

NEUROEVOLUÇÃO APLICADA À GERAÇÃO DE DADOS SINTÉTICOS PARA DIAGNÓSTICO DE FALHAS EM MÁQUINAS ROTATIVAS EM CENÁRIOS DE ESCASSEZ DE DADOS

NEUROEVOLUTION APPLIED TO SYNTHETIC DATA GENERATION FOR
FAULT DIAGNOSIS IN ROTATING MACHINERY UNDER DATA-SCARCE
SCENARIOS

Ciências Exatas e da Terra, Engenharias • 31/08/2026

REGISTRO DOI: [10.70773/revistatopicos/788063363](https://doi.org/10.70773/revistatopicos/788063363)

Hygor Santiago Lara¹

Milton Dias Júnior²

RESUMO

O diagnóstico de falhas em máquinas rotativas por meio da análise de vibrações é amplamente utilizado em sistemas de manutenção preditiva. Entretanto, a aplicação de técnicas de aprendizado de máquina nessa área é frequentemente limitada pela escassez de dados rotulados, característica comum em ambientes industriais. Neste contexto, este trabalho tem como objetivo desenvolver uma metodologia baseada em neuroevolução para melhorar o desempenho de modelos preditivos de diagnóstico de falhas mecânicas sob condições de baixa disponibilidade de dados. A pesquisa possui caráter experimental e utiliza uma abordagem híbrida composta por um Classificador baseado em Redes Neurais profundas, um Gerador de dados sintéticos baseado em Variational Autoencoder e um processo de otimização evolutiva responsável pela definição das arquiteturas dos modelos. A metodologia adota um processo recursivo no qual o desempenho do Classificador orienta a evolução do Gerador e a incorporação progressiva de novas amostras sintéticas ao conjunto de treinamento. Os resultados demonstram ganhos consistentes de acurácia, precisão e recall em comparação com modelos treinados apenas com dados reais ou com técnicas convencionais de aumento de dados, especialmente em falhas combinadas e classes sub-representadas. Como principal contribuição, o trabalho demonstra que a combinação de neuroevolução e geração sintética de dados constitui uma estratégia eficaz para reduzir os efeitos da escassez de dados em sistemas de diagnóstico de falhas mecânicas.

Palavras-chave: Diagnóstico de falhas; Máquinas rotativas; Neuroevolução; Geração de dados sintéticos; Manutenção preditiva.

ABSTRACT

Fault diagnosis in rotating machinery through vibration analysis is

widely used in predictive maintenance systems. However, the application of machine learning techniques in this field is often limited by the scarcity of labeled data, a common characteristic of industrial environments. In this context, this work aims to develop a neuroevolution-based methodology to improve the performance of predictive models for mechanical fault diagnosis under conditions of limited data availability. The research is experimental in nature and employs a hybrid approach composed of a deep neural network classifier, a synthetic data generator based on a Variational Autoencoder, and an evolutionary optimization process responsible for defining the architectures of the models. The proposed methodology adopts a recursive process in which the classifier's performance guides the evolution of the generator and the progressive incorporation of new synthetic samples into the training dataset. The results demonstrate consistent improvements in accuracy, precision, and recall when compared to models trained solely on real data or using conventional data augmentation techniques, particularly for combined faults and underrepresented classes. As its main contribution, this work demonstrates that the combination of neuroevolution and synthetic data generation constitutes an effective strategy for mitigating the effects of data scarcity in mechanical fault diagnosis systems.

Keywords: Fault diagnosis; Rotating machinery; Neuroevolution; Synthetic data generation; Predictive maintenance.

1. INTRODUÇÃO

O diagnóstico de falhas em máquinas rotativas por meio da análise de vibrações desempenha papel fundamental na manutenção preditiva, permitindo a identificação precoce de defeitos e a redução de paradas não programadas. Com o avanço das técnicas de

aprendizado de máquina, modelos baseados em Redes Neurais têm sido amplamente utilizados para automatizar esse processo, alcançando elevados níveis de desempenho em diferentes aplicações. Entretanto, a eficácia desses modelos depende fortemente da disponibilidade de dados rotulados, condição que raramente é encontrada em ambientes industriais, onde falhas reais são eventos pouco frequentes e de difícil obtenção.

Nesse contexto, este trabalho investiga o uso de neuroevolução como estratégia para melhorar o desempenho de modelos preditivos em cenários de escassez de dados. A metodologia proposta integra um Classificador baseado em Redes Neurais Profundas, um Gerador de dados sintéticos fundamentado em *Variational Autoencoders* e um Algoritmo Genético responsável pela otimização das arquiteturas e pelo processo evolutivo de geração de amostras. Por meio de ciclos recursivos de treinamento, geração sintética e reavaliação, o sistema busca produzir dados cada vez mais representativos e úteis para a tarefa de diagnóstico.

O objetivo principal é desenvolver e validar uma metodologia neuroevolutiva capaz de aumentar a capacidade de generalização de modelos de diagnóstico de falhas mecânicas em máquinas rotativas sob condições de dados limitados. Como objetivos específicos, busca-se avaliar a qualidade das amostras sintéticas geradas, analisar a organização do espaço latente dos modelos generativos, investigar o comportamento do processo evolutivo e quantificar os ganhos de desempenho obtidos pelo Classificador em comparação com abordagens convencionais de treinamento e aumento de dados.

2. FUNDAMENTAÇÃO

Máquinas rotativas estão entre os equipamentos mais utilizados em ambientes industriais, sendo empregadas em sistemas de geração de energia, processos de manufatura, transporte e bombeamento. Devido à sua importância operacional, falhas nesses equipamentos podem causar perdas econômicas significativas, redução da produtividade e riscos à segurança operacional. Nesse contexto, o monitoramento de condições tornou-se uma ferramenta essencial para a implementação de estratégias de manutenção preditiva.

Entre as diversas técnicas de monitoramento disponíveis, a análise de vibrações destaca-se como uma das mais utilizadas para diagnóstico de falhas mecânicas. O princípio dessa técnica baseia-se no fato de que defeitos mecânicos alteram a dinâmica do sistema, produzindo modificações mensuráveis nos sinais de vibração. Essas alterações podem ser identificadas por meio da análise temporal ou espectral dos sinais adquiridos por sensores instalados na máquina ((INMAN, 2014); (RANDALL, 2011)).

Os defeitos mais frequentemente encontrados em sistemas rotativos incluem desbalanceamento, desalinhamento, folgas mecânicas e falhas em rolamentos. Cada um desses defeitos produz assinaturas vibratórias características, frequentemente associadas a componentes harmônicas específicas da frequência de rotação da máquina. Entretanto, em aplicações reais, múltiplos defeitos podem ocorrer simultaneamente, tornando o processo de diagnóstico significativamente mais complexo devido à sobreposição de padrões vibratórios ((MUSZYNSKA, 2005); (RAO, 2017)).

O processamento dos sinais de vibração tem como objetivo extrair informações relevantes capazes de caracterizar o estado operacional

da máquina. Para isso, os sinais podem ser analisados tanto no domínio do tempo quanto no domínio da frequência.

