictive maintenance systems and quantitatively advances the application of Artificial Intelligence techniques in industrial fault diagnosis.

Keywords: Fault detection; Industry 4.0; Artificial intelligence; Online monitoring.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

1.1 Contextualização

A detecção de defeitos em máquinas industriais é um processo essencial para manter a eficiência operacional, reduzir custos de manutenção e prevenir paradas inesperadas que impactam severamente a produtividade e a qualidade (Zonta *et al*., 2020). Com o aumento da concorrência, o crescimento da automação e a complexidade dos sistemas produtivos, torna-se cada vez mais crucial que os equipamentos operem de forma confiável. Por isso, as indústrias buscam soluções mais precisas e proativas para maximizar a eficiência global de seus ativos (Cakir; Guvenc; Mistikoglu, 2021). Métodos eficazes de detecção de falhas desempenham um papel fundamental na estabilidade das operações, permitindo a identificação precoce de problemas e viabilizando intervenções preventivas com o mínimo impacto no processo produtivo.

A manutenção preditiva representa um avanço significativo na evolução da manutenção industrial, trazendo benefícios substanciais para as empresas e configurando-se como um dos principais desafios da Indústria 4.0 (Zonta *et al*., 2020). De acordo com Poór et al. (2019), mais de 55% das atividades de manutenção em instalações industriais ainda são reativas, ou seja, realizadas apenas após a ocorrência de falhas. Os autores destacam que a manutenção preditiva pode gerar economias de 12% a 18% nos custos de manutenção (podendo chegar a 40% dependendo do tipo de máquina, materiais e condições de operação), além de prolongar a vida útil dos ativos e reduzir o número de avarias. Ressaltam, ainda, que, quando implantados de forma eficaz, os sistemas de manutenção preditiva podem proporcionar um retorno sobre o investimento de até 10 vezes, bem como elevar a produção em até 25%.

Conforme apontam Zonta *et al*. (2020), historicamente a detecção de falhas esteve vinculada a abordagens tradicionais no que se refere à coleta e ao tratamento de sinais provenientes dos equipamentos em operação. De acordo com os autores, embora esses métodos tenham representado avanços em seus contextos, apresentaram também limitações práticas. Entre elas, destaca-se a dependência da intervenção humana na análise e interpretação dos dados, o que aumentava a suscetibilidade a erros e, em muitos casos, resultava em longos períodos de inatividade dos equipamentos. Essa condição restringia a antecipação de falhas e dificultava a tomada de decisões de manutenção em tempo hábil. Diante desse cenário, tornam-se necessários métodos mais eficazes e automatizados.

No contexto da detecção de falhas em ambientes industriais, a análise de sinais acústicos emitidos por máquinas durante a operação tem se consolidado como uma estratégia eficaz para

identificar alterações indicativas de problemas mecânicos, como desgaste, fissuras ou falhas estruturais (Tagawa; Maskeliūnas; Damaševičius, 2021). Essa técnica apresenta vantagens relevantes, pois é não invasiva, possibilita o monitoramento em tempo real e reduz a necessidade de interrupções no processo produtivo (Yao *et al*., 2021). Complementarmente, os sinais de vibração também são amplamente empregados na detecção de defeitos, sobretudo em máquinas rotativas, devido à sua sensibilidade a variações dinâmicas e anomalias mecânicas (Altaf *et al*., 2022). Estudos indicam ainda que a combinação de sinais acústicos e de vibração potencializa a eficácia dos modelos de diagnóstico, proporcionando resultados superiores em comparação ao uso isolado de cada tipo de sinal (Li *et al*., 2016, 2024; Praveen Kumar *et al*., 2019).

O avanço da Inteligência Artificial (IA) ampliou significativamente as possibilidades de análise, permitindo o desenvolvimento de aplicações capazes de identificar padrões de anomalia a partir de dados históricos. Algoritmos tradicionais, como *Random Forest* e *Support Vector Machines*, têm sido amplamente empregados nesse contexto, apresentando resultados consistentes em diferentes cenários industriais (Inturi et al., 2023; Raouf; Lee; Kim, 2022). A integração dessas tecnologias com sistemas de monitoramento em tempo real viabiliza estratégias de manutenção preditiva mais eficazes, permitindo a antecipação de falhas e a otimização dos processos de manutenção (Zonta *et al*., 2020).

Mais recentemente, técnicas de aprendizado profundo (*Deep Learning*) têm sido exploradas no contexto da detecção de falhas, especialmente por sua capacidade de aprender representações discriminativas diretamente a partir dos dados brutos ou de representações tempo-frequência, reduzindo a dependência de engenharia manual de atributos. Redes neurais convolucionais (*Convolutional Neural Networks* – CNNs) destacam-se nesse cenário por sua eficácia na análise de sinais e imagens, como espectrogramas, sendo capazes de capturar padrões locais e hierárquicos associados a diferentes condições operacionais. Entretanto, tais modelos demandam maior volume de dados, maior custo computacional e apresentam menor interpretabilidade quando comparados a métodos clássicos. Dessa forma, torna-se relevante investigar, de forma comparativa, o desempenho de abordagens baseadas em extração explícita de atributos e algoritmos de aprendizado de máquina tradicionais frente a modelos de aprendizado profundo, avaliando seus trade-offs em termos de acurácia, robustez e complexidade.

Na literatura, diversos trabalhos recentes têm demonstrado o potencial da fusão de sinais acústicos e vibracionais para o diagnóstico de falhas em máquinas industriais. Li et al. (2016, 2024) evidenciam que a combinação multissensorial aumenta a precisão da classificação em

comparação ao uso isolado de cada tipo de sinal. De forma semelhante, Praveen Kumar et al. (2019) aplicaram técnicas de fusão de dados em caixas de engrenagens, alcançando ganhos significativos na robustez do diagnóstico. Estudos como o de Yao et al. (2021) e Altaf et al. (2022) reforçam a eficácia individual de sinais acústicos e de vibração, respectivamente, enquanto pesquisas mais recentes têm integrado tais sinais a algoritmos de aprendizado de máquina, incluindo *Random Forest*, *Support Vector Machines* e redes neurais, obtendo resultados promissores em diferentes cenários industriais (Inturi et al., 2023). Esses trabalhos evidenciam a relevância da abordagem multissensorial e fundamentam a proposta desta pesquisa, que aplica e avalia tais técnicas em um contexto ainda pouco explorado: máquinas de fabricação de copos de papel.

O crescimento da demanda por embalagens sustentáveis tem impulsionado transformações significativas na indústria de alimentos e bebidas. Pressões ambientais, regulatórias e mudanças no comportamento do consumidor têm favorecido a substituição de materiais derivados de petróleo por alternativas renováveis, recicláveis e biodegradáveis. Nesse cenário, as embalagens à base de papel destacam-se como uma das principais soluções adotadas pelo setor, especialmente em aplicações de consumo rápido e descartável, consolidando-se como protagonistas na transição para sistemas produtivos mais sustentáveis (Adibi; Trinh; Mekonnen, 2023).

O aumento da demanda por produtos à base de papel implica a necessidade de processos industriais cada vez mais eficientes e confiáveis. A produção em larga escala de copos de papel requer equipamentos de alta precisão e sincronismo, nos quais falhas mecânicas podem comprometer a qualidade do produto e a continuidade operacional. Nesse contexto, estratégias de monitoramento e diagnóstico de falhas tornam-se fundamentais para garantir produtividade e competitividade no setor.

Esta pesquisa concentra-se em máquinas industriais de fabricação de copos de papel, amplamente utilizadas nos setores de alimentos e bebidas e fundamentais para atender à crescente demanda por embalagens sustentáveis. Esses equipamentos realizam a transformação de bobinas de papel em copos por meio de etapas sucessivas, como dobra, selagem, corte, colagem do fundo e formação da borda. Entre as falhas operacionais mais recorrentes destacam-se a folga em correntes, a perda de pressão no regravador e falhas no corte do fundo. O foco experimental deste estudo foi delimitado à detecção de falhas na Corrente 1 do sistema de transmissão, reconhecida como elemento crítico para o sincronismo da máquina. Os demais defeitos são mencionados apenas como parte do contexto geral.

No contexto desta pesquisa, são avaliados diferentes algoritmos de aprendizado de

máquina, cuja escolha fundamenta-se em suas características complementares e na ampla aplicação em problemas de diagnóstico de falhas, conforme detalhado na Seção 2.4. Esses algoritmos contemplam distintos paradigmas de aprendizado, incluindo métodos baseados em margens de separação, *Ensemble*s de árvores de decisão e redes neurais artificiais, permitindo uma análise comparativa sob diferentes perspectivas de desempenho, complexidade e robustez. As Máquinas de Vetores de Suporte (*Support Vector Machines* – SVM) destacam-se pela eficácia na separação de padrões sutis em conjuntos de dados de pequeno e médio porte e em espaços de alta dimensionalidade, embora apresentem sensibilidade à escolha do *kernel* e ao ajuste de hiperparâmetros. O *Random Forest*, por sua vez, é reconhecido pela robustez a ruídos e pela estabilidade de desempenho em cenários industriais com variabilidade operacional, ainda que apresente perda de interpretabilidade e aumento do custo computacional à medida que o número de árvores cresce. O *Extreme Gradient Boosting* (*XGBoost*) oferece elevado desempenho preditivo e capacidade de modelar relações complexas, porém requer ajuste cuidadoso de seus parâmetros para mitigar riscos de sobreajuste, especialmente em bases de dados limitadas. Complementarmente, o Perceptron Multicamadas (MLP) possibilita a captura de padrões não lineares mais complexos, ao custo de menor interpretabilidade e maior demanda computacional.

Adicionalmente, são avaliadas redes neurais convolucionais (*Convolutional Neural Networks* – CNNs), empregadas na análise de representações tempo-frequência dos sinais, as quais permitem o aprendizado automático de padrões discriminativos, reduzindo a dependência da engenharia manual de atributos. Por último, a utilização de um modelo de *Ensemble* por votação busca explorar a diversidade entre os classificadores baseados em atributos, combinando suas previsões individuais para aumentar a estabilidade e a capacidade de generalização do diagnóstico, embora implique maior custo computacional global.

Além dos algoritmos de aprendizado de máquina amplamente empregados na literatura recente, é fundamental considerar modelos estatísticos tradicionais como referência (*baseline*s) para avaliação de desempenho. Métodos como a Regressão Logística e a Análise do Discriminante Linear, embora apresentem menor complexidade computacional e capacidade limitada de modelar relações altamente não lineares, oferecem vantagens importantes, como maior interpretabilidade e fundamentação estatística consolidada. A inclusão desses modelos permite estabelecer um patamar mínimo de desempenho, contribuindo para uma análise comparativa mais rigorosa e transparente dos ganhos efetivamente proporcionados por abordagens mais sofisticadas, como os modelos de aprendizado de máquina e redes neurais artificiais

Nesse contexto, este trabalho adota uma abordagem comparativa estruturada, caracterizando-se como uma pesquisa experimental, de natureza quantitativa e com abordagem aplicada, conduzida em ambiente controlado a partir de dados reais de operação. São avaliados desde modelos estatísticos tradicionais até algoritmos de aprendizado de máquina e redes neurais profundas, de modo a analisar não apenas o desempenho preditivo, mas também aspectos como robustez, complexidade computacional e viabilidade de aplicação em ambientes industriais reais.

As contribuições esperadas deste trabalho incluem o desenvolvimento de um método robusto e eficaz para o diagnóstico precoce de falhas em máquinas industriais, promovendo maior eficiência operacional e reduzindo os impactos negativos sobre a produtividade e a qualidade dos produtos. A pesquisa não apenas oferece uma solução prática, escalável e potencialmente replicável em diferentes contextos produtivos, como também contribui para o avanço do conhecimento acerca da aplicação de técnicas de Inteligência Artificial no setor industrial, fortalecendo os processos de transformação digital e manutenção preditiva, dentro do contexto da indústria 4.0.

1.2 Objetivos

Diante disso, o presente trabalho tem como objetivo geral desenvolver e avaliar uma abordagem para a detecção de falhas em máquinas de fabricação de copos de papel, por meio da aplicação e comparação de algoritmos de aprendizado de máquina baseados tanto na extração explícita de atributos quanto no aprendizado automático de representações a partir de sinais acústicos e de vibração, visando aumentar a acurácia e a robustez do diagnóstico em diferentes condições de operação.

1.3 Limitações

Cabe destacar que o presente trabalho possui limitações relacionadas ao escopo experimental, à instrumentação empregada e à fase de aplicação prática do método proposto. A pesquisa foi conduzida em um único modelo de máquina de fabricação de copos de papel (PT80), com foco exclusivo na detecção de falhas associadas à folga na Corrente 1 do sistema de transmissão, não abrangendo experimentalmente outros subsistemas relevantes do equipamento. Os dados foram coletados sob condições controladas de simulação de falhas, o que pode não refletir integralmente a variabilidade de um ambiente industrial operando de forma contínua. Quanto à instrumentação, os sinais de vibração foram adquiridos por meio de

um dispositivo móvel, com taxa de amostragem limitada a 100 Hz, adequada para a análise de oscilações mecânicas de baixa frequência, mas insuficiente para a detecção de defeitos de alta frequência, que demandariam sensores industriais dedicados. Por fim, em consonância com a metodologia CRISP-DM, esta pesquisa não contempla a etapa de implantação do sistema em ambiente produtivo, restringindo-se ao desenvolvimento e à avaliação experimental dos modelos.

2 Fundamentação Teórica

2.1 Contexto da manutenção preditiva na Indústria 4.0

A manutenção preditiva, também denominada manutenção baseada em condição (*condition-based maintenance*), surgiu na década de 1980 como alternativa mais eficiente aos métodos reativos e preventivos (Cakir; Guvenc; Mistikoglu, 2021). Enquanto a manutenção reativa ocorre apenas após a falha e a preventiva segue cronogramas periódicos, a preditiva busca identificar anomalias antes que atinjam estágio crítico, possibilitando que a intervenção ocorra somente quando necessária. Essa abordagem resulta em redução direta dos custos de manutenção e aumento da confiabilidade dos equipamentos. Tal avanço está diretamente associado ao contexto da Indústria 4.0, entendida como a quarta revolução industrial e caracterizada pela integração de sistemas ciberfísicos, Internet das Coisas (IoT) e Inteligência Artificial (Zonta *et al*., 2020). Ao contrário das revoluções anteriores, a Indústria 4.0 promove a convergência entre o mundo físico e o virtual em tempo real, elevando os níveis de automação e confiabilidade dos processos. Suas principais características incluem conectividade em rede, uso de robôs colaborativos, redução do consumo de insumos e maior segurança ocupacional (Cakir; Guvenc; Mistikoglu, 2021).

No campo da manutenção, a Indústria 4.0 substitui abordagens tradicionais por sistemas de monitoramento de condição (CMS) integrados ao IoT, capazes de coletar dados em tempo real e antecipar falhas. Nessa lógica, algoritmos de *Machine Learning* (como SVM e *Random Forest*) analisam informações de sensores de vibração, som, corrente e temperatura, fornecendo diagnósticos automáticos e acionando alarmes preventivos (Cakir; Guvenc; Mistikoglu, 2021). Diante disso, abre-se espaço para novas práticas de monitoramento que utilizam diretamente os sinais emitidos pelas máquinas. Entre eles, os sinais acústicos e de vibração se destacam por fornecer informações valiosas sobre o comportamento dinâmico dos equipamentos, permitindo diagnósticos ainda mais robustos e precisos.

2.2 Previsão de falha por sinais sonoros e de vibração

Os sinais de vibração e acústicos têm se consolidado como fontes fundamentais de informação para o diagnóstico de falhas em máquinas rotativas, especialmente em rolamentos, engrenagens e sistemas de transmissão. Os sinais de vibração estão presentes de forma natural em qualquer equipamento desse tipo, resultantes do movimento de seus componentes, do atrito e das interações dinâmicas internas. Quando ocorre um defeito, como em pistas ou esferas de rolamentos, esses sinais passam a apresentar padrões anômalos, caracterizados por choques de

alta frequência que se propagam pela estrutura mecânica. Tais alterações no espectro vibracional permitem identificar falhas incipientes ou em evolução (Wang; Mao; Li, 2021). A coleta desses dados é tradicionalmente realizada por meio de acelerômetros fixados na máquina, em altas taxas de amostragem, o que faz da vibração uma técnica consolidada e amplamente utilizada no monitoramento de condição (Pacheco-Chérrez *et al*., 2022)

Já os sinais acústicos, diferentemente dos de vibração, não dependem de contato direto com a máquina. Eles correspondem a ondas de pressão propagadas pelo ar ou por outros meios físicos, captadas por microfones ou sensores de emissão acústica (Wang; Mao; Li, 2021). Assim como as vibrações, refletem o funcionamento normal do equipamento, mas também apresentam alterações de padrão em situações de anormalidade. Sua principal vantagem está na característica não invasiva e no potencial de detectar falhas em estágios iniciais, muitas vezes antes que se tornem críticas, configurando-se como uma alternativa de baixo custo e alta sensibilidade para aplicações de manutenção preditiva. Além de não invasivos e de baixo custo, os sinais acústicos exigem menor preparação do *hardware* e podem ser coletados mesmo quando a instalação de acelerômetros não é viável (Liu *et al*., 2020).

Cada técnica apresenta vantagens e limitações: os sinais de vibração oferecem maior precisão em cenários controlados, mas demandam instalação física e são sensíveis a ruídos externos; por outro lado, os sinais acústicos têm melhor desempenho em condições de baixo custo e acesso restrito, embora também possam ser impactados por ambientes ruidosos *(Xu et al*., 2021; Pham; Kim; Kim, 2020). A Tabela 1 mostra um comparativo entre os dois tipos de sinal.

Tabela 1 - Comparação entre sinais de vibração e acústicos

Aspecto	Vibração	Acústico
Origem do sinal	Resulta do movimento, atrito e choques entre componentes da máquina, propagando-se pela estrutura mecânica.	Resulta da propagação de ondas de pressão no ar (ruído) ou ondas elásticas de alta frequência (emissão acústica).
Coleta	Necessita de contato direto com a máquina, geralmente por acelerômetros fixados na estrutura.	Realizada de forma **não** invasiva, por microfones ou sensores de emissão acústica próximos ao equipamento.
Taxa de amostragem típica	Alta (kHz a dezenas de kHz), dependendo do tipo de análise de falha.	Variável: desde frequências audíveis (Hz–kHz) até faixas ultrassônicas (>100 kHz).
Sensibilidade a falhas	Detecta falhas estruturais e mecânicas consolidadas; muito usado para rolamentos, engrenagens e eixos.	Mais sensível a falhas incipientes ou sutis, detectando mudanças de padrão sonoro em estágios iniciais.
Vantagens	Técnica consolidada e confiável, ampla literatura, boa precisão em cenários controlados.	Não invasivo, de baixo custo, fácil instalação, útil em situações em que sensores de vibração não podem ser aplicados.
Limitações	Requer instalação física (invasivo); suscetível a ruídos estruturais; difícil em pontos de acesso restrito.	Pode ser afetado por ruído ambiente; maior necessidade de técnicas de filtragem em ambientes industriais.
Aplicações típicas	Monitoramento de condição de máquinas rotativas, diagnóstico de rolamentos e engrenagens.	Monitoramento de motores, correias transportadoras, turbinas, e detecção precoce de falhas incipientes.

Aspecto	Vibração	Acústico
Complementaridade	Fornece informação estrutural detalhada sobre a máquina.	Capta informações adicionais que a vibração pode não registrar, enriquecendo o diagnóstico quando usados em conjunto.

Estudos recentes mostram que a fusão vibro-acústica supera as limitações individuais, combinando a robustez da vibração com a sensibilidade da acústica. Métodos híbridos, como o uso de CNNs unidimensionais, algoritmos de *Random Forest* ou técnicas de redução de dimensionalidade como LDA, alcançam acurácias superiores, chegando a mais de 98% na classificação de falhas (Wang; Mao; Li, 2021) . Essa fusão se mostra especialmente eficaz em condições adversas, com baixos níveis de sinal-ruído, ampliando a generalização dos modelos e fortalecendo a resiliência frente a diferentes cenários operacionais (Li *et al*., 2024)

Dessa forma, a utilização conjunta de sinais de vibração e acústicos não apenas aumenta a acurácia dos diagnósticos, como também promove maior confiabilidade na manutenção preditiva. A complementaridade desses sinais permite que falhas incipientes sejam identificadas com antecedência e que padrões complexos de degradação sejam reconhecidos de forma mais estável, consolidando o uso de dados multissensoriais como tendência central no desenvolvimento de soluções de manutenção inteligente (Praveen Kumar *et al*., 2019)

Diversos estudos recentes reforçam essa tendência, aplicando diferentes arquiteturas de aprendizado de máquina sobre sinais fusionados de vibração e acústica em diferentes contextos industriais. A Tabela 2 sintetiza alguns desses trabalhos, destacando os sistemas analisados, os tipos de fusão empregados, os algoritmos utilizados e os principais resultados obtidos. Observa-se que, independentemente do setor ou da abordagem adotada, a fusão multissensorial tem se mostrado superior à análise unimodal, seja pela elevação das taxas de acurácia, pela maior robustez em ambientes ruidosos ou pela capacidade de generalização em diferentes condições operacionais.

Tabela 2 - Resumo de pesquisas sobre fusão multissensorial com técnicas de IA

Autores / Ano	Sistema estudado	Algoritmos aplicados	Principais resultados
Praveen Kumar et al. (2019)	*Gearbox*	ANN, SVM, PSVM	Fusão multi-sensor elevou acurácia em relação a sinais isolados
Inturi et al. (2023)	*Gearbox* multiestágio	Árvores de decisão, SVM, DL	Acurácia 96,4%; fusão > sinais individuais
Pacheco-Chérrez et al. (2022)	Rolamentos	LDA, ML supervisionado	Acurácia até 98,28% com TSFDR-LDA
Wang et al. (2021)	Rolamentos	CNN 1D	Acurácia >98% em baixo SNR; robustez superior
Zhang et al. (2022)	Rolamentos	CNN 2D + AdaBoost	Melhor generalização e resistência a ruído
Li et al. (2016)	*Gearbox*	*Deep Random Forest Fusion*	97,68% para 11 condições
Xu et al. (2021)	Rolamentos de trens HSTWSB	Modelo híbrido prognóstico	Melhora no diagnóstico e prognóstico em ambiente ruidoso
Al Mamun et al. (2022)	Máquinas rotativas	MPCA (freq. domain)	Fusão mais eficaz que sinais isolados
Li et al. (2024)	Rolamentos (turbina eólica)	*Viewable Neural Nets*	Maior acurácia conforme aumenta nº de sinais fusionados
Gültekin et al. (2022)	Veículo autônomo industrial	CNN + STFT	Acurácia e robustez superiores em detecção de falhas

2.3 Coleta e análise dos sinais acústicos e de vibração

A utilização de sinais acústicos e de vibração no diagnóstico de falhas exige um processo estruturado de aquisição e análise de dados. O primeiro passo consiste na coleta dos sinais brutos, que pode ser realizada por sensores de vibração (acelerômetros) acoplados diretamente à estrutura da máquina ou por sensores acústicos (microfones e sensores de emissão acústica) posicionados próximos ao equipamento, sem necessidade de contato físico. A escolha e posicionamento adequado dos sensores é fundamental para garantir que os sinais coletados representem de forma fidedigna as condições reais de operação.

Uma vez captados, esses sinais passam por conversão analógico-digital, de modo que as variações contínuas de vibração ou pressão sonora possam ser representadas numericamente

em um sistema computacional. Esse processo é realizado por conversores A/D (ADC – *Analog to Digital Converter*), cuja resolução e precisão impactam diretamente na qualidade dos dados adquiridos.

Outro aspecto crítico é a frequência de amostragem, definida de acordo com a faixa de frequências relevantes ao fenômeno investigado. A teoria de Nyquist estabelece que a taxa de amostragem deve ser, no mínimo, o dobro da frequência máxima presente no sinal, para evitar *aliasing* e preservar a integridade da informação (Marks, 1991). Em estudos de diagnóstico de falhas em máquinas rotativas, são comuns frequências de amostragem variando de alguns kHz até centenas de kHz, dependendo se o foco é a análise de vibração de baixa frequência ou de emissão acústica em alta frequência.

Após a aquisição e digitalização, realiza-se a extração de características (*features*), etapa essencial para transformar os sinais brutos em representações compactas e informativas. No domínio do tempo, utilizam-se parâmetros estatísticos como valor RMS, média, desvio padrão, assimetria e curtose, que descrevem a forma da onda e sua variabilidade. No domínio da frequência, a aplicação da Transformada Rápida de Fourier (FFT) permite identificar componentes espectrais associados a falhas específicas, como as frequências características de rolamentos. Já no domínio tempo-frequência, métodos como a STFT (*Short-Time Fourier Transform*) e a *Wavelet Packet Transform* possibilitam analisar sinais não estacionários, capturando variações espectrais ao longo do tempo.

A seleção adequada dessas características é determinante para a eficácia dos modelos de aprendizado de máquina empregados no diagnóstico (Raouf; Lee; Kim, 2022; Xu et al., 2021). Além disso, com o avanço das técnicas de *Deep Learning*, cresce o interesse em arquiteturas capazes de realizar a extração automática de atributos diretamente a partir dos sinais brutos, reduzindo a necessidade de engenharia manual e ampliando a capacidade de generalização dos sistemas de manutenção preditiva (Janssens et al., 2016a; Wang; Mao; Li, 2021). No presente trabalho, a partir dos sinais acústicos e de vibração foram extraídas características tanto no domínio do tempo quanto no da frequência.

2.3.1 Características (*Features*) Domínio do tempo

A análise no domínio do tempo foi realizada a partir da extração de um conjunto de características que, conforme demonstrado na literatura, são sensíveis a alterações nas condições de operação de máquinas. (Altaf et al., 2022; Liu et al., 2020; Ramteke; Chelladurai; Amarnath, 2022; Xu et al., 2021).

A média (μ) corresponde ao valor médio do sinal dentro de uma janela de análise e

representa sua tendência central (Eq. 1).

O desvio-padrão (σ) expressa a dispersão dos valores em torno da média, indicando a variabilidade do sinal no intervalo considerado (Eq. 2).

O valor quadrático médio (RMS) quantifica a energia média do sinal ao longo da janela, sendo amplamente utilizado por sua sensibilidade a variações de amplitude (Eq. 3).

O valor de pico (x_peak) representa a maior amplitude observada na janela, útil para capturar excursões instantâneas associadas a choques, impactos ou eventos transitórios (Eq. 4).

A assimetria (*Skewness*, Sk) descreve a simetria da distribuição do sinal: valores positivos indicam cauda à direita e negativos, cauda à esquerda; valores próximos de zero sugerem distribuição aproximadamente simétrica (Eq. 5).

A curtose (*Kurtosis*, Ku) avalia a concentração de energia em torno da média e a presença de "picos" na distribuição; valores elevados podem sinalizar impulsividade e são úteis para detecção de falhas incipientes (Eq. 6).

