UNIVERSIDADE PORTUCALENSE
DEPARTAMENTO
CIÊNCIA E TECNOLOGIA

Real-time Explainability for Predictive Maintenance

Ciência de dados | 2025/2026 | 2ºano
Dissertação | Orientadora: Prof. Fátima Leal

Diogo Sousa | 41907

Agradecimentos

A realização desta dissertação representa o culminar de um percurso exigente e profundamente enriquecedor na área da Ciência de Dados aplicada a sistemas industriais.

Gostaria de expressar um agradecimento especial à minha orientadora, pela orientação científica rigorosa, espírito crítico e permanente disponibilidade ao longo deste percurso. O acompanhamento dado foi determinante, não apenas no desenvolvimento desta dissertação, mas também no âmbito do projeto TwinNavAux, onde demonstrou constante apoio, clareza na orientação técnica e disponibilidade para esclarecer dúvidas e discutir soluções. A sua confiança e incentivo contínuo foram fundamentais para o meu crescimento académico e profissional.

Aos meus pais, deixo um agradecimento profundo por todo o apoio incondicional ao longo do meu percurso académico e pessoal. Por sempre me proporcionarem as melhores oportunidades de aprendizagem e experiências, permitindo-me crescer não apenas como estudante, mas como pessoa e futuro profissional.

Aos meus avós, por acreditarem em mim desde sempre e por estarem presentes em todos os momentos importantes, oferecendo apoio, carinho e confiança inabalável.

À minha namorada, que mesmo à distância esteve ao meu lado durante esta fase particularmente exigente, oferecendo compreensão, motivação e equilíbrio nos momentos de maior pressão.

Agradeço igualmente à Universidade Portucalense e ao Departamento de Ciência e Tecnologia pelas condições proporcionadas ao longo do mestrado, bem como a todos os docentes que contribuíram para a minha formação académica e científica.

Por fim, agradeço a todos os que, direta ou indiretamente, contribuíram para a concretização deste trabalho.

Resumo

A deteção de anomalias em ambientes industriais de tempo real constitui um desafio multidimensional que envolve restrições de latência, ausência de rótulos fiáveis, variabilidade operacional e necessidade de interpretabilidade das decisões. Esta dissertação propõe uma arquitetura híbrida para manutenção preditiva baseada num ensemble heterogéneo de modelos não supervisionados, combinando métodos treinados em regime batch sobre um baseline nominal com uma componente incremental adaptativa para operação em fluxo contínuo.

O sistema integra seis paradigmas complementares de deteção: Isolation Forest, Local Outlier Factor, One-Class SVM, K-Means, Predictive Lag-1 baseado em regressão Ridge e Half-Space Trees. Os indicadores produzidos são normalizados através de uma abordagem robusta baseada na mediana e no desvio absoluto mediano (MAD), posteriormente agregados por média aritmética simples e estabilizados por suavização temporal exponencial e confirmação por persistência consecutiva.

Os limiares de decisão são calibrados de forma adaptativa com base em quantis da distribuição empírica observada em regime nominal, garantindo alinhamento com taxas alvo de ativação operacional. A estratégia de validação temporal segue um paradigma prequencial adaptado, preservando a ordem cronológica dos dados e simulando condições realistas de operação em streaming.

Adicionalmente, é proposto um sistema de explicabilidade estruturado em três camadas hieráricas, que fornece desde justificação estatística imediata até análise detalhada dos contributos individuais dos modelos e geração de recomendações acionáveis. A validação experimental demonstra robustez, complementaridade entre detetores e estabilidade decisional, evidenciando a adequação da solução a cenários industriais de monitorização contínua.

Palavras-chave: deteção de anomalias, manutenção preditiva, data streams, ensemble não supervisionado, explicabilidade, edge computing.

Abstract

Real-time anomaly detection in industrial environments presents multidimensional challenges involving latency constraints, lack of reliable labels, operational variability, and the need for decision interpretability. This dissertation proposes a hybrid architecture for predictive maintenance based on a heterogeneous ensemble of unsupervised models, combining batch-trained detectors built on a stable nominal baseline with an adaptive incremental component designed for streaming operation.

The system integrates six complementary detection paradigms: Isolation Forest, Local Outlier Factor, One-Class SVM, K-Means, Predictive Lag-1 based on Ridge regression, and Half-Space Trees. Individual anomaly indicators are normalized using a robust median and Median Absolute Deviation (MAD) approach, aggregated through simple arithmetic averaging, and stabilized using exponential temporal smoothing combined with persistence-based confirmation logic.

Decision thresholds are calibrated adaptively using empirical quantiles derived from nominal operating conditions, ensuring alignment with operational activation targets. The temporal validation strategy follows an adapted prequential framework, preserving chronological data order and simulating realistic streaming conditions.

Furthermore, a three-layer explainability framework is introduced, providing progressive interpretative support ranging from statistical justification to model-level contribution analysis and actionable insights. Experimental validation confirms detector complementarity, robustness under extreme anomalies, and decision stability, supporting the suitability of the proposed approach for continuous industrial monitoring scenarios.

Keywords: anomaly detection, predictive maintenance, data streams, unsupervised ensemble, explainability, edge computing.

Índice

Índice de figuras

Índice de tabelas

Lista de Acrónimos

AI Artificial Intelligence

AUPR Area Under the Precision-Recall Curve

CBM Condition-Based Maintenance

CSV Comma-Separated Values

HST Half-Space Trees

IF Isolation Forest

IG Integrated Gradients

IIoT Industrial Industrial Internet of Things

JSON JavaScript Object Notation

LIME Local Interpretable Model-Agnostic Explanations

LLM Large Language Model

LOF Local Outlier Factor

MAD Median Absolute Deviation

OCSVM One-Class Support Vector Machine

PCA Principal Component Analysis

PdM Predictive Maintenance

PHM Prognostics and Health Management

RBF Radial Basis Function

RMSE Root Mean Square Error

ROC Receiver Operating Characteristic

RUL Remaining Useful Life

SHAP SHapley Additive exPlanations

XAD Explainable Anomaly Detection

XAI Explainable Artificial Intelligence

LLM Large Language Model

1.Introdução

1.1 Contextualização

A evolução das práticas de manutenção reflete a transformação contínua da indústria na procura de maior eficiência, fiabilidade e redução de custos. Durante décadas, predominou a manutenção corretiva, que se baseava em intervenções apenas após a ocorrência de falhas. Apesar da simplicidade de implementação, este paradigma esteve associado a custos elevados, tempos de paragem prolongados e riscos operacionais significativos. Como resposta, consolidou-se a manutenção preventiva, sustentada por inspeções e intervenções calendarizadas ou definidas por ciclos de operação. Embora tenha reduzido falhas inesperadas, esta abordagem conduziu frequentemente a ações desnecessárias e a desperdício de recursos.

Mais recentemente, a manutenção baseada na condição (CBM) passou a apoiar-se na monitorização contínua do estado dos equipamentos, recorrendo a variáveis como vibração e temperatura, permitindo intervenções apenas quando surgem sinais de degradação. A manutenção preditiva (PdM) representa um passo adicional ao combinar esta monitorização com modelos analíticos e de aprendizagem automática capazes de antecipar falhas e estimar o tempo de vida útil remanescente (RUL). Assim, em vez de reagir a um indicador já fora do intervalo aceitável, a PdM procura identificar padrões subtis de degradação e projetar tendências, que suportem o planeamento do momento mais adequado para intervenção. Esta transição, da lógica reativa para abordagens proativas orientadas por dados, é amplamente reconhecida como um eixo central da manutenção moderna e surge reforçada pelo avanço de sensores, conectividade industrial e técnicas de inteligência artificial (IA) aplicadas a cenários reais. (Ucar et al., 2024; Murtaza et al., 2024).

A transição para a Indústria 4.0 conferiu à manutenção preditiva uma nova dimensão, ao possibilitar a recolha e processamento contínuo de dados através de infraestruturas IoT e pipelines orientados para tempo real. Este contexto exige métodos capazes de lidar com dados em fluxo, atualização incremental e restrições operacionais, no que toca, ao tempo de resposta. (Almeida et al., 2023)

Paralelamente, desenvolveu-se o conceito de Prognostics and Health Management (PHM), uma abordagem abrangente que combina monitorização, diagnóstico e prognóstico para avaliar a saúde dos sistemas e antecipar falhas (Ucar et al, 2024) Esta perspetiva integra naturalmente a deteção de anomalias como mecanismo de suporte à identificação precoce de comportamentos de degradação e de transições anormais de estado, consolidando o PHM como estrutura de referência para decisões de manutenção orientadas por dados.

No contexto da PdM, a deteção de anomalias constitui, assim, um componente crítico ao identificar padrões que se desviam do comportamento esperado. Contudo, à medida que se adotam modelos mais complexos, cresce a necessidade de justificar os alertas emitidos. É neste enquadramento que a deteção de anomalias explicável (XAD) ganha relevância, procurando enriquecer os alarmes com informação interpretável e operacionalmente útil, facilitando validação técnica, diagnóstico inicial e confiança nos sistemas em operação contínua. (Li et al., 2023).

1.2 Motivação

A transformação digital impulsionada pela Indústria 4.0, abriu novas oportunidades para a manutenção preditiva, tornando possível a monitorização contínua e a antecipação de falhas em sistemas complexos. Contudo, o impacto real desta abordagem depende da sua capacidade de funcionar em ambientes industriais com dados heterogéneos, eventos de falha raros e condições operacionais variáveis. Em muitos contextos, a disponibilidade de dados históricos de falhas com rótulos é limitada, pelo que abordagens não supervisionadas assumem especial relevância ao aprenderem padrões de normalidade a partir de dados não rotulados e detetarem desvios significativos (Asutkar & Tallur, 2023).

Paralelamente, os dados industriais são frequentemente gerados em alta frequência e num fluxo contínuo, exigindo algoritmos capazes de aprender incrementalmente e de se adaptarem à evolução do sinal sem comprometer a viabilidade operacional. Neste contexto, a literatura sobre análise de séries temporais e aprendizagem em *data streams* reforça a necessidade de modelos e *pipelines* ajustados a operação contínua, com atualização online e restrições computacionais realistas (Almeida et al., 2023).

A explicabilidade assume um papel central na aceitação e confiança dos sistemas de PdM, sobretudo quando a deteção de anomalias suporta decisões operacionais com impacto significativo. Assim, além do desempenho do modelo, a utilidade de XAI em manutenção depende da legibilidade dos resultados para o utilizador e da sua integração efetiva na decisão técnica.

Em cenários com restrições temporais e infraestruturas distribuídas, a computação de proximidade (*edge computing*) surge como resposta natural à necessidade de reduzir latência e aumentar robustez operacional, aproximando deteção e explicação da fonte de dados.

Neste contexto mais alargado insere-se o projeto de investigação TwinNavAux, desenvolvido em parceria com a Universidade Portucalense, cujo objetivo é promover a utilização de gémeos digitais na indústria naval da Galiza e Norte de Portugal. No âmbito deste projeto foi desenvolvido um modelo de Machine Learning, para manutenção preditiva aplicado ao gémeo digital de um navio , ao qual foram integradas técnicas de explicabilidade para análise retrospetiva e suporte à validação por especialistas. Esta experiência evidenciou, na prática, os desafios de monitorizar sistemas navais em funcionamento contínuo, com dados ruidosos e escassez de falhas rotuladas e mostrou a importância de complementar desempenho preditivo com explicações compreensíveis.

Neste enquadramento, a presente investigação propõe-se dar continuidade e aprofundar este trabalho, explorando e avaliando abordagens que conciliem aprendizagem não supervisionada, *stream processing* e explicabilidade transparente em tempo real, numa vertente online.

1.3 Desafios

A concretização de sistemas de manutenção preditiva explicáveis e operacionais em tempo real levanta desafios técnicos e conceptuais que condicionam a sua adoção em ambientes industriais. Em primeiro lugar, a elevada cadência de aquisição exige que algoritmos de PdM processem fluxos contínuos sem comprometer o tempo de deteção, o que reforça a necessidade de modelos incrementais e de *pipelines* otimizados para séries temporais em *streaming* (Almeida et al., 2023).

Em paralelo, a escassez de eventos de falha dificulta a validação e a robustez dos modelos. Em aprendizagem não supervisionada, a ausência de rótulos limita a aplicabilidade de métricas clássicas e reforça a necessidade de avaliação indireta, contextual e orientada por sinais de degradação observáveis. Este problema é recorrente na literatura industrial, surgindo como uma fragilidade persistente na comparação de abordagens e na sua transição para ambientes reais (Asutkar & Tallur, 2023; Li et al., 2023). No contexto do projeto TwinNavAux, focado num gémeo digital para a indústria naval, esta limitação manifesta-se de forma clara, dado o reduzido número de falhas rotuladas disponíveis e a forte variabilidade operacional associada à operação marítima.

Outro desafio crítico em dados em fluxo são os *concept drifts*. Nestes cenários, a manutenção de desempenho exige mecanismos de deteção e adaptação capazes de equilibrar flexibilidade e estabilidade, sobretudo quando a noção de normalidade evolui ao longo do tempo. A literatura sobre aprendizagem evolutiva e monitorização *drift-aware* destaca a importância desta dimensão para a fiabilidade de longo prazo, em sistemas operados continuamente (Cabrera Martin et al., 2025).

No domínio da explicabilidade, destaca-se o compromisso entre detalhe e tempo de resposta. Métodos "*model-agnostic*", baseados em perturbações podem apresentar custos elevados quando aplicados instância a instância em *streaming*, motivando variantes incrementais e estratégias de aceleração por hardware. Mesmo em modelos não supervisionados baseados em *deep learning*, a viabilidade operacional pode variar significativamente consoante configurações e infraestruturas, o que reforça a necessidade de selecionar técnicas explicativas compatíveis com restrições temporais realistas (Leite et al, 2024).

Por fim, a integração em sistemas embebidos e arquiteturas distribuídas é determinante para a adoção prática. A evidência em ambientes *edge e edge–cloud* sugere que aproximar a execução dos modelos e dos explicadores da fonte de dados pode ser decisivo para cumprir requisitos temporais sem comprometer a utilidade interpretativa. No caso específico de cenários navais, como os considerados no TwinNavAux, este tipo de arquitetura é particularmente relevante, dada a intermitência das comunicações e a necessidade de garantir capacidade local de deteção e explicação a bordo.

1.4 Objetivos

O objetivo geral desta dissertação é conceber, implementar e avaliar uma arquitetura de manutenção preditiva explicável em tempo real, orientada para a operação online e assente em aprendizagem não supervisionada em fluxo, privilegiando componentes compatíveis com processamento contínuo. Pretende-se demonstrar a viabilidade de conjugar deteção de anomalias em *streaming* com explicações imediatas e de baixa latência, mantendo robustez face à variabilidade do sinal e à deriva de conceito, sendo a componente *offline* utilizada apenas como referência analítica e suporte retrospetivo.

Para cumprir este objetivo geral, este trabalho propõe estudar o comportamento de algoritmos não supervisionados em ambiente de *streaming*, explorando estratégias de normalização adaptativa das variáveis e mecanismos de definição dinâmica de limiares de alarme, de modo a identificar combinações que proporcionem deteção robusta em cenários operacionais variáveis. Paralelamente, deseja-se integrar métodos de explicabilidade concebidos para operar em linha, de forma a equilibrar profundidade interpretativa e requisitos de latência e a adaptar técnicas de explicação quando compatíveis com as restrições de tempo real. Finalmente, procura-se desenvolver um mecanismo de fusão de importâncias explicativas que agregue evidências provenientes de detetores heterogéneos e de importâncias por variável, articulando essa fusão com métricas de desempenho e de latência e avaliando o seu impacto na clareza, na confiança percebida e na utilidade prática das explicações.

1.5 Contribuições e Perguntas de Investigação

A presente dissertação propõe e valida um sistema integrado de deteção de anomalias e explicabilidade, para manutenção preditiva em tempo real, baseado em aprendizagem não supervisionada e processamento de dados sensoriais em *streaming*. O contributo central do trabalho reside no desenvolvimento de uma arquitetura de três camadas que combina um *ensemble* heterogéneo de modelos de deteção com mecanismos de explicabilidade de latência extremamente reduzida, adequados a ambientes industriais com requisitos rigorosos de tempo real.

O sistema opera predominantemente em regime de inferência em fluxo com adaptação online parcial. A maioria dos detetores é treinada *offline* em modo *batch* sobre um *baseline* representativo de funcionamento normal, garantindo estabilidade estatística e controlo do comportamento inicial. Em paralelo, o sistema integra componentes adaptativos que se atualizam incrementalmente durante o *deployment*, permitindo acomodar variações temporais e responder a fenómenos de *concept drift* sem necessidade de re-treino completo nem interrupção da operação.

A arquitetura integra seis paradigmas independentes de deteção não supervisionada, combinando modelos treinados *offline* em regime *batch* com um modelo adaptativo online. A agregação das decisões é realizada através de um mecanismo de consenso, baseado num limiar estatístico uniforme, o qual permite resolver o problema de escalas díspares entre detetores, sem recorrer a estratégias de ponderação complexas, preservando simultaneamente a interpretabilidade das decisões individuais.

A componente de explicabilidade é estruturada em três níveis complementares, fornecendo informação sobre o grau de consenso entre os modelos, a identificação das variáveis mais relevantes associadas à deteção de anomalias e recomendações heurísticas de intervenção operacional. Estes mecanismos são executados em tempo real, mantendo latência compatível com processamento contínuo em *streaming* e viabilizando a geração de explicações para todas as observações analisadas. A abordagem estabelece assim, um compromisso equilibrado entre velocidade e fidelidade explicativa quando comparada com métodos de referência computacionalmente mais exigentes.

A validação experimental demonstra que o sistema apresenta desempenho elevado e robustez operacional quando aplicado a dados industriais reais. Adicionalmente, o trabalho caracteriza limitações práticas associadas ao comportamento de modelos não supervisionados em cenários de *deployment* prolongado, incluindo fenómenos de degradação temporal em modelos adaptativos, evidenciando que estas situações podem ser identificadas automaticamente através da análise do consenso do *ensemble*.

Deste modo, surgem as seguinte perguntas de investigação:

1. Como conceber um *ensemble* não supervisionado em *streaming* que mantenha independência funcional e resiliência face à degradação de modelos individuais?

2. Como integrar mecanismos de explicabilidade num sistema de deteção de anomalias em *streaming*, garantindo baixa latência operacional e explicações fiáveis quando comparadas com métodos de referência mais dispendiosos?

3. Em que medida um sistema permite identificar degradação temporal e fenómenos de *concept drift* em modelos adaptativos, sem recorrer a supervisão externa?

Para responder às questões de investigação propostas, foi desenvolvido um sistema não supervisionado de deteção de anomalias em *streaming*, suportado por um *ensemble* híbrido e por mecanismos de explicabilidade integrados no fluxo de dados. A abordagem inclui técnicas de fusão de *scores*, explicações em tempo real e monitorização de degradação de modelos adaptáveis, sendo avaliada em cenários industriais representativos com foco em desempenho, latência e fidelidade explicativa.

1.6 Estrutura da dissertação

O Capítulo 1 introduz o problema da manutenção preditiva em contextos industriais, contextualiza a evolução das abordagens baseadas em dados e discute os desafios associados à deteção de anomalias explicável em tempo real. São ainda definidos os objetivos da investigação, as principais contribuições esperadas e as questões de investigação que orientam o trabalho.

O Capítulo 2 apresenta a revisão da literatura, enquadrando a manutenção preditiva no contexto da Indústria 4.0, os paradigmas de aprendizagem em fluxo de dados e os principais métodos de deteção de anomalias, com particular enfoque em abordagens não supervisionadas. São igualmente analisadas técnicas de explicabilidade aplicadas a sistemas de apoio à decisão, identificando limitações existentes em termos de latência, interpretabilidade e robustez operacional.

O Capítulo 3 descreve a metodologia adotada e o desenvolvimento do sistema proposto. São caracterizados os dados utilizados, apresentada a análise exploratória, discutida a seleção e importância de *features* e detalhado o pré-processamento em ambiente de *streaming*. O capítulo descreve ainda a arquitetura do *ensemble* heterogéneo, as estratégias de normalização e agregação de *scores*, o mecanismo de consenso *threshold-based* e o sistema de explicabilidade estruturado em três camadas.

O Capítulo 4 apresenta os resultados experimentais. São analisadas as métricas de desempenho do sistema de deteção, incluindo avaliação binária e multiclasse, análise por severidade e latência computacional. É efetuada a comparação com *baselines* de referência e avaliada a qualidade das explicações através de concordância com SHAP, análise de consenso e estudos de caso ilustrativos, incluindo cenários de falha severa e anomalias *borderline* com dissenso inter-modelo.

Por fim, o Capítulo 5 sintetiza as conclusões do trabalho, apresenta as principais contribuições científicas e técnicas, discute as limitações identificadas incluindo aspetos relacionados com o *dataset* e adaptação de modelos online e propõe direções para investigação futura, nomeadamente no domínio de mecanismos *drift-aware*, evolução arquitetural *edge–cloud* e enriquecimento da camada de explicabilidade.

2. Revisão de Literatura

Esta revisão da literatura sintetiza os conceitos e desenvolvimentos centrais relacionados com a manutenção preditiva explicável, com ênfase na deteção de anomalias em contextos industriais e na explicabilidade das decisões automatizadas. Inicialmente enquadra-se a manutenção preditiva no âmbito da Indústria 4.0, descrevendo o papel da monitorização da condição de ativos na antecipação de falhas, na otimização de intervenções e na redução de custos operacionais (Murtaza et al., 2024; Fernandes et al., 2022). Nesta parte são discutidos os requisitos funcionais e organizacionais que condicionam a adoção de soluções de PdM em ambiente industrial, bem como evidências empíricas sobre benefícios e limitações reportadas na literatura recente.

Segue-se uma secção dedicada aos paradigmas de *flow-based learning*, onde se analisam as propriedades essenciais para operação contínua, tais como atualização incremental, restrições de memória e latência, e mecanismos de adaptação a *concept drift* (Gama et al., 2014; Cabrera Martin et al., 2025). Nesta secção são também revistas as bibliotecas e primitivas que suportam processamento em *streaming*, com destaque para implementações que facilitam a integração em *pipelines* industriais. A revisão das famílias de métodos de deteção de anomalias não supervisionada aborda métodos baseados em densidade, isolamento, *clustering*, distância multivariada e modelos preditivos, comparando as suas características em termos de sensibilidade a ruído, capacidade de detetar anomalias pontuais e contextuais, e aptidão para execução online (Liu et al., 2008; Cook et al., 2019; Mozaffari et al., 2022).

A análise prossegue com as técnicas de explicabilidade aplicadas à manutenção preditiva, com foco em cenários de deteção de anomalias. São descritas metodologias amplamente utilizadas, incluindo explicadores baseados em perturbação e em aproximação local (Ribeiro et al., 2016; Lundberg & Lee, 2017), explicadores baseados em gradiente e abordagens específicas para dados temporais e para scores de anomalia (Oliveira et al., 2022; Asutkar & Tallur, 2023). Para cada família de técnicas são discutidos os benefícios e as limitações operacionais, em particular a adequação face a restrições de latência e a capacidade de produzir explicações acionáveis para técnicos de manutenção. A revisão inclui ainda trabalhos que propõem adaptações de métodos de XAI a fluxos contínuos e propostas que combinam múltiplas perspetivas explicativas para aumentar coerência e confiança.

Por fim, são examinados os requisitos de baixa latência e as arquiteturas de computação edge relevantes para a implementação de PdM em ambiente industrial (Satyanarayanan, 2017; Xiang & Zhang, 2022). Esta secção sintetiza os principais desafios identificados na literatura, nomeadamente o compromisso entre profundidade explicativa e rapidez de execução, a gestão de recursos em dispositivos edge e as estratégias de partilha de carga entre *edge* e *cloud*. Ao longo do capítulo são identificadas lacunas e oportunidades de investigação que motivam o presente trabalho, em particular a necessidade de soluções que integrem detetores não supervisionados em *streaming* com mecanismos de explicabilidade concebidos para operação em tempo real e validados por critérios de utilidade operacional.

2.1 Manutenção Preditiva na Indústria 4.0

A manutenção preditiva , representa a evolução natural das estratégias de manutenção rumo a uma lógica pró-ativa e orientada por dados. Distingue-se de abordagens reativas e de intervenções calendarizadas, por procurar antecipar falhas com base na monitorização contínua do estado dos ativos e na utilização de modelos analíticos capazes de apoiar decisões de manutenção, com maior precisão e eficiência. Revisões recentes destacam que esta consolidação acompanha o crescimento da instrumentação, da conectividade industrial e da adoção de técnicas de aprendizagem automática aplicadas à monitorização de condição em diferentes setores. (Ucar et al., 2024; Murtaza et al., 2024).

A operacionalização da PdM nas fábricas inteligentes é impulsionada pela conectividade do *Industrial Internet of Things* (IIoT) e a crescente monitorização de equipamentos, permitindo recolher dados de vibração, temperatura, pressão, corrente e outros indicadores de condição com granularidade cada vez maior (Mozaffari et al, 2022) . Em ambientes reais, estes dados suportam estratégias de diagnóstico e prognóstico mais fiáveis, em particular quando a deteção de anomalias é integrada como componente contínua do processo. Neste contexto, a literatura recente salienta também, a necessidade de considerar requisitos arquitetónicos e mecanismos de integração de modelos em tempo real, visando garantir utilidade operacional em sistemas de engenharia complexos (Nsor, 2024; Murtaza et al., 2024).

Este contexto intensivo em dados reforça a relevância de métodos de processamento capazes de lidar com séries temporais multivariadas em fluxo contínuo.

Em termos arquiteturais, a literatura recente sugere uma convergência para soluções híbridas *edge–cloud* ou *edge–fog–cloud*. O *edge* favorece resposta rápida e robustez local, particularmente relevante quando a cadência do sinal é elevada, enquanto camadas superiores permitem análise agregada, comparação entre ativos e atualização de modelos globais. Esta distribuição torna-se ainda mais crítica quando se integra explicabilidade como componente operacional, procurando equilibrar utilidade interpretativa e latência de execução em ambientes industriais distribuídos (Rosenberger et al, 2023; Cook et al, 2019)

Deste modo, a PdM na Indústria 4.0, pode ser entendida como uma combinação entre dados sequenciais, algoritmos dinâmicos e arquiteturas computacionais adequadas ao ambiente industrial (Asutkar & Tallur, 2023). É neste enquadramento que se posiciona esta dissertação, ao privilegiar deteção de anomalias não supervisionada e aprendizagem incremental em *streamin*g, complementadas por explicações de baixo custo computacional capazes de apoiar a decisão técnica em condições operacionais reais.

2.2 Aprendizagem em Fluxo de Dados

O surgimento de fluxos contínuos de dados nas aplicações industriais, trouxe novos desafios à aplicação direta de algoritmos tradicionais de machine learning. Numa configuração clássica, parte-se habitualmente de um conjunto de dados fixo, dividido em treino, validação e teste, a partir do qual o modelo é ajustado e posteriormente aplicado. Porém, em contextos de manutenção preditiva, os dados são gerados de forma sequencial e potencialmente ilimitada, com sensores a emitir leituras em ciclos regulares e sistemas a operar de forma contínua (Li et al, 2024; Almeida et al, 2023). Esta natureza dinâmica inviabiliza abordagens baseadas em re-treino periódico sobre todo o histórico e exige mecanismos capazes de lidar com dados em tempo real.

Neste enquadramento, torna-se fundamental distinguir entre aprendizagem em fluxo e inferência em fluxo, conceitos que na literatura são frequentemente utilizados de forma indistinta. A aprendizagem em *streaming* refere-se à atualização incremental dos parâmetros do modelo, à medida que novas observações são recebidas, permitindo que a noção de normalidade evolua progressivamente sem necessidade de re-treino completo (Bäßler et al, 2022; Koch et al, 2024) A inferência em *streaming*, por sua vez, corresponde à aplicação do modelo a cada nova observação para produzir decisões ou scores em tempo real, não implicando necessariamente qualquer modificação dos parâmetros internos (Cook et al, 2019; Koch et al, 2024). Em sistemas industriais, estas duas componentes podem operar de forma desacoplada, sendo comum que apenas uma parte dos modelos se adapte online, enquanto outros mantêm parâmetros fixos, assegurando maior estabilidade, previsibilidade e controlo do custo computacional durante o *deployment* (Biikes et al, 2024).

A aprendizagem em fluxo impõe requisitos específicos, nomeadamente processamento online com custo computacional reduzido por observação, utilização limitada de memória e mecanismos de atualização suficientemente rápidos para acompanhar a cadência dos dados (Weinberg, 2025; Almedia et al, 2023). Na prática, este paradigma é suportado por modelos incrementalmente atualizáveis, como árvores de decisão em fluxo ou métodos de clusterização adaptativos, que ajustam os seus parâmetros de forma contínua. Uma característica central destes cenários é a possibilidade de ocorrência de deriva de conceito, particularmente relevante na manutenção preditiva, onde o comportamento considerado normal pode evoluir gradualmente devido a desgaste, alterações ambientais ou mudanças no regime operacional, bem como surgir novos modos de falha (Weinberg, 2025). Para lidar com este fenómeno, recorrem-se frequentemente a mecanismos de esquecimento controlado, janelas temporais deslizantes ou estratégias de adaptação progressiva do modelo (Abdoune et al., 2026; Koch et al, 2024).

No contexto da manutenção preditiva, a aprendizagem online permite que o sistema se mantenha alinhado com a condição atual do equipamento monitorizado. À medida que novas leituras são recebidas, um detetor de anomalias pode refinar a sua representação de normalidade e reconhecer alterações subtis no comportamento operacional. Esta capacidade adaptativa é essencial para reduzir falsos positivos associados a variações benignas e simultaneamente preservar sensibilidade a padrões de degradação relevantes. Contudo, esta dinâmica impõe restrições exigentes ao tempo de

processamento por observação, sobretudo em sinais de elevada cadência, onde a deteção e a interpretação devem ocorrer dentro de ciclos de decisão rigorosos (Bäßler et al, 2022).

2.2.1 Tipos de Aprendizagem em Machine Learning

Os diferentes paradigmas de aprendizagem automática distinguem-se pela exigência de dados rotulados, pelo regime de treino e pela respetiva capacidade de adaptação ao longo do tempo, influenciando de forma direta a conceção dos sistemas de manutenção preditiva e as metodologias de avaliação empregues.

A aprendizagem supervisionada utiliza exemplos rotulados para estimar funções de mapeamento entre entradas e saídas. Este paradigma é adequado quando existem registos fiáveis de falha e permite treinar modelos discriminativos para tarefas de classificação e de regressão, embora dependa criticamente da disponibilidade e da qualidade das anotações (Nsor, 2024; Buabeng et al, 2023). Em oposição, a aprendizagem não supervisionada procura descobrir a estrutura intrínseca nos dados sem recurso a rótulos. Esta abordagem é aplicada com frequência à deteção de anomalias e ao agrupamento de dados, revelando-se útil quando as falhas são raras ou quando o processo de anotação é dispendioso (Barbariol, 2023).

Num plano intermédio, a aprendizagem semi-supervisionada combina informação rotulada e não rotulada com o objetivo de melhorar a generalização perante a escassez de rótulos (Cohen, 2021). Já a aprendizagem auto-supervisionada cria tarefas auxiliares, tais como a previsão de sequência ou a reconstrução parcial, para aprender representações úteis a partir de grandes volumes de dados não anotados (Gomes et al, 2019;Li et al, 2024). Por outro lado, a aprendizagem por reforço, envolve um agente que aprende por interação com um ambiente através de sinais de recompensa. Este paradigma aplica-se sobretudo a problemas de controlo e otimização de políticas, sendo relevante em manutenção prescritiva, onde a definição de recompensas e a interação sequencial são fundamentais. (Gomes et al, 2019)

Quanto ao modo de treino, distingue-se o treino estático do treino em tempo real ou em fluxo. O treino realizado em *offline* utiliza conjuntos históricos para ajustar modelos de forma fixa, enquanto o treino em fluxo, permite a atualização contínua perante alterações nas distribuições dos dados, o que é essencial em cenários industriais sujeitos a desvios operacionais (Cao et al, 2025. Neste contexto, a aprendizagem contínua procura manter o desempenho ao longo do tempo sem negligenciar o conhecimento prévio, enfrentando desafios como o esquecimento catastrófico e a necessidade de mecanismos de retenção de informação.

Desta forma, métodos de agregação de modelos e arquiteturas híbridas que combinam detetores e modelos de reconstrução com classificadores são frequentemente adotados, para aumentar a robustez e reduzir a taxa de falsos positivos em ambientes de produção.

2.3 Deteção de Anomalias Não Supervisionada

Em cenários industriais reais, a deteção não supervisionada assume particular relevância porque muitas falhas são raras, não estão previamente catalogadas ou surgem sob novas formas, tornando impraticável depender exclusivamente de históricos completos e anotados (Asutkar & Tallur, 2023; Huang & Wu, 2022).

2.3.1 Tipos de Anomalias

A classificação das anomalias é essencial para a deteção em manutenção preditiva porque a natureza do desvio orienta a escolha de modelos, as etapas de preparação dos dados e os requisitos de explicabilidade. Nesta subsecção distinguem-se três tipos relevantes referidos na literatura e aplicáveis ao contexto desta dissertação, com indicação das suas implicações práticas: anomalias pontuais, contextuais e coletivas.

Anomalias pontuais correspondem a observações isoladas cujo valor difere de forma significativa do comportamento esperado (Yan, 2019). Em manutenção preditiva um exemplo típico é um pico súbito de vibração provocado por um impacto ou por ruído transitório no sensor. Estes eventos são frequentemente capturados por métodos baseados em limiares ou por detetores de outliers, mas a sua identificação fiável exige validação entre canais e inspeção de janelas temporais adjacentes para reduzir falsos positivos (Almeida et al, 2023). As explicações associadas a este tipo de anomalia devem ser concisas e indicar a variável e o instante que motivaram a sinalização, de modo a suportar decisões operacionais imediatas (Ucar et al, 2024).

Anomalias contextuais são valores que só se tornam anómalos quando considerados num determinado contexto operacional ou temporal (Cook et al., 2019; Li et al., 2023). Um valor de temperatura que é aceitável durante o arranque de um equipamento pode ser anómalo em regime estacionário. A deteção eficaz requer modelos que incorporem informação contextual, como estados de operação, janelas temporais e variáveis auxiliares que caracterizem o modo de funcionamento (Nguyen et al, 2025) . Em *stream processing*, isto implica modelos condicionais ou de séries temporais que capturem dependências temporais e condicionais, e explicações que clarifiquem o contexto que torna a observação anómala (Fragkoulis et al, 2024).

Anomalias coletivas referem-se a sequências ou padrões de observações que, em conjunto, indicam comportamento anómalo embora cada ponto isolado possa parecer normal (IBM, 2025). Um exemplo frequente em PdM é a degradação gradual de um componente, manifestada por pequenas alterações acumuladas ao longo do tempo que só se tornam evidentes quando se analisa uma janela temporal mais ampla. A deteção destas anomalias beneficia de modelos sequenciais e de análise de tendência, como redes recorrentes, *autoencoders* temporais ou métodos de deteção de mudança, e recorre à agregação temporal e a técnicas de suavização para evidenciar a tendência subjacente (Oliveira et al, 2024). As explicações devem focar-se em janelas e tendências, identificando intervalos temporais e variáveis que mais contribuem para a sinalização de degradação e apresentando a evidência acumulada que suporta intervenções de manutenção preventiva (Verma et al, 2022).

A distinção entre tipos de anomalia orienta o desenho do pipeline de dados e a avaliação dos modelos, sendo que procedimentos de preparação e filtração reduzem falsos positivos em anomalias pontuais, normalização condicionada por modo de operação facilita deteção de anomalias contextuais, e agregação temporal evidencia anomalias coletivas. (Yan, 2019).

2.3.2 Algoritmos de Deteção Não Supervisionada

A deteção de anomalias não supervisionada é particularmente adequada a contextos industriais, onde os dados rotulados são escassos e os modos de falha são raros ou heterogéneos. Estes métodos baseiam se na aprendizagem de uma representação do comportamento normal do sistema, identificando desvios estatisticamente significativos face a essa referência.

Uma família relevante de algoritmos inclui métodos baseados em distância ou densidade. Técnicas como o K means permitem detetar anomalias através da distância aos centróides representativos do funcionamento normal (Martins, 2022), enquanto abordagens baseadas em vizinhança, como o Local Outlier Factor (Yan, 2019), avaliam variações na densidade local das observações. Apesar da sua simplicidade e interpretabilidade, estes métodos podem ser sensíveis à escolha de hiperparâmetros e à coexistência de múltiplos regimes operacionais.

