CLASSIFICAÇÃO E DETECÇÃO DE ANOMALIAS EM SINAIS DE VIBRAÇÃO: UMA ABORDAGEM HÍBRIDA DE MACHINE LEARNING

CLASSIFICATION AND ANOMALY DETECTION IN VIBRATION SIGNALS: A HYBRID MACHINE LEARNING APPROACH

Alan Diek Guimarães[1],
Juan Viana Lorenzo[2],
Marcos Vinicius Oliveira dos Santos[3],
Vinicius de Jesus Silva[4],
Vinicius Schilieve Santos[5],

RESUMO

A manutenção preditiva, apoiada por Inteligência Artificial (IA), é um vetor estratégico da Indústria 4.0/IIot, permitindo a antecipação de falhas e a redução de custos operacionais. Este trabalho insere-se no contexto da Indústria 4.0 e IIoT, aplicando técnicas de aprendizado de máquina para detecção de falhas em motores elétricos a partir de sinais de vibração. Utilizou-se um dataset laboratorial fornecido pela Tractian, contém 4.416 amostras de 1 motor, distribuídas em quatro condições: HEALTHY (37,3%), INNER_RACEWAY (10,9%), OUTER_RACEWAY (10,3%) e STRUCTURAL_LOOSENESS (41,5%), com waveforms amostrados a 16 kHz e 32 kHz em três canais de aceleração. Foram extraídas 14 features finais no domínio do tempo e frequência (RMS, Crest Factor, frequência de pico, energia espectral, centroide, MFCC). Resultados: RF 99,3% ±0,3%, SVM 98,7% ±0,5%, Autoencoder AUC ROC 0,956 ±0,01; métricas adicionais (F1, precision, recall, PR-AUC) e matrizes de confusão foram avaliadas. Limitações: único ativo, ausência de séries temporais, e risco de confusão por variáveis operacionais (RPM/carga). O pipeline permite cálculo de score de saúde em regime estacionário e suporte a monitoramento near real-time via dashboards interativos. Trabalhos futuros devem expandir para múltiplos ativos, incorporar marcação temporal e desenvolver modelos de prognóstico (LSTM, TCN, PHM) para RUL.

Palavras-chave: Manutenção preditiva; Machine Learning; Vibração; Internet das Coisas (IoT); Visual analytics; Detecção de falhas

ABSTRACT

Predictive maintenance supported by Artificial Intelligence (AI) is a strategic driver of Industry 4.0/IIoT, enabling early fault detection and operational cost reduction. This

[1] Graduando em Ciência de Dados pela Faculdade Senai São Paulo Campus Paulo Antônio Skaf. E-mail: alandiekguimaraes@gmail.com
[2] Graduando em Ciência de Dados pela Faculdade Senai São Paulo Campus Paulo Antônio Skaf. E-mail: juanviana2577@gmail.com
[3] Graduando em Ciência de Dados pela Faculdade Senai São Paulo Campus Paulo Antônio Skaf. E-mail: mvinicius.oliveira04@gmail.com
[4] Graduando em Ciência de Dados pela Faculdade Senai São Paulo Campus Paulo Antônio Skaf. E-mail: viniciusdejesussilva12@gmail.com
[5] Graduando em Ciência de Dados pela Faculdade Senai São Paulo Campus Paulo Antônio Skaf. E-mail: viniciusschilive@gmail.com

study applies machine learning techniques for fault detection in electric motors using vibration signals. The laboratory dataset provided by Tractian comprises 4,416 samples from 1 motor, across four conditions: HEALTHY (37.3%), INNER_RACEWAY (10.9%), OUTER_RACEWAY (10.3%), and STRUCTURAL_LOOSENESS (41.5%), with waveforms sampled at 16 kHz and 32 kHz on three acceleration channels. 14 final features were extracted from the time and frequency domains (RMS, Crest Factor, peak frequency, spectral energy, centroid, MFCC). Results: Random Forest 99.3% ±0.3%, SVM 98.7% ±0.5%, Autoencoder ROC AUC 0.956 ±0.01; additional metrics (F1, precision, recall, PR-AUC) and confusion matrices were evaluated. Limitations include a single asset, absence of temporal information, and potential confounding from operational variables (RPM/load). The pipeline supports health score calculation in stationary regimes and near real-time monitoring via interactive dashboards. Future work includes multi-asset expansion, temporal labeling, and development of prognostic models (LSTM, TCN, PHM) for RUL prediction.