Por fim, a energia (E) é a soma dos quadrados das amostras na janela e reflete a energia acumulada do sinal (proporcional à potência) (Eq. 7).

$$\mu = \left(\frac{1}{N}\right) \sum_{\{i=1\}}^{N} x_i \quad \textbf{(1)}$$

$$\sigma = \sqrt{\left(\frac{1}{N}\right) \sum_{\{i=1\}}^{N} (x_i - \mu)^2} \quad \textbf{(2)}$$

$$x_{RMS} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2} \quad \textbf{(3)}$$

$$x_{\text{peak}} = \max_{i \in \{1,\ldots,N\}} |x_i| \quad \textbf{(4)}$$

$$Sk = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3} \quad \textbf{(5)}$$

$$Ku = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} \quad \textbf{(6)}$$

$$E = \sum_{i=1}^{N} x_i^2 \quad \textbf{(7)}$$

2.3.2 Características (*Features*) Domínio da frequência

No domínio da frequência, a frequência de pico (f_{peak}) foi considerada como principal parâmetro para este trabalho, conforme indicado em estudos que demonstram sua relevância para o diagnóstico de falhas em máquinas rotativas (Janssens et al., 2016; Ramteke; Chelladurai; Amarnath,2022). Essa característica corresponde à componente espectral de maior magnitude, refletindo a frequência predominante do sinal. As frequências discretas associadas aos bins do espectro, obtidas pela FFT de uma janela com *N* amostras e frequência de amostragem f_s, são dadas por (8); para sinais reais, considera-se o espectro unilateral até k =*N*/2. A frequência de pico é então definida por (9), em que $X(f_k)$ denota a magnitude do espectro na frequência f_k e k é o índice do *bin* de maior magnitude.

$$f_k = \frac{k \cdot f_s}{N}, \quad k = 0,1,\ldots,\frac{N}{2} \quad \textbf{(8)}$$

$$f_{\text{peak}} = f_{k^*}, \quad \text{com } k^* = \arg\max_k(|X(f_k)|) \quad \textbf{(9)}$$

Onde**:**

$\boldsymbol{x_i}$: i-ésima amostra do sinal em uma janela de análise.

N : número total de amostras na janela.

μ : valor médio (média aritmética) do sinal na janela.

σ : desvio-padrão do sinal na janela.

$x_{(RMS)}$: valor quadrático médio *(Root Mean Square*) do sinal.

$x_{(peak)}$: valor de pico, isto é, a maior amplitude observada no sinal.

Sk : coeficiente de assimetria (Skewness).

Ku : coeficiente de curtose (Kurtosis).

E : energia acumulada do sinal (proporcional à potência).

f_s: frequência de amostragem (Hz).

$X(f_k)$: magnitude do espectro de Fourier no bin de frequência f_k.

f_k = frequência associada ao índice k na FFT.

$f_{(peak)}$: frequência de pico, correspondente ao bin de frequência com maior magnitude no espectro.

$k*$: índice no qual ocorre o valor máximo de $|X(f_k)|$.

2.3.3 Representações no domínio tempo-frequência

Embora as análises no domínio do tempo e da frequência forneçam informações relevantes para o diagnóstico de falhas, sinais vibroacústicos oriundos de sistemas mecânicos reais são, em geral, não estacionários, isto é, suas características espectrais variam ao longo do tempo em função do regime de operação, da carga e do estado de degradação do equipamento. Nesses casos, abordagens puramente temporais ou puramente espectrais podem não capturar de forma adequada a dinâmica completa do sinal (Tran; Lundgren, 2020).

As representações no domínio tempo-frequência surgem, portanto, como uma alternativa capaz de descrever simultaneamente quando e em quais frequências determinadas componentes energéticas ocorrem, permitindo uma análise mais rica e informativa do comportamento do sistema(Inturi et al., 2023). Essa característica torna tais representações particularmente adequadas para o diagnóstico de falhas incipientes, que frequentemente se manifestam como eventos transitórios ou padrões localizados no tempo e na frequência (Siddique et al., 2023).

Além disso, as representações tempo-frequência desempenham um papel central em abordagens modernas baseadas em *Deep Learning*, uma vez que possibilitam a conversão de sinais unidimensionais em estruturas bidimensionais, análogas a imagens, que podem ser exploradas de forma eficiente por Redes Neurais Convolucionais (CNNs) (Islam; Kim, 2019).

2.3.3.1 Transformada de Fourier de Curto Tempo (STFT)

A Transformada de Fourier de Curto Tempo (*Short-Time Fourier Transform* – STFT) é

uma das técnicas mais utilizadas para análise tempo-frequência de sinais não estacionários. Diferentemente da FFT convencional, que fornece apenas uma visão global do conteúdo espectral, a STFT aplica a Transformada de Fourier em segmentos sucessivos do sinal, obtidos por meio de uma janela deslizante ao longo do tempo (Ramteke; Chelladurai; Amarnath, 2022).

Matematicamente, a STFT de um sinal $\boldsymbol{x(t)}$ pode ser expressa como:

$$\mathrm{STFT}\{x(t)\}(\tau, f) = \int_{-\infty}^{\infty} x(t)\; w(t-\tau)\, e^{-j2\pi f t}\, dt \qquad \textbf{(10)}$$

onde $\boldsymbol{w(t)}$representa a função janela centrada no instante $\boldsymbol{\tau}$, e $\boldsymbol{f}$é a frequência analisada. A escolha do tipo e do tamanho da janela implica um compromisso entre resolução temporal e resolução espectral, conhecido como princípio de incerteza tempo-frequência.

O módulo ao quadrado da STFT resulta no espectrograma, uma representação bidimensional que expressa a distribuição de energia do sinal em função do tempo e da frequência. Espectrogramas têm sido amplamente empregados no diagnóstico de falhas em máquinas rotativas, pois permitem visualizar padrões característicos associados a impactos, modulações e excitações periódicas decorrentes de defeitos mecânicos (Liu; Li; Ma, 2016).

2.3.3.2 Espectrogramas e aplicações em CNNs

O espectrograma constitui uma representação especialmente relevante no contexto de aprendizado profundo, pois pode ser interpretado como uma imagem, na qual um dos eixos corresponde ao tempo, o outro à frequência, e a intensidade de cada pixel representa a energia espectral. Essa estrutura é particularmente adequada para o uso de CNNs bidimensionais, que exploram correlações locais por meio de filtros convolucionais (Siddique et al., 2023).

Diversos estudos demonstram que CNNs aplicadas a espectrogramas de sinais acústicos e de vibração apresentam desempenho superior em comparação a métodos baseados exclusivamente em *features* manuais, especialmente em ambientes ruidosos e em condições operacionais variáveis. A capacidade das CNNs de aprender automaticamente padrões discriminantes em diferentes escalas torna essas abordagens mais robustas à variabilidade do processo e menos dependentes do conhecimento prévio do domínio (Hasan; Islam; Kim, 2019; Janssens et al., 2016b; Wang; Mao; Li, 2021)..

No caso de sinais acústicos, é comum a utilização de variações do espectrograma tradicional, como o Mel-espectrograma, que organiza as frequências segundo a escala Mel, aproximando-se da percepção auditiva humana. Essa representação tem se mostrado

particularmente eficaz na análise de sinais sonoros industriais, contribuindo para a detecção precoce de falhas (Tran; Lundgren, 2020).

2.3.3.2.1 Escala Mel e Mel-espectrogramas

A escala Mel é uma escala perceptual de frequência proposta originalmente para aproximar a forma como o ouvido humano percebe diferenças entre tons sonoros. Diferentemente da escala linear de frequência, na qual incrementos constantes em Hertz correspondem a variações igualmente espaçadas, a escala Mel apresenta maior resolução em baixas frequências e resolução progressivamente menor em altas frequências. Essa característica reflete o fato de que o sistema auditivo humano é mais sensível a variações frequenciais em faixas baixas do espectro (Stevens; Volkmann; Newman, 1937).

A conversão de uma frequência linear f, expressa em Hertz, para a escala Mel pode ser aproximada por:

$$\text{Mel}(f) = 2595 \cdot \log_{10}\left(1+\frac{f}{700}\right) \quad \textbf{(11)}$$

A partir dessa transformação, é possível definir um banco de filtros Mel, composto por filtros triangulares sobrepostos, distribuídos de forma aproximadamente uniforme na escala Mel. Cada filtro atua como um ponderador de energia em uma faixa específica de frequências, sendo aplicado ao espectro de potência obtido via STFT.

Matematicamente, o m-ésimo filtro Mel $H_m(f)$pode ser definido como:

$$H_m(f) = \begin{cases} 0, & f < f_{m-1} \\ \dfrac{f - f_{m-1}}{f_m - f_{m-1}}, & f_{m-1} \le f < f_m \\ \dfrac{f_{m+1} - f}{f_{m+1} - f_m}, & f_m \le f < f_{m+1} \\ 0, & f \ge f_{m+1} \end{cases} \quad \textbf{(12)}$$

onde f_{m-1}, f_me f_{m+1}representam as frequências centrais adjacentes, definidas a partir de uma discretização aproximadamente uniforme da escala Mel e posteriormente mapeadas de

volta para a escala linear de frequência.

A energia associada ao m-ésimo filtro Mel é então obtida pela soma ponderada do espectro de potência:

$$E_m = \sum_f | X(f) |^2 \cdot H_m(f) \quad \textbf{(13)}$$

em que $| X(f) |^2$corresponde ao espectro de potência do sinal. O conjunto dos valores E_m, para $m = 1,2, \dots, M$, constitui a representação espectral na escala Mel, sendo M o número total de filtros adotados.

Essa formulação permite enfatizar regiões de frequência mais relevantes do ponto de vista perceptual, ao mesmo tempo em que reduz a dimensionalidade do espectro original, resultando em uma representação compacta e adequada ao uso em métodos de aprendizado profundo.

O Mel-espectrograma é, portanto, obtido pela combinação da STFT com esse banco de filtros perceptuais, seguido, usualmente, da aplicação de uma escala logarítmica sobre a energia espectral. Essa representação reduz a dimensionalidade do espectro original, atenua variações irrelevantes de alta frequência e destaca padrões globais de energia ao longo do tempo.

Embora a escala Mel tenha sido originalmente desenvolvida para aplicações em reconhecimento de fala e processamento de áudio perceptual, diversos estudos recentes demonstram sua eficácia também na análise de sinais acústicos industriais. Em particular, Mel-espectrogramas têm se mostrado adequados para tarefas de diagnóstico de falhas baseadas em aprendizado profundo, uma vez que fornecem representações compactas, robustas a ruído e compatíveis com Redes Neurais Convolucionais bidimensionais(Natesha; Guddeti, 2021; Tagawa; Maskeliūnas; Damaševičius, 2021; Tran; Lundgren, 2020).

Além disso, o uso de Mel-espectrogramas contribui para a padronização da entrada da rede neural, reduzindo a sensibilidade a pequenas variações espectrais e favorecendo a generalização do modelo em diferentes condições operacionais. Dessa forma, essa representação estabelece um compromisso adequado entre fidelidade espectral, robustez e custo computacional, justificando sua adoção na vertente baseada em aprendizado profundo desenvolvida neste trabalho.

2.3.3.3 Outras técnicas de representação tempo-frequência

Além da STFT e dos espectrogramas derivados, outras técnicas de análise tempo-frequência têm sido amplamente investigadas na literatura de diagnóstico de falhas, com destaque para as transformadas *wavelet*, como a Transformada *Wavelet* Contínua (CWT) e a *Wavelet Packet Transform* (WPT). Essas abordagens utilizam funções base localizadas no tempo e na frequência, permitindo melhor resolução temporal em altas frequências e melhor resolução espectral em baixas frequências, característica particularmente útil para a detecção de eventos impulsivos associados a falhas incipientes em rolamentos e engrenagens (Praveen Kumar et al., 2019; Siddique et al., 2023). Assim como os espectrogramas, as representações obtidas por *wavelets* podem ser organizadas na forma de mapas bidimensionais (scalograms), viabilizando sua aplicação como entrada para Redes Neurais Convolucionais.

Apesar de suas vantagens, a adoção de técnicas baseadas em *wavelets* envolve escolhas adicionais, como o tipo de *wavelet* mãe, o nível de decomposição e os critérios de reconstrução, o que pode aumentar a complexidade do processo de análise e dificultar a comparação direta entre abordagens (Inturi et al., 2023). Dessa forma, embora reconhecidamente eficazes, as representações *wavelet* não foram exploradas em profundidade neste trabalho, sendo consideradas uma alternativa relevante para investigações futuras.

No contexto desta pesquisa, as representações no domínio tempo-frequência desempenham o papel de ponte conceitual entre os métodos baseados em engenharia manual de atributos e as abordagens baseadas em *Deep Learning*. Enquanto as *features* extraídas nos domínios do tempo e da frequência alimentam modelos clássicos de aprendizado de máquina, as representações tempo-frequência, em especial os espectrogramas obtidos via STFT, permitem a aplicação de Redes Neurais Convolucionais bidimensionais, possibilitando a extração automática de características discriminantes diretamente dos sinais vibroacústicos.

Assim, a escolha pela STFT e por espectrogramas fundamenta-se em sua ampla adoção na literatura recente, simplicidade de implementação e adequação direta ao uso com CNNs, permitindo uma comparação consistente entre as duas vertentes investigadas neste trabalho: modelos baseados em engenharia de atributos e modelos baseados em aprendizado profundo. Essa abordagem possibilita avaliar, de forma sistemática, os ganhos e limitações de cada estratégia no diagnóstico de falhas em máquinas de fabricação de copos de papel.

2.4 Modelos de Classificação Supervisionada aplicado a detecção de falhas

2.4.1 Modelos Estatísticos Tradicionais

Os modelos estatísticos tradicionais desempenham um papel fundamental no desenvolvimento e na avaliação de sistemas de classificação supervisionada, especialmente em estudos voltados ao diagnóstico de falhas. Essas técnicas constituem a base histórica do reconhecimento de padrões (Cortes; Vapnik, 1995; Haykin, 2009) e são amplamente utilizadas como modelos de referência (*baseline*), devido à sua simplicidade, baixo custo computacional e elevada interpretabilidade (Cakir; Guvenc; Mistikoglu, 2021; Pacheco-Chérrez et al., 2022).

No contexto desta pesquisa, a inclusão de modelos estatísticos clássicos tem como objetivo fornecer uma linha de base consistente para comparação com algoritmos de aprendizado de máquina mais complexos, além de contribuir para a análise exploratória dos atributos extraídos dos sinais acústicos e de vibração. Dentre esses modelos, destacam-se a Regressão Logística e a Análise Discriminante Linear (LDA), amplamente empregadas na literatura de diagnóstico de falhas em sistemas mecânicos e industriais.

2.4.1.1 Regressão Logística

A regressão logística é um modelo estatístico clássico amplamente utilizado em problemas de classificação supervisionada, especialmente em cenários nos quais se deseja estimar a probabilidade de ocorrência de um determinado evento a partir de um conjunto de variáveis explicativas (Natesha; Guddeti, 2021). Diferentemente da regressão linear, cuja saída assume valores contínuos em todo o conjunto dos números reais, a regressão logística emprega uma função de ligação não linear, conhecida como função logística ou sigmoide, que mapeia a combinação linear dos atributos de entrada para o intervalo [0,1], permitindo a interpretação probabilística da saída do modelo(Haykin, 2009).

Matematicamente, a probabilidade de uma observação pertencer à classe positiva é expressa por:

$$P(y = 1|x) = \frac{1}{\left(1 + e^{-(\beta_0 + \beta^T x)}\right)} \quad \textbf{(14)}$$

Onde x representa o vetor de atributos extraídos dos sinais, β_0 é o termo de intercepto e β corresponde ao vetor de coeficientes associados a cada atributo. Uma das formas de se obter o ajuste dos parâmetros da regressão logística consiste na minimização da entropia cruzada, também conhecida como *log loss*. Essa função de custo quantifica a discrepância entre as

probabilidades previstas pelo modelo e os rótulos reais das amostras, sendo amplamente utilizada em problemas de classificação binária.

No caso da regressão logística, a minimização da entropia cruzada é matematicamente equivalente à maximização da função de verossimilhança, o que confere ao modelo uma fundamentação estatística sólida (Haykin, 2009). Dessa forma, os coeficientes do modelo são estimados de modo a maximizar a probabilidade dos dados observados, dadas as probabilidades previstas pela função sigmoide.

Para um conjunto de N amostras, a função de custo pode ser expressa como:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)] \quad \textbf{(15)}$$

Onde$y_i \in \{0,1\}$ representa o rótulo real da i-ésima amostra e $\hat{p}_i = P(y = 1 \mid x_i)$ corresponde à probabilidade estimada pelo modelo. A minimização dessa função permite obter os parâmetros β_0e βque melhor se ajustam aos dados de treinamento.

2.4.1.2 Análise do Discriminante Linear

A Análise Discriminante Linear (*Linear Discriminant Analysis* – LDA) é uma técnica estatística clássica de classificação supervisionada amplamente empregada em problemas de reconhecimento de padrões. Derivada do discriminante linear proposto por Fisher, a LDA tem como objetivo encontrar uma transformação linear dos dados que maximize a separação entre classes distintas, ao mesmo tempo em que minimiza a dispersão intra-classe (Haykin, 2009; Raouf; Lee; Kim, 2022).

Diferentemente da regressão logística, que modela diretamente a probabilidade condicional de pertencimento a uma classe, a LDA baseia-se em uma abordagem gerativa, assumindo que os dados de cada classe seguem uma distribuição aproximadamente gaussiana, com médias distintas e matrizes de covariância semelhantes(Cortes; Vapnik, 1995; Haykin, 2009). A partir dessas hipóteses, o método busca um vetor de projeção que maximize a razão entre a variância entre classes e a variância dentro das classes.

Matematicamente, o critério de Fisher pode ser expresso como:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \quad \textbf{(16)}$$

Onde S_B representa a matriz de dispersão entre classes (*between-class scatter matrix*) e S_W a matriz de dispersão intra-classe (*within-class scatter matrix*). O vetor w que maximiza esse critério define a direção na qual os dados projetados apresentam maior separabilidade linear.

No contexto do diagnóstico de falhas em sistemas mecânicos e industriais, a LDA é frequentemente utilizada como método de referência e comparação, aparecendo de forma recorrente ao lado da regressão logística em estudos que avaliam algoritmos mais avançados(Pacheco-Chérrez et al., 2022). Além disso, a técnica pode ser empregada tanto como classificador direto quanto como etapa de redução de dimensionalidade, antecedendo classificadores mais complexos, especialmente em aplicações envolvendo sinais de vibração e acústicos (Altaf et al., 2022).

2.4.2 Modelos de Aprendizado de Máquina

A utilização de algoritmos de *Machine Learning* (ML) é essencial no diagnóstico automatizado de falhas, uma vez que esses métodos permitem classificar padrões extraídos de sinais acústicos e de vibração e distinguir entre condições normais e defeituosas de operação. Os modelos de ML supervisionados, em especial, têm se destacado por sua capacidade de aprender com conjuntos de dados rotulados, generalizando o conhecimento adquirido para novos cenários. Nesta pesquisa, foram empregados algoritmos clássicos e avançados de classificação, incluindo *Support Vector Machines* (SVM*), Extreme Gradient Boosting* (*XGBoost*), *Random Forest* (RF), redes neurais do tipo *Multilayer Perceptron* (MLP) e abordagens de *Ensemble Learning* por votação suave (*Soft Voting*).

A seleção desses algoritmos fundamenta-se em suas características complementares e na ampla aplicação em problemas de diagnóstico de falhas em sistemas industriais. As Máquinas de Vetores de Suporte (SVM) são reconhecidas pela robustez em conjuntos de dados de pequeno e médio porte e pela capacidade de separar padrões sutis em espaços de alta dimensionalidade, sendo frequentemente empregadas como modelos de referência em tarefas de classificação (Praveen Kumar et al., 2019). Os métodos baseados em árvores, como *Random Forest* e *Extreme Gradient Boosting* (*XGBoost*), destacam-se pela habilidade de modelar relações não lineares complexas e pela robustez a ruídos típicos de sinais industriais, ainda que

apresentem diferentes compromissos entre interpretabilidade, custo computacional e risco de sobreajuste (Liu et al., 2020). As redes neurais do tipo *Multilayer Perceptron* (MLP), por sua vez, permitem capturar padrões não lineares mais complexos, ao custo de maior demanda computacional e menor interpretabilidade(Gültekin et al., 2022; Haykin, 2009). Por fim, a adoção de uma abordagem de *Ensemble Learning* por votação suave (*Soft Voting*) busca explorar a diversidade entre esses classificadores, combinando suas previsões individuais com o objetivo de aumentar a estabilidade e a capacidade de generalização do diagnóstico, aspectos particularmente relevantes em aplicações industriais reais (Mohammed; Kora, 2023).

2.4.2.1 Support Vector Machines (SVM)

As Máquinas de Vetores de Suporte (*Support Vector Machines – SVM*), propostas inicialmente por Cortes e Vapnik (1995), constituem uma das técnicas mais consolidadas e versáteis do aprendizado de máquina, aplicáveis não apenas à classificação, mas também à regressão e à detecção de anomalias. A ideia central do classificador SVM é encontrar um hiperplano ótimo que separe as classes, maximizando a margem — isto é, a distância entre o hiperplano e as instâncias de treinamento mais próximas, chamadas vetores de suporte. Essa maximização da margem reduz o risco de sobreajuste e confere ao modelo maior capacidade de generalização.

Em casos linearmente separáveis, a SVM busca a chamada margem rígida, na qual todas as instâncias ficam do lado correto da fronteira de decisão. Entretanto, em bases reais, com ruído e sobreposição entre classes, adota-se a margem suave, que permite violações controladas por meio do hiperparâmetro C, que é detalhado adiante. Para lidar com conjuntos de dados não linearmente separáveis, as SVM utilizam o chamado truque do *kernel* (*kernel trick*), que projeta os dados em espaços de maior dimensão, nos quais a separação linear é possível, por meio de funções que satisfaçam o Teorema de Mercer (Cortes; Vapnik, 1995).

No diagnóstico de falhas, as SVM destacam-se pela robustez em conjuntos de dados de pequeno e médio porte, pela boa performance em espaços de alta dimensionalidade e pela capacidade de separar padrões sutis associados a condições normais e defeituosas. Na prática, sua implementação é amplamente disponibilizada em pacotes de aprendizado de máquina em *Python*, como o *Scikit-Learn* (Géron, 2023) , o que possibilita sua aplicação em cenários industriais. Do ponto de vista formal, de acordo com Cortes e Vapnik (1995), um classificador SVM linear busca resolver o seguinte problema de otimização, de acordo com (17).

$$\min_{\boldsymbol{w},b} \frac{1}{2} \| \boldsymbol{w} \|^2 \quad \text{sujeito a} \quad y_i(\boldsymbol{w}^\top xi + b) \geq 1,\ i = 1, \ldots, m \tag{17}$$

Onde:

$x_i \in R^n$ representa a i-ésima instância de entrada,

$y_i \in \{-1, +1\}$ é o rótulo da classe,

w é o vetor de pesos do hiperplano,

b é o termo de viés, que desloca o hiperplano em relação à origem do espaço de atributos.

Quando adotada a margem suave, introduzem-se variáveis de folga $\boldsymbol{\xi_i \geq 0}$ que permitem violações controladas (Eq. 18):

$$\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} \xi_i \quad \text{sujeito a} \quad y_i(w^\top xi + b) \geq 1 - \xi_i, \tag{18}$$

Onde o hiperparâmetro C > 0 regula o compromisso entre maximizar a margem e minimizar as violações.

No caso das SVM com *kernel*s, em vez de trabalhar diretamente no espaço original dos dados, utiliza-se uma função de similaridade $K(\cdot,\cdot)$ que calcula o produto interno em um espaço transformado $\phi(x), (Eq. 19)$:

$$K(x_i, x_j) = \langle \phi(x_i),\ \phi(x_j) \rangle \tag{19}$$

Essa formulação permite que o modelo aprenda fronteiras de decisão não lineares de forma eficiente, sem necessidade de calcular explicitamente a transformação $\phi(\cdot)$. Adiante são mostrados os principais hiperparâmetros relacionados a SVM's de acordo com Géron, (2023):

- **C (parâmetro de penalização):** O parâmetro C, também chamado de penalização, define o equilíbrio entre maximizar a margem e minimizar os erros de classificação. Valores altos de C levam o modelo a priorizar a classificação correta de todas as instâncias, resultando em uma margem mais estreita e aumentando o risco de *overfitting*. Por outro lado, valores baixos de C permitem margens mais largas com algumas violações, favorecendo a capacidade de generalização, ainda que isso possa aumentar a taxa de erro no treinamento. No contexto de manutenção preditiva, um C muito alto pode capturar ruídos do sinal, enquanto valores mais moderados ajudam a identificar padrões gerais de falha.

- ***Kernel* (função de mapeamento**): É um dos pontos centrais da SVM. Conforme mencionado anteriormente, permite que dados não linearmente separáveis no espaço original sejam projetados em um espaço de maior dimensão, no qual a separação linear se torna possível. Os *kernel*s mais comuns são

    - **Linear**: adequado quando os dados já são aproximadamente separáveis por um hiperplano no espaço original; a função de similaridade reduz-se ao produto interno entre vetores, conforme Equação (20). É simples, rápido e interpretável, servindo bem como *baseline*.

$$K(x_i, x_j) = x_i^\top xj \quad \textbf{(20)}$$

    - **Polinomial**: captura interações de ordem superior entre atributos (termos quadráticos, cúbicos etc.), conforme Equação (21). O grau *d* controla a complexidade do limite de decisão e coef(r) ajusta a influência dos termos de ordem mais alta; valores muito elevados de *d* aumentam o risco de *overfitting* e o custo computacional, exigindo normalização prévia.

$$K(x_i, x_j) = (\gamma x_i^\top \mathrm{xj} + r)^d \quad \textbf{(21)}$$

    - **RBF (*Radial Basis Function*) Gaussiano**: mapeia os dados para um espaço de dimensão efetivamente infinita, permitindo limites altamente não

lineares (ver Equação (22)). O parâmetro γ controla o alcance de influência de cada amostra: γ alto tende a *overfitting*; γ baixo pode subajustar. Em prática industrial, costuma ser a opção padrão junto com um ajuste cuidadoso de C e γ.

$$K(xi, xj) = exp(-\gamma \parallel xi - xj \parallel^2) \quad \textbf{(22)}$$

- **Sigmoide:** inspirado em neurônios artificiais, utiliza uma função tipo tanh para medir similaridade, conforme Equação (23). Pode funcionar bem em cenários específicos, mas é menos estável e, para certos pares (γ,r), pode não satisfazer plenamente as condições de *kernel* positivo-definido; na prática, é menos utilizado que RBF ou polinomial.

$$K(x_i, x_j) = \tanh(\gamma x_i^{\top} x_j + r) \quad \textbf{(23)}$$

- **γ**: O parâmetro γ, que determina a largura de alcance tanto no *kernel* RBF quanto no polinomial, regula a influência de cada ponto de treino no modelo. Valores elevados de γ fazem com que cada instância exerça uma influência restrita, levando o modelo a aprender fronteiras altamente complexas e aumentando o risco de *overfitting*. Por outro lado, valores baixos de γ ampliam o raio de influência das instâncias, resultando em fronteiras mais suaves e o risco de *underfitting*.

- ***degree* (d):**Utilizado apenas no *kernel* polinomial, define o grau do polinômio. Graus mais altos capturam relações complexas, mas aumentam o risco de *overfitting* e o custo computacional.

- **coef0 (r)**:Presente no *kernel* polinomial e no sigmoide, controla a contribuição dos termos de ordem mais alta. Valores diferentes de zero permitem que os atributos interajam de maneira mais flexível, mas exigem cuidado para não gerar instabilidade.

2.4.2.2 Modelos baseados em árvores

As Árvores de Decisão constituem uma das abordagens mais conhecidas e intuitivas do aprendizado supervisionado. Sua lógica baseia-se em dividir recursivamente o espaço de atributos em regiões homogêneas, construindo uma estrutura hierárquica em forma de árvore (Breiman *et al*., 2017). Em cada nó interno é feita uma pergunta sobre uma variável de entrada (por exemplo, "a vibração RMS é maior que um determinado limiar?"), e o percurso segue para o nó filho correspondente. Esse processo continua até que se atinja um nó folha, no qual é atribuída uma classe (na tarefa de classificação) ou um valor médio (na tarefa de regressão).

