Handling Noisy Labels in Industrial Scenarios: A Data-Centric Approach for Lubrication Fault Diagnosis

Bruno F. L. Pinto[1], Thiago C. Bittencourt[1], Victor Afonso Bauler[1], João Paulo Vieira[1], Karoline da Rocha[2], Diego Marques do Carmo[2], Danilo Silva[1]

[1]Universidade Federal de Santa Catarina (UFSC)

[2]Dynamox

bruno.forte@grad.ufsc.br,

{t.bittencourt, victor.bauler, j.p.vieira}@posgrad.ufsc.br

Abstract. *Industrial fault diagnosis frequently encounters inconsistencies in labels extracted from maintenance reports. To address this issue, this work investigates a data-centric pipeline for lubrication fault diagnosis in conveyor belt bearings using real vibration signals. The approach adapts to vibration data a noisy-label detection method previously proposed in the literature, using a convolutional neural network to rank suspicious labels before automatic correction and filtering. The best configuration raised the Random Forest AUROC from 87.8% to 91.3%, demonstrating that label refinement can improve lubrication fault diagnosis in industrial data.*

Resumo. *O diagnóstico de falhas industriais esbarra frequentemente na inconsistência de rótulos extraídos de laudos de manutenção. Para contornar este problema, este trabalho investiga uma pipeline centrada nos dados para o diagnóstico de falhas de lubrificação em mancais de transportadores de correia, usando sinais reais de vibração. A abordagem adapta ao domínio de vibração um método de detecção de rótulos ruidosos previamente proposto na literatura, usando uma rede neural convolucional para ranquear rótulos suspeitos antes da correção e filtragem automáticas. A melhor configuração elevou a AUROC da Random Forest de 87,8% para 91,3%, mostrando que o tratamento dos rótulos pode melhorar o diagnóstico de falhas de lubrificação em dados industriais.*

1. Introdução

A manutenção baseada em condição, ou *condition-based maintenance* (CBM), é uma estratégia importante em sistemas industriais contínuos, nos quais falhas inesperadas podem causar indisponibilidade de ativos, perdas de produção e aumento dos custos de manutenção [Jardine et al. 2006]. Nesses cenários, o monitoramento de condição permite acompanhar a evolução dos equipamentos e apoiar intervenções antes que defeitos se agravem. Entre os ativos críticos presentes em plantas industriais, destacam-se os transportadores de correia, que operam de forma contínua e dependem do bom funcionamento de componentes mecânicos sujeitos a desgaste, como os mancais dos tambores.

O monitoramento por vibração é uma das técnicas mais consolidadas para o diagnóstico de falhas em mancais e rolamentos, permitindo identificar alterações no comportamento dinâmico dos componentes a partir de sinais medidos durante a operação

[Randall and Antoni 2011]. No caso de falhas associadas à lubrificação, mudanças no regime de contato entre superfícies podem modificar a resposta vibratória do mancal, produzindo padrões relacionados ao aumento de atrito, impulsividade e energia em determinadas faixas de frequência. Por isso, sua detecção precoce é relevante para evitar a progressão do defeito e reduzir o risco de falhas mais severas.

Apesar da maior disponibilidade de sensores e sistemas de monitoramento em ambientes industriais, a construção de modelos de aprendizado de máquina ainda enfrenta uma limitação importante, que é a qualidade dos rótulos. Em aplicações reais, esses rótulos normalmente não são obtidos em condições controladas de laboratório, mas a partir de laudos técnicos, inspeções de rotina e registros textuais produzidos por diferentes analistas ao longo do tempo. Como consequência, o processo de rotulagem pode incorporar subjetividade, inconsistências temporais e diferenças de interpretação, dando origem ao problema de rótulos ruidosos [Northcutt et al. 2021], em que o diagnóstico registrado pode divergir da condição física do equipamento.

Nesse contexto, este artigo investiga o problema de diagnóstico de falhas de lubrificação em mancais de transportadores de correia, utilizando sinais reais de vibração coletados em ambiente industrial e laudos técnicos em texto livre elaborados manualmente por analistas parceiros. A partir desses laudos, os rótulos de falha foram extraídos automaticamente por meio de expressões regulares aplicadas ao texto livre. A abordagem utilizada segue uma perspectiva *data-centric* [Zha et al. 2025], na qual o foco não está no aumento da complexidade do modelo, mas na melhoria da qualidade do conjunto de treinamento. Assim, por meio da identificação, filtragem e correção automática de rótulos potencialmente errados, busca-se melhorar a confiabilidade do diagnóstico final, mesmo em um cenário com forte desbalanceamento entre as classes.