No domínio do tempo, métricas estatísticas como valor eficaz (RMS), curtose, fator de crista e desvio padrão são amplamente utilizadas como indicadores de condição. Já no domínio da frequência, a Transformada Discreta de Fourier permite decompor o sinal em suas componentes harmônicas, facilitando a identificação de frequências associadas a defeitos específicos (RANDALL; ANTONI, 2011).

A utilização conjunta de informações temporais e espectrais tem se mostrado particularmente eficiente para o diagnóstico de falhas mecânicas, uma vez que diferentes defeitos podem manifestar-se de forma mais evidente em domínios distintos. Por essa razão, abordagens modernas frequentemente exploram representações híbridas capazes de combinar informações provenientes de múltiplos domínios de análise.

O avanço do aprendizado profundo possibilitou o desenvolvimento de modelos capazes de extrair automaticamente características relevantes diretamente dos sinais de entrada. Diferentemente das abordagens tradicionais, que dependem da seleção manual de atributos, as Redes Neurais Profundas aprendem representações hierárquicas dos dados durante o processo de treinamento (GOODFELLOW; BENGIO; COURVILLE, 2016).

As Redes Neurais Convolucionais destacam-se pela capacidade de identificar padrões locais e correlações espaciais em sinais e imagens. Originalmente propostas para tarefas de visão computacional, essas redes passaram a ser amplamente utilizadas em diagnóstico de falhas devido à sua eficiência na extração

automática de características espectrais e temporais ((LECUN et al., 1998); (KRIZHEVSKY; SUTSKEVER; HINTON, 2012)).

Por outro lado, as Redes Neurais Recorrentes como as *Long Short-Term Memory* (LSTM) foram desenvolvidas para modelar dependências temporais de longo prazo em dados sequenciais. Sua arquitetura incorpora mecanismos de memória que permitem preservar informações relevantes ao longo do tempo, tornando-as adequadas para análise de sinais de vibração e séries temporais industriais (HOCHREITER; SCHMIDHUBER, 1997).

A combinação entre Redes Convolucionais e Recorrentes permite explorar simultaneamente características locais e dependências temporais, resultando em modelos particularmente eficazes para classificação de falhas em máquinas rotativas.

Embora modelos de aprendizado profundo apresentem elevado potencial de desempenho, seu treinamento normalmente requer grandes quantidades de dados rotulados. Em aplicações industriais, entretanto, a disponibilidade de dados de falha costuma ser limitada, uma vez que defeitos reais são eventos relativamente raros e frequentemente indesejáveis.

Esse cenário conduz a problemas de sobreajuste (*overfitting*), nos quais o modelo aprende excessivamente as particularidades dos dados de treinamento e perde capacidade de generalização. Como consequência, o desempenho observado durante o treinamento pode não se reproduzir quando o modelo é aplicado a novos dados.

Diversas estratégias têm sido propostas para mitigar esse problema, incluindo transferência de aprendizado, regularização, validação cruzada (*cross validation*) e técnicas de aumento de dados (*data*

augmentation). Entre essas abordagens, a geração de dados sintéticos tem recebido atenção crescente devido à sua capacidade de ampliar artificialmente o conjunto de treinamento sem necessidade de novos experimentos físicos.

Autoencoders são modelos não supervisionados projetados para aprender representações compactas dos dados. Sua arquitetura é composta por duas partes principais: um Codificador (*encoder*), responsável por projetar os dados em um Espaço Latente de menor dimensionalidade, e um Decodificador (*decoder*), responsável por reconstruir os dados originais a partir dessa representação comprimida (LI et al., 2023).

Apesar de sua eficiência na compressão de dados, *Autoencoders* convencionais produzem representações determinísticas, o que limita sua utilização para geração de novas amostras.

Os *Variational Autoencoders* estendem essa abordagem ao introduzir uma formulação probabilística para o Espaço Latente. Em vez de aprender pontos específicos, o modelo aprende distribuições de probabilidade caracterizadas por médias e variâncias. Durante o treinamento, uma regularização baseada na divergência de Kullback-Leibler força essas distribuições a se aproximarem de uma distribuição normal padrão, produzindo um espaço latente contínuo e estruturado (KINGMA; WELLING, 2014).

Essa propriedade torna este modelo particularmente adequado para geração de dados sintéticos, uma vez que novos pontos podem ser amostrados diretamente do espaço latente e convertidos em amostras realistas por meio do Decodificador.

Os Algoritmos Genéticos são técnicas de otimização inspiradas nos mecanismos de evolução natural descritos por Darwin. Nessas abordagens, potenciais soluções são representadas como indivíduos que evoluem ao longo de gerações por meio de operadores de seleção, cruzamento e mutação ((HOLLAND, 1975); (GOLDBERG, 1989)).

Uma das principais vantagens dos Algoritmos Genéticos é sua capacidade de explorar espaços de busca complexos, não lineares e multimodais, sem necessidade de informações sobre gradientes. Essa característica os torna particularmente adequados para problemas envolvendo otimização de arquiteturas neurais e hiperparâmetros.

Quando aplicados ao desenvolvimento de Redes Neurais, esses métodos compõem a área conhecida como Neuroevolução. Nessa abordagem, mecanismos evolutivos são utilizados para otimizar parâmetros estruturais dos modelos, como número de camadas, quantidade de neurônios, funções de ativação e dimensões do espaço latente.

No presente trabalho, a Neuroevolução é empregada para otimizar a arquitetura do Gerador de dados sintéticos e orientar um processo recursivo de aprimoramento do Classificador. Dessa forma, a geração de amostras sintéticas deixa de ser uma etapa isolada e passa a integrar um mecanismo adaptativo capaz de produzir dados cada vez mais úteis para a tarefa de diagnóstico.

A integração entre modelos generativos e algoritmos evolutivos representa uma abordagem promissora para cenários de escassez de dados. Enquanto o Variational *Autoencoder* é responsável por

aprender a estrutura estatística dos sinais de vibração e gerar novas amostras, o Algoritmo Genético atua como mecanismo de otimização, selecionando arquiteturas e regiões do espaço latente que maximizam o desempenho do Classificador.

Essa estratégia cria um ciclo recursivo de aprendizado no qual a qualidade das amostras sintéticas é avaliada indiretamente por sua contribuição ao desempenho do modelo supervisionado. Como consequência, o sistema evolui progressivamente para produzir dados mais informativos e representativos das condições operacionais de interesse.

Essa abordagem constitui a base conceitual da metodologia proposta nesta tese, cuja hipótese central é que a combinação entre geração sintética de dados e Neuroevolução pode reduzir os impactos da escassez de dados e melhorar a capacidade de generalização de modelos de diagnóstico de falhas mecânicas em máquinas rotativas.