De acordo com Breiman *et al*. (2017), a divisão em cada nó busca maximizar o ganho de pureza (redução de impureza) dos subconjuntos formados. Para quantificar a impureza utilizam-se o índice de Gini (Equação 24) e a entropia (Equação 25). A entropia mede a incerteza do nó: vale 0 quando o nó é puro e atinge o máximo quando as classes estão uniformemente distribuídas. O índice de Gini, por sua vez, corresponde à probabilidade de classificação incorreta ao se escolher aleatoriamente segundo a distribuição do nó; também é 0 em nós puros e máximo quando as classes estão equilibradas. O algoritmo escolhe a divisão que produz a maior queda nessas medidas, gerando partições progressivamente mais homogêneas. Esse mecanismo captura interações não lineares entre variáveis e resulta em modelos interpretáveis, expressos como regras de decisão simples.

$$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2 \tag{24}$$

$$Entropy(t) = -\sum_{i=1}^{C} p_i \, log_2(p_i) \tag{25}$$

Onde,

C = número de classes

p_i = proporção de elementos da classe *i* no nó *t*

Entre as principais vantagens das Árvores de Decisão estão a capacidade de lidar com variáveis categóricas e numéricas (Breiman, 2001), a robustez a transformações lineares (como normalização ou padronização, que muitas vezes não são necessárias) e a clareza interpretativa, que as torna modelos de "caixa branca" . Além disso, o custo computacional de previsão é baixo, uma vez que basta percorrer os nós até a folha correspondente (Breiman *et al*., 2017; Géron, 2023).

Contudo, as Árvores de Decisão apresentam limitações relevantes. De acordo com Chacón *et al*., (2021), modelos individuais são altamente sensíveis a pequenas variações nos dados de treinamento, o que pode resultar em fronteiras de decisão instáveis. Essa instabilidade decorre da heurística gulosa (*greedy*) utilizada no processo de divisão, no qual cada nó é escolhido para maximizar a pureza localmente, sem garantia de que a estrutura final seja globalmente ótima (Géron, 2023). Além disso, o sobreajuste (*overfitting*) constitui um problema recorrente, especialmente quando não se impõem restrições à profundidade da árvore, levando o modelo a memorizar ruídos específicos dos dados em vez de capturar padrões gerais (Breiman *et al*., 2017).

Essas limitações levaram ao desenvolvimento de métodos baseados em *Ensemble Learning*, nos quais diversas árvores são combinadas para formar modelos mais robustos. Nesse contexto, Breiman (2001) propôs as Florestas Aleatórias (*Random Forest*), baseadas em *bagging*, enquanto Chen e Guestrin (2016) apresentaram o *Extreme Gradient Boosting* (*XGBoost*), um dos algoritmos mais eficientes da família de *boosting*. Ambos se destacam como evoluções das Árvores de Decisão, explorando diferentes estratégias para aumentar a estabilidade e o poder preditivo.

Assim, as Árvores de Decisão, embora simples e interpretáveis, se consolidam como o alicerce para métodos mais avançados e eficazes, amplamente utilizados em aplicações industriais, inclusive no monitoramento e diagnóstico de falhas em máquinas rotativas.

2.4.2.2.1 Random Forest (RF)

Proposto por Breiman (2001), o modelo *Random Forest* é um algoritmo de *Ensemble Learning* que utiliza Árvores de Decisão como base, combinando-as para formar modelos mais robustos e acurados. Seu funcionamento é baseado no método de *bagging* (*bootstrap aggregation*), no qual múltiplas árvores são treinadas a partir de subconjuntos amostrados com reposição dos dados originais. Além disso, a cada divisão interna, em vez de considerar todas as variáveis disponíveis, apenas um subconjunto aleatório de atributos é utilizado, introduzindo

diversidade adicional entre os classificadores. A previsão final é obtida pela agregação das saídas individuais: no caso de classificação, pela votação majoritária; e, no caso de regressão, pela média das predições.

De acordo com Breiman (2001), essa estratégia reduz substancialmente a variância em comparação com uma árvore isolada, aumentando a capacidade de generalização e tornando o modelo menos sensível a ruídos ou variações nos dados de treinamento. Por essa razão, o *Random Forest* costuma apresentar desempenho superior em bases de dados reais e é amplamente reconhecido como uma das técnicas mais eficazes de aprendizado supervisionado.

Entre as principais vantagens do *Random Forest* destacam-se a robustez frente à instabilidade das Árvores de Decisão individuais, que são altamente sensíveis a pequenas alterações nos dados de treino, além do desempenho preditivo elevado, com fronteiras de decisão mais suaves e generalizáveis. O modelo também se destaca pela escalabilidade, já que cada árvore pode ser treinada em paralelo, tornando o algoritmo eficiente mesmo em grandes volumes de dados. Outro ponto relevante é a capacidade de interpretação relativa, pois a importância das variáveis pode ser estimada com base na redução média da impureza provocada por cada atributo em todas as árvores do *Ensemble* (Breiman, 2001).

Apesar dessas vantagens, o *Random Forest* não está livre de limitações. Quando a diversidade entre as árvores é insuficiente ou quando os dados de treinamento são escassos, pode ocorrer sobreajuste (*overfitting*). Além disso, embora mais interpretável que métodos, como de rede neural profunda, perde-se parte da simplicidade explicativa de uma única árvore de decisão (Géron, 2023).

Um aspecto relevante é o papel dos hiperparâmetros de regularização, que permitem controlar o equilíbrio entre viés e variância. De acordo com Géron (2023), destacam-se:

- ***n_estimators*(número de árvores):** Define quantas árvores serão geradas no *Ensemble*. Quanto maior esse número, mais estável e robusto é o modelo, pois reduz a variância das predições. Entretanto, o custo computacional cresce proporcionalmente, e ganhos adicionais tendem a se estabilizar após certo ponto.

- ***max_depth* (profundidade máxima):** Controla até que nível cada árvore pode se expandir. Árvores muito profundas capturam relações complexas, mas correm risco de **sobreajuste**; profundidades limitadas reduzem a complexidade e favorecem a generalização.

- **min_samples_split (mínimo de amostras para divisão):** Especifica o número

mínimo de instâncias necessário para que um nó seja dividido. Valores maiores forçam nós a terem mais exemplos antes de se subdividir, gerando árvores mais "rasas" e menos suscetíveis a ruído.

- **min_samples_leaf (mínimo de amostras em folhas):** Determina a quantidade mínima de instâncias que um nó folha deve conter. Evita folhas muito pequenas (com poucos exemplos), que tendem a capturar variações aleatórias dos dados.

- **max_*features* (número máximo de atributos por divisão):** Controla quantas variáveis são selecionadas aleatoriamente a cada divisão. Valores pequenos aumentam a aleatoriedade e reduzem a correlação entre as árvores, fortalecendo o *Ensemble*. Já valores maiores podem levar as árvores a se parecerem mais entre si, reduzindo a diversidade.

- **max_leaf_nodes (número máximo de folhas):** Restringe o número de nós terminais de cada árvore. É um controle direto sobre a complexidade do modelo, funcionando como forma de regularização.

Do ponto de vista formal, com base no trabalho de Breiman (2001), seja $h_b(x)$ a predição da $b-\acute{e}sima$ árvore do *Ensemble* para a entrada x. A saída do Floresta Aleatória é dada pelo Equação 26, onde *B* é o número de árvores no *Ensemble*.

$$\hat{y} = \text{mode}\{h_b(x),\ \text{b} = 1,2,\dots,\text{B}\} \quad \textbf{(26)}$$

No contexto da manutenção preditiva, o *Random Forest* é particularmente atrativo por aliar elevado poder preditivo com robustez a ruídos típicos de sinais industriais (Natesha; Guddeti, 2021). Essa combinação o torna adequado para o diagnóstico de falhas em máquinas rotativas a partir de múltiplas *features* extraídas de sinais acústicos e de vibração.

2.4.2.2.2 Extreme Gradient Boosting (XGBoost)

Desenvolvido por Chen e Guestrin (2016), o *Extreme Gradient Boosting* (*XGBoost*) constitui uma evolução dos algoritmos de *boosting* baseados em Árvores de Decisão, amplamente reconhecido por seu alto desempenho em competições de aprendizado de máquina (como as do *Kaggle*) e em aplicações industriais de larga escala. Diferente do *bagging* utilizado

no *Random Forest*, o *boosting* consiste em treinar árvores de forma sequencial, em que cada nova árvore busca corrigir os erros residuais cometidos pelas anteriores. Dessa forma, constrói-se um modelo aditivo no qual diversas árvores fracas (rasas e de baixa complexidade) são combinadas para formar um preditor robusto e de alta capacidade preditiva. Além disso, o *XGBoost* introduz avanços relevantes, como regularização explícita, otimização paralela e técnicas de prevenção de *overfitting*, o que explica sua popularidade e superioridade em diversos cenários práticos.

Segundo Chen e Guestrin (2016), o *XGBoost* aprimora o *boosting* sequencial de árvores ao incorporar regularização explícita, otimizações numéricas eficientes e suporte à paralelização, resultando em um método robusto e escalável. Formalmente, o treinamento minimiza a função objetivo que combina a perda empírica com a penalização da complexidade do conjunto de árvores, conforme Equação (27). A regularização é detalhada na Equação (28), onde O termo γT penaliza a complexidade estrutural, desestimulando árvores com muitas folhas (T), ao passo que $\frac{1}{2}\lambda \parallel w \parallel^2$ aplica encolhimento L^2 aos escores das folhas (w), reduzindo a variância do estimador. Em termos práticos, γ atua como um "custo mínimo de divisão" (exige maior ganho para criar novas folhas), e λ controla o amortecimento dos valores nas folhas, tornando as predições mais estáveis e menos propensas a *overfitting*.

:

$$Obj(\theta) = \sum_{i=1}^{m} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \qquad \textbf{(27)}$$

Onde:

$l(y_i, \hat{y_i})$ é a função de perda (por exemplo, erro quadrático ou log loss),

f_k representa a k-ésima árvore,

$\Omega(f_k)$ é o termo de regularização que penaliza a complexidade da árvore.

A regularização $\Omega(f_k)$ é definida como:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \quad \textbf{(28)}$$

Onde:

T é o número de folhas da árvore,

w são os pesos atribuídos às folhas,

γ controla a penalização pelo número de folhas (complexidade estrutural),

λ controla a regularização L^2 (ridge) sobre os pesos.

O *XGBoost*, segundo Chen e Guestrin (2016), destaca-se pelo elevado desempenho preditivo, frequentemente superando outros algoritmos em competições e aplicações práticas. Sua regularização integrada reduz consideravelmente o risco de sobreajuste em relação ao *Gradient Boosting* tradicional. Além disso, apresenta grande eficiência computacional, com suporte à paralelização e otimizações em memória, e oferece flexibilidade, podendo ser aplicado a tarefas de classificação, regressão e ranking. Outra vantagem importante é a capacidade de capturar padrões complexos, o que o torna especialmente útil em conjuntos de dados com alta dimensionalidade.

Principais hiperparâmetros do *XGBoost* (Chen; Guestrin, 2016; Géron, 2023):

- **n_estimators (número de árvores):** define o número de árvores adicionadas sequencialmente ao *Ensemble*, é fundamental para o desempenho do *XGBoost*. Quando utilizado em valores elevados, permite que o modelo capture padrões mais complexos; entretanto, pode aumentar o risco de sobreajuste caso não seja ajustado em conjunto com um *learning rate* apropriado. Por outro lado, valores muito baixos podem simplificar excessivamente o modelo, levando ao subajuste. Portanto, é comum combinar um número maior de árvores com uma taxa de aprendizado mais baixa, garantindo um aprendizado progressivo e estável.

- **max_depth (profundidade máxima das árvores):** Define a profundidade máxima das árvores no *XGBoost*. Árvores mais profundas são capazes de capturar interações complexas entre variáveis, mas apresentam maior risco de memorizar os dados de treino, levando ao sobreajuste. Por outro lado, árvores mais rasas tendem a

favorecer a generalização, embora possam não conseguir representar relações mais sofisticadas entre os dados. No contexto de manutenção preditiva, é recomendável optar por valores moderados de profundidade, pois isso evita que ruídos presentes nos sinais vibroacústicos sejam interpretados como padrões relevantes pelo modelo.

- ***learning*_rate (η):** Controla a escala da contribuição de cada árvore adicionada ao *Ensemble*. Quando são utilizados valores baixos (por exemplo, η entre 0,01 e 0,1), o aprendizado do modelo ocorre de forma mais lenta, porém tende a ser mais estável, normalmente exigindo um número maior de árvores (n_estimators alto) para alcançar boa performance. Por outro lado, valores mais elevados de *learning_rate* (acima de 0,3) aceleram o aprendizado, mas aumentam o risco de sobreajuste. Assim, o *learning_rate* atua como um "passo de *Gradiente*" no *boosting*, equilibrando a velocidade de convergência do modelo com sua capacidade de generalização.

- **colsample_bytree (fração de atributos por árvore):** define a fração de atributos considerada em cada árvore construída pelo *XGBoost*. Valores elevados fazem com que cada árvore utilize praticamente todas as variáveis, o que pode aumentar a correlação entre elas. Por outro lado, valores mais baixos (geralmente entre 0.3 e 0.8) promovem maior diversidade, já que cada árvore trabalha com subconjuntos diferentes de atributos. Essa abordagem é especialmente útil em cenários com grande quantidade de *features* derivadas de sinais, pois evita que todas as árvores dependam dos mesmos atributos e contribui para um modelo mais robusto.

- **γ:** Representa a mínima perda necessária para divisão, controla o ganho mínimo de informação para que um nó seja dividido em uma árvore de decisão. Valores baixos para γ (por exemplo, próximos de zero) fazem com que o modelo crie árvores bastante detalhadas, capazes de capturar até padrões irrelevantes dos dados, enquanto valores mais altos tornam o modelo mais conservador, exigindo divisões apenas quando há ganhos realmente significativos. Dessa forma, γ atua como uma forma de poda preventiva, reduzindo a quantidade de divisões desnecessárias e contribuindo para um modelo mais robusto e generalizável.

- $\boldsymbol{\lambda}$(**regularização L^2**): Representada por λ, penaliza grandes valores nos pesos das folhas utilizando a norma quadrática. Esse mecanismo mantém os valores das

folhas mais "encolhidos", evitando que uma única árvore atribua pontuações extremas. Dessa forma, a regularização L^2 contribui para reduzir a variância e aumentar a estabilidade das predições do modelo.

- $\boldsymbol{\alpha}(\boldsymbol{regularização\ L^1})$: Representada por α, penaliza o valor absoluto dos pesos das folhas, utilizando a norma Manhattan. Esse mecanismo pode forçar alguns pesos a zero, promovendo a eliminação de variáveis irrelevantes do modelo e atuando como um método de seleção de atributos dentro do processo de *boosting*.

2.4.2.3 Ensemble Learning (Soft Voting)

O *Ensemble Learning* é uma estratégia em que múltiplos modelos individuais (*base learners*) são combinados para produzir uma predição final mais robusta e precisa. A intuição subjacente é que, ao integrar modelos com diferentes vieses e pontos fortes, o *Ensemble* consegue reduzir erros de generalização e superar o desempenho dos classificadores individuais. Hansen e Salamon (1990) demonstraram pioneiramente que a combinação de classificadores, como redes neurais, reduz o erro de generalização e aumenta a estabilidade do aprendizado. Posteriormente, Dietterich (2000) sistematizou os principais métodos de *Ensemble*, incluindo *bagging*, *boosting*, *stacking* e votações (*hard* e *Soft Voting*), tornando-se uma referência fundamental na área.

Uma forma simples, porém eficaz, é o *voting classifier*, no qual a predição final é feita a partir da agregação das saídas de vários modelos. No *hard volting*, cada classificador fornece apenas a classe prevista, e a decisão final é dada pela moda dessas predições (maioria simples). Já no *Soft Voting,* cada classificador fornece probabilidades de classe, e a saída final é obtida pela média ponderada dessas probabilidades, selecionando-se a classe com maior valor agregado.

Formalmente, seja $\hat{y}_{m,k}(x)$ a probabilidade atribuída ao exemplo x pela m-ésima base, e M o número total de classificadores no *Ensemble*. A predição final pelo *Soft Voting* é dada pela Equação 29, onde $\hat{y}_{m,k}(x)$ é a probabilidade prevista pelo classificador *m* para a classe *k*.

$$\hat{y} = \arg\max_{k} \left(\frac{1}{M} \sum_{m=1}^{M} \hat{y}_{m,k}(x) \right) \quad \textbf{(29)}$$

O *Soft Voting* apresenta várias vantagens, conforme destacado por Dietterich (2000): trata-se de uma abordagem que confere maior robustez ao combinar diferentes algoritmos, como SVM, RF, *XGBoost* e MLP. Além disso, contribui para uma melhor generalização do modelo ao reduzir tanto o viés quanto a variância. Outro ponto positivo é a sua simplicidade, sendo fácil de implementar e interpretar quando comparado a outros métodos de *Ensemble* mais sofisticados.

Apesar de suas vantagens, o *Soft Voting* apresenta algumas limitações importantes, como destacado por Mohammed e Kora (2023). A eficácia dessa abordagem depende fortemente da diversidade entre os modelos base, de modo que, quando os classificadores são muito semelhantes, o ganho em desempenho tende a ser reduzido. Além disso, o *Soft Voting* pode implicar maior custo computacional, uma vez que requer o treinamento e a execução simultânea de múltiplos algoritmos.

2.4.3 Modelos de Aprendizado Profundo.

A literatura estabelece uma distinção clara entre os conceitos de Aprendizado de Máquina (*Machine Learning*) e Aprendizado Profundo (*Deep Learning*), posicionando este último como uma subárea especializada dentro de um campo mais amplo (Géron, 2023). O Aprendizado Profundo é caracterizado pelo uso de redes neurais artificiais profundas, isto é, arquiteturas compostas por múltiplas camadas capazes de aprender representações hierárquicas dos dados. O termo *Deep Learning* consolidou-se a partir dos avanços apresentados por Hinton et al. (2006), que demonstraram a viabilidade do treinamento eficaz de redes neurais profundas, alcançando desempenhos superiores aos das técnicas tradicionais da época, especialmente em tarefas complexas como reconhecimento de padrões. Em contrapartida, essas arquiteturas distinguem-se por demandarem grandes volumes de dados, elevado poder computacional e maior esforço de ajuste, sendo particularmente indicadas para problemas de alta complexidade estrutural, como reconhecimento de imagens, sinais acústicos, linguagem natural e, no contexto deste trabalho, o diagnóstico de falhas a partir de sinais vibroacústicos e representações tempo-frequência.

2.4.3.1 Multilayer Perceptron (MLP)

As redes do tipo Perceptron Multicamadas (MLP) constituem um dos modelos mais clássicos de redes neurais profundas, sendo atualmente enquadradas no escopo do Aprendizado Profundo, embora tenham sido propostas antes da consolidação formal dessa subárea. Esse modelo começou a se popularizar a partir do trabalho Rumelhart, Hinton e Williams, (1986), que introduziu o algoritmo de retropropagação (*backpropagation*). Essas redes são compostas por uma sequência de camadas de neurônios artificiais interconectados, em que cada neurônio aplica uma transformação linear sobre os dados de entrada, seguida por uma função de ativação não linear, o que permite ao modelo capturar padrões complexos e aprender representações não lineares dos dados.

De acordo com Haykin, (2009), para uma entrada $x \in R^n$, a saída de uma camada escondida por ser expressa de acordo com a Equação 30.

$$h^{(l)} = \sigma\left(W^{(l)}h^{(l-1)} + b^{(l)}\right) \quad \textbf{(30)}$$

Onde:

$\boldsymbol{h^{(l-1)}}$ é a saída da camada anterior (com $\boldsymbol{h^{(0)}}$**= x**),

$\boldsymbol{W^{(l)}}$ é a matriz de pesos da camada $\boldsymbol{l}$,

$\boldsymbol{b^{(l)}}$ é o vetor de vieses,

$\boldsymbol{\sigma}(.)$ é a função de ativação (ReLU, sigmoide, tanh, etc.).

A camada de saída determina a dimensionalidade, a função de ativação e a função de perda do modelo, variando conforme o tipo de problema. Na classificação binária, emprega-se um único neurônio que produz o logito z = $W^{(L)}h^{(L-1)} + b^{(L)}$, convertido em probabilidade pela sigmoide $\hat{y} = \sigma$ (z) (ver Equação (31)). Os rótulos são y ∈ {0,1} e a perda típica é a entropia cruzada binária; a decisão de classe decorre da aplicação de um limiar τ (padrão 0,5, ajustável via ROC/PR ou custos de erro) (Haykin, 2009).

Já na classificação multiclasse mutuamente exclusiva (K classes), utiliza-se uma camada com K neurônios; o vetor de *logits* z = $W^{(L)}h^{(L-1)} + b^{(L)}$ é transformado pela softmax $\hat{y} = softmax(z_i)$ (ver Equação (32)).A softmax converte *logits* em probabilidades não negativas

que somam 1, conforme a Equação (33) (Géron, 2023).

$$\hat{y} = \sigma\left(W^{(L)} h^{(L-1)} + b^{(L)}\right) \quad \textbf{(31)}$$

$$\hat{y} = softmax\left(W^{(L)} h^{(L-1)} + b^{(L)}\right) \quad \textbf{(32)}$$

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}, \quad i = 1, \ldots, K \quad \textbf{(33)}$$

O MLP apresenta como vantagens a capacidade de capturar padrões não lineares complexos, algo impossível para modelos lineares, além de oferecer flexibilidade para diferentes tarefas, como classificação, regressão e redução de dimensionalidade. Destaca-se ainda pelo bom desempenho em conjuntos de dados de alta dimensionalidade, como sinais vibroacústicos convertidos em *features* ou espectrogramas (Haykin, 2009; Inturi *et al*., 2023; Praveen Kumar *et al*., 2019):

Entre as principais limitações do MLP destacam-se a necessidade por elevado poder computacional e o ajuste minucioso de hiperparâmetros, além da tendência ao sobreajuste em conjuntos de dados pequenos, caso não seja empregada uma regularização adequada. O modelo também apresenta menor interpretabilidade quando comparado a técnicas baseadas em árvores, o que pode dificultar a compreensão dos critérios de decisão empregados pela rede (Al Mamun et al., 2023; Chen; Jin; Jiri, 2018; Haykin, 2009)

A seguir, apresentam-se os principais hiperparâmetros frequentemente ajustados na prática (Géron, 2023; Haykin, 2009)

- **n_hidden_layers e n_neurons:** A configuração de n_hidden_layers e n_neurons determina o número de camadas ocultas e a quantidade de neurônios presentes em cada uma delas. O aumento do número de camadas e de neurônios eleva a

capacidade de representação do modelo, podendo capturar padrões mais complexos; porém, esse acréscimo também aumenta o risco de sobreajuste e exige maior poder computacional. Por outro lado, utilizar poucas camadas pode simplificar demais a arquitetura, levando ao subajuste e à limitação do desempenho.

- **activation (função de ativação):**Determina a não linearidade aplicada em cada neurônio.

    - **ReLU (Rectified Linear Unit):** mais utilizada por sua simplicidade e bom desempenho em redes profundas.

    - **Sigmoide e tanh:** clássicas, mas sofrem com *Gradient*es saturados em redes maiores. A escolha da ativação impacta diretamente a **capacidade de aprendizado** do modelo.

- **learning_rate:** A *learning_rate*, ou taxa de aprendizado, é utilizada pelo algoritmo de otimização para controlar o tamanho dos ajustes feitos nos pesos durante o treinamento. Valores altos de *learning*_rate aceleram o processo de aprendizagem, porém podem dificultar a convergência do modelo, levando a oscilações ou ultrapassagens do ponto ótimo. Já valores baixos tornam o aprendizado mais estável e preciso, mas exigem um número maior de épocas de treino para que o modelo atinja um desempenho satisfatório.

  • **optimizer (algoritmo de otimização):** O algoritmo de otimização define a forma como os pesos da rede neural são atualizados a partir do *Gradient*e da função de perda. Entre os métodos clássicos destaca-se o *Stochastic Gradient Descent* (SGD)**,** que realiza atualizações iterativas com base em subconjuntos dos dados, podendo ser combinado com momentum para acelerar a convergência e reduzir oscilações em regiões de vale da superfície de erro. Outras variações amplamente utilizadas incluem o RMSProp, que ajusta a taxa de aprendizado de forma adaptativa com base na média móvel dos *Gradient*es ao quadrado, e o Adagrad, que adapta o passo de aprendizado individualmente para cada parâmetro, sendo eficaz em problemas esparsos, porém suscetível a taxas de aprendizado excessivamente pequenas ao longo do treinamento. Entre os métodos mais adotados em aplicações práticas encontra-se o Adam (*Adaptive Moment Estimation*)**,** que combina os princípios do

momentum e do RMSProp por meio da estimativa adaptativa de primeira e segunda ordem dos *Gradient*es. Esse algoritmo apresenta convergência mais rápida e estável, especialmente em problemas de alta dimensionalidade e com *Gradient*es ruidosos, sendo amplamente empregado no treinamento de MLPs e redes neurais profundas em aplicações industriais e de processamento de sinais(Géron, 2023; Haykin, 2009).

- **batch_size:** O parâmetro batch_size refere-se ao número de amostras utilizado em cada atualização dos pesos durante o treinamento. Mini-batches pequenos tendem a gerar maior variabilidade no *Gradient*e, favorecendo a generalização do modelo, enquanto mini-batches grandes proporcionam um treinamento mais estável e rápido, embora possam levar o modelo a encontrar mínimos locais menos favoráveis..

- **max_iter / epochs:** O hiperparâmetro max_iter, também conhecido como número máximo de épocas, define a quantidade de ciclos de treinamento realizados pelo modelo. Um número maior de épocas pode permitir um ajuste mais refinado dos parâmetros, mas também aumenta o risco de sobreajuste. Por isso, o valor desse parâmetro deve ser cuidadosamente equilibrado, e muitas vezes é combinado com técnicas como o *early stopping* para evitar que o modelo se ajuste excessivamente aos dados de treinamento.

- **regularização (L^2, *dropout*):** Entre as principais estratégias de regularização para mitigar o sobreajuste destacam-se o L^2 (ridge) e o *dropout*. O L^2 penaliza a presença de pesos muito elevados na rede, promovendo uma solução mais generalizável, enquanto o *dropout* atua desativando aleatoriamente neurônios durante o treinamento, o que força o modelo a aprender representações mais robustas e menos dependentes de unidades específicas.

- **early stopping:** Interrompe o treinamento quando o desempenho no conjunto de validação deixa de melhorar após certo número de épocas. Evita sobreajuste e economiza tempo computacional.

2.4.3.2 Redes Neurais Convolucionais (CNN)

As Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNNs) constituem uma classe de arquiteturas de *Deep Learning* projetadas para processar dados com estrutura topológica local, como imagens, sinais bidimensionais e representações tempo-frequência. A formulação moderna das CNNs foi consolidada por Lecun et al., (1998) no trabalho seminal *Gradient-Based Learning Applied to Document Recognition*, amplamente reconhecido como o marco fundador dessa arquitetura.

Nesse trabalho, os autores definem as CNNs como arquiteturas especializadas que incorporam, em sua própria estrutura, conhecimento prévio sobre invariâncias espaciais, como translação, pequenas distorções e variações locais de escala. Diferentemente de redes totalmente conectadas, que ignoram a organização espacial da entrada e demandam um número excessivo de parâmetros, as CNNs exploram explicitamente a correlação local entre elementos vizinhos, reduzindo a complexidade do modelo e melhorando sua capacidade de generalização.

Segundo Lecun et al., (1998), a arquitetura das CNNs baseia-se na combinação de três princípios fundamentais: campos receptivos locais, compartilhamento de pesos e subamostragem. Esses elementos definem não apenas a estrutura da rede, mas também sua motivação conceitual.

A materialização clássica desses princípios arquiteturais é a arquitetura *LeNet-5*, também proposta por Lecun et al., (1998). Originalmente desenvolvida para o reconhecimento de dígitos manuscritos, a *LeNet-5* consolidou o uso de camadas convolucionais intercaladas com etapas de subamostragem, seguidas por camadas totalmente conectadas responsáveis pela classificação. Essa arquitetura tornou-se um referencial histórico e conceitual no desenvolvimento das CNNs, influenciando uma ampla gama de modelos posteriores e permanecendo como base para diversas adaptações empregadas em aplicações contemporâneas, incluindo o diagnóstico de falhas a partir de representações tempo-frequência.