Keywords: Predictive maintenance; Machine Learning; Vibration; IoT; Visual analytics; Fault detection

1 INTRODUÇÃO

A manutenção preditiva baseada em dados tem se consolidado como uma das estratégias centrais da Indústria 4.0 para aumentar a confiabilidade operacional e reduzir custos com falhas não planejadas. Segundo a ABRAMAN (2023), os gastos com manutenção podem representar até 5% do faturamento das empresas industriais, reforçando o impacto econômico do tema. Segundo um relatório da Siemens AG (2022, The true cost of downtime), no setor automotivo as linhas de produção automatizadas sofrem grandes perdas financeiras devido a falhas inesperadas em equipamentos críticos, resultando em aproximadamente 29 horas de downtime por mês, com custo médio estimado em US$ 1 milhão por hora.

Este trabalho apresenta o desenvolvimento e a avaliação de uma plataforma híbrida de predição de falhas em motores elétricos, utilizando dados de vibração coletados em laboratório pela Tractian. O dataset é composto por duas famílias de anomalias: rolamentos (INNER_RACEWAY – IR, OUTER_RACEWAY – OR) e folgas estruturais (STRUCTURAL_LOOSENESS – SL), além da condição saudável (HEALTHY – H). Para cada amostra, foram extraídas features temporais (RMS, Kurtosis, Skewness, Fator de Crista) e espectrais (Energia, Frequência de Pico, Centroide Espectral), totalizando 14 variáveis após seleção e tratamento de multicolinearidade.

O estudo foi conduzido utilizando um único ativo, o que limita a generalização para outros equipamentos, mas permite validar de forma controlada o desempenho do pipeline. O objetivo principal é verificar se a inclusão de uma camada não supervisionada (Autoencoder) reduz a ocorrência de falsos negativos em comparação com pipelines puramente supervisionados (Random Forest e SVM). Buscando preencher a lacuna da combinação entre diagnóstico supervisionado e detecção de anomalias não supervisionada.

O trabalho se insere o escopo da Indústria 4.0, IIoT e Machine Learning, com a plataforma sendo concebida para uso em cenários near real-time, com pipeline de extração de features FFT em tempo real em hardware convencional

para ambientes industriais críticos. Destaca-se que os dados utilizados são experimentais/laboratoriais, ainda não provenientes de operação contínua em campo, o que será considerado como limitação na interpretação dos resultados e na extrapolação para cenários de múltiplos ativos.

2 REVISÃO DE LITERATURA

A manutenção preditiva em máquinas rotativas tem avançado principalmente em duas frentes. A primeira envolve modelos supervisionados clássicos, como Random Forest e Support Vector Machines (SVM), que apresentam bom desempenho em cenários de diagnóstico quando há dados rotulados disponíveis (Magena, 2024). Embora robustos na classificação de falhas conhecidas, esses métodos mostram limitações na detecção de degradação inicial ou de condições anômalas não previamente observadas.

Nos últimos anos, abordagens de Deep Learning aplicadas diretamente a sinais de vibração têm ganhado destaque, permitindo que os modelos aprendam automaticamente características discriminativas a partir de séries temporais brutas. Redes convolucionais unidimensionais (1D-CNN) extraem padrões locais no tempo, especialmente quando combinadas com mecanismos de atenção e convoluções causais, garantindo uso exclusivo de dados passados para diagnóstico em tempo real (ZHANG et al., 2023). Redes recorrentes, como LSTM, capturam dependências temporais de longo prazo, enquanto modelos híbridos que combinam BiLSTM, Multi-Head Self-Attention e 1D Conv-ResNet demonstram alta robustez a ruído e excelente generalização entre diferentes modos de falha. Redes Convolucionais Temporais (TCN) preservam a ordem temporal em sequências curtas, permitindo diagnóstico rápido e prognóstico de vida útil residual, e arquiteturas baseadas em Transformers integram sinais multimodais (vibração, corrente e som) via autoatenção, atingindo acurácia superior a 99% em diagnósticos complexos (XU et al., 2025).