Com isso, este artigo apresenta três contribuições principais:

- Investigação e desenvolvimento de um modelo para diagnóstico de falhas de lubrificação em um cenário industrial real, com rótulos ruidosos extraídos de laudos de manutenção e forte desbalanceamento entre as classes;
- Aplicação e avaliação de uma abordagem de detecção de rótulos ruidosos baseada na dinâmica de treinamento de redes neurais para priorizar amostras potencialmente inconsistentes em sinais reais de vibração;
- Avaliação de estratégias automáticas de correção e filtragem de rótulos ruidosos, analisando seu impacto no diagnóstico de falhas de lubrificação.

2. Trabalhos relacionados

2.1. Diagnóstico de Falhas de Lubrificação

A literatura sobre diagnóstico de falhas de lubrificação em mancais e rolamentos ainda se apoia, em grande parte, em dados obtidos em bancadas de teste controladas [Duan et al. 2018], frequentemente utilizando sensores de Emissão Acústica de alta frequência que elevam o custo de implantação em sistemas contínuos [Elforjani 2018]. Esse cenário limita a avaliação de métodos em condições reais de operação, principalmente em ativos como transportadores de correia, para os quais há poucos *datasets* públicos voltados a esse modo de falha [Alharbi et al. 2023]. Embora avanços recentes demonstrem o sucesso de modelos clássicos no diagnóstico de outras falhas nesses ativos

a partir de sinais reais de vibração, a própria literatura aponta que a precisão desses modelos esbarra na presença de rótulos inconsistentes nos dados de campo [Bauler et al. 2024].

2.2. Detecção e Tratamento de Rótulos Ruidosos

A presença de rótulos errados é um problema recorrente em bases reais de aprendizado de máquina e tem motivado diferentes estratégias de aprendizado robusto a ruído de rótulos [Song et al. 2023]. Uma etapa fundamental consiste em identificar as amostras com maior probabilidade de inconsistência antes de qualquer intervenção no conjunto de dados. Para essa finalidade, adotamos a abordagem de melhor desempenho no benchmark de detecção considerado [Pickler et al. 2025], baseada na dinâmica de treinamento de uma rede neural, e avaliamos sua aplicação a sinais industriais de vibração.

Uma vez identificadas as amostras suspeitas, o tratamento pode ser realizado por correção, filtragem ou pela combinação das duas estratégias. A literatura sugere que exemplos com forte evidência de erro tendem a se beneficiar mais da correção, enquanto casos de incerteza intermediária podem ser tratados de forma mais segura por filtragem [Jiang et al. 2024]. Inspirado por essa discussão, este trabalho investiga uma estratégia automática de refinamento dos rótulos. Diferentemente de estudos baseados principalmente em ruído sintético, o ruído analisado aqui surge do próprio fluxo operacional da planta, a partir de inconsistências entre os rótulos obtidos por extração automática e o padrão vibratório observado nos equipamentos.

3. Metodologia

A metodologia utilizada avalia se a identificação e o tratamento automático de rótulos potencialmente errados podem melhorar o diagnóstico de falhas de lubrificação em dados industriais de vibração. Para isso, a pipeline é organizada em dois estágios principais, conforme ilustrado na Figura 1. No primeiro, uma rede neural convolucional é treinada com os rótulos originais para gerar um ranqueamento de amostras suspeitas de inconsistência. No segundo, esse ranqueamento é usado para orientar estratégias de correção e filtragem dos rótulos antes do treinamento e avaliação dos classificadores finais.

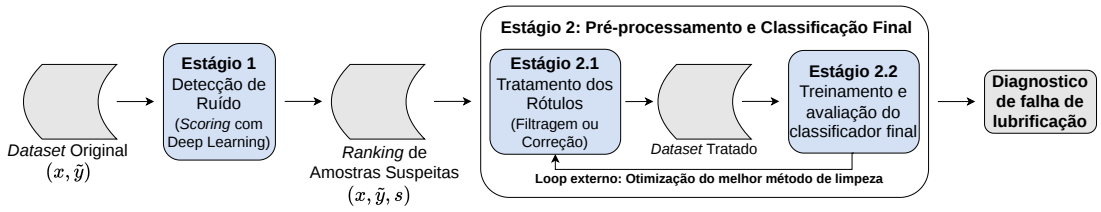


Figura 1. Arquitetura geral da pipeline utilizada, organizada em dois estágios principais: geração dos *scores* de suspeita com *deep learning* e refinamento dos rótulos seguido da classificação final com modelos clássicos.

3.1. *Dataset* e Revisão Manual

Os dados utilizados neste estudo foram obtidos a partir de mancais de tambores de correias transportadoras. O conjunto original reúne 2458 sinais de vibração triaxial, coletados em 300 máquinas distintas entre março de 2022 e outubro de 2025. Por se tratar de dados reais de operação, as medições apresentam elevada variabilidade, abrangendo taxas de

amostragem entre 5 e 26,3 kHz, rotações de 15 a 140 RPM e durações de aquisição de 0,6 a 12,5 segundos.