2.4.3.2.1 Campos Receptivos Locais

O primeiro princípio arquitetural das CNNs é o uso de campos receptivos locais (local receptive fields). Em contraste com redes totalmente conectadas, nas quais cada neurônio recebe informações de toda a camada anterior, nas CNNs cada neurônio é conectado apenas a uma pequena vizinhança local da entrada. Essa estratégia permite que a rede aprenda padrões elementares locais, como bordas, transições abruptas e estruturas simples, que são posteriormente combinadas em representações mais complexas nas camadas profundas.

Esse mecanismo explora a propriedade de localidade presente em dados estruturados, como imagens e sinais tempo-frequência, nos quais valores próximos tendem a ser mais

correlacionados do que valores distantes. Dessa forma, a CNN consegue capturar características relevantes preservando a topologia da entrada.

2.4.3.2.2 Compartilhamento de Pesos e Operação de Convolução

O segundo princípio fundamental é o compartilhamento de pesos (*shared weights*). Lecun et al., (1998) definem que neurônios pertencentes a um mesmo plano ou mapa de características utilizam o mesmo conjunto de pesos. Isso implica que um mesmo detector de padrão é aplicado em diferentes posições da entrada, permitindo que a rede reconheça a mesma característica independentemente de sua localização espacial.

Matematicamente, essa operação equivale a uma convolução entre a entrada e um filtro (*kernel*), seguida da adição de um termo de viés e da aplicação de uma função de ativação não linear. É justamente essa equivalência que dá nome às Redes Neurais Convolucionais. O compartilhamento de pesos reduz drasticamente o número de parâmetros do modelo, tornando o treinamento viável e contribuindo para maior robustez frente a variações de posição e pequenas distorções.

2.4.3.2.3 Subamostragem (Pooling)

O terceiro componente arquitetural destacado no trabalho fundador é a subamostragem (*subsampling*), atualmente conhecida como *Pooling*. Camadas de subamostragem são intercaladas com camadas convolucionais e têm como objetivo reduzir progressivamente a resolução espacial dos mapas de características. Esse processo torna as representações internas menos sensíveis a pequenas variações de posição, ruído e distorções locais.

Além de contribuir para a invariância espacial, a subamostragem reduz o custo computacional e ajuda a controlar o sobreajuste, ao forçar a rede a aprender representações mais compactas e abstratas.

2.4.3.2.4 Definição conceitual de CNNs

De forma conceitual, Lecun et al., (1998) definem as CNNs como arquiteturas capazes de sintetizar automaticamente seus próprios extratores de características diretamente a partir dos dados de entrada, dispensando a necessidade de engenharia manual de atributos. Ao impor restrições estruturais, como campos receptivos locais e compartilhamento de pesos, a rede aprende representações hierárquicas cada vez mais abstratas, mantendo eficiência computacional e boa capacidade de generalização.

Embora originalmente desenvolvidas para reconhecimento de padrões visuais, diversos estudos mostram que as CNNs passaram a ser amplamente aplicadas a outros tipos de dados

estruturados, incluindo sinais temporais e representações tempo-frequência. Em particular, espectrogramas de sinais acústicos e de vibração podem ser interpretados como imagens bidimensionais, tornando as CNNs especialmente adequadas para aplicações de diagnóstico (Di Maggio, 2023; Gültekin et al., 2022; Hasan; Islam; Kim, 2019; Islam; Kim, 2019; Siddique et al., 2023)

2.4.3.2.5 Redes Neurais Convolucionais Bidimensionais (CNN 2D)

As Redes Neurais Convolucionais bidimensionais (CNN 2D) são particularmente adequadas para o processamento de dados organizados em grades bidimensionais, como imagens e representações tempo-frequência. No contexto do diagnóstico de falhas, espectrogramas obtidos a partir de sinais acústicos e de vibração podem ser interpretados como imagens, nas quais um eixo representa o tempo, o outro a frequência e a intensidade dos pixels corresponde à energia espectral. Essa estrutura permite a aplicação direta de CNNs 2D para a extração automática de padrões discriminantes (Siddique et al., 2023).

A operação fundamental de uma CNN 2D é a convolução bidimensional, que consiste na aplicação de filtros locais sobre a entrada, preservando a topologia espacial e explorando correlações locais entre regiões adjacentes do sinal.

- **Operação de convolução bidimensional:**

Considere uma entrada bidimensional $X \in \mathbb{R}^{H \times W}$, representando um espectrograma com altura H(frequências) e largura W(tempo), e um filtro convolucional (*kernel*) $K \in \mathbb{R}^{h \times w}$. A saída da operação de convolução, denominada mapa de características (*feature map*), é dada por:

$$Y(i,j) = \sum_{m=0}^{h-1} \sum_{n=0}^{w-1} X(i+m, j+n)\ K(m,n) \quad \textbf{(34)}$$

onde (i, j)denota a posição espacial do filtro sobre a entrada. Na prática, adiciona-se um termo de viés be aplica-se uma função de ativação não linear $\sigma(\cdot)$, resultando em:

$$Z(i,j) = \sigma\left(\sum_{m=0}^{h-1} \sum_{n=0}^{w-1} X(i+m, j+n)\ K(m,n) + b \right) \quad \textbf{(35)}$$

Essa operação é repetida ao longo de toda a entrada, produzindo um mapa de ativações que indica a presença do padrão aprendido pelo filtro em diferentes regiões do espectrograma.

- **Mapas de características e aprendizado hierárquico**

Cada filtro convolucional aprende a responder a um padrão específico presente nos dados de entrada. Em camadas iniciais, esses padrões tendem a corresponder a estruturas simples, como bordas, transições abruptas de energia ou bandas espectrais dominantes. À medida que a profundidade da rede aumenta, os mapas de características passam a representar combinações mais complexas desses padrões elementares, capturando assinaturas associadas a regimes operacionais ou estados de falha.

Formalmente, considerando uma camada convolucional com F filtros, a saída pode ser representada como um tensor tridimensional:

$$Z \in R^{H' \times W' \times F}$$

onde cada fatia $\mathbf{Z}^{(f)}$corresponde ao mapa de características produzido pelo filtro f.

- **Compartilhamento de pesos e redução de complexidade.**

Um aspecto central da CNN 2D é o compartilhamento de pesos, no qual o mesmo *kernel* Ké aplicado em todas as posições da entrada. Isso reduz drasticamente o número de parâmetros do modelo quando comparado a redes totalmente conectadas. Enquanto uma camada densa exigiria $H \times W$pesos por neurônio, uma camada convolucional requer apenas $h \times w$pesos por filtro, independentemente do tamanho da entrada.

Essa propriedade é particularmente relevante no processamento de espectrogramas, que podem apresentar alta dimensionalidade, tornando inviável o uso de arquiteturas totalmente conectadas sem risco elevado de sobreajuste.

- **Camadas de *Pooling***

Após camadas convolucionais, é comum empregar camadas de *pooling*, cujo objetivo é reduzir a dimensionalidade espacial dos mapas de características e tornar a representação menos sensível a pequenas variações locais. No *pooling* máximo (*max pooling*), por exemplo, a saída

é definida como:

$$Y_{\text{pool}}(i,j) = \max_{(m,n)\in\Omega} Z\,(i+m, j+n) \tag{36}$$

onde **Ω**representa a região de *pooling*, tipicamente uma janela $\boldsymbol{p} \times \boldsymbol{p}$. Essa operação preserva as ativações mais relevantes, reduzindo o custo computacional e contribuindo para maior invariância a deslocamentos locais no espectrograma**.**

- **Camada de classificação**

Após a extração hierárquica de características por meio de camadas convolucionais e de *pooling*, os mapas de características são reorganizados em um vetor unidimensional (*flattening*) e fornecidos a camadas totalmente conectadas ou diretamente a uma camada de saída. Em problemas de classificação, a camada final produz um vetor de logits **z**, convertido em probabilidades por meio de funções como a sigmoide (classificação binária) ou a softmax (classificação multiclasse):

$$\widehat{y_k} = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}, \quad k = 1, \ldots, K \tag{37}$$

O treinamento da CNN é realizado por retropropagação do erro, minimizando uma função de perda apropriada, como a entropia cruzada, ajustando iterativamente os pesos dos filtros convolucionais e das camadas finais.

- **Aplicação a espectrogramas vibroacústicos**

No presente trabalho, as CNNs 2D são aplicadas a espectrogramas de sinais acústicos e de vibração, explorando simultaneamente padrões temporais e espectrais associados a diferentes condições operacionais da máquina. A capacidade das CNNs de aprender automaticamente representações discriminantes a partir dessas imagens tempo-frequência reduz a dependência de engenharia manual de atributos e aumenta a robustez do diagnóstico em cenários ruidosos e com variabilidade operacional**.**

- **Hiperparâmetros de Redes Neurais Convolucionais Bidimensionais**

Assim como em outras arquiteturas de aprendizado profundo, o desempenho das Redes Neurais Convolucionais bidimensionais (CNN 2D) depende fortemente da escolha adequada de seus hiperparâmetros, os quais controlam tanto a capacidade de representação do modelo quanto seu comportamento durante o treinamento. A seguir, apresentam-se os principais hiperparâmetros frequentemente ajustados na prática (Géron, 2023; Haykin, 2009):

- **n_conv_layers e n_filters:** A configuração de *n_conv_layers* define o número de camadas convolucionais empilhadas na arquitetura, enquanto n_filters especifica a quantidade de filtros (*kernel*s) em cada camada. O aumento do número de camadas e de filtros eleva a capacidade do modelo de capturar padrões hierárquicos e estruturas complexas presentes nos espectrogramas. Entretanto, arquiteturas muito profundas ou com excesso de filtros aumentam o risco de sobreajuste e o custo computacional. Por outro lado, configurações excessivamente simples podem levar ao subajuste, limitando a capacidade discriminativa da rede.
- **kernel_size (tamanho do filtro):** O parâmetro *kernel*_size define as dimensões espaciais dos filtros convolucionais (por exemplo, $3 \times 3, 5 \times 5$). Filtros menores tendem a capturar padrões locais finos, como transições abruptas de energia, enquanto filtros maiores agregam informações em regiões mais amplas do espectrograma. A escolha do tamanho do *kernel* influencia diretamente o tipo de padrão aprendido e deve considerar a resolução tempo-frequência da representação utilizada.
- **stride e padding**: O *stride* determina o deslocamento do filtro a cada aplicação da convolução. *Strides* maiores reduzem a dimensionalidade espacial mais rapidamente, diminuindo o custo computacional, mas podem resultar em perda de informação. O *padding* controla a adição de bordas artificiais à entrada, permitindo preservar as dimensões espaciais dos mapas de características e evitar a perda de informações nas regiões periféricas do espectrograma.
- **activation (função de ativação):** A função de ativação introduz não linearidade ao modelo. A função ReLU (*Rectified Linear Unit*) é a mais empregada em CNNs profundas, devido à sua simplicidade computacional e à mitigação do problema do *Gradient*e desvanescente. Funções clássicas como sigmoide e tanh podem ser utilizadas em camadas específicas, porém tendem a apresentar

saturação em arquiteturas profundas, o que dificulta o aprendizado eficiente.

- ***Pooling*_type e *Pooling*_size**: O tipo de *pooling* (máximo ou médio) e o tamanho da janela de *pooling* controlam a redução espacial dos mapas de características. O *max pooling* é amplamente utilizado por preservar as ativações mais relevantes, enquanto o *average Pooling* produz representações mais suavizadas. A escolha desses parâmetros impacta diretamente o grau de invariância a deslocamentos locais e o nível de abstração das *features* aprendidas.
- ***learning*_rate:** A taxa de aprendizado regula o tamanho das atualizações dos pesos durante o treinamento. Valores elevados podem acelerar a convergência, porém aumentam o risco de instabilidade e divergência do processo de otimização. Valores muito baixos tornam o treinamento mais estável, mas exigem maior número de épocas. Assim como em MLPs, a *learning*_rate exerce papel central no equilíbrio entre velocidade de aprendizado e qualidade da solução.
- **optimizer (algoritmo de otimização):** O algoritmo de otimização define como os *Gradient*es da função de perda são utilizados para atualizar os pesos da rede. Métodos clássicos como o *Stochastic Gradient Descent* (SGD) podem ser combinados com momentum para acelerar a convergência. Em aplicações práticas, otimizadores adaptativos como RMSProp e Adam são amplamente utilizados em CNNs, pois ajustam automaticamente a taxa de aprendizado para cada parâmetro, apresentando convergência mais rápida e estável, especialmente em problemas de alta dimensionalidade e dados ruidosos.
- **batch_size**: O parâmetro *batch_size* corresponde ao número de amostras processadas antes da atualização dos pesos. *Mini-batches* menores tendem a introduzir maior variabilidade no *Gradient*e, favorecendo a generalização, enquanto batches maiores proporcionam treinamento mais estável e eficiente do ponto de vista computacional, podendo, contudo, levar a mínimos locais menos favoráveis.
- **epochs (ou max_iter):** Define o número máximo de ciclos de treinamento realizados pelo modelo. Um número elevado de épocas permite ajustes mais refinados dos parâmetros, mas aumenta o risco de sobreajuste. Por esse motivo, esse hiperparâmetro é frequentemente combinado com estratégias de parada

antecipada.

- **regularização (L^2 e dropout):** Para mitigar o sobreajuste, são frequentemente empregadas técnicas de regularização. A regularização L^2 penaliza pesos excessivamente elevados, promovendo soluções mais suaves e generalizáveis. O *dropout* atua desativando aleatoriamente neurônios durante o treinamento, reduzindo a coadaptação entre unidades e aumentando a robustez do modelo.
- **early stopping:** O *early stopping* interrompe o treinamento quando o desempenho no conjunto de validação deixa de melhorar após determinado número de épocas. Essa estratégia previne o sobreajuste e reduz o custo computacional, sendo amplamente utilizada no treinamento de CNNs aplicadas a problemas industriais.

3 Metodologia

A metodologia deste trabalho foi organizada de forma sequencial, inspirada no CRISP-DM (Cross Industry Standard Process for Data Mining), modelo amplamente consolidado para condução de projetos de ciência de dados (CHAPMAN et al., 2000). Essa abordagem estruturada permite narrar o desenvolvimento da pesquisa em etapas cronológicas, partindo da compreensão do problema até a avaliação dos modelos preditivos. Assim, as fases de compreensão do problema, compreensão e preparação dos dados, modelagem e avaliação foram adaptadas ao contexto da detecção de falhas em máquinas de fabricação de copos de papel, compondo um fluxo metodológico claro e replicável. A figura 1, ilustra as fases do modelo.

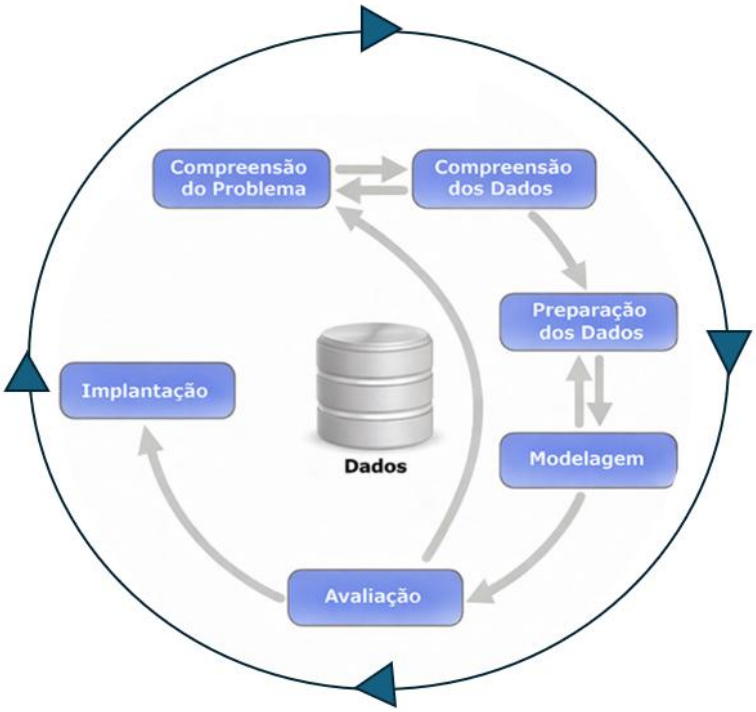

Figura 1 - Etapas do CRISP-DM

Adaptado de (CHAPMAN, et al., 2000)

3.1 Compreensão do problema

O ponto de partida desta pesquisa foi a identificação de um desafio recorrente em ambientes industriais: a ocorrência de falhas mecânicas em máquinas de fabricação de copos de papel, que podem comprometer a qualidade do produto e interromper o processo produtivo.

No caso específico da máquina formadora de potes de papel PT80 (Zhejiang New Debao), observou-se que determinados subsistemas apresentam maior suscetibilidade a defeitos, em especial o sistema de transmissão por correntes, responsável pelo sincronismo entre as etapas do processo.

Entre os modos de falha relevantes estão: (i) a folga em correntes de transmissão, que eleva os níveis de vibração e pode causar perda de sincronismo; (ii) a baixa pressão no regravador, que compromete a colagem do fundo ao corpo do copo; e (iii) o desalinhamento ou desgaste da faca de corte, que gera cortes assimétricos e risco de travamentos. Esses cenários são importantes para compreender o funcionamento da máquina e ilustrar a diversidade de problemas potenciais no processo produtivo.

Entretanto, o escopo deste trabalho foi delimitado à análise da Corrente 1 do sistema de transmissão, escolhida por seu papel fundamental na estabilidade e sincronismo do equipamento. Dessa forma, embora outros defeitos sejam reconhecidos e descritos como parte do contexto, apenas a detecção de falhas na Corrente 1 (ver figura 5) foi efetivamente investigada nas etapas experimentais e de modelagem preditiva.

3.2 Compreensão dos dados

Após a definição do problema, a etapa seguinte consistiu em compreender o sistema estudado e os dados disponíveis para a investigação. O equipamento utilizado foi a máquina formadora de potes de papel modelo PT80, fabricada em 2012 pela Zhejiang New Debao Machine Co., Ltd. Trata-se de um modelo de baixa rotação (*Low Speed Paper Bowl Machine*), projetado para operar a velocidades nominal entre 20 e 35 peças por minuto, com potência nominal de 3,8 kW e alimentação trifásica de 380 V / 50–60 Hz. O equipamento pesa aproximadamente 1.800 kg e é capaz de produzir potes de papel de 80 mL. A figura 2 mostra uma foto da máquina, e a 3, um exemplo de produto final.

O processo produtivo, ilustrado na figura 4, ocorre de forma sequencial e sincronizada, abrangendo cinco etapas principais: (i) conformação do corpo do copo por meio da dobra do papel, (ii) selagem lateral por ultrassom, (iii) corte do fundo por faca circular, (iv) colagem do fundo ao corpo com aquecimento e pressão em dois estágios, e (v) formação da borda para acabamento e resistência. O sincronismo dessas etapas é garantido por um sistema de transmissão acionado por motor trifásico, regulado por um inversor de frequência (Omron SYSDRIVE 3G3JZ, 400 V / 1,5 kW), cujo movimento é distribuído a todo o sistema por meio de três correntes principais. A figura 5 mostra os principais elementos da máquina.

Nesta pesquisa, o foco da análise foi a detecção de folga na Corrente 1, considerada elemento crítico para o sincronismo da máquina. Para enriquecer os experimentos e avaliar a robustez do diagnóstico, foram realizadas manipulações controladas nas Correntes 2 e 3, que também puderam ser configuradas em condição normal ou com folga. Essa estratégia permitiu analisar interações entre as correntes, verificando como falhas simultâneas ou combinadas impactam a assinatura vibro-acústica do equipamento.

A coleta de dados foi viabilizada por meio de dois tipos de sinais multissensoriais:

- Áudio: captado pelo microfone interno de um notebook HP, operando a 44,1 kHz (configuração padrão do sistema operacional), com gravação realizada no software *Audacity* 3.6.3. Os arquivos foram exportados em formato WAV sem perdas, assegurando fidelidade ao sinal registrado.
- Vibração: registrada pelo acelerômetro triaxial de um iPhone 13, utilizando o aplicativo *Physics Toolbox Sensor Suite*, com taxa de amostragem de 100 Hz (valor máximo disponível na versão utilizada do aplicativo) e exportação em arquivos CSV com marca temporal.

Para assegurar variabilidade e representatividade, os experimentos foram conduzidos em duas velocidades principais (20 e 35 Hz), ajustadas por meio do inversor de frequência, que compuseram a base utilizada nas etapas de treinamento e validação. Adicionalmente, foi incluída uma condição intermediária a 27,5 Hz, destinada exclusivamente à etapa de teste, permitindo avaliar a capacidade de generalização dos modelos frente a uma situação inédita, não observada durante o processo de modelagem. O planejamento experimental resultante contemplou a Corrente 1 como alvo do diagnóstico e as Correntes 2 e 3 como variáveis de interação, abrangendo cenários isolados e combinados de falha em diferentes regimes operacionais.

Embora a coleta de dados tenha sido realizada em nível de ensaio, totalizando 19 sessões experimentais, o conjunto de dados utilizado na modelagem não se restringe a esse número de observações. Os sinais acústicos e de vibração obtidos em cada ensaio foram posteriormente segmentados em janelas temporais, conforme detalhado na etapa de preparação dos dados, resultando em cerca de 75 instâncias por ensaio. Dessa forma, a modelagem foi conduzida em nível de janela, e não de ensaio, o que define o tamanho efetivo da amostra utilizada pelos algoritmos de aprendizado de máquina.

Figura 2 - Máquina de Fabricação de Copo (PT80)

Figura 3 - Exemplo de Produto

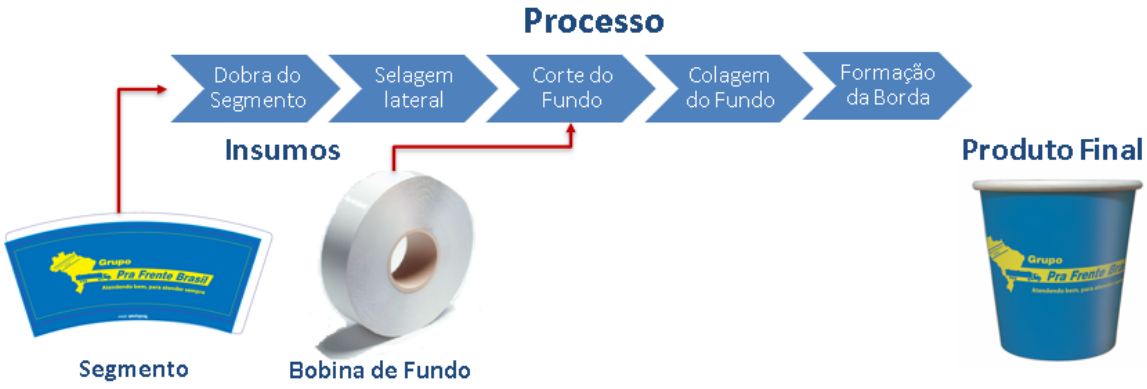

Figura 4 - Processo de operação da máquina

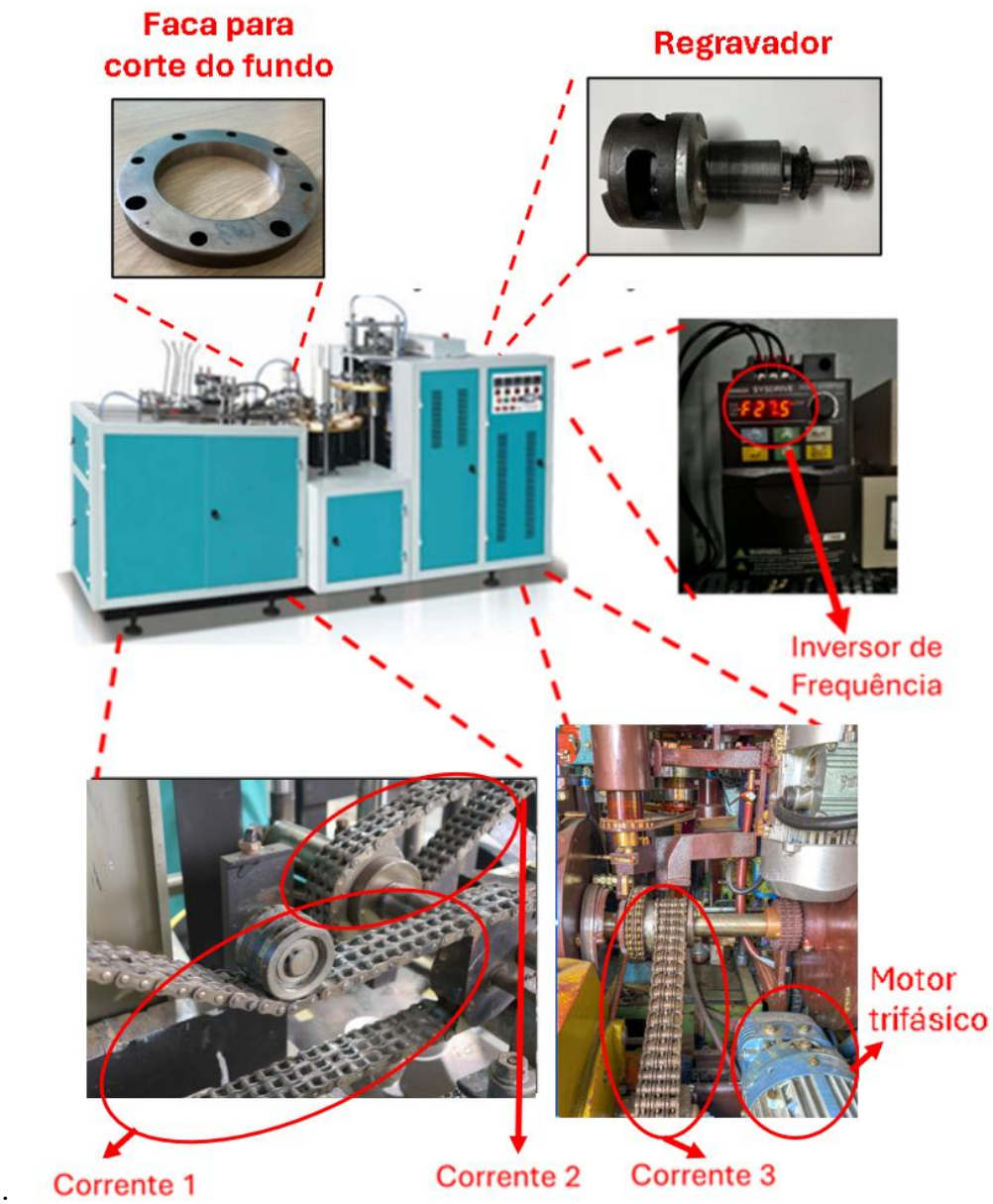

Figura 5 - Principais Elementos da máquina

3.3 Preparação dos dados

Com a compreensão do problema e dos sinais disponíveis, a etapa seguinte consistiu na coleta estruturada e organização dos dados, assegurando padronização e consistência para posterior análise.