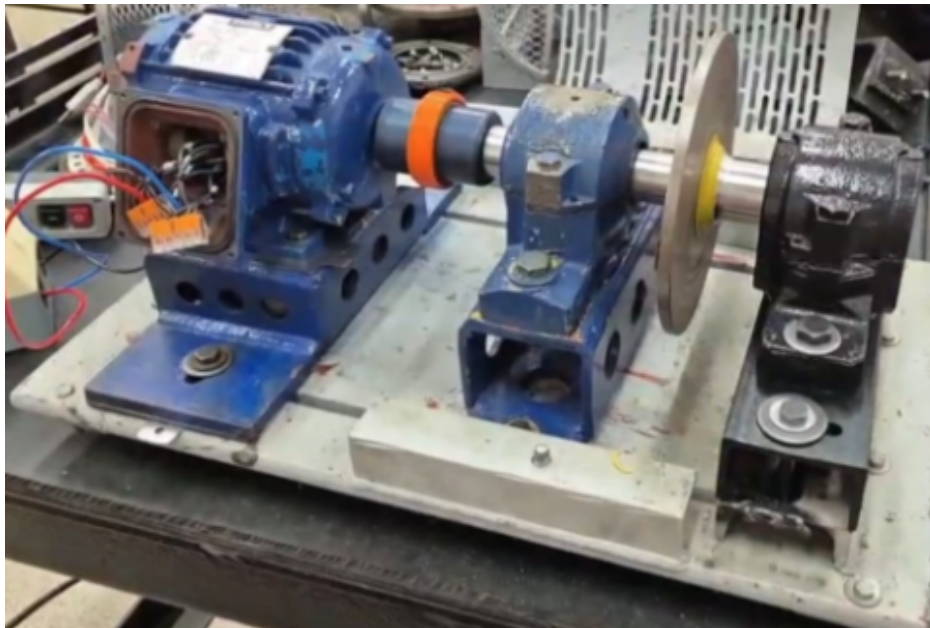
3. METODOLOGIA

A metodologia proposta neste trabalho tem como objetivo desenvolver um sistema híbrido para diagnóstico automático de falhas em sistemas rotativos, capaz de operar adequadamente mesmo em cenários de escassez de dados. Para isso, são integradas técnicas de aprendizado supervisionado, aprendizado não supervisionado e otimização evolutiva, combinando Redes Neurais Artificiais, Variational Autoencoders (VAE) e Algoritmos Genéticos (AG).

O banco de dados utilizado foi obtido a partir de uma bancada experimental composta por um motor de indução trifásico acoplado

a um eixo suportado por mancais de rolamento, como apresentado na Figura 1. O sistema opera sob carga constante aplicada por um freio de Prony e rotação próxima de 1800 rpm. Os sinais de vibração foram adquiridos por meio de um acelerômetro triaxial instalado em um dos mancais, permitindo medições nas direções axial, horizontal e vertical. As aquisições foram realizadas com frequência de amostragem de 8000 Hz e duração de 1 segundo, resultando em 8000 amostras por canal. Cada registro experimental foi organizado em uma matriz contendo os três sinais de vibração sincronizados.

Figura 1. Bancada desenvolvida.



Fonte: Autor.

O banco de dados contempla oito condições operacionais: 1) condição normal, 2) desbalanceamento, 3) desalinhamento, 4) falha em rolamento, 5) falha em rolamento associada ao desbalanceamento, 6) falha em rolamento associada ao desalinhamento, 7) falha em rolamento associada simultaneamente ao desbalanceamento e desalinhamento, e 8) desbalanceamento associado ao desalinhamento. Foram adquiridas 30 amostras para cada condição, totalizando 240 registros experimentais.

O sistema de diagnóstico foi estruturado em quatro módulos principais: Classificador, Autoencoder, Gerador de Amostras Sintéticas e Modelo Híbrido Integrado.

O Classificador é responsável pela identificação das condições operacionais do sistema rotativo. Sua arquitetura é definida automaticamente por um Algoritmo Genético, que realiza a busca simultânea pela melhor combinação de camadas convolucionais, recorrentes do tipo LSTM e densas, além dos principais hiperparâmetros de treinamento. Cada indivíduo da população representa uma arquitetura completa de Rede Neural, sendo avaliado de acordo com seu desempenho em validação, complexidade estrutural e custo computacional. O processo evolutivo utiliza operadores de seleção, cruzamento e mutação para explorar diferentes configurações até a obtenção de modelos com elevada capacidade de generalização.

Paralelamente, é desenvolvido um *Autoencoder* destinado ao aprendizado não supervisionado das características dos sinais de vibração. Assim como no Classificador, sua arquitetura é definida automaticamente por Algoritmo Genético. O modelo é composto por um Codificador responsável por reduzir a dimensionalidade dos dados e um Decodificador encarregado de reconstruir os sinais originais. O treinamento busca minimizar o erro de reconstrução, permitindo que o Espaço Latente aprenda representações compactas das diferentes condições operacionais.

A partir do Autoencoder é construído um *Variational Autoencoder*. Diferentemente do modelo convencional, o VAE aprende uma representação probabilística do Espaço Latente, modelando a distribuição dos dados experimentais. Essa característica permite

gerar novas amostras por meio da amostragem de vetores latentes e posterior reconstrução pelo Decodificador. O treinamento do VAE é realizado utilizando exclusivamente dados reais, garantindo que as características físicas dos sinais sejam preservadas na distribuição latente aprendida.

O módulo Gerador utiliza o Decodificador do *Variational Autoencoder* associado a um Algoritmo Genético para produzir novas amostras sintéticas. Nesse processo, cada indivíduo da população corresponde a um vetor do Espaço Latente do VAE. Os operadores genéticos atuam diretamente sobre esses vetores, explorando regiões promissoras do Espaço Latente. As amostras geradas são avaliadas por um Classificador previamente treinado, que verifica sua coerência em relação à classe desejada. Além disso, critérios de diversidade e similaridade são empregados para evitar a geração de amostras redundantes ou excessivamente próximas dos dados reais.