Os rótulos originais definem uma tarefa binária voltada especificamente à falha de lubrificação. Assim, a classe positiva indica a presença desse modo de falha nos laudos, enquanto a classe negativa indica apenas sua ausência, sem excluir a possibilidade de outros modos de falha. Como esperado em um cenário industrial, o conjunto é fortemente desbalanceado, com 94,51% das amostras na classe negativa e 5,49% na classe positiva. Além disso, a rotulagem é inerentemente ruidosa, pois os laudos estão sujeitos à subjetividade e a diferenças de interpretação entre os analistas da planta.

Para obter uma referência mais confiável, um subconjunto de 322 sinais foi revisado manualmente por um especialista em vibração. As amostras foram distribuídas entre as máquinas selecionadas e buscou-se aumentar a representação da classe minoritária, com seleção aleatória dentro desses critérios. Enquanto os laudos operacionais são registrados em texto livre e podem refletir interpretações ou recomendações preventivas, a revisão foi realizada com foco nas formas de onda brutas. Dessa forma, ela foi adotada como referência para avaliar tanto os scores de suspeita quanto os classificadores finais. Como mostra a Tabela 1, 91 amostras tiveram seus rótulos alterados, correspondendo a 28,26% de discordância.

		Rótulo revisado		
		Negativo	**Positivo**	**Total**
Rótulo original	**Negativo**	220	28	248
	Positivo	63	11	74
	Total	283	39	322

Tabela 1. Matriz de concordância entre os rótulos originais e revisados.

3.2. Divisão de Dados, Protocolo de Avaliação e Métricas

Para simular uma aplicação real e evitar vazamento de informação, adotou-se a divisão *machine-wise* [Vieira et al. 2026], na qual sinais de uma mesma máquina não podem aparecer simultaneamente nos conjuntos de treino e teste. Assim, quando uma máquina integra o subconjunto revisado de avaliação, seus demais sinais são excluídos do treinamento da classificação final.

O uso dos dados depende do objetivo de cada estágio. No Estágio 1, a rede neural é treinada exclusivamente com os rótulos originais. Os hiperparâmetros e a época de parada são selecionados em três divisões *machine-wise* de treino e validação, com aproximadamente 70% e 30% das amostras, respectivamente. Em seguida, a configuração selecionada é retreinada sobre todos os dados disponíveis para gerar os *scores* de suspeita. As anotações do especialista são usadas somente depois, para avaliar se o ranking prioriza rótulos de fato inconsistentes.

No Estágio 2, os *scores* orientam a correção ou filtragem dos rótulos de treino. A avaliação final ocorre no subconjunto de 322 sinais revisados manualmente, enquanto o treinamento utiliza as 1313 amostras restantes, descartando-se 823 sinais para garantir a separação *machine-wise*. Por fim, um classificador clássico baseado em *features* é treinado com os dados originais ou refinados. O desempenho é medido pela *Area Un-*

der the Receiver Operating Characteristic (AUROC), e a diferença entre as abordagens é avaliada pelo teste pareado de DeLong [DeLong et al. 1988].

Conjunto	Total	Negativo	Positivo	Uso na metodologia
Base completa	2458	2323 (94,51%)	135 (5,49%)	Estágio 1
Base de treino	1313	1282 (97,64%)	31 (2,36%)	Estágio 2
Subconjunto revisado	322	283 (87,89%)	39 (12,11%)	Avaliação dos *scores* no Estágio 1 e do classificador no Estágio 2
Demais sinais	823	793 (96,35%)	30 (3,65%)	Não utilizados

Tabela 2. Conjuntos de dados utilizados na metodologia. As contagens e porcentagens correspondem aos rótulos originais, exceto no subconjunto revisado, que apresenta a distribuição após a revisão manual. As contagens originais desse subconjunto, necessárias para a decomposição da base completa, são apresentadas na Tabela 1.

3.3. Estágio 1: Geração dos *scores* de suspeita

O primeiro estágio tem como objetivo atribuir um *score* de suspeita a cada amostra do conjunto de dados. Esse *score* indica o quanto o rótulo parece discordar do comportamento observado no sinal de vibração. Assim, a rede neural não é usada como classificador final, mas como uma ferramenta para ranquear os rótulos mais prováveis de conter erro.

Formalmente, considera-se o conjunto $\tilde{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$, em que x_i representa o sinal de vibração da amostra i e $\tilde{y}_i \in \{0, 1\}$ é o rótulo observado no laudo. Como o rótulo verdadeiro y_i não é conhecido para todo o conjunto, assume-se que há ruído quando o rótulo observado diverge da condição real do equipamento, isto é, quando $\tilde{y}_i \neq y_i$.

