57º SBPO
SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
Gramado • RS
DE 5 A 9 DE OUTUBRO DE 2025

Detecção de Anomalias em Poços de Petróleo com LSTM: Uma Comparação entre Classificação Plana e Hierárquica

Bernardo da Silva Puppim
Universidade Federal Fluminense
Rua Recife, s/n, Lotes 1-7, Jardim Bela Vista, Rio das Ostras – RJ, 28895-532
bernardo_puppim@id.uff.br

Dalto Garcia Borges de Souza, Iara Tammela
Universidade Federal Fluminense
Rua Recife, s/n, Lotes 1-7, Jardim Bela Vista, Rio das Ostras – RJ, 28895-532
daltonborges@id.uff.br, iaratammela@id.uff.br

RESUMO

A detecção automatizada de anomalias em poços de petróleo surgentes representa um desafio crítico para a operação segura e eficiente em ambientes offshore. Este estudo investiga o impacto da estruturação hierárquica das classes na tarefa de classificação de anomalias com base em séries temporais multivariadas extraídas do dataset 3W. Foram comparadas abordagens plana e hierárquica utilizando modelos LSTM sob as mesmas condições experimentais. Os resultados mostram que a classificação hierárquica supera consistentemente a abordagem plana em métricas de desempenho, especialmente na detecção de eventos raros ou ambíguos. Além disso, a estrutura hierárquica favorece maior coerência semântica nas previsões e permite inferências parciais úteis à tomada de decisão operacional. Esses achados reforçam o potencial da classificação hierárquica como alternativa robusta e interpretável para sistemas inteligentes de monitoramento em tempo real no setor de óleo e gás.

PALAVRAS CHAVE. Classificação Hierárquica, LSTM, Anomalias em Poços de Petróleo.

Tópicos (EM&OG – PO na Área de Energia, Óleo e Gás, IA – PO e IA)

ABSTRACT

The automated detection of anomalies in naturally flowing oil wells is a critical challenge for safe and efficient offshore operations. This study investigates the impact of hierarchical class structuring on the task of anomaly classification using multivariate time series data from the 3W dataset. Flat and hierarchical classification approaches were compared using LSTM models under identical experimental conditions. Results show that the hierarchical approach consistently outperforms the flat model across all performance metrics, particularly in detecting rare or ambiguous events. Furthermore, the hierarchical structure improves the semantic consistency of predictions and enables partial inferences that support operational decision-making. These findings highlight the potential of hierarchical classification as a robust and interpretable solution for real-time intelligent monitoring systems in the oil and gas industry.

KEYWORDS. Hierarchical Classification. LSTM. Oil Well Anomalies.

Paper topics (EM&OG – OR in the Energy, Oil & Gas Sector, AI – OR and AI)

1. Introdução

A crescente complexidade das operações offshore e a necessidade de maior confiabilidade nos sistemas de produção de petróleo têm impulsionado o desenvolvimento de soluções baseadas em inteligência artificial (IA) para monitoramento e diagnóstico de falhas. A detecção precoce de anomalias em poços surgentes é essencial para mitigar riscos operacionais, evitar perdas econômicas e reduzir impactos ambientais (VARGAS, 2019).

Entre as abordagens de aprendizado de máquina aplicadas a esse domínio, os modelos de séries temporais multivariadas têm se destacado por sua capacidade de capturar o comportamento dinâmico de variáveis operacionais. No contexto da elevação natural de petróleo, o trabalho de Vargas (2019) apresentou o *dataset 3W*, um conjunto de dados realista e desafiador com instâncias reais, simuladas e desenhadas por especialistas, destinado ao treinamento e avaliação de modelos de prognóstico de anomalias. As anomalias contempladas incluem fenômenos como aumento abrupto de BSW, fechamento espúrio de válvula de segurança, instabilidade de fluxo, entre outras.

Nos últimos anos, modelos baseados em redes neurais recorrentes, especialmente os do tipo Long Short-Term Memory (LSTM), têm demonstrado desempenho superior na modelagem de séries temporais com dependências de longo prazo. Estudos recentes, como o de Bayazitova et al. (2024), aplicaram LSTM e GRU combinados com algoritmos genéticos para otimizar a detecção de falhas em poços offshore, alcançando resultados expressivos.

Apesar dos avanços, a maioria dos estudos ainda adota uma abordagem de classificação plana, em que cada instância é atribuída a uma classe de anomalia de forma mutuamente exclusiva, ignorando relações estruturais entre os rótulos. A classificação hierárquica, por sua vez, considera a organização das classes em níveis, podendo representar melhor a progressão e semelhança entre tipos de anomalias, além de permitir uma classificação mais interpretável e robusta a erros.

Neste trabalho, propõe-se uma comparação entre as abordagens de classificação plana e hierárquica aplicadas a séries temporais multivariadas no contexto do *dataset 3W*. Ambas as abordagens são implementadas com modelos LSTM, e são avaliadas quanto à sua acurácia, robustez e coerência com o processo decisório em ambiente industrial. O objetivo é demonstrar os ganhos potenciais da estrutura hierárquica na tarefa de detecção de anomalias em poços de petróleo offshore.