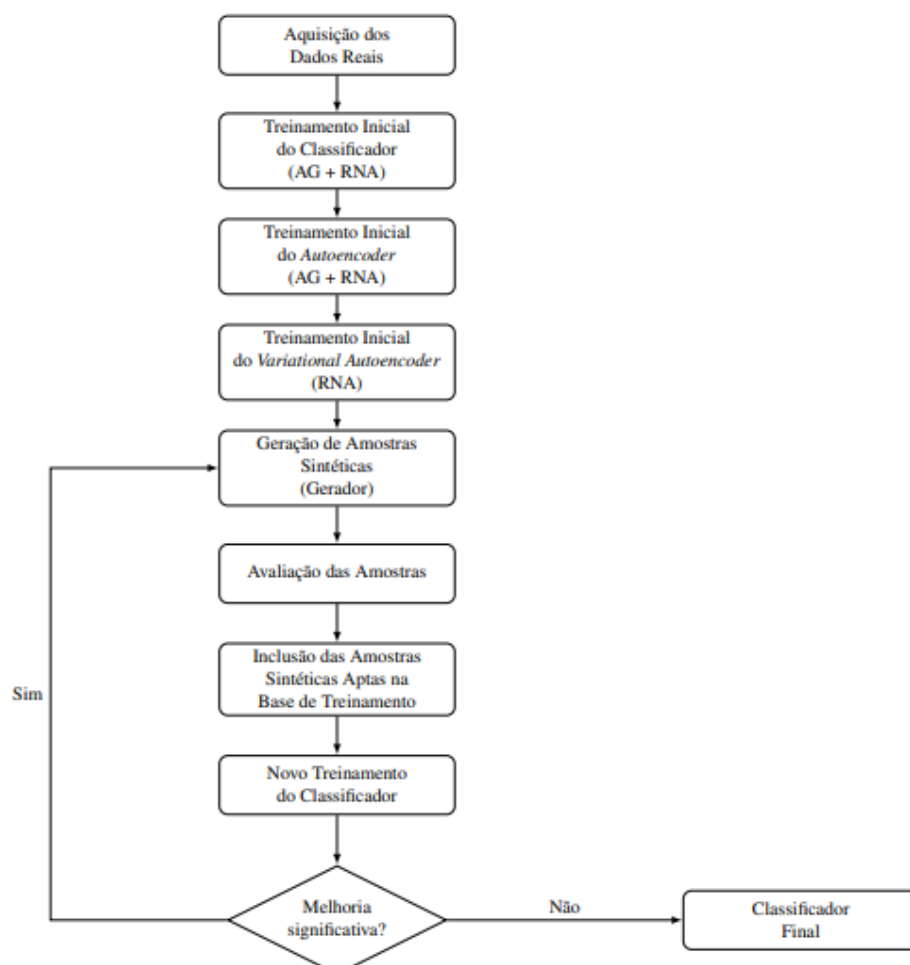
As amostras sintéticas consideradas válidas são incorporadas ao conjunto de treinamento de forma controlada. Os conjuntos de validação e teste permanecem compostos exclusivamente por dados reais, garantindo uma avaliação imparcial do desempenho do sistema. Com a nova base de dados ampliada, um novo Classificador é treinado e avaliado.

O processo completo é executado de forma iterativa. Após cada ciclo de geração e incorporação de amostras sintéticas, o Classificador é novamente otimizado e avaliado. O procedimento é encerrado quando não são observadas melhorias significativas no desempenho, quando o gerador deixa de produzir amostras adequadas ou quando é atingido o número máximo de ciclos

definido experimentalmente. A Figura 2 apresenta um fluxograma da metodologia proposta.

Dessa forma, a metodologia proposta estabelece uma integração entre aprendizado supervisionado, aprendizado não supervisionado e otimização evolutiva, permitindo a ampliação controlada do conjunto de treinamento e o desenvolvimento de modelos de diagnóstico mais robustos. A abordagem busca reduzir os efeitos da escassez de dados experimentais, problema recorrente em aplicações de manutenção preditiva e monitoramento de condição de sistemas rotativos.

Figura 2. Fluxograma da metodologia proposta.



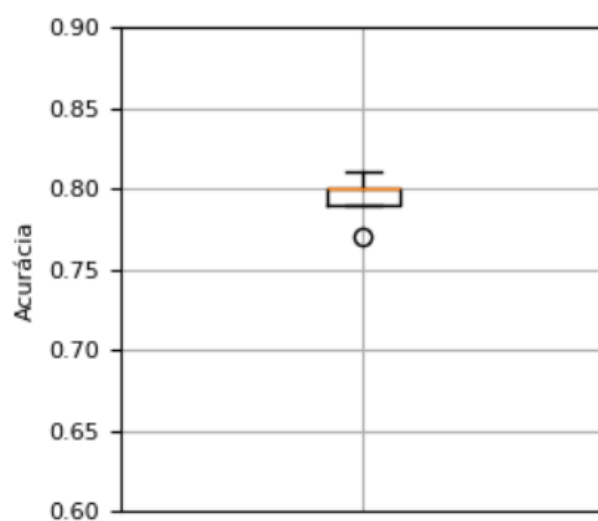
Fonte: Autor.

4. RESULTADOS E DISCUSSÕES

Os experimentos foram conduzidos utilizando o banco de dados obtido na bancada experimental descrita anteriormente, composto por 240 amostras distribuídas entre oito condições operacionais distintas. O conjunto de dados foi dividido em subconjuntos de treinamento, validação e teste, preservando a proporção entre as classes e garantindo que o conjunto de teste permanecesse composto exclusivamente por dados reais.

Devido ao caráter estocástico dos Algoritmos Genéticos e do treinamento das Redes Neurais, cada experimento foi executado múltiplas vezes de forma independente. Essa estratégia permitiu avaliar a robustez do processo de otimização e a consistência dos resultados obtidos, Figura 3.

Figura 3. Variabilidade do desempenho entre execuções independentes.



Fonte: Autor.

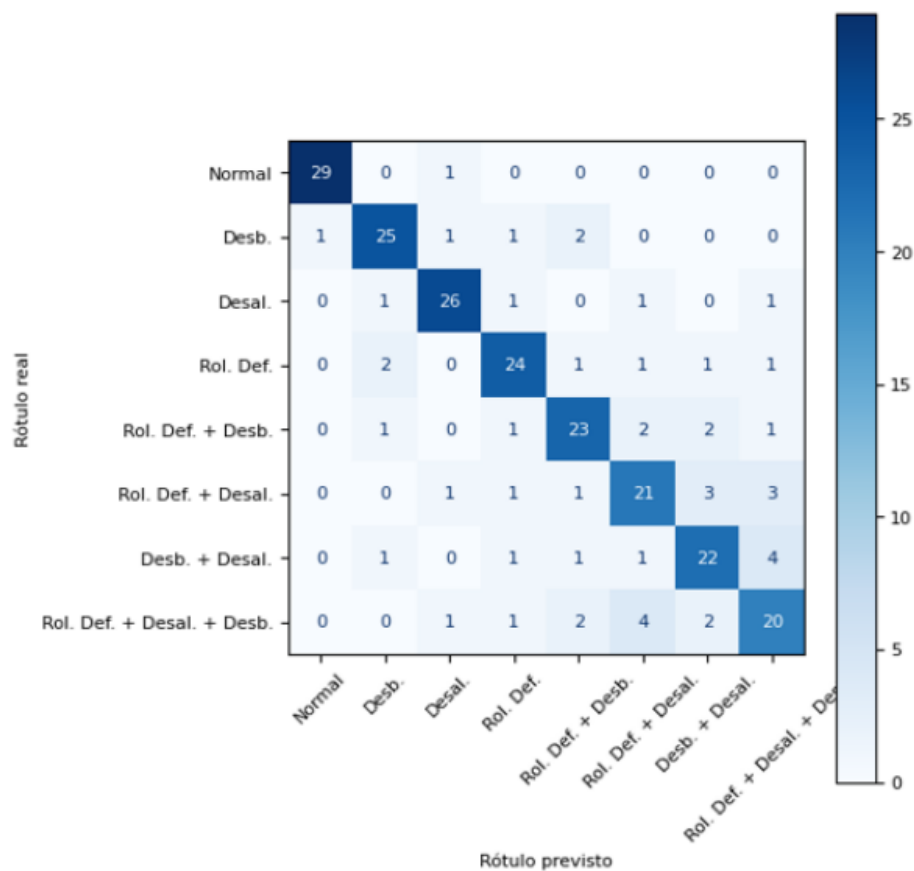
Os resultados evidenciaram baixa dispersão entre as execuções, indicando que a metodologia proposta apresenta comportamento estável e reproduzível mesmo diante das variações inerentes ao processo evolutivo.