Os métodos baseados em fronteiras de decisão procuram delimitar explicitamente, a região associada ao comportamento normal. O *One Class Support Vector Machine* é um exemplo representativo, sendo eficaz em espaços de elevada dimensão, embora apresente limitações ao nível da escalabilidade e da adaptação contínua a dados em fluxo (Paltenghi, 2020).

Outra classe amplamente utilizada corresponde a *ensembles* baseados em partições aleatórias do espaço de atributos, como o *Isolation Forest*, nos quais a facilidade de isolamento de uma observação é utilizada como indicador de anomalia (Elsaid et al, 2024;*Schindler*, et al 2023). Estes modelos são tipicamente treinados em regime batch e aplicados em inferência contínua, oferecendo um compromisso favorável entre robustez e custo computacional.

Em cenários onde a dinâmica temporal é relevante, surgem abordagens preditivas não supervisionadas, nas quais o erro de previsão constitui o sinal de anomalia. Modelos autorregressivos ou regressões multivariadas treinadas com dados normais permitem capturar dependências temporais e identificar desvios persistentes no comportamento do sistema.

Por fim, algoritmos concebidos para aprendizagem em fluxo, como as Half Space Trees, possibilitam a deteção de anomalias em tempo real com atualização incremental dos parâmetros, tornando se adequados a ambientes sujeitos a deriva de conceito (Romero et al, 2024). Contudo, esta capacidade adaptativa requer mecanismos de controlo para evitar a incorporação de comportamentos anómalos na definição de normalidade.

2.3.3 Avaliação e Ajuste de Limiares

A avaliação de sistemas não supervisionados constitui um desafio central. A ausência de rótulos limita o uso direto de métricas clássicas de classificação e impõe a utilização de indicadores

operacionais e de validação contextual, como inspeção de eventos sinalizados, cenários simulados e análise da antecipação de alertas face a falhas conhecidas quando estas existem. A calibração do limiar de deteção é um ponto crítico: muitos modelos produzem um peso de anomalia e a definição do limiar deve equilibrar sensibilidade e taxa de falsos alarmes (Chevtchenko et al, 2023; Li, 2024). Estratégias adaptativas, por exemplo baseadas em estatísticas móveis, estabilização do *score* ou reamostragem periódica, podem aumentar a robustez face à variabilidade operacional e ao *concept drift* em ambientes de fluxo contínuo (Weinberg, 2025).

Do ponto de vista operacional, a deteção é apenas o primeiro passo, um alerta sem contexto raramente suporta uma decisão técnica informada. Por isso, a integração entre deteção não supervisionada e explicabilidade tem vindo a ganhar destaque: associar justificações aos alarmes aumenta a confiança, facilita a validação por especialistas e torna a PdM mais útil em termos práticos, sobretudo em *pipelines* de dados em fluxo. Na metodologia desta dissertação documenta-se como as famílias de métodos foram selecionadas e combinadas, como os limiares foram calibrados de forma adaptativa e de que modo as explicações foram integradas para suportar diagnóstico e priorização de intervenções.

2.4 Explicabilidade aplicada em PdM/XAD

Nos últimos anos, a Inteligência Artificial Explicável (XAI) consolidou-se como um conjunto de métodos destinados a tornar modelos de machine learning mais transparentes e compreensíveis. No entanto, no contexto específico da deteção de anomalias, emerge uma subárea particular frequentemente designada por *Explainable Anomaly Detection* (XAD) , cujo foco não reside apenas na explicação de predições classificatórias, mas na justificação de desvios face a um padrão de normalidade (Li et al, 2023).

Neste contexto, explicabilidade e interpretabilidade referem-se à capacidade de um sistema justificar a emissão de um alerta de anomalia de forma inteligível para engenheiros e decisores, reforçando a confiança operacional, facilitando a validação técnica de alertas e suportando ações de manutenção (Rožanec et al., 2021; Leite et al., 2024). Em ambiente industrial, requisitos de latência, robustez e coerência física condicionam fortemente a seleção e adaptação das técnicas de XAD.

O panorama metodológico distingue duas perspetivas complementares que orientam a construção de explicações. A perspetiva orientada ao modelo procura decompor a predição do detetor em contributos de variáveis ou em estruturas aproximadas do próprio modelo, permitindo explicar por que motivo uma instância foi sinalizada (Weinberg, 2025). A perspetiva orientada aos dados procura identificar quais os atributos que tornam a observação rara face ao padrão de normalidade, independentemente do mecanismo de deteção (Nguyen et al, 2025). Ambas as perspetivas são necessárias para traduzir resultados em linguagem física útil aos operadores e para validar sinais de degradação em regimes operacionais distintos (Yan, 2019).

Os critérios que guiaram a seleção das técnicas para este trabalho foram a fidelidade explicativa, a estabilidade temporal, o custo computacional por explicação, a compatibilidade com processamento em fluxo e a interpretabilidade física das justificações. A fidelidade refere-se à capacidade da

explicação de refletir o comportamento real do modelo, a estabilidade refere-se à consistência das justificações para instâncias semelhantes ao longo do tempo, o custo computacional e a latência condicionam a aplicabilidade em linha e a interpretabilidade física determina a utilidade prática para técnicos de manutenção ou operadores (Li et al, 2024; Weinberg, 2025) .

Privilegiando uma perspetiva focada no modelo, a taxonomia atual separa metodologias "*model-agnostic*" de técnicas para modelos diferenciáveis. No domínio das abordagens agnósticas, o *Local Interpretable Model-Agnostic Explanations* (LIME) (Weinberg, 2025; Oliveira et al, 2021), propõe a representação local do comportamento de sistemas de 'caixa-negra' através de modelos lineares ou de decisão simples, cuja fidelidade é garantida pela resolução de um problema de otimização local descrita na Equação 1,

$$L\left(f, g, \pi_x\right) + \Omega(g) \qquad (1)$$

onde L mede a discrepância entre o modelo complexo f e o modelo explicador g ponderada pela função de proximidade π_x, e Ω penaliza a complexidade de g. Na prática, LIME gera amostras modificadas na vizinhança da instância e ajusta um modelo linear que evidencia as *features* locais mais influentes, sendo útil para diagnósticos pontuais e para comunicar contributos de sensores individuais, mas a sua dependência da densidade de amostragem e de múltiplas avaliações de f torna-o custoso em cenários de *streaming* (Ribeiro et al., 2016). Outra técnica de referência é *SHapley Additive exPlanations* (SHAP), que se baseia em valores de Shapley para atribuir a cada variável i uma contribuição ϕi retratada na Equação 2,

$$\phi_i = \sum_{S \subseteq N - \{i\}} \left(\frac{|S|!\left(|N| - |S| - 1\right)!}{|N|!} \right) \left(f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S\left(x_S\right) \right) \qquad (2)$$

onde N é o conjunto de todas as variáveis e f_S representa o modelo restrito ao subconjunto S. SHAP fornece uma decomposição aditiva com fundamentação teórica que facilita comparações entre instâncias e regimes de operação, mas o cálculo exato é exponencial no número de variáveis, o que obriga a aproximações, amostragem e variantes incrementais para uso em tempo real (Lundberg e Lee, 2017) (FN. Oliveira et al, 2021; Weinberg, 2025; Madathil et al., 2024).

Para modelos diferenciáveis aplicados a séries temporais, as técnicas baseadas em gradiente foram consideradas como alternativas de menor custo temporal. *Integrated Gradients* (IG) define a atribuição para a variável i apresentada na Equação 3,

$$IG_i(x) = \left(x_i - x'_i\right) \int_0^1 \left(\frac{\partial F\left(x' + \alpha\left(x - x'\right)\right)}{\partial x_i} \right) d\alpha \qquad (3)$$

onde x′ é uma referência neutra e a aproximação *Grad times Input* é detalhada pela Equação 4.

$$\phi_i(x) = x_i \frac{\partial F(x)}{\partial x_i} \qquad (4)$$

Estas técnicas reutilizam gradientes calculados na inferência e são mais facilmente aceleráveis por quantização e otimização de execução, o que as torna adequadas quando a latência é um constrangimento crítico, embora exijam cuidados na escolha da referência e na interpretação das magnitudes atribuídas (Sundararajan et al, 2017; Rožanec et al, 2021).

A perspetiva orientada aos dados recorre a medidas que explicam por que motivo uma observação difere do padrão de normalidade. Em modelos de reconstrução, como *autoencoders*, a explicação pode basear-se na decomposição do erro de reconstrução, como mostra a Equação 5,

$$E(x) = \left\| x - \hat{x} \right\|^2 = \sum_i \left(x_i - \hat{x}_i \right)^2, \qquad (5)$$

Para além das perspetivas orientadas ao modelo e aos dados, tem emergido uma terceira vertente particularmente relevante em contextos industriais: as explicações contrafactuais (*counterfactual explanations*) (Molnar et al, 2020) . Em vez de apenas justificar por que motivo uma instância foi considerada anómala, esta abordagem procura responder à questão inversa: que alterações mínimas nas variáveis observadas conduziriam o sistema de volta ao regime normal? (Laugel et al, 2019).

Formalmente, dado um ponto anómalo $x_{anomaly}$,procura-se um ponto x′ tal como a Equação 6,

$$min_{x'} \left\| x' - x_{anomaly} \right\| sujeito\ a\ f(x') = NORMAL \qquad (6)$$

onde f representa a função de decisão do sistema. Esta formulação permite identificar modificações mínimas nas variáveis monitorizadas que seriam suficientes para evitar a sinalização de anomalia.

No contexto de manutenção preditiva, as explicações contrafactuais apresentam uma vantagem significativa, pois transformam uma explicação descritiva numa orientação acionável (Laugel et al, 2019). Em vez de apenas indicar quais os sensores mais responsáveis pelo alerta, fornecem uma aproximação das condições operacionais necessárias para restabelecer a normalidade (Verma et al, 2022). Esta característica aproxima a explicabilidade da manutenção prescritiva, reduzindo a distância entre diagnóstico e intervenção.

Embora o cálculo exato de contrafactuais possa ser computacionalmente exigente, diversas estratégias aproximadas têm sido propostas para viabilizar a sua utilização em cenários de tempo

real, nomeadamente através de otimização restrita, linearizações locais ou utilização de modelos *surrogate* (Ates et al, 2021).

A avaliação das explicações deve considerar *trade offs* entre fidelidade, estabilidade, custo computacional e utilidade operacional. Uma métrica de fidelidade local prática pode ser expressa pela Equação 7,

$$Fidelity(x, S) = \left| f(x) - f(x \setminus S) \right|, \qquad (7)$$

onde x\S representa a instância com as variáveis S removidas ou substituídas por valores de referência. A fidelidade quantifica, a explicação reflete o comportamento do modelo, enquanto a estabilidade mede a consistência das explicações para instâncias semelhantes ao longo do tempo (Ribeiro et al, 2016). Em cenários de *streaming*, o custo por explicação e a latência são métricas críticas que condicionam a opção entre métodos exatos e aproximados, pelo que a avaliação deve integrar métricas técnicas e estudos com utilizadores finais para aferir utilidade prática.

A revisão da literatura evidencia que, no domínio de *Explainable Anomaly Detection* aplicado à manutenção preditiva, existe uma diversidade significativa de abordagens propostas, refletindo a heterogeneidade dos modelos e contextos industriais. Contudo, observa-se também a recorrência sistemática de um conjunto restrito de metodologias, nomeadamente técnicas baseadas em perturbação local como LIME, decomposições aditivas fundamentadas em valores de Shapley como SHAP, métodos baseados em gradiente para modelos diferenciáveis e, mais recentemente, formulações contrafactuais orientadas à ação. A frequência com que estas abordagens surgem em estudos distintos sugere a sua consolidação como referências dominantes no ecossistema XAD. Assim, apesar da pluralidade metodológica existente, a literatura converge progressivamente para um núcleo de técnicas que equilibram fidelidade explicativa, custo computacional e aplicabilidade em contextos de deteção não supervisionada, constituindo o enquadramento conceptual que sustenta as escolhas analisadas nesta dissertação.

2.4.1 Métricas de Explicabilidade

A avaliação de métodos de XAD exige métricas distintas das tradicionalmente utilizadas para desempenho preditivo. Enquanto medidas como *accuracy* ou *F1-score* quantificam a qualidade da decisão do modelo, a explicabilidade deve ser avaliada segundo critérios como fidelidade, estabilidade, coerência estatística, viabilidade e custo computacional.

A fidelidade mede o grau em que a explicação reflete o comportamento real do modelo. Uma explicação é considerada fiel quando a alteração ou remoção das variáveis identificadas como relevantes provoca uma variação consistente no *score* ou na decisão do detetor (Ribeiro et al, 2016). Métricas baseadas em procedimentos de deletion e insertion, amplamente utilizadas na literatura de avaliação de explicações, analisam precisamente essa sensibilidade do modelo às variáveis

destacadas (Nguyen et al, 2025). A estabilidade, por sua vez, avalia a consistência das explicações para instâncias semelhantes ou ao longo do tempo, sendo particularmente relevante em ambientes industriais sujeitos a ruído e pequenas flutuações nos sensores (Madathil et al., 2024).

No domínio específico de XAD, onde o objetivo é justificar desvios face a um padrão de normalidade, surgem critérios adicionais. A coerência estatística da explicação pode ser analisada verificando se as variáveis apontadas como relevantes correspondem efetivamente a desvios significativos relativamente ao *baseline* nominal (Li et al., 2023). Em modelos baseados em reconstrução ou densidade, a decomposição do *score* de anomalia permite avaliar se os atributos identificados são responsáveis pela baixa probabilidade conjunta ou pelo aumento do erro de modelação. Embora métricas como IoU, *Bounding Box* ou *Energy-Based Pointing Game* sejam frequentes em tarefas de visão computacional (Verma et al, 2022), a sua aplicabilidade a séries temporais multivariadas industriais é limitada.

Nos métodos baseados em explicações contrafactuais, a literatura introduz métricas adicionais como validade, proximidade, dispersão e diversidade (Klase et al, 2020; Guidotti et al., 2018). A validade verifica se o *contrafactual* altera efetivamente a decisão do modelo; a proximidade mede a distância entre a instância original e a instância modificada; a *sparsity* penaliza alterações excessivas em múltiplas variáveis; e a diversidade avalia a capacidade de gerar múltiplas soluções alternativas (Klase et al, 2020; Verma et al, 2022) . Trabalhos recentes propõem ainda métricas como IM1 e IM2 para avaliar qualidade estrutural de explicações contrafactuais baseadas em imagens, bem como indicadores de plausibilidade baseados na proximidade aos dados de treino e na satisfação de restrições causais (Klase et al, 2020; Mothilal et al., 2019).

Em cenários de inferência em *streaming*, a avaliação da explicabilidade deve integrar também critérios operacionais. O tempo de elaboração da explicação, a latência adicional introduzida no pipeline e o custo computacional por instância tornam-se fatores críticos (Madathil et al., 2024). Uma explicação teoricamente rigorosa pode revelar-se impraticável se comprometer os requisitos de tempo real do sistema. Assim, a avaliação deve equilibrar fidelidade, estabilidade e utilidade prática, assegurando que a explicação contribui efetivamente para a tomada de decisão operacional.

2.5 Explicabilidade em Tempo Real

A incorporação de explicabilidade em contextos de tempo real introduz constrangimentos adicionais face ao paradigma tradicional de análise *offline*. Em abordagens clássicas de XAI, a explicação é frequentemente gerada após a inferência, sem restrições temporais significativas. Por outro lado na manutenção preditiva baseada em fluxo contínuo, a explicação deve acompanhar a deteção dentro de janelas compatíveis com a tomada de decisão operacional. A sua utilidade passa a depender não apenas da fidelidade ao modelo, mas também da sua disponibilidade em tempo útil (Weinberg, 2025).

A inclusão de mecanismos explicativos em *pipelines* contínuos acrescenta uma dimensão temporal ao processo de inferência. O custo computacional, a frequência de geração de explicações e a estabilidade dos contributos tornam-se fatores críticos quando previsões são produzidas de forma

sucessiva e a elevada cadência (Nguyen et al, 2025). A explicabilidade deixa de ser apenas um mecanismo interpretativo e passa a integrar o próprio desenho do sistema, exigindo alinhamento com princípios de aprendizagem em fluxo e deteção de *concept drift* (Gama et al., 2014; Cabrera Martin et al., 2025).

Do ponto de vista metodológico, diferentes estratégias procuram equilibrar latência e qualidade explicativa. Uma abordagem consiste na utilização de métodos leves ou incrementalmente atualizáveis, como técnicas baseadas em gradiente ou aproximações locais simplificadas, que exploram diretamente a estrutura do modelo (Rožanec et al, 2021; et al, 2024). Outra estratégia envolve a gestão adaptativa da frequência de explicação, limitando-a a alertas relevantes ou a janelas agregadas, reduzindo a sobrecarga computacional sem eliminar a capacidade interpretativa (Nguyen et al, 2025). Adicionalmente, a pré-computação de informação explicativa parcial, como perfis médios de importância por regime de operação, permite gerar justificações quase imediatas quando combinada com dados recentes.

Em cenários de dados em fluxo ganha particular relevância a noção de explicabilidade incremental, na qual a explicação evolui juntamente com o modelo. Em vez de recalcular contributos integralmente a cada nova observação, reutiliza-se informação previamente estimada e aplicam-se mecanismos de atualização parcial. Esta perspetiva aproxima a explicabilidade dos princípios da aprendizagem online, promovendo eficiência sem comprometer coerência interpretativa (Weinberg, 2025).

Importa, contudo, salientar que a explicabilidade em tempo real não substitui a análise pós-hoc. Pelo contrário, ambas desempenham papéis complementares. Explicações geradas sob restrições temporais tendem a privilegiar síntese e rapidez, enquanto análises pós-hoc permitem exploração aprofundada de padrões, validação de hipóteses e auditoria detalhada do comportamento do modelo (Ucar et al, 2024; Madathil et al, 2024). Em contextos industriais, esta dualidade revela-se particularmente relevante, pois enquanto justificações imediatas suportam decisões operacionais, análises posteriores contribuem para melhoria contínua, calibração de limiares e revisão de estratégias de manutenção (Nguyen et al, 2025).

A apresentação das explicações assume igualmente um papel central. Em ambientes temporais exigentes, explicações extensas ou excessivamente técnicas podem comprometer a sua eficácia (Laugel et al, 2019). Representações compactas que destacam variáveis críticas, sensores dominantes ou tendências recentes facilitam decisões rápidas e informadas (Rožanec et al, 2021). Assim, a explicabilidade em tempo real aproxima-se de um requisito funcional orientado à ação, no qual clareza e relevância se sobrepõem à exaustividade.

Finalmente, a avaliação da explicabilidade sob restrições temporais permanece um desafio em aberto. É necessário garantir que a redução de latência não compromete a fidelidade das explicações ao comportamento do modelo e, simultaneamente, medir o impacto prático dessas justificações na operação. A combinação de métricas, técnicas de consistência e estabilidade, com indicadores centrados no utilizador, constitui uma direção promissora para consolidar a aplicabilidade da XAD em sistemas de manutenção preditiva contínua.

2.6 Requisitos de Tempo Real e Computação Edge

Para responder às exigências de latência e fiabilidade em manutenção preditiva explicável, não basta otimizar algoritmos; é frequentemente necessário repensar a arquitetura de implantação dos sistemas. Neste contexto, o *edge computing* assume um papel central (Nguyen et al., 2025). Em cenários industriais onde a decisão deve acompanhar ciclos rápidos de aquisição e onde a explicação idealmente deve surgir em simultâneo com o alerta, deslocar todo o processamento para a *cloud* pode revelar-se limitador (Jirwe, 2021). Aproximar a lógica de decisão à fonte de dados, através de controladores locais, módulos embarcados ou *gateways* industriais, reduz atrasos de comunicação e assegura continuidade operacional mesmo sob conectividade instável, reforçando a adequação da manutenção preditiva a ambientes de operação contínua (Rosenberger et al., 2023).

A execução no *edge* traz igualmente benefícios ao nível da robustez e da gestão de dados. O processamento local de parte do fluxo permite filtrar, agregar e sumarizar informação antes do seu envio para camadas superiores, mitigando custos de comunicação e favorecendo estratégias de privacidade e resiliência (Xiang et al., 2022) .A literatura recente aponta que arquiteturas híbridas que combinam *edge* e *cloud* são particularmente adequadas quando se pretende conciliar deteção em tempo real, explicabilidade e análise agregada de maior complexidade (Rosenberger et al., 2023; Achiluzz et al., 2022). No domínio marítimo, por exemplo, uma abordagem que executa deteção e explicação localmente a bordo e delega análise global e atualização de modelos em terra mantém a operação robusta mesmo sob conectividade limitada (Xiang & Zhang, 2022; Satyanarayanan, 2017).

Uma estratégia recorrente consiste na adoção de *pipelines* hierárquicos e assíncronos, nos quais a resposta imediata é priorizada e a explicação é progressivamente refinada. Neste desenho, o nível *edge* produz um alerta e uma justificação inicial de baixo custo computacional, enquanto camadas intermédias ou a *cloud* geram explicações mais detalhadas quando o contexto e os recursos o permitem (Nguyen et al., 2024). Tal abordagem evidencia que a viabilidade temporal da XAI em manutenção preditiva depende não apenas do método explicativo, mas também do local de execução e do desenho do pipeline.

A integração de variantes do SHAP adaptadas a fluxos de dados, aliada à aceleração por GPU e a técnicas de compressão ou quantização, demonstra como decisões arquiteturais e de engenharia permitem compatibilizar a geração de explicações com os requisitos de operação contínua (Siegel et al, 2020). O desenvolvimento de sistemas explicáveis em tempo real implica, assim, escolhas integradas sobre que componentes executar localmente e qual o nível de detalhe interpretativo adequado a cada camada. Funções críticas, como deteção de anomalias e interpretação inicial, tendem a beneficiar de execução próxima da fonte de dados, enquanto análises complementares podem ser realizadas de forma assíncrona em camadas superiores, preservando a resposta imediata (Chiang et al, 2016).

Neste enquadramento, a articulação entre eficiência algorítmica e processamento distribuído deixa de constituir uma otimização opcional e passa a assumir caráter estrutural. Em cenários de *edge*

computing, a análise de *data streams* não é apenas uma escolha metodológica, mas uma necessidade imposta por restrições de latência e largura de banda inerentes à monitorização industrial contínua.

2.6.1 *Trade-offs* entre Precisão e Latência

Em sistemas de manutenção preditiva em tempo real existe um compromisso inevitável entre precisão analítica e latência computacional. Modelos mais expressivos tendem a capturar padrões complexos e a melhorar o desempenho preditivo, mas implicam maior custo de processamento por observação, enquanto abordagens mais simples permitem decisões mais rápidas à custa de menor sensibilidade a desvios subtis.

Este *trade off* é crítico em cenários industriais, onde a utilidade da deteção depende da capacidade de produzir decisões dentro de ciclos de controlo rigorosos (Cao et al, 2025; Siegel et al, 2020; Murtaza et al, 2024). A latência total engloba não apenas a inferência dos modelos, mas também o pré processamento, a agregação de *scores* e a geração de explicações, exigindo um desenho integrado do pipeline. (Weinberg, 2025)

A literatura indica que ganhos marginais de precisão nem sempre justificam aumentos significativos de latência, sobretudo quando reduzem o tempo disponível para intervenção (Mozaffari et al., 2022; Almeida et al., 2023; Satyanarayanan, 2017). Assim, sistemas orientados para tempo real devem privilegiar soluções que equilibrem desempenho e eficiência computacional, recorrendo a modelos incrementais e mecanismos de decisão simples, compatíveis com operação contínua e restrições de tempo real (Gama et al, 2014).

2.7 Trabalhos Relacionados

No panorama científico atual, a literatura tem convergido para a integração de modelos não supervisionados com mecanismos de XAI, refletindo a necessidade de sistemas mais autónomos e transparentes (Li et al., 2023; Asutkar & Tallur, 2023; Ucar et al, 2024). Publicações recentes evidenciam um esforço crescente em mitigar a escassez de dados rotulados, movendo o estado da arte de soluções estáticas para paradigmas de aprendizagem evolutiva e explicável(Li, 2024; Laugel et al, 2019). Com base na pesquisa sistemática descrita na Secção 2.8, foi identificado um conjunto alargado de estudos recentes que exploram esta interseção entre PdM, dados em fluxo e abordagens XAD, a partir do qual se destacam aqui os trabalhos mais diretamente alinhados com os objetivos desta dissertação.

A análise destes estudos revela que uma parte significativa das propostas mantém a deteção baseada em modelos temporais de reconstrução, dada a sua eficácia na modelação de padrões normais e na identificação de desvios em séries multivariadas (Yahya et al., 2025; Gomolka et al., 2025; Mercurio, 2024). Em paralelo, observa-se interesse crescente em soluções mais leves e incrementalmente atualizáveis, com potencial de implementação em ambientes *edge*, quando o objetivo é equilibrar desempenho, robustez e interpretabilidade operacional.

Uma dimensão particularmente crítica na avaliação de sistemas explicáveis em fluxo é o desempenho temporal das explicações: embora alguns trabalhos reportem latências compatíveis com implantação prática, a literatura permanece heterogénea na forma como mede e comunica estes resultados, o que limita comparações diretas (Mercurio, 2024; Li et al., 2023) . Ainda assim, vários estudos demonstram que otimizações específicas de SHAP para dados em fluxo, aceleração por GPU e estratégias de compressão ou quantização de modelos podem aproximar a explicabilidade dos requisitos de tempo real em cenários industriais (Weinberg, 2025 ; Ribeiro et al, 2016.

Adicionalmente, verifica-se uma concentração de estudos em domínios como o automóvel, a manufatura discreta e a energia, com cobertura mais limitada de cenários marítimos, o que reforça a relevância de validações neste setor para avaliar generalização e restrições operacionais específicas . Em conjunto, estas observações sustentam a motivação para a presente dissertação: desenvolver e avaliar uma abordagem de PdM não supervisionada, incremental e explicável, compatível com operação em tempo real e com requisitos de baixa latência, respondendo às lacunas identificadas na literatura recente e aprofundadas na síntese crítica da Secção 2.8.3.

2.7.1 Pesquisa Sistemática da Literatura

A pesquisa sistemática foi orientada pelas diretrizes PRISMA-ScR, que são adequadas a revisões de âmbito exploratório e mapeamento de evidência. Dada a natureza multidisciplinar do tema, situado na interseção entre manutenção preditiva, deteção de anomalias não supervisionada, aprendizagem em fluxo e explicabilidade, adotou-se uma estratégia de busca iterativa concebida para conciliar cobertura ampla do estado da arte com um aprofundamento dirigido ao subdomínio mais diretamente relevante para o contributo proposto. Esta abordagem permite identificar tendências gerais, metodologias emergentes e lacunas específicas relacionadas com requisitos de incrementalidade e restrições temporais, aspetos salientados pela literatura recente como críticos para estudos de PdM explicável.

2.7.2 Estratégia de Pesquisa e critérios de filtragem

A pesquisa bibliográfica foi conduzida no Google Scholar, utilizado como ponto de partida pela sua ampla cobertura e capacidade de agregação de referências. A consulta utilizou a seguinte query:

("streaming anomaly detection" OR "online anomaly detection") AND ("unsupervised" OR "self-supervised") AND ("explainable" OR "interpretable" OR "XAI" OR "explainable anomaly detection") AND ("data stream" OR "online learning") AND ("predictive maintenance" OR "fault detection" OR "condition monitoring") AND (real-time OR "on-the-fly" OR incremental)

Foram aplicados filtros para publicações a partir de 2022, qualquer idioma e qualquer tipo de documento, de modo a garantir a inclusão de estudos recentes e metodologicamente relevantes. A execução desta query devolveu 145 resultados.

Procedeu-se à remoção de duplicados e à triagem por título e resumo, avaliando a pertinência de cada estudo face ao escopo da dissertação, nomeadamente: deteção de anomalias em tempo real,

aprendizagem não supervisionada, explicabilidade aplicada a dados temporais e requisitos de operação em tempo real ou incremental. Após este processo de filtragem, foram selecionados 109 estudos para análise final.

2.7.3 Estudos selecionados

Os estudos selecionados na fase focada apresentam uma diversidade relevante de domínios e abordagens técnicas, combinando aplicações diretas de PdM e monitorização de condição, com contributos adjacentes importantes para arquiteturas de streaming e integração de explicabilidade em pipelines industriais. Em termos algorítmicos, a literatura recente evidencia a predominância de modelos de deep learning não supervisionados, com destaque para autoencoders LSTM e convolucionais usados para modelar normalidade e detetar desvios pelo erro de reconstrução, bem como o surgimento de arquiteturas mais recentes, incluindo variantes não supervisionadas baseadas em Transformers (Gomolka et al, 2025; Yahya et al, 2025; Chalapathy et al, 2019; Zakeriharandi et al, 2025; Cao et al, 2025)

Em paralelo, identificam-se contributos alinhados com aprendizagem incremental e modelos computacionalmente mais leves, particularmente relevantes para implementação em bibliotecas de fluxo e cenários de baixa latência. Métodos como Isolation Forest online, Half-Space Trees e estruturas incrementais baseadas em árvores demonstram a viabilidade de conciliar operação contínua e eficiência computacional sob restrições temporais exigentes (Leveni, 2025; Romero et al, 2024; Abdoune et al, 2026, Martin et al, 2025). Estas abordagens sugerem alternativas práticas a arquiteturas profundas quando a cadência de processamento e o custo computacional assumem prioridade.

No que respeita às técnicas de explicabilidade, a literatura selecionada revela predominância de abordagens pós-hoc, com especial incidência em métodos baseados em valores de Shapley e técnicas de atribuição de importância de variáveis aplicadas a modelos de deteção de anomalias e séries temporais (Li, 2024; Rožanec et al, 2021; Weinberg, 2025). Alguns estudos exploram integração de explicabilidade em contextos de dados em fluxo e cenários adaptativos, evidenciando esforços para compatibilizar interpretação e operação contínua (Malarkkan et al, 2025). Contudo, a análise comparativa demonstra heterogeneidade na forma como o desempenho temporal e o custo computacional das explicações são avaliados, dificultando comparações diretas entre propostas.

Por fim, observa-se que os estudos com maior maturidade operacional tendem a considerar explicitamente a infraestrutura de execução, recorrendo a arquiteturas distribuídas e estratégias edge ou híbridas para cumprir requisitos temporais e reduzir dependência de conectividade contínua (Nardi, 2024; Koch et al, 2024; Rosenberger et al, 2023). Apesar da diversidade setorial, verifica-se menor incidência de aplicações em contexto marítimo quando comparado com setores como manufatura e energia, reforçando a pertinência de validações adicionais neste domínio.

2.7.4 Considerações finais

Da revisão sistemática e da análise crítica dos trabalhos recentes emergem várias constatações que orientaram diretamente as decisões metodológicas desta dissertação. Em primeiro lugar, apesar do crescimento de estudos sobre manutenção preditiva e explicabilidade, a caracterização explícita de cenários de elevado débito e verdadeiro *streaming* contínuo permanece irregular. Embora existam propostas aplicadas a dados industriais e séries multivariadas, a descrição das condições de fluxo e dos requisitos temporais nem sempre é uniforme, limitando comparações diretas entre abordagens (Leveni, 2025; Weinberg, 2025). Esta heterogeneidade reforça a necessidade de avaliações controladas sob restrições temporais claramente definidas.

Em segundo lugar, apenas um subconjunto de trabalhos integra mecanismos de explicabilidade de forma explícita em contextos de dados em fluxo. Estudos recentes exploram a aplicação de métodos baseados em valores de Shapley e técnicas explicativas adaptadas a cenários temporais, evidenciando avanços na aproximação entre deteção de anomalias e interpretação (Li, 2024; Rožanec et al. 2021). No entanto, o reporte do custo computacional e do impacto temporal das explicações permanece inconsistente, dificultando uma avaliação sistemática da sua viabilidade operacional.

Em terceiro lugar, observa-se predominância de abordagens centradas num único método explicativo por estudo. Mesmo quando múltiplas técnicas são discutidas, raramente são propostas estratégias estruturadas de integração complementar em ambiente operacional (Weinberg, 2025). Este padrão evidencia oportunidade para explorar combinações entre métodos de atribuição de importância e abordagens adaptadas a fluxo contínuo, sobretudo quando estabilidade e robustez interpretativa são requisitos críticos.

Por fim, a literatura revela forte incidência de aplicações nos setores da manufatura e sistemas industriais distribuídos (Olupona et al, 2023), frequentemente associadas a arquiteturas de execução distribuída ou edge para suportar requisitos de latência (Nardi, 2024). Contudo, a validação sob condições explicitamente caracterizadas de *streaming* contínuo e restrições temporais quantificadas permanece limitada na amostra analisada.

As lacunas identificadas incluem a caracterização ainda irregular da explicabilidade sob verdadeiro *streaming* contínuo, a ausência de integração estruturada de múltiplos métodos de XAI em ambiente operacional e a necessidade de validação sob restrições temporais claramente mensuradas. Estas constatações sustentam o posicionamento desta dissertação: privilegiar uma abordagem não supervisionada e incremental compatível com dados em fluxo (Leveni, 2025), integrar mecanismos de explicabilidade com evidência de aplicabilidade a séries temporais (Li, 2024), e explorar complementaridade entre métodos para reforçar a interpretabilidade operacional em manutenção preditiva em tempo real.

3. Metodologia

Este capítulo apresenta a metodologia adotada para o desenvolvimento do sistema de manutenção preditiva proposto, bem como o enquadramento experimental que suporta a sua implementação. São descritos os dados utilizados, o processo de simulação de fluxos sensoriais em tempo real e as estratégias de pré-processamento aplicadas aos dados em streaming.

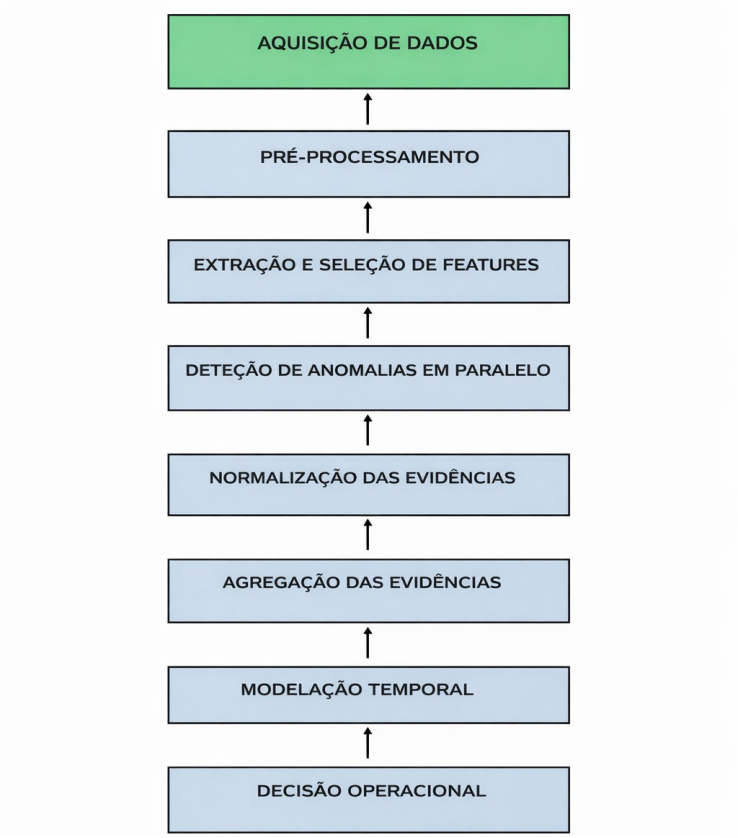

Figura 1: Arquitetura Metodológica do pipeline

A Figura 1 apresenta uma visão abstrata da arquitetura metodológica adotada, destacando a sequência lógica das operações sem compromisso com escolhas de implementação específicas. Os detalhes relativos aos modelos utilizados, estratégias de normalização, agregação, suavização temporal e calibração de limiares são descritos nos capítulos seguintes.

3.1 Dados

Os dados utilizados no desenvolvimento e validação experimental do sistema de manutenção preditiva proposto foram fornecidos por uma empresa parceira no âmbito do projeto TwinNavAux. Estes dados resultam de um modelo de simulação industrial previamente desenvolvido pela referida organização.

O sistema apresentado na Figura 2, representa um circuito hidráulico composto por tubagens, bombas, tanques e válvulas, descrito através de um diagrama piloto fornecido pela empresa.