Este estágio é dividido em duas etapas. Na primeira, a rede neural é otimizada em divisões de treino e validação, buscando uma configuração que aprenda padrões úteis sem memorizar os erros dos rótulos. Na segunda, a rede é treinada novamente com essa configuração, e suas saídas ao longo das épocas são armazenadas para calcular o *score* de suspeita. Ao final, cada amostra passa a ser representada por $(x_i, \tilde{y}_i, s_i)$, em que s_i é o *score* usado no pré-processamento do Estágio 2.

3.3.1. Arquitetura e Pré-processamento

Para gerar o ranqueamento de suspeita dos rótulos, adotou-se a *Wide Deep Convolutional Neural Network* (WDCNN) como modelo base. A escolha justifica-se pelo seu amplo uso no processamento de sinais de vibração industrial e por apresentar bons resultados no diagnóstico de falhas em cenários com ruído [Zhang et al. 2017]. Antes do treinamento, os sinais passaram por uma etapa de pré-processamento na qual a média de cada amostra foi subtraída, a fim de diminuir os componentes de *offset* do sensor, e todos os sinais foram reamostrados para uma frequência uniforme de 12 kHz. Em seguida, sinais com duração inferior a 1 segundo foram descartados. Durante o treinamento, aplica-se um recorte aleatório de 1 segundo (*random crop*), enquanto na validação os sinais são segmentados em janelas consecutivas fixas de mesma dimensão.

3.3.2. Otimização da rede neural

A otimização da rede buscou selecionar a configuração de treinamento e a época de parada usadas na geração dos *scores* de suspeita. Essa etapa é necessária porque, se o treinamento for prolongado demais, a rede pode começar a memorizar rótulos incorretos em vez de capturar apenas os padrões mais consistentes dos sinais. O espaço de busca e a configuração selecionada são apresentados na Tabela 3.

Hiperparâmetro	**Espaço de Busca**	**Configuração Selecionada**
Representação de entrada	Tempo, FFT, Envelope temporal, Envelope espectral	FFT
Otimizador	Adam, AdamW	AdamW [Loshchilov and Hutter 2019]
Learning rate (*lr*)	10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}	10^{-4}
Batch size	16, 32, 64, 128, 256	128
Weight decay	0, 10^{-2}, $5 \cdot 10^{-2}$	10^{-2}
Normalização (Z-score)	Sem, Por dataset, Por waveform	Sem normalização
Filtro passa-alta (*cutoff*)	0, 1000, 1500, 2000 Hz	Sem filtro (0 Hz)
Random gain (*Data Aug.*)	0; 0,1; 0,3; 0,5	Desativado (0)
Época de parada	Curva de validação	10

Tabela 3. Espaço de busca avaliado e configuração final de hiperparâmetros selecionada após a etapa de otimização.

3.3.3. Criação dos *scores* (*Mean Probabilities + Logit Margin*)

Na segunda fase, a rede WDCNN é treinada novamente com a configuração selecionada na etapa de otimização. Para essa execução, utiliza-se o todo o conjunto disponível para a detecção de rótulos ruidosos, e o treinamento é conduzido até o ponto de parada definido anteriormente. Ao longo das épocas, os *logits* gerados para cada amostra são armazenados, permitindo calcular posteriormente um *score* de suspeita de ruído baseado na dinâmica de aprendizagem do modelo.

Seja $z_i^{(t)} \in \mathbb{R}^K$ o vetor de *logits* produzido para a amostra i na época t, em que $K = 2$ corresponde às duas classes do problema: condição sem e com falha de lubrificação. Seguindo a formulação de [Pickler et al. 2025], aplica-se inicialmente uma agregação temporal das probabilidades preditas pelo modelo, denominada *Temporal Ensemble*. Para isso, os *logits* de cada época são convertidos em probabilidades por meio da função *softmax* e, em seguida, calcula-se a média dessas probabilidades ao longo das E épocas consideradas:

$$\bar{p}_i = \frac{1}{E} \sum_{t=1}^{E} \text{softmax}(z_i^{(t)}). \tag{1}$$

Como a métrica *Logit Margin* opera em escala de *logits*, aplica-se a função logarítmica componente a componente para mapear as probabilidades agregadas de volta ao domínio dos números reais, gerando o vetor transformado $\bar{z}_i = \log \bar{p}_i$. A partir de $\bar{z}_i$, o *score* final de suspeita L_{LM} é definido diretamente pela margem entre a maior ativação atribuída a

uma classe diferente do rótulo observado e a ativação correspondente ao próprio rótulo observado $\tilde{y}_i$:

$$L_{\text{LM}}(\bar{z}_i, \tilde{y}_i) = \max_{k \neq \tilde{y}_i} \bar{z}_{i,k} - \bar{z}_{i,\tilde{y}_i}. \tag{2}$$

Valores mais altos de L_{LM} indicam maior discordância entre o rótulo e o comportamento preditivo médio da rede ao longo do treinamento. Assim, as amostras com maiores *scores* são interpretadas como mais suspeitas de conter ruído de rotulagem.