3.3.1 Procedimento de Aquisição

A coleta foi realizada diretamente na máquina PT80 em operação real, seguindo um protocolo padronizado:

- **Preparação do ambiente**: antes de cada ensaio, verificou-se que a máquina operava em condições normais de produção. Foram minimizados ruídos externos abruptos (ex.: conversas próximas ou batidas metálicas).
- **Fixação dos dispositivos**: o notebook e o smartphone foram posicionados e fixados sobre a estrutura da máquina, de forma a evitar deslocamentos durante a coleta e garantir reprodutibilidade entre sessões. A escolha desse ponto deveu-se a três fatores principais: (i) acessibilidade, permitindo fácil manuseio dos equipamentos durante o processo de coleta; (ii) qualidade do sinal, uma vez que o local se mostrou adequado para a captação tanto de vibração quanto de áudio; e (iii) viabilidade prática de fixação, oferecendo uma superfície estável que facilitou a montagem do arranjo experimental. (ver figura 6)
- **Captação de áudio**: realizada pelo microfone interno do *notebook* HP, configurado em 44,1 kHz (padrão do sistema operacional), com gravação no software *Audacity* 3.6.3. Os arquivos foram exportados em formato WAV sem perdas, assegurando fidelidade do sinal.
- **Captação de vibração**: conduzida pelo acelerômetro triaxial do iPhone 13 (100 Hz), utilizando o aplicativo *Physics Toolbox Sensor Suite*. Os dados foram exportados em CSV com marca temporal, possibilitando sincronização com o áudio.
    - *Observação técnica:* a taxa de amostragem do acelerômetro utilizado (100 Hz) restringe a análise a componentes até 50 Hz, conforme o teorema de Nyquist. Estudos como o de Tsutada (2007) indicam que falhas associadas a correntes se manifestam predominantemente em baixas frequências (<50 Hz), embora possam existir componentes adicionais próximos a 100 Hz relacionados ao engrenamento. Dessa forma, o arranjo experimental adotado mostrou-se adequado para captar a dinâmica essencial da Corrente 1. Em contrapartida, falhas incipientes em rolamentos apresentam assinaturas vibracionais em faixas de kHz, o que demandaria sensores industriais com maior capacidade de aquisição (Tandon; Choudhury, 1999). Reconhece-se, ainda, a possibilidade teórica de distorção da informação coletada devido a efeitos de *aliasing* associados às harmônicas de rotação, sobretudo na condição de operação a 35 Hz. Entretanto, a análise previa dos sinais no domínio da frequência

indicou que a maior parte da energia espectral concentra-se em frequências inferiores a 10 Hz, sugerindo que a influência prática de componentes *aliased* é limitada no contexto experimental analisado.

- **Duração das gravações**: cada sessão de coleta teve duração entre 2 a 3 minutos, tanto para os arquivos de áudio quanto para os de vibração. Esse tempo foi considerado suficiente para capturar a variabilidade natural da operação da máquina e gerar um número expressivo de janelas de análise. Ressalta-se que cada janela de 5 segundos corresponde aproximadamente a um ciclo completo da máquina, o que garante que a segmentação adotada seja compatível com a dinâmica real do processo produtivo.
- **Sincronização manual**: início e término das gravações de áudio e vibração foram acionados simultaneamente. A marca temporal e a padronização da duração dos ensaios facilitaram a correspondência entre os sinais.
- **Organização dos arquivos**: todos os registros foram nomeados de acordo com a condição experimental (corrente normal ou com folga; velocidade de operação), e organizados em diretórios específicos.
- **Critérios de consistência**: foram descartadas amostras com ruídos externos significativos ou registros incompletos.

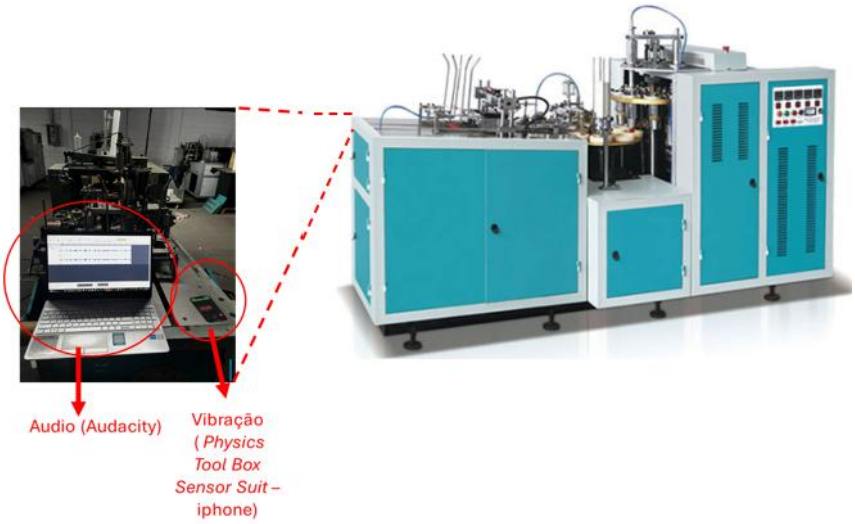

Figura 6 - Coleta de dados: disposição de notebook e smartphone sobre a máquina.

3.3.2 Planejamento Experimental

O delineamento experimental teve como objetivo avaliar a capacidade dos modelos de diagnóstico em identificar folga na Corrente 1, considerada o alvo central desta pesquisa. Para enriquecer a análise e verificar interações, as Correntes 2 e 3 foram manipuladas de forma controlada, assumindo condições normais ou com folga. Dessa forma, foi possível observar não apenas o efeito direto da folga na Corrente 1, mas também o impacto combinado de múltiplas correntes em condição anormal.

Além da configuração das correntes, a velocidade de operação da máquina foi utilizada como variável adicional de ajuste operacional, doravante denominada AJUSTE nas análises. Essa variável corresponde diretamente às condições definidas na matriz de experimentos e representa o regime de funcionamento sob o qual cada ensaio foi realizado, não sendo derivada dos sinais acústicos ou de vibração.

Foram adotadas duas velocidades principais (20 Hz e 35 Hz), destinadas à composição da base de treinamento e validação. Adicionalmente, incluiu-se uma velocidade intermediária (27,5 Hz), empregada exclusivamente na etapa de teste externo, com o objetivo de avaliar a capacidade de generalização dos modelos diante de uma condição operacional inédita, não observada durante o processo de modelagem. A inclusão da variável AJUSTE permite fornecer contexto operacional aos classificadores, reduzindo ambiguidades associadas às variações naturais do comportamento dinâmico da máquina em diferentes regimes de operação, sem caracterizar vazamento de informação (*data leakage*).

A Tabela 3 sintetiza a matriz experimental construída, apresentando os 19 ensaios realizados com diferentes combinações das correntes e das velocidades de operação.

Tabela 3 - Matriz de Experimentos

Ensaio	Corrente 1	Corrente 2	Corrente 3	Velocidade (Hz)
1	Normal	Normal	Normal	20,00
2	Normal	Normal	Normal	35,00
3	Normal	Anormal	Normal	20,00
4	Normal	Anormal	Normal	35,00
5	Anormal	Normal	Normal	20,00
6	Anormal	Normal	Normal	35,00
7	Anormal	Anormal	Normal	20,00
8	Anormal	Anormal	Normal	35,00

9	Normal	Normal	Anormal	20,00
10	Normal	Normal	Anormal	35,00
11	Normal	Anormal	Anormal	20,00
12	Normal	Anormal	Anormal	35,00
13	Anormal	Normal	Anormal	20,00
14	Anormal	Normal	Anormal	35,00
15	Anormal	Anormal	Anormal	20,00
16	Anormal	Anormal	Anormal	35,00
17	Anormal	Anormal	Anormal	27,50
18	Normal	Normal	Normal	20,00
19	Normal	Normal	Normal	35,00

(Nota: “Normal” indica condição sem defeito e “Anormal” refere-se à introdução de folga na corrente correspondente.)

É importante destacar que os Ensaios 1 a 16 correspondem a um arranjo fatorial completo ($2^3 \times 2$ velocidades), cobrindo todas as combinações possíveis de estados das três correntes em dois regimes de operação distintos. Já os Ensaios 17 a 19 foram incluídos como extensões: o Ensaio 17 para a etapa de teste em condição inédita (27,5 Hz) e os Ensaios 18 e 19 como repetições adicionais da condição “Normal”, reforçando a consistência dos dados de referência.

3.4 Modelagem

A etapa de modelagem teve como objetivo desenvolver e comparar diferentes estratégias de classificação para a detecção de folga na Corrente 1 da máquina PT80, a partir de sinais acústicos e de vibração previamente segmentados e preparados. Para esse fim, a modelagem foi estruturada em duas vertentes complementares.

A primeira vertente contemplou modelos baseados em atributos extraídos dos sinais de áudio e vibração. Nessa abordagem, foram considerados tanto modelos estatísticos tradicionais, utilizados como *baseline*, quanto modelos clássicos de aprendizado de máquina com maior capacidade de modelagem não linear.

A segunda vertente, apresentada posteriormente, consiste em uma abordagem de aprendizado profundo por meio de Redes Neurais Convolucionais (CNNs) aplicadas a representações tempo-frequência dos sinais. Nessa etapa, optou-se por utilizar exclusivamente

o sinal de áudio, em razão do menor custo de instrumentação e do caráter não invasivo desse tipo de aquisição, além do seu potencial para viabilizar soluções de monitoramento mais simples e facilmente escaláveis em ambientes industriais.

3.4.1 Pré-processamento.

A etapa de pré-processamento contempla a preparação dos dados brutos para utilização nas duas vertentes da pesquisa. Ambas compartilham a etapa inicial de segmentação dos sinais, na qual os dados contínuos de áudio e vibração são divididos em janelas temporais, sendo cada janela tratada como uma instância individual de análise.

A partir dessa etapa comum, as vertentes seguem fluxos distintos. Na primeira, procede-se à extração de atributos estatísticos e espectrais dos sinais de áudio e vibração, com posterior consolidação em uma base multissensorial rotulada. Na segunda vertente, cada janela de áudio é convertida em uma imagem de espectrograma Mel, organizada em uma base específica com as respectivas rotulações.

Em ambas as vertentes, adotou-se o mesmo critério de separação entre os conjuntos de treinamento e teste, realizado no nível de ensaio. Dessa forma, garante-se que a avaliação dos modelos ocorra exclusivamente com sinais provenientes de experimentos completamente inéditos, mitigando o risco de vazamento de informações entre etapas. Além disso, a utilização de um conjunto de teste comum possibilita a comparação direta e consistente do desempenho entre as duas abordagens propostas. O conjunto de teste externo foi composto pelos Ensaios 5, 8, 17, 18 e 19, enquanto os demais ensaios foram utilizados nas etapas de treinamento e validação dos modelos.

Os Ensaios 5 e 8 foram incluídos por conterem a condição-alvo de falha (folga na Corrente 1) nas velocidades de 20 Hz e 35 Hz, garantindo a presença de amostras anormais em regimes já contemplados durante a modelagem, porém obtidas em coletas independentes. Em conjunto, os Ensaios 18 e 19 (condição normal) e o Ensaio 17 (27,5 Hz, condição operacional não observada no treinamento) permitem avaliar, simultaneamente, o desempenho em operação normal, a detecção de falha em regimes conhecidos e a capacidade de generalização para uma velocidade intermediária inédita.

3.4.1.1. Segmentação dos sinais

A primeira etapa do pré-processamento consistiu na segmentação dos arquivos de gravação de áudio e vibração em janelas temporais, de modo a converter os sinais contínuos em instâncias individuais de análise. Para isso, foi empregada a técnica de *sliding window*, com

janelas de 5 segundos e sobreposição de 50%. A figura 7, ilustra esse processo.

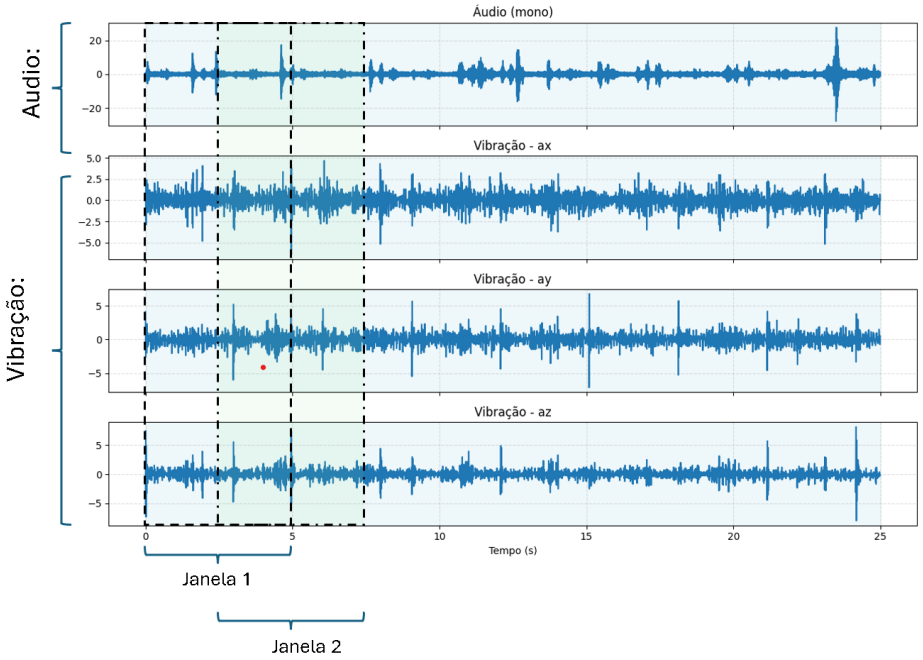

Figura 7 - Processo de Janelamento (*sliding window*) aplicado aos sinais de Áudio e Vibração

A escolha do comprimento da janela foi motivada por critérios físicos e operacionais do processo estudado. O intervalo de 5 segundos abrange mais de um ciclo completo de operação da máquina na menor velocidade de funcionamento, assegurando que cada janela contenha a dinâmica integral do sistema. Dessa forma, evita-se a fragmentação de eventos relevantes, que poderia ocorrer caso janelas excessivamente curtas fossem adotadas.

A utilização de uma sobreposição de 50% teve como objetivo aumentar o número de instâncias disponíveis para treinamento sem perda significativa de informação, além de mitigar efeitos de borda e reduzir a dependência da posição exata de início das janelas, assegurando que eventos transitórios não sejam perdidos.

O procedimento de segmentação resultou na geração de múltiplas janelas temporais por ensaio, ampliando significativamente o número de instâncias disponíveis para a modelagem. Considerando que os ensaios tiveram duração aproximada de 3 minutos, cada experimento produziu, em média, cerca de 75 janelas temporais.

Cada janela segmentada foi tratada como uma instância associada ao respectivo rótulo experimental, definido pela condição da Corrente 1 e pelo arranjo das Correntes 2 e 3 e da Velocidade da máquina (ver Tabela 3). Ao final do processo de segmentação, o conjunto de dados totalizou 1.453 instâncias, derivadas dos 19 ensaios experimentais realizados, das quais

1.077 constituíram a base de treinamento e 376 a base de teste,

3.4.1.2 Modelos baseados em atributos.

Nesta vertente, optou-se pela utilização de modelos de classificação baseados em atributos, nos quais os sinais de áudio e vibração não são utilizados diretamente em sua forma bruta, mas sim representados por meio de descritores estatísticos e espectrais. Esse tipo de abordagem pressupõe a extração prévia de características relevantes dos sinais, as quais são então fornecidas como entrada para modelos estatísticos tradicionais e algoritmos clássicos de aprendizado de máquina supervisionado. Tal estratégia permite maior controle sobre as informações utilizadas na modelagem, além de favorecer a interpretabilidade dos resultados e a comparação entre diferentes paradigmas de classificação.

3.4.1.2.1 Extração de atributos (features)

De cada janela temporal segmentada foram extraídos atributos no domínio do tempo e da frequência, com o objetivo de capturar características estatísticas e espectrais relevantes para a distinção entre as diferentes condições de operação da máquina.

Para os sinais de vibração, analisados nos três eixos ortogonais (x, y e z), foram inicialmente calculadas métricas no domínio do tempo, incluindo valor eficaz (RMS), valor de pico, média, desvio-padrão, *skewness*, *kurtosis* e energia do sinal. Adicionalmente, foi considerada a magnitude resultante do vetor de aceleração, a partir da qual foram extraídos o RMS e o valor de pico. No domínio da frequência, foi extraída a frequência de pico para cada eixo, obtida por meio da Transformada Rápida de Fourier (FFT).

Para os sinais de áudio, previamente convertidos para um único canal (mono), foram extraídas métricas análogas no domínio do tempo, a saber: RMS, média, desvio-padrão, *skewness*, *kurtosis*, energia e taxa de cruzamento por zero (*Zero Crossing Rate* – ZCR), atributo amplamente empregado na caracterização de sinais acústicos. No domínio da frequência, foi igualmente considerada a frequência de pico, obtida por meio da FFT. A tabela 4 mostra a relação das *features* extraídas por tipo de sinal e domínio.

Tabela 4 - *Features* extraídas Sinal x Domínio

Sinal	Domínio	Atributos Extraídos
Áudio	Tempo	RMS, Média, Desvio-padrão, *Skewness*, *Kurtosis*, Energia, ZCR
	Frequência	Frequência de pico (FFT)
Vibração	Tempo	RMS, Pico, Média, Desvio-padrão, *Skewness*, *Kurtosis*, Energia, RMS_res, Pico_res
	Frequência	Frequência de pico (FFT)

Esse conjunto de atributos permitiu representar, de forma compacta e informativa, tanto a dinâmica temporal quanto o conteúdo espectral dos sinais, constituindo a base de entrada para os modelos de classificação baseados em atributos apresentados na sequência.

Por fim, todas as instâncias de áudio e vibração foram pareadas janela a janela e consolidadas em uma única base de dados multissensorial. Essa base foi posteriormente dividida em conjuntos de treinamento e teste no nível de ensaio, conforme descrito na seção 3.4.1.1. Vale observar que a variável de ajuste (Velocidade) foi adicionada ao vetor de características como entradas do modelo.

3.4.1.2.3 Definição e configuração dos modelos

Com a base de dados preparada, a etapa seguinte consistiu na definição e configuração dos algoritmos de aprendizado supervisionado empregados neste trabalho. A seleção dos modelos teve como objetivo contemplar diferentes paradigmas de classificação — baseados em modelos estatísticos, árvores de decisão, margens de separação e redes neurais artificiais — amplamente utilizados em problemas de diagnóstico de falhas e adequados ao tratamento de dados multivariados e ruidosos.

Foram selecionados os seguintes algoritmos, os quais foram discutidos em maior detalhe na Seção 2:

- **Modelos estatísticos tradicionais (*baseline*):**
  - **Regressão Logística (LR)** – modelo estatístico probabilístico baseado em combinações lineares dos atributos, amplamente empregado como referência em problemas de classificação binária;

- **Análise do Discriminante Linear (LDA)** – técnica estatística supervisionada que busca maximizar a separação entre classes por meio de combinações lineares das variáveis, assumindo distribuições gaussianas com covariâncias semelhantes entre as classes.

- **Modelos de aprendizado de máquina:**
  - ***Random Forest* (RF)** – modelo de *Ensemble* baseado em árvores de decisão, reconhecido por sua robustez a ruídos e variações nos dados;
  - ***Support Vector Machine* (SVM)** com *kernel* RBF – classificador baseado em margens de separação, adequado para problemas com fronteiras de decisão não lineares;
  - **Perceptron Multicamadas (MLP)** – rede neural artificial composta por camadas densas, capaz de capturar relações não lineares complexas;
  - ***Ensemble* por *Soft Voting*** – combinação dos modelos RF, SVM e MLP, explorando a complementaridade entre os classificadores com o objetivo de aumentar a robustez e a estabilidade das predições.

A implementação dos modelos foi realizada em *Python* 3.10, utilizando as bibliotecas especializadas *Scikit-Learn* (LR, LDA, RF, SVM, MLP e *VotingClassifier*).

Considerando o caráter exploratório da pesquisa e a necessidade de estabilidade nos resultados, optou-se por não realizar busca exaustiva para otimização de hiperparâmetros. Em vez disso, foram utilizados valores padrão ou moderadamente ajustados, com base em literatura prévia e testes preliminares. A Tabela 5 resume os principais hiperparâmetros adotados.

Tabela 5 - Hiperparâmetros configurados para os modelos de classificação

Modelo	Hiperparâmetros
Regressão Logística (LR)	max_iter=1000, solver='lbfgs', n_jobs=-1
Análise do Discriminante Linear (LDA)	Utilizou parâmetros padrão.
Random Forest (RF)	n_estimators = 200,
Extreme Gradient Boosting (*XGBoost*)	n_estimators = 200, use_label_encoder = False, eval_metric = 'logloss',
Support Vector Machine (SVM)	*kernel* = 'rbf', probability = True,
Perceptron Multicamadas (MLP)	hidden_layer_sizes = (128, 32), activation = 'relu', max_iter = 50000,

Ensemble (*Soft Voting*)	Combinou RF, *XGBoost*, SVM e MLP, voting = 'soft'

Essa configuração buscou garantir equilíbrio entre desempenho, interpretabilidade e custo computacional, permitindo avaliar de forma consistente o impacto da integração multissensorial (áudio + vibração) na qualidade dos diagnósticos.

3.4.1.2.4 Treinamento, validação e teste

Com os modelos de aprendizado de máquina definidos e parametrizados, procedeu-se ao treinamento supervisionado e à avaliação de desempenho seguindo uma estratégia estruturada de particionamento dos dados, composta por validação interna e posterior teste externo. Essa abordagem teve como objetivo garantir uma estimativa robusta da capacidade de generalização dos modelos, reduzindo vieses associados a ajustes excessivos aos dados de treinamento.

O processo de avaliação foi organizado em três subconjuntos distintos:

- **Treinamento** – utilizado para o ajuste dos parâmetros internos dos modelos;
- **Validação** – empregado para o acompanhamento do desempenho durante o treinamento, permitindo a identificação de sobreajuste (*overfitting*);
- **Teste** – reservado exclusivamente para a avaliação final, sendo composto por ensaios não utilizados em nenhuma etapa anterior do processo, assegurando a verificação da capacidade de generalização.

Conforme mencionado na 3.4.1.2., a base multissensorial foi separada em base de treinamento e base de teste, após a extração das *features*.

A base de treinamento, composta por 1.077 instâncias, foi posteriormente subdividida em subconjuntos de treinamento e validação, utilizando uma proporção de 80% para treinamento e 20% para validação. Com o intuito de avaliar a influência de efeitos estocásticos inerentes ao processo de treinamento, tais como a inicialização dos modelos e a composição dos subconjuntos de treinamento e validação, foi adotado um procedimento de avaliação repetida. Para isso, foi implementado um laço de repetição no qual, a cada iteração, uma nova semente aleatória (*seed*) era definida e utilizada tanto no particionamento dos dados (*train_test_split*) quanto nos parâmetros *random_state* dos modelos que o suportam.

Na base de treinamento, após o *split*, composta inicialmente por 862 instâncias (~80% de 1077), observou-se um leve desbalanceamento da variável de saída, referente ao estado da Corrente 1, com 511 instâncias (59%) classificadas como NORMAL e 351 instâncias (41%)

como ANORMAL. Para mitigar esse desbalanceamento, foi aplicada a técnica de *oversampling* por meio do algoritmo *Synthetic Minority Over-sampling Technique* (SMOTE), utilizando a implementação da biblioteca *imblearn*.

Testes adicionais foram conduzidos sem a aplicação de técnicas de balanceamento, tendo sido observado um desempenho médio ligeiramente inferior no conjunto de teste, com diferenças, na acurácia, da ordem de 0,5 a 1 ponto percentual, em grande parte dentro do intervalo de confiança das métricas avaliadas. Ainda assim, optou-se pelo uso do SMOTE visando reduzir viés do classificador em relação à classe minoritária e garantir maior estabilidade no processo de aprendizado, sem impacto adverso significativo na capacidade de generalização.

Adicionalmente, os atributos foram normalizados com o objetivo de eliminar efeitos de escala entre as variáveis, procedimento particularmente relevante para modelos sensíveis à magnitude dos dados, como SVM e MLP. Para esse fim, foi utilizado o *StandardScaler* da biblioteca *Scikit-Learn*. O ajuste do normalizador (*fit*) foi realizado exclusivamente sobre a base de treinamento, após a separação dos subconjuntos de treinamento e validação, sendo posteriormente aplicado de forma consistente aos conjuntos de validação e de teste.

O procedimento completo de treinamento e avaliação foi então repetido 50 vezes, permitindo a obtenção de métricas médias de desempenho, bem como a construção de intervalos de confiança associados aos resultados obtidos, os quais são analisados na seção de resultados.

3.4.1.3 Modelo baseado em aprendizado automático de atributos

Na segunda vertente da modelagem, foi adotada uma abordagem de aprendizado profundo, baseada em Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNNs), aplicadas a representações tempo-frequência dos sinais de áudio. Diferentemente da primeira vertente, na qual os sinais são representados por atributos previamente extraídos, essa abordagem utiliza diretamente representações bidimensionais derivadas do sinal bruto, permitindo que o próprio modelo aprenda automaticamente padrões discriminativos relevantes para o diagnóstico de falhas.

Nesta vertente, optou-se por utilizar exclusivamente o sinal de áudio, em função de três fatores principais:

(i) o menor custo de instrumentação em comparação aos sensores de vibração;

(ii) o caráter não invasivo da aquisição acústica;
(iii) o potencial de escalabilidade dessa solução para aplicações industriais reais, nas quais a simplicidade do sistema de monitoramento é um requisito relevante.

3.4.1.3.1 Representação tempo-frequência dos sinais de áudio

Após a etapa de segmentação temporal dos sinais de áudio, descrita na Seção 3.4.1.1, cada janela resultante foi tratada como uma instância independente e convertida em uma representação tempo-frequência, utilizada como entrada para os modelos da segunda vertente. Nessa abordagem, não foram extraídos atributos estatísticos ou espectrais de forma explícita. Em vez disso, buscou-se preservar a estrutura completa do sinal por meio de uma representação bidimensional, permitindo que o próprio modelo realizasse o aprendizado das características discriminativas relevantes para a tarefa de classificação.

Para cada janela de áudio, foi gerado um espectrograma Mel, obtido a partir da aplicação da Transformada de Fourier de Curto Tempo (STFT), seguida da projeção do espectro de potência em um banco de filtros distribuídos na escala Mel. Essa transformação foi realizada utilizando a função *melspectrogram* da biblioteca *librosa*, amplamente empregada em aplicações de processamento de sinais acústicos e aprendizado de máquina.

Foram utilizados 128 filtros Mel, valor escolhido por representar um compromisso entre resolução espectral e custo computacional. Esse número é comumente adotado na literatura de classificação acústica e mostrou-se suficiente para capturar variações relevantes no conteúdo frequencial do sinal, sem introduzir dimensionalidade excessiva que pudesse dificultar o treinamento da rede neural.

A potência espectral resultante foi convertida para a escala logarítmica (decibéis), com normalização em relação ao valor máximo de cada espectrograma. Esse procedimento reduz a faixa dinâmica dos dados, contribui para maior estabilidade numérica durante o treinamento da rede neural e torna mais evidentes padrões de baixa energia potencialmente associados a condições anormais de operação.

Os espectrogramas obtidos foram então convertidos em imagens bidimensionais por meio de sua renderização gráfica, utilizando funções de visualização da biblioteca *librosa* em conjunto com a biblioteca *Matplotlib*. Para esse fim, a matriz tempo-frequência em escala logarítmica foi representada como um mapa de cores bidimensional e exportada como arquivo de imagem, com remoção de eixos, legendas e margens, de modo que a imagem final contivesse exclusivamente a informação tempo-frequência do sinal.

Inicialmente, as imagens foram geradas com resolução controlada de 224 × 224 pixels, assegurada por meio da definição explícita das dimensões da figura e da resolução (DPI) no momento da renderização. Posteriormente, após análise exploratória preliminar, constatou-se que resoluções superiores a 64 × 64 pixels não resultaram em ganhos significativos de desempenho preditivo, implicando apenas aumento do custo computacional. Assim, as imagens foram redimensionadas para 64 × 64 pixels antes de serem fornecidas como entrada às redes neurais convolucionais. Essa resolução foi adotada por representar um equilíbrio adequado entre o detalhamento espacial da informação tempo-frequência e o custo computacional associado ao treinamento, além de ser compatível com arquiteturas padrão de Redes Neurais Convolucionais.

A padronização do tamanho das imagens assegura consistência na entrada do modelo e evita distorções associadas a variações dimensionais entre amostras. A Figura 8 ilustra, como exemplo, o espectrograma correspondente à janela 18 do ensaio 5.