2. Fundamentação Teórica

A detecção de anomalias em processos industriais é um campo consolidado da pesquisa operacional e da inteligência computacional, especialmente em contextos críticos como a produção de petróleo offshore. Nesta seção, são apresentados os principais conceitos e métodos que embasam este estudo: séries temporais multivariadas, classificação de anomalias (plana e hierárquica) e o uso de redes LSTM na modelagem desses dados.

2.1 Séries Temporais Multivariadas em Poços de Petróleo

Poços de petróleo naturalmente surgentes, especialmente em ambiente offshore, possuem sistemas de instrumentação que permitem o monitoramento contínuo de variáveis como pressão, temperatura e vazão. Esses sinais, quando coletados ao longo do tempo, constituem séries temporais multivariadas, caracterizadas pela dependência temporal entre múltiplas variáveis correlacionadas (VARGAS, 2019).

O *dataset 3W*, desenvolvido por Vargas (2019), representa um marco na área ao oferecer um conjunto público, realista e desafiador com medições reais, simuladas e desenhadas por especialistas. Ele contém nove classes de anomalias, além do estado normal, incluindo eventos como instabilidade de fluxo, *slugging* severo, hidrato na linha de produção e perda rápida de produtividade. Esses dados são especialmente relevantes por refletirem o comportamento típico dos sistemas de elevação natural de petróleo, nos quais a instrumentação é limitada e os eventos anômalos são raros e de difícil rotulagem.

2.2 Classificação Plana vs. Classificação Hierárquica

A classificação comum, também chamada de classificação plana, é o paradigma tradicional em tarefas de aprendizado supervisionado, no qual cada classe é tratada como independente e mutuamente exclusiva. Nesse modelo, o classificador é treinado para associar diretamente uma instância a um único rótulo, ignorando qualquer relação estrutural ou semântica entre as classes (SILLA JUNIOR; FREITAS, 2011). Essa abordagem é amplamente utilizada devido à sua simplicidade e pela ampla disponibilidade de algoritmos e métricas compatíveis. Diversos estudos recentes aplicaram essa estrutura ao *3W Dataset* da Petrobras, utilizando redes neurais profundas, como LSTM e GRU, para classificar anomalias em poços naturalmente fluentes (BAYAZITOVA et al., 2024; GATTA et al., 2022).

Entretanto, em muitos domínios, especialmente naqueles com natureza ontológica, como biologia, medicina, taxonomia e engenharia industrial, a relação entre as classes é intrinsecamente hierárquica. Nesse contexto, a relação hierárquica entre classes geralmente segue o padrão "IS-A", no qual uma classe filha é um tipo mais específico de sua classe ancestral (por exemplo, “hidrato” é um tipo de “falha química”). A abordagem plana pode levar a erros semânticos significativos, como classificar uma instância em uma classe altamente específica sem reconhecer sua associação com categorias mais amplas e generalistas (SILLA JUNIOR; FREITAS, 2011).

A classificação hierárquica, por sua vez, trata as classes como elementos inter-relacionados em uma estrutura em árvore ou grafo acíclico direcionado (DAG). Esse tipo de modelagem permite que o sistema de classificação leve em consideração níveis de generalidade, relações de herança e coerência semântica.

Esse tipo de abordagem é especialmente útil em ambientes operacionais complexos, como o monitoramento de poços offshore, onde diferentes falhas podem apresentar sintomas semelhantes ou estar relacionadas a mecanismos físicos correlatos (VARGAS et al., 2019). A classificação hierárquica permite que o modelo seja mais tolerante a ruídos, facilitando uma escalabilidade de decisões operacionais: mesmo quando a classe específica não é identificada corretamente, é possível que o modelo acerte ao menos a categoria superior, o que ainda oferece valor operacional significativo para ações corretivas iniciais (SILLA JUNIOR; FREITAS, 2011; VARGAS et al., 2019).

2.3 LSTM na Modelagem de Séries Temporais

Redes neurais do tipo Long Short-Term Memory (LSTM) são uma variação das redes recorrentes (RNNs), projetadas para lidar com dependências de longo prazo em séries temporais. Seu uso tem sido amplamente documentado em aplicações industriais, inclusive na detecção de anomalias em produção de petróleo (BAYAZITOVA et al., 2024). As LSTM utilizam mecanismos de portas (gates) para controlar o fluxo de informações, evitando o problema de desvanecimento do gradiente típico das RNNs clássicas.

No estudo de Bayazitova et al. (2024), a combinação de LSTM com algoritmos genéticos para ajuste de hiperparâmetros demonstrou resultados superiores aos obtidos por abordagens tradicionais, com *F1-score* de até 0,97 na detecção multiclasse de anomalias no *dataset 3W*. Além disso, essa abordagem mostrou-se robusta mesmo diante da escassez de dados anômalos rotulados, um desafio frequente na indústria de óleo e gás.