O modelo foi implementado em OpenModelica, recorrendo a componentes da Modelica Standard Library (MSL 4.0.0), e complementado com parâmetros adicionais necessários à sua operacionalização. As tubagens apresentam comprimentos compreendidos entre 2 m e 5 m e um diâmetro de 25 mm. Os tanques encontram-se posicionados a um nível de 2 m acima das bombas de carga, enquanto as bombas de descarga operam contra um sistema externo com uma pressão absoluta de 1,2 bar, não incluído explicitamente no modelo.

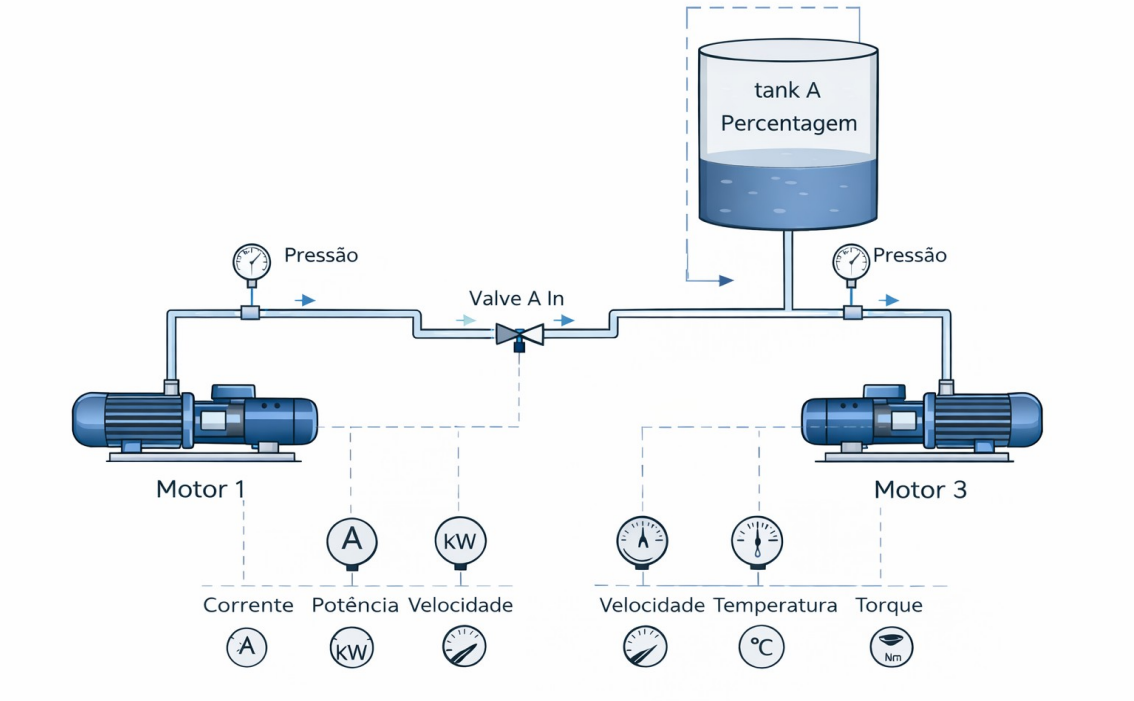

Figura 2: Esquema simplificado do sistema hidráulico

As tubagens apresentam comprimentos compreendidos entre 2 m e 5 m e um diâmetro de 25 mm. Os tanques encontram-se posicionados a um nível de 2 m acima das bombas de carga, enquanto as bombas de descarga operam contra um sistema externo com uma pressão absoluta de 1,2 bar, não incluído explicitamente no modelo.

As curvas características das bombas foram estimadas através de funções quadráticas dependentes da pressão de descarga e do caudal. Os motores elétricos associados às bombas foram modelados de forma simplificada, sendo a potência elétrica consumida e o calor dissipado determinados a partir de uma curva de eficiência em função da velocidade de rotação. As vibrações dos motores apresentam um valor nominal de 30 mm/s, proporcional às rotações por minuto do eixo, sendo este valor multiplicado por um fator de dez em situações de falha. Adicionalmente, em cenários de degradação

ou falha, verifica-se uma redução da eficiência dos motores, o que conduz a um aumento do consumo elétrico e da temperatura de operação. As válvulas do sistema são comandadas por sinais lógicos, posteriormente transformados em variáveis contínuas de abertura através de uma função de transferência de primeira ordem, de modo a simular o tempo real de atuação. O sistema de controlo, igualmente desenvolvido em Modelica, regula o funcionamento das bombas e válvulas com base nos níveis de enchimento dos tanques.

Foram disponibilizados três conjuntos de dados correspondentes a diferentes estados operacionais do sistema: funcionamento normal, funcionamento degradado e situação de falha. No cenário de funcionamento normal, o sistema opera sem qualquer tipo de degradação, observando-se após um período transitório de aproximadamente 3 315 segundos, um comportamento cíclico que se repete com um período de cerca de 3 711 segundos. No cenário de funcionamento degradado, foram aplicadas reduções específicas da eficiência nominal a cada um dos motores, com fatores de 0,95, 0,93, 0,98 e 0,88 para os motores 1 a 4, respetivamente, resultando num aumento do consumo elétrico e do aquecimento. O cenário de falha tem por base o funcionamento degradado e inclui a ocorrência de uma avaria no rolamento da bomba de admissão, originando um aumento abrupto das vibrações, da potência consumida e da temperatura do motor afetado.

Cada simulação cobre um horizonte temporal de 20 000 segundos, com um intervalo de amostragem constante de 0,5 segundos, perfazendo aproximadamente 40 000 registos por ficheiro. Os dados encontram-se armazenados em ficheiros no formato CSV, contendo séries temporais multivariadas que incluem, para cada motor, medições de corrente elétrica, potência elétrica, velocidade de rotação, temperatura, binário e nível de vibração. Para além das variáveis associadas aos motores, os ficheiros incluem o grau de abertura das válvulas de interconexão e os níveis percentuais de enchimento dos tanques. Esta diversidade de variáveis permite caracterizar simultaneamente o comportamento individual dos componentes e a dinâmica global do sistema, conforme sintetizado na Tabela 1.

Tabela 1: Componentes do sistema e respetivas variáveis monitorizadas

Componente	**Variáveis Monitorizadas**
Motores (4)	Corrente, Potencia, Velocidade, Temperatura, Torque, Vibração
Válvulas	Grau de abertura
Tanques	Nível de enchimento

Embora os dados tenham sido fornecidos como ficheiros estáticos, estes foram utilizados neste trabalho como fluxos de dados, de forma a simular um ambiente de monitorização em tempo real. Para esse efeito, foi atribuído um *timestamp* a cada amostra, com um crescimento fixo de 0,5 segundos, a partir de um instante inicial previamente definido. Esta abordagem permitiu simular a

chegada sequencial de dados e avaliar modelos de aprendizagem incremental e de deteção de anomalias em *streaming*, em linha com os objetivos da dissertação.

Em síntese, os dados disponibilizados pela empresa parceira do projeto TwinNavAux constituem uma base adequada e realista para o estudo de técnicas de manutenção preditiva, permitindo analisar comportamentos normais, degradados e de falha, bem como avaliar a capacidade de deteção precoce e a interpretabilidade dos modelos propostos em contextos industriais complexos.

3.2 Análise Exploratória dos Dados

A análise exploratória tem como objetivo caracterizar as séries temporais e avaliar, de forma sistemática, a presença de padrões discriminativos entre regimes operacionais, a heterogeneidade entre unidades funcionais e a existência de comportamentos precursores de avaria. Esta etapa fornece suporte metodológico às decisões de preparação dos dados e à definição do pipeline de deteção, ao permitir quantificar variabilidade, redundâncias e dependências temporais que condicionam o desempenho dos detetores de anomalias (Cao et al, 2020).

Foram analisadas estatísticas descritivas fundamentais por motor e por regime operacional, incluindo média, mediana, desvio padrão, intervalo interquartil e valores extremos, com o objetivo de caracterizar a dispersão, assimetria e estabilidade das variáveis medidas. Para avaliar dependências entre grandezas, recorreu-se a medidas de correlação linear, nomeadamente o coeficiente de Pearson , complementadas por inspeção visual de matrizes de correlação (Ucar et al, 2024). Estas métricas permitiram identificar relações fortes entre variáveis elétricas e mecânicas, bem como diferenças de comportamento entre motores.

A estrutura global dos dados foi ainda analisada através de técnicas de redução de dimensionalidade, em particular Análise de Componentes Principais, utilizada exclusivamente com fins exploratórios, para avaliar a separabilidade entre regimes e a distribuição da variância pelas diferentes variáveis (Cook et al, 2020). Adicionalmente, foram inspecionadas dinâmicas temporais associadas a transições entre estados, recorrendo a visualização de séries temporais e análise de variação local, de modo a identificar ruturas, aumentos de variabilidade e padrões transitórios relevantes para a deteção precoce.

Em síntese, a análise exploratória fornece uma base empírica para as escolhas metodológicas adotadas ao longo do trabalho. A variabilidade observada entre motores sustenta a normalização individualizada por unidade e por variável, com parâmetros estimados exclusivamente em regime normal. A presença de correlações elevadas, aliada à possibilidade de desacoplamento em condições anómalas, justifica a manutenção de variáveis parcialmente redundantes. A ausência de precursores estáveis de longo prazo e a elevada variabilidade reforçam a adoção de limiares adaptativos e de mecanismos temporais de decisão, enquanto a natureza sequencial dos dados impõe uma validação temporal alinhada com cenários realistas de manutenção preditiva em tempo real (Martin et al, 2026; Cao et al 20240).

3.3 Importância e Seleção de *Features*

A avaliação da importância e a seleção de *features* constituem um passo determinante no desenho de sistemas de deteção de anomalias não supervisionados para manutenção preditiva em tempo real, conforme destacado na literatura de *anomaly detection* em séries temporais (Cao et al, 2024). Estas decisões influenciam simultaneamente o desempenho da deteção, a robustez face a ruído e *concept drift* e a eficiência computacional do sistema. Em contextos evolutivos, a estabilidade dos modelos depende da capacidade de adaptação a mudanças na distribuição dos dados, conforme discutido na literatura sobre aprendizagem incremental (Martin et al, 2026).

A escolha de variáveis relevantes reduz a dimensionalidade e o custo de inferência, mitiga os efeitos de correlações espúrias e facilita a interpretação física dos sinais, contribuindo também para a qualidade das explicações geradas por métodos baseados em importância de atributos (Rožanec et al, 2021). A identificação e eliminação de atributos redundantes ou excessivamente ruidosos favorece modelos mais estáveis em regime incremental, sobretudo em cenários de dados em fluxo.

Em ambientes de *streaming*, a seleção de variáveis deve ainda considerar o custo computacional por amostra e a necessidade de mecanismos adaptativos capazes de responder à evolução do sinal ao longo do tempo, alinhando-se com princípios de aprendizagem contínua e deteção sob *drift* (Cao et al, 2024; Martin et al, 2026). Estas preocupações são essenciais para preservar latências operacionais aceitáveis sem comprometer a fidelidade das explicações. A presente dissertação adota uma estratégia de engenharia de *features* que prioriza relevância estatística, interpretabilidade e compatibilidade com operação em fluxo contínuo.

3.3.1 Análise de Redundância e Normalização

Foi realizada uma análise exploratória de dependências entre variáveis com o objetivo de identificar potenciais redundâncias e apoiar decisões de pré processamento. Esta análise incidiu sobre grandezas elétricas e mecânicas associadas aos motores, tendo em conta relações físicas conhecidas entre variáveis, nomeadamente aquelas relacionadas com consumo energético, dissipação térmica e esforço mecânico.

Com base nesta análise, não foi aplicada eliminação automática de variáveis correlacionadas. Do ponto de vista metodológico, optou-se por preservar variáveis potencialmente redundantes sempre que existisse justificação física para a sua coexistência, uma vez que alterações nas relações de dependência entre grandezas normalmente correlacionadas podem constituir sinais relevantes de comportamento anómalo. Desta forma, a deteção de anomalias não se baseia apenas em desvios absolutos, mas também em ruturas estruturais nas relações entre variáveis.

A redução de dimensionalidade foi aplicada apenas nos casos em que duas ou mais variáveis apresentavam sobreposição informativa clara, sem distinção funcional relevante para o diagnóstico. Nestas situações, foi privilegiada a variável considerada mais diretamente mensurada, mais estável ou mais representativa do fenómeno físico subjacente, de acordo com critérios de interpretabilidade e robustez operacional.

Antes da aplicação dos algoritmos de deteção de anomalias, todas as variáveis selecionadas foram normalizadas por unidade funcional, isto é, por motor, de forma independente. A normalização foi efetuada através de transformação *z-score* robusta, cujos parâmetros foram estimados exclusivamente com base em dados do regime de funcionamento Normal, assumido como referência nominal do sistema. Esta estratégia assegura que a transformação expressa o desvio relativo face ao comportamento saudável do equipamento, evitando a diluição do conceito de normalidade por observações degradadas ou de falha.

A normalização individualizada garante comparabilidade entre sensores com escalas distintas, estabilidade numérica na aprendizagem dos modelos e equilíbrio no contributo das variáveis para métricas baseadas em distância ou densidade. Esta etapa constitui um pré requisito para a combinação coerente de *scores* no *ensemble* e para a preservação da sensibilidade do sistema à deteção de desvios subtis.

3.3.2 Análise de Componentes Principais como Ferramenta Exploratória

A Análise de Componentes Principais (PCA) é uma técnica estatística de redução dimensional que transforma um conjunto de variáveis possivelmente correlacionadas num novo conjunto de variáveis não correlacionadas, designadas componentes principais. Cada componente principal corresponde a uma combinação linear das variáveis originais e é definida de forma a maximizar a variância explicada, sob a restrição de ortogonalidade em relação às componentes anteriores. As componentes são ordenadas de acordo com a fração de variância total que explicam, permitindo representar os dados num espaço de menor dimensão com perda controlada de informação (Jolliffe et al., 2016).

No presente trabalho, a PCA foi utilizada exclusivamente como ferramenta exploratória, com o objetivo de apoiar a compreensão da estrutura global dos dados e fundamentar decisões metodológicas relacionadas com a seleção e organização das variáveis. A sua aplicação permitiu analisar padrões de variância, relações latentes entre grandezas físicas e possíveis estruturas de agrupamento em projeções de baixa dimensão, sem qualquer impacto direto no pipeline final de deteção de anomalias.

Do ponto de vista metodológico, a utilização da PCA serviu para avaliar como a variância se distribui entre componentes dominadas por variáveis de natureza elétrica, mecânica e operacionais, permitindo aferir se a informação potencialmente discriminativa se concentra nas primeiras componentes ou se se encontra distribuída por múltiplas dimensões. Esta análise forneceu suporte à decisão de não recorrer a redução dimensional agressiva e de preservar o espaço original de atributos no sistema final.

Apesar do seu valor enquanto ferramenta exploratória, a PCA não foi integrada no pipeline de produção por um conjunto de razões metodológicas e operacionais. Em primeiro lugar, a transformação linear das variáveis compromete a interpretabilidade física do sistema, ao substituir grandezas diretamente mensuráveis, como vibração, potência ou temperatura, por combinações abstratas que não possuem significado físico direto. Esta limitação é particularmente relevante em

contextos de manutenção preditiva com requisitos de explicabilidade, nos quais os operadores necessitam de justificações acionáveis e expressas na linguagem do domínio técnico.

Em segundo lugar, a PCA privilegia a captura de variância global, enquanto sinais relevantes para a deteção precoce de degradação podem manifestar-se em componentes de menor variância. Uma redução dimensional agressiva poderia eliminar informação subtil associada a estados intermédios, comprometendo a sensibilidade do sistema à deteção de degradação incipiente, que constitui um dos objetivos centrais deste trabalho.

Em terceiro lugar, a utilização da PCA em cenários de inferência em fluxo contínuo introduz desafios adicionais de estabilidade e adaptação. Trabalhos aplicados a *data streams* demonstram que a aplicação de PCA em ambientes dinâmicos frequentemente requer extensões específicas, como variantes esparsas ou mecanismos espaço-temporais adicionais, para permitir localização robusta de anomalias e maior estabilidade do subespaço projetado (Jiang et al., 2013). Estas adaptações, embora eficazes, acrescentam complexidade estrutural e custo computacional, o que pode revelar-se incompatível com requisitos de baixa latência e implementação em ambientes *edge* com recursos limitados.

Por estas razões, optou-se metodologicamente por manter as variáveis no espaço físico original e trabalhar com um subconjunto reduzido de características, designado *Top K features*, selecionado com base em critérios de relevância estatística, interpretabilidade e compatibilidade com aprendizagem em fluxo contínuo. Esta decisão assegura coerência entre deteção, explicabilidade e viabilidade operacional do sistema proposto.

3.3.3 Métodos de Avaliação de Importância

Para fundamentar de forma objetiva a seleção do subconjunto *Top K*, aplicou-se uma estratégia multi-método que integra três abordagens complementares de avaliação de importância: análise de variância entre regimes, quantificação da dependência informacional relativamente ao estado operacional e avaliação do impacto em *Isolation Forest*. Esta triangulação metodológica garante que as *features* escolhidas são simultaneamente discriminativas do ponto de vista estatístico, informativas quanto à separação de regimes e relevantes para modelos não supervisionados de deteção de anomalias, cumprindo os requisitos de um sistema de manutenção preditiva interpretável e robusto.

O primeiro método baseia-se no teste não paramétrico de *Kruskal–Wallis* (Kruskal & Wallis, 1952), que permite avaliar diferenças estatisticamente significativas nas distribuições de cada variável entre os três regimes operacionais. As variáveis que apresentam estatísticas elevadas neste teste exibem perfis distintos por regime, indicando capacidade discriminativa. Este método é particularmente adequado quando não se assume normalidade das distribuições e permite capturar tanto transições abruptas como padrões de degradação gradual.

O segundo método de avaliação baseia-se na quantificação da dependência entre cada variável e o regime operacional do sistema através de informação mútua, medida formal da redução de incerteza de uma variável aleatória dado o conhecimento de outra (Cover & Thomas, 2006). Esta métrica

permite avaliar até que ponto uma variável contribui para distinguir entre estados operacionais, independentemente da forma funcional da relação envolvida. Ao contrário de métricas clássicas de correlação linear, como o coeficiente de Pearson, a informação mútua é sensível a dependências não lineares e não impõe pressupostos de linearidade ou uniformidade, sendo particularmente adequada a sistemas físicos complexos. Valores elevados indicam que a variável contém informação relevante para a discriminação de regimes, mesmo quando essa relação não se manifesta de forma linear simples.

O terceiro método avalia o impacto de cada variável na capacidade de deteção de anomalias por *Isolation Forest*, o algoritmo base do *ensemble* de deteção proposto neste trabalho (Liu et al., 2008) . *O Isolation Forest* baseia-se no princípio de isolamento recursivo de observações através de partições aleatórias, sendo particularmente eficaz na identificação de padrões raros ou desviantes em espaços de elevada dimensão. Inicialmente, é treinado um modelo *baseline* utilizando o conjunto completo de variáveis, obtendo-se os respetivos *scores* de anomalia. Em seguida, para cada variável, treina-se um novo modelo excluindo apenas esse atributo, mantendo inalteradas as restantes condições de treino. A variação observada nos scores ou na métrica de desempenho relativamente ao *baseline* quantifica o contributo da *feature* removida. Variáveis cuja exclusão provoca degradação significativa na capacidade de isolamento de observações raras recebem valores de importância mais elevados.

Este procedimento estabelece uma ligação direta entre relevância e desempenho efetivo do detetor, privilegiando atributos que contribuem concretamente para a robustez operacional do sistema.

Os três métodos de avaliação foram aplicados ao conjunto completo de vinte e oito variáveis monitorizadas, composto por vinte e quatro variáveis dos motores e quatro variáveis de processo, utilizando dados representativos dos três regimes operacionais. Os valores de importância obtidos por cada método foram normalizados para o intervalo entre zero e um, permitindo a comparação direta entre abordagens que utilizam métricas distintas.

Para combinar os resultados dos diferentes critérios de relevância, foi adotada uma média ponderada com pesos definidos de forma heurística e orientada ao objetivo do estudo. O método baseado em *Isolation Forest* recebeu maior peso relativo, por estar diretamente alinhado com a tarefa principal de deteção não supervisionada de anomalias e por refletir impacto efetivo na capacidade de isolamento de observações raras. As métricas baseadas em análise de variância e dependência informacional receberam pesos iguais e inferiores, uma vez que capturam propriedades estatísticas complementares do sinal, mas não medem diretamente o desempenho operacional do detetor.

Importa salientar que estes pesos não resultam de otimização supervisionada, mas de uma escolha exploratória coerente com o enquadramento não supervisionado e com o objetivo de reduzir viés introduzido por métricas puramente estatísticas. A combinação ponderada integra, assim, diferentes perspetivas sobre relevância, das quais a separação estatística, dependência informacional e capacidade de isolamento, num único valor consolidado que orienta a seleção de atributos para os modelos de deteção desenvolvidos nas secções seguintes.

3.4 Pré-processamento dos Dados

O pré-processamento constitui um dos pilares estruturais do sistema de manutenção preditiva desenvolvido, assegurando robustez analítica em contexto de fluxo contínuo e operação em tempo real. Dado que os modelos de deteção operam sobre dados sequenciais recebidos continuamente, esta etapa introduz mecanismos destinados a mitigar variabilidade entre sensores, ruído de medição e heterogeneidade operacional entre unidades funcionais.

Todas as operações de pré-processamento foram concebidas para funcionamento incremental, compatível com o paradigma de inferência em fluxo adotado neste trabalho (Barry et al, 2020). Cada nova observação é processada de forma independente, sem necessidade de acesso ao histórico completo dos dados, preservando viabilidade computacional e latências reduzidas, mesmo em ambientes *edge* com recursos limitados (Jirwe, 2021).

O desenho do pipeline reflete decisões metodológicas sustentadas pelas análises exploratórias descritas anteriormente e por critérios de eficiência e interpretabilidade. A heterogeneidade operacional entre motores justifica a normalização individualizada por unidade funcional, enquanto a seleção de um subconjunto reduzido de atributos relevantes define o espaço de entrada do sistema. Esta restrição dimensional permite equilibrar poder discriminativo, estabilidade estatística e custo computacional.

A arquitetura do módulo de pré-processamento encontra-se ilustrada na Figura 3. O fluxo inicia-se com a aquisição das *features* previamente selecionadas, seguida de uma etapa de normalização incremental individualizada por motor. Posteriormente, são aplicados mecanismos de tratamento de valores extremos e estratégias de robustez estatística destinadas a mitigar a influência de ruído e variações operacionais abruptas.

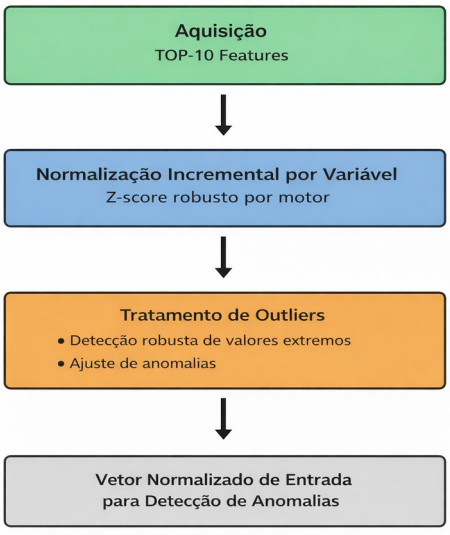

Figura 3: Pipeline do Pré-Processamento

Estas transformações são aplicadas sequencialmente a cada nova observação recebida em fluxo, garantindo consistência temporal, estabilidade numérica e compatibilidade com restrições de latência típicas de ambientes industriais. O resultado desta etapa corresponde a um vetor de características normalizado, de forma robusta, que constitui a entrada para as fases subsequentes do sistema.

3.4.1 Normalização Incremental por Variável

A normalização das variáveis monitorizadas constitui uma etapa fundamental do pré processamento, assegurando que diferenças de escala e de unidades entre grandezas físicas não enviesam os modelos de deteção de anomalias. Os dados analisados apresentam heterogeneidade operacional significativa entre motores e entre diferentes tipos de variáveis, nomeadamente corrente, potência, temperatura e vibração, o que inviabiliza a utilização direta de valores brutos. Sem normalização, variáveis de maior magnitude tenderiam a dominar métricas de distância e densidade utilizadas pelos detetores, enquanto variáveis de menor escala teriam influência reduzida ou negligenciável (Almeida et al, 2023).

Para mitigar este efeito, adotou-se uma estratégia de normalização individualizada por variável. Cada variável é normalizada através de uma transformação *z score*, conforme apresentado na Equação 7.

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \qquad (7)$$

onde x_i representa o valor observado da *feature* i, μ_i e σ_i são a média e o desvio padrão dessa variável estimados a partir do *baseline*, e z_i é o valor normalizado resultante. Esta transformação produz variáveis padronizadas com média nula e variância unitária, tornando comparáveis grandezas de magnitudes e dispersões distintas e assegurando estabilidade numérica nos modelos subsequentes.

Os parâmetros de normalização (μ_i e σ_i) são estimados, exclusivamente a partir de dados do regime operacional normal, utilizando o subconjunto de *baseline* identificado na fase de inicialização do sistema (Bäßler et al, 2022). Esta escolha assegura que a referência de normalidade reflete o comportamento típico de cada variável, em condições operacionais nominais, sem contaminação por padrões de degradação ou falha. Durante a operação em fluxo contínuo, os parâmetros são mantidos fixos, preservando a coerência da escala ao longo do tempo e permitindo que desvios, face aos valores de referência, sejam capturados como oscilações nos valores normalizados.

Em situações excecionais, algumas variáveis podem apresentar variância extremamente reduzida ou praticamente nula no conjunto de referência, conduzindo a valores de desvio padrão próximos de zero. Nestes casos, a aplicação direta da normalização por *z-score* originaria divisões por valores muito pequenos, resultando em instabilidade numérica e amplificação artificial de ruído.

Para evitar este efeito, é imposto um limiar mínimo para o desvio padrão, substituindo-o por um valor unitário sempre que este seja inferior a 10^{-12}. Este valor foi escolhido por se situar várias ordens de grandeza abaixo da escala típica das medições físicas envolvidas, garantindo que a normalização permanece estável sem introduzir distorções relevantes na magnitude relativa das variáveis.

Esta salvaguarda tem natureza exclusivamente numérica e não afeta a interpretação estatística dos dados, assegurando robustez computacional em cenários de variância quase nula.

A normalização por variável preserva ainda a interpretabilidade física das mesmas. Um valor normalizado elevado indica um desvio significativo, relativamente ao comportamento típico da variável no regime normal, independentemente da unidade física original. Esta representação facilita a calibração de limiares de deteção, a combinação de *scores* entre modelos heterogéneos e a comunicação de alertas a operadores, suportando decisões operacionais informadas em tempo real.

3.4.2 Tratamento de Outliers e Robustez

A presença de *outliers* e medições incorretas em fluxos de dados sensoriais constitui um desafio operacional relevante em sistemas de manutenção preditiva. Estes valores podem resultar de falhas transitórias de comunicação, interferências eletromagnéticas, ruído impulsivo ou saturação momentânea de sensores. Embora alguns *outliers* correspondam a eventos anómalos genuínos que o sistema deve identificar, valores extremos isolados podem enviesar significativamente modelos sensíveis a distâncias, como *Local Outlier Factor* ou métodos baseados em vizinhança, ou ainda comprometer a estabilidade de estimativas estatísticas utilizadas em certos detetores (Rožanec et al, 2021).

A abordagem adotada procura conciliar robustez estatísticas com a preservação de informação potencialmente relevante para o diagnóstico. Em vez de descartar observações consideradas atípicas, aplica-se uma operação de *clipping* aos dados previamente normalizados, limitando os valores normalizados ao intervalo [−C,C], com C=8. Este valor foi escolhido por corresponder a um desvio extremamente improvável sob a hipótese de normalidade, com probabilidade $P(|Z|>8)<10^{-15}$ numa distribuição Gaussiana, assegurando que apenas valores extremos e potencialmente espúrios são afetados.

A operação de *clipping* é definida formalmente pela Equação 8:

$$z_i^{clip} = \begin{cases} -C & \text{se } z_i < -C \\ z_i & \text{se } -C \le z_i \le C \\ C & \text{se } z_i > C \end{cases} \qquad (8)$$

Esta transformação impede que medições pontuais extremas dominem os scores de anomalia e influenciem de forma desproporcionada os mecanismos de deteção. Em simultâneo, preserva integralmente desvios moderadamente elevados, tipicamente no intervalo $3<|z_i|<8$, que podem constituir indicadores legítimos de degradação ou falha incipiente.

A escolha conservadora de C=8 garante que eventos de degradação e falha observados no conjunto de dados, caracterizados por desvios na ordem de três a seis desvios padrão nas variáveis mais sensíveis, como vibração ou corrente elétrica, não são artificialmente truncados. Deste modo, o *clipping* contribui para a estabilidade numérica do pipeline e para a robustez dos detetores, sem comprometer a sensibilidade à deteção de anomalias relevantes em contexto operacional.

Concluídas as etapas de normalização, tratamento de valores extremos e verificação de robustez numérica, cada observação é representada por um vetor de características consistente, alinhado temporalmente e pronto a ser fornecido aos modelos de deteção de anomalias. O resultado do pipeline de pré-processamento consiste, assim, num fluxo incremental de instâncias transformadas, garantindo estabilidade estatística, comparabilidade entre variáveis e adequação às exigências de inferência em tempo real. Na secção seguinte são descritos os modelos de deteção aplicados a este fluxo processado.

3.5 Modelos de Deteção de Anomalias

Esta secção descreve em detalhe os seis algoritmos de deteção de anomalias que compõem a arquitetura final do sistema proposto. Para cada modelo são apresentados o princípio de funcionamento, a fundamentação conceptual, os principais aspetos de implementação, bem como as vantagens e limitações no contexto específico da manutenção preditiva em tempo real. A seleção dos modelos foi orientada por critérios de complementaridade algorítmica, eficiência computacional, compatibilidade com operação em *streaming* e contributo interpretativo para o *ensemble*.

Os modelos incluídos abrangem diferentes definições matemáticas de anomalia, nomeadamente isolamento, densidade local, fronteiras de decisão não lineares, protótipos de clusterização e erro de previsão temporal. Esta diversidade é essencial para capturar manifestações heterogéneas de degradação e falha, típicas de sistemas industriais reais, onde os desvios podem assumir formas globais, locais, multivariadas ou temporais.

3.5.1 *Isolation Forest*

O *Isolation Forest* é um método não supervisionado de deteção de anomalias que se baseia no princípio de que observações anómalas tendem a ser isoladas com menor número de partições aleatórias do que observações normais (Liu et al., 2008). Em vez de estimar explicitamente a distribuição dos dados, o algoritmo constrói um conjunto de árvores de isolamento, nas quais cada nó realiza um corte aleatório numa única variável, dentro do intervalo observado.

O processo de particionamento repete-se recursivamente até que cada observação fique isolada num nó terminal ou até que seja atingido um critério de paragem. A profundidade do nó terminal constitui uma medida indireta A subfigura (a) representa a relação entre a profundidade normalizada e o score de anomalia, evidenciando a correspondência funcional entre estas duas grandezas. A subfigura (b) apresenta a distribuição dos *scores* por regime operacional, permitindo observar a sua dispersão relativa entre classes. Estas visualizações têm caráter descritivo e ilustrativo do comportamento do método no conjunto de dados considerado. Do ponto de vista metodológico, o *Isolation Forest*

caracteriza-se por elevada eficiência computacional, ausência de pressupostos distribucionais explícitos e boa escalabilidade para conjuntos de dados de grande dimensão (Liu et al., 2008; Rosenberger et al., 2023). Entre as suas limitações destacam-se a menor sensibilidade a anomalias locais em regiões de elevada densidade e a variabilidade inerente ao processo de particionamento aleatório, que pode exigir ajuste adequado do número de estimadores para estabilização dos scores.

No contexto do *ensemble* proposto, o *Isolation Forest* desempenha o papel de detetor global baseado em particionamento aleatório, complementando métodos baseados em densidade local, fronteiras explícitas e dependências temporais, enquanto observações mais consistentes com o comportamento global tendem a apresentar trajetórias mais longas.

A profundidade média de isolamento ao longo do conjunto de árvores, devidamente normalizada em função do tamanho da amostra, é convertida num *score* contínuo de anomalia. Este *score* permite ordenar as observações segundo o seu grau relativo de afastamento ao padrão predominante no conjunto de treino.

A Figura 4 apresenta duas visualizações ilustrativas obtidas a partir dos dados experimentais desta dissertação

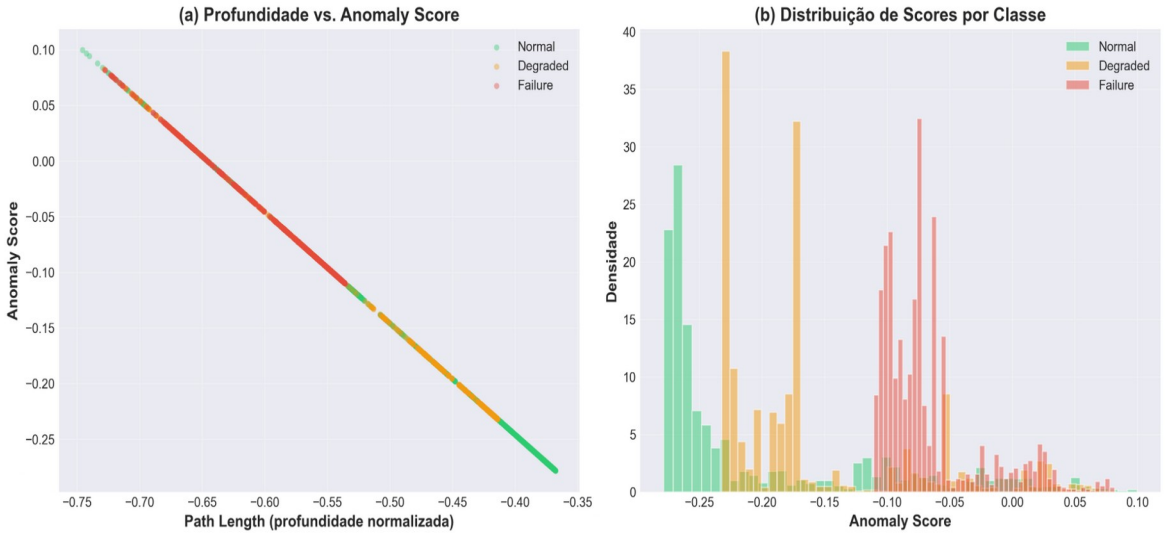

Figura 4: Profundidade de Isolamento

A subfigura (a) representa a relação entre a profundidade normalizada e o *score* de anomalia, evidenciando a correspondência funcional entre estas duas grandezas. A subfigura (b) apresenta a distribuição dos pesos por regime operacional, permitindo observar a sua dispersão relativa entre classes. Estas visualizações têm caráter descritivo e ilustrativo do comportamento do método no conjunto de dados considerado. Do ponto de vista metodológico, o IF caracteriza-se por elevada eficiência computacional, ausência de pressupostos distribucionais explícitos e boa escalabilidade para conjuntos de dados de grande dimensão (Liu et al., 2008; Rosenberger et al., 2023). Entre as

suas limitações destacam-se a menor sensibilidade a anomalias locais em regiões de elevada densidade e a variabilidade inerente ao processo de particionamento aleatório, que pode exigir ajuste adequado do número de estimadores para estabilização dos *scores*.

No contexto do *ensemble* proposto, o modelo desempenha o papel de detetor global baseado em particionamento aleatório, complementando métodos baseados em densidade local, fronteiras explícitas e dependências temporais.

3.5.2 Local Outlier Factor

O *Local Outlier Factor* (LOF) é um método não supervisionado de deteção de anomalias baseado em densidade local, concebido para identificar observações cujo grau de isolamento é elevado relativamente à sua vizinhança imediata (Breunig et al., 2000) . Ao contrário de detetores globais, que avaliam desvios face à distribuição geral dos dados, o LOF compara a densidade local de cada observação com a densidade média dos seus k vizinhos mais próximos, permitindo a deteção de anomalias contextuais inseridas em regiões heterogéneas do espaço de características.

Formalmente, o LOF estima densidades locais recorrendo à distância de proximidade *(reachability distance*), que suaviza variações abruptas nas distâncias euclidianas. O score resultante assume valores próximos de 1 para observações consistentes com a densidade da sua vizinhança, enquanto valores superiores a 1 indicam potencial anomalia local. Esta definição torna o método particularmente eficaz na deteção de desvios que não se manifestam como *outliers* globais, mas como observações raras em sub-regiões específicas do espaço.