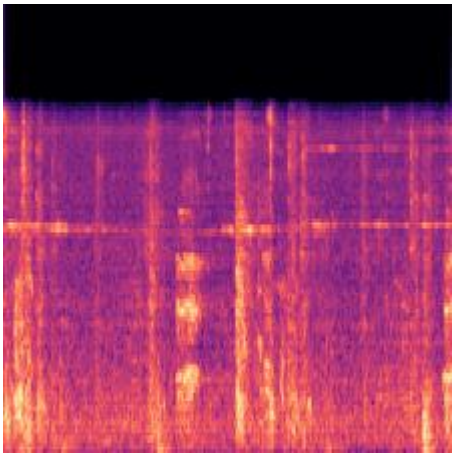

Figura 8 - Exemplo de Espectograma para uma Janela de Áudio

Cada imagem foi armazenada juntamente com seus metadados correspondentes, incluindo o identificador do arquivo de origem, o índice da janela e o instante inicial relativo dentro do ensaio, além das variáveis experimentais associadas à condição de operação da máquina. Essa estrutura possibilitou a organização de uma base de dados específica para a etapa de aprendizado profundo, mantendo o vínculo entre cada espectrograma e o respectivo rótulo experimental.

Ao final do processo, o conjunto de dados baseado em espectrogramas totalizou 1.453 instâncias, correspondendo ao mesmo número de janelas obtidas na etapa de segmentação dos sinais de áudio. Desse total, 1.077 instâncias foram destinadas às etapas de treinamento e validação, enquanto 376 instâncias compuseram o conjunto de teste, assegurando equivalência

quantitativa com a base utilizada na vertente baseada em atributos.

A separação entre os conjuntos de treinamento e teste seguiu o mesmo critério adotado na Vertente 1, sendo realizada no nível de ensaio. Dessa forma, os Ensaios 5, 8, 17, 18 e 19 foram reservados exclusivamente para teste, enquanto os demais ensaios foram utilizados para treinamento e validação do modelo. Essa estratégia garante que a avaliação do desempenho ocorra sobre dados completamente inéditos, evitando o compartilhamento de janelas altamente correlacionadas entre os conjuntos e possibilitando uma comparação direta e consistente entre as duas abordagens investigadas neste trabalho.

Ao adotar essa estratégia, a etapa tradicional de engenharia manual de atributos é substituída por um processo integrado de aprendizado de representações, no qual a extração de padrões relevantes e a classificação são realizadas de forma conjunta pela rede neural. Tal abordagem reduz a dependência de conhecimento prévio sobre o sinal e favorece a generalização do modelo em cenários operacionais distintos.

3.4.1.3.2 Definição e configuração do modelo baseado em Redes Neurais Convolucionais

Grande parte dos estudos que empregam Redes Neurais Convolucionais (CNNs) para a detecção automática de falhas adota arquiteturas relativamente compactas, frequentemente inspiradas ou derivadas da *LeNet-5*, sobretudo em cenários caracterizados por volumes de dados limitados ou restrições computacionais, como evidenciado nos trabalhos de Islam e Kim (2019) e Gültekin et al. (2022). Embora arquiteturas mais profundas e modelos pré-treinados também sejam explorados na literatura, o panorama geral indica que modelos rasos ou moderadamente profundos permanecem amplamente eficazes, sendo que ganhos expressivos de desempenho estão mais associados à qualidade da representação tempo-frequência e à fusão multissensorial do que ao aumento indiscriminado da complexidade da rede.

Diante desse cenário, a arquitetura proposta neste estudo foi definida de forma inspirada em modelos clássicos do tipo *LeNet-like*, amplamente utilizados na literatura para classificação de falhas a partir de representações tempo-frequência, como no trabalho de Islam e Kim (2019). Esses modelos se caracterizam por uma hierarquia de camadas convolucionais e de *pooling*, seguida por camadas totalmente conectadas responsáveis pela decisão final.

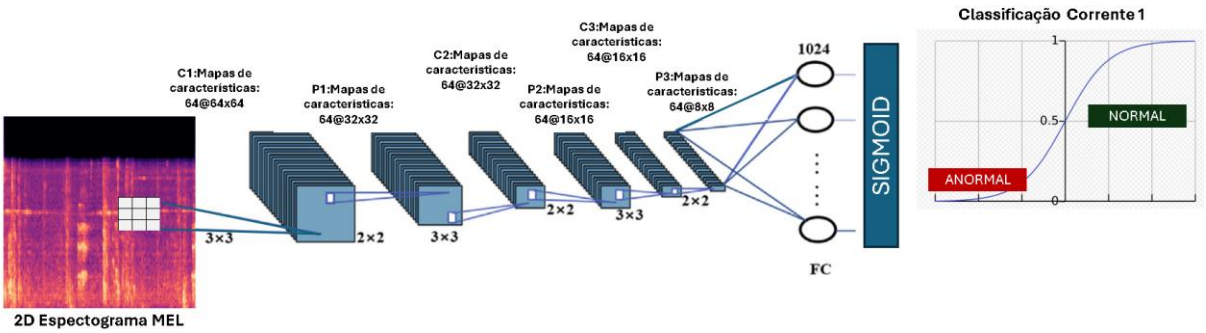

Figura 9 - Arquitetura da CNN do tipo *LeNet-like* adotada neste estudo para classificação binária da condição da Corrente 1 a partir de espectrogramas Mel 2D

Conforme ilustrado na Figura 9, o modelo adotado é composto por três blocos convolucionais sequenciais, cada um formado por uma camada convolucional bidimensional seguida de uma camada de *max pooling*. Todas as camadas convolucionais utilizam 64 filtros de tamanho 3×3, com preenchimento do tipo *same* e função de ativação ReLU. A escolha por manter o número de filtros constante ao longo da profundidade da rede visa favorecer a estabilidade na extração de padrões locais da representação tempo-frequência, ao mesmo tempo em que se preserva uma arquitetura simples e bem estabelecida na literatura.

As operações de *max pooling* com janelas 2×2 são empregadas após cada camada convolucional, promovendo a redução progressiva da resolução espacial das ativações e o aumento do campo receptivo efetivo dos filtros, resultando em mapas de características progressivamente mais abstratos.

Após os blocos convolucionais, as ativações são *flattened* e conectadas a uma camada totalmente conectada com 1024 neurônios e função de ativação ReLU, responsável por integrar as características extraídas nas etapas anteriores. Para mitigar o risco de sobreajuste, é aplicada uma camada de *dropout* com taxa de 0,4.

A camada de saída é composta por um único neurônio com função de ativação sigmoide, adequada a problemas de classificação binária, fornecendo como saída a probabilidade associada à condição da Corrente 1.

A implementação da CNN foi realizada em *Python* 3.10, utilizando a biblioteca *TensorFlow/Keras*. Os experimentos foram conduzidos em um computador equipado com processador *Intel® Core™* i5-1135G7 (11ª geração), 8 GB de memória RAM, executando o sistema operacional *Windows* 11 (64 bits). O treinamento dos modelos foi realizado sem o uso de aceleração por GPU dedicada, utilizando apenas a capacidade computacional da CPU, uma vez que o sistema dispõe apenas de gráficos integrados *Intel Iris Xe*. A Tabela 6 apresenta a

configuração completa da arquitetura da CNN adotada neste estudo.

Tabela 6 - Arquitetura e hiperparâmetros da CNN

Componente	Configuração
Entrada	Imagens RGB 64 × 64 × 3
C1:Camada Conv 1	64 filtros, *kernel* 3×3, ativação ReLU
P1:Max *Pooling* 1	2×2
C2:Camada Conv 2	64 filtros, *kernel* 3×3, ativação ReLU
P2:Max *Pooling* 2	2×2
C3:Camada Conv 3	64 filtros, *kernel* 3×3, ativação ReLU
P3:Max *Pooling* 3	2×2
Flatten	Vetorização dos mapas de características (64 × 8 × 8 = 4096)
Camada Densa	1024 neurônios, ativação ReLU
Dropout	taxa = 0,4
Camada de Saída	1 neurônio, ativação sigmoide
Função de perda	*Binary cross-entropy*
Otimizador	*Adam (learning rate = 0,001)*
Batch size	32
Número máximo de épocas	30

3.4.1.3.3 Treinamento, validação e teste

Com o modelo convolucional definido e configurado, procedeu-se ao treinamento supervisionado e à avaliação de desempenho seguindo uma estratégia estruturada de particionamento dos dados, análoga à adotada na vertente baseada em atributos. Essa estratégia teve como objetivo garantir uma estimativa robusta da capacidade de generalização do modelo, bem como possibilitar uma comparação direta e consistente entre as duas abordagens investigadas.

O processo de avaliação foi organizado em três subconjuntos distintos, de maneira semelhante à vertente 1:

Treinamento – utilizado para o ajuste dos pesos e parâmetros internos da rede neural;

Validação – empregado para o acompanhamento do desempenho ao longo do treinamento, permitindo a identificação de sobreajuste (*overfitting*) e subsidiando mecanismos de controle do processo de otimização;

Teste – reservado exclusivamente para a avaliação final do modelo, sendo composto por ensaios não utilizados em nenhuma etapa anterior do treinamento, assegurando a verificação da capacidade de generalização.

Conforme descrito na Seção 3.4.2.1, a base de dados composta por espectrogramas Mel foi inicialmente separada em conjuntos de treinamento e teste no nível de ensaio, mantendo o mesmo critério adotado na Vertente 1. Dessa forma, os Ensaios 5, 8, 17, 18 e 19 foram reservados exclusivamente para teste, enquanto os demais ensaios compuseram a base utilizada para treinamento e validação do modelo.

A base de treinamento, composta por 1.077 instâncias, foi posteriormente subdividida em subconjuntos de treinamento e validação, utilizando uma proporção de 80% para treinamento e 20% para validação, com estratificação em relação à variável de saída. Esse procedimento assegura a preservação da proporção entre as classes em ambos os subconjuntos, permitindo um acompanhamento mais confiável do desempenho do modelo durante o treinamento.

Diferentemente da vertente baseada em atributos, não foi aplicada a técnica de *oversampling* na base utilizada pela CNN. Essa decisão deve-se ao fato de que técnicas como o SMOTE são definidas no espaço de atributos e não se aplicam diretamente a dados organizados na forma de imagens. Além disso, o uso de validação estratificada e de mecanismos de regularização durante o treinamento contribui para mitigar efeitos adversos do desbalanceamento moderado observado na base.

Como etapa de pré-processamento adicional, as imagens de entrada foram normalizadas por reescala dos valores de intensidade, de modo que os pixels fossem mapeados para o intervalo [0,1]. Esse procedimento é amplamente adotado em modelos de aprendizado profundo e contribui para a estabilidade numérica do processo de otimização, especialmente em redes neurais treinadas por métodos baseados em gradiente.

Durante o treinamento, foram empregados mecanismos de interrupção antecipada (*early stopping*), baseados na evolução da perda no conjunto de validação, bem como a redução adaptativa da taxa de aprendizado, acionada quando não observada melhora no desempenho de validação ao longo de épocas consecutivas. Esses mecanismos permitem controlar o sobreajuste e conduzir o treinamento a soluções mais estáveis.

Seguindo a mesma lógica adotada na Vertente 1, com o objetivo de avaliar a influência de efeitos estocásticos inerentes ao processo de treinamento, tais como a inicialização aleatória dos pesos da rede neural e o particionamento interno entre os subconjuntos de treinamento e validação, o procedimento completo de treinamento e avaliação do modelo convolucional foi repetido ao longo de 50 ciclos independentes, cada um associado a uma semente aleatória distinta.

Em todas as repetições, o treinamento foi realizado por um número máximo de 30

épocas em cada execução, sendo interrompido automaticamente quando os critérios de parada foram atendidos. O desempenho final do modelo foi então avaliado exclusivamente sobre o conjunto de teste externo, composto por ensaios inéditos. As métricas apresentadas na seção de resultados correspondem a valores médios e intervalos de confiança obtidos ao longo das múltiplas execuções realizadas.

3.5 Avaliação

A etapa de avaliação foi planejada com o objetivo de mensurar a eficácia, a robustez e a capacidade de generalização dos modelos de classificação frente às diferentes condições experimentais investigadas neste trabalho. Para isso, foram definidos critérios quantitativos e procedimentos de análise que permitiram comparar o desempenho dos algoritmos de forma consistente e diretamente comparável entre as duas vertentes de modelagem adotadas.

A avaliação foi estruturada em dois níveis complementares, comuns a ambas as vertentes:

Validação interna: análise de desempenho realizada sobre subconjuntos de validação obtidos a partir da base de treinamento, por meio de particionamento interno estratificado, com 80% dos dados destinados ao treinamento e 20% à validação. Esse nível de avaliação permite acompanhar o comportamento dos modelos em dados já observados durante a fase de modelagem, bem como identificar possíveis indícios de sobreajuste.

Teste externo: análise de desempenho realizada exclusivamente sobre ensaios completamente inéditos — Ensaios 5, 8, 17, 18 e 19 — não utilizados em nenhuma etapa de treinamento ou validação. Esse procedimento assegura uma avaliação rigorosa da capacidade de generalização dos modelos em cenários não observados, mais próximos de uma aplicação prática em ambiente industrial.

Seguindo a mesma lógica nas duas vertentes, tanto a validação interna quanto o teste externo foram repetidos ao longo de 50 execuções independentes, associadas a diferentes sementes aleatórias. Dessa forma, os resultados apresentados refletem valores médios de desempenho, acompanhados de intervalos de confiança, reduzindo a influência de efeitos estocásticos inerentes aos processos de treinamento e proporcionando uma avaliação mais robusta e estável dos modelos.

3.5.1 Métricas de avaliação

O desempenho dos modelos foi avaliado a partir das seguintes métricas e ferramentas

de análise:

- **Acurácia**, adotada como indicador principal de desempenho global, por refletir a proporção de classificações corretas em relação ao total de instâncias avaliadas. Essa métrica mostrou-se adequada ao contexto do estudo, considerando o balanceamento das classes realizado na base de treinamento. A acurácia (Acc) é definida por:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(38)}$$

em que TP, TN, FPe FN representam, respectivamente, os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

- **Matriz de confusão**, utilizada para a análise detalhada dos padrões de acerto e erro, permitindo identificar assimetrias na classificação entre as classes NORMAL (OK) e ANORMAL (Folga). Em função das múltiplas repetições realizadas, adotou-se uma análise consolidada das matrizes de confusão, obtidas a partir da agregação dos resultados de todas as execuções independentes. Para essa análise, os valores correspondentes a cada elemento da matriz de confusão (TP, TN, FPe FN) foram somados ao longo das 50 execuções e, posteriormente, normalizados pelo total de amostras classificadas, sendo então expressos em valores percentuais. Essa abordagem possibilita uma interpretação média e estatisticamente mais representativa do comportamento dos classificadores no conjunto de teste externo, além de manter coerência direta com as acurácias médias apresentadas ao longo desta seção, uma vez que a acurácia média associada a cada matriz corresponde à soma dos elementos da sua diagonal principal, isto é, ao traço da matriz normalizada.

3.5.2 Critérios de interpretação dos resultados

A interpretação dos resultados foi conduzida com base nos seguintes critérios:

- **Comparação entre as duas vertentes de modelagem** — baseada em engenharia de atributos e baseada em aprendizado profundo — avaliando ganhos, limitações e compromissos associados a cada abordagem no diagnóstico

de falhas.

- **Análise do impacto da fusão multissensorial**, por meio da comparação entre modelos treinados com sinais de áudio isolados e aqueles que incorporam simultaneamente sinais de áudio e vibração, investigando ganhos de desempenho e estabilidade.
- **Comparação entre desempenho em validação interna e teste externo**, como indicador da robustez dos modelos e de sua capacidade de generalização frente a dados inéditos.
- **Análise qualitativa dos erros sistemáticos**, identificados a partir das matrizes de confusão, buscando compreender padrões recorrentes de classificação incorreta associados a determinadas condições operacionais da máquina.

3.5.3 Intervalos de confiança

Com o objetivo de avaliar a estabilidade estatística dos resultados obtidos e reduzir a influência de variações estocásticas associadas ao processo de treinamento, as métricas de desempenho foram analisadas a partir de intervalos de confiança bilaterais de 98% para a média. Para cada modelo, as métricas foram calculadas ao longo de $n = 50$ execuções independentes e, em seguida, foi estimado o intervalo de confiança da média utilizando a distribuição t de Student, considerando variância populacional desconhecida e $n - 1$ graus de liberdade.

O intervalo de confiança foi calculado conforme**:**

$$IC_{98\%} = \bar{x} \pm t_{0.99,\,49} \cdot \frac{s}{\sqrt{n}} \qquad \textbf{(39)}$$

em que $\bar{x}$representa a média da métrica ao longo das 50 execuções, so desvio-padrão amostral e $t_{0.99,49} \approx 2{,}405$**.** Os intervalos de confiança foram estimados separadamente para os conjuntos de validação interna e de teste externo**,** sendo apresentados de forma gráfica ao longo do capítulo de resultados.

3.6 Implantação

De acordo com o modelo CRISP-DM, a etapa final de um projeto de ciência de dados corresponde à implantação, fase na qual os modelos desenvolvidos são integrados a sistemas produtivos ou incorporados a processos de apoio à decisão. No presente trabalho, entretanto, essa etapa não foi contemplada, uma vez que o escopo da pesquisa esteve restrito ao desenvolvimento, treinamento e avaliação experimental de modelos para diagnóstico de falhas.

Ainda assim, a metodologia proposta apresenta elevado potencial de aplicação prática em ambientes industriais. Os modelos desenvolvidos podem ser incorporados a sistemas de monitoramento preditivo, operando de forma contínua ou quase em tempo real, por meio da utilização de sensores fixos de vibração e microfones direcionados, associados a rotinas automatizadas de aquisição, pré-processamento e classificação dos sinais vibroacústicos.

Essa perspectiva é considerada como trabalho futuro, visando a transposição dos resultados obtidos nesta pesquisa para soluções implementáveis em ambiente industrial, capazes de apoiar estratégias de manutenção preditiva em máquinas de fabricação de copos de papel e em outros equipamentos industriais de características operacionais semelhantes.

4 Resultados

Esta seção apresenta os resultados obtidos na etapa de experimentação com os modelos de aprendizado de máquina para identificação de folga na Corrente 1, conforme descrito na Seção 3. Os experimentos foram organizados em duas vertentes principais: (i) modelos baseados em engenharia de atributos (Vertente 1) e (ii) um modelo baseado em Redes Neurais Convolucionais, com aprendizado automático de representações (Vertente 2).

Para ambas as vertentes, os modelos foram treinados a partir da base de treinamento e avaliados por meio de validação interna estratificada, com particionamento de 80% dos dados para treinamento e 20% para validação, além de teste externo composto por ensaios completamente inéditos (Ensaios 5, 8, 17, 18 e 19). Seguindo a mesma lógica metodológica, todo o processo de treinamento e avaliação foi repetido ao longo de 50 execuções independentes, associadas a diferentes sementes aleatórias, permitindo a obtenção das acurácias médias, acompanhadas de seus respectivos intervalos de confiança.

Adicionalmente, em função das múltiplas repetições realizadas, adotou-se uma análise consolidada das matrizes de confusão, obtidas a partir da agregação dos resultados de todos os

ensaios. Para essa análise, os valores correspondentes a cada elemento da matriz de confusão foram somados ao longo das 50 execuções e, posteriormente, normalizados pelo total de amostras classificadas, sendo então expressos em valores percentuais. Essa representação possibilita uma interpretação média e estatisticamente mais representativa do comportamento dos classificadores no conjunto de teste externo, além de manter coerência direta com as acurácias médias apresentadas ao longo desta seção, de modo que a acurácia média associada a cada matriz corresponde à soma dos elementos da sua diagonal principal (traço).

Os resultados são apresentados a seguir de forma organizada por vertente e por configuração de entrada, permitindo a análise comparativa do desempenho, da robustez e da capacidade de generalização dos modelos em função do tipo de sinal e da abordagem de aprendizado adotada.

4.1 Modelos baseados em atributos

Esta seção apresenta os resultados obtidos com os modelos baseados em engenharia de atributos, correspondentes à Vertente 1 da metodologia proposta. Nessa abordagem, os sinais vibroacústicos são representados por descritores estatísticos e espectrais extraídos nos domínios do tempo e da frequência, utilizados como entrada para algoritmos clássicos de aprendizado de máquina supervisionado.

A análise foi conduzida considerando três configurações distintas de entrada, com o objetivo de investigar de forma sistemática a contribuição individual e combinada de cada tipo de sinal: (i) utilização exclusiva de atributos extraídos dos sinais de áudio, (ii) utilização exclusiva de atributos extraídos dos sinais de vibração e (iii) combinação conjunta de atributos de áudio e vibração, caracterizando uma abordagem multissensorial.

Os resultados são apresentados nas subseções a seguir, organizados de acordo com cada configuração de entrada, permitindo a análise comparativa do desempenho, da robustez e da capacidade de generalização dos modelos em função do tipo de sinal utilizado.

4.1.1 – Áudio

Esta subseção apresenta os resultados obtidos com os modelos de classificação treinados exclusivamente a partir de atributos extraídos dos sinais de áudio, combinados ás diferentes velocidades de operação (Ajuste) consideradas na matriz de experimentos (Tabela 3). O objetivo dessa etapa é avaliar o potencial diagnóstico do sinal acústico de forma isolada, estabelecendo uma linha de base para as análises subsequentes que incorporam sinais de

vibração e abordagens multissensoriais.

A figura 10 apresenta o comparativo das acurácias médias obtidas nos conjuntos de validação e teste externo para os diferentes modelos avaliados, considerando a configuração baseada exclusivamente em atributos acústicos e variável Ajuste.

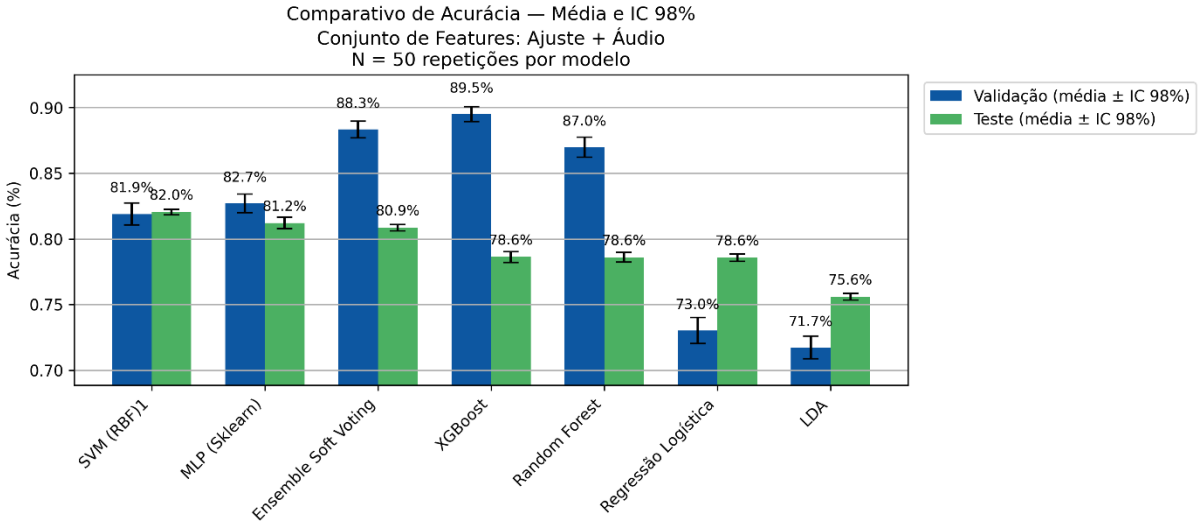

Figura 10 - Comparativo de Acurácia Média entre Modelos utilizando somente Áudio

Entre os modelos estatísticos tradicionais (*baseline*), a Regressão Logística apresentou desempenho inferior no conjunto de validação, mas apresentou acurácia de teste superior à observada na validação. Esse comportamento reforça a interpretação de um modelo de baixa capacidade de ajuste, porém com decisões mais conservadoras e relativamente estáveis frente a dados inéditos, ainda que com desempenho global limitado. A Análise do Discriminante Linear (LDA) obteve os menores valores médios de acurácia, tanto na validação quanto no teste, sugerindo que as hipóteses de separabilidade linear e homogeneidade das covariâncias não são plenamente satisfeitas.

Entre os modelos de aprendizado de máquina, o *XGBoost* apresentou a maior acurácia média no conjunto de validação, aproximando-se de 90%. Entretanto, esse ganho não se refletiu de forma equivalente no conjunto de teste, no qual a acurácia caiu para valores em torno de 79%, caracterizando um comportamento típico de sobreajuste. Tendência semelhante foi observada no *Random Forest*, que apresentou bom desempenho na validação, mas redução perceptível no teste.

O MLP (*Scikit-Learn*) e o SVM com *kernel* RBF destacaram-se pelo equilíbrio entre validação e teste, apresentando diferenças reduzidas entre os dois conjuntos e acurácias de teste próximas ou superiores a 81%. Esse padrão indica maior robustez desses modelos frente à

variabilidade dos dados e sugere melhor capacidade de generalização no cenário que combina informações acústicas e de ajuste operacional. O *Ensemble Soft Voting*, por sua vez, apresentou desempenho intermediário: embora tenha alcançado elevada acurácia na validação, não superou de forma consistente os melhores modelos individuais no teste, mantendo, contudo, comportamento estável.

A análise conjunta desses resultados indica que parte dos modelos é sensível a variações entre os conjuntos de validação e teste, evidenciando que esse conjunto de atributos não captura integralmente a complexidade do comportamento dinâmico da máquina.

Conforme descrito no início da Seção 4, a avaliação do desempenho dos modelos foi complementada por meio da análise das matrizes de confusão consolidadas, obtidas a partir da agregação dos resultados das 50 execuções independentes no conjunto de teste externo.

Essa análise permite observar, de forma média e estatisticamente representativa, o comportamento dos classificadores em relação às taxas de acerto e erro por classe, complementando as métricas globais apresentadas anteriormente. A Figura 11 apresenta as matrizes de confusão percentuais médias para os modelos MLP (*Scikit-Learn*), SVM com *kernel* RBF e *Ensemble Soft Voting*, considerando o conjunto de atributos de Ajuste + Áudio

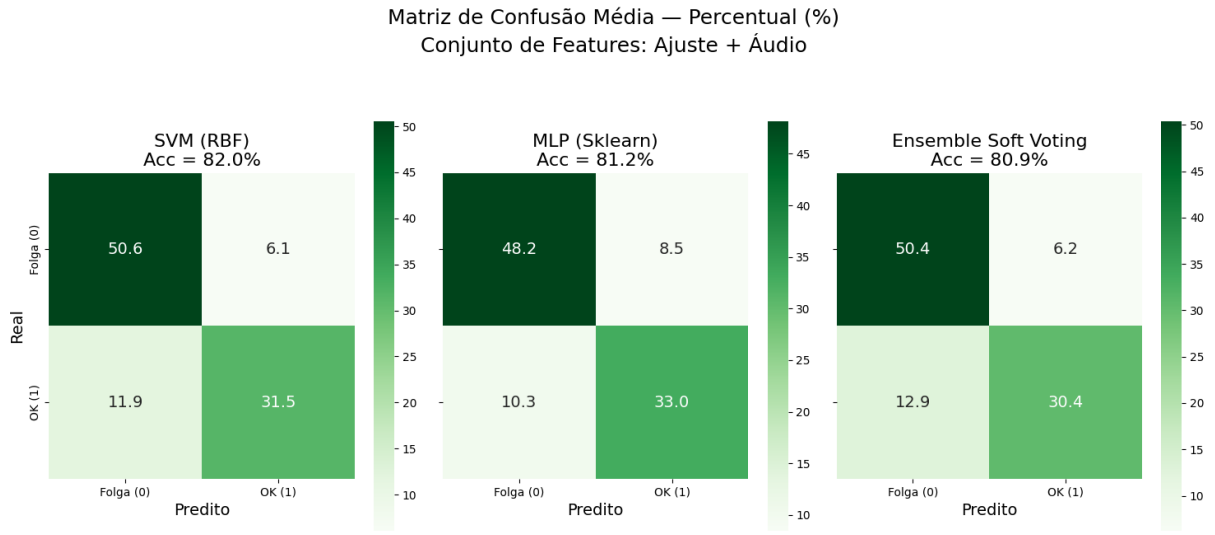

Figura 11 - Matriz de Confusão Média (percentual) referente à base de teste (Top 3 - Áudio)

De modo geral, os três modelos apresentam desempenho consistente na identificação da condição de Folga (0), com elevada concentração de decisões corretas na diagonal principal das matrizes. Observa-se que a proporção das falhas detectadas corretamente situa-se entre aproximadamente 48% e 51% do total de classificações, indicando que os atributos acústicos, combinados às variáveis de ajuste, capturam de forma relativamente robusta padrões associados

à condição de falha. Esse resultado sugere que irregularidades sonoras decorrentes da folga na corrente permanecem bem representadas pelos descritores utilizados.