3 Modelagem do Problema

3.1 Conjunto de Dados

Este estudo utiliza o *3W Dataset*, proposto por Vargas (2019), um dos primeiros conjuntos públicos e realistas voltados à detecção de anomalias em sistemas de elevação natural de petróleo offshore. O nome “3W” refere-se às três fontes de dados que compõem o *dataset*: dados reais, dados simulados e dados desenhados por especialistas. O objetivo do *dataset* é oferecer uma base padronizada e desafiadora para avaliação de modelos de aprendizado de máquina em contextos de alta complexidade operacional e baixa frequência de falhas.

O conjunto de dados contém séries temporais multivariadas de 42 poços offshore naturalmente fluentes, rotulados em nove classes: uma correspondente à operação normal (classe 0) e nove associadas a tipos distintos de anomalias operacionais (classes 1 a 9). Cada instância compreende um intervalo temporal específico (janela de observação) com leituras contínuas de sensores instalados em pontos estratégicos do sistema de produção, como:

1. P-TPT / T-TPT: pressão e temperatura na árvore de natal molhada;
2. P-PDG: pressão medida no *downhole gauge*;
3. P-MON-CKP / T-JUS-CKP: dados da linha de produção após o *choke*;
4. Vazão instantânea, condições do escoamento, entre outras.

As anomalias presentes no *dataset* foram rotuladas manualmente com base em critérios técnicos e operacionais, contemplando os seguintes eventos:

Tabela 1 - Classes de Falhas Mapeadas

Classe	Tipo de Falha	Descrição resumida
0	Operação estável	Poço operando normalmente, sem anomalias.
1	Aumento abrupto do teor de água (BSW)	Crescimento rápido do percentual de água na produção, indicando possível rompimento de zona ou canal preferencial.
2	Fechamento espúrio da válvula DHSV	Atuação incorreta ou não comandada da válvula de segurança da produção.
3	*Slugging* severo	Presença de escoamento intermitente, com grandes variações de pressão e vazão.
4	Instabilidade de fluxo leve/moderada	Variações frequentes, mas menos intensas, típicas de início de regime instável.
5	Perda rápida de produtividade	Redução abrupta na vazão do poço, sem causa evidente imediata.
6	Restrição repentina no *choke*	Restrição de escoamento devido a obstruções súbitas no sistema de controle de fluxo.
7	Anomalia de temperatura	Variações anômalas na temperatura de produção, muitas vezes relacionadas à entrada de fluidos externos.
8	Formação de hidratos na linha de produção	Indícios de bloqueio parcial devido à formação de cristais de hidrato, especialmente sob condições de baixa temperatura e alta pressão.
9	Formação de hidratos na linha de serviço	Formação de hidratos na linha de serviço, comumente relacionada a operações de intervenção ou pressurização que envolvem injeção de fluidos.

3.2 Estrutura Hierárquica de Falhas

A classificação de anomalias em poços de petróleo tradicionalmente adota uma abordagem plana, em que cada tipo de falha é tratado como uma classe independente. No entanto, falhas operacionais em sistemas de elevação natural frequentemente compartilham características físicas, químicas ou funcionais, o que sugere uma estrutura subjacente mais rica entre as classes. Para capturar essas relações, este trabalho propõe a organização das classes do *3W Dataset* em uma hierarquia do tipo “IS-A”, conforme discutido por Silla Jr. e Freitas (2011).

A taxonomia proposta foi construída a partir das classes de falhas mapeadas no *3W Dataset* (Tabela 1), que descrevem o fenômeno físico ou operacional associado a cada evento. Por exemplo, a “classe 1” (aumento abrupto de BSW) foi agrupada como uma falha de natureza contaminante, enquanto as “classes 8 e 9” (formação de hidratos) foram organizadas sob a categoria de falhas químicas. Da mesma forma, falhas relacionadas à dinâmica do escoamento, como *slugging* severo (“classe 3”) e instabilidade de fluxo (“classe 4”), foram agrupadas na categoria hidráulica.

Esse processo resultou em uma estrutura hierárquica de dois níveis:

- Nível 1: categorias gerais fundamentadas na natureza da falha (mecânica, hidráulica, química, produtividade ou contaminante).

- Nível 2: os tipos específicos de falhas já definidos no 3W Dataset (classes 0 a 9 da Tabela 1).

A Figura 1 ilustra essa hierarquia de classes, evidenciando a relação semântica entre as categorias gerais e suas especializações.

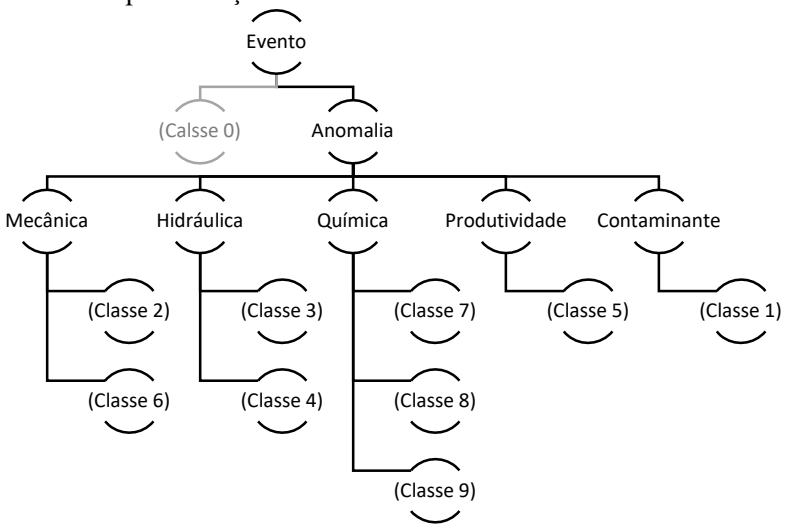


Figura 1 - Taxonomia hierárquica de falhas operacionais no *dataset 3W*.