3.4. Estágio 2: Pré-processamento dos rótulos e classificação final

No segundo estágio, os *scores* gerados pela rede neural são usados para ajustar o conjunto de treinamento antes da classificação final. A partir da representação $(x_i, \tilde{y}_i, s_i)$ obtida no Estágio 1, o objetivo desta etapa é usar s_i para decidir quais amostras devem ser mantidas, removidas ou ter o rótulo corrigido antes do treinamento do modelo.

3.4.1. Pré-processamento baseado nos *scores*

No pré-processamento, as amostras do conjunto de treino são ordenadas de forma decrescente, de modo que as primeiras posições do ranking correspondem aos rótulos considerados mais suspeitos. Em seguida, aplicam-se as frações de intervenção δ_{corr} e δ_{filt}, que definem as proporções de amostras a serem corrigidas ou removidas, respectivamente. Esses parâmetros não representam valores absolutos do *score*, mas proporções do conjunto ordenado, calculadas sobre todo o conjunto de treino quando todas as classes são consideradas ou apenas sobre as amostras originalmente negativas quando o refinamento é restrito à classe majoritária.

Foram avaliadas quatro estratégias de pré-processamento:

- ***Baseline*:** mantém o conjunto de treino sem alterações;
- **Filtragem:** remove as amostras com maiores *scores*, dentro da fração definida por δ_{filt};
- **Correção:** inverte os rótulos das amostras com maiores *scores*, dentro da fração definida por δ_{corr};
- **Hibrido:** aplica as duas operações em sequência. Primeiro, os rótulos das amostras mais suspeitas são invertidos, de acordo com δ_{corr}. Em seguida, as próximas amostras do ranking são removidas, conforme δ_{filt}.

Assim, na estratégia híbrida, as duas frações atuam sobre partes diferentes do ranking do conjunto-alvo. Por exemplo, se $\delta_{corr} = 15\%$ e $\delta_{filt} = 25\%$, os primeiros 15% têm seus rótulos invertidos, os 25% seguintes são descartados e o restante do conjunto permanece inalterado.

Como o conjunto é bastante desbalanceado, aplicar essa regra em todas as classes poderia afetar muitos exemplos positivos, que são minoria. Por isso, as estratégias também foram avaliadas restringindo o refinamento às amostras originalmente negativas. Dessa forma, os exemplos positivos são preservados e a intervenção é concentrada na classe majoritária, permitindo avaliar se uma limpeza mais direcionada é mais adequada ao problema.

3.4.2. Classificação Final

Após o pré-processamento dos rótulos, o diagnóstico é feito com classificadores baseados em extração de *features*. Essa escolha permite avaliar se a limpeza do conjunto de treino melhora a classificação sem depender de uma arquitetura específica. Foram considerados modelos de menor custo computacional e maior interpretabilidade, como Regressão Logística, *Support Vector Machine* (SVM), Random Forest e Gradient Boosting.

Os classificadores usam uma representação tabular dos sinais de vibração, construída a partir de *features* nos domínios de aceleração e envelope, comuns no diagnóstico de falhas em mancais [Randall and Antoni 2011]. As métricas da Tabela 4 foram escolhidas por resumirem variações de energia e impulsividade, como RMS, curtose, fator de crista e RMS carpete [Sikora 2024], que podem estar associadas à degradação da lubrificação.

Features	**Domínio**	**Bandas disponíveis (Hz)**
RMS	Aceleração	1000 a 3000, 3000 a 6395, 6400 a 13000
RMS Carpete	Aceleração e Envelope	1000 a 6395 , 6400 a 13000
Fator de Crista	Aceleração	1000 a 6395 , 3000 a 13000
Curtose	Aceleração	1000 a 6395 , 3000 a 13000

Tabela 4. *Features* extraídas para os modelos clássicos.

A busca pelos melhores resultados foi organizada em dois níveis, conforme resumido na Figura 1. No nível externo, são testadas diferentes combinações de δ_{corr} e δ_{filt}. Para cada combinação, o conjunto de treino é primeiro ajustado com base nos *scores* de suspeita, gerando uma nova versão refinada dos dados. No nível interno, os hiperparâmetros do classificador são otimizados por validação cruzada no conjunto de treino já refinado. Após essa escolha, o modelo é retreinado com todo o conjunto de treino refinado e avaliado no subconjunto revisado manualmente.