Outra abordagem relevante envolve Autoencoders Variacionais (VAE), que permitem modelar o estado saudável do equipamento e detectar desvios sem necessidade de rótulos. Ibrahim et al. (2024) demonstraram que a inclusão de um termo de desejabilidade no custo do VAE padroniza a localização de clusters no espaço latente, agrupando corretamente 97% das amostras de mesma falha e aumentando a consistência do diagnóstico mesmo em sinais ruidosos.

Um aspecto crítico da literatura diz respeito aos protocolos de validação. Estratégias ingênuas, como k-fold aleatório, podem introduzir vazamento de dados em séries temporais, sobretudo quando amostras correlacionadas de um mesmo ativo aparecem simultaneamente em treino e teste. Para contornar esse problema, têm sido recomendados protocolos como time-based split, que respeitam a ordem cronológica dos dados, e group k-fold por ativo, que mantém todos os dados de um mesmo equipamento juntos em treino ou teste (Bagri et al., 2024). Mesmo em cenários com dados provenientes de um único ativo, essas práticas são importantes para garantir que os resultados sejam consistentes e representativos de condições futuras, evitando superestimação de métricas.

Além disso, a robustez frente ao drift de condição operacional é outro fator crítico. Mudanças graduais na distribuição dos sinais, provocadas por desgaste dos componentes, variações de carga ou envelhecimento de sensores,

podem degradar o desempenho do modelo ao longo do tempo. Li (2025) ressalta que a manutenção preditiva em ambientes reais requer monitoramento contínuo dos dados e atualização dos modelos via aprendizado incremental ou adaptação de domínio, de modo a minimizar falsos positivos e negativos e manter a confiança no sistema.

Apesar dos avanços recentes, ainda existe uma lacuna científica na realização de comparativos sistemáticos entre pipelines híbridos supervisionado + não supervisionado aplicados a sinais vibracionais de motores elétricos. Este estudo busca integrar um Autoencoder como etapa inicial de detecção de anomalias, seguido de classificadores supervisionados (Random Forest e SVM), com o objetivo de reduzir falsos negativos e aumentar a confiabilidade em cenários industriais near real-time, utilizando protocolos de validação robustos que aproximem a pesquisa das condições reais de operação.

3 METODOLOGIA

O dataset utilizado neste estudo foi disponibilizado pela Tractian Tecnologia Ltda e compreende quatro condições de operação: H (1647 amostras), IR, OR e SL (total de 2769 amostras anômalas), resultando em 4416 janelas após o pré-processamento. Os sinais foram adquiridos por sensores de vibração industriais, com taxas de amostragem de 16 kHz para folgas estruturais e 32 kHz para rolamentos, variando conforme o ensaio. Além dos sinais de vibração, foram registrados os parâmetros operacionais de rotação do motor (RPM) e carga aplicada (Load em kW), que variam entre condições, representando diferentes regimes de operação e descartadas para modelagem. Ressalta-se que o dataset não possui dados de *datetime* e ordem cronológica de coleta, o que impossibilita análises temporais diretas ou o uso de modelos prognósticos de séries temporais. Para cada janela, foram extraídas features estatísticas no domínio do tempo (RMS, Kurtosis, Skewness, Crest Factor), espectrais (Energia de banda, Centroide Espectral, Envelope) e cepstrais (MFCC), resultando em um conjunto final de 14 variáveis após análise de correlação e remoção de multicolinearidade. A rotulagem dos dados foi realizada em ambiente controlado de laboratório, seguindo o protocolo: (i) motor em condição normal → coleta de dados H; (ii) indução de falha no rolamento → coleta de dados IR e OR; (iii) indução de folga estrutural nos mancais → coleta de dados SL. Dessa forma, considerou-se como falha qualquer janela associada às classes anômalas.

O pré-processamento gerou uma linha de features agregadas, resultando em 4416 janelas distribuídas conforme a tabela de condições: H (1647), IR (482), OR (453) e SL (1834). Cabe destacar que todos os dados provêm de um único ativo, o que limita a generalização dos resultados para outros equipamentos, mas permite a avaliação controlada do pipeline.

Os modelos supervisionados (Random Forest e SVM) foram avaliados utilizando validação cruzada estratificada, separando treino, validação e teste. A separação cronológica não foi utilizada porque o dataset não possui uma coluna datetime ou qualquer outra informação que indique a ordem temporal das amostras, impossibilitando a realização de uma divisão baseada no tempo. O Autoencoder foi treinado exclusivamente com dados da classe H, utilizando uma arquitetura densa simétrica (14–12–7–12–14) e função de perda MSE, com limiar de decisão definido pelo percentil 99 do erro de reconstrução, permitindo a detecção de

anomalias sem supervisão.