É importante destacar que a construção da hierarquia não altera os rótulos originais do dataset e não adiciona informações externas que possam enviesar os resultados. Todos os agrupamentos foram baseados exclusivamente na descrição técnica fornecida por Vargas et al. (2019) e em critérios operacionais amplamente aceitos pela engenharia de produção de petróleo. Dessa forma, a taxonomia reflete a natureza intrínseca de cada falha, garantindo consistência sem introduzir viés artificial que pudesse favorecer as métricas de avaliação.

O modelo hierárquico resultante pode ser explorado por diferentes estratégias de modelagem, como classificação local por nível, por nó, ou abordagem global com adaptação da função de perda (Silla Jr.; Freitas, 2011).

3.3 Modelagem com Redes LSTM

Foram utilizados modelos baseados em Long Short-Term Memory (LSTM) com a mesma arquitetura base para ambos os experimentos (comum e hierárquico), a fim de isolar o impacto da estrutura de classificação. A arquitetura utilizada consiste em:

i. Uma camada LSTM com 64 unidades
ii. Duas camada densa intermediária com 32 neurônios e função ReLU
iii. Uma camada de saída:
    a. Com 9 neurônios e softmax na classificação comum
    b. Variando entre sigmoid e softmax na classificação hierárquica.

Os modelos foram treinados por 20 épocas com *early stopping*, utilizando o otimizador *Adam* (learning rate de 0.001) e função de perda categórica cruzada (*sparse categorical crossentropy*).

É importante destacar que, em todos modelos, as variáveis de entrada (atributos) são as mesmas em todos os níveis da hierarquia, consistindo das séries temporais multivariadas descritas na Seção 3.1. Não há inclusão de rótulos de nível 1 como atributos adicionais.

O processo ocorre em cascata: o classificador de Nível 1 identifica a categoria geral da falha (por exemplo, hidráulica, química, mecânica), e essa decisão direciona a instância para o classificador especialista correspondente no Nível 2. Assim, o modelo hierárquico difere do plano não pelas variáveis de entrada, mas pela estratégia de decomposição da tarefa em submodelos especializados por grupo.

3.4 Avaliação dos Modelos

A avaliação dos modelos foi realizada por meio de validação estratificada hold-out, com divisão de 50% dos dados para treinamento e 50% para teste, preservando a proporção de classes no conjunto original. A Figura 2 apresenta a distribuição de frequência entre as classes nos conjuntos de treino (Grupo A) e teste (Grupo B), resultantes da divisão estratificada. Observa-se que a proporção de amostras foi preservada em todas as classes, inclusive nas de baixa representatividade, o que garante que tanto o modelo plano quanto o hierárquico foram avaliados sob condições equivalentes e sem vieses de amostragem.

O desempenho preditivo foi mensurado com base em métricas clássicas de classificação, incluindo:

i. Precision: proporção de predições corretas entre todas as predições positivas realizadas para cada classe;
ii. Recall: proporção de instâncias corretamente identificadas entre todas as instâncias reais de uma classe;
iii. F1-score macro: média harmônica entre precision e recall calculada separadamente para cada classe e depois agregada igualmente, sem ponderação por frequência;
iv. Matriz de confusão: utilizada para análise qualitativa dos erros, identificando padrões de confusão entre classes específicas.

No caso da abordagem hierárquica, a avaliação considerou dois cenários complementares:

i. acerto estrito na classe folha, exigindo que o modelo previsse exatamente o tipo de falha específico; e
ii. acerto parcial por caminho hierárquico, em que também se consideravam como corretas as predições situadas em níveis ancestrais da classe real, desde que estivessem no caminho hierárquico correto. Essa segunda forma de avaliação reflete uma lógica operacional mais tolerante ao erro, especialmente relevante em sistemas de monitoramento em tempo real, nos quais a identificação do grupo geral de anomalia já pode ser suficiente para acionar respostas corretivas iniciais.