Esta etapa teve como objetivo estabelecer uma referência de desempenho para o problema estudado utilizando exclusivamente dados reais, sem qualquer mecanismo de aumento de dados. Esse cenário representa uma situação frequentemente encontrada em aplicações industriais reais, nas quais a obtenção de amostras rotuladas é limitada pelo alto custo de instrumentação, indisponibilidade de falhas reais ou restrições operacionais.

Os resultados demonstram que o processo de neuroevolução foi capaz de selecionar arquiteturas adequadas para o problema, produzindo modelos com elevada capacidade de discriminação mesmo diante de um conjunto de treinamento relativamente pequeno. A combinação automática de blocos convolucionais, recorrentes e densos permitiu explorar simultaneamente características espectrais locais e dependências temporais de longo alcance presentes nos sinais de vibração.

Apesar desse desempenho satisfatório, a matriz de confusão evidencia limitações importantes associadas à escassez de dados, Figura 4. As maiores taxas de erro concentram-se nas classes compostas por múltiplas falhas simultâneas, situação esperada devido à elevada sobreposição de assinaturas vibratórias.

Figura 4. Matriz de confusão do Classificador primário.



Fonte: Autor.

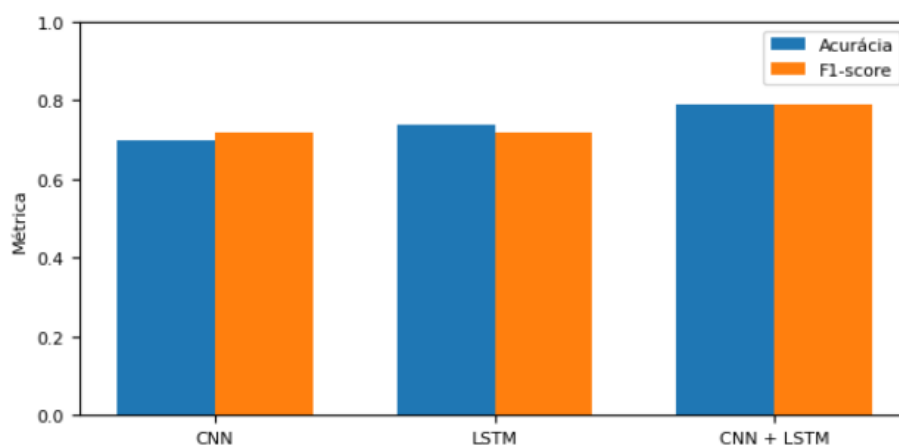
Do ponto de vista físico, falhas combinadas produzem espectros que contêm simultaneamente componentes associadas ao desbalanceamento, desalinhamento e defeitos em rolamentos. Como consequência, diferentes classes passam a compartilhar características semelhantes no espaço de atributos, reduzindo a separabilidade entre elas. Em bancos de dados reduzidos, essa dificuldade torna-se ainda mais pronunciada, pois o modelo dispõe de poucas amostras para aprender adequadamente a variabilidade interna de cada condição operacional.

A análise do *recall* por classe reforça essa interpretação. As condições normais de operação e os defeitos isolados apresentam elevados índices de reconhecimento, indicando que os padrões vibratórios dominantes dessas falhas são suficientemente distintos para serem identificados mesmo com poucos exemplos. Em contrapartida, as falhas compostas apresentam redução significativa

do *recall*, evidenciando aumento na ocorrência de falsos negativos e maior dificuldade de generalização.

Outro aspecto relevante é a comparação entre arquiteturas encontradas pelo Algoritmo Genético. Observa-se que as arquiteturas híbridas superam tanto as arquiteturas exclusivamente convolucionais quanto aquelas compostas apenas por blocos LSTM. A Figura 5 apresenta o desempenho de diferentes composições de arquiteturas.

Figura 5. Desempenho de diferentes arquiteturas encontradas.



Fonte: Autor.

Esse resultado sugere que a tarefa de diagnóstico estudada depende simultaneamente da identificação de padrões locais no espectro e da modelagem temporal das séries de vibração. Enquanto as CNN apresentam maior eficiência na extração de características locais, as LSTM contribuem para capturar relações temporais associadas à evolução dinâmica das falhas. A combinação dessas duas abordagens produz representações mais ricas e discriminativas.

Entretanto, mesmo as melhores arquiteturas encontradas apresentaram desempenho limitado pelas características do banco

de dados disponível. Os ganhos obtidos por refinamento arquitetural mostraram-se relativamente modestos quando comparados às limitações impostas pela quantidade reduzida de amostras. Esse comportamento sugere que o principal gargalo do problema não está apenas na capacidade do modelo, mas principalmente na representatividade estatística dos dados disponíveis para treinamento.

A análise do Espaço Latente constitui uma etapa fundamental da metodologia proposta, pois a qualidade das representações aprendidas influencia diretamente a capacidade de geração de novas amostras sintéticas. Um espaço latente adequado deve apresentar simultaneamente separabilidade entre classes, continuidade geométrica e estabilidade estatística.

Os resultados da Tabela 1 demonstram superioridade consistente do *Variational Autoencoder* em todas as métricas avaliadas. O aumento do *Fisher Score* indica maior separação entre classes, enquanto a redução da métrica *Smoothness* evidencia transições mais suaves e previsíveis no Espaço Latente.

Tabela 1. Comparação entre o espaço latente de *Autoencoder* e *Variational Autoencoder*.

Métrica	<i>Autoencoder</i>	<i>Variational Autoencoder</i>
<i>Fisher Score</i>	0,42	0,91
<i>Smoothness</i>	0,83	0,21
<i>KS Normality</i>	0,38	0,09
<i>Linear Accuracy</i>	71%	85%

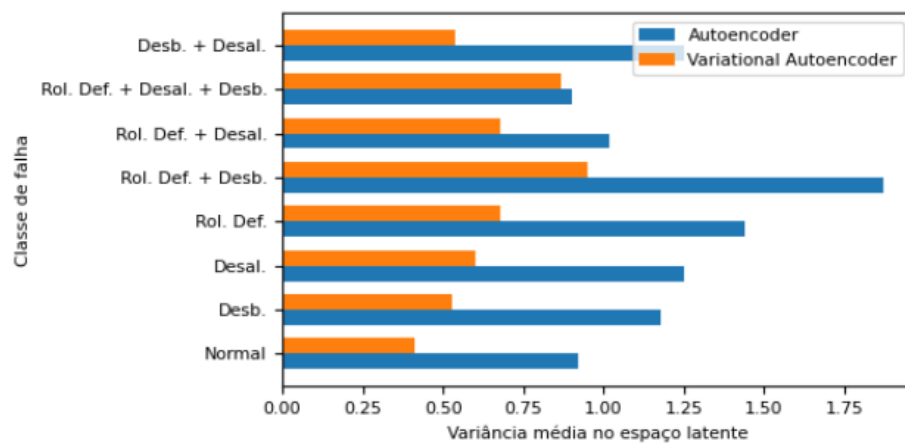
Fonte: Autor.