O desbalanceamento de classes foi tratado com ajustes de class_weight nos modelos supervisionados, e o tuning de limiar foi realizado para otimizar a métrica PR-AUC, priorizando-a como indicador principal. As métricas de desempenho incluíram PR-AUC, ROC-AUC, F1-score, precisão, recall, matriz de confusão e intervalos de confiança de 95% obtidos via bootstrap (n=1000), garantindo robustez estatística.

Adicionalmente, o pipeline de pré-processamento e extração de features foi estruturado para permitir implementação near real-time, com capacidade de atualização contínua à medida que novos dados sejam coletados, servindo como base para estudos futuros que envolvam múltiplos ativos, validação cross-asset e análise de progressão temporal de falhas.

Quanto à ética e conformidade, todos os dados utilizados foram anonimizados, consistindo exclusivamente em medições laboratoriais de vibração de motores elétricos, sem informações pessoais. A anonimização garante confidencialidade e conformidade com a LGPD. Os arquivos processados e scripts de pré-processamento permitem reprodução dos resultados com bases sintéticas ou simuladas, assegurando transparência e reprodutibilidade da pesquisa. Falhas induzidas seguiram protocolos controlados, sem risco para operadores ou equipamentos.

4 RESULTADOS E DISCUSSÕES

A avaliação dos modelos demonstrou elevado desempenho dos algoritmos supervisionados para classificação das condições da máquina. O Random Forest apresentou acurácia de 99,32% e o SVM 98,7%, ambos com baixa variância, evidenciada pelas curvas de aprendizado e validação cruzada. O Autoencoder, treinado exclusivamente com dados da classe H, apresentou desempenho satisfatório na detecção de anomalias não rotuladas, com AUC de 96,33%, reforçando sua aplicabilidade como filtro inicial no pipeline de diagnóstico.

Além do desempenho em classificação, os modelos permitiram calcular um score de saúde contínuo para o ativo. Nos algoritmos supervisionados, a probabilidade de cada classe foi utilizada como indicador de proximidade de falha, enquanto no Autoencoder o erro de reconstrução foi interpretado como medida de degradação, possibilitando monitoramento do estado do equipamento ao longo do tempo. Essa abordagem permite identificar alterações no comportamento do motor antes que falhas se tornem críticas, oferecendo suporte à tomada de decisão preventiva.

As métricas foram avaliadas considerando intervalos de confiança de 95% obtidos via bootstrap (n=1000), e incluíram PR-AUC, F1-score, ROC-AUC, precisão, recall e matriz de confusão, fornecendo uma visão robusta do desempenho do sistema mesmo frente ao desbalanceamento das classes.

5 CONCLUSÃO

Os resultados confirmam a viabilidade de integrar algoritmos supervisionados e não supervisionados em um pipeline único de manutenção preditiva. O Random Forest se destacou pela robustez e interpretabilidade, adequado à classificação de condições conhecidas, enquanto o SVM apresentou boa separação de classes em espaços de

alta dimensionalidade. O Autoencoder atuou como filtro inicial de anomalias, reduzindo falsos negativos e fornecendo um indicador contínuo de degradação. A combinação dos modelos permite calcular um score de saúde contínuo, integrando probabilidades de classe (RF, SVM) e erro de reconstrução (AE), ampliando a capacidade de monitoramento e suporte à decisão preventiva em cenários industriais near real-time.

Do ponto de vista prático, a plataforma de visual analytics facilita a interpretação por meio de dashboards interativos e consultas em linguagem natural. Entre as limitações, destaca-se o uso de dados de um único motor em ambiente laboratorial, restringindo a generalização, e a ausência de dados temporais, que impede estratégias de prognóstico de vida útil residual (RUL). Futuras pesquisas devem explorar modelos baseados em PHM, Weibull, LSTM e TCN, ampliar a base para múltiplos ativos, incluir métricas de RUL e avaliar escalabilidade em operação contínua.

Conclui-se que a solução é promissora para ambientes industriais críticos, pois combina alta acurácia, interpretabilidade e integração visual, demonstrando que pipelines híbridos podem reduzir falsos negativos e fornecer scores de saúde mais confiáveis.