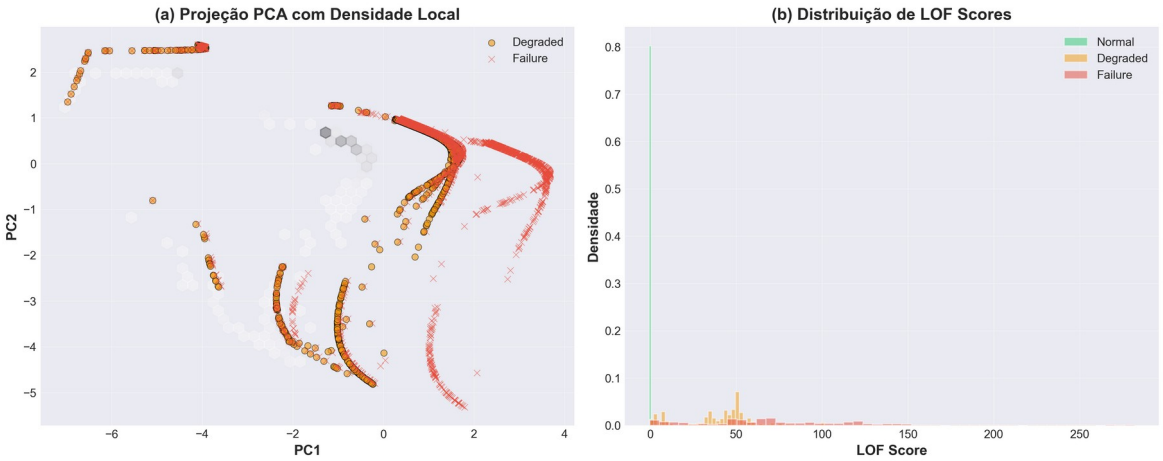

Figura 5: Densidade Local vs Score

A Figura 5 ilustra o comportamento típico do LOF aplicado ao conjunto de dados experimental, permitindo observar a distribuição dos *scores* por regime. Na projeção bidimensional por *PCA* com codificação de densidade local (Figura 5a), observa-se que as observações normais se concentram em regiões densas do espaço latente, enquanto os estados degradados ocupam zonas intermédias e as falhas tendem a surgir em regiões claramente mais dispersas. A distribuição dos LOF *scores* por

classe (Figura 5b) reforça esta interpretação, mostrando uma separação progressiva entre Normal, Degradado e Falha, ainda que com sobreposição nas zonas intermédias. Esta sobreposição evidencia a necessidade de calibração cuidadosa de limiares e de integração do LOF com outros detetores para decisões robustas.

Do ponto de vista computacional, o custo dominante do LOF reside na busca dos k vizinhos mais próximos. Embora a implementação elementar apresente complexidade quadrática, o uso de estruturas de indexação adequadas permite reduzir o custo médio para aproximadamente O(n log n) (Breunig et al., 2000). . A memória cresce linearmente com o número de observações e com a dimensionalidade do espaço de características. Em dimensões elevadas, a eficácia das métricas de distância tende a degradar-se, fenómeno conhecido como *curse of dimensionality*, o que limita a discriminabilidade do método isoladamente (Jolliffe & Cadima, 2016).

As principais vantagens do LOF residem na sua capacidade de detetar anomalias locais e de adaptar-se a distribuições multimodais e heterogéneas, características frequentes em sistemas industriais reais. Em contrapartida, o método é sensível à escolha de k, depende fortemente de um baseline limpo e apresenta limitações em espaços de elevada dimensão. Em termos operacionais, recomenda-se calibrar limiares por motor e por regime com curvas ROC (Curva de Característica de Operação do Receptor) e *precision-recall*, validar a escolha de k através de validação temporal e bootstrap, adotar políticas de confirmação temporal e por múltiplos sensores e integrar o LOF num ensemble com detetores globais e modelos temporais. Esta integração permite explorar o seu contributo específico na deteção de desvios contextuais, mitigando simultaneamente as suas limitações e reforçando a robustez global do sistema.

3.5.3 One-class SVM

O *One-Class Support Vector Machine* (*OCSVM*) é um método não supervisionado de deteção de anomalias que modela explicitamente a região associada ao comportamento normal, aprendendo uma fronteira de decisão que envolve a maioria das observações do *baseline* nominal. O treino é realizado exclusivamente com dados representativos do regime normal (Almeida et al, 2023).

O método recorre a funções *kernel* para projetar os dados num espaço transformado onde relações não lineares se tornam separáveis. Neste trabalho é considerado o *kernel radial* de base gaussiana (RBF), que permite modelar fronteiras não lineares no espaço original das variáveis. O parâmetro γ controla a largura do *kernel* e, consequentemente, a complexidade da fronteira aprendida, enquanto o parâmetro ν define um limite superior para a fração de observações do treino admitidas como desvios, regulando o compromisso entre inclusão do comportamento nominal e tolerância a variações(Li et al, 2023).

A Figura 6 apresenta uma projeção bidimensional por *PCA* do espaço de características, juntamente com a fronteira de decisão induzida pelo *OCSVM*. A linha tracejada representa a separação aprendida entre a região considerada normal e as regiões exteriores Embora a projeção reduza a dimensionalidade original, a visualização ilustra qualitativamente o comportamento do modelo, evidenciando a capacidade do *kernel RBF* para capturar estruturas não lineares no espaço de dados.

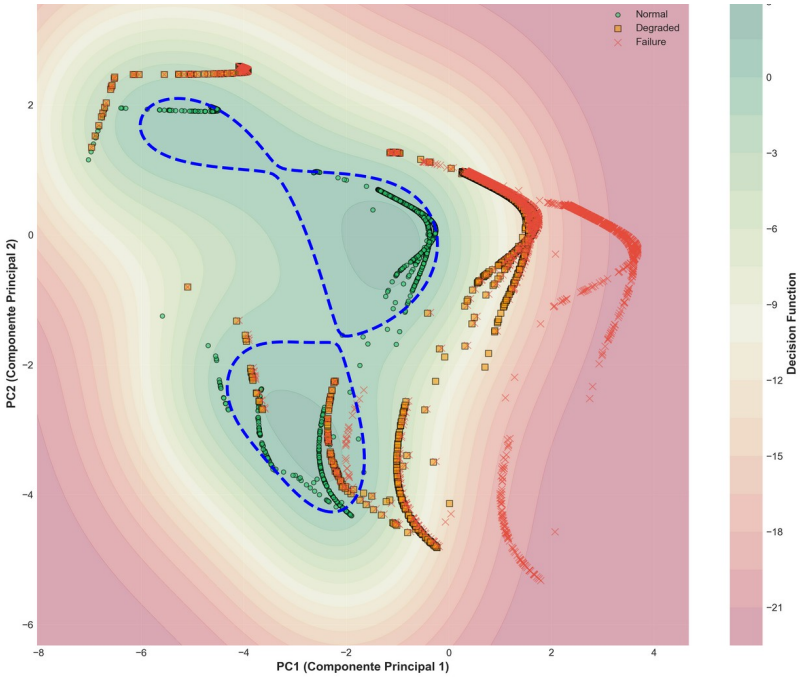

Figura 6: Fronteira de Decisão (Projeção PCA)

Entre as principais vantagens deste método destacam-se a sua elevada expressividade para modelar fronteiras complexas e a fundamentação teórica sólida baseada em otimização convexa. Entre as limitações figuram a sensibilidade à escolha dos hiperparâmetros, o custo computacional do treino em conjuntos de grande dimensão e a reduzida interpretabilidade da fronteira no espaço induzido pelo kernel (Almeida et al, 2023). No contexto do ensemble proposto, o OCSVM contribui com uma perspetiva baseada em fronteira global do comportamento nominal, complementando métodos baseados em densidade, isolamento e dependência temporal.

3.5.4 K-Means

O *K-Means*, embora concebido originalmente como um algoritmo de *clustering* não supervisionado, pode ser eficazmente reutilizado como detetor de anomalias ao interpretar a distância de cada observação ao ponto representativo mais próximo, como um *score* contínuo de anomalia. Formalmente, para uma observação x, o *score* é definido como:

$$s(x) = min_j \left|\left| x - c_j \right|\right|^2 \qquad (11)$$

onde c_j representa o centróide do *cluster* j. Esta formulação assenta na premissa de que o comportamento normal tende a formar protótipos compactos no espaço de características, enquanto observações anómalas se manifestam como desvios com maior distância relativamente a esses protótipos.

O número de *clusters* foi definido como k=8, com base em análise exploratória preliminar e critérios de estabilidade geométrica dos centróides. Esta escolha procura equilibrar granularidade representativa dos diferentes regimes operacionais com robustez estatística dos protótipos aprendidos.

A Figura 7 apresenta uma visualização bidimensional por PCA dos *clusters* obtidos a partir dos dados experimentais desta dissertação, incluindo a localização dos centróides e a dispersão das observações em torno destes. A visualização tem caráter ilustrativo e permite observar a estrutura de agrupamento no espaço latente, bem como a distribuição relativa dos diferentes regimes operacionais.

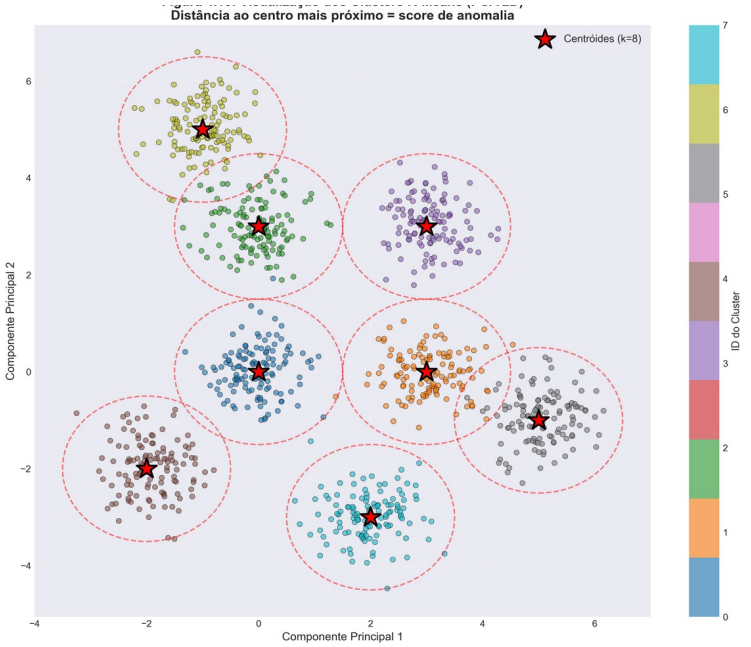

Figura 7: Visualização dos Clusters K-Means (PCA 2D)

Do ponto de vista metodológico, o *K-Means* apresenta simplicidade conceptual, interpretabilidade direta dos centróides como protótipos representativos e facilidade de integração num esquema de *ensemble* através de um score contínuo baseado em distância (Barbariol, 2023; Asutkar & Tallur, 2023) .

As principais vantagens do *K-Means* neste enquadramento residem na sua simplicidade conceptual, elevada interpretabilidade e eficiência computacional. Os *clusters centers* podem ser inspecionados diretamente, associados a estados físicos do sistema e utilizados como protótipos representativos para diagnóstico, o que facilita a comunicação dos resultados a engenheiros e operadores. Adicionalmente, o método fornece um *score* contínuo de fácil calibração e integração em estratégias de agregação no *ensemble*.

Por outro lado, o *K-Means* apresenta limitações estruturais relevantes. O algoritmo assume implicitamente *clusters* aproximadamente esféricos e de variância semelhante, o que pode não

refletir fielmente distribuições reais de dados industriais. É igualmente sensível à escala das variáveis, tornando indispensável uma normalização prévia consistente, e a métrica de distância ao centróide não incorpora informação explícita sobre densidade local, podendo conduzir a falsos negativos em regiões naturalmente dispersas mas ainda normais (Mozaffari et al., 2022).

3.5.5 *Predictive Lag 1 (Ridge Regression)*

O *Predictive Lag-1* é o único modelo do *ensemble* que explora explicitamente a dependência temporal entre observações consecutivas. Parte-se do princípio de que, em funcionamento normal, o estado do sistema num instante pode ser bem estimado a partir do instante imediatamente anterior. Quando a diferença entre o valor observado e o valor previsto aumenta de forma significativa, isso indica uma rutura na dinâmica normal do sistema, frequentemente associada ao início de degradação ou à aproximação de uma falha (Yan, 2019; Gomes et al., 2019)

Para operacionalizar esta ideia, é treinado um regressor linear regularizado (*Ridge Regression*) que estima o vetor de observações no instante t a partir do vetor observado em t−1, segundo a relação expressa na equação 12,

$$\hat{x}(t) = Wx(t-1) + b \qquad (12)$$

onde W representa a matriz de pesos e b o termo de bias. Os parâmetros são ajustados minimizando o erro quadrático regularizado por norma L2, o que controla a complexidade do modelo e previne *overfitting* em presença de colinearidade entre variáveis sensoriais (Yan, 2019) .

Como *score* de anomalia, utiliza-se o erro quadrático médio de previsão por instância, expresso como o *RMSE* entre o vetor observado e o vetor previsto:

$$s(t) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}\left(x_i(t) - \hat{x}_i(t)\right)^2} \qquad (13)$$

Valores elevados deste erro indicam discrepâncias relevantes entre a dinâmica esperada e o comportamento observado.

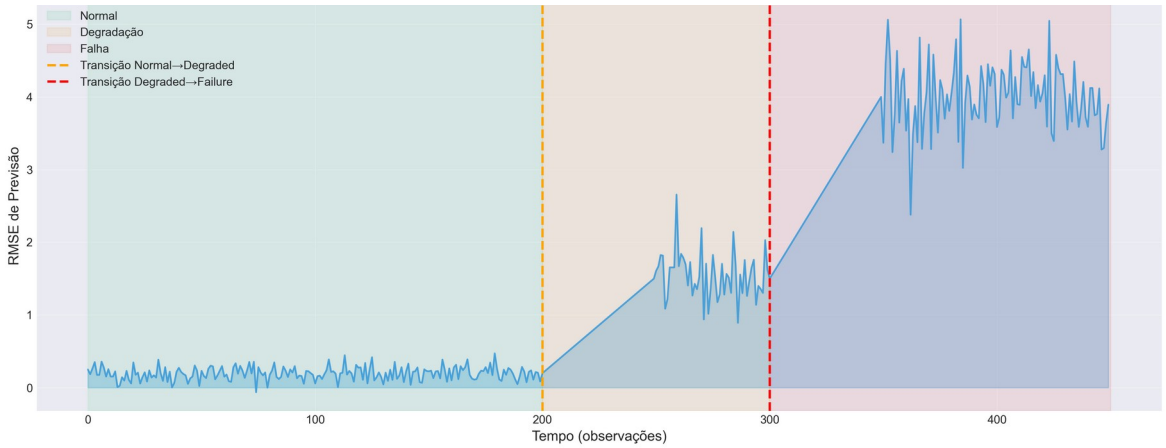

Figura 8: Erro de Previsão Temporal

A Figura 8 apresenta uma representação temporal do erro de previsão ao longo das diferentes fases operacionais do sistema. Observam-se variações no *RMSE* associadas a mudanças de regime, evidenciando a sensibilidade do modelo a alterações na dinâmica subjacente, mesmo quando os valores absolutos das variáveis permanecem dentro de intervalos aparentemente normais.

Entre as principais vantagens do *Predictive Lag-1* destacam-se a capacidade de capturar anomalias de natureza temporal que escapam a métodos puramente espaciais, a interpretabilidade direta do *RMSE* como medida de desvio dinâmico e a robustez conferida pela regularização *Ridge*. Entre as limitações figuram a impossibilidade de avaliação na primeira observação do fluxo, a sensibilidade a ruído de curto prazo (mitigada neste trabalho por suavização temporal subsequente) e a hipótese implícita de linearidade, que pode não modelar adequadamente dinâmicas fortemente não lineares.

No contexto do *ensemble*, o *Predictive Lag-1* atua como um detetor complementar aos métodos baseados em densidade, distância ou fronteiras de decisão, acrescentando uma perspetiva temporal explícita à análise de anomalias.

3.5.6 Half Space Trees

As *Half Space Trees* constituem um algoritmo de deteção de anomalias concebido especificamente para aprendizagem em fluxo contínuo, combinando particionamento aleatório do espaço de atributos com atualização incremental das estruturas internas. Cada árvore divide recursivamente o espaço através de hiperplanos aleatórios até uma profundidade fixa, sendo que cada nó acumula a massa de observações que por ele passam ao longo do tempo (Gomes et al, 2019) . O *score* de anomalia de uma observação é definido de forma inversamente proporcional à massa da folha onde essa observação é projetada, pelo que folhas com baixa massa correspondem a regiões esparsas do espaço e são interpretadas como potenciais anomalias.

Na implementação adotada recorreu se à biblioteca *River*, utilizando um *ensemble* de árvores com profundidade controlada e inicialização aleatória. O algoritmo opera segundo o paradigma *test then train*, isto é, para cada nova observação o *score* de anomalia é calculado com base no estado atual do modelo e apenas depois a observação é utilizada para atualizar as massas das árvores (Gomes et al., 2019; Rožanec et al., 2021) . Esta estratégia permite realizar inferência e aprendizagem num único ciclo, garantindo adaptação contínua ao fluxo de dados sem necessidade de retreino em *batch*.

A Figura 9 ilustra o comportamento dinâmico das *Half Space Trees* ao longo do tempo. A comparação entre as distribuições de pesos antes e após o período inicial de aprendizagem evidencia a estabilização progressiva das massas acumuladas nas folhas. Adicionalmente, a evolução temporal do *score* médio mostra a influência do período inicial de adaptação na sensibilidade do detetor, refletindo a natureza incremental do método.

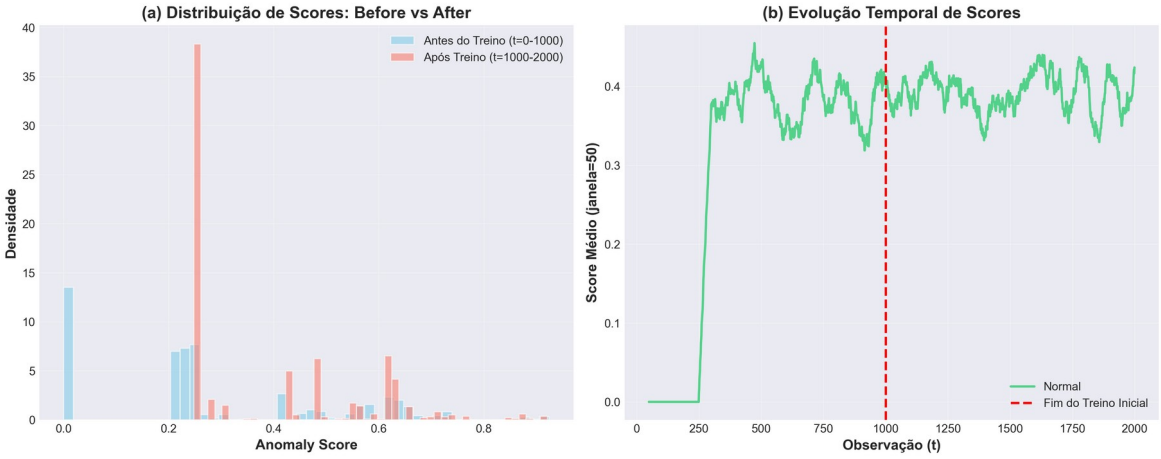

Figura 9: Distribuição e evolução temporal de scores de anomalia (HST)

Entre as principais vantagens destacam-se a verdadeira incrementalidade, que elimina a necessidade de retreino periódico, e a capacidade de adaptação gradual a alterações na distribuição dos dados (Rožanec et al., 2021) . Entre as limitações incluem-se a instabilidade inicial dos *scores* até ser acumulada massa suficiente nas folhas, a sensibilidade aos parâmetros estruturais e a menor interpretabilidade interna quando comparada com métodos baseados em protótipos ou fronteiras explícitas.

No contexto do *ensemble* proposto, as *Half Space Trees* introduzem uma componente adaptativa orientada a fluxo, complementando os detetores treinados em *batch* e contribuindo para aumentar a diversidade de perspetivas na avaliação de anomalia.

3.5.7 Comparação dos modelos de deteção de anomalias

Os modelos que compõem o *ensemble* adotam uma estratégia híbrida que combina treino batch sobre um *baseline* estável com inferência em *streaming*. Esta abordagem é amplamente utilizada em

manutenção preditiva industrial por equilibrar robustez estatística, eficiência computacional e viabilidade operacional (Almeida et al, 2023; Ucar et al, 2024; Murtaza et al, 2024).

Tabela 2: Síntese funcional dos modelos do ensemble de deteção de anomalias

Modelo	Paradigma	Manifestação de Anomalia	Complexidade Inferência	Interpretabilidade	Capacidade de Adaptação em Streaming
Isolation Forest	Particionamento	Global	O(log n)	Média	Não (*batch*)
Local Outlier Factor	Densidade local	Local	O(k log n)	Média-Alta	Não (*batch*)
One-Class SVM	Fronteira não linear	Contextual	O(n_sv·d)	Baixa	Não (*batch*)
K-Means	Protótipos	Distância a *clusters*	O(k·d)	Alta	Parcial (*mini-batch*)
Predictive Lag-1	Previsão temporal	Temporal	O(d²)	Média-Alta	Não (*batch*)
Half-Space Trees	Particionamento incremental	Global + *Drift*	O(T log h)	Baixa-Média	Sim (*online*)

Após o treino inicial, todos os modelos operam em modo de inferência incremental, processando observações individualmente com latência reduzida e sem necessidade de acesso ao histórico completo dos dados. Esta característica permite integrar o pipeline em cenários de monitorização contínua em tempo real, mantendo previsibilidade computacional (Biikes et al., 2024; Fragkoulis et al., 2024).

As *Half-Space Trees* constituem a única componente verdadeiramente online do *ensemble*, atualizando a sua estrutura interna a cada nova observação segundo o paradigma *test-then-train* (Gomes et al., 2019; Rožanec et al., 2021). A sua inclusão funciona como mecanismo de adaptação a deriva gradual de longo prazo, enquanto os restantes modelos preservam a memória do *baseline* original. Esta redundância estratégica aumenta a robustez global do sistema, mitigando o risco de degradação progressiva do desempenho em contextos de operação prolongada.

Em ambientes industriais de longa duração, recomenda-se adicionalmente o retreino periódico *offline* dos modelos *batch* sobre janelas temporais recentes classificadas como normais (Nsor, 2024; Koch et al., 2024). Este processo não interfere com a inferência em tempo real, mas constitui uma prática de manutenção preventiva do próprio modelo, garantindo alinhamento com a evolução lenta do sistema físico.

De forma global, a estratégia híbrida adotada combina a qualidade estatística de modelos *batch* calibrados com a adaptabilidade incremental das *Half-Space Trees*, resultando numa solução pragmática, robusta e alinhada com as restrições operacionais típicas de sistemas industriais de manutenção preditiva.

3.6 *Ensembles* na Deteção de Anomalias

Esta secção descreve a estratégia adotada para combinar múltiplos detetores de anomalias e converter as suas saídas individuais numa decisão operacional única (Cabrera Martin et al., 2025; Almeida et al., 2023). A abordagem foi concebida para ambientes de monitorização em tempo real, caracterizados por dados em fluxo contínuo, ausência de rótulos durante o treino e elevada exigência de robustez face a ruído e variação operacional.

A estratégia integra quatro componentes principais: a utilização de um *ensemble* heterogéneo de detetores, a normalização robusta das evidências produzidas, a agregação por consenso com filtragem temporal e um mecanismo de decisão baseado em limiares calibrados de forma adaptativa.

3.6.1 *Ensemble* Heterogéneo

A adoção de um ensemble heterogéneo de detetores de anomalias fundamenta-se, em primeiro lugar, na complementaridade entre diferentes paradigmas de deteção, permitindo capturar desvios de natureza diversa num único sistema (Cao et al., 2025). Métodos baseados em isolamento e partição revelam-se eficazes na identificação de *outliers* globais, enquanto técnicas de vizinhança são particularmente adequadas à deteção de anomalias locais. De forma complementar, abordagens multivariadas e baseadas em fronteira permitem capturar anomalias contextuais, ao passo que modelos preditivos evidenciam desvios temporais face ao comportamento esperado. Adicionalmente, algoritmos baseados em protótipos possibilitam a identificação de padrões anómalos de natureza coletiva. Esta diversidade de perspetivas é especialmente relevante em manutenção preditiva, onde os processos de degradação podem manifestar-se de forma heterogénea ao longo do tempo e através de diferentes variáveis do sistema.

Em segundo lugar, a combinação de múltiplos algoritmos contribui para mitigar o viés inerente a cada método individual, uma vez que cada técnica incorpora pressupostos específicos sobre a estrutura dos dados. Alguns modelos assumem fronteiras suaves no espaço de características, outros baseiam-se em representações por protótipos ou em densidade local, enquanto outros exploram mecanismos de isolamento ou continuidade temporal. Ao integrar detetores com hipóteses distintas, reduz-se a probabilidade de falhas sistemáticas, permitindo que limitações de um método sejam compensadas pela sensibilidade de outros a padrões complementares (Cabrera Martin et al., 2025; Almeida et al., 2023).

Por fim, a agregação por consenso constitui um complemento natural à adoção de um *ensemble* heterogéneo. A combinação das evidências individuais reforça desvios corroborados por múltiplos detetores e atenua respostas isoladas associadas a ruído ou a sensibilidades específicas de um único

modelo. Esta estratégia é particularmente adequada a contextos industriais, caracterizados por elevada heterogeneidade operacional, incerteza quanto aos modos de falha e uma assimetria clara de custos, na qual os falsos negativos tendem a ser mais penalizadores do que os falsos positivos (Verma et al., 2022). . Assim, a introdução de redundância e diversidade visa aumentar a robustez e a fiabilidade do sistema de deteção em cenários de operação prolongada.

3.6.2 Normalização dos Indicadores de Anomalias

Os diferentes detetores que compõem o *ensemble* produzem indicadores de anomalia com naturezas matemáticas, escalas numéricas e distribuições estatísticas distintas. Enquanto alguns métodos se baseiam em profundidade de isolamento, outros recorrem a densidade local, distância a protótipos ou erro de previsão temporal. Esta heterogeneidade inviabiliza a combinação direta dos valores brutos, uma vez que indicadores de maior magnitude tenderiam a dominar a decisão agregada, comprometendo o princípio de consenso que sustenta a estratégia de *ensemble* (Verma et al., 2022; Cao et al., 2025).

Para garantir comparabilidade entre os indicadores produzidos pelos diferentes modelos, é adotada uma normalização prévia que transforma cada *score* para uma escala comum. Em vez da normalização clássica baseada em média e desvio padrão, opta-se por uma abordagem robusta baseada na mediana e no desvio absoluto mediano (*Median Absolute Deviation*, MAD) (Cover & Thomas, 2006; Molnar et al., 2020; Jolliffe & Cadima, 2016). Esta escolha é motivada pela maior resistência destas estatísticas à presença de valores extremos ou anomalias residuais no conjunto de referência, evitando distorções na escala normalizada.

Dado um conjunto de *scores* S={s1,…,sn} obtidos para um determinado detetor em regime nominal, define-se a mediana como:

$$\tilde{s} = median(S)$$

A partir dos valores S produzidos por um determinado detetor em regime nominal, a mediana é definida como *median*(S). A MAD é calculada como a mediana dos desvios absolutos relativamente a essa mediana, de acordo com a Equação 14:

$$MAD = median\left(\left|s - median(S)\right|\right) \qquad (14)$$

Para tornar esta medida comparável ao desvio padrão no caso de distribuições aproximadamente Gaussianas, a MAD é multiplicada por um fator de escala igual a 1.4826 (Jolliffe & Cadima, 2016). A normalização robusta de um valor individual s é então definida pela Equação 15:

$$z = \frac{s - median(S)}{MAD \cdot 1.4826} \qquad (15)$$

O valor resultante pode ser interpretado como o número de desvios robustos relativamente ao comportamento típico observado em funcionamento nominal. Valores próximos de zero indicam consistência com o padrão normal, enquanto valores de maior magnitude correspondem a desvios progressivamente mais severos.

A normalização é realizada em duas fases distintas. Numa primeira fase, a calibração, a mediana e a MAD escalada são estimadas exclusivamente a partir de dados representativos do regime nominal, estabelecendo uma referência estável de normalidade. Estes parâmetros permanecem fixos durante a fase de operação.

Na fase de inferência em *streaming*, cada nova observação é processada individualmente. Cada detetor produz o seu *score* bruto, que é transformado utilizando os parâmetros previamente estimados. Para garantir estabilidade numérica e evitar amplificação excessiva de desvios, são aplicados limites de saturação simétricos (tipicamente no intervalo [−8,8]), restringindo valores extremos que possam resultar de dispersão muito reduzida (Molnar et al., 2020). .

Do ponto de vista metodológico, esta normalização robusta constitui um elemento central do sistema proposto. Ao assegurar escalas homogéneas entre detetores heterogéneos e ao reduzir a influência de valores extremos, esta etapa permite uma agregação equilibrada dos indicadores, preservando a diversidade informativa dos modelos e reforçando a estabilidade global do *ensemble*.

A Figura 10 apresenta a comparação entre os pesos brutos produzidos pelos diferentes detetores e os respetivos *scores* após normalização robusta, evidenciando a redução da heterogeneidade de escala e a obtenção de distribuições diretamente comparáveis.

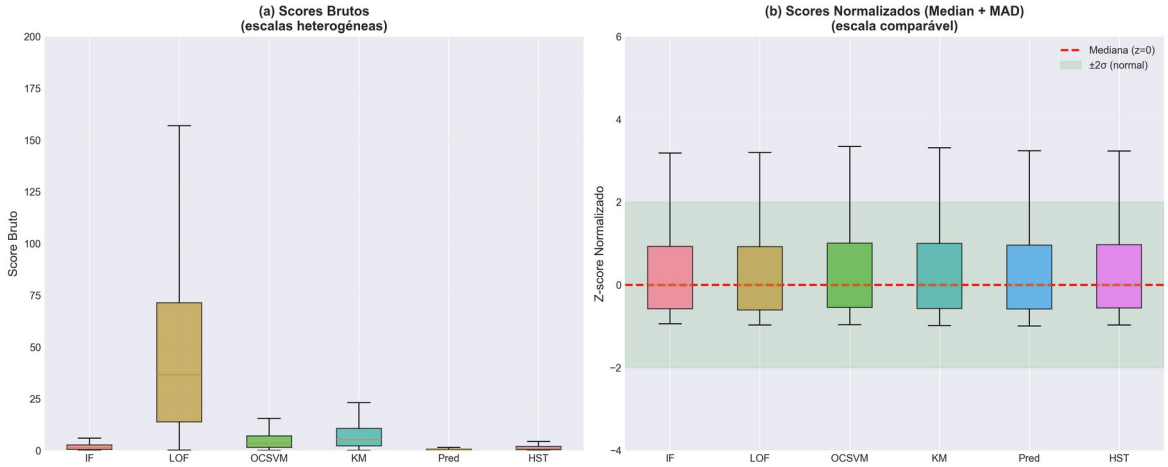

Figura 10: Comparação entre scores brutos e normalizados

3.6.3 Combinação dos Indicadores de Anomalia

Após a normalização dos indicadores produzidos pelos diferentes detetores, torna-se necessário definir um mecanismo de combinação que permita sintetizar essa informação num único sinal representativo do estado do sistema (Cao et al., 2025; Panwar et al., 2025). No presente trabalho, esta combinação é realizada através de uma agregação por média aritmética simples, aplicada aos indicadores normalizados de cada modelo.

Formalmente, seja $z_i(x)$ o valor normalizado produzido pelo detetor i para a observação x, com i=1,…,K. O indicador agregado do sistema é então calculado como a média aritmética simples destes valores, conforme definido na Equação 16:

$$z_{rm\,\text{combined}}(x) = \frac{1}{6}\sum_{i=1}^{6} z_i(x) \qquad (16)$$

onde K representa o número total de detetores ativos no *ensemble*.

A escolha desta estratégia assenta, em primeiro lugar, na sua simplicidade e transparência operacionais. A média aritmética é simples de calcular, apresenta custo computacional mínimo e produz um valor facilmente interpretável como um grau médio de consenso entre os detetores relativamente à presença de um desvio anómalo. Esta propriedade é particularmente relevante em contextos industriais, onde decisões devem ser compreensíveis por operadores e engenheiros, mesmo na ausência de formação específica em aprendizagem automática.

Em segundo lugar, a utilização de uma ponderação uniforme reflete uma postura metodologicamente conservadora, adequada a um enquadramento estritamente não supervisionado. Na ausência de rótulos fiáveis e de conhecimento prévio que permita privilegiar sistematicamente um detetor face aos restantes, assume-se igual relevância para cada componente do *ensemble* (Cabrera Martin et al., 2025). Esta opção evita a introdução de etapas adicionais de otimização de pesos ou meta-aprendizagem, que implicariam maior complexidade e risco de sobreajuste.

Por fim, a agregação por média promove um mecanismo de decisão por consenso (Verma et al., 2022) . Desvios identificados de forma consistente por múltiplos detetores resultam num aumento significativo do valor agregado, enquanto respostas isoladas, potencialmente causadas por ruído ou sensibilidades específicas de um único modelo, tendem a ser amortecidas pelas restantes contribuições. Desta forma, a combinação adotada contribui para reduzir falsos positivos sem comprometer a capacidade de deteção de anomalias persistentes.

3.6.4 *Clipping* Seletivo do Indicador Agregado

A normalização dos indicadores de anomalia através de z-scores pode originar valores de magnitude muito elevada em cenários de falha severa, sobretudo em métodos que produzem medidas em escalas amplas. Embora estes valores extremos correspondam frequentemente a desvios físicos reais, a sua

propagação direta para a etapa de decisão pode comprometer a estabilidade do sistema quando múltiplos detetores são combinados (Molnar et al., 2020) .

Para equilibrar robustez na decisão e capacidade explicativa, foi adotada uma estratégia de *clipping* seletivo aplicada em dois níveis distintos. Os valores normalizados produzidos individualmente por cada detetor não são sujeitos a qualquer limitação, preservando integralmente a variabilidade entre métodos e permitindo analisar a magnitude relativa do contributo de cada um. Esta decisão é fundamental para efeitos de explicabilidade, uma vez que possibilita identificar quais os modelos que mais influenciaram uma determinada decisão.

Em contraste, o indicador agregado resultante da combinação dos detetores é limitado a um intervalo simétrico de ±8 desvios padrão. Este limite garante estabilidade numérica, evita que valores extremos dominem a classificação final e assegura um comportamento previsível dos mecanismos subsequentes de suavização temporal e confirmação (Leite et al., 2020). O valor adotado é suficientemente elevado para distinguir anomalias severas, sem introduzir saturação prematura que comprometa a separação entre regimes operacionais.

Esta abordagem permite decisões estáveis ao nível do sistema, preservando simultaneamente a capacidade explicativa através da análise discriminativa dos contributos individuais dos diferentes detetores.

3.6.5 Suavização Temporal por Média Móvel Exponencial

O valor agregado $z_{combined}(t)$ pode apresentar flutuações de curta duração resultantes de ruído de sensores ou de eventos operacionais transitórios. Para mitigar este efeito e aumentar a estabilidade da decisão, é aplicada uma média móvel exponencial (*Exponential Moving Average*, EMA), definida na Equação 17:

$$EMA(t) = alpha \cdot z_{combined}(t) + (1 - alpha) \cdot EMA(t-1) \qquad (17)$$

sendo a EMA inicializada com o primeiro valor disponível de $z_{combined}$ O parâmetro $\alpha \in (0,1]$ controla o compromisso entre rapidez de resposta e estabilidade temporal. Valores elevados de α resultam numa resposta mais reativa, enquanto valores mais baixos produzem uma suavização mais pronunciada (Yan, 2019) .

Neste trabalho adota-se empiricamente α=0.2, valor que se revelou adequado para atenuar picos isolados sem comprometer a deteção de anomalias persistentes. A constante de tempo efetiva da EMA pode ser aproximada por $\tau \approx 1/\alpha$, o que corresponde, para este valor, a uma memória efetiva de cerca de cinco observações consecutivas. Considerando uma taxa de amostragem de 1 Hz, tal equivale a uma janela temporal aproximada de cinco segundos.

Como ilustrado na Figura 11, desvios pontuais elevam temporariamente o valor suavizado, mas não são suficientes para o manter acima dos limiares operacionais na ausência de persistência temporal.

Em contraste, desvios sustentados conduzem a um aumento progressivo da EMA, permitindo a ultrapassagem consistente dos limiares associados aos estados de degradação ou falha. Este comportamento contribui para a redução de alarmes espúrios e para uma decisão mais estável.