O *Ensemble Soft Voting* apresentou elevada taxa de acertos para a classe Folga, porém também exibiu maior incidência de classificações incorretas da condição OK (1) como falha, caracterizando um comportamento mais conservador. Tal padrão é típico de modelos que priorizam a detecção de anomalias, o que pode ser desejável em aplicações de manutenção preditiva, embora implique maior taxa de alarmes falsos.

O MLP (*Scikit-Learn*) apresentou comportamento intermediário, com melhor equilíbrio entre a identificação da condição normal e da condição de falha quando comparado ao *Ensemble*. Ainda assim, observa-se uma parcela relevante de confusão entre as classes, evidenciando limitações na separação dos estados operacionais quando se utiliza predominantemente informação acústica.

O SVM com *kernel* RBF destacou-se por apresentar a maior acurácia média global, resultado de uma distribuição mais equilibrada entre os acertos das classes Folga e OK, bem como menores proporções relativas de erro fora da diagonal principal. Esse comportamento reforça a interpretação de um modelo com compromisso mais adequado entre sensibilidade à falha e reconhecimento da condição normal, em consonância com a estabilidade observada nas métricas agregadas.

Em síntese, a análise das matrizes de confusão percentuais confirma que os modelos baseados em Ajuste + Áudio são capazes de identificar padrões associados à folga na corrente de forma consistente, porém apresentam limitações na distinção entre condições normais e anômalas em dados inéditos. Esses resultados corroboram as conclusões obtidas a partir das acurácias médias e reforçam a motivação para a incorporação de sinais de vibração, visando ampliar a robustez e a confiabilidade do diagnóstico, conforme explorado nas subseções seguintes.

4.1.2 – Vibração

Nesta subseção são apresentados os resultados obtidos com os modelos de classificação treinados exclusivamente a partir de atributos extraídos dos sinais de vibração, combinados à variável Ajuste. Diferentemente da análise anterior, baseada apenas em informações acústicas, o foco aqui recai sobre variáveis diretamente associadas à resposta dinâmica da máquina, tais como acelerações, picos, métricas estatísticas e componentes espectrais, amplamente empregadas no diagnóstico de falhas mecânicas.

O objetivo desta etapa é avaliar o potencial diagnóstico dos sinais de vibração de forma isolada,

bem como analisar seu desempenho relativo em comparação à abordagem baseada exclusivamente em áudio.

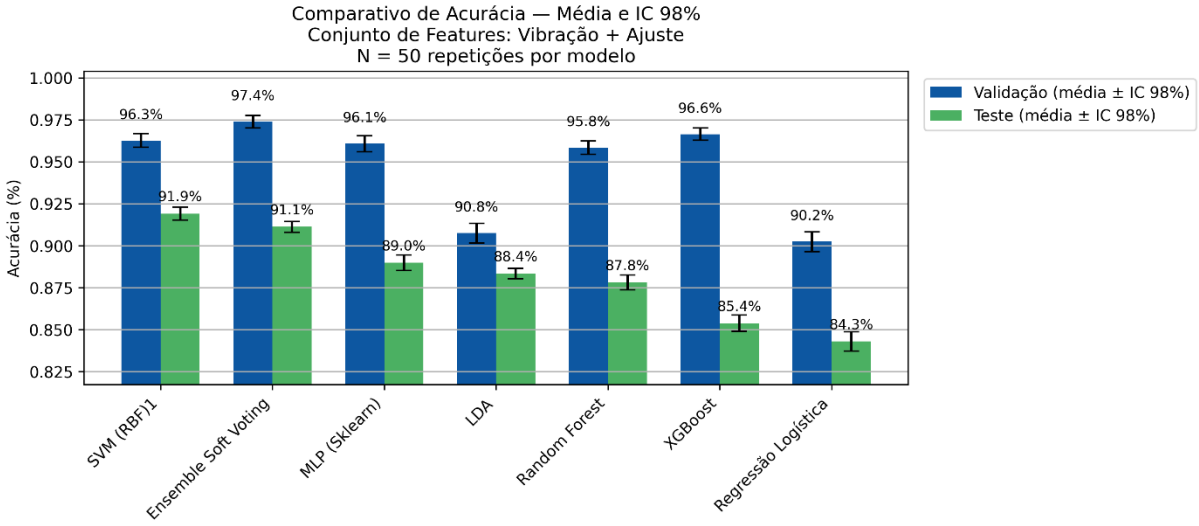

Figura 12 - Comparativo de Acurácia Média entre Modelos utilizando somente Vibração

A Figura 12 apresenta o comparativo das acurácias médias obtidas pelos modelos treinados com atributos de vibração combinados às variáveis de ajuste, considerando os conjuntos de validação e teste, juntamente com seus respectivos intervalos de confiança de 98%. De forma geral, observa-se que o uso dos sinais de vibração resulta em um ganho expressivo de desempenho em relação às abordagens baseadas em áudio, tanto em termos de acurácia absoluta quanto de robustez frente a dados inéditos.

Os modelos baseados em árvores e *Ensemble*s destacaram-se de maneira consistente. O *XGBoost* apresentou uma das maiores acurácias médias no conjunto de validação (≈ 96–97%), mantendo desempenho robusto no conjunto de teste (≈ 85%). Embora se observe um *gap* mais pronunciado entre validação e teste, esse comportamento ocorre em um patamar de desempenho substancialmente superior ao observado na abordagem acústica, indicando alta capacidade de ajuste aliada a boa discriminação dos estados operacionais.

O *Random Forest* apresentou comportamento semelhante, com acurácia próxima de 96% na validação e cerca de 88% no teste, evidenciando uma capacidade de generalização ligeiramente superior à do *XGBoost* nesse cenário. Esse resultado sugere que a estrutura de múltiplas árvores favorece a captura de padrões relevantes nos sinais vibracionais, com menor sensibilidade a variações específicas dos dados.

O *Ensemble Soft Voting* obteve o maior desempenho médio na validação (≈ 97%), com acurácia de teste em torno de 91%, configurando um dos melhores compromissos entre desempenho e estabilidade. Esse resultado indica que a combinação de classificadores se beneficia da complementaridade entre modelos quando aplicada aos atributos de vibração, embora ainda apresente alguma perda de desempenho ao generalizar para dados inéditos.

Entre os modelos individuais, o MLP (*Scikit-Learn*) apresentou resultados consistentes, com acurácia próxima de 96% na validação e cerca de 89% no teste, demonstrando boa capacidade de capturar relações não lineares presentes nos atributos vibracionais. O SVM com *kernel* RBF também se destacou, alcançando acurácia em torno de 96% na validação e aproximadamente 92% no teste, sendo o modelo com melhor desempenho médio no conjunto de teste, o que evidencia elevada robustez e estabilidade.

No que diz respeito aos modelos *baseline*, a Regressão Logística e a Análise do Discriminante Linear (LDA) apresentaram desempenhos inferiores aos modelos mais complexos no conjunto de validação, porém ainda elevados em termos absolutos. Ambos atingiram acurácias superiores a 90% na validação e entre 84% e 88% no teste, indicando que os atributos de vibração apresentam boa separabilidade linear. A menor diferença entre validação e teste nesses modelos sugere que, embora mais simples, eles generalizam de forma consistente, enquanto os ganhos adicionais obtidos pelos modelos mais complexos decorrem da exploração de padrões não lineares e interações entre variáveis. Com relação ao conjunto de teste, vale observar que o LDA apresentou desempenho similar ao *Random Forest* e superior ao *XGBoost*.

De maneira geral, os resultados evidenciam que os sinais de vibração, em conjunto com a variável de ajuste, oferecem informação altamente discriminativa para o diagnóstico de falhas, superando de forma clara o desempenho observado com o uso exclusivo de áudio. Além disso, como será evidenciado na análise das matrizes de confusão (figura 13), a principal contribuição dos atributos de vibração está associada à melhoria do reconhecimento da condição sem falha. Em comparação à abordagem baseada exclusivamente em áudio, os modelos treinados com sinais de vibração apresentaram uma redução expressiva das classificações incorretas da classe OK como falha, indicando menor incidência de alarmes falsos. Esse comportamento sugere que os sinais de vibração promovem uma separação mais nítida entre os estados operacionais, especialmente no que se refere à confirmação da condição normal de funcionamento da máquina.

.

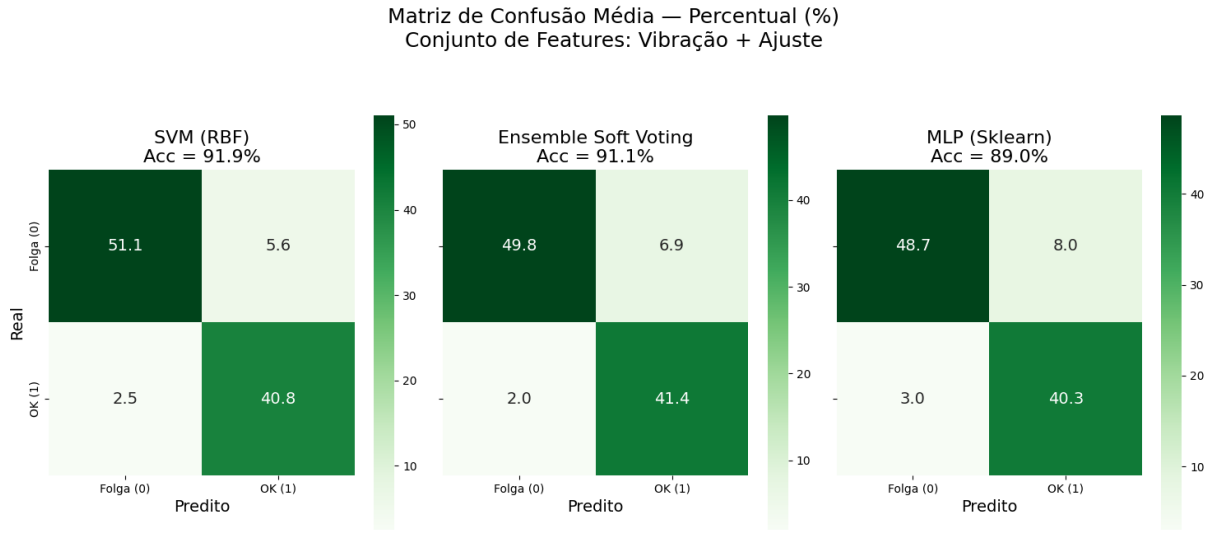

Figura 13 - Matriz de Confusão Média (percentual) referente à base de teste (Top 3 - Vibração)

A Figura 13 apresenta as matrizes de confusão percentuais médias obtidas no conjunto de teste externo para os três modelos com melhor desempenho global na configuração baseada em atributos de vibração combinados às variáveis de ajuste: SVM com *kernel* RBF, *Ensemble Soft Voting* e MLP (*Scikit-Learn*). A análise dessas matrizes permite uma avaliação mais detalhada dos padrões de acerto e erro por classe, complementando as métricas globais de acurácia discutidas anteriormente.

De forma geral, observa-se que os três modelos apresentam elevadas taxas de classificação correta em ambas as classes, com forte concentração dos valores na diagonal principal, evidenciando o alto poder discriminativo dos atributos vibracionais. Em particular, nota-se que a identificação da condição de Folga (0) é realizada de forma consistente, com percentuais de verdadeiros positivos superiores a 48% em todos os casos, enquanto a correta identificação da condição OK (1) também se mantém elevada.

O *Ensemble Soft Voting* apresentou um perfil de erro caracterizado por baixa incidência de falsos negativos, indicando elevada sensibilidade à detecção da condição de falha. Esse comportamento é desejável em aplicações de diagnóstico de falhas, nas quais a não detecção de uma anomalia pode acarretar impactos operacionais relevantes, ainda que implique uma leve elevação na taxa de falsos positivos.

O MLP (*Scikit-Learn*) apresentou comportamento equilibrado entre as classes, com boa capacidade de reconhecimento tanto da condição de falha quanto da condição normal. Observa-se, contudo, uma proporção ligeiramente maior de confusão entre as classes quando comparado

ao *Ensemble*, refletindo limitações residuais na separação dos estados operacionais, mesmo diante do elevado conteúdo informacional dos sinais de vibração.

O SVM com *kernel* RBF destacou-se por apresentar a maior acurácia média global no conjunto de teste, resultado de uma distribuição mais equilibrada entre os acertos das duas classes e de baixas proporções relativas de erro fora da diagonal principal. Observa-se, entretanto, um perfil levemente mais conservador, com maior tendência a classificar a condição normal como falha, priorizando a detecção de estados anômalos.

Em síntese, a análise das matrizes de confusão confirma os resultados observados nas métricas agregadas: os modelos baseados em vibração apresentam alto poder discriminativo e boa robustez no conjunto de teste, com diferenças sutis nos perfis de erro entre os algoritmos. Enquanto o *Ensemble Soft Voting* e o MLP exibem comportamento mais equilibrado, o SVM tende a privilegiar a detecção de falhas. Esses resultados reforçam o potencial dos sinais de vibração como fonte primária de informação para o diagnóstico de falhas e estabelecem uma base sólida para a análise da abordagem multissensorial apresentada na próxima subseção.

A análise conjunta dos resultados obtidos com sinais de áudio e vibração, considerados de forma isolada, permite estabelecer uma comparação direta entre as duas modalidades sensoriais. De modo geral, os modelos treinados com atributos de vibração apresentaram acurácias médias significativamente superiores às observadas na abordagem baseada exclusivamente em áudio, tanto nos conjuntos de validação quanto de teste.

Esse resultado indica que os sinais de vibração carregam informações mais diretamente associadas à dinâmica mecânica da máquina e à manifestação física da folga na corrente, permitindo uma separação mais clara entre as classes. Tal comportamento é consistente com a literatura de diagnóstico de falhas mecânicas, na qual a vibração é amplamente reconhecida como uma das fontes de informação mais ricas para a identificação de defeitos estruturais.

Dessa forma, a comparação entre áudio e vibração indica uma superioridade clara da abordagem baseada em sinais de vibração no contexto experimental avaliado, refletida por acurácias mais elevadas e menores taxas de erro no conjunto de teste. No entanto, a análise isolada das modalidades sensoriais não permite, por si só, concluir sobre a existência de complementaridade entre as informações acústicas e vibracionais. A investigação dessa possível complementaridade é, portanto, objeto da subseção seguinte, na qual se avalia a fusão dos atributos de áudio e vibração com o objetivo de maximizar a robustez e a capacidade de generalização dos modelos de diagnóstico.

4.1.3 – Áudio + Vibração

Nesta subseção são apresentados os resultados obtidos com os modelos de classificação treinados a partir da fusão dos atributos extraídos dos sinais de áudio e vibração, caracterizando uma abordagem multissensorial para o diagnóstico de folga na corrente. Conforme discutido nas subseções anteriores, os sinais de vibração demonstraram poder discriminativo superior ao áudio. A combinação dessas duas fontes de informação, juntamente com a variável Ajuste, busca explorar a complementaridade sensorial, integrando aspectos dinâmicos estruturais e manifestações acústicas do sistema.

O objetivo desta etapa é avaliar se a fusão multissensorial é capaz de elevar simultaneamente o desempenho médio e a robustez dos modelos, reduzindo a discrepância entre validação e teste observada nas abordagens unissensoriais. Para garantir uma comparação direta e consistente, os modelos foram treinados e avaliados seguindo exatamente o mesmo protocolo metodológico adotado nas subseções anteriores, incluindo validação estratificada, teste externo composto por ensaios inéditos e repetição do procedimento ao longo de 50 execuções independentes, permitindo a análise das métricas médias de desempenho e de seus respectivos intervalos de confiança.

A Figura 14 apresenta o comparativo das acurácias médias obtidas nos conjuntos de validação e teste para os diferentes modelos avaliados na configuração multissensorial, servindo como base para a análise do impacto da fusão de áudio e vibração no diagnóstico de falhas.

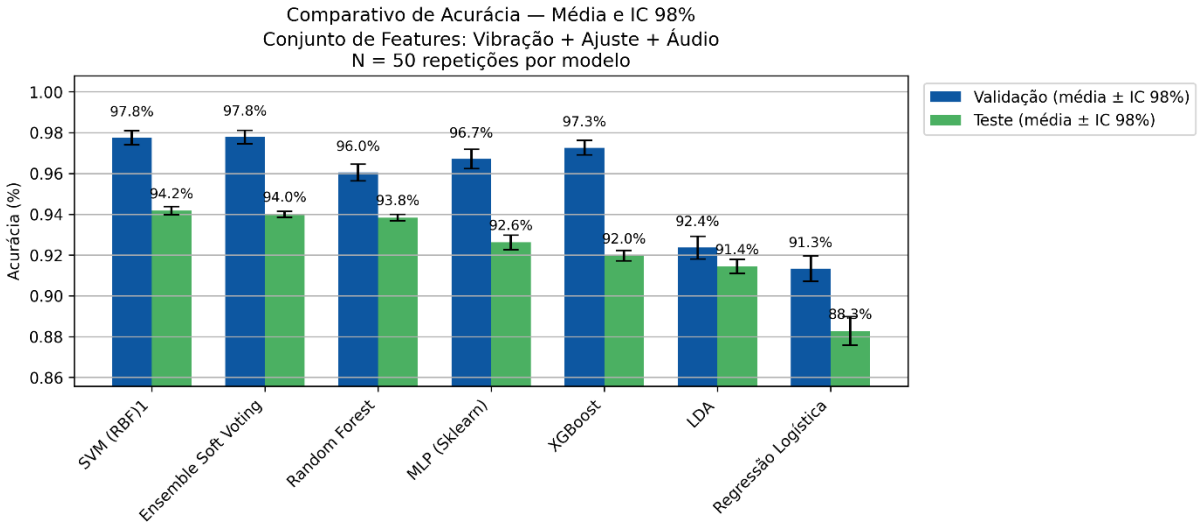

Figura 14 - Comparativo de Acurácia Média entre Modelos utilizando Áudio + Vibração,

De forma geral, observa-se que a abordagem multissensorial promoveu um ganho consistente de desempenho em relação às configurações unissensoriais, tanto em termos de acurácia absoluta quanto de estabilidade entre os conjuntos de validação e teste. A combinação de atributos de vibração, ajuste operacional e áudio resultou nos melhores níveis de desempenho observados em todo o estudo.

Todos os modelos avaliados apresentaram acurácias médias de validação superiores a 91%, com destaque para o SVM com *kernel* RBF e o *Ensemble Soft Voting*, que atingiram valores próximos a 98%. O *XGBoost* e o MLP (*Scikit-Learn*) também apresentaram desempenho elevado na validação, com acurácias acima de 96%, evidenciando a alta capacidade de ajuste dos modelos quando múltiplas fontes sensoriais são exploradas de forma conjunta.

Mais relevante, diferentemente do comportamento observado nas análises baseadas exclusivamente em áudio ou exclusivamente em vibração, a queda de desempenho no conjunto de teste foi significativamente reduzida. No teste externo, o SVM com *kernel* RBF apresentou a maior acurácia média (≈ 94%), seguido de perto pelo *Ensemble Soft Voting* (≈ 94%), *Random Forest* (≈ 94%) e MLP (Sklearn) (≈ 93%). Esses valores superam de forma consistente os resultados obtidos com áudio isolado e se mantêm em patamares ligeiramente superiores aos observados na abordagem baseada apenas em vibração, evidenciando o efeito positivo da fusão sensorial sobre a capacidade de generalização dos modelos.

O *XGBoost*, embora tenha apresentado excelente desempenho na validação, obteve acurácia de teste em torno de 92%, mantendo um pequeno gap entre os conjuntos. Ainda assim, esse comportamento ocorre em um patamar de desempenho elevado e não compromete a superioridade global da abordagem multissensorial.

Os modelos *baseline*, Regressão Logística e Análise do Discriminante Linear (LDA), também se beneficiaram da fusão sensorial, apresentando ganhos claros de desempenho em relação às abordagens unissensoriais. Embora permaneçam abaixo dos modelos de aprendizado de máquina mais complexos, esses resultados indicam que a combinação de atributos acústicos e vibracionais melhora a separabilidade das classes mesmo sob hipóteses lineares, reforçando a complementaridade das informações extraídas.

Em síntese, os resultados demonstram que a abordagem multissensorial baseada na fusão de áudio e vibração proporciona o melhor compromisso entre desempenho, estabilidade e capacidade de generalização dentre todas as configurações avaliadas. A redução consistente do *gap* entre validação e teste, aliada ao estreitamento dos intervalos de confiança, indica que os modelos passam a capturar padrões mais robustos e menos dependentes de condições

específicas de ensaio. Esses achados consolidam a fusão sensorial como a estratégia mais adequada para o diagnóstico de folga na corrente, preparando o terreno para a análise detalhada das matrizes de confusão apresentada na sequência.

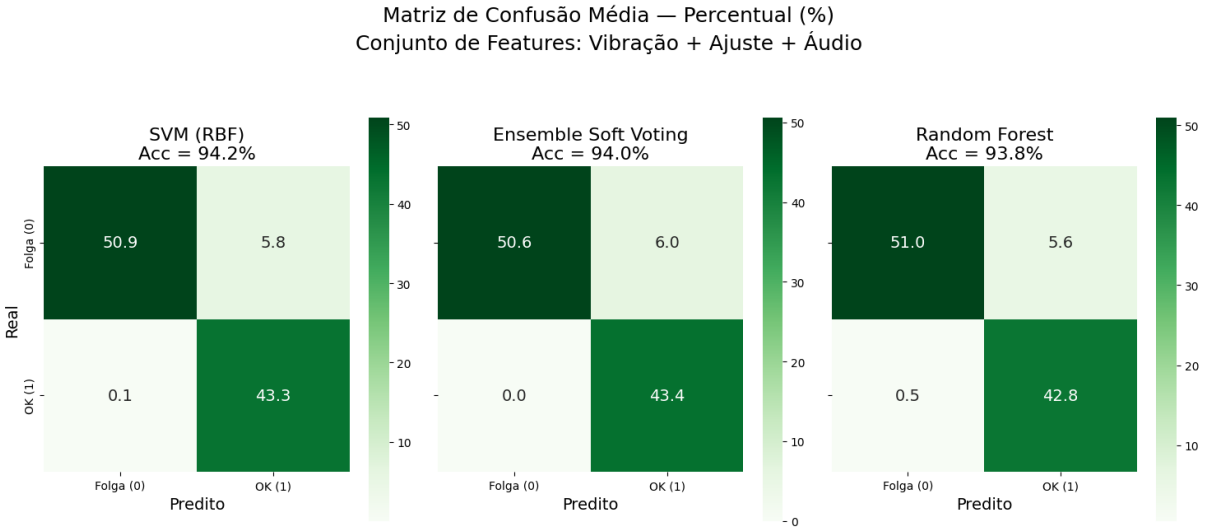

Figura 15 - Matriz de Confusão Média (percentual) referente à base de teste (Top 3 – Áudio + Vibração)

A figura 15 apresenta as matrizes de confusão percentuais médias obtidas no conjunto de teste externo para os três modelos com melhor desempenho global na configuração baseada na fusão dos atributos de vibração, ajuste operacional e áudio: SVM com *kernel* RBF, *Ensemble Soft Voting* e *Random Forest*. A análise dessas matrizes permite uma avaliação detalhada do comportamento dos classificadores em dados inéditos, complementando as métricas agregadas de acurácia apresentadas anteriormente.

De forma geral, observa-se um desempenho altamente consistente dos três modelos, com elevadas taxas de acerto em ambas as classes e forte concentração dos valores na diagonal principal. Destaca-se, em particular, a baixíssima incidência de falsos negativos para a classe OK (1), com valores próximos de zero em todos os modelos, indicando que condições normais raramente são classificadas incorretamente como falha. Esse comportamento é altamente desejável em sistemas de diagnóstico de falhas, nos quais alarmes falsos podem gerar custos operacionais desnecessários.

O *Ensemble Soft Voting* apresentou um perfil de erro especialmente equilibrado, combinando elevada taxa de acertos para a classe Folga (0) com ausência total de falsos negativos para a classe OK (1) no conjunto de teste analisado. Esse resultado reforça a robustez da estratégia de *Ensemble* quando aplicada a uma base multissensorial, evidenciando sua capacidade de integrar informações complementares de forma eficaz.

O *Random Forest* apresentou desempenho semelhante, com altas taxas de acerto em ambas as classes e apenas uma fração residual de erros fora da diagonal principal. Observa-se uma ligeira presença de falsos negativos para a classe OK, porém em níveis extremamente baixos, o que indica boa capacidade de generalização e estabilidade do modelo frente a dados inéditos.

O SVM com *kernel* RBF destacou-se por apresentar o maior equilíbrio entre sensibilidade à falha e reconhecimento da condição normal, refletido na distribuição quase simétrica dos acertos entre as classes. Embora apresente uma proporção mínima de falsos negativos, o modelo mantém elevada capacidade de detecção de padrões associados à folga, comportamento coerente com o perfil mais conservador observado em análises anteriores.

Em síntese, a análise das matrizes de confusão confirma que a abordagem multissensorial resulta em classificadores altamente confiáveis, com desempenho superior ao observado nas configurações unissensoriais. A combinação de atributos acústicos e vibracionais permite reduzir de forma expressiva os erros de classificação, ao mesmo tempo em que preserva um equilíbrio adequado entre detecção de falhas e reconhecimento da condição normal. Esses resultados consolidam a fusão sensorial como a estratégia mais robusta e eficaz para o diagnóstico de folga na corrente, encerrando de forma consistente a análise dos modelos baseados em atributos e preparando o terreno para a comparação com a vertente baseada em modelos convolucionais.

4.2 Rede Neural Convolucional (CNN)

Nesta seção são apresentados os resultados obtidos com os modelos baseados em Redes Neurais Convolucionais (CNNs), que diferem conceitualmente da abordagem anterior ao operar diretamente sobre representações tempo-frequência dos sinais, dispensando a etapa explícita de extração manual de atributos. Enquanto os modelos da Seção 4.1 baseiam-se em *features* estatísticas e espectrais previamente definidas, a vertente convolucional busca aprender automaticamente padrões discriminativos a partir dos dados transformados em espectrogramas Mel.

O objetivo desta etapa é avaliar se a capacidade de aprendizado hierárquico das CNNs permite alcançar desempenho comparável aos melhores modelos baseados em atributos, bem como analisar sua robustez frente a dados inéditos. A avaliação foi conduzida seguindo a mesma estratégia metodológica adotada anteriormente, incluindo separação por ensaios, validação estratificada, teste externo e repetição do processo ao longo de 50 execuções independentes,

assegurando uma comparação direta e consistente entre as duas vertentes.

.