REFERÊNCIAS

BAGRI, Ikram; TAHIRY, Karim; HRAIBA, Aziz; TOUIL, Achraf; MOUSRIJ, Ahmed. Vibration signal analysis for intelligent rotating machinery diagnosis and prognosis: a comprehensive systematic literature review. Vibration, v. 7, n. 4, p. 1013–1062, 2024. DOI: 10.3390/vibration7040054.

IBRAHIM, Rony; ZEMOURI, Ryad; TAHAN, Antoine; KEDJAR, Bachir; MERKHOUF, Arezki; AL-HADDAD, Kamal. Fault detection based on vibration measurements and variational autoencoder-desirability function. IEEE Open Journal of Industry Applications, v. 5, p. 106–116, 2024. DOI: 10.1109/OJIA.2024.3380249.

LI, Wenjun; LI, Ting. Comparison of deep learning models for predictive maintenance in industrial manufacturing systems using sensor data. Scientific Reports, v. 15, p. 23545, 2025. DOI: 10.1038/s41598-025-08515-z.

MAGENA, C. Machine learning models for predictive maintenance in industrial engineering. International Journal of Computing and Engineering, v. 6, n. 3, p. 1–14, 2024.

SENSEYE PREDICTIVE MAINTENANCE. The true cost of downtime 2022. Siemens AG, 2023.

XU, Bo; LI, Huipeng; DING, Ruchun; ZHOU, Fengxing. Fault diagnosis in electric motors using multi-mode time series and ensemble transformers network. Scientific Reports, v. 15, p. 23545, 2025. DOI: 10.1038/s41598-025-89695-6.

ZHANG, Hui; GE, Baojun; HAN, Bin. Real-time motor fault diagnosis based on TCN and attention. Machines, v. 10, n. 4, p. 249, 2022. DOI: 10.3390/machines10040249.

SOBRE O(S)AUTOR(ES)

[i] **ALAN DIEK GUIMARÃES**

Graduando em Ciência de Dados pela Faculdade SENAI São Paulo – Campus Paulo Antônio Skaf. Graduado em Gestão da tecnologia da informação pela Uninter.
https://orcid.org/0009-0004-0738-9212

[ii] **JUAN VIANA LORENZO**

Graduando em Ciência de Dados pela Faculdade SENAI São Paulo – Campus Paulo Antônio Skaf, com experiência prévia como desenvolvedor júnior, estagiário em automação de processos e assistente de monitoramento de infraestrutura. Possui formação técnica em Redes de Computadores e em Análise e Desenvolvimento de Sistemas pelo SENAI, além de conclusão do Ensino Médio pelo SESI. Une sólida base acadêmica a vivência prática em tecnologia, com foco em soluções eficientes e inovação.
https://orcid.org/0009-0003-7532-7198

[ii] **MARCOS VINICIUS OLIVEIRA DOS SANTOS**

Graduando em Ciência de Dados pela Faculdade SENAI São Paulo – Campus Paulo Antônio Skaf. Atualmente, atuo como Instrutor de Formação Profissional no SENAI, com experiência anterior como desenvolvedor Full Stack e professor em ONG voltada ao atendimento de crianças em situação de vulnerabilidade social. Possuo formação técnica em Desenvolvimento de Sistemas pelo SENAI, aliando sólida base tecnológica a experiência prática em educação e desenvolvimento de soluções digitais.
https://orcid.org/my-orcid?orcid=0009-0008-9151-9612

[ii] **VINICIUS DE JESUS SILVA**

Graduando em Ciência de Dados pela Faculdade SENAI São Paulo – Campus Paulo Antônio Skaf. Atualmente sou estagiário na área de Ciência de Dados. Sou formado como Técnico em Logística e possuo experiência prévia em automação, machine learning, Python e análise de dados, atuando no desenvolvimento de soluções voltadas ao tratamento, modelagem e visualização de informações para apoio à tomada de decisão.
https://orcid.org/0009-0002-8151-9439

[ii] **VINICIUS SCHILIEVE SANTOS**

Graduando em Ciência de Dados pela Faculdade SENAI São Paulo Campus Paulo Antônio Skaf (2024–2025), do Serviço Nacional de Aprendizagem Industrial (SENAI) de Tecnologia. Formado como Técnico em Informática pela Escola Técnica Estadual (ETEC) (2014–2015). https://orcid.org/0009-0002-7713-7075