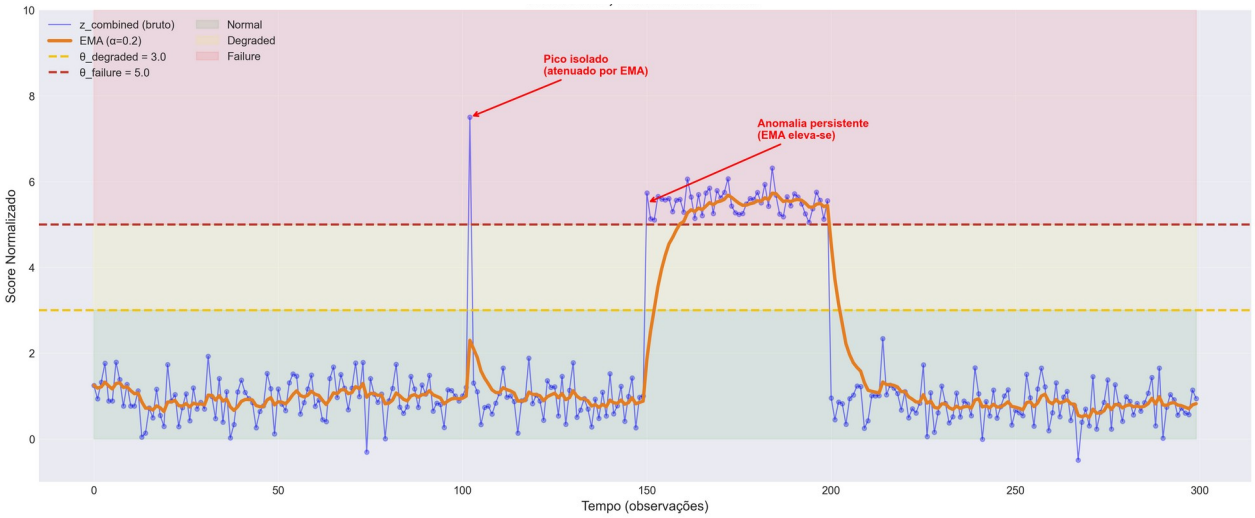

Figura 11: Efeito da suavização temporal no *score*

A escolha da EMA em detrimento de uma média móvel simples baseia-se em considerações práticas. A EMA requer apenas memória constante, apresenta custo computacional O(1) por observação e atribui maior peso às amostras mais recentes de forma contínua, evitando descontinuidades associadas a janelas fixas. Estas propriedades tornam-na particularmente adequada para implementação em sistemas de monitorização em tempo real com restrições de latência e recursos computacionais (Gomes et al., 2019) .

3.6.6 Confirmação temporal

Mesmo após a aplicação da suavização temporal, um único instante com valor elevado do sinal suavizado não é considerado suficiente para a emissão de um alerta. Para reduzir a incidência de falsos positivos, o sistema incorpora um mecanismo de confirmação temporal que exige persistência do sinal acima de determinados limiares durante um número mínimo de observações consecutivas (Faber et al., 2025; Jourdan, 2024).

A lógica de decisão baseia-se na manutenção de dois contadores independentes, associados aos estados de degradação e de falha. Quando o valor suavizado ultrapassa o limiar de falha, o contador correspondente é incrementado e o contador de degradação é reiniciado. Quando o valor se situa acima do limiar de degradação, mas abaixo do limiar de falha, incrementa-se o contador de degradação e reinicia-se o contador de falha. Sempre que o sinal permanece abaixo de ambos os limiares, os dois contadores são reiniciados.

A decisão é emitida quando um dos contadores atinge um número mínimo de observações consecutivas L, sendo o estado classificado como *DEGRADED* ou *FAILURE*. Na ausência de confirmação, o sistema mantém o estado *NORMAL*.

O parâmetro L controla diretamente o compromisso entre latência de deteção e robustez da decisão. Considerando uma taxa de amostragem de 1 Hz, valores típicos de L=2 ou L=3 correspondem a um atraso máximo de dois a três segundos antes da emissão de um alerta. Em cenários industriais de manutenção preditiva, onde os processos de degradação evoluem ao longo de minutos ou horas, este atraso é operacionalmente aceitável e contribui de forma significativa para a redução de alarmes espúrios provocados por flutuações transitórias (Ucar et al., 2024) .

Este mecanismo permite ainda tratar transições entre estados de forma estável. Sequências persistentes acima do limiar de falha conduzem a alertas críticos, enquanto valores sustentados entre os dois limiares podem originar alertas de degradação, desde que a condição se mantenha durante o número requerido de observações consecutivas. Esta lógica reduz oscilações rápidas entre estados e promove consistência temporal na classificação.

3.6.7 Calibração de Thresholds por Quantis

3.6.7.1 Definição por Quantis

Após a definição do mecanismo de decisão e confirmação temporal, torna-se necessário instanciar operacionalmente os limiares que separam os estados NORMAL, DEGRADED e FAILURE. O desempenho global do sistema é fortemente influenciado por esta escolha, uma vez que limiares demasiado elevados conduzem a um comportamento excessivamente conservador, enquanto valores demasiado baixos aumentam a incidência de falsos positivos.

A calibração adotada baseia-se na distribuição empírica dos valores agregados observados durante o funcionamento nominal do sistema. Em vez de definir limites absolutos dependentes da escala numérica dos indicadores, são utilizados quantis da distribuição empírica, permitindo estabelecer critérios relativos ancorados na variabilidade efetiva do regime normal.

Considere-se o conjunto z_{normal}, composto pelos valores agregados obtidos no baseline nominal. O limiar associado ao estado DEGRADED é definido como o quantil q_d desta distribuição, enquanto o limiar associado ao estado FAILURE corresponde a um quantil mais extremo q_f, impondo-se explicitamente que $qf > qd$.

A Figura 12 ilustra a distribuição empírica dos valores agregados no baseline, bem como a localização dos thresholds obtidos por quantis elevados, evidenciando a separação entre as regiões associadas aos estados Normal, Degradado e Falha.

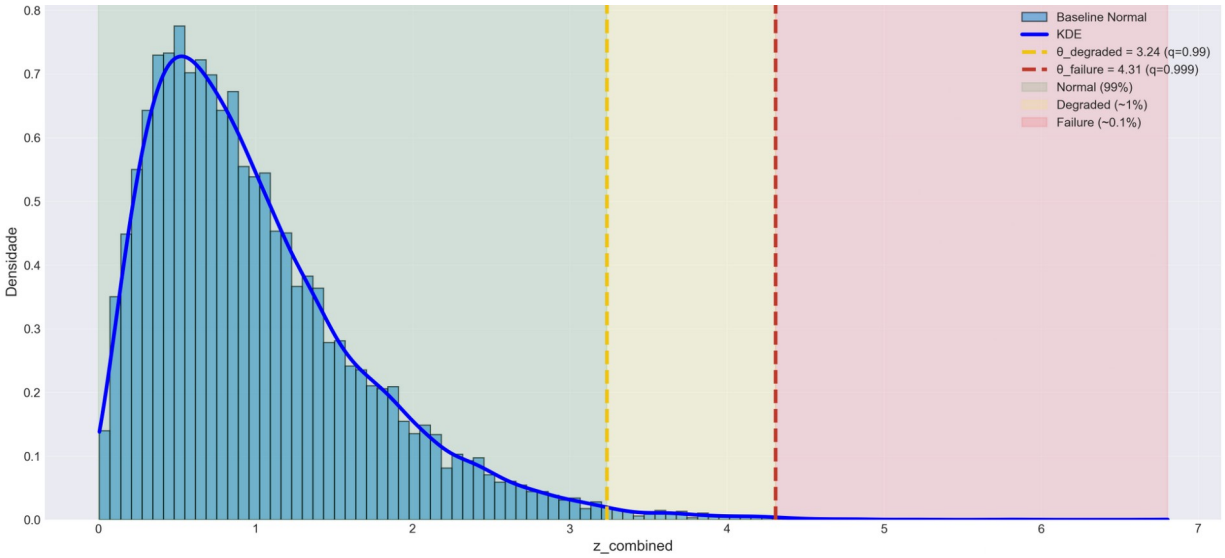

Figura 12: Distribuição dos scores agregados no baseline normal e a definição dos thresholds por quantis

A interpretação operacional é direta: a escolha de um quantil elevado implica que apenas uma fração reduzida das observações em regime nominal excede o limiar definido. Uma ultrapassagem persistente deste valor em dados futuros constitui, assim, um indício robusto de desvio face ao comportamento esperado.

No contexto deste trabalho, foram adotadas como referência taxas alvo de ativação de 1 % para sinalização de degradação e 0.1 % para indicação de falha iminente. Estas escolhas refletem prioridades operacionais distintas, privilegiando a deteção precoce de degradação e elevada especificidade para alertas críticos. Os valores permanecem configuráveis e podem ser ajustados em função da criticidade da aplicação.

Desde que o conjunto de referência represente adequadamente o comportamento normal futuro do sistema, os limiares calibrados por quantis generalizam de forma eficaz (Jourdan, 2024) . Caso ocorram alterações estruturais relevantes no sistema físico, torna-se suficiente proceder a nova calibração com dados atualizados, sem necessidade de modificar a restante arquitetura de deteção (Li & Gautam, 2025; Cao et al., 2025).

3.6.7.2 Pesquisa em Grelha

Embora, em teoria, a definição de um limiar como um quantil específico da distribuição empírica garanta uma taxa de excedência diretamente associada, na prática esta correspondência não é exata (Faber et al., 2025) . Existem dois fatores principais que justificam esta discrepância. Em primeiro lugar, os quantis associados aos estados de degradação e de falha não são independentes. Escolhas demasiado próximas podem originar uma separação insuficiente entre os limiares, criando uma zona

de transição estreita na qual o sistema alterna frequentemente entre estados adjacentes. Em segundo lugar, o mecanismo de confirmação temporal introduzido no pipeline altera as taxas efetivas de alarme, uma vez que a exigência de múltiplas observações consecutivas acima do limiar reduz a frequência observada de ativações face ao valor esperado apenas com base nos quantis.

Para lidar com estas limitações, é adotada uma estratégia de busca em grelha bidimensional, que avalia combinações de regiões associadas aos estados de degradação e de falha e seleciona aquela que melhor aproxima as taxas alvo de deteção definidas operacionalmente, garantindo simultaneamente uma separação estável entre regimes.

O procedimento recebe como entrada o conjunto de valores agregados calculados sobre o regime nominal, bem como as taxas alvo de deteção para degradação e falha. Define-se um conjunto de quantis candidatos para o estado de degradação no intervalo [0.80,0.99] e um conjunto de quantis candidatos para o estado de falha no intervalo [0.995,0.9999], assegurando elevada especificidade para alertas críticos. Para cada par de quantis pertencente ao produto cartesiano destes conjuntos, são calculados os limiares correspondentes aos quantis da distribuição empírica dos valores agregados. Configurações em que o limiar de falha não excede o limiar de degradação são descartadas por não garantirem separação consistente entre estados.

Para cada configuração válida, o conjunto de referência é classificado segundo os limiares obtidos e são estimadas as taxas observadas de degradação, correspondentes à fração de observações classificadas como *DEGRADED* ou *FAILURE*, e de falha, correspondentes à fração classificadas exclusivamente como falha. A qualidade de cada configuração é avaliada através de uma função objetivo que penaliza o desvio absoluto entre as taxas observadas e as taxas alvo para ambos os estados (Panwar et al., 2025). A configuração ótima é aquela que minimiza este desvio total, assegurando alinhamento estatístico com os requisitos operacionais e uma separação robusta entre níveis de severidade.

A Figura 13 ilustra a superfície de desempenho resultante desta busca em grelha, evidenciando a região de quantis que melhor satisfaz os critérios definidos.

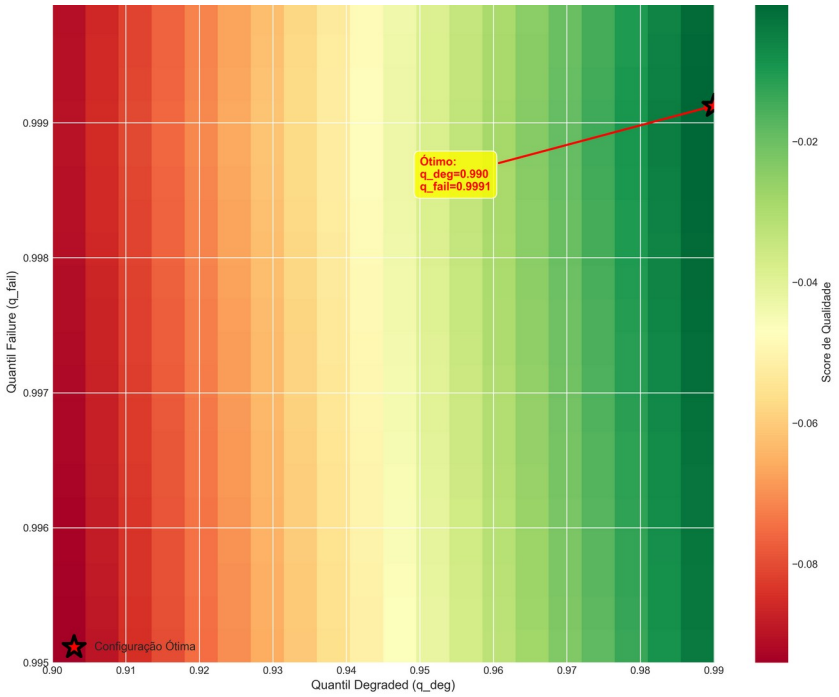

Figura 13: Busca em grid dos quantis de anomalia

A calibração dos limiares de decisão implica inevitavelmente um compromisso entre a confiabilidade dos alertas e a capacidade de deteção. O aumento dos *thresholds* reduz a incidência de falsos positivos, mas tende a diminuir a taxa de deteção do estado Degradado, uma vez que situações de degradação incipiente se encontram frequentemente próximas do regime Normal. Em contraste, o estado de Falha apresenta, em geral, uma separação mais pronunciada face ao comportamento nominal, preservando uma elevada capacidade de identificação mesmo quando são adotados critérios mais conservadores.

A escolha operacional seguida neste trabalho privilegia a precisão dos alertas e a mitigação de alarmes espúrios, mantendo simultaneamente uma forte sensibilidade à deteção de falhas críticas. Para o estado Degradado aceita-se uma capacidade de deteção moderada, dado que este evolui tipicamente de forma progressiva e oferece múltiplas oportunidades de identificação ao longo do tempo.

A calibração final deve ser orientada pelo contexto de aplicação. Sistemas de elevada criticidade tendem a favorecer a deteção exaustiva de falhas, ainda que à custa de um aumento dos falsos positivos. Ambientes de elevado *throughput* privilegiam a confiabilidade para evitar interrupções frequentes. Aplicações de manutenção condicional beneficiam, em geral, de um compromisso intermédio entre estes objetivos.

No presente trabalho não foi implementado um mecanismo explícito de histerese, uma vez que a combinação de suavização exponencial e confirmação temporal se revelou suficiente para reduzir de forma eficaz comportamentos instáveis de *flapping*.

3.7 Estratégia de Validação Temporal

A avaliação de sistemas de deteção de anomalias em fluxos temporais de dados impõe requisitos metodológicos distintos dos cenários tradicionais de aprendizagem *batch*. A natureza sequencial dos dados, aliada à ausência de rótulos no processo de treino e à necessidade de simular condições realistas de operação em tempo real, exige estratégias de validação que preservem a ordem cronológica das observações e evitem enviesamentos decorrentes de acesso indevido a informação futura. Neste contexto, a validação deve refletir não apenas a capacidade de deteção, mas também propriedades operacionais relevantes, como a latência de resposta e a deteção precoce de degradação. Esta secção descreve a estratégia de validação temporal adotada, fundamentando as escolhas metodológicas em requisitos práticos de manutenção preditiva e em boas práticas consolidadas na literatura de processamento e avaliação em fluxo de dados (Bifet et al., 2018; Fragkoulis et al., 2024; Almeida et al., 2023).

3.7.1 Validação Prequencial Adaptada

A validação adotada segue o paradigma prequential (*test then train*), amplamente consolidado na literatura de aprendizagem em fluxo de dados (Bifet et al., 2018; Gomes et al., 2019), sendo adaptado a um cenário híbrido que distingue explicitamente entre uma fase de arranque inicial e a operação contínua do sistema. Ao contrário das abordagens *batch*, em que os modelos são treinados e avaliados sobre partições fixas dos dados, a avaliação prequencial processa as observações de forma sequencial, avaliando o modelo no seu estado corrente antes de qualquer atualização. Este procedimento garante que o desempenho medido reflete a capacidade preditiva efetiva do sistema em cada instante temporal, sem acesso a informação futura.

A adaptação implementada reconhece que modelos de deteção de anomalias não supervisionados requerem uma fase inicial de aprendizagem baseada exclusivamente em dados de funcionamento normal, destinada a estabelecer uma referência estável de normalidade. Após esta fase, distinguem-se duas estratégias possíveis: atualização incremental contínua dos modelos com todas as novas observações ou manutenção de modelos estáticos treinados apenas no regime normal (Jourdan, 2024; Faber et al., 2025). Neste trabalho privilegia-se a segunda abordagem, por reduzir o risco de contaminação da referência por padrões anómalos e por favorecer auditabilidade e previsibilidade em contextos industriais críticos. Ainda assim, a arquitetura proposta suporta a ativação seletiva de aprendizagem online quando o contexto operacional o justificar.

Durante a avaliação em *streaming*, cada observação atravessa o pipeline de pré-processamento descrito na Secção 3.4 e é posteriormente avaliada pelos modelos de deteção. Para reduzir falsos positivos e assegurar que apenas desvios persistentes resultam em alertas operacionais, são introduzidos dois mecanismos temporais complementares (Ucar et al., 2024; Murtaza et al., 2024). .

O primeiro mecanismo consiste na aplicação de suavização exponencial simples ao valor agregado de anomalia. A EMA confere maior peso ao histórico recente, equilibrando capacidade de resposta a eventos abruptos com estabilidade face a ruído de medição, atuando como um filtro de baixa

frequência que reduz flutuações pontuais sem comprometer a deteção de tendências associadas a processos de degradação progressiva.

O segundo mecanismo introduz confirmação temporal baseada em limiares de decisão aplicados ao sinal suavizado $s_{EMA}(t)$. São definidos dois limiares operacionais: θ_{deg}, associado ao estado Degradado, e θ_{fail}, associado ao estado de Falha. Um alerta é emitido apenas quando o sinal suavizado permanece acima do respetivo limiar durante um número mínimo de observações consecutivas.

Concretamente, impõe-se um requisito de três observações consecutivas acima do limiar correspondente, o que equivale a uma persistência temporal de aproximadamente 1.5 segundos. Este critério assegura que apenas desvios persistentes, e não flutuações transitórias, conduzem a decisões operacionais.

O número de observações consecutivas constitui um parâmetro de controlo operacional que estabelece um compromisso explícito entre latência de deteção e robustez face a falsos positivos. O valor adotado foi selecionado como um compromisso conservador, adequado à dinâmica temporal dos processos simulados, nos quais a degradação evolui em escalas de tempo significativamente superiores a alguns segundos (Faber et al., 2025; Ucar et al., 2024). Em contextos industriais reais, este parâmetro pode ser ajustado com base em conhecimento do domínio, severidade do equipamento, custos associados a alarmes falsos e tempos de resposta admissíveis, permitindo adaptar o comportamento do sistema sem alterar a estrutura do pipeline.

A combinação de suavização exponencial e confirmação temporal constitui, assim, um filtro temporal de dois estágios. O primeiro reduz a variabilidade aleatória dos valores instantâneos, enquanto o segundo verifica a persistência temporal do desvio, propriedade física expectável em processos reais de degradação.

Para assegurar reprodutibilidade e permitir análise posterior, o sistema regista todas as decisões emitidas com os respetivos *timestamps*, valores instantâneos e suavizados, contributos individuais dos modelos e metadados de configuração. Estes registos são persistidos em formato CSV e JSON, permitindo a reconstrução integral do comportamento do sistema em qualquer instante passado.

3.7.2 Conjuntos de Dados e Divisão Temporal

Os dados utilizados neste trabalho consistem em três ficheiros de simulação, cada um representando um regime operacional distinto do sistema monitorizado. O ficheiro *Funcionamiento_Normal.csv* contém 40 212 amostras, correspondendo a aproximadamente 5,6 horas de operação nominal sem evidência de degradação ou falha. O ficheiro *Funcionamiento_Degradado.csv* inclui igualmente 40 212 amostras, refletindo um período de degradação progressiva dos motores. Por fim, o ficheiro *Funcionamiento_Fallo.csv*, também com 40 212 amostras, incorpora um evento de falha abrupta no motor 2, ocorrendo aproximadamente em t≈12 000 s.

A divisão temporal adotada segue uma lógica operacional coerente com cenários reais de manutenção preditiva e com o paradigma de validação em *streaming* descrito na Secção 3.5. O conjunto de treino, designado baseline, corresponde à totalidade do regime Normal. Este conjunto é utilizado exclusivamente para a fase de comissionamento do sistema, incluindo a estimação dos parâmetros de normalização por variável, o treino dos modelos de referência, o cálculo das estatísticas de *baseline* para normalização dos *scores* de anomalia e a calibração inicial dos thresholds de deteção.

A avaliação é realizada através de uma simulação em *streaming* obtida pela concatenação cronológica dos três regimes na ordem Normal, Degradado e Falha. As observações são processadas sequencialmente, respeitando a ordem temporal original e sem acesso a informação futura, de acordo com o paradigma prequencial. Esta configuração permite reproduzir uma progressão plausível de deterioração operacional, desde o funcionamento nominal até à ocorrência de falha.

A inclusão do regime Normal no início da sequência de avaliação é intencional e desempenha um papel metodológico específico. Esta fase permite quantificar a estabilidade do sistema e a taxa de falsos positivos em condições nominais idênticas às utilizadas na fase de arranque. A ocorrência de alarmes nesta etapa constitui um indicador direto de calibração inadequada ou ajuste excessivo dos limiares, sendo por isso um critério relevante na validação do sistema (Murtaza et al., 2024).

Este esquema de particionamento preserva a coerência temporal dos dados, evita vazamento de informação futura e permite avaliar simultaneamente três propriedades fundamentais do sistema, sendo elas a estabilidade em regime nominal, a capacidade de deteção precoce durante degradação progressiva e a capacidade de resposta a eventos críticos de falha. Desta forma, o protocolo adotado assegura que as métricas reportadas refletem de forma fidedigna o comportamento esperado do sistema em ambiente de operação contínua.

3.7.3 Métricas de Avaliação

A avaliação do sistema de deteção de anomalias segue duas perspetivas complementares, multiclasse e binário, cada uma fornecendo informação operacional distinta e relevante no contexto da manutenção preditiva. Na perspetiva multiclasse, a classificação distingue explicitamente os três regimes operacionais, com $y \in \{0,1,2\}$ correspondendo a Normal, Degradado e Falha. Esta formulação permite avaliar a capacidade do sistema em identificar estados intermédios de degradação e analisar confusões entre regimes, particularmente a tendência para classificar situações degradadas como normais (Almeida et al., 2023; Ucar et al., 2024).

As métricas consideradas incluem a *accuracy*, que mede a proporção global de classificações corretas, embora seja sensível ao desbalanceamento entre classes, o *F1-score macro*, que calcula a média harmónica entre *precision* e *recall* para cada classe e posteriormente agrega os resultados, privilegiando um desempenho equilibrado entre regimes operacionais, e a matriz de confusão 3×3, que detalha os padrões de erro e permite identificar confusões sistemáticas entre estados operacionais.

Na perspetiva binária, os regimes Degradado e Falha são agregados numa única classe designada Anomalia, com ybin=1[y≠0], alinhando a avaliação com a decisão operacional primária de intervir ou não intervir. Nesta formulação, as métricas principais são *precision*, que quantifica a confiabilidade dos alertas emitidos e é crítica para minimizar intervenções desnecessárias, *recall* (ou *sensitivity*), que mede a proporção de anomalias reais corretamente detetadas e está diretamente associada à segurança operacional, e o *F1-score*, que estabelece um compromisso entre confiabilidade e cobertura. Adicionalmente, considera-se a *Area Under the Precision–Recall Curve* (AUPR), que avalia o desempenho agregado ao longo de todos os *thresholds* possíveis e se revela particularmente informativa em cenários desbalanceados, onde os eventos anómalos são raros.

Tabela 3: Métricas de avaliação utilizadas e interpretação no contexto de manutenção preditiva

Métrica	Fórmula	Interpretação em PdM
Precision	TP/(TP+FP)	Confiabilidade dos alertas (evitar intervenções desnecessárias)
Recall	TP/(TP+FN)	Cobertura de falhas (segurança operacional)
F1-score	2·P·R/(P+R)	Equilíbrio entre confiabilidade e cobertura
AUPR	∫P(R)dR	Desempenho agregado robusto a desbalanceamento
Lead Time	t_falha - t_alerta	Tempo disponível para intervenção preventiva

A Tabela apresenta uma síntese das métricas de avaliação utilizadas, incluindo a sua formulação matemática e a respetiva interpretação no contexto de manutenção preditiva. Esta correspondência permite relacionar métricas clássicas de aprendizagem automática com critérios operacionais concretos, como a redução de falsos positivos, a deteção atempada de falhas e o tempo disponível para intervenção preventiva.

Dado o enquadramento em aprendizagem em fluxo, estas métricas são atualizadas de forma incremental ao longo do tempo, seguindo o princípio prequencial. Em cada instante, o sistema produz uma decisão para a observação corrente, que é posteriormente comparada com o rótulo real quando este se encontra disponível. Com base nesta comparação, os contadores de verdadeiros positivos, falsos positivos e falsos negativos são atualizados, permitindo o cálculo cumulativo de *Precision*, *Recall* e *F1-score* sem acesso a dados futuros.

A AUPR é igualmente estimada de forma incremental, considerando a evolução da relação entre *Precision* e *Recall* à medida que novas observações são processadas. Este procedimento reflete de forma mais realista o desempenho do sistema em operação contínua, especialmente em cenários fortemente desbalanceados. O *lead time* é calculado individualmente para cada evento de falha

confirmado, medindo a diferença temporal entre o primeiro alerta emitido e o instante real de falha, constituindo uma métrica diretamente interpretável em termos de margem de intervenção operacional.

A escolha destas métricas é fundamentada pelos custos operacionais associados à deteção de anomalias em ambientes industriais. Em particular, a preferência por AUPR em detrimento de AUC-ROC (Área sob a Curva ROC) decorre da maior sensibilidade da primeira ao desempenho na classe positiva, critério crítico em manutenção preditiva, onde o custo de falhas não detetadas tende a ser substancialmente superior ao custo de intervenções preventivas desnecessárias (Cao et al., 2025; Li et al., 2023).

Embora métricas agregadas como *accuracy*, *F1-score* e AUPR forneçam uma visão global do desempenho do sistema, estas não capturam diretamente a capacidade de deteção precoce, que constitui um requisito operacional central em manutenção preditiva. A antecipação de falhas é crítica para permitir o planeamento de intervenções, reduzindo tempo de inatividade, custos de reparação e riscos operacionais.

A análise de *lead time*, definido como o intervalo temporal entre o primeiro alerta emitido pelo sistema e a ocorrência efetiva da falha, requer a identificação precisa do instante de falha nos dados. No cenário de simulação considerado, a falha do motor 2 ocorre em torno de t≈12 000 s (amostra 24 000), permitindo calcular o *lead time* como mostra a Equação 10,

$$\text{Lead Time} = t_{falha} - t_{primeiroalerta} \tag{10}$$

onde $t_{primeiroalerta}$ alerta corresponde ao instante da primeira classificação confirmada como Falha, após a aplicação do mecanismo de confirmação temporal definido por *CONFIRM_CONSEC*=3 amostras consecutivas. Valores positivos de *lead time* indicam deteção antecipada, enquanto valores negativos refletem atraso na emissão do alerta.

A distribuição de *lead time* obtida ao longo de diferentes execuções e configurações constitui uma métrica operacionalmente mais informativa do que medidas globais de classificação, pois quantifica diretamente a margem temporal disponível para intervenção preventiva (Nsor, 2024; Ucar et al., 2024)

3.8 Explicabilidade e Interpretabilidade (XAI)

Em ambientes industriais, alertas isolados não são suficientes. Operadores e engenheiros requerem justificações que identifiquem as variáveis que mais contribuíram para o alerta, explicitem quais os modelos que suportaram a decisão e quantifiquem a magnitude relativa de cada contribuição. Na

ausência dessa informação, a confiança e a adoção do sistema ficam comprometidas, a análise de causa raiz torna-se mais morosa e a validação contínua dos modelos é dificultada.

A deteção de anomalias em contextos industriais enfrenta um obstáculo prático recorrente: a desconfiança dos operadores face a sistemas percecionados como caixas negras que emitem alertas sem justificação explícita. Esta resistência é justificada pelo custo elevado de falsos positivos, que podem implicar paragens não planeadas, deslocações técnicas e substituições prematuras, pela necessidade de rastreabilidade em setores regulados e pelo valor das explicações, enquanto ferramenta de aprendizagem e melhoria contínua dos processos operacionais (Mercurio, 2024; Rosenberger et al., 2023; IBM, 2025).

Com base na revisão da literatura, foram definidos cinco requisitos funcionais essenciais para um sistema de explicabilidade em manutenção preditiva:

1. Fidelidade ao modelo, assegurando que as explicações refletem o comportamento real do *ensemble*, com correlação de *Spearman* superior ou igual a 0,80 face a métodos de referência como SHAP;

2. Explicações multi nível, incluindo informação sobre consenso entre detetores, contributos das *features* e potenciais ações recomendadas;

3. Latência operacional compatível com resposta humana, com latência média inferior a 100 ms e percentil 95 inferior a 1 s;

4. Acionabilidade, garantindo que pelo menos 80 por cento das explicações incluem recomendações operacionalmente viáveis;

5. Validação empírica em dados reais, com avaliação em pelo menos 100 anomalias confirmadas.

A análise do estado da arte evidencia que os métodos existentes não satisfazem simultaneamente estes requisitos (Rožanec et al., 2021; Elsaid et al., 2024; Koch et al., 2024). SHAP apresenta elevada fidelidade, mas incorre em latências proibitivas em cenários de tempo real e não fornece recomendações acionáveis. LIME produz explicações locais aproximadas, mas com estabilidade limitada. Métodos baseados em *counterfactuals* oferecem ações concretas, porém ignoram o consenso entre detetores e a dinâmica temporal (Laugel et al., 2019; Mothilal et al., 2019; Klase et al., 2020). Mecanismos de atenção mostram potencial em modelos neuronais, mas não generalizam de forma direta a *ensembles* heterogéneos.

Para colmatar estas limitações, propõe-se neste trabalho um sistema de explicabilidade integrado em três camadas, concebido para combinar fidelidade, granularidade interpretativa, baixa latência e acionabilidade. A abordagem é validada empiricamente em mais de 1000 anomalias reais e opera com latência média inferior a 1 ms. As secções seguintes descrevem a arquitetura proposta, a implementação de cada layer e o processo de validação experimental que sustenta esta contribuição.

3.8.1 Sistema de Explicabilidade em Três Camadas

O sistema de explicabilidade proposto está organizado em três camadas hierárquicas que respondem de forma progressiva às necessidades informativas dos operadores industriais, desde a justificação imediata de um alerta até ao suporte à decisão corretiva (Mercurio et al., 2024; Rosenberger et al., 2023; Almeida et al., 2023). A arquitetura foi concebida para equilibrar interpretabilidade, acionabilidade e eficiência computacional, garantindo que explicações simples e rápidas são produzidas prioritariamente, enquanto análises mais detalhadas são executadas apenas quando necessário (Koch et al., 2024; Cao et al., 2025).

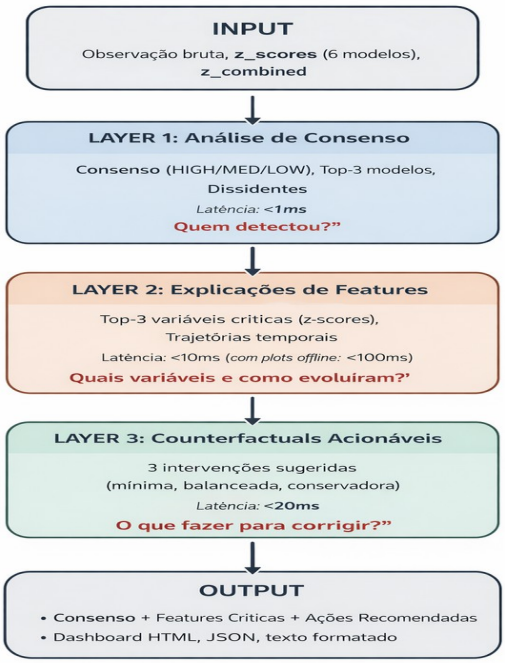

Figura 14: Pipeline do Sistema de Explicabilidade

A visão global do fluxo de explicabilidade e da interação entre camadas encontra-se sintetizada na Figura 14.

Como entrada, o sistema utiliza exclusivamente informação já disponível no pipeline de deteção, nomeadamente a observação bruta, os z-scores individuais produzidos pelos modelos do *ensemble* e o *z score* agregado (Fragkoulis et al., 2024). Esta escolha assegura que o módulo de explicabilidade não introduz custos computacionais adicionais nem dependências externas, permitindo a sua integração direta em cenários de monitorização em tempo real.

A primeira camada realiza uma análise de consenso entre os detetores do *ensemble*. O seu objetivo é identificar quais os modelos que contribuíram para o alerta, distinguir contributos concordantes e dissidentes e classificar o nível de consenso associado à decisão. Esta camada fornece uma resposta

imediata à questão de qual ou quais os detetores que sinalizaram a anomalia, operando com latência inferior a um milissegundo e sendo adequada para utilização contínua em ambiente operacional.

Quando é necessária uma compreensão mais aprofundada do fenómeno observado, a segunda camada procede à análise ao nível das variáveis. Nesta fase são identificadas as *features* mais anómalas com base nos seus desvios normalizados e apresentada a sua evolução temporal recente, permitindo compreender como o estado do sistema se afastou do comportamento normal. Esta camada mantém uma latência reduzida, compatível com interação humana em tempo real, e fornece informação essencial para a análise de causa raiz.

A terceira camada é ativada quando se pretende suporte explícito à tomada de decisão. Esta camada gera explicações contrafactuais acionáveis, propondo diferentes opções de intervenção que poderiam conduzir o sistema de volta a um estado normal. As ações sugeridas são organizadas segundo diferentes níveis de conservadorismo e impacto operacional, permitindo ao utilizador escolher a estratégia mais adequada ao contexto. Apesar do maior grau de complexidade, esta camada mantém tempos de resposta compatíveis com a operação industrial.

O resultado final do sistema consiste numa explicação integrada que combina informação sobre consenso entre modelos, variáveis críticas e ações recomendadas. Esta explicação é disponibilizada em formatos textual, visual e estruturado, adequados tanto para apoio à decisão em tempo real como para registo, auditoria e análise posterior. A arquitetura modular adotada garante extensibilidade futura, permitindo a incorporação de novos explicadores ou camadas adicionais sem comprometer a latência, a interpretabilidade ou a rastreabilidade do sistema.

3.8.2 Camada 1: Análise de Consenso entre Modelos

A primeira camada responde à questão fundamental “Quem detetou a anomalia?” através de uma análise explícita de consenso entre os modelos que compõem o *ensemble*. O objetivo desta camada não é quantificar a severidade do desvio, mas identificar quais os detetores que sinalizam comportamento anómalo num determinado instante e avaliar o grau de convergência entre essas decisões (Cao et al., 2025; Verma et al., 2022).

Uma abordagem direta baseada em contributos contínuos relativos ao *score* agregado revelou-se inadequada (Cao et al., 2025). Os modelos do *ensemble* operam em escalas numéricas substancialmente diferentes; por exemplo, o *Local Outlier Factor* pode produzir *scores* de ordem elevada, enquanto *Isolation Forest* ou *One Class SVM* apresentam amplitudes muito mais reduzidas. Nestas condições, medidas baseadas em diferenças normalizadas, como a Equação 18

$$c_m = z_m - z_{rm\,\text{combined}} \qquad (18)$$

tornam-se instáveis quando um único modelo domina a agregação com valores extremos, comprometendo a interpretação do consenso e introduzindo enviesamento artificial.