Como o número de laudos revisados por especialista é limitado, não foi possível reservar um conjunto separado apenas para escolher os valores de δ_{corr} e δ_{filt}. Por isso, a seleção desses limiares usa o mesmo subconjunto revisado empregado na avaliação final. Assim, os resultados associados à melhor combinação de limiares devem ser interpretados como uma estimativa otimista do ganho possível com o refinamento dos rótulos, e não como uma medida totalmente independente de generalização.

4. Resultados e Discussão

Os resultados são organizados em dois estágios experimentais principais. Primeiro, avaliamos a capacidade do modelo de rede neural de atribuir *scores* de suspeita aos rótulos e, a partir deles, priorizar amostras com maior probabilidade de inconsistência em relação à revisão manual. Em seguida, avaliamos se o uso desses *scores* para refinamento do conjunto de treinamento, por meio de correção e/ou filtragem de rótulos, melhora o desempenho da tarefa final de diagnóstico de falha de lubrificação com modelos clássicos de aprendizado de máquina. Como parte do protocolo experimental, todos os experimentos foram executados com cinco sementes aleatórias distintas, e as métricas quantitativas são reportadas em termos de média e desvio padrão, de modo a representar a variabilidade associada ao treinamento e à divisão dos dados.

4.1. Detecção e Triagem de Rótulos Ruidosos

A capacidade do primeiro estágio de identificar inconsistências nos rótulos foi avaliada no subconjunto revisado, descrito na Seção 3.1. Como mostra a Figura 2, as amostras cujos rótulos foram alterados pelo especialista tendem a aparecer nas primeiras posições do ranking de suspeita. Esse resultado indica que o ranqueamento consegue priorizar rótulos com maior chance de inconsistência, mesmo em um cenário no qual o ruído de rotulagem não é introduzido de forma controlada, mas surge do processo de anotação.

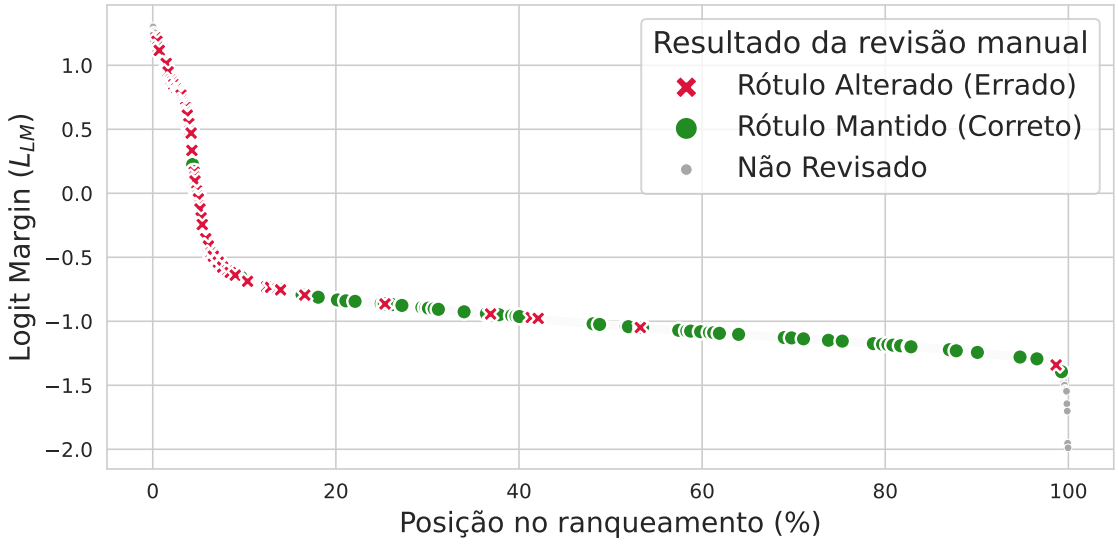


Figura 2. Distribuição dos *scores* de suspeita das amostras, ordenadas de forma decrescente por L_{LM}.

4.2. Baseline de Classificação Final

Antes de aplicar as técnicas de refinamento, estabelecemos uma referência de desempenho treinando os classificadores tradicionais diretamente com os rótulos originais. Os resultados obtidos no subconjunto revisado manualmente constam na Tabela 5.

Modelo	**AUROC**	**Tempo de Otimização (HP)**
SVM	$83{,}27\% \pm 0{,}0\%$	300 s
Regressão Logística	$80{,}35\% \pm 0{,}0\%$	34 s
Random Forest (Yggdrasil)	$\mathbf{87{,}82\% \pm 0{,}8\%}$	87 s
XGBoost	$86{,}17\% \pm 0{,}0\%$	149 s

Tabela 5. Desempenho baseline dos modelos clássicos treinados com os rótulos originais e avaliados no subconjunto revisado manualmente.