Essa melhoria decorre da regularização probabilística introduzida pelo termo de divergência KL. Diferentemente do *Autoencoder* convencional, que aprende representações determinísticas, o *Variational Autoencoder* força as representações a seguirem uma distribuição contínua e estruturada. Como consequência, regiões próximas do Espaço Latente tendem a produzir sinais fisicamente semelhantes, propriedade essencial para geração de dados sintéticos consistentes.

A melhoria observada na métrica *Linear Accuracy* também possui importância prática significativa. O aumento da separabilidade linear sugere que informações relevantes para o diagnóstico de falhas foram organizadas de forma mais eficiente no Espaço Latente. Isso facilita tanto tarefas de classificação quanto processos de busca evolutiva conduzidos posteriormente pelo Algoritmo Genético.

A distribuição das classes no Espaço Latente mostra agrupamentos coerentes com a física do problema. As condições normais e os defeitos isolados ocupam regiões relativamente compactas, enquanto as falhas combinadas distribuem-se em áreas intermediárias. A Figura 6 apresenta a variância entre as classes.

Figura 6. Variância do Espaço Latente por falha.



Fonte: Autor.

Esse comportamento é particularmente interessante porque reproduz relações físicas reais entre os defeitos. Falhas compostas efetivamente compartilham características de múltiplos mecanismos de degradação, sendo esperado que ocupem regiões intermediárias do espaço de representação. Dessa forma, o modelo não apenas separa as classes, mas também preserva relações estruturais coerentes entre elas.

Outro aspecto importante refere-se ao custo computacional dos modelos, Tabela 2. Embora o *Variational Autoencoder* apresente maior custo de treinamento e consumo de memória, os ganhos obtidos na organização do espaço latente justificam sua utilização como núcleo do sistema gerador.

Tabela 2. Comparação entre o custo computacional dos modelos.

Métrica	<i>Autoencoder</i>	<i>Variational Autoencoder</i>
Tempo médio por época [s]	0,42	0,78
Número médio de épocas até a convergência	65	110
Tempo total de treinamento [s]	27,3	85,8

Uso relativo de memória GPU	1,0x	1,6x
Estabilidade de convergência	Alta	Moderada

Fonte: Autor.

A qualidade das amostras sintéticas representa um dos aspectos mais críticos desta pesquisa, uma vez que dados artificiais inadequados podem introduzir vieses e comprometer o desempenho do Classificador.

A comparação estatística apresentada na Tabela 3 demonstra elevada concordância entre sinais reais e sintéticos para todas as classes avaliadas. RMS, energia espectral e amplitude na frequência de rotação apresentam diferenças reduzidas, indicando que o Gerador foi capaz de preservar as características vibratórias fundamentais de cada condição operacional.

A preservação do RMS possui relevância particular porque essa métrica está diretamente associada ao nível global de vibração da máquina. Da mesma forma, a manutenção da energia espectral indica que a distribuição energética dos sinais foi adequadamente reproduzida. Já a preservação da amplitude em 1x f_r demonstra que os mecanismos associados ao desbalanceamento e outras excitações síncronas permaneceram representados de forma consistente.

Tabela 3. Avaliação da similaridade intra-classes das amostras sintéticas.

Classe	RMS [m/s ²]	Energia Espectral	A
--------	-------------------------	-------------------	---

	Real	Sintético	Real	Sintético	Rea
Normal	0,42	0,44	1,12	1,16	0,18
Desb.	0,86	0,89	2,45	2,51	0,74
Desal.	0,91	0,94	2,68	2,72	0,61
Rol. Def.	1,12	1,08	3,92	3,85	0,33

⚠ Esta tabela possui muitas colunas e foi cortada para impressão. Para visualizá-la completa, acesse o artigo original em:

<https://revistatopicos.com.br/artigos/neuroevolucao-aplicada-a-geracao-de-dados-sinteticos-para-diagnostico-de-falhas-em-maquinas-rotativas-em-cenarios-de-escassez-de-dados?noblockage>

Fonte: Autor.

Os resultados da Tabela 4 reforçam essa interpretação. As métricas de Wasserstein, Divergência KL e MMD apresentam valores reduzidos para todas as classes, indicando elevada proximidade estatística entre distribuições reais e sintéticas.

Tabela 4. Avaliação da similaridade intra-classes das amostras sintéticas.

Classe	Wasserstein	Divergência KL	MMD
Normal	0,21	0,08	0,04
Desb.	0,27	0,11	0,06
Desal.	0,29	0,13	0,07
Rol. Def.	0,34	0,18	0,09
Rol. Def. + Desb.	0,38	0,22	0,11
Rol. Def. + Desal.	0,41	0,25	0,12

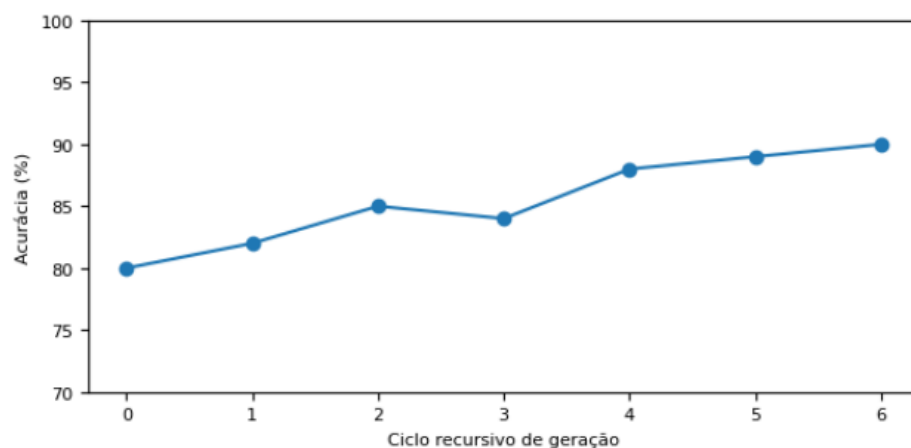
Desb. + Desal.	0,36	0,20	0,10
Rol. Def. + Desb. + Desal.	0,45	0,29	0,14

Fonte: Autor.

Esses resultados indicam que o Gerador não está apenas reproduzindo padrões estatísticos, mas também capturando relações físicas reais do sistema monitorado.

O principal objetivo da metodologia proposta consiste em utilizar geração sintética para aumentar a capacidade de generalização do Classificador. Os resultados obtidos demonstram que esse objetivo foi alcançado, Figura 7.