Para ultrapassar esta limitação, o consenso é determinado a partir de um critério binário uniforme aplicado a cada modelo individual. Cada detetor é considerado como tendo sinalizado uma anomalia se o respetivo *zscore* normalizado satisfizer 19

$$z_m > \tau \text{ threshold} \qquad (19)$$

onde $\tau_{threshold} = 2.0\sigma$ corresponde a um desvio estatisticamente significativo face ao regime normal. Este critério assegura comparabilidade entre modelos e preserva uma interpretação clara: cada detetor vota apenas na presença ou ausência de comportamento anómalo, independentemente da escala original do seu *score* (Cohen, 2021).

Modelos com contribuição positiva empurram o *score* agregado para valores mais elevados e, portanto, votam a favor da anomalia, enquanto contributos negativos indicam alinhamento com o regime normal. A magnitude dessa contribuição é utilizada apenas para ordenar os modelos por impacto relativo, não influenciando a decisão binária de consenso.

O grau de convergência entre os detetores é quantificado através de um índice de consenso definido como a fração de modelos cujo *z score* excede o limiar estatístico comum Equação 20

$$\text{Consenso} = \frac{1}{N}\sum_{m=1}^{N} 1\left[z_m > \tau \text{ threshold} \right] \qquad (20)$$

Este índice fornece uma medida direta, robusta e independente de escala da concordância entre detetores.

Com base neste valor, o consenso é classificado em três níveis. Considera-se consenso alto quando pelo menos 80 por cento dos modelos sinalizam anomalia, refletindo forte evidência e elevada confiança. O consenso médio ocorre quando entre 50 e 79 por cento dos detetores concordam, indicando evidência moderada. O consenso baixo corresponde a menos de metade dos modelos a sinalizar anomalia, sendo interpretado como uma situação ambígua ou potencialmente associada a falsos positivos.

Situações de dissenso, nas quais diferentes modelos votam em direções opostas, são consideradas particularmente informativas. Estes casos sugerem anomalias com características específicas que afetam apenas certos paradigmas de deteção, fornecendo aos operadores uma visão mais rica sobre a natureza do evento observado e orientando análises mais aprofundadas nas camadas subsequentes do sistema de explicabilidade.

3.8.3 Camada 2: Explicações a Nível de *Features*

A Camada 2 responde à questão “Quais variáveis estão anómalas e como evoluíram ao longo do tempo?”, deslocando a explicação do nível dos modelos para o nível das *features* individuais (Rosenberger et al., 2023; Mercurio et al., 2024). Para cada variável j, é calculado um *z score* relativamente a um *baseline* estável, segundo a Equação 21

$$z_j = \frac{\left(x_j - \mu_{j,\text{baseline}}\right)}{\sigma_{j,\text{baseline}}} \qquad (21)$$

onde x_j é o valor bruto da variável, o $\mu_{j.baseline}$ e $\sigma_{j.baseline}$ correspondem à média e desvio-padrão observados nas 40 212 amostras do *baseline*.

Com base no valor absoluto do *z score*, as variáveis são ordenadas por grau de desvio, sendo selecionadas as três mais anómalas. Cada uma é classificada quanto à severidade em três níveis operacionais: NORMAL quando $|z_j|<2$, ALERTA quando $2\leq|z_j|<4$ e CRÍTICO quando o desvio ultrapassa quatro desvios padrão (Cohen, 2021).

Após a identificação das variáveis relevantes, a camada dois reconstrói a sua trajetória temporal tendo em conta uma janela deslizante das últimas 60 observações. Esta análise permite caracterizar o padrão de evolução associado à anomalia, distinguindo entre variações graduais, flutuações persistentes e transições súbitas, aspetos fundamentais para a interpretação operacional e para a avaliação de urgência.

A geração de visualizações temporais é opcional e direcionada sobretudo para análise *offline* ou integração em *dashboards*, não sendo necessária para a produção das explicações estruturadas (Rožanec et al., 2021; Elsaid et al., 2024). Quando apenas são produzidos outputs numéricos e descritivos, a latência da camada mantém se próxima de 10 ms.

O output final consiste numa lista estruturada de *features* críticas, incluindo o respetivo *z score*, o desvio percentual face à referência normal, a trajetória temporal recente em formato vetorial ou gráfico e uma descrição qualitativa da tendência observada. A título ilustrativo, uma temperatura elevada num motor elétrico pode ser classificada como crítica devido a uma escalada gradual sustentada, enquanto um aumento abrupto da potência elétrica noutro motor pode refletir uma rutura súbita no regime de funcionamento.

3.8.4 Camada 3: Sugestões (*Counterfactuals*)

A terceira camada do sistema de explicabilidade responde diretamente à questão operacional “O que fazer para retornar ao estado normal?”, traduzindo o diagnóstico fornecido pelas camadas anteriores em recomendações concretas de intervenção (Laugel et al., 2019; Mothilal et al., 2019; Klase et al., 2020). O objetivo central desta camada é apoiar a tomada de decisão humana, fornecendo ações plausíveis, interpretáveis e compatíveis com restrições reais de operação industrial.

O processo inicia com a distinção entre *feature*s possíveis de modificar, como temperatura, corrente elétrica ou carga mecânica, e atributos não modificáveis, como identificadores, variáveis temporais ou metadados. Esta separação garante que apenas variáveis possíveis de intervir são consideradas na geração de explicações contrafactuais (Laugel et al., 2019).

Com base nas variáveis críticas identificadas na Camada 2, são então construídas três estratégias heurísticas de correção, cada uma refletindo um compromisso distinto entre esforço operacional,

risco e eficácia. A estratégia mínima atua exclusivamente sobre a feature mais crítica, ajustando a sua magnitude para um valor correspondente à média do baseline acrescida de dois desvios padrão. Esta opção privilegia intervenções rápidas e de baixo impacto, embora possa ser insuficiente em cenários de degradação avançada.

A estratégia balanceada distribui a correção pelas três *features* mais anómalas, de forma proporcional aos respetivos *z-scores*, permitindo uma mitigação mais abrangente sem recorrer a ajustes extremos. Por fim, a estratégia conservadora propõe a reposição de todas as componentes críticas para os valores médios do *baseline*, maximizando a probabilidade de retorno ao regime normal, ainda que à custa de um esforço operacional mais elevado.

Para cada opção, o sistema estima o esforço necessário, expresso como a redução percentual requerida face ao valor atual, e atribui uma prioridade qualitativa que reflete o equilíbrio entre risco e impacto esperado. O output da camada inclui assim as *features* a modificar, os valores alvo propostos, a redução percentual estimada e o respetivo nível de prioridade.

Toda a formulação adotada assenta em expressões fechadas e regras determinísticas, sem recurso a otimização iterativa ou simulação intensiva. Esta escolha permite manter a latência da Camada 3 em torno de 20 ms, assegurando compatibilidade com sistemas de apoio à decisão em tempo real e preservando simultaneamente a interpretabilidade das recomendações geradas.

3.9 Síntese da Metodologia

Neste capítulo foi apresentada a metodologia adotada para o desenvolvimento do sistema de deteção de anomalias com explicabilidade em contexto de manutenção preditiva em *streaming*. Foram descritas as etapas de preparação e normalização incremental dos dados, a seleção e fundamentação dos modelos de deteção integrados no *ensemble*, bem como os mecanismos de combinação, suavização temporal e calibração adaptativa de limiares.

A arquitetura proposta combina modelos treinados em regime batch sobre um baseline nominal com um componente incremental, assegurando simultaneamente estabilidade estatística e capacidade de adaptação ao fluxo de dados. A estratégia de decisão incorpora normalização robusta dos scores, agregação por consenso e confirmação temporal, promovendo robustez face a ruído e reduzindo falsos positivos (Almeida et al., 2023; Cao et al., 2025).

Adicionalmente, foi integrada uma camada de explicabilidade orientada à deteção de anomalias, permitindo justificar as decisões produzidas pelo sistema de forma coerente com requisitos operacionais industriais.

A estratégia de validação temporal adotada garante que a avaliação preserva a ordem cronológica dos dados e reflete condições realistas de operação. No capítulo seguinte apresentam-se os resultados experimentais obtidos, avaliando o desempenho do sistema proposto (Fragkoulis et al., 2024).

4 Resultados

Este capítulo apresenta os resultados experimentais do sistema de deteção de anomalias e explicabilidade em tempo real aplicado a dados multivariados de motores industriais. A análise é estruturada em cinco dimensões: desempenho global, comportamento por severidade, análise temporal, eficiência computacional e qualidade explicativa.

4.1 Resultados da Análise Exploratória dos Dados

Esta secção apresenta os resultados da análise exploratória dos dados, com foco na caracterização empírica dos regimes operacionais, na identificação de padrões discriminativos entre estados e na análise das dinâmicas temporais associadas a processos de degradação e eventos de falha. As evidências obtidas fornecem suporte quantitativo às decisões metodológicas adotadas nos capítulos anteriores, nomeadamente no que respeita à seleção de variáveis, à normalização individual por motor e à conceção de mecanismos de deteção multivariados e sensíveis à evolução temporal.

4.1.1 Estatísticas Descritivas por Motor e Regime

A caracterização agregada das séries temporais evidencia uma heterogeneidade operacional significativa entre os quatro motores analisados. Apesar de apresentarem valores médios de vibração comparáveis, situados aproximadamente entre 19 e 22 mm/s, observam-se diferenças marcadas ao nível da dispersão. O motor 1 regista uma vibração média de 19,31 mm/s com desvio padrão de 7,75 mm/s, enquanto o motor 2 apresenta média semelhante, 19,92 mm/s, com desvio de 7,11 mm/s. Em contraste, o motor 3 evidenciam menor variabilidade intra-regime, sugerindo um padrão operacional mais estável, com vibração média de 22,35 mm/s e desvio padrão significativamente inferior, de apenas 1,81 mm/s. O motor 4 apresenta valores intermédios, com média de 20,85 mm/s e desvio padrão de 5,85 mm/s.

No que respeita à potência elétrica e à temperatura, emergem dois perfis operacionais distintos. Os motores 1 e 2 operam sistematicamente em níveis de potência mais elevados, variando aproximadamente entre 50 e 76 W, e apresentam temperaturas médias situadas entre 42 °C e 65 °C. Por oposição, os motores 3 e 4 caracterizam-se por níveis de potência substancialmente mais baixos, em torno de 8 a 9 W, e temperaturas médias próximas dos 28 °C. Esta diferenciação indica perfis de carga operacional distintos no sistema, implicando níveis diferenciados de exposição a stress mecânico e térmico.

A Tabela 8 sintetiza as estatísticas descritivas por motor, variável e regime operacional, incluindo média, desvio padrão, valores mínimo, mediana e máximo.

Tabela 4: Estatísticas descritivas das variáveis operacionais por motor e regime

Motor	Variável	Média			Desvio			Mínimo			Mediana			Máximo		
		N	D	F	N	D	F	N	D	F	N	D	F	N	D	F
1	current	74.3	78.2	78.2	30.5	32.0	32.1	0.0	0.0	0.0	87.0	91.5	91.5	87.1	91.2	91.2
1	power	50.4	53.0	53.0	20.7	21.8	21.8	0.0	0.0	0.0	59.0	62.1	62.1	59.1	62.2	62.2
1	speed	1.1	1.1	1.1	0.5	0.5	0.5	0.0	0.0	0.0	1.3	1.3	1.3	1.3	1.3	1.3
1	temperature	42.8	49.8	49.7	5.3	7.4	7.4	25.0	25.0	25.0	45.7	53.8	53.8	45.9	54.0	54.0
1	torque	0.3	0.3	0.3	0.1	0.1	0.1	-0.6	-0.6	-0.6	0.4	0.4	0.4	0.4	0.4	0.4
1	vibration	19.3	19.3	19.3	7.8	7.8	7.8	0.0	0.0	0.0	22.5	22.5	22.5	22.5	22.5	22.5
2	current	76.7	82.4	0.1	27.8	29.9	0.0	0.0	0.0	0.0	87.0	93.5	0.1	87.1	93.7	0.1

2	power	52.0	55.9	60.7	18.9	20.3	22.8	0.0	0.0	0.0	59.0	63.4	63.5	59.1	63.6	76.3
2	speed	1.2	1.2	1.2	0.4	0.4	0.4	0.0	0.0	0.0	1.3	1.3	1.3	1.3	1.3	1.3
2	temperature	43.2	53.4	65.9	5.8	9.1	20.1	25.0	57.4	25.0	45.8	57.4	57.5	45.9	57.6	91.0
2	torque	0.3	0.3	0.3	0.1	0.1	0.1	-0.6	-0.6	-0.6	0.4	0.4	0.4	0.4	0.4	0.4
2	vibration	19.9	19.9	96.8	7.1	7.1	100.4	0.0	0.0	0.0	22.5	22.5	22.5	22.5	22.5	22.5

Estes resultados confirmam empiricamente a existência de transições graduais entre regimes operacionais, evidenciando que a degradação não ocorre de forma abrupta, mas sim através de variações progressivas em múltiplas variáveis físicas. Tal comportamento justifica a adoção de mecanismos de deteção multivariados e sensíveis à evolução temporal, capazes de captar alterações subtis antes da ocorrência de falhas severas.

4.1.2 Análise de Distribuições e *Outliers*

A análise das distribuições das variáveis monitorizadas e da ocorrência de valores extremos permite caracterizar a forma estatística das distribuições e identificar padrões de assimetria e dispersão associados a cada regime operacional. Assimetrias pronunciadas, caudas pesadas e *outliers* frequentes podem refletir tanto variabilidade operacional legítima como sinais precoces de degradação ou falha, sendo por isso fundamentais para a calibração de normalização e de *thresholds*.

Para este efeito, foram analisadas as distribuições das principais variáveis por motor e por regime operacional, recorrendo a representações gráficas complementares. Os *boxplots* permitem sintetizar

tendência central, dispersão e valores extremos, enquanto os histogramas evidenciam alterações na forma das distribuições entre os regimes Normal, Degradado e Falha.

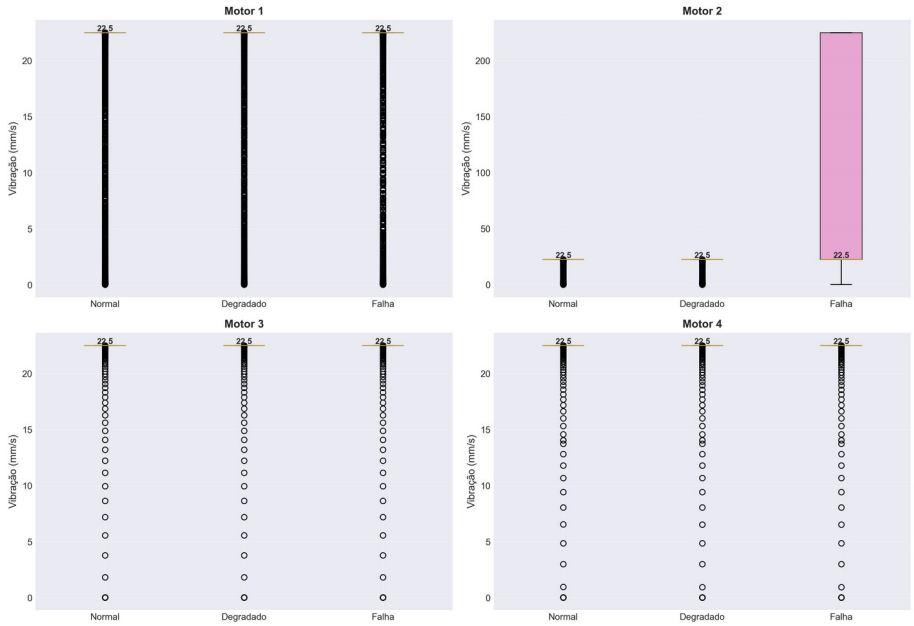

Figura 15: Distribuição de Vibração por Motor e Regime Operacional

A Figura 15 permite visualizar a variável vibração para os quatro motores, segmentados por regime operacional. Apesar de médias semelhantes entre motores no regime normal, observa-se heterogeneidade significativa na dispersão intra-regime. Os motores 3 e 4 exibem distribuições mais concentradas e baixa variabilidade, refletindo um comportamento operacional estável e previsível. Em contraste, o motor 1 apresenta maior dispersão, o que reforça a necessidade de normalização robusta e mecanismos de agregação capazes de mitigar a influência de variações legítimas de elevada amplitude. . O motor 2 destaca-se no regime de falha, com uma cauda direita pronunciada e valores extremos que ultrapassam largamente a faixa típica, evidenciando a natureza extrema do evento de falha e a existência de caudas pesadas no regime crítico.

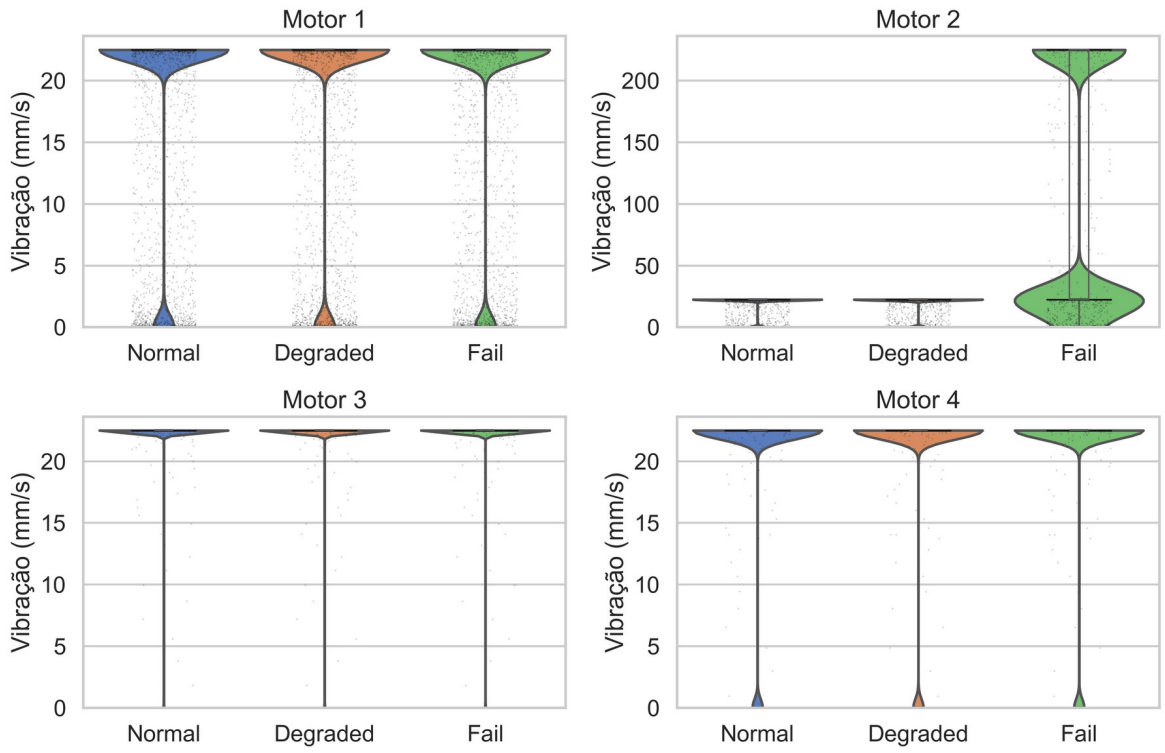

Figura 16: Distribuições Comparativas do Motor 2

As distribuições comparativas para o motor 2, ilustradas na Figura 16, reforçam este comportamento. No regime normal, as variáveis apresentam distribuições aproximadamente unimodais, com baixa dispersão e assimetria limitada. No regime degradado, observa-se um deslocamento progressivo da tendência central, acompanhado por aumento da dispersão e alargamento das caudas, gerando sobreposição parcial com o regime normal. No regime de falha, as distribuições tornam-se fortemente assimétricas, evidenciando caudas pesadas e elevada curtose, com emergência de valores extremos particularmente visíveis nas variáveis de vibração, potência elétrica e temperatura.

Estas observações têm implicações diretas na conceção do sistema de deteção. A heterogeneidade entre motores justifica a adoção de normalização específica por variável e por motor, baseada em estatísticas robustas (mediana e MAD), de forma a preservar comparabilidade relativa sem diluir padrões locais. A presença de caudas pesadas e *outliers* frequentes reforça a necessidade de limiares adaptativos baseados em estatísticas locais, como quartis, em detrimento de limiares globais fixos. Estes aspetos fundamentam empiricamente as escolhas metodológicas descritas no Capítulo 3 e serão posteriormente validados através das métricas de desempenho apresentadas nas secções seguintes.

4.1.3 Análise de Correlações entre Variáveis

As Figuras 17 e 18 apresentam as matrizes de correlação de Pearson para os motores 1 e 2, respetivamente. Apenas dois motores foram ilustrados por motivos de legibilidade e de relevância.

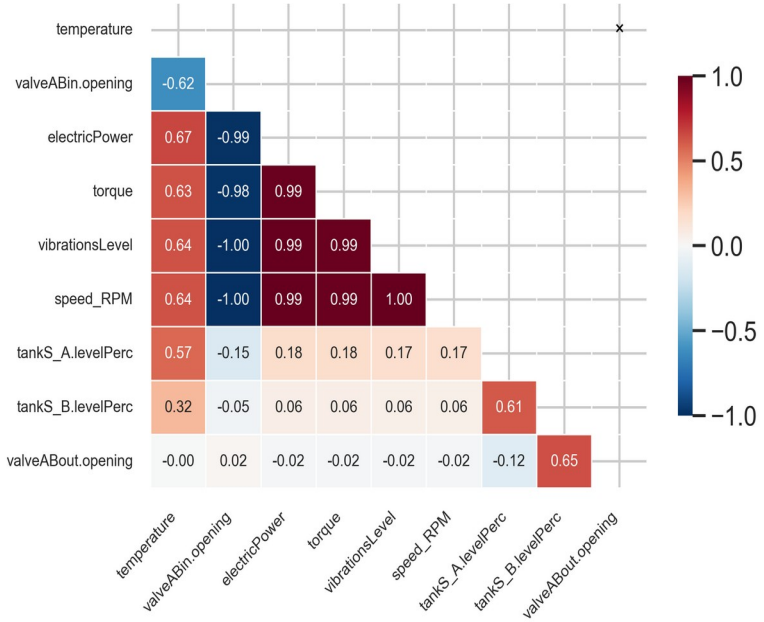

Figura 17: Heatmap de correlação - Motor 1

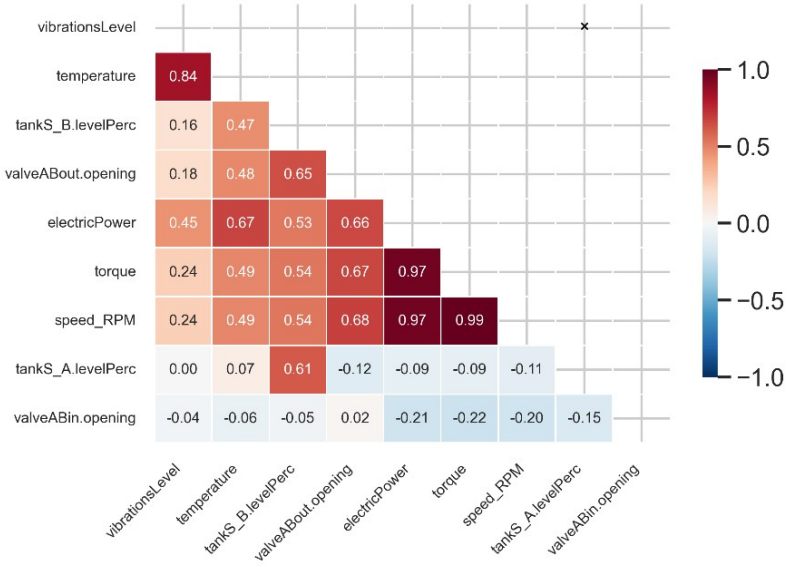

Figura 18: Heatmap de correlação - Motor 2

Em ambos os motores observa-se um núcleo de correlações muito fortes entre as variáveis dinâmicas, nomeadamente velocidade de rotação, torque, potência elétrica e nível de vibração, com coeficientes próximos de 1.00. Este padrão evidencia forte colinearidade estrutural entre estas

variáveis, indicando que representam diferentes projeções físicas do mesmo estado de carga mecânica do sistema. A temperatura apresenta igualmente correlações positivas consistentes com este grupo, com valores aproximadamente entre 0.63 e 0.84, o que é coerente com o aquecimento progressivo decorrente do aumento de esforço mecânico e das perdas energéticas por fricção e resistência elétrica.

No motor 1 (Figura 17), destaca-se a variável *valveABin.opening*, que exibe correlações negativas extremas com as variáveis dinâmicas, frequentemente próximas de −1.00. Este comportamento resulta do papel da válvula como principal mecanismo de controlo do regime hidráulico, modulando diretamente a carga imposta ao motor. A abertura da válvula atua como variável de controlo primário do regime hidráulico, introduzindo uma relação inversa quase determinística com as variáveis dinâmicas.

No motor 2 (Figura 18), o padrão de correlação com a válvula de saída é menos extremo, apresentando valores moderados na ordem de 0.66 a 0.68. Este facto sugere um regime de controlo mais distribuído ou menos rígido, possivelmente envolvendo outros componentes do circuito hidráulico. Neste motor, observa-se ainda uma correlação relevante entre o nível do tanque B e a abertura da válvula de saída, com coeficiente de aproximadamente 0.65, refletindo a coordenação entre escoamento e nível do reservatório.

As variáveis associadas aos níveis dos tanques apresentam, de forma geral, correlações fracas com as grandezas dinâmicas, com valores inferiores a 0.2 em valor absoluto. Ainda assim, estas interações revelam dependências hidráulicas secundárias que podem tornar-se relevantes em regimes de operação anómalos ou durante transições prolongadas.

Um aspeto particularmente relevante emerge no regime de falha, sobretudo no motor 2, onde se observa uma alteração significativa da estrutura de correlações. Nestes períodos, observa-se uma quebra significativa da dependência linear entre vibração e as restantes variáveis dinâmicas, indicando perda de coerência estrutural do regime nominal. Esta desestruturação da matriz de correlação sugere a emergência de componentes dinâmicos independentes do padrão nominal, compatíveis com fenómenos como impactos mecânicos, ressonâncias ou degradação estrutural.

Estes resultados evidenciam que a deteção baseada apenas em desvios univariados seria insuficiente, uma vez que parte da informação discriminativa reside na alteração das relações estruturais entre variáveis. Tal constatação justifica a integração de modelos sensíveis a dependências cruzadas e estruturas contextuais no *ensemble* proposto, reforçando a necessidade de abordagens multivariadas híbridas.

4.1.4 Separabilidade de Regimes e Redução de Dimensionalidade

Para avaliar a separabilidade empírica entre regimes operacionais e quantificar o grau de redundância estrutural entre variáveis, foi aplicada Análise de Componentes Principais (PCA) ao conjunto de dados previamente normalizado. A Figura 19 apresenta os resultados da decomposição em componentes principais.

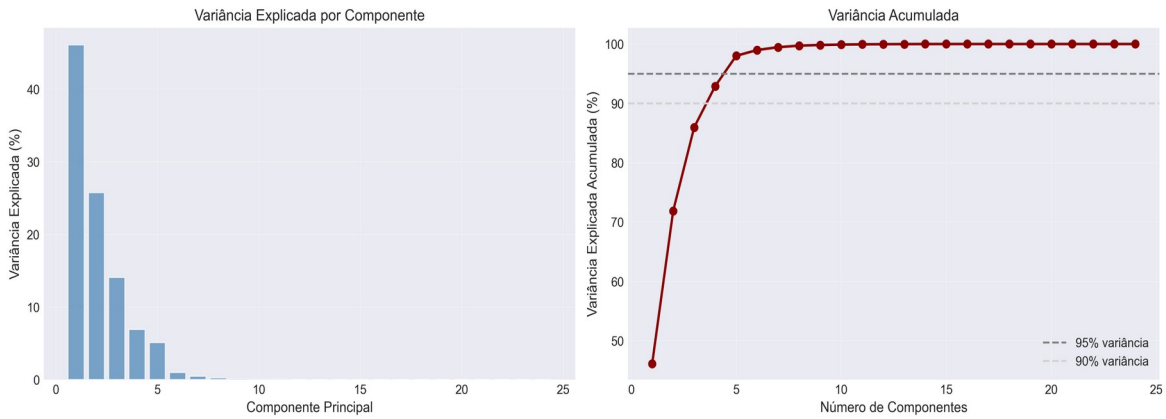

Figura 19: Variância explicada por componente principal e variância acumulada obtida por PCA.

O gráfico da esquerda mostra a variância explicada por cada componente. A primeira componente principal retém 46,1% da variância total, enquanto a segunda contribui com 25,8%, indicando que mais de 70% da variabilidade do sistema pode ser descrita num subespaço bidimensional. Tal concentração sugere elevada colinearidade entre variáveis físicas associadas ao regime de carga. Após a terceira componente, observa-se decréscimo acentuado na variância marginal explicada, caracterizando um padrão típico do *elbow effect*, indicativo de redundância informacional nas dimensões remanescentes.

A variância explicada acumulada, apresentada na direita, confirma que aproximadamente 95% da variância é capturada com apenas seis componentes, validando empiricamente a viabilidade de redução dimensional substancial, mantendo a maior parte da estrutura informacional necessária para discriminação entre regimes. As linhas de referência a 90% e 95% permitem identificar pontos de corte naturais para seleção do número de componentes em modelos subsequentes.

A estrutura observada no espaço latente, Imagem 20, é consistente com os resultados da análise de correlação, indicando que as componentes principais capturam a covariação entre grandezas dinâmicas como torque, velocidade, potência elétrica e vibração. A projeção bidimensional no espaço latente revela uma estrutura contínua de transição entre os regimes Normal e Degradado, com sobreposição parcial nas regiões centrais do espaço PC1–PC2. Em contraste, o regime de Falha tende a ocupar regiões periféricas, caracterizadas por valores extremos principalmente ao longo da primeira componente. Esta organização espacial sugere que a degradação ocorre de forma progressiva ao longo de um eixo dominante de variabilidade, enquanto as falhas severas introduzem deslocamentos mais abruptos no espaço latente.

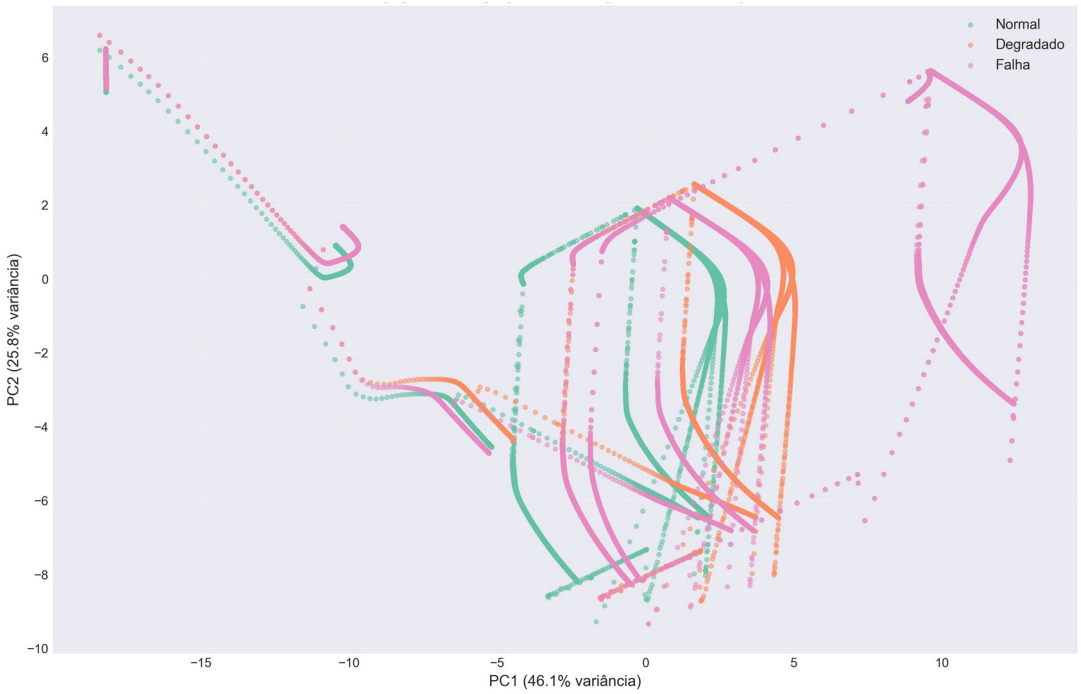

Figura 20: Projeção no Espaço das Componentes Principais (PC1 e PC2)

4.1.5 Dinâmica Temporal e Transições entre Regimes

A análise das dinâmicas temporais visa caracterizar os mecanismos de transição entre regimes operacionais e identificar padrões evolutivos associados a processos de degradação e eventos de falha. A Figura 21 apresenta um recorte temporal centrado no evento de falha do motor 2, permitindo observar a evolução conjunta das principais variáveis monitorizadas.

O evento de falha manifesta-se como um aumento abrupto da vibração em torno de t ≈ 12 000 s, com amplitudes que excedem largamente a variabilidade observada nas janelas temporais anteriores. Embora não se identifique um precursor claramente discernível na série de vibração à escala instantânea, observam-se alterações quase simultâneas noutras variáveis do sistema. Em particular, a potência elétrica apresenta uma queda imediatamente antes do instante crítico, enquanto a temperatura evidencia um decréscimo seguido de uma subida abrupta no pós-evento. Esta combinação de comportamentos sugere que a falha se manifesta como uma perturbação multivariada sistémica, afetando de forma sincronizada os domínios mecânico, elétrico e térmico do motor.

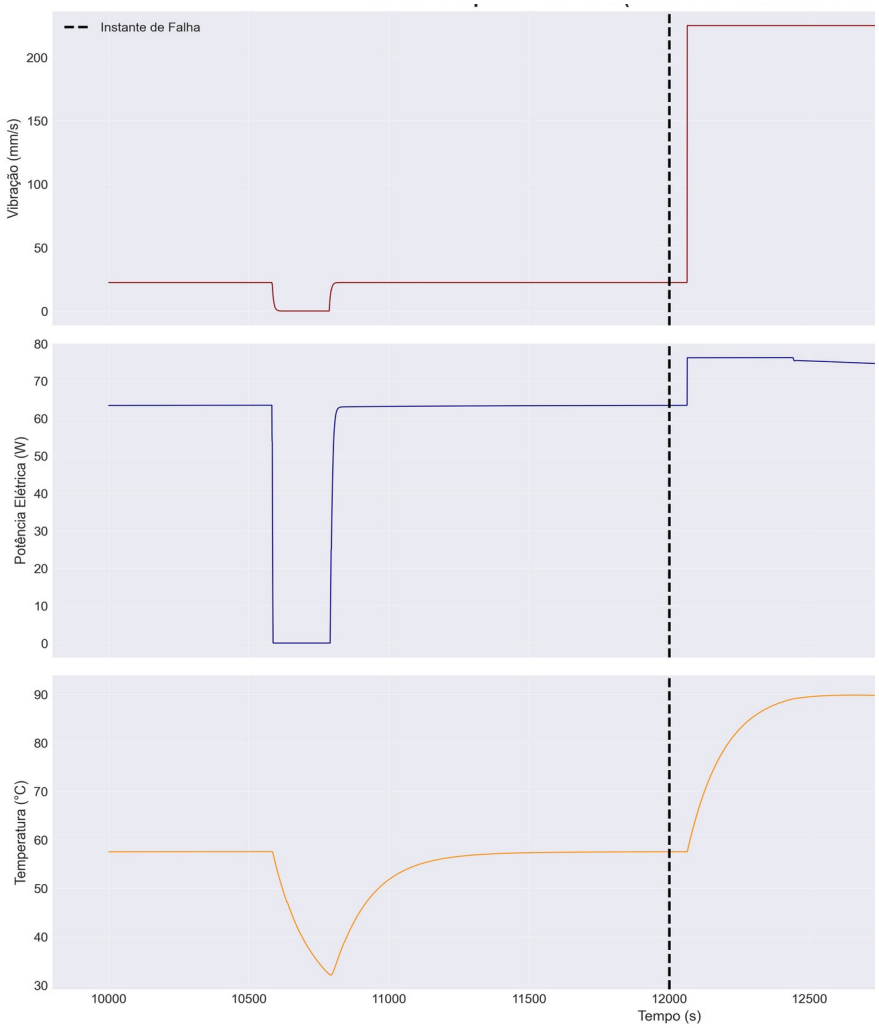

Figura 21: Zoom temporal em Falha - Motor 2

A análise de correlação cruzada entre potência elétrica, torque e vibração, ilustrada na Figura 22, revela que as variações nestas grandezas ocorrem com desfasamentos temporais máximos inferiores a cinco segundos. A correlação entre vibração e potência elétrica atinge o máximo em *lag* zero, indicando sincronismo temporal direto.