Os resultados da Tabela 5 mostram que todos os classificadores ficaram acima do acaso, indicando que os modelos aprenderam relações úteis entre as *features* de vibração mesmo com rótulos errados. A Random Forest apresentou a maior AUROC, de 87,82%, mantendo um tempo de otimização moderado em relação aos demais modelos. Desse modo, por combinar melhor desempenho e custo computacional adequado, ela foi adotada como referência na análise da Seção 4.3, cuja configuração selecionada é posteriormente avaliada nos demais classificadores.

4.3. Impacto do Refinamento dos Rótulos no Diagnóstico

A busca pela melhor combinação de tratamento de rótulos mostrou que o forte desbalanceamento do dataset exige uma estratégia mais direcionada. O mapeamento sistemático indicou que aplicar as regras de limpeza de forma indistinta em todas as classes pode degradar o desempenho, principalmente pela remoção de exemplos da classe minoritária. A melhor configuração identificada nessa varredura é resumida na Tabela 6.

Estratégia	**Alvo da Correção**	**Alvo da Filtragem**	δ_{corr}	δ_{filt}	**AUROC**
Filtragem	–	Todas as classes	–	0%	$87{,}8\% \pm 0{,}8\%$
Filtragem	–	Apenas negativas	–	20%	$90{,}1\% \pm 0{,}2\%$
Correção	Todas as classes	–	0%	–	$87{,}8\% \pm 0{,}8\%$
Correção	Apenas negativas	–	4%	–	$90{,}7\% \pm 1{,}2\%$
Híbrida	Todas as classes	Apenas negativas	0%	20%	$90{,}1\% \pm 0,2\%$
Híbrida	Apenas negativas	Apenas negativas	4%	12%	$\mathbf{91{,}3\% \pm 0{,}9\%}$

Tabela 6. Comparação de desempenho das configurações ótimas para as diferentes estratégias de refinamento dos rótulos.

Para detalhar a dinâmica interna da estratégia selecionada, a Figura 3 apresenta a superfície de desempenho do modelo à medida que variamos os limiares de correção (δ_{corr}) e filtragem (δ_{filt}) sobre a classe majoritária. O mapa térmico reforça a tendência observada na Tabela 6, indicando que uma abordagem híbrida tende a produzir melhores resultados do que o uso isolado de apenas uma das etapas. Além disso, a região de maior desempenho indica que a configuração escolhida não depende de um único ponto específico da grade, mas de uma faixa relativamente consistente de limiares.

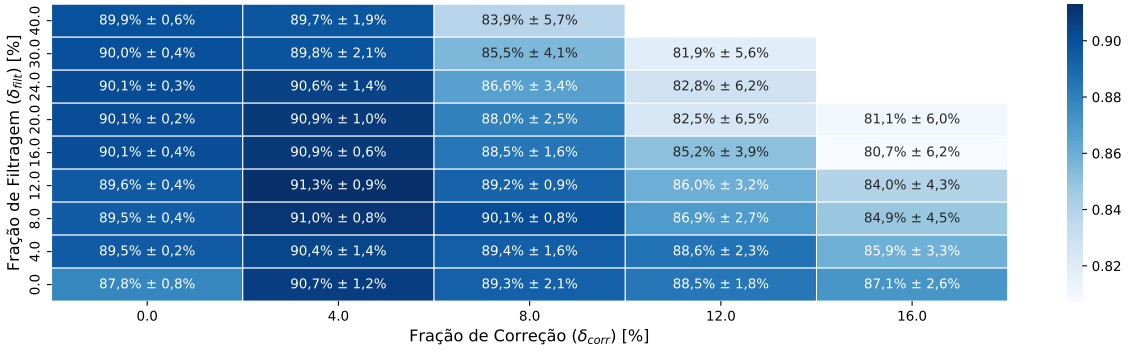


Figura 3. AUROC da RF para diferentes combinações de δ_{corr} e δ_{filt} aplicadas exclusivamente sobre as amostras inicialmente rotuladas como negativas.

Adicionalmente, o teste de DeLong [DeLong et al. 1988] confirmou a significância estatística do ganho médio de 3,48% da estratégia híbrida sobre a Random Forest original ($p = 0,044$; IC95% 0,08%–6,56%), corroborando o ganho observado da abordagem.

Por fim, a configuração selecionada foi testada nos demais classificadores para verificar se o ganho se estendia a outros modelos (Tabela 7). O objetivo é avaliar se a limpeza automática melhora a separabilidade da tarefa e reduz a dependência do algoritmo

escolhido. Com isso, o tratamento deixa de ser um simples pré-processamento para atuar como uma intervenção *data-centric* que impacta diretamente o diagnóstico.