Figura 7. Evolução do desempenho do Classificador ao longo dos ciclos recursivos.



Fonte: Autor.

A evolução da acurácia ao longo dos ciclos evidencia crescimento consistente do desempenho durante as primeiras iterações, seguido por estabilização gradual.

Esse comportamento sugere que as primeiras amostras sintéticas adicionam informações relevantes ainda ausentes no conjunto de

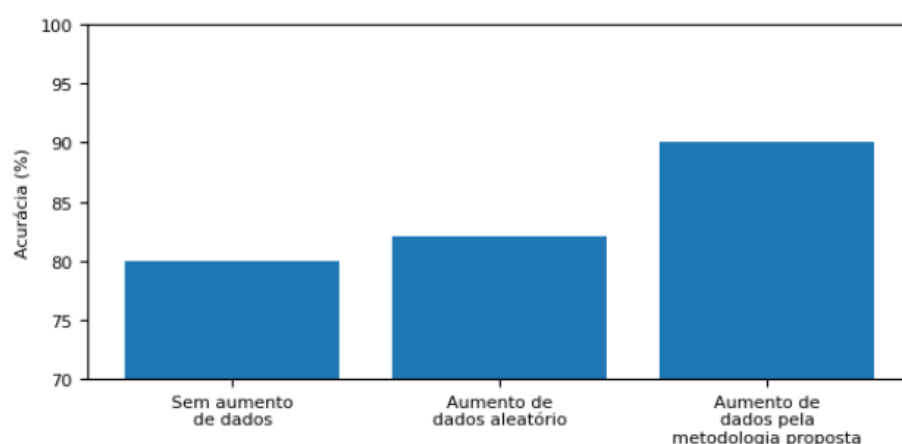
treinamento original. À medida que o espaço de características torna-se mais bem representado, os ganhos passam naturalmente a diminuir, indicando aproximação de um limite de desempenho para o problema estudado.

A comparação com técnicas convencionais de aumento de dados mostra que o simples acréscimo de perturbações aleatórias produz benefícios limitados. Em contraste, a metodologia proposta gera melhorias significativamente superiores e mais estáveis, como apresentado na Figura 8.

Esse resultado evidencia que não basta aumentar a quantidade de dados disponíveis. A qualidade e a representatividade das amostras geradas desempenham papel determinante na capacidade de generalização do modelo.

Os maiores ganhos foram observados justamente nas classes mais complexas. As falhas combinadas apresentaram aumentos expressivos em *recall* e precisão, reduzindo substancialmente as confusões observadas no Classificador inicial.

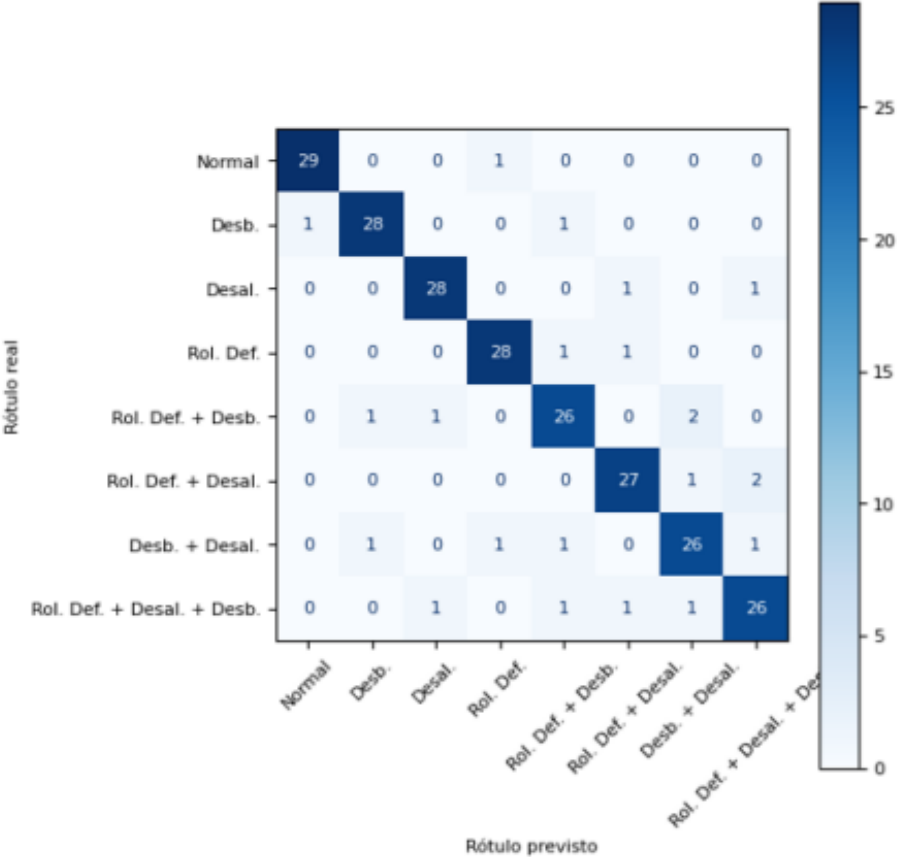
Figura 8. Comparação entre técnicas de aumento de dados.



Fonte: Autor.

A matriz de confusão final, Figura 9, confirma essa tendência, demonstrando melhora significativa na capacidade de separação entre classes com elevada sobreposição espectral.

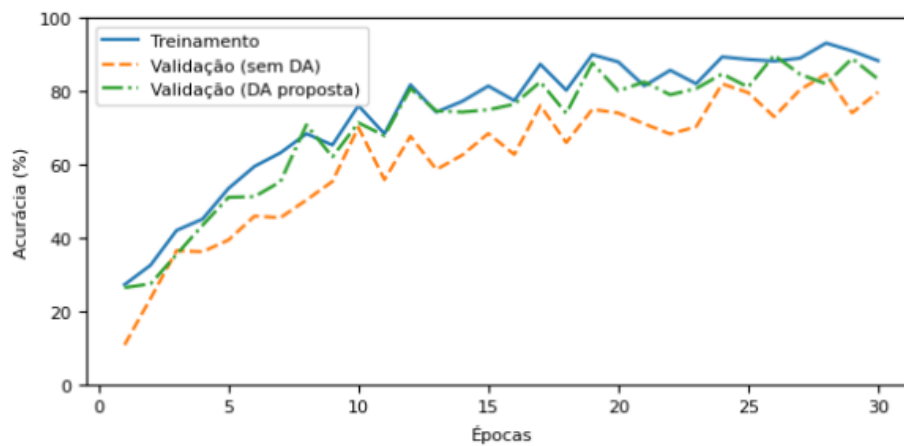
Figura 9. Matriz de confusão do Classificador.



Fonte: Autor.

Outro resultado importante refere-se à redução do sobreajuste. As curvas de treinamento tornam-se progressivamente mais estáveis, com menor divergência entre treinamento e validação, Figura 10. Isso indica que o aumento da diversidade amostral permitiu ao modelo aprender representações mais generalizáveis e menos dependentes das particularidades do conjunto original.