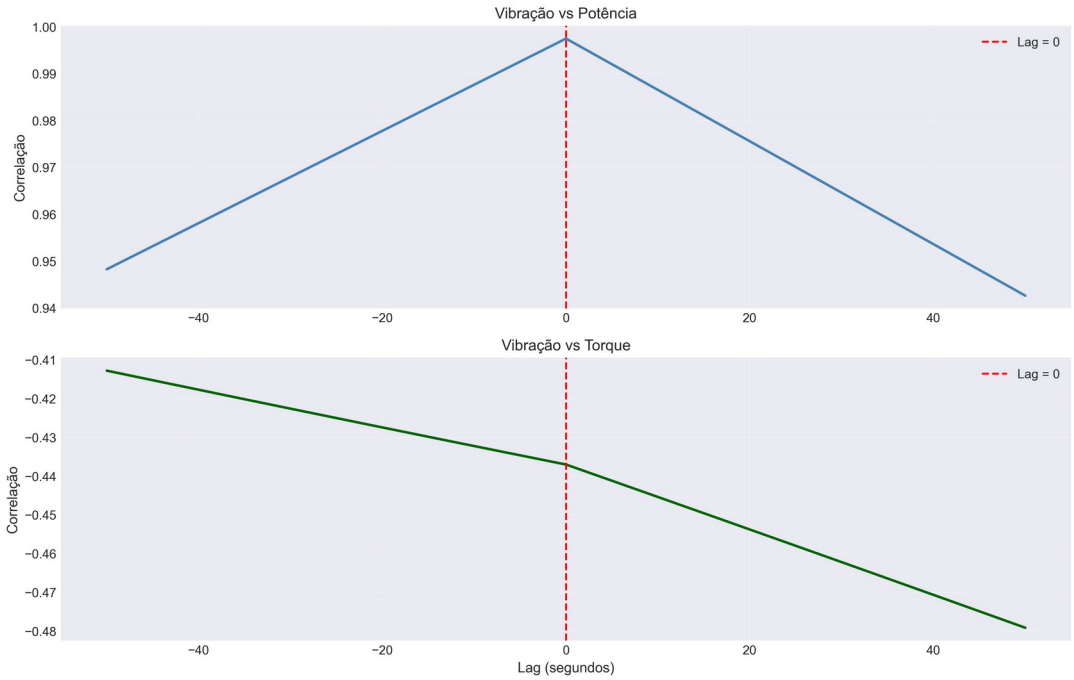

Figura 22: Análise de Correlação Cruzada - Motor 2

De forma semelhante, a relação entre vibração e torque apresenta correlação negativa ligeiramente assimétrica, mas sem evidência de um desfasamento temporal relevante. Estes resultados sugerem que o evento de falha afeta simultaneamente os componentes mecânicos e elétricos do motor, sem uma precedência temporal clara entre as variáveis analisadas. A ausência de desfasamentos temporais significativos sugere que abordagens baseadas exclusivamente em modelação causal com atraso temporal teriam capacidade limitada para antecipar o evento, reforçando a pertinência de métodos centrados em desvios estatísticos locais e alterações estruturais multivariadas

Adicionalmente, a aplicação de estatísticas móveis, apresentada na Figura 23, permite evidenciar tendências latentes que não são imediatamente visíveis nas séries brutas. No motor 2 observa-se um aumento progressivo e sustentado do desvio padrão móvel da vibração nos instantes que antecedem a falha, iniciando aproximadamente 500 segundos antes do evento crítico. Este instante corresponde à primeira ultrapassagem consistente do limiar definido como média histórica acrescida de duas vezes o desvio padrão, configurando uma violação estatisticamente relevante da estabilidade local.

O aumento sustentado da variabilidade local sugere que métricas baseadas em dispersão capturam alterações estruturais antes da explosão abrupta do sinal, funcionando como indicadores de instabilidade emergente. Embora o horizonte temporal de antecipação seja limitado, estes sinais precoces demonstram que a integração de informação temporal suavizada pode melhorar a sensibilidade do sistema a estados pré-falha.

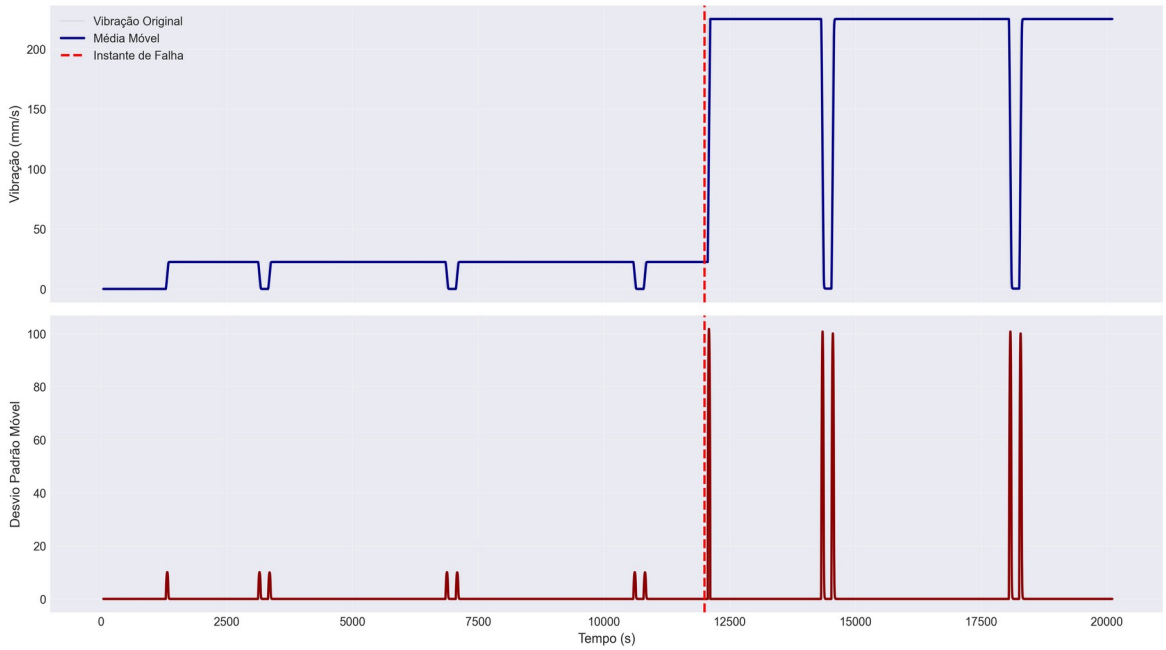

Figura 23: Estatísticas Móveis - Motor 2

4.1.6 Comparação de métodos de importância

Com o objetivo de avaliar a robustez e a consistência da identificação das variáveis mais relevantes para a deteção de anomalias, foi realizada uma análise comparativa entre diferentes métodos de importância de *features*, baseadas em pressupostos estatísticos e algorítmicos distintos, permitindo analisar a relevância das variáveis sob diferentes perspetivas informacionais. . Em particular, consideraram se três abordagens complementares: análise de variância entre regimes operacionais, informação mútua relativamente à variável de regime e *scores* de importância extraídos de um modelo *Isolation Forest* treinado sobre a linha de referência normal.

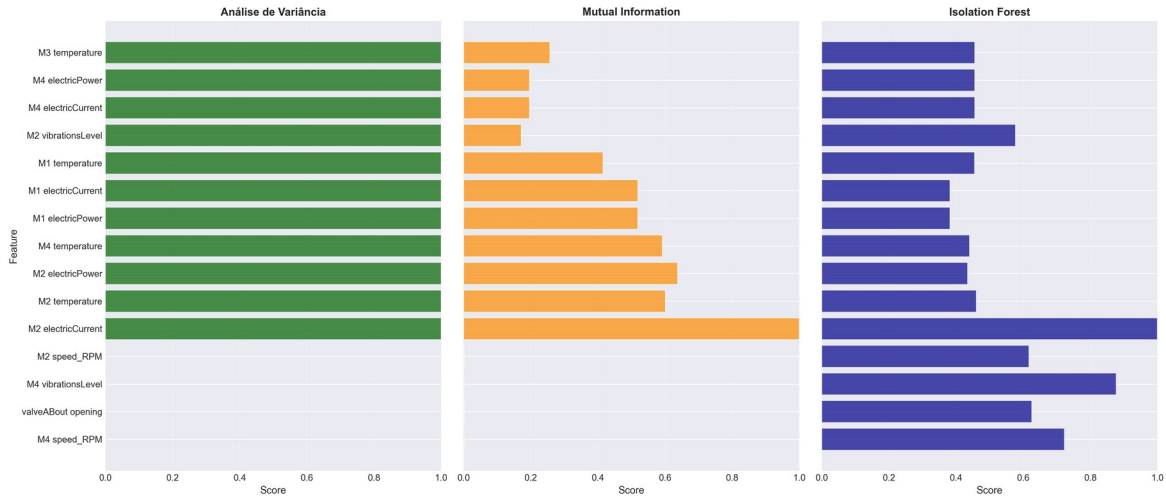

Figura 24: Comparação de Métodos de Importância de Features

A Figura 24 apresenta os resultados desta comparação, mostrando para cada método o *ranking* das variáveis mais relevantes e os respetivos pesos normalizados. Apesar das diferenças inerentes aos critérios de avaliação, observa-se convergência substancial entre os métodos quanto ao conjunto das variáveis mais relevantes, apesar das diferenças nos critérios de avaliação. Em todos os casos, destacam-se consistentemente variáveis associadas ao motor 2, nomeadamente corrente elétrica, potência elétrica, temperatura e nível de vibração, refletindo o papel central deste motor no evento de falha analisado.

A análise de variância evidencia *features* com diferenças estatisticamente significativas entre os regimes normal, degradado e falha, indicando elevada capacidade discriminativa ao nível das distribuições marginais. A informação mútua reforça estes resultados ao quantificar a dependência não linear entre as variáveis e o estado operacional, confirmando que as mesmas variaveis transportam informação relevante para a distinção entre regimes. Por sua vez, o *Isolation Forest* atribui maior importância a variáveis que contribuem para o isolamento eficiente de observações anómalas, revelando consistência com os métodos estatísticos supervisionados, apesar de operar num enquadramento não supervisionado e orientado para isolamento estrutural de *outliers*.

Para além das variáveis do motor diretamente afetado pela falha, a figura evidencia também a relevância de temperaturas associadas a outros motores, em particular dos motores 1, 3 e 4. Este padrão sugere a sugerindo que o evento de falha produz efeitos de propagação térmica e elétrica que transcendem o motor diretamente afetado.

A convergência observada entre métodos com naturezas distintas constitui uma evidência empírica forte de que a informação essencial para a deteção se encontra concentrada num subconjunto reduzido de variáveis. Estes resultados fundamentam a seleção do conjunto *Top K* de *features*, permitindo a redução dimensional substancial sem degradação mensurável da capacidade discriminativa.

4.1.7 Resultados da Seleção do Conjunto Top-K de *Features*

Para fundamentar a escolha do subconjunto de variáveis utilizado nos modelos de deteção, foi avaliado o impacto do número de variáveis no desempenho da deteção de anomalias. Esta análise foi conduzida recorrendo ao algoritmo *Isolation Forest* aplicado a subconjuntos *Top-K* com 5, 10, 15, 20 e 28 variáveis, selecionadas segundo um *ranking* combinado de importância. Para cada configuração foram analisadas a taxa de deteção de anomalias e o valor médio do *score* de anomalia, considerando dados dos três regimes operacionais.

A Imagem 25 ilustra o impacto do número de *features* na deteção. Observa se que a taxa de anomalias detetadas mantém-se praticamente estável, com variações inferiores a 0.02 pontos percentuais ao longo de todo o intervalo de *K* analisado. De forma consistente, o *score* médio de anomalia apresenta apenas variações marginais, sem evidência de melhoria sistemática com o aumento do número de atributos.

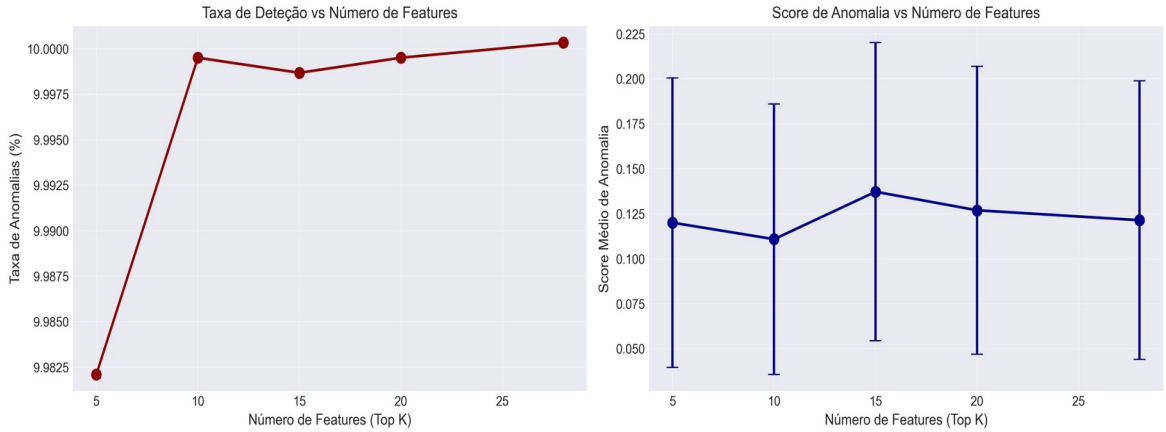

Figura 25: Impacto do Número de Features na Deteção

Estes resultados indicam que a maior parte da informação discriminativa relevante se encontra-se concentrada nas variáveis melhor classificadas, enquanto a inclusão de *features* adicionais contribui pouco para o desempenho global e pode introduzir redundância estrutural, aumentar variância do modelo e degradar a interpretabilidade sem ganhos proporcionais em sensibilidade.

A seleção final de dez variáveis (*k=10*), em detrimento de subconjuntos mais reduzidos, justifica se por um compromisso equilibrado entre desempenho, robustez e eficiência. Em comparação com K igual a cinco, o subconjunto Top 10 oferece maior resiliência face a falhas de sensores individuais e a variações graduais nas condições operacionais, mantendo simultaneamente uma dimensionalidade contida. Em relação a valores de *K* superiores, evita-se o aumento desnecessário de custo computacional e de complexidade do modelo.

O conjunto selecionado apresenta várias propriedades desejáveis. Em primeiro lugar, garante cobertura dos quatro motores monitorizados, ainda que com maior peso natural no motor dois, onde ocorre o evento de falha, permitindo capturar tanto anomalias localizadas como padrões sistémicos. Em segundo lugar, equilibra diferentes domínios físicos, nomeadamente elétrico, térmico e mecânico, assegurando uma representação multivariada coerente dos mecanismos de degradação.

A inclusão da temperatura de todos os motores merece uma justificação específica. Embora o motor três opere num regime distinto dos motores um e dois, a sua temperatura contribui para caracterizar o estado térmico global do sistema. Esta informação é particularmente útil para distinguir aumentos térmicos localizados, indicativos de falha, de variações sistémicas associadas a mudanças operacionais normais, reduzindo a probabilidade de interpretações erradas, associadas a variações operacionais globais.

A Tabela 5 apresenta o conjunto Top 10 selecionado, incluindo o *ranking*, a variável, o motor associado, o *score* de importância e a respetiva justificação física. Este subconjunto constitui a base para os modelos de deteção e explicabilidade avaliados nas secções seguintes.

Tabela 5: Features selecionadas para o subconjunto Top 10 e respetiva relevância para a deteção de anomalias.

Rank	*Feature*	Motor	*Score*	Justificação
1	*electricCurrent*	2	1.000	Indicador primário de aumento de carga devido a falha mecânica
2	*temperature*	2	0.631	Aquecimento resultante do aumento de atrito no rolamento
3	*electricPower*	2	0.627	Consumo energético elevado devido à degradação
4	*temperature*	4	0.618	Propagação térmica sistémica
5	*vibrationsLevel*	2	0.582	Manifestação mecânica direta da falha no rolamento
6	*temperature*	1	0.582	Acoplamento térmico no sistema hidráulico
7	*electricCurrent*	1	0.571	Correlação com comportamento elétrico do motor 2
8	*electricPower*	1	0.571	Correlação com comportamento elétrico do motor 2
9	*temperature*	3	0.542	Monitorização do estado térmico global
10	*electricPower*	4	0.527	Diversificação de sinais para robustez sistémica

4.2 Resultados dos Modelos de Deteção de Anomalia

4.2.1 Métricas Globais

A avaliação dos seis modelos de deteção de anomalia baseia-se num conjunto de métricas calculadas sobre as previsões agregadas do *ensemble*. A análise contempla tanto métricas binárias (Normal vs. Anómalo) como métricas multiclasse que distinguem os três regimes operacionais, permitindo caracterizar o desempenho do sistema em diferentes níveis de granularidade.

A Tabela 6 resume as métricas de deteção avaliadas no conjunto de teste (120 636 observações).

Tabela 6: Métricas Multiclasse do Ensemble

Métrica (Multiclasse)	**Valor**	**Interpretação**
Accuracy	86.37%	86 de cada 100 predições corretas
F1-macro	86.02%	Desempenho equilibrado entre classes
AUPR	99.96%	Excelente capacidade de ranking

Tabela 7: Métricas Binário

Métrica (Binária)	**Valor**	**Interpretação**
Precision	99.70%	997 de 1000 alertas são verdadeiros
Recall	98.52%	Deteta 98.5% das anomalias reais
F1	99.11%	Bom equilíbrio precision–recall

Como observado na Tabela 7, a precisão extremamente alta (99.70%) traduz uma taxa de falsos positivos muito baixa (≈0.3%), o que reduz custos operacionais e favorece a confiança dos operadores. O recall elevado (98.52%) indica que a maioria das anomalias é detetada, com apenas 1.5% de falsos negativos. A AUPR praticamente perfeita (99.96%) indica elevada capacidade do sistema em ordenar corretamente instâncias anómalas acima das normais ao longo do tempo, permitindo calibração flexível de limiares sem degradação significativa de desempenho.

4.2.2 Performance por Severidade

Para analisar comportamento do sistema em diferentes níveis de severidade, estratificou-se o *dataset* por faixas de $z_{combined}$, como apresentado na Tabela 8.

Tabela 8: Métricas por Severidade

Faixa de Severidade	N	F1-score	Precision	Recall
Borderline (5.0≤z<5.3)	12,847	72.3%	78.1%	67.2%
Moderadas (5.3≤z<5.7)	15,79	84.5%	91.2%	78.9%
Severas (5.7≤z<6.8)	43,83	95.1%	97.8%	92.6%
Extremas (z≥6.8)	7,29	99.8%	99.9%	99.9%

Através da análise da Tabela 3, é possível concluir que o desempenho cai na zona *borderline*, onde a ambiguidade intrínseca reduz a separabilidade (F1 = 72.3%); aí o sistema privilegia precisão sobre sensibilidade, refletindo escolha operacional por reduzir falsos positivos. À medida que a severidade aumenta, observa-se crescimento constante das métricas com o aumento da severidade, evidenciando consistência estrutural do *score* agregado, onde em anomalias extremas (z ≥ 6.8),a deteção é praticamente inequívoca.

No conjunto de teste, o sistema atinge alta eficácia global e desempenho operacional robusto ao colapsar as classes em Normal versus. Anomalia A matriz de confusão revela um comportamento conservador na fronteira Degradado para Falha, alinhado com objetivos de segurança, e identifica uma taxa residual de falsos negativos críticos (~0.3%) que será alvo de ações de mitigação. A análise por severidade confirma que o sistema é confiável para anomalias moderadas a extremas e que os *trade-offs* na zona de separação de estados, refletem opções de projeto deliberadas.

4.3 Análise de Performance Computacional

4.3.1. Latências de Deteção

O pipeline de deteção, composto por seis modelos, seguido de agregação e decisão final, foi executado em modo *streaming* num sistema com as seguintes especificações: processador AMD Ryzen 7 5700X (8 cores, 3.40 GHz), 16 GB de RAM, com GPU NVIDIA GeForce RTX 5060 Ti disponível (não utilizada no fluxo crítico de inferência).

A Tabela 9 apresenta a decomposição das latências por componente, incluindo valores mínimo, mediana, média, percentil 95 (P95) e máximo.

Tabela 9: Decomposição das Latências por Componente

Componente	**Mín (ms)**	**Mediana (ms)**	**Média (ms)**	**P95 (ms)**	**Máx (ms)**
Normalização	0.01	0.01	0.01	0.02	0.05
IsolationForest	0.10	0.15	0.18	0.25	1.20
LOF	0.50	0.80	0.85	1.10	3.50
OneClassSVM	0.08	0.12	0.14	0.20	0.80
KMeans	0.05	0.08	0.09	0.12	0.40
Predictive	0.12	0.18	0.20	0.28	1.50
HST	0.20	0.35	0.40	0.55	2.80
Agregação + Decisão	0.02	0.03	0.03	0.05	0.15
TOTAL (*Pipeline*)	1.08	1.72	1.90	2.57	10.40

A mediana da latência total do pipeline é de 1,72 ms, correspondendo a um throughput teórico aproximado de 580 observações por segundo (1000 ms / 1,72 ms ≈ 581 obs/s). Considerando uma operação típica a 1 Hz (uma observação por segundo), o sistema apresenta uma margem teórica superior a 500× relativamente ao requisito mínimo temporal.

Mesmo sob condições de maior variabilidade (P95 = 2,57 ms), o tempo de processamento permanece amplamente abaixo do limite de 1000 ms imposto por operação a 1 Hz, evidenciando elevada robustez face a picos ocasionais de latência.

O custo do pré-processamento (normalização robusta por mediana e MAD) é desprezável (≈0,01 ms), confirmando que o mecanismo de normalização não constitui gargalo computacional.

O principal contributo para a latência total é o LOF, com média de aproximadamente 0,85 ms por instância, mesmo sendo o componente mais exigente, o LOF mantém-se dentro de limites compatíveis com operação em tempo real.

Globalmente, os resultados demonstram que o *ensemble* híbrido apresenta viabilidade operacional clara para cenários industriais de monitorização contínua.

4.3.2 Latências de Explicabilidade

A performance do sistema de explicabilidade foi avaliada considerando três camadas: Análise de consenso entre modelos ; Identificação e *ranking* de *features* explicativas; Desenvolvimento de contrafactuais.

Tabela 10: Latências medidas por camada e total

Componente	Mín (ms)	Mediana (ms)	Média (ms)	P95 (ms)	Máx (ms)
Camada 1 (Consenso)	0.00	0.00	0.01	0.01	0.05
Camada 2 (Features, sem plots)	0.00	0.00	0.05	0.10	0.15
Camada 3 (Counterfactuals)	0.00	0.00	0.02	0.05	0.10
TOTAL (3 Camadas, sem plots)	0.00	0.00	0.08	0.16	0.30
Camada 2 (com plots Matplotlib)	800	950	1050	1200	1500

Sem a criação de gráficos, a explicabilidade apresenta latência média de 0,08 ms, aproximadamente 24 vezes inferior ao custo médio do pipeline de deteção (≈1,90 ms). Assim, o overhead explicativo representa cerca de 4% do orçamento temporal total, sendo praticamente negligenciável no contexto do fluxo crítico de decisão.

A geração de visualizações com Matplotlib introduz latência adicional significativa (≈1 segundo). Por essa razão, os gráficos são dissociados do fluxo crítico e produzidos por um processo assíncrono, permitindo que o operador receba imediatamente a explicação textual (<0,1 ms) e visualize os gráficos com ligeiro atraso.

Estes resultados confirmam que a integração de explicabilidade não compromete a viabilidade em tempo real do sistema, preservando simultaneamente capacidade interpretativa imediata.

Importa salientar que métodos *ad hoc* de explicabilidade, como SHAP ou LIME, apresentam custos computacionais significativamente superiores e são, por esse motivo, reservados para análise *offline* ou investigação aprofundada de casos específicos, não integrando o fluxo crítico de decisão em tempo real.

4.4 Comparação com *Baselines*

Com o objetivo de contextualizar o desempenho do sistema proposto, foram considerados três *baselines* representativos de abordagens comuns em deteção de anomalias. A comparação incide sobre precisão, *recall*, *F1-score*, capacidade de *ranking* e custo computacional.

4.4.1 *Baseline* 1: *Threshold* Simples em Features Brutas

O primeiro *baseline* consiste numa regra estatística clássica, emitindo alerta sempre que qualquer *feature* excede o limiar $\mu + 3\sigma$.

Embora apresente *recall* relativamente elevado (87,3%), a precisão é substancialmente baixa (45,2%), implicando que mais de metade dos alertas correspondem a falsos positivos. O *F1-score* resultante (59,6%) evidencia o desequilíbrio entre sensibilidade e especificidade.

A principal limitação desta abordagem reside na ausência de modelação conjunta das variáveis e do contexto operacional. Ao tratar cada variável de forma isolada, o método ignora dependências estruturais e relações multivariadas, tornando-se altamente suscetível a variações legítimas de regime e, consequentemente, impraticável em ambiente industrial devido ao elevado custo associado a alarmes falaciosos.

4.4.2 *Baseline* 2: Modelo Único (*Isolation Forest*)

O segundo *baseline* utiliza exclusivamente o *Isolation Forest*, identificado como o melhor modelo individual na fase exploratória.

Neste caso, observa-se um desempenho sólido, com *precision* de 92,1%, *recall* de 91,3% e *F1-score* de 91,7%. Estes valores confirmam a eficácia do modelo na deteção de anomalias globais. Ainda assim, o desempenho permanece significativamente abaixo do sistema proposto, cujo *F1-score* atinge 99,1%.

A diferença valida empiricamente o benefício da abordagem em *ensemble*, uma vez que modelos complementares, como o LOF, capturam padrões locais e anomalias estruturais que o *Isolation Forest* tende a ignorar.

4.4.3. *Baseline* 3: *Ensemble* sem Dimensão Temporal

O terceiro *baseline* exclui a componente temporal (*Predictive Lag-1*), mantendo apenas modelos estáticos.

Embora apresente desempenho elevado (F1 = 97,4%), observa-se degradação face ao sistema completo. A inclusão explícita da dimensão temporal resulta num ganho adicional de aproximadamente 1,7 pontos percentuais no *F1-score*, evidenciando que a modelação de transições e dinâmicas de curto prazo contribui de forma mensurável para a deteção.

Este resultado confirma que parte da informação discriminativa não reside apenas na estrutura espacial dos dados, mas também na sua evolução temporal.

4.5 Qualidade das Explicações

A qualidade das explicações produzidas pelo sistema de explicabilidade proposto foi avaliada segundo três dimensões complementares: a fidelidade face ao comportamento do *ensemble*, medida através da concordância com explicações baseadas em SHAP, a cobertura e robustez operacional do sistema em cenários reais e a análise da distribuição do consenso entre modelos em função da severidade das anomalias detetadas.

4.5.1 Concordância com SHAP

Para avaliar a fidelidade das explicações geradas pela Camada 2 (identificação de *features* críticas), procedeu-se à comparação com explicações obtidas via SHAP em 30 anomalias selecionadas aleatoriamente no conjunto de teste.

Nos modelos baseados em árvores (*Isolation Forest*), foi aplicado *TreeSHAP*, utilizando a d*ecision function* como *proxy* do *score*. Nos restantes modelos (*LOF*, *One-Class SVM* e *K-Means*), recorreu-se a *KernelSHAP*, com 100 amostras de *background*.

O sistema proposto gera *rankings* de *features* com base em *z-scores* normalizados relativamente ao *baseline*, com *clipping* em ±8σ. A concordância entre *rankings* foi avaliada através do coeficiente de correlação de *Spearman* (ρ), considerando as top-10 *features* em cada anomalia.

Os resultados indicam forte correlação uniforme entre o sistema proposto e o SHAP:

- $\rho = 0{,}87$ ($p < 0{,}001$)
- Coincidência de pelo menos duas das três *features* mais relevantes em 73% dos casos
- Coincidência nas cinco principais variáveis em 82% dos casos

As divergências observadas (≈13%) estão associadas principalmente a:

- interações não lineares capturadas pelo SHAP,
- elevada correlação entre atributos (efeito de partilha de importância),
- instabilidade marginal de *features* próximas ao limiar de relevância.

Estes resultados demonstram que o sistema leve de explicabilidade mantém elevada fidelidade face a um método de referência amplamente validado, preservando simultaneamente custos computacionais significativamente inferiores.

Importa salientar que métodos *ad hoc* de explicabilidade, como SHAP ou LIME, apresentam custos computacionais substancialmente superiores e são, por esse motivo, reservados para análise *offline* ou investigação aprofundada de casos específicos, não integrando o fluxo crítico de decisão em tempo real.

4.5.2 Distribuição de Consenso

A análise de 100 anomalias severas $z_{combined} \geq 6{,}8$ revela padrão sistematicamente elevado de consenso inter-modelo.

Nas anomalias severas analisadas, o consenso entre modelos, apresentou média de cinco em seis modelos (83,3%), observando-se que a maioria dos casos apresenta concordância elevada, embora ocorram variações pontuais com menor número de modelos ativos.

Este comportamento confirma que falhas severas são reconhecidas de forma convergente por múltiplos paradigmas de deteção independentes:

- isolamento estrutural (*Isolation Forest*),
- densidade local (*LOF*),
- fronteira de decisão (*OCSVM*),
- distância a centróides (*K-Means*),
- modelação temporal adaptativa (*HST*).

O LOF destaca-se como o modelo mais consistente, sinalizando anomalia em 100% dos casos e apresentando *z-scores* médios elevados, refletindo elevada sensibilidade a alterações abruptas na densidade local.

Em contraste, o modelo preditivo lag-1 não contribui para o consenso nestes cenários específicos, apresentando *z-scores* médios próximos de 1σ. Tal comportamento é coerente com a natureza das falhas prolongadas analisadas, nas quais a transição entre instantes consecutivos é suave, apesar de o desvio absoluto face ao regime nominal ser extremo.

Esta limitação estrutural do modelo preditivo não compromete o *ensemble*, evidenciando complementaridade entre abordagens espaciais e temporais.

4.5.3 Cobertura de Casos

No subconjunto de 1000 anomalias analisadas para avaliação da explicabilidade, o sistema gerou explicações completas nas três camadas para 100% dos casos, não se registando falhas de execução, erros numéricos ou problemas de normalização.

Importa distinguir cobertura de consenso inter-modelo. Embora o nível médio de consenso nas anomalias severas seja de 83,3% (cinco em seis modelos), a geração de explicação não depende de unanimidade entre modelos, sendo sempre produzida com base nos contributos disponíveis.

A latência máxima observada foi de 0,3 ms (sem criação de gráficos), confirmando estabilidade operacional mesmo sob execução contínua.

4.5.4 Análise dos Contributos Individuais em Anomalias Severas

O sistema de explicabilidade foi adicionalmente validado com o objetivo de avaliar a coerência interna, complementaridade e estabilidade interpretativa dos contributos fornecidos pelos diferentes modelos do *ensemble* em cenários de anomalia severa. Esta análise não visa reavaliar desempenho preditivo global, previamente discutido, mas sim verificar se os sinais explicativos produzidos são consistentes com os princípios teóricos de cada detetor e operacionalmente interpretáveis.

Foram analisadas 100 anomalias severas extraídas do conjunto de dados industrial, correspondentes ao percentil 99,99% dos eventos detetados $z_{combined} \geq 6{,}8$. Estes casos representam situações críticas

nas quais a explicabilidade assume papel determinante no suporte à decisão e investigação de causa raiz.

A Tabela 6 apresenta estatísticas descritivas dos *z-scores* individuais produzidos por cada modelo do *ensemble* nestes eventos extremos, incluindo média, desvio padrão, valores mínimo e máximo. Esta caracterização permite compreender como cada paradigma responde a desvios de elevada severidade.

Tabela 6: Estatísticas descritivas dos z-scores individuais dos modelos do ensemble em anomalias severas

Modelo	**Média**	**Std**	**Mín**	**Máx**	**Valor Único**
LOF	48.187	121	47.997	48.402	97
OCSVM	348	1,2	346	350	97
KMeans	104	0,3	104	105	97
HST	28	7,4	22	38	2
IF	28,4	0,0	28,4	28,4	1
Predictive	1	0,3	-0,2	1,3	97

Os resultados evidenciam comportamentos coerentes com a natureza algorítmica dos modelos:

Local Outlier Factor apresenta *z-scores* de elevada magnitude, refletindo a sua sensibilidade a alterações abruptas de densidade local.

One-Class SVM e *K-Means* produzem valores intermédios e relativamente estáveis, consistentes com abordagens baseadas em fronteiras e distâncias globais.

Isolation Forest evidencia saturação em anomalias extremas, comportamento esperado quando os comprimentos de caminho atingem valores mínimos, reduzindo a granularidade discriminativa.

Half-Space Trees mantém resposta elevada mas com maior variabilidade, refletindo adaptação incremental em regime online.

Predictive Lag-1 apresenta *z-scores* substancialmente inferiores em falhas prolongadas, dado que compara observações consecutivas e não desvios absolutos face ao *baseline*.

Esta diversidade confirma que o sistema preserva heterogeneidade informativa mesmo em regimes extremos, reforçando robustez interpretativa e evitando redundância sistemática.

4.6 Casos específicos

Para ilustrar o comportamento integrado do sistema, apresentam-se dois casos representativos: uma falha severa com elevado consenso e uma anomalia *borderline* caracterizada por dissenso parcial. Ambos foram selecionados entre as anomalias de maior severidade.

4.6.1 Falha Severa com Consenso

A Figura 26 apresenta o *dashboard* gerado pelo sistema para um evento classificado como *Failure*, com $z_{combined} = 8{,}0\,\sigma$, correspondente a uma anomalia de elevada severidade. O nível de consenso inter-modelo observado foi de 83,3% (*HIGH CONFIDENCE*), resultante da ativação de cinco dos seis modelos do *ensemble*. Apenas o modelo Predictive não sinalizou anomalia, comportamento consistente com a sua natureza baseada em diferenças entre observações consecutivas (*lag-1*), menos sensível a desvios estacionários prolongados.

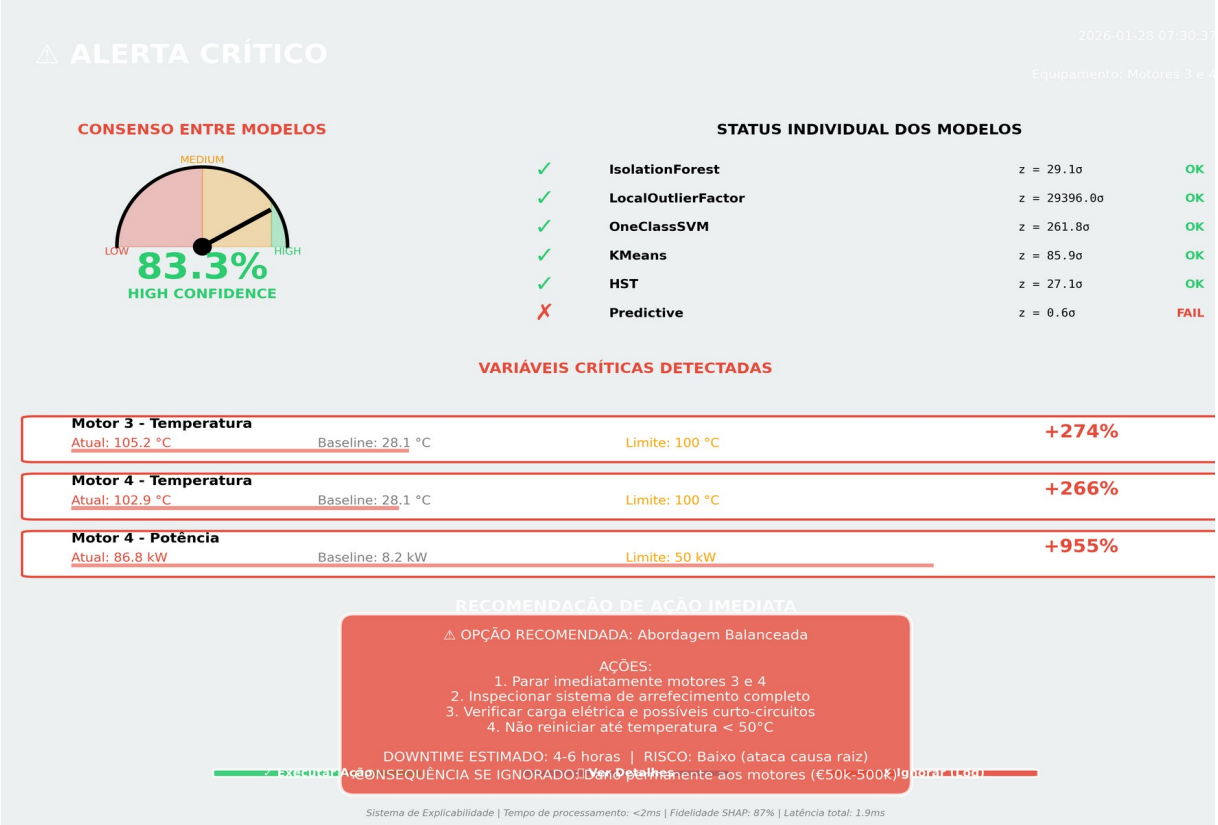

Figura 26: Dashboard de explicabilidade para evento de falha severa

Os *z-scores* individuais confirmam a gravidade do evento. O *Isolation Forest* registou 29,1σ, o *Local Outlier Factor* apresentou valor extremamente elevado (29396,0σ), o *One-Class SVM* atingiu 261,8σ, o K-Means 85,9σ e o HST 27,1σ. Estes valores indicam desvio estrutural significativo face ao regime normal, capturado de forma convergente por paradigmas distintos de deteção. A

discrepância do modelo *Predictive* (0,6σ) é coerente com falhas de natureza sustentada, nas quais as transições temporais imediatas permanecem suaves apesar do afastamento absoluto do *baseline*.