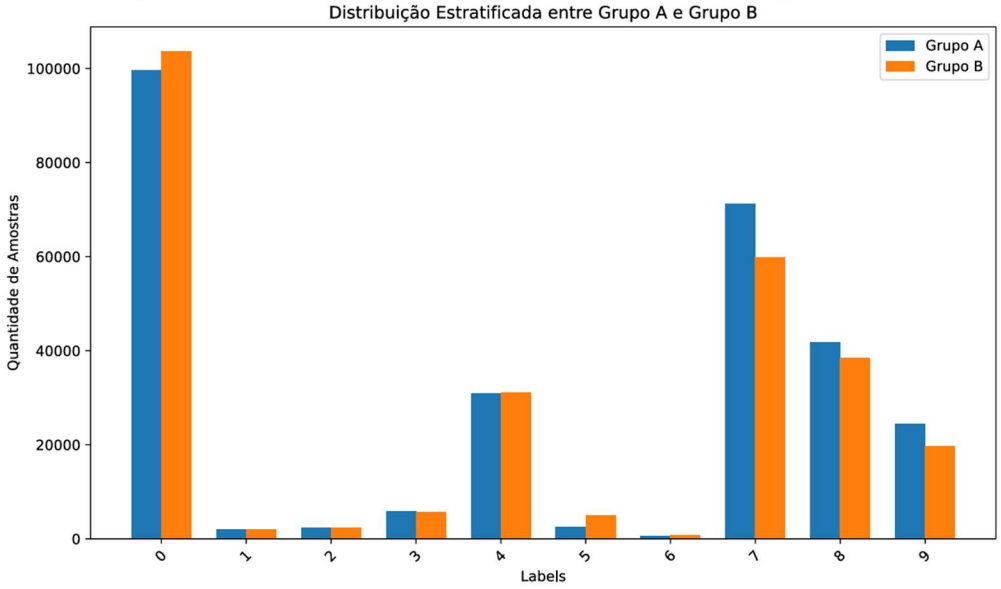


Figura 2 - Distribuição estratificada entre treino e teste

Todo o código, scripts de pré-processamento e experimentos realizados neste estudo estão disponíveis publicamente no repositório "*github.com/bernardopuppim/sbpo2025-oilwell-anomaly-hierarchical-lstm*".

4. Resultados e Discussão

Neste capítulo, apresentam-se os resultados obtidos a partir dos experimentos realizados com os modelos LSTM nas abordagens de classificação plana e hierárquica. São discutidos os impactos da estrutura hierárquica sobre a acurácia preditiva, a robustez frente a eventos raros e a coerência com o processo de decisão operacional.

4.1 Desempenho Geral dos Modelos

A Tabela 2 apresenta os resultados comparativos entre as abordagens de classificação comum (plana) e classificação hierárquica nos níveis definidos para este trabalho. Os modelos foram avaliados com base nas métricas de *precision*, *recall* e *F1-score*, calculadas como a média aritmética simples entre as classes, sem ponderação pelo suporte (número de instâncias por classe), de modo a evidenciar o desempenho geral do classificador independentemente do desbalanceamento dos dados. Também foi incluído o tempo médio de previsão por janela, como indicador da eficiência computacional durante a inferência. Observa-se que os modelos hierárquicos, tanto no Nível 1 (categorias funcionais e físicas) quanto no Nível 2 (tipos específicos de falhas), superaram consistentemente o modelo plano em todas as métricas de desempenho, ainda que com custo computacional marginalmente maior.

Tabela 2 - Comparação de desempenho entre classificações comum e hierárquica

Abordagem	Precision[1]	Recall[1]	F1-score[1]	Tempo[2] (s)
Classificação plana	0,954	0,947	0,946	0.000124
Classificação hierárquica (Nível 1)	0,994	0,989	0,992	0.000511
Classificação hierárquica (Nível 2)	0,995	0,985	0,989	0.000511

[1]média aritmética de todas as classes, sem considerar o tamanho (peso) de cada classe.
[2]Tempo médio de previsão por janela.

Observa-se que o modelo hierárquico obteve desempenho superior em todas as métricas, com destaque para a *F1-macro*, que avalia o equilíbrio entre as classes, especialmente relevante em um conjunto de dados com forte desbalanceamento como o 3W (VARGAS, 2019). A melhoria sugere que a estrutura hierárquica contribui para maior generalização e melhor aproveitamento das relações entre classes.

A análise dos resultados apresentados na Tabela 2 evidencia ganhos relevantes no desempenho preditivo ao se empregar uma abordagem hierárquica em detrimento da classificação plana. O modelo hierárquico no Nível 1, que considera categorias funcionais e mecanismos físicos das falhas, obteve *precision* de 0,994, *recall* de 0,989 e *F1-score* de 0,992 — valores superiores aos da abordagem plana, que apresentou 0,954, 0,947 e 0,946, respectivamente. Isso demonstra que a introdução de estrutura semântica na modelagem permite ao classificador capturar melhor as relações entre as classes, reduzindo erros de categorização.

No Nível 2, correspondente à previsão dos tipos específicos de falha conforme definidos no *dataset 3W*, o desempenho também foi superior ao da abordagem plana, com F1-score de 0,989, mesmo considerando um número maior de classes e maior granularidade na saída do modelo. Isso indica que o classificador hierárquico foi capaz de manter a precisão mesmo em cenários mais desafiadores, possivelmente por se beneficiar da organização taxonômica que orienta o aprendizado e a inferência.

Em termos de tempo médio de previsão por janela, ambos os modelos hierárquicos apresentaram 0,000511 segundos, contra 0,000124 segundos da classificação plana. Embora o tempo seja aproximadamente quatro vezes maior, a diferença é operacionalmente desprezível, sobretudo em aplicações de monitoramento quase em tempo real, como é o caso da gestão de anomalias em poços offshore. Essa relação custo-benefício sugere que o pequeno aumento em tempo de processamento é amplamente compensado pela melhoria na acurácia e interpretabilidade dos resultados.