Figura 10. Treinamento do Classificador.



Fonte: Autor.

A análise do processo evolutivo demonstra que o Algoritmo Genético desempenhou papel fundamental na construção de modelos eficientes para o problema estudado. A Tabela 5 demonstra tal afirmação.

Tabela 5. Complexidade do Gerador a cada ciclo evolutivo.

Ciclo evolutivo	Dimensão do espaço latente	Número de parâmetros ($\times 10^3$)	Profundidade da rede
1	64	420	6
2	48	360	6
3	32	290	5
4	24	240	4
5	16	210	4

Fonte: Autor.

Os resultados indicam clara tendência de convergência para arquiteturas progressivamente mais compactas. A dimensionalidade do espaço latente reduziu-se de 64 para 16 dimensões ao longo dos

ciclos, enquanto o número de parâmetros caiu aproximadamente pela metade.

Esse comportamento sugere que a evolução não favoreceu simplesmente modelos maiores e mais complexos. Pelo contrário, o processo foi capaz de identificar representações mais eficientes, capazes de preservar informações relevantes utilizando menor quantidade de parâmetros.

A redução da profundidade das redes também indica que parte da complexidade inicial era redundante. À medida que o Classificador passou a receber conjuntos de treinamento mais ricos, tornou-se possível obter desempenho semelhante utilizando arquiteturas mais compactas e especializadas.

A recursividade introduz outro aspecto importante. Em vez de gerar dados sintéticos apenas uma vez, o sistema cria um ciclo de retroalimentação entre Gerador e Classificador. Cada nova geração de amostras influencia o treinamento do Classificador, que por sua vez passa a orientar futuras etapas de geração.

Esse mecanismo transforma o processo de geração em uma estratégia adaptativa. As amostras produzidas deixam de ser avaliadas apenas por critérios de reconstrução e passam a ser julgadas pelo impacto efetivo que exercem sobre o desempenho da tarefa final de diagnóstico.

Os resultados demonstram que essa coevolução favorece simultaneamente desempenho, estabilidade e eficiência computacional, produzindo modelos capazes de operar em cenários de escassez de dados sem crescimento excessivo da complexidade estrutural.

Essa característica constitui uma das principais contribuições da metodologia proposta, diferenciando-a de abordagens convencionais de aumento de dados baseadas exclusivamente em geração estatística ou perturbações aleatórias.

5. CONCLUSÃO

Este trabalho apresentou uma metodologia híbrida para diagnóstico de falhas em máquinas rotativas baseada na integração entre neuroevolução, geração sintética de dados e aprendizado supervisionado. A proposta foi desenvolvida para enfrentar um dos principais desafios da manutenção preditiva: a escassez de dados rotulados, frequentemente observada em aplicações industriais.

A metodologia emprega Algoritmos Genéticos para otimizar automaticamente arquiteturas de Redes Neurais e de *Variational Autoencoders*, permitindo a geração de amostras sintéticas representativas dos diferentes estados operacionais do sistema. Essas amostras são incorporadas ao processo de treinamento de forma controlada, ampliando o conjunto de dados disponível e contribuindo para o desenvolvimento de Classificadores mais robustos.

Os resultados obtidos demonstraram que a utilização de dados sintéticos gerados pelo modelo proposto proporcionou melhorias consistentes no desempenho do Classificador, refletidas no aumento da acurácia, precisão, *recall* e *F1-score*. Os ganhos foram particularmente relevantes para as classes associadas a falhas combinadas, que apresentam maior complexidade e menor representatividade no banco de dados original. Além disso, observou-se redução do sobreajuste e melhoria da capacidade de

generalização dos modelos, evidenciando a eficácia da estratégia proposta para cenários de dados limitados.

A análise do processo evolutivo também indicou a convergência para arquiteturas mais compactas e eficientes, capazes de manter elevado desempenho com menor complexidade estrutural. Esse resultado reforça o potencial da neuroevolução como ferramenta para automatizar o projeto de modelos de aprendizado profundo aplicados à engenharia.

Dessa forma, conclui-se que a integração entre *Variational Autoencoders*, Algoritmos Genéticos e Redes Neurais constitui uma abordagem promissora para o desenvolvimento de sistemas inteligentes de diagnóstico de falhas. A metodologia proposta mostrou-se capaz de ampliar artificialmente bases de dados reduzidas e aumentar a confiabilidade dos modelos preditivos, contribuindo para a evolução das técnicas de manutenção preditiva e monitoramento de condição.

Como trabalhos futuros, recomenda-se a investigação de estratégias de coevolução entre gerador e classificador, a inclusão de mecanismos para estimativa da severidade dos defeitos, a avaliação de falhas simultâneas mais complexas e a validação da metodologia em diferentes sistemas mecânicos e condições operacionais.

REFERÊNCIAS BIBLIOGRÁFICAS

GOLDBERG, D. E. Genetic Algorithms in Search, Optimization and Machine Learning. [S.l.]: Addison-Wesley, 1989. ISBN 9780201157673.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep Learning. MIT Press, 2016. ISBN 9780262035613.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. Neural Computation, MIT Press, v. 9, n. 8, p. 1735–1780, 11 1997. ISSN 0899-7667.

HOLLAND, J. H. Adaptation in Natural and Artificial Systems. Ann Arbor, USA: University of Michigan Press, 1975.

INMAN, D. J. Engineering Vibration. 4. ed. Upper Saddle River, NJ: Pearson, 2014. ISBN 9780273768449.

KINGMA, D. P.; WELLING, M. Auto-encoding variational bayes. International Conference on Learning Representations (ICLR), 2014.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. NeurIPS, 2012. v. 25, p. 1097–1105.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, IEEE, v. 86, n. 11, p. 2278–2324, 11 1998. ISSN 0018-9219.

LI, D.; LIU, F.; SHEN, T.; CHEN, L.; ZHAO, D. Data augmentation method for underwater acoustic target recognition based on underwater acoustic channel modeling and transfer learning. Applied Acoustics, Elsevier, v. 208, p. 109344, 2023. Conex. Ci. e Tecnol. Fortaleza/CE, v. 8, n. 1, p. 28-35, mar. 2014.

MUSZYNSKA, A. Rotordynamics. Boca Raton, FL, USA: CRC Press, 2005. ISBN 9780824723989.

RANDALL, R. B. Vibration-Based Condition Monitoring: Industrial, Aerospace and Automotive Applications. Chichester, United Kingdom: Wiley, 2011. ISBN 9780470747858.

RANDALL, R. B.; ANTONI, J. Rolling element bearing diagnostics - a tutorial. Mechanical Systems and Signal Processing, Elsevier, v. 25, n. 2, p. 485–520, 2 2011. ISSN 0888-3270.

¹ Discente do Doutorado em Engenharia Mecânica da Universidade de Campinas. E-mail: [acesse o artigo original para visualizar o e-mail](#)

² Docente do Doutorado em Engenharia Mecânica da Universidade de Campinas. E-mail: [acesse o artigo original para visualizar o e-mail](#)