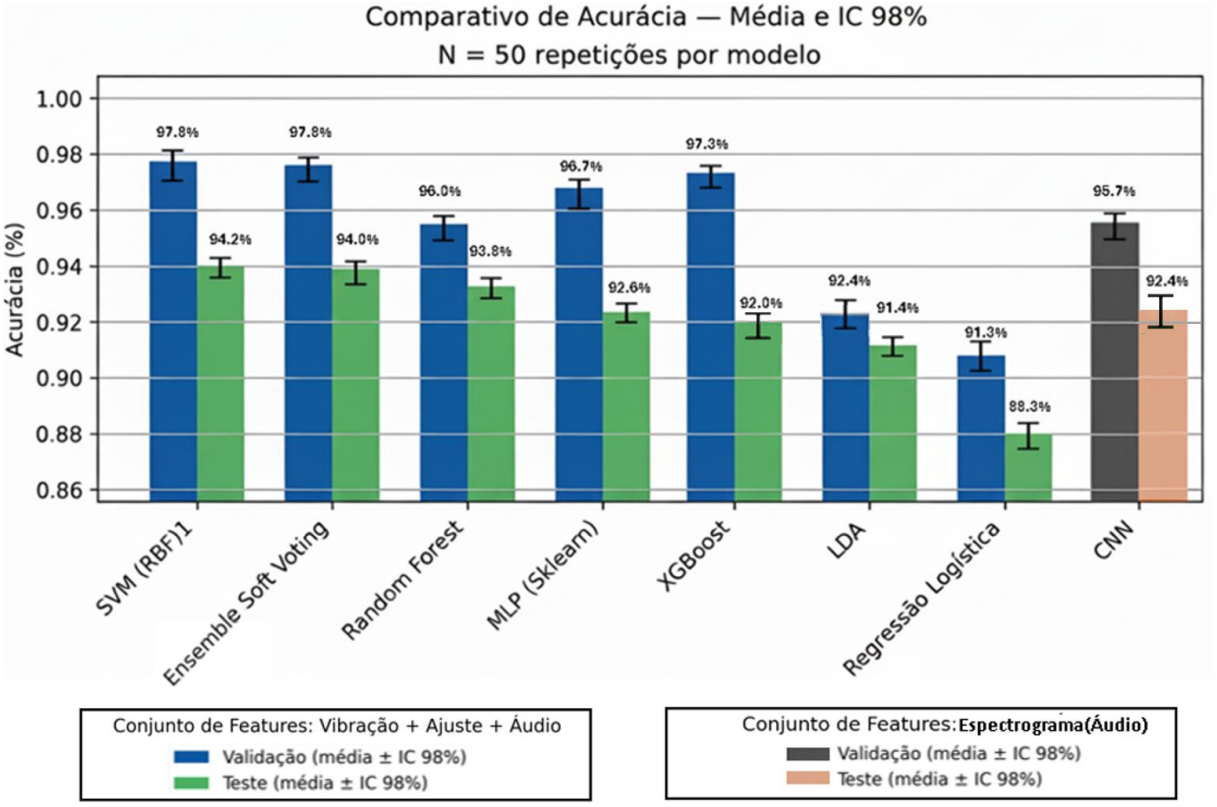

Figura 16 - Comparativo de Acurácia Média entre CNN e Modelos utilizando Áudio + Vibração

A figura 16 apresenta o comparativo das acurácias médias obtidas pelos modelos baseados em atributos na configuração Vibração + Ajuste + Áudio e pelo modelo convolucional, considerando os conjuntos de validação e teste, juntamente com seus respectivos intervalos de confiança de 98%. Essa análise permite avaliar o desempenho da abordagem convolucional frente aos melhores resultados obtidos com engenharia explícita de atributos.

Observa-se que o modelo CNN apresenta acurácia média elevada tanto na validação (≈ 95,7%) quanto no teste (≈ 92,4%), posicionando-se de forma competitiva em relação aos modelos baseados em atributos. Embora alguns classificadores, como SVM e *Ensemble Soft Voting*, apresentem valores ligeiramente superiores no conjunto de teste, a CNN mantém um gap reduzido entre validação e teste, indicando boa capacidade de generalização.

Outro aspecto relevante refere-se aos intervalos de confiança, que se mostram relativamente estreitos para o modelo convolucional, sugerindo estabilidade estatística ao longo

das execuções independentes. Esse comportamento indica que a CNN aprende representações consistentes do sinal, menos dependentes de variações específicas do conjunto de treinamento, mesmo sem recorrer a engenharia manual de atributos.

De forma geral, os resultados indicam que a abordagem convolucional é capaz de alcançar desempenho próximo ao dos melhores modelos multissensoriais baseados em atributos, com a vantagem metodológica de operar diretamente sobre representações tempo-frequência, reduzindo a dependência de conhecimento prévio na definição das *features*.

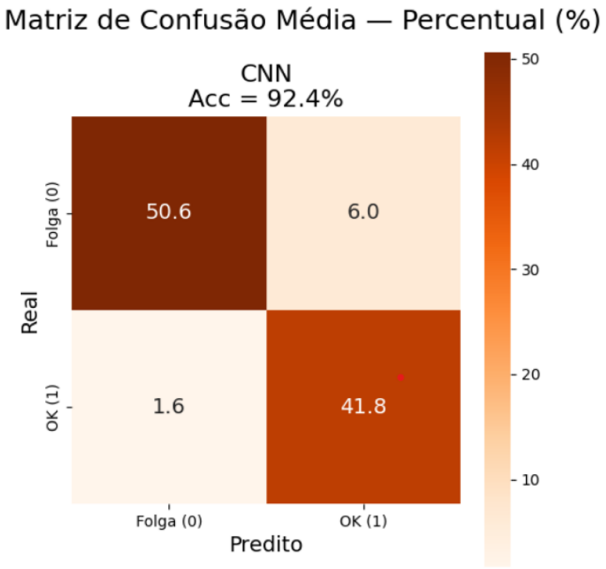

Figura 17 - Matriz de Confusão Média (percentual) na base de teste - CNN

A figura 17 apresenta a matriz de confusão percentual média obtida pelo modelo baseado em Redes Neurais Convolucionais no conjunto de teste externo. A análise dessa matriz permite avaliar de forma detalhada o comportamento do classificador, evidenciando não apenas a taxa global de acerto, mas também a natureza dos erros por classe.

Observa-se que o modelo convolucional apresenta elevada capacidade de identificação da condição OK (1), com aproximadamente 42% de verdadeiros positivos e baixa incidência de falsos positivos, indicando alta confiabilidade na identificação do estado normal de operação da máquina. Esse comportamento é desejável em aplicações práticas, pois reduz a probabilidade de alarmes indevidos.

Em relação à classe Folga (0), a CNN apresenta taxa elevada de acertos, porém com uma fração não desprezível de falsos negativos, correspondentes a casos de folga classificados como condição normal. Esse padrão indica que, embora o modelo capture padrões relevantes

associados à falha, ainda enfrenta dificuldades na identificação de manifestações mais sutis no sinal acústico.

Esse comportamento sugere um viés em favor da classe OK, priorizando a correta identificação da operação normal em detrimento da máxima sensibilidade à falha. Do ponto de vista operacional, esse trade-off reduz alarmes falsos, mas pode resultar na não detecção de determinadas condições anômalas quando estas apresentam baixa assinatura acústica.

Em síntese, a análise da matriz de confusão confirma que o modelo convolucional apresenta boa capacidade de generalização e elevada confiabilidade na identificação da condição normal, mas com sensibilidade limitada para certos padrões de falha quando apenas o sinal de áudio é utilizado. Esse resultado está em consonância com as métricas globais de acurácia e reforça a motivação para a incorporação de informações complementares, como os sinais de vibração, ou para estratégias multissensoriais, capazes de reduzir a taxa de falsos negativos e enriquecer a representação do estado da máquina.

4.3 Síntese dos Resultados

Esta subseção apresenta uma síntese consolidada dos principais resultados obtidos ao longo da Seção 4, possibilitando uma comparação direta entre as diferentes vertentes e configurações de entrada avaliadas. São destacados os melhores desempenhos médios de acurácia no conjunto de teste externo, acompanhados de seus respectivos intervalos de confiança, de modo a fornecer uma visão global e objetiva do comportamento dos modelos analisados.

A Tabela 7 sintetiza os melhores desempenhos obtidos em cada vertente e configuração de entrada no conjunto de teste externo, oferecendo uma visão consolidada dos resultados experimentais discutidos nesta seção. Como se pode observar, os modelos baseados em engenharia de atributos atingem seus melhores resultados quando empregada a fusão multissensorial de áudio e vibração. Vale destacar que, nessa vertente, o SVM com *kernel* RBF obteve os maiores valores médios de acurácia de teste tanto para os sinais utilizados individualmente quanto para a abordagem multissensorial.

Por sua vez, a abordagem baseada em Redes Neurais Convolucionais apresenta desempenho competitivo, mesmo utilizando exclusivamente sinais acústicos, situando-se abaixo apenas da configuração multissensorial da Vertente 1 em termos de acurácia média no conjunto de teste externo. Outro aspecto relevante refere-se à estabilidade estatística dos

resultados, uma vez que, mesmo considerando um intervalo de confiança de 98%, as estimativas apresentaram intervalos relativamente estreitos.

Tabela 7 - Síntese dos melhores desempenhos de acurácia no conjunto de teste externo por vertente e tipo de sinal (IC 98%)

Vertente	Tipo de Sinal	Melhor Modelo	Acurácia no Conjunto de Teste		
			IC 98% (inf.)	Média	IC 98% (sup.)
*Modelos baseados em atributos (**1**)*	Áudio	SVM(RBF)	81,7%	82,0%	82,3%
	Vibração	SVM(RBF)	91,5%	91,9%	92,3%
	Áudio + Vibração	SVM(RBF)	94,0%	94,2%	94,4%
*Modelo baseado em aprendizado automático de atributos (**2**)*	Áudio	CNN	91,7%	92,4%	93,1%

5 Conclusão

O presente trabalho demonstrou que a utilização de sinais acústicos e de vibração possibilita, de forma robusta, o diagnóstico de folga na Corrente 1 em máquinas formadoras de copos de papel. Tanto os modelos baseados em extração manual de atributos (Vertente 1) quanto aqueles baseados em aprendizado automático de representações (Vertente 2) apresentaram desempenhos satisfatórios quando avaliados no conjunto de teste externo, sendo que o melhor modelo alcançou acurácia média de aproximadamente 94%, com intervalo de ±0,2 p.p. para um nível de confiança de 98%.

Esse resultado foi obtido pelo SVM com *kernel* RBF, utilizando a fusão vibroacústica na Vertente 1. No contexto dessa vertente, os experimentos evidenciaram que, embora os sinais considerados de forma isolada apresentem potencial razoável para a classificação do defeito, é a combinação de informações acústicas e vibracionais que proporciona os melhores resultados. Quando utilizados apenas atributos de áudio, o melhor modelo atingiu acurácia média de teste em torno de 82%, enquanto o uso exclusivo de atributos de vibração elevou esse valor para aproximadamente 92%.

De forma consistente, o SVM destacou-se como o classificador com os maiores valores médios de acurácia em todas as configurações avaliadas nessa vertente, tanto nas abordagens unissensoriais quanto na fusão dos sinais, apresentando intervalos estreitos mesmo para um nível de confiança de 98%. Esses resultados indicam que, apesar do ganho expressivo obtido

com a substituição de atributos acústicos por vibracionais, o desempenho máximo é alcançado apenas com a fusão de ambas as modalidades, evidenciando um efeito claro de complementariedade sensorial.

Esses achados estão em consonância com a literatura recente sobre fusão vibroacústica. Trabalhos como os de Praveen Kumar et al. (2019), Inturi et al. (2023) e Pacheco-Chérrez et al., (2022) também reportam ganhos expressivos de desempenho ao combinar sinais acústicos e vibracionais, especialmente quando associados a classificadores baseados em árvores e redes neurais. Dessa forma, os resultados obtidos reforçam a evidência de que abordagens multissensoriais tendem a superar estratégias monossensoriais, oferecendo maior robustez frente a variações operacionais e dados inéditos

Na segunda vertente, baseada em aprendizado profundo e sem a necessidade de extração manual de atributos, observou-se que o desempenho obtido, embora ligeiramente inferior ao alcançado pela abordagem multissensorial da Vertente 1, mostrou-se adequado para aplicações práticas. A CNN utilizada, operando exclusivamente sobre sinais acústicos convertidos em espectrogramas Mel, atingiu acurácia média de teste em torno de 92%, com intervalo de ±0,7 p.p. para 98% de confiança, ainda considerado estreito. Esse resultado é particularmente relevante, pois, apesar da maior complexidade do modelo, sua implementação prática é mais simples do que a da Vertente 1, uma vez que dispensa a etapa de engenharia de atributos e utiliza apenas sinais acústicos, cuja instrumentação apresenta menor custo, maior simplicidade de instalação e caráter não invasivo.

Como limitações do estudo, destaca-se que os experimentos foram conduzidos em um único modelo de máquina (PT80), sob condições controladas de simulação de falhas e com foco restrito à folga na Corrente 1 do sistema de transmissão. Embora esse componente seja crítico para o sincronismo da máquina, outros subsistemas relevantes, como o regravador e a faca de corte, não foram abordados experimentalmente. Adicionalmente, a taxa de amostragem de 100 Hz utilizada para os sinais de vibração mostrou-se adequada para a análise de fenômenos de baixa frequência associados a folgas em correntes, conforme discutido por Tsutada et al. (2007), mas não seria suficiente para a investigação de falhas de alta frequência, como defeitos incipientes em rolamentos, que demandam sensores com capacidade de aquisição na faixa de quilohertz (Tandon; Choudhury, 1999).

De modo geral, sob o ponto de vista prático, os resultados indicam a viabilidade de implementação de soluções de monitoramento preditivo em equipamentos industriais por meio de técnicas consolidadas de aprendizado de máquina utilizando sinais de áudio e vibração, em ambas as vertentes avaliadas. A Vertente 1 apresenta maior robustez e desempenho, ao custo

de maior complexidade de implementação associada à fusão sensorial e à engenharia de atributos. Já a Vertente 2 configura-se como uma alternativa mais simples e de menor custo operacional, embora com desempenho inferior ao da abordagem multissensorial baseada em atributos. Ainda assim, ambas representam estratégias acessíveis e escaláveis para indústrias que buscam aumentar a confiabilidade operacional, reduzir custos de manutenção e mitigar paradas não programadas. No contexto específico desta dissertação, as duas vertentes mostraram-se eficazes para a detecção da condição de folga na Corrente 1 do sistema de transmissão da máquina formadora de copos de papel, permitindo classificar de forma robusta estados de operação normal e anormal. Embora o foco experimental tenha sido delimitado à Corrente 1, a metodologia proposta apresenta potencial de extensão para a análise das Correntes 2 e 3 e de outros subsistemas críticos da máquina, desde que observadas as devidas adaptações de instrumentação e coleta de dados.

5.1 Etapas Futuras

Com base nos resultados alcançados até o momento, algumas etapas adicionais estão previstas para fortalecer e consolidar a pesquisa, bem como para posicionar este trabalho como base para investigações futuras em nível de doutorado.

5.1.1 Ampliação do escopo.

Como continuidade natural deste estudo, prevê-se a ampliação do escopo experimental, tanto no que diz respeito à investigação de outros componentes críticos da máquina formadora de copos de papel — tais como o regravador, a faca de corte e as demais correntes do sistema de transmissão — quanto à aplicação da metodologia em outros modelos de máquinas além da PT80. Adicionalmente, pretende-se avaliar o uso de sensores de vibração com maior taxa de amostragem, possibilitando a detecção de outros tipos de defeitos, como falhas em rolamentos.

Essa ampliação permitirá avaliar se a abordagem multissensorial proposta mantém desempenho robusto em defeitos de natureza distinta da folga na Corrente 1, bem como analisar a capacidade dos modelos em lidar com múltiplos modos de falha coexistentes. Tal investigação contribui para o aumento da capacidade de generalização do diagnóstico e para a aproximação do método a cenários industriais mais complexos e realistas.

5.1.2 Aprimoramento da vertente de aprendizado profundo (CNN)

No âmbito deste trabalho, foi implementada e avaliada uma vertente de aprendizado

profundo baseada em Redes Neurais Convolucionais aplicadas a espectrogramas Mel dos sinais acústicos, a qual demonstrou bom equilíbrio entre desempenho e capacidade de generalização quando comparada a modelos baseados exclusivamente em engenharia manual de atributos.

Como etapa futura, pretende-se aprofundar a análise dessa vertente, investigando de forma mais sistemática o impacto de diferentes configurações arquiteturais, parâmetros de geração das representações tempo-frequência e estratégias de regularização. Essa análise visa consolidar o entendimento dos *trade-offs* entre desempenho preditivo, custo computacional e simplicidade de instrumentação, especialmente em contextos industriais nos quais soluções não invasivas e escaláveis são desejáveis. Além disso, com a adoção de sensores de vibração com maior taxa de amostragem, pretende-se incorporar esses sinais à abordagem convolucional, viabilizando a exploração de arquiteturas verdadeiramente multissensoriais baseadas em aprendizado profundo.

5.1.3 Perspectiva de continuidade em nível de doutorado

Os resultados obtidos neste trabalho indicam que diferentes abordagens — modelos baseados em engenharia de atributos, métodos de *Ensemble* e redes neurais profundas — apresentam desempenhos complementares, capturando padrões distintos associados às condições operacionais da máquina. Essa constatação sugere que o diagnóstico de falhas pode se beneficiar de estratégias que explorem não apenas modelos isolados, mas também estruturas de decisão mais organizadas e adaptativas.

Nesse contexto, uma perspectiva natural de continuidade em nível de doutorado consiste na investigação de estratégias de aprendizado progressivo e multissensorial, nas quais o processo de diagnóstico seja conduzido de forma hierárquica, guiado por diferentes níveis de confiança nas previsões. Tal abordagem permite o descascamento progressivo do espaço de dados, direcionando amostras mais complexas ou incertas para modelos especializados, enquanto padrões mais simples são tratados por classificadores de menor complexidade.

A incorporação de mecanismos de aprendizado residual e a formação incremental de submodelos especializados ao longo do espaço de dados configuram caminhos promissores para melhorar simultaneamente a robustez, a interpretabilidade e a capacidade de generalização dos sistemas de diagnóstico. Além disso, a integração explícita de múltiplas modalidades sensoriais nesse processo progressivo pode ampliar a confiabilidade das decisões, especialmente em cenários industriais caracterizados por variabilidade operacional e presença de ruído.

Dessa forma, o presente trabalho estabelece uma base metodológica e experimental

sólida para pesquisas futuras voltadas ao desenvolvimento de sistemas de diagnóstico inteligentes mais adaptativos e estruturados, alinhados aos princípios da manutenção preditiva avançada e aos desafios atuais da Indústria 4.0, ao mesmo tempo em que fornece subsídios conceituais consistentes para investigações de maior profundidade em nível de doutorado.

7 Referências

ADIBI, Azin; TRINH, Binh Minh; MEKONNEN, Tizazu H. Recent progress in sustainable barrier paper coating for food packaging applications. **Progress in Organic Coatings**, v. 181, 2023.

AL MAMUN, A. *et al.* Multi-channel sensor fusion for real-time bearing fault diagnosis by frequency-domain multilinear principal component analysis. **International Journal of Advanced Manufacturing Technology**, v. 124, n. 3–4, p. 1321–1334, 2023.

ALTAF, M. *et al.* A New Statistical Features Based Approach for Bearing Fault Diagnosis Using Vibration Signals. **Sensors**, v. 22, n. 5, 2022.

BREIMAN, Leo. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, Leo *et al.* **Classification and regression trees**. *[S.l.]*: CRC Press, 2017.

CAKIR, M.; GUVENC, M. A.; MISTIKOGLU, S. The experimental application of popular machine learning algorithms on predictive maintenance and the design of IIoT based condition monitoring system. **Computers and Industrial Engineering**, v. 151, 2021.

CHACÓN, J. L. F. *et al.* A novel machine learning-based methodology for tool wear prediction using acoustic emission signals. **Sensors**, v. 21, n. 17, 2021.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R., . **CRISP-DM 1.0: Step-by-step data mining guide**. *[S.l.: S.n.]*. Disponível em: <https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman/54bad20bbc7938991bf34f86dde0babfbd2d5a72>. Acesso em: 5 set. 2025.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A scalable tree boosting system. *In*: PROCEEDINGS OF THE ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. **Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.** Association for Computing Machinery, 2016.

CHEN, Y.; JIN, Y.; JIRI, G. Predicting tool wear with multi-sensor data using deep belief networks. **International Journal of Advanced Manufacturing Technology**, v. 99, n. 5–8, p. 1917–1926, 2018.

CORTES, Corinna; VAPNIK, Vladimir. Support-Vector Networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995.

DI MAGGIO, L. G. Intelligent Fault Diagnosis of Industrial Bearings Using Transfer Learning and CNNs Pre-Trained for Audio Classification. **Sensors**, v. 23, n. 1, 2023.

DIETTERICH, Thomas G. Ensemble methods in machine learning. *In*: LECTURE NOTES IN COMPUTER SCIENCE (INCLUDING SUBSERIES LECTURE NOTES IN ARTIFICIAL INTELLIGENCE AND LECTURE NOTES IN BIOINFORMATICS). **Lect. Notes Comput. Sci.** Springer Verlag, 2000.

GÉRON, Aurélien. **Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems**. Third edition ed. Beijing: O'Reilly Media, Inc, 2023.

GÜLTEKIN, Ö. *et al.* Multisensory data fusion-based deep learning approach for fault diagnosis of an industrial autonomous transfer vehicle. **Expert Systems with Applications**, v. 200, 2022.

HANSEN, Lars Kai; SALAMON, Peter. Neural Network Ensembles. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 12, n. 10, p. 993–1001, 1990.

HASAN, M. J.; ISLAM, M. M. M.; KIM, J. M. Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions. **Measurement: Journal of the International Measurement Confederation**, v. 138, p. 620–631, 2019.

HAYKIN, Simon S. **Neural networks and learning machines**. 3. ed ed. New York Munich: Prentice-Hall, 2009.

HINTON, Geoffrey E.; OSINDERO, Simon; TEH, Yee-Whye. A fast learning algorithm for deep belief nets. **Neural Computation**, v. 18, n. 7, p. 1527–1554, 2006.

INTURI, V. *et al.* An integrated condition monitoring scheme for health state identification of a multi-stage gearbox through Hurst exponent estimates. **Structural Health Monitoring**, v. 22, n. 1, p. 730–745, 2023.

ISLAM, M. M. M.; KIM, J. M. Automated bearing fault diagnosis scheme using 2D representation of wavelet packet transform and deep convolutional neural network. **Computers in Industry**, v. 106, p. 142–153, 2019.

JANSSENS, Olivier *et al.* Convolutional Neural Network Based Fault Detection for Rotating Machinery. **Journal of Sound and Vibration**, v. 377, p. 331–345, set. 2016a.

JANSSENS, Olivier *et al.* Convolutional Neural Network Based Fault Detection for Rotating Machinery. **Journal of Sound and Vibration**, v. 377, p. 331–345, set. 2016b.

LECUN, Y. *et al.* Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, nov. 1998.

LI, C. *et al.* Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals. **Mechanical Systems and Signal Processing**, v. 76–77, p. 283–293, 2016.

LI, X. *et al.* Multi-sensor fusion fault diagnosis method of wind turbine bearing based on adaptive convergent viewable neural networks. **Reliability Engineering and System Safety**, v. 245, 2024.

LIU, H.; LI, L.; MA, J. Rolling Bearing Fault Diagnosis Based on STFT-Deep Learning and Sound Signals. **Shock and Vibration**, v. 2016, 2016.

LIU, X. *et al.* Acoustic signal based fault detection on belt conveyor idlers using machine learning. **Advanced Powder Technology**, v. 31, n. 7, p. 2689–2698, 2020.

MARKS, Robert J. **Introduction to Shannon Sampling and Interpolation Theory**. New York, NY: Springer, 1991.

MOHAMMED, Ammar; KORA, Rania. A comprehensive review on ensemble deep learning: Opportunities and challenges. **Journal of King Saud University - Computer and**

Information Sciences, v. 35, n. 2, p. 757–774, 2023.

NATESHA, B. V.; GUDDETI, R. M. R. Fog-Based Intelligent Machine Malfunction Monitoring System for Industry 4.0. **IEEE Transactions on Industrial Informatics**, v. 17, n. 12, p. 7923–7932, 2021.

PACHECO-CHÉRREZ, J. *et al.* Bearing fault detection with vibration and acoustic signals: Comparison among different machine leaning classification methods. **Engineering Failure Analysis**, v. 139, 2022.

PHAM, M. T.; KIM, J. M.; KIM, C. H. Deep learning-based bearing fault diagnosis method for embedded systems. **Sensors (Switzerland)**, v. 20, n. 23, p. 1–15, 2020.

POÓR, P.; BASL, J.; ZENISEK, D. Predictive Maintenance 4.0 as next evolution step in industrial maintenance development. *In*: 2019 INTERNATIONAL RESEARCH CONFERENCE ON SMART COMPUTING AND SYSTEMS ENGINEERING (SCSE). **2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)**. mar. 2019. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8842659?casa_token=o7nGFmkFJm0AAAAA:oD5vA_7bF9a6dA8T_3J2EtTswu3CGF6FtWOEAwjVhAmRPQH85ZRdszq-pYdXRmgiReT7mrPIdbYY>. Acesso em: 5 nov. 2024

PRAVEEN KUMAR, T. *et al.* A multi-sensor information fusion for fault diagnosis of a gearbox utilizing discrete wavelet features. **Measurement Science and Technology**, v. 30, n. 8, 2019.

RAMTEKE, S. M.; CHELLADURAI, H.; AMARNATH, M. Diagnosis and Classification of Diesel Engine Components Faults Using Time–Frequency and Machine Learning Approach. **Journal of Vibration Engineering and Technologies**, v. 10, n. 1, p. 175–192, 2022.

RAOUF, I.; LEE, H.; KIM, H. S. Mechanical fault detection based on machine learning for robotic RV reducer using electrical current signature analysis: A data-driven approach. **Journal of Computational Design and Engineering**, v. 9, n. 2, p. 417–433, 2022.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. **Nature**, v. 323, n. 6088, p. 533–536, 1986.

SIDDIQUE, M. F. *et al.* A Hybrid Deep Learning Approach: Integrating Short-Time Fourier Transform and Continuous Wavelet Transform for Improved Pipeline Leak Detection. **Sensors**, v. 23, n. 19, 2023.

STEVENS, S. S.; VOLKMANN, J.; NEWMAN, E. B. A Scale for the Measurement of the Psychological Magnitude Pitch. **The Journal of the Acoustical Society of America**, v. 8, n. 3, p. 185–190, 1 jan. 1937.

TAGAWA, Yuki; MASKELIŪNAS, Rytis; DAMAŠEVIČIUS, Robertas. Acoustic anomaly detection of mechanical failures in noisy real-life factory environments. **Electronics (Switzerland)**, v. 10, n. 19, 2021.

TANDON, N.; CHOUDHURY, A. A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings. **Tribology International**, v. 32, n. 8, p. 469–480, ago. 1999.

TRAN, T.; LUNDGREN, J. Drill fault diagnosis based on the scalogram and MEL spectrogram of sound signals using artificial intelligence. **IEEE Access**, v. 8, p. 203655–203666, 2020.

TSUTADA, HIROYUKI; HIRAI, TAKASHI; ITOH, YUTAKA; SHIGA, SATOSHI. Chain fault detection of escalator using handrail vibration. *In*: 14TH INTERNATIONAL CONGRESS ON SOUND AND VIBRATION (ICSV 2007). **Anais**... 2007.

WANG, X.; MAO, D.; LI, X. Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network. **Measurement: Journal of the International Measurement Confederation**, v. 173, 2021.

XU, G. *et al.* High-speed train wheel set bearing fault diagnosis and prognostics: A new prognostic model based on extendable useful life. **Mechanical Systems and Signal Processing**, v. 146, 2021.

YAO, Jiachi *et al.* Fault detection of complex planetary gearbox using acoustic signals. **Measurement**, v. 178, p. 109428, jun. 2021.

ZONTA, Tiago *et al.* Predictive maintenance in the Industry 4.0: A systematic literature review. **Computers & Industrial Engineering**, v. 150, p. 106889, 1 dez. 2020.