4.2 Análise por Classe

A Tabela 3 compara o desempenho por classe entre as abordagens plana e hierárquica no Nível 2, que contempla os tipos específicos de falha. Nota-se que a classificação hierárquica supera a abordagem plana em quase todas as classes, especialmente nas de menor representatividade e maior complexidade.

Destaque para as classes 1 e 2, nas quais a abordagem plana apresentou baixa consistência entre *precision* e *recall*, resultando em *F1-scores* significativamente inferiores (0,8239 e 0,8449, respectivamente). Já a abordagem hierárquica obteve *F1-scores* de 0,9976 e 0,9546, indicando um ganho expressivo de equilíbrio e confiabilidade nessas previsões.

Outras classes, como 7, 8 e 9, também tiveram F1-scores mais elevados com a abordagem hierárquica, reforçando a capacidade do modelo em manter desempenho robusto mesmo em casos desafiadores ou pouco frequentes. Uma exceção foi observada na classe 6, onde houve queda no recall da versão hierárquica, impactando o *F1-score* (0,8203), o que sugere que certas estruturas da hierarquia ainda podem induzir ambiguidade em categorias específicas.

Tabela 3 - F1-score por classe (Nível 2)

	Classificação plana			Classificação hierárquica		
Nível 2	**Precision[1]**	**Recall[1]**	**F1-score[1]**	**Precision[1]**	**Recall[1]**	**F1-score[1]**
1	0.9869	0.7071	0.8239	1,0000	0,9952	0,9976
2	0.7490	0.9690	0.8449	0,9398	0,9702	0,9546
3	0.9907	0.9779	0.9843	0,9893	0,9896	0,9894
4	0.9842	0.9915	0.9878	0,9916	0,9919	0,9917
5	0.9904	0.9995	0.9949	0,9961	0,9798	0,9879
6	0.9963	0.9864	0.9913	0,9659	0,7113	0,8203
7	0.9577	0.8920	0.9237	0,9927	0,9959	0,9942
8	0.9979	0.9509	0.9739	0,9924	0,9957	0,9940
9	0.8896	0.9940	0.9389	0,9899	0,9885	0,9891

A Tabela 4 detalha o desempenho da classificação hierárquica no Nível 1, que corresponde às categorias superiores da taxonomia de falhas, agrupadas por origem física ou funcional. Como esperado para um nível mais geral da hierarquia, os resultados mostram *F1-scores* elevados e consistentes em todas as classes, uma vez que a granularidade é menor e os limites entre categorias são menos sutis do que no Nível 2.

A classe 0 (operação normal) mantém desempenho excelente (F1-score de 0,9983), indicando robustez na detecção de estados normais. Todas classes de falha apresentam *F1-scores* superiores a 0,97, reforçando a capacidade do modelo em capturar falhas amplas e bem definidas.

Em síntese, o alto desempenho no Nível 1 está de acordo com a expectativa para níveis superiores da hierarquia "IS-A", onde a abstração reduz a ambiguidade entre classes, favorecendo acertos e reforçando a utilidade da abordagem hierárquica na modelagem de falhas industriais.

Tabela 4 - Desempenho por classe no Nível 1

Nível 1	**VP**	**FP**	**FN**	**Precision**	**Recall**	**F1-Score**
0	152755	259	250	0.9983	0.9984	0.9983
1	1322	19	37	0.9858	0.9728	0.9793
2	25134	195	187	0.9923	0.9926	0.9924

Nível 1	VP	FP	FN	Precision	Recall	F1-Score
3	92552	65	36	0.9930	0.9961	0.9945
4	1789	7	37	0.9961	0.9798	0.9879
5	418	0	2	1.0000	0.9952	0.9976

As Tabelas 5 e 6 apresentam, respectivamente, as matrizes de confusão obtidas pela abordagem plana e pela classificação hierárquica no Nível 1. Essas matrizes permitem visualizar o desempenho do modelo em termos de acertos e erros de classificação por categoria, destacando a capacidade de distinguir corretamente cada classe e identificar os padrões de confusão mais frequentes.

A comparação entre as duas abordagens evidencia que a estrutura hierárquica no Nível 1 promove uma distribuição de erros mais controlada e semanticamente coerente, com menor ocorrência de classificações cruzadas entre categorias funcionais distintas. Isso reforça os ganhos de interpretabilidade e robustez observados nas métricas globais, especialmente no agrupamento correto das falhas em categorias superiores mesmo diante de incertezas.

4.3 Robustez em Eventos Ambíguos

A robustez do modelo frente a eventos ambíguos, situações em que diferentes falhas compartilham sintomas operacionais semelhantes, é um dos principais achados deste estudo. Em sistemas offshore, esse tipo de ambiguidade é comum e levou, na classificação plana, a confusões recorrentes entre classes próximas, como 1, 2 e 9, com queda no F1-score.

Na abordagem hierárquica, o desempenho se manteve elevado mesmo nesses casos, com *F1-scores* acima de 0,95 na maioria das classes. Isso mostra que o modelo foi capaz de ancorar previsões em categorias mais amplas e coerentes, fornecendo respostas úteis mesmo quando a falha exata não era claramente distinguível.