Modelo	Rótulos originais	Rótulos refinados
SVM	$\mathbf{83{,}27\% \pm 0{,}0\%}$	$83{,}07\% \pm 3{,}7\%$
Regressão Logística	$80{,}35\% \pm 0{,}0\%$	$\mathbf{89{,}62\% \pm 0{,}4\%}$
Random Forest (Yggdrasil)	$87{,}82\% \pm 0{,}8\%$	$\mathbf{91{,}3\% \pm 0{,}9\%}$
XGBoost	$86{,}17\% \pm 0{,}0\%$	$\mathbf{88{,}67\% \pm 2{,}27\%}$

Tabela 7. Comparação entre o baseline treinado com rótulos originais e o desempenho após o refinamento automático dos rótulos.

Os resultados mostram um efeito predominantemente positivo, ainda que variável. Embora o SVM tenha-se mantido praticamente estável, com uma leve redução nominal, a Regressão Logística, a Random Forest e o XGBoost apresentaram ganhos de desempenho significativos. Isso confirma que os benefícios do refinamento não se limitam ao classificador de referência, validando a utilidade do método para além de um único modelo, ainda que o impacto dependa da arquitetura utilizada.

5. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma pipeline *data-centric* para lidar com rótulos errados no diagnóstico de falhas de lubrificação em dados industriais de vibração. Em vez de focar apenas na arquitetura do classificador, a abordagem buscou melhorar o conjunto de treinamento por meio da detecção e do refinamento automático de rótulos.

A revisão manual identificou 28,26% de discordância entre os rótulos originais e os revisados no subconjunto avaliado. Além disso, o refinamento direcionado elevou a AUROC da Random Forest de 87,8% para 91,3%, indicando que a melhoria da qualidade dos rótulos pode beneficiar o diagnóstico mesmo com classificadores clássicos e *features* especializadas.

Algumas limitações devem ser consideradas. O subconjunto revisado manualmente é limitado em tamanho e também foi utilizado na seleção dos limiares de refinamento, de modo que o melhor resultado representa uma estimativa otimista do ganho potencial. Além disso, as correções assumem que grandes discordâncias entre a rede neural e os rótulos indicam erros de anotação, o que pode não ocorrer em todos os casos.

Como trabalhos futuros, pretende-se ampliar o conjunto de sinais revisados por especialistas, avaliar a abordagem em outros modos de falha e comparar sistematicamente diferentes métodos de detecção e tratamento de rótulos ruidosos. Esses avanços podem contribuir para uma metodologia mais sólida de tratamento de rótulos industriais inconsistentes, aumentando a confiança nas métricas reportadas e nos modelos treinados a partir desses dados.

Referências

Alharbi, F., Luo, S., Zhang, H., Shaukat, K., Yang, G., Wheeler, C. A., and Chen, Z. (2023). A brief review of acoustic and vibration signal-based fault detection for belt conveyor idlers using machine learning models. *Sensors*, 23(4).

Bauler, V., Cordioli, J., Silva, D., and Braga, D. (2024). Bearing fault detection in conveyor belt drums using machine learning. *Annual Conference of the PHM Society*, 16.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Duan, J., Xia, Z., Xu, G., Dan, Z., Zhang, S., and Liang, L. (2018). Lubrication condition monitoring and evaluation of rolling bearing based on acoustic emission. In *2018 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 1–7.

Elforjani, M. (2018). Diagnosis and prognosis of slow speed bearing behavior under grease starvation condition. *Structural Health Monitoring*, 17(3):532–548.

Jardine, A. K. S., Lin, D., and Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7):1483–1510.

Jiang, G., Zhang, J., Bai, X., Wang, W., and Meng, D. (2024). Which is more effective in label noise cleaning, correction or filtering? *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:12866–12873.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Northcutt, C. G., Jiang, L., and Chuang, I. L. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

Pickler, H., Kamassury, J. K., and Silva, D. (2025). Benchmarking noisy label detection methods. *arXiv preprint arXiv:2510.16211*.

Randall, R. B. and Antoni, J. (2011). Rolling element bearing diagnostics—a tutorial. *Mechanical Systems and Signal Processing*, 25(2):485–520.

Sikora, M. V. (2024). Comparison of machine learning binary classifiers for detection of gear defects. Master's thesis, Universidade Federal de Santa Catarina, Florianópolis, Brazil.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. (2023). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34:8135–8153.

Vieira, J. P., Bauler, V. A., Rosa, R. K., and Silva, D. (2026). Towards a more realistic evaluation of machine learning models for bearing fault diagnosis. *Mechanical Systems and Signal Processing*, 258:114640.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., and Hu, X. (2025). Data-centric artificial intelligence: A survey. *ACM Comput. Surv.*, 57(5).

Zhang, W., Peng, G., Li, C., Chen, Y., and Zhang, Z. (2017). A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors*, 17:425.