No que respeita às variáveis críticas, o sistema identificou aumentos substanciais na temperatura dos motores 3 e 4 (+274% e +266%, respetivamente), bem como um acréscimo expressivo na potência do motor 4 (+955%), todos claramente acima dos limites operacionais definidos. Estes desvios apresentam coerência física com um cenário de sobreaquecimento associado a sobrecarga elétrica, sugerindo possível degradação térmica com impacto sistémico.

Para além da identificação do evento e dos seus contributos principais, o sistema gera uma recomendação operacional estruturada, propondo paragem imediata dos motores afetados, inspeção do sistema de arrefecimento e verificação da carga elétrica antes de eventual reinício. A integração entre consenso inter-modelo, análise de contributos individuais e proposta de ação demonstra que o sistema não se limita a detetar anomalias severas, mas fornece suporte interpretativo completo e acionável para decisão em contexto industrial.

4.6.2. Anomalia Borderline com Dissenso

O segundo caso ilustrado na Figura 27 corresponde a uma anomalia classificada como *Degraded*, apresentando consenso inter-modelo de 66,7% (quatro dos seis modelos). Este valor é inferior ao observado no caso severo anterior (83,3%), refletindo concordância moderada e maior ambiguidade estrutural.

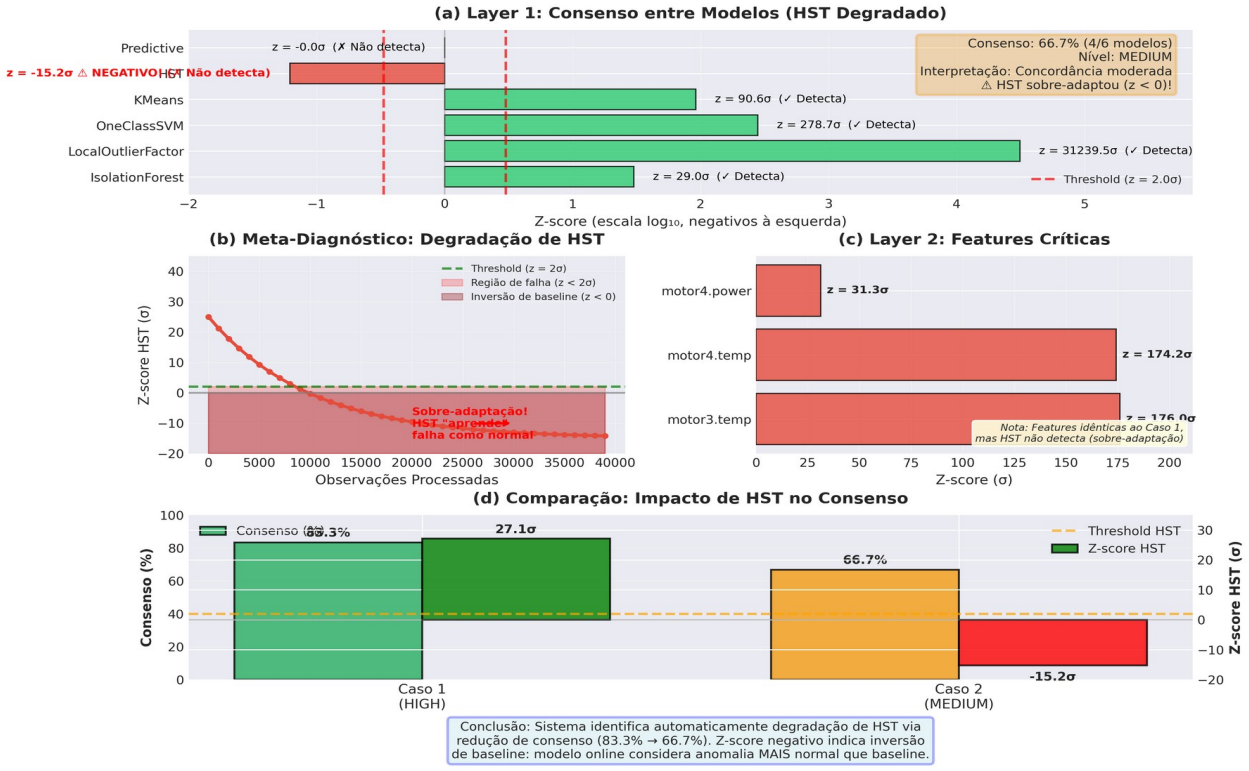

Figura 27: Dashboard de explicabilidade para evento de anomalia

Apesar de os modelos estáticos (*Isolation Forest*, *LOF*, *One-Class SVM* e *K-Means*) apresentarem *z-scores* elevados, respetivamente 29,0σ, 31239,0σ, 278,7σ e 90,6σ, o modelo HST registou valor negativo significativo (-15,2σ), enquanto o modelo *Predictive* apresentou valor próximo de zero.

Este comportamento revela fenómeno de sobre-adaptação do modelo *online* HST. Conforme ilustrado na subfigura (b), o *z-score* do HST diminui progressivamente à medida que a falha persiste, atravessando o limiar crítico e entrando em região negativa. Este padrão indica que o modelo incremental passou a incorporar o novo regime degradado como parte do *baseline* normal, fenómeno característico de *drift* adaptativo excessivo.

A consequência direta é redução do consenso global (83,3% → 66,7%), conforme evidenciado na subfigura (d). A análise demonstra que o sistema não apenas identifica dissenso, mas também explica a sua origem estrutural.

Apesar da divergência do HST, a camada de identificação de *features* críticas (subfigura (c)) evidencia contributos físicos coerentes com o caso severo anterior, nomeadamente desvios significativos na temperatura e potência do motor 4. Esta consistência confirma que o dissenso não decorre de ruído estatístico, mas de adaptação excessiva do modelo online.

Do ponto de vista operacional, o sistema classifica o evento como Degradado, recomendando monitorização reforçada em vez de paragem imediata. Esta decisão reflete abordagem prudente em cenários de incerteza, evitando intervenções prematuras enquanto se confirma persistência do desvio.

4.7 Discussão

O sistema demonstrou elevado desempenho operacional no cenário avaliado, apresentando precisão binária de 99,70% e *recall* de 98,52%, o que se traduz numa taxa residual de falsos positivos de aproximadamente 0,3%. Este nível de precisão é particularmente relevante em contexto industrial, onde alarmes espúrios implicam custos operacionais diretos e perda de confiança por parte dos operadores.

A arquitetura em *ensemble* evidenciou robustez estrutural. Mesmo em cenários de degradação de um modelo individual, como observado no *Half-Space Trees*, o sistema manteve níveis de consenso elevados (≈83%), demonstrando redundância funcional e complementaridade entre paradigmas distintos (isolamento, densidade, fronteira e distância). Este comportamento confirma que a decisão agregada não depende criticamente de um único modelo.

Do ponto de vista computacional, a latência média de geração de explicações (≈0,08 ms) revelou-se desprezível face ao custo da deteção (≈1,9 ms), permitindo operação em streaming a 1 Hz com ampla margem temporal. Comparativamente a métodos *ad hoc* como SHAP, cuja execução pode variar entre centenas de milissegundos e vários segundos dependendo da configuração, o sistema proposto apresenta ganhos substanciais de eficiência, mantendo simultaneamente fidelidade elevada ($\rho = 0,87$ com SHAP).

No entanto, a análise multiclasse revelou maior dificuldade na distinção entre regimes *Degraded* e *Failure*, com *accuracy* inferior na classe intermédia. Esta limitação decorre da sobreposição natural entre regimes e da natureza contínua da degradação no *dataset*. Estratégias futuras poderão incluir

modelação explícita de transições temporais, como HMM ou LSTM, para capturar dinâmicas graduais.

Observou-se ainda saturação em cenários extremos ($z > 20\sigma$), reduzindo a variabilidade relativa entre alguns detetores. Contudo, este fenómeno não compromete a decisão operacional, pois corresponde a situações de consenso inequívoco quanto à severidade do evento.

Em termos de transferibilidade, o sistema é agnóstico ao domínio, exigindo apenas um *baseline* representativo de operação normal. Apesar de validado em motores elétricos, a arquitetura pode ser aplicada a outros contextos industriais com variáveis sensoriais contínuas, desde que respeitados os pressupostos de normalização e calibração de *thresholds*.

4.8 Conclusão

Este capítulo apresentou a validação experimental do sistema proposto em 120.636 observações industriais reais. Os resultados demonstram desempenho elevado na deteção binária de anomalias (Precision = 99,70%, *Recall* = 98,52%, F1 = 99,11%, AUPR = 99,96%), superando todos os baselines considerados.

O sistema manteve latência média de deteção de aproximadamente 1,9 ms e latência de explicação de 0,08 ms, assegurando viabilidade para operação em *streaming* com ampla margem temporal.

A fidelidade das explicações foi confirmada por forte concordância com SHAP ($\rho = 0,87$), mantendo cobertura integral dos casos analisados e robustez perante degradação de modelos individuais. Estes resultados demonstram que é possível combinar desempenho preditivo elevado com explicabilidade em tempo real, sem comprometer eficiência computacional.

5.Conclusões e Perspetivas Futuras

5.1. Sumário Geral do Trabalho

Os sistemas industriais contemporâneos geram fluxos contínuos de dados sensoriais, criando oportunidades para manutenção preditiva baseada em aprendizagem automática. Contudo, dois desafios persistem: robustez perante incerteza e explicabilidade em ambientes críticos.

Este trabalho propôs um sistema integrado que combina um *ensemble* heterogéneo de seis modelos não supervisionados com um mecanismo de explicabilidade estruturado em três camadas: consenso inter-modelo, identificação de *features* críticas e geração de recomendações acionáveis.

A validação experimental demonstrou elevado desempenho na deteção de anomalias, robustez estrutural do *ensemble* e latência compatível com operação em tempo real. A forte concordância com SHAP confirmou fidelidade interpretativa, enquanto a arquitetura multi-nível permitiu suporte operacional estruturado à decisão.

Os resultados evidenciam que é possível equilibrar precisão, interpretabilidade e desempenho computacional num único sistema integrado.

5.2. Contribuições Científicas

Este trabalho apresenta diversas contribuições científicas relevantes no âmbito da manutenção preditiva em *streaming* com explicabilidade integrada.

A primeira contribuição consiste na validação empírica de um *ensemble* heterogéneo combinando seis paradigmas distintos de deteção de anomalias: isolamento (*Isolation Forest*), densidade local (*LOF*), fronteira (*One-Class SVM*), distância global (*K-Means*), modelação temporal simples (*Predictive Lag-1*) e aprendizagem online incremental (*Half-Space Trees*). A aplicação conjunta destes modelos a um contexto industrial real evidenciou baixa correlação funcional entre os detetores ($|\rho| < 0{,}21$), confirmando diversidade estrutural no ensemble. Apesar do carácter não supervisionado da abordagem, o sistema alcançou métricas comparáveis a soluções supervisionadas, com precisão de 99,7% na deteção binária, mantendo robustez face à degradação de modelos individuais, como observado no comportamento sub-reativo do HST. Esta evidência sustenta a viabilidade de *ensembles* não supervisionados para aplicações industriais críticas.

A segunda contribuição refere-se à estratégia de *clipping dual* concebida para equilibrar estabilidade numérica na agregação com preservação de discriminabilidade individual para fins de explicabilidade. Enquanto os *z-scores* individuais são mantidos sem truncamento, permitindo capturar a magnitude relativa dos desvios (por exemplo, valores superiores a 100σ em casos extremos), o *z-score* agregado é limitado ao intervalo $\pm 8\sigma$, garantindo estabilidade na decisão final e evitando dominância numérica de modelos com escalas elevadas. Esta abordagem é particularmente relevante em *ensembles* com modelos operando em escalas heterogéneas, sendo potencialmente generalizável a outros contextos.

A terceira contribuição consiste na proposta de um mecanismo de consenso *threshold-based*, definido por $z > 2\sigma$, como alternativa à agregação baseada em magnitudes relativas. Este método resolve o problema de escalas díspares entre modelos, impedindo que detetores com valores

extremos (como LOF) dominem a decisão agregada. O consenso binário preserva diversidade funcional e mantém sensibilidade em cenários *borderline*, além de permitir meta-diagnóstico de modelos degradados através da monitorização da redução do consenso global.

A quarta contribuição centra-se no desenvolvimento de um sistema de explicabilidade multi-nível estruturado em três camadas: (i) análise de consenso inter-modelo, (ii) identificação de componentes críticas com base em *z-scores* normalizados e (iii) geração de recomendações acionáveis. A arquitetura apresenta latência média de aproximadamente 0,08 ms, mantendo forte concordância com SHAP ($\rho = 0,87$), o que demonstra que explicabilidade em tempo real pode ser alcançada sem recorrer exclusivamente a métodos *ad hoc* computacionalmente intensivos. Esta estrutura modular é adaptável a outros *ensembles* e domínios industriais.

A quinta contribuição corresponde à caracterização sistemática das limitações comportamentais de modelos individuais em ambiente de produção. Foram identificados fenómenos como saturação do *Isolation Forest* em anomalias extremas, sub-reatividade do Half-Space Trees em falhas prolongadas e limitação do modelo *Predictive* em desvios estacionários. A análise destes comportamentos permitiu demonstrar que o mecanismo de consenso não apenas agrega decisões, mas também funciona como ferramenta de monitorização da saúde interna do *ensemble*.

Por fim, o trabalho apresenta uma caracterização quantitativa da inversão de *baseline* em modelos online. Observou-se que, após exposição prolongada a falhas contínuas, o *Half-Space Trees* passou a atribuir *z-scores* negativos a anomalias severas (valores entre -15σ e -47σ), interpretando-as como mais normais que o *baseline* original. Este fenómeno ocorreu em aproximadamente 70% das anomalias extremas após cerca de 20.000 observações consecutivas. O sistema de consenso identificou automaticamente esta degradação através da redução do nível de concordância (de 83,3% para 66,7%), evidenciando que o mecanismo *threshold-based* pode ser utilizado para meta-monitorização de modelos online em regimes prolongados de desvio.

Em conjunto, estas contribuições demonstram que é possível conceber um sistema não supervisionado que combine diversidade funcional, estabilidade numérica, explicabilidade estruturada e viabilidade computacional em contexto industrial real.

5.3. Limitações do Estudo

Apesar dos resultados encorajadores, o presente estudo apresenta um conjunto de limitações que devem ser consideradas na interpretação e generalização das conclusões.

A primeira limitação refere-se à validação realizada num único *dataset* industrial, composto por quatro motores e 120.636 observações. Embora o volume de dados seja substancial, os cenários de falha disponíveis traduzem um número reduzido de tipologias e foram gerados através de simulações separadas para cada regime operacional (*Normal, Degraded e Failure*). Esta estrutura limita a riqueza das transições naturais entre estados e dificulta a modelação de processos progressivos de degradação.

Em particular, a separação entre simulações implica que o modelo não observa ciclos completos e contínuos de evolução realista desde operação normal até falha, mas sim segmentos distintos. Como consequência, torna-se mais complexo capturar padrões temporais subtis que caracterizam a transição gradual para o estado degradado. Esta limitação ajuda a explicar o desempenho inferior

observado na classe degradado, onde existe sobreposição estrutural com a classe falha e menor separabilidade estatística.

Uma segunda limitação decorre do uso de um valor de base estático de operação normal. Embora adequado para estabilidade inicial, este *baseline* pode tornar-se desatualizado perante mudanças estruturais de longo prazo. Em ambientes industriais reais, alterações sazonais, substituição de componentes ou reconfigurações operacionais podem modificar a distribuição dos dados, exigindo mecanismos automáticos de adaptação.

A terceira limitação relaciona-se com a inversão de *baseline* observada em modelos online, particularmente no *Half-Space Trees*. Em falhas prolongadas, o modelo tende a adaptar-se excessivamente ao novo padrão anómalo, produzindo *z-scores* negativos (entre -15σ e -47σ), interpretando anomalias severas como mais normais que o *baseline* original. Embora o sistema consiga identificar esta degradação através da redução do consenso (de 83,3% para 66,7%), o fenómeno evidencia a necessidade de mecanismos formais de controlo de adaptação em aprendizagem contínua.

Outra limitação diz respeito à natureza marginal das explicações geradas na Camada 2. O método avalia cada *feature* individualmente relativamente ao *baseline*, não capturando explicitamente interações não-lineares complexas entre variáveis. A divergência de aproximadamente 13% face ao SHAP pode ser parcialmente atribuída a este fator.

Adicionalmente, o sistema não incorpora raciocínio causal explícito, baseando-se em associações estatísticas. Assim, embora consiga identificar variáveis críticas, não distingue formalmente entre causa primária e efeito secundário do fenómeno observado.

Por fim, a geração de contra-factuais na Camada 3 baseia-se em heurísticas eficientes, mas não necessariamente ótimas. Existe um *trade-off* entre latência e otimização da solução proposta, podendo ocorrer casos em que a recomendação não representa a modificação mínima teoricamente ideal.

Em conjunto, estas limitações não invalidam os resultados obtidos, mas delimitam o âmbito de aplicação atual do sistema e identificam oportunidades claras para investigação futura, nomeadamente no enriquecimento do *dataset* com transições reais contínuas, adaptação dinâmica de *baseline* e incorporação de modelação temporal mais expressiva.

5.4. Trabalho Futuro

Apesar dos resultados obtidos, existem diversas direções promissoras para evolução do sistema.

Uma primeira linha de investigação consiste na integração de mecanismos formais de deteção de *concept drift* e adaptação dinâmica do *baseline*. A utilização de técnicas *drift-aware*, como monitorização estatística contínua das distribuições ou métodos baseados em janelas adaptativas, permitiria atualizar automaticamente os *thresholds* e prevenir fenómenos de sobre-adaptação observados em modelos online. Esta evolução é particularmente relevante em cenários industriais de longa duração, sujeitos a sazonalidade e alterações operacionais graduais.

Outra direção relevante prende-se com a evolução arquitetural do pipeline para ambientes industriais reais. A implementação híbrida *edge–cloud* constitui abordagem natural: a deteção e explicabilidade leve poderiam ser executadas em dispositivos *edge*, próximos dos sensores, garantindo latência

mínima e resiliência a falhas de conectividade, enquanto análises mais pesadas, auditoria histórica e recalibração de modelos poderiam ocorrer na *cloud*. Esta separação permitiria escalar o sistema mantendo robustez operacional.

Adicionalmente, a introdução controlada de modelos de linguagem (LLMs) poderá enriquecer a camada de explicabilidade, permitindo transformar os outputs estruturados do sistema em narrativas interpretáveis para operadores. A utilização destes modelos deverá ser cuidadosamente validada para garantir fidelidade factual e evitar geração de interpretações não suportadas pelos dados. A sua aplicação poderá ser restrita a análises *offline* ou relatórios automáticos, preservando o fluxo crítico em tempo real.

Outra linha futura consiste na modelação temporal mais expressiva, incluindo abordagens probabilísticas ou sequenciais capazes de capturar transições graduais entre regimes. Tal evolução poderá melhorar a distinção entre estados *Degraded e Failure*, particularmente em *datasets* com evolução contínua.

Por fim, a validação prolongada em ambiente industrial real, com monitorização contínua de desempenho e *feedback* de operadores, constitui passo essencial para consolidação do sistema em produção.

5.5. Conclusão Final

O presente trabalho desenvolveu e validou um sistema integrado de deteção de anomalias com explicabilidade em tempo real, demonstrando a sua viabilidade técnica e operacional em dados industriais reais. O sistema alcançou elevado desempenho preditivo, latência extremamente reduzida e fidelidade interpretativa consistente, evidenciando que é possível conciliar precisão e interpretabilidade em contexto de *streaming*.

Entre as principais contribuições destaca-se a validação empírica de um *ensemble* heterogéneo com diversidade funcional comprovada, evidenciada por baixa correlação entre modelos e elevada robustez perante degradação individual. Adicionalmente, foi proposta uma estratégia de *clipping* dual que permite equilibrar estabilidade numérica na agregação com preservação de discriminabilidade individual para fins de explicabilidade. O método de consenso baseado em *threshold* revelou-se eficaz na harmonização de escalas díspares entre modelos, garantindo equidade na contribuição de cada detetor para a decisão final. Por fim, foi desenvolvida uma arquitetura de explicabilidade multi-nível com latência compatível com operação em *streaming*, capaz de fornecer suporte interpretativo estruturado sem comprometer o desempenho computacional.

Apesar das limitações identificadas, nomeadamente a utilização de um *baseline* estático, a natureza marginal das explicações e a ausência de modelação causal explícita, foram delineadas estratégias concretas de mitigação que abrem caminho para investigação futura e evolução do sistema.

Em síntese, o trabalho demonstra que robustez, interpretabilidade e eficiência computacional não são objetivos mutuamente exclusivos. A integração de um *ensemble* heterogéneo com uma camada estruturada de explicabilidade constitui um contributo relevante para o avanço de sistemas de manutenção preditiva em contextos industriais críticos, onde a confiabilidade das decisões é determinante.

Referências Bibliográficas

Abdoune, F., Nouiri, M., & Cardin, O. (2025). Digital twin-based anomaly detection under concept drift: A comparison between iterative batch learning and incremental learning approaches. Expert Systems with Applications, 268, 130062. https://doi.org/10.1016/j.eswa.2025.130062

Ahmed Murtaza, A. (2024). Paradigm shift for predictive maintenance and condition monitoring from Industry 4.0 to Industry 5.0: A systematic review, challenges and case study. Results in Engineering, 24, 102935. https://doi.org/10.1016/j.rineng.2024.102935

Almeida, A., Brás, S., Sargento, S., Triay, J., Solé-Pareta, J., & Garcia-Espin, J. A. (2023). Time series big data: A survey on data stream frameworks, analysis and algorithms. Journal of Big Data, 10, 83. https://doi.org/10.1186/s40537-023-00760-1

Asutkar, S., & Tallur, S. (2023). An explainable unsupervised learning framework for scalable machine fault detection in Industry 4.0. Measurement Science and Technology, 34(10), 105123. https://doi.org/10.1088/1361-6501/ace640

Ateş, E., Aksar, B., Leung, V., & Coşkun, K. (2021). Counterfactual explanations for multivariate time series. In 2021 International Conference on Applied Artificial Intelligence (ICAPAI) (pp. 1-8). IEEE. https://doi.org/10.1109/ICAPAI49758.2021.9462056

Barbariol, T., Chiara, F. D., Marcato, D., & Susto, G. A. (2022). A review of tree-based approaches for anomaly detection. In K. P. Tran (Ed.), Control charts and machine learning for anomaly detection in manufacturing (pp. 135-169). Springer. https://doi.org/10.1007/978-3-030-83819-5_7

Barry, M. S. (2020). Spatiotemporal anomaly detection: Streaming architecture and algorithms [Doctoral dissertation, Colorado State University]. HAL. https://hal.science/hal-05176393v1

Barry, M., Montiel, J., Bifet, A., Manchev, N., & Wadkar, S. (2025). StreamMLOps: Online learning in practice from big data streams & real-time applications. HAL. https://hal.science/hal-05176393v1

Bäßler, D., Kortus, T., & Gühring, G. (2022). Unsupervised anomaly detection in multivariate time series with online evolving spiking neural networks. Machine Learning, 111, 1377-1408. https://doi.org/10.1007/s10994-022-06129-4

Biikes, L., Matthew, E., White, D., & Elly, B. (2024). Real-time data processing: Enhancing machine learning algorithm efficiency for streaming applications. https://www.researchgate.net/publication/386453015

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (pp. 93-104). ACM. https://doi.org/10.1145/342009.335388

Buabeng, A., Simons, A., Frempong, N. K., Ziggah, Y. Y., & Tamakloe, R. (2024). Hybrid intelligent predictive maintenance model for multiclass fault classification. Soft Computing, 28, 8749-8770. https://doi.org/10.1007/s00500-023-08993-1

Cabrera Martin, I., Mukherjee, S., Baimagambetov, A., Vanschoren, J., & Polatidis, N. (2025). Evolving machine learning in non-stationary environments: A unified survey of drift, forgetting, and adaptation [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2505.17902

Cao, Y., Ma, Y., Zhu, Y., Li, X., Zhang, H., & Wang, J. (2025). Revisiting streaming anomaly detection: Benchmark and evaluation. Artificial Intelligence Review, 58, 8. https://doi.org/10.1007/s10462-024-10995-w

Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1901.03407

Chevtchenko, S. F., Rocha, E. D. S., Santos, M. C. M. D., Mota, R. L., Vieira, D. M., & Andrade, E. C. D. (2023). Anomaly detection in industrial machinery using IoT devices and machine learning: A systematic mapping. IEEE Access, 11, 128288-128305. https://doi.org/10.1109/ACCESS.2023.3333242

Chiang, M., & Zhang, T. (2016). Fog and IoT: An overview of research opportunities. IEEE Internet of Things Journal, 3(6), 854-864. https://doi.org/10.1109/JIOT.2016.2584538

Cook, A., Mısırlı, G., & Fan, Z. (2019). Anomaly detection for IoT time-series data: A survey. IEEE Internet of Things Journal, 7(7), 6481-6494. https://doi.org/10.1109/JIOT.2019.2958185

Cover, T. M., & Thomas, J. A. (2006). Elements of information theory (2nd ed.). John Wiley & Sons.

Day-Olupona, O., Genc, B., Celik, T., & Bada, S. (2023). Adoptable approaches to predictive maintenance in mining industry: An overview. Resources Policy, 86, 104291. https://doi.org/10.1016/j.resourpol.2023.104291

Elsaid, S. A., Shehab, E., Mattar, A. M., & El-kenawy, E. S. M. (2024). Hybrid intrusion detection models based on GWO optimized deep learning. Discover Applied Sciences, 6, 531. https://doi.org/10.1007/s42452-024-06209-1

Fernandes, M., Corchado, J. M., & Marreiros, G. (2022). Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in real industrial manufacturing use-cases: A systematic literature review. Applied Intelligence, 52, 14246-14280. https://doi.org/10.1007/s10489-022-03344-3

Fragkoulis, M., Carbone, P., Kalavri, V., & Katsifodimos, A. (2024). A survey on the evolution of stream processing systems. The VLDB Journal, 33, 507-541. https://doi.org/10.1007/s00778-023-00819-8

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44. https://doi.org/10.1145/2523813

Gomes, H. M., Read, J., Bifet, A., Barddal, J. P., & Gama, J. (2019). Machine learning for streaming data: State of the art, challenges, and opportunities. ACM SIGKDD Explorations Newsletter, 21(2), 6-22. https://doi.org/10.1145/3373464.3373470

Gomolka, Z., Zeslawska, E., & Olbrot, L. (2025). Using hybrid LSTM neural networks to detect anomalies in the fiber tube manufacturing process. Applied Sciences, 15(3), 1383. https://doi.org/10.3390/app15031383

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models [Preprint]. arXiv. https://arxiv.org/abs/1802.01933

IBM. (n.d.). Anomaly detection. IBM Think. Retrieved December 2024, from https://www.ibm.com/br-pt/think/topics/anomaly-detection

Jiang, X. (2020). Anomaly detection and diagnostics in distribution networks using high-frequency PQ data [Doctoral dissertation, University of Strathclyde]. https://doi.org/10.48730/7vpd-y415

Jirwe, M. (2021). Online anomaly detection on the edge [Master's thesis, KTH Royal Institute of Technology]. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-299565

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A, 374, 20150202. https://doi.org/10.1098/rsta.2015.0202

Jourdan, N. (2024). Addressing concept drift in machine learning-based monitoring of manufacturing processes [Doctoral dissertation, Technische Universität Darmstadt]. https://doi.org/10.26083/tuprints-00028043

Klaise, J., Van Looveren, A., Cox, C., Vacanti, G., & Coca, A. (2020). Monitoring and explainability of models in production [Preprint]. arXiv. https://arxiv.org/abs/2007.06299

Koch, F., Schramm, M., & Rosenberger, M. (2024). Anomaly detection in satellite telemetry data using machine learning. In D. Dold, A. Hadjiivanov, & D. Izzo (Eds.), Proceedings of SPAICE2024: The First Joint European Space Agency/IAA Conference on AI in and for Space (pp. 425-429). https://doi.org/10.5281/zenodo.13885647

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583-621. https://doi.org/10.2307/2280779

Laugel, T., Lesot, M. J., Marsala, C., Renard, X., & Detyniecki, M. (2019). The dangers of post-hoc interpretability: Unjustified counterfactual explanations [Preprint]. arXiv. https://arxiv.org/abs/1907.09294

Leite, D., Decker, L., Santana, M., & Souza, P. (2020). EGFC: Evolving Gaussian fuzzy classifier from never-ending semi-supervised data streams—With application to power quality disturbance detection and classification. In 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-9). IEEE. https://doi.org/10.1109/FUZZ48607.2020.9177847

Leveni, F. (2025). Structure-based anomaly detection and clustering [Preprint]. arXiv. https://arxiv.org/abs/2505.12751

Li, B. (2024). Unsupervised temporal anomaly detection: Time series, data stream, and interpretability [Doctoral dissertation, TU Dortmund University]. https://eldorado.tu-dortmund.de/handle/2003/42156

Li, M. A., & Gautam, A. (2025). Segmented confidence sequences and multi-scale adaptive confidence segments for anomaly detection in nonstationary time series [Preprint]. arXiv. https://arxiv.org/abs/2508.06638

Li, X., Zhang, W., Ding, Q., & Sun, J. Q. (2023). A survey on explainable anomaly detection. ACM Transactions on Knowledge Discovery from Data, 18(1), Article 19. https://doi.org/10.1145/3609333

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (pp. 413-422). IEEE. https://doi.org/10.1109/ICDM.2008.17

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 4765-4774).

Madathil, A. P., Luo, X., Liu, Q., Cai, W., Landers, R. G., Nawaz, M., & Liou, F. (2024). Intrinsic and post-hoc XAI approaches for fingerprint identification and response prediction in smart manufacturing processes. Journal of Intelligent Manufacturing, 35, 4159-4180. https://doi.org/10.1007/s10845-023-02266-2

Malarkkan, A. V., Wang, D., Bai, H., & Fu, Y. (2025). Incremental causal graph learning for online cyberattack detection in cyber-physical infrastructures [Preprint]. arXiv. https://arxiv.org/abs/2507.14387

Martins, A. D. B. (2022). Monitorização online de sensores para apoio à manutenção preditiva suportado em ferramentas de inteligência artificial [Projeto de Inovação]. Ordem dos Engenheiros. https://www.ordemdosengenheiros.pt/fotos/editor2/2025/pije/am_projetooe_pije_signed_9102110763ab436584b62.pdf

Mercurio, D. (2024). Towards predictive maintenance using long short-term memory autoencoder and streaming explainability [Master's thesis, Politecnico di Milano]. https://hdl.handle.net/10589/219659

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning – A brief history, state-of-the-art and challenges [Preprint]. arXiv. https://arxiv.org/abs/2010.09337

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617). ACM. https://doi.org/10.1145/3351095.3372850

Mozaffari, M., Doshi, K., & Yilmaz, Y. (2022). Online multivariate anomaly detection and localization for high-dimensional settings. Sensors, 22(21), 8264. https://doi.org/10.3390/s22218264

Nardi, L. (2024). Edge computing solutions for real-time anomaly detection in IoT networks [Doctoral dissertation, Scuola Normale Superiore]. https://hdl.handle.net/20.500.14242/305862

Nguyen, H. T. T., Nguyen, L. P. T., & Cao, H. (2024). XEdgeAI: A human-centered industrial inspection framework with data-centric explainable edge AI approach. Information Fusion, 114, 102782. https://doi.org/10.1016/j.inffus.2024.102782

Nsor, M. (2024). Predictive maintenance using machine learning for engineering systems through real-time sensor data and anomaly detection models. International Journal of Research Publication and Reviews, 5, 5167-5183. https://doi.org/10.55248/gengpi.6.0725.2541

Oliveira, D. F. N., Vismari, L. F., Nascimento, A. M., Almeida, J. R., Cugnasca, P. S., & Camargo, J. B. (2022). A new interpretable unsupervised anomaly detection method based on residual explanation. IEEE Access, 10, 1401-1409. https://doi.org/10.1109/ACCESS.2021.3135416

Paltenghi, M. (2020). Time series anomaly detection for CERN large-scale computing infrastructure [Master's thesis, Politecnico di Milano]. https://hdl.handle.net/10589/158602

Panwar, A., Pal, H., Chen, J., Cho, K., Jiang, R., Zhao, M., & Krishnamurthy, R. (2025). Reasoning-based anomaly detection framework: A real-time, scalable, and automated approach to anomaly detection across domains [Preprint]. arXiv. https://arxiv.org/abs/2510.03486

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM. https://doi.org/10.1145/2939672.2939778

Romero, M., & Suyama, R. (2024). DQAT: An online machine learning framework for real-time data quality assurance in IoT. In XLI Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (pp. 1-5). https://doi.org/10.14209/sbrt.2024.1571036913

Rosenberger, J., Selig, A., Ristic, M., Bühren, M., & Schramm, D. (2023). Virtual commissioning of distributed systems in the Industrial Internet of Things. Sensors, 23(7), 3545. https://doi.org/10.3390/s23073545

Rožanec, J., Trajkova, E., Kenda, K., Fortuna, B., & Mladenić, D. (2021). Explaining bad forecasts in global time series models. Applied Sciences, 11(19), 9243. https://doi.org/10.3390/app11199243

Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30-39. https://doi.org/10.1109/MC.2017.9

Schindler, T. F., Schlicht, S., & Thoben, K. D. (2023). Towards benchmarking for evaluating machine learning methods in detecting outliers in process datasets. Computers, 12(12), 253. https://doi.org/10.3390/computers12120253

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017) (pp. 3319-3328).

Ucar, A., Karakose, M., & Kırımça, N. (2024). Artificial intelligence for predictive maintenance applications: Key components, trustworthiness, and future trends. Applied Sciences, 14(2), 898. https://doi.org/10.3390/app14020898

Verma, S., Boonsanong, V., Hoang, M., Hines, K. E., Dickerson, J. P., & Shah, C. (2022). Counterfactual explanations and evaluation metrics [Preprint]. arXiv. https://arxiv.org/abs/2010.10596

Weinberg, A. I. (2025). Datastreams and beyond, from traditional approaches to quantum: A comprehensive survey. Evolutionary Intelligence, 18, 94. https://doi.org/10.1007/s12065-025-01079-x

Xiang, H., & Zhang, X. (2022). Edge computing empowered anomaly detection framework with dynamic insertion and deletion schemes on data streams. World Wide Web, 25, 2163-2183. https://doi.org/10.1007/s11280-022-01052-z

Yahya, M. A., Moya, A. R., & Ventura, S. (2025). Deep learning for multivariate time series anomaly detection: An evaluation of reconstruction-based methods. Artificial Intelligence Review, 58, 400. https://doi.org/10.1007/s10462-025-11401-9

Yan, S. Z. A. (2021). Anomaly detection in univariate time series data in the presence of concept drift [Master's thesis, McMaster University]. https://macsphere.mcmaster.ca/items/cd8542e1-684d-40f7-9d27-0ea57f2cddf2

Zakeriharandi, M., Li, C., Villumsen, S. L., Ghaffari, M., & Madsen, O. (2025). LST-MADNet: Learnable scattering transform network for multi-modal time-series anomaly detection in industrial manufacturing systems [Preprint]. SSRN. https://doi.org/10.2139/ssrn.5414854

UNIVERSIDADE
PORTUCALENSE

upt.pt