4.4 Limitações e Considerações

Apesar dos resultados promissores obtidos com a aplicação da classificação hierárquica sobre o *dataset 3W*, este estudo apresenta algumas limitações que devem ser consideradas na interpretação dos achados.

Em primeiro lugar, a estrutura hierárquica adotada neste estudo foi construída manualmente, com base em critérios funcionais e físicos das falhas. Embora essa organização siga uma lógica técnica alinhada ao contexto industrial, trata-se de uma taxonomia proposta pelo pesquisador, e não por um especialista diretamente envolvido na curadoria dos eventos do *dataset*. Como alternativa, seria possível adotar abordagens automáticas baseadas em relações estatísticas entre as classes. A ausência de uma validação especializada ou derivação empírica da hierarquia pode impactar o desempenho em cenários nos quais a estrutura definida não representa fielmente a complexidade dos fenômenos reais.

Tabela 5 - Matriz de Confusão: Classificação plana

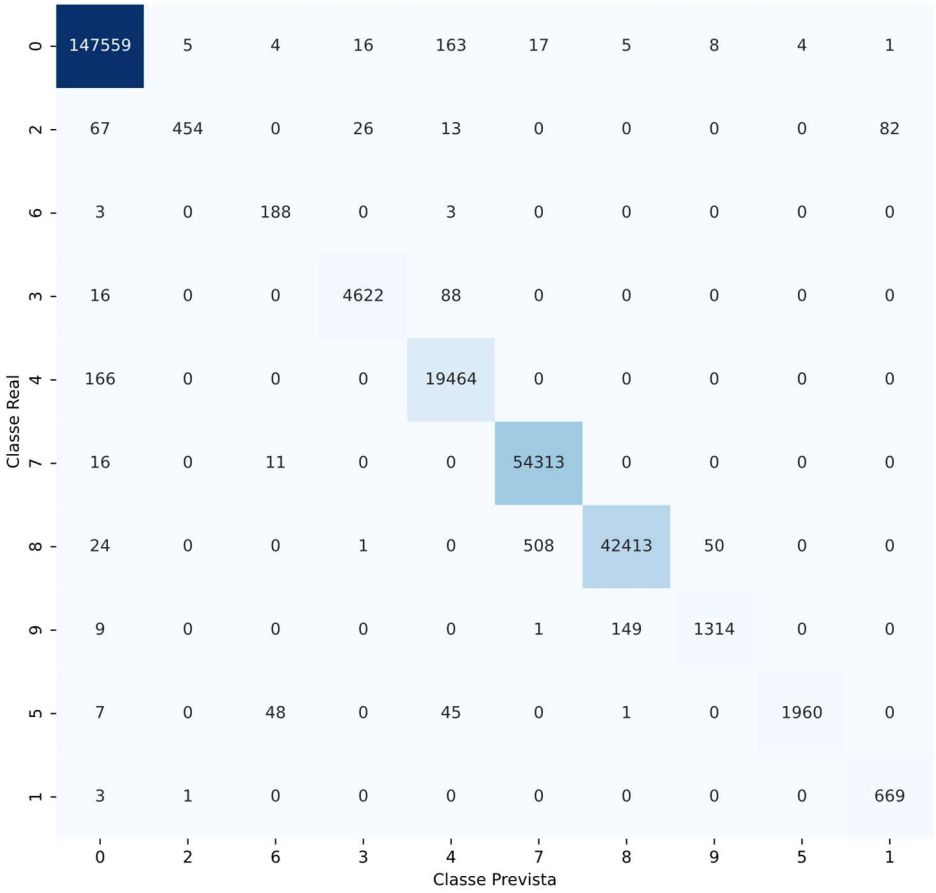
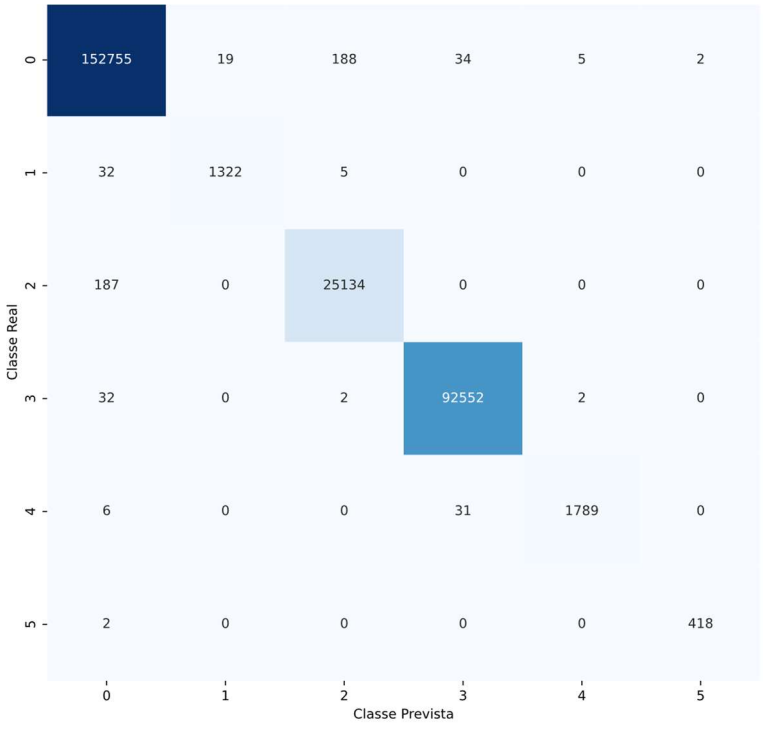


Tabela 6 - Matriz de Confusão: Classificação Hierárquica (Nível 1)



Além disso, a avaliação foi conduzida com base em um único conjunto de dados, ainda que realista e relevante para a indústria de petróleo e gás. A generalização dos resultados para outros contextos operacionais ou tipos de ativos (como poços com elevação artificial, por exemplo) requer validação adicional.

Outro ponto a destacar é que, embora a classificação hierárquica tenha apresentado maior tolerância a ruídos e ambiguidade, o custo computacional, embora baixo, foi superior ao da abordagem plana, o que pode se tornar relevante em aplicações com restrições de latência ou processamento em borda.

5. Conclusões

A detecção automatizada de anomalias em poços de petróleo surgentes é uma tarefa crítica para a operação segura e eficiente na indústria de óleo e gás. Este estudo investigou a aplicação de classificação hierárquica como alternativa à abordagem comum (plana) na identificação de falhas a partir de séries temporais multivariadas extraídas do *dataset 3W* (VARGAS, 2019), combinando modelagem com redes LSTM e uma taxonomia estruturada de falhas baseada em critérios físicos e funcionais.

Os resultados mostraram que a classificação hierárquica superou a abordagem plana em todas as métricas avaliadas, com destaque para o *F1-score macro*, o que indica melhor equilíbrio na detecção de anomalias, inclusive nas classes com menor representatividade. A análise por classe evidenciou maior precisão em categorias raras e ambíguas, e menor dispersão nos erros, confirmada também pelas matrizes de confusão.

Além do ganho quantitativo, a hierarquia de classes proporcionou melhor coerência semântica nas previsões e respostas mais úteis sob a ótica da tomada de decisão operacional. Mesmo quando a classe específica não era corretamente identificada, o modelo hierárquico frequentemente acertava a categoria funcional superior, permitindo inferências parciais confiáveis, algo fundamental em ambientes industriais dinâmicos, onde decisões precisam ser tomadas mesmo diante de incertezas.

A estrutura hierárquica mostrou-se especialmente robusta frente à ambiguidade entre eventos com sintomas operacionais semelhantes, como variações de pressão e vazão, e também frente à escassez de rótulos confiáveis, cenário comum em campos offshore. A representação multiescalar do problema (Nível 1 e Nível 2) refletiu com maior fidelidade a lógica técnica da engenharia de produção e potencialmente favorece decisões operacionais progressivas.

Portanto, os achados reforçam que a classificação hierárquica é não apenas mais precisa, mas também mais alinhada ao contexto operacional, permitindo diagnósticos gradativos, interpretações mais consistentes e maior tolerância a ruídos e falhas de anotação. Isso consolida seu potencial como componente estratégico em sistemas inteligentes de monitoramento em tempo real, especialmente quando integrados a soluções como gêmeos digitais.

Para estudos futuros, recomenda-se:

i. Avaliar estruturas hierárquicas mais profundas ou derivadas automaticamente por agrupamento semântico;
ii. Investigar a integração do modelo a ambientes com feedback contínuo;
iii. Explorar mecanismos de explicabilidade e visualizações orientadas ao usuário final, de modo a ampliar a transparência e a utilidade das inferências para as equipes de campo.

Agradecimentos: Este trabalho foi apoiado pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) – Bolsa de Pesquisa nº 408313/2023-4 e pela Petrobras – Projeto de Pesquisa nº 23069.168903/2023-01.

Referências

BAYAZITOVA, Guzel; ANASTASIADOU, Maria; SANTOS, Vitor Duarte dos. Oil and gas flow anomaly detection on offshore naturally flowing wells using deep neural networks. Geoenergy Science and Engineering, v. 242, p. 213240, 2024. Disponível em: https://doi.org/10.1016/j.geoen.2024.213240.

GATTA, Ricardo et al. Combining convolutional autoencoder and genetic algorithm for anomaly detection in offshore oil wells. In: 2022 IEEE International Conference on Big Data (Big Data). IEEE, 2022. p. 3962–3971.

VARGAS, Ricardo Emanuel Vaz. Base de dados e benchmarks para prognóstico de anomalias em sistemas de elevação de petróleo. 2019. Tese (Doutorado em Engenharia Elétrica) – Universidade Federal do Espírito Santo, Vitória, 2019. Disponível em: https://doi.org/10.13140/RG.2.2.35980.13447.

SILLA JUNIOR, C. N.; FREITAS, A. A. A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery, [S.l.], v. 22, n. 1–2, p. 31–72, 2011. DOI: https://doi.org/10.1007/s10618-010-0173-3.