UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Maira Farias de Andrade Lira

Abordagens para Seleção de Limiares de Decisão e Filtros de Suavização em Detecção de Anomalias

Recife

2024

Maira Farias de Andrade Lira

Abordagens para Seleção de Limiares de Decisão e Filtros de Suavização em Detecção de Anomalias

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Orientador: Prof. Ricardo Bastos Cavalcante Prudêncio

Recife

2024

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

L768a Lira, Maira Farias de Andrade
Abordagens para seleção de limiares de decisão e filtros de suavização em detecção de anomalias / Maira Farias de Andrade Lira – 2024.
88 f.: il., fig., tab.

Orientador: Ricardo Bastos Cavalcante Prudêncio.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2024.
Inclui referências.

1. Inteligência computacional. 2. Detecção de anomalias. 3. *Sparse Autoencoder*. 4. *Threshold*. 5. Filtro passa-baixa. I. Prudêncio, Ricardo Bastos Cavalcante (orientador). II. Título

006.31 CDD (23. ed.) UFPE - CCEN 2024 – 83

Maira Farias de Andrade Lira

“Abordagens para Seleção de Limiares de Decisão e Filtros de Suavização em Detecção de Anomalias”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 25 de janeiro de 2024.

BANCA EXAMINADORA

_____

Prof. Dr. Paulo Salgado Gomes de Mattos Neto
Centro de Informática / UFPE

_____

Prof. Dr. João Paulo Pordeus Gomes
Departamento de Computação / UFC

_____

Prof. Dr. Ricardo Bastos Cavalcante Prudêncio
Centro de Informática / UFPE
(**orientador**)

Dedico esse trabalho aos meus pais e avós que desde pequena me ensinaram a não aceitar um 'porque sim' ou 'porque não' como resposta, por mais que muitas vezes tenham me respondido assim.

AGRADECIMENTOS

Primeiramente, agradeço aos meus pais, minha avó e meu avó (*in memoriam*) por sempre acreditarem em mim e serem os melhores exemplos que alguém poderia sonhar em ter. Em especial, obrigada à minha mãe por me ensinar a sempre ser gentil; ao meu pai por me provar que ninguém morreu afogado em suor (pelo menos até o momento); aos meus avôs por todas as conversas que sempre me abriram os olhos para outros pontos de vista. Vocês enxergaram meu potencial quando nem eu mesma sabia e me incentivaram a sonhar cada vez mais alto. Se hoje cheguei aqui, sei que foi por vocês, e espero deixá-los sempre orgulhosos.

Ao meu parceiro, Elias Amancio, por me apoiar incansavelmente e me ajudar cada dia a ser um pouquinho melhor do que no dia anterior. Obrigada por todas as nossas conversas e, especialmente, por nunca me deixar desistir dos meus sonhos. Além disso, obrigada por sempre estar presente por mim, mesmo nos momentos que eu não sou a melhor companhia do mundo. Eu te amo mais que demais, espero que nunca esqueças disso!

Ao meu orientador Professor Ricardo Prudêncio que desde a primeira reunião me acolheu e acreditou em mim. Sua orientação e conselhos amigos foram fundamentais para o desenvolvimento desta dissertação e para minha formação acadêmica.

A todas minhas famílias: a genética, a tabirense, a amiga, a literária e a do IATI. Obrigada por comemorarem cada pequena conquista e sempre demonstrarem apoio e carinho à sua maneira.

À minha afilhada Aurora, a quem eu desejo toda a felicidade e saúde do mundo. Tia Maira e Tio Elias sempre estarão aqui para te apoiar e, espero, te inspirar.

Ao meu irmãozinho de quatro patas, Floquinho (*in memoriam*), por sempre saber o que eu estava sentindo, mesmo sem entender o que eu estava falando. Eu nunca vou te esquecer.

À Universidade Federal de Pernambuco e todos que contribuíram, direta ou indiretamente, para minha formação acadêmica.

Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, por investir na pequisa. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

"Só aqueles que experimentaram podem imaginar as tentações da ciência. Em outros estudos, vai-se tão longe quanto outros foram antes de você e não há mais nada a saber; mas numa busca científica, há alimento contínuo para a descoberta e o maravilhamento."(SHELLEY, 2023)

RESUMO

Técnicas de detecção de anomalias são amplamente utilizadas para identificar instâncias com padrões distintos do comportamento geral de um conjunto de dados. O desenvolvimento de novas técnicas, como as baseadas em aprendizado profundo, e a maior disponibilidade de dados têm alavancado ainda mais o uso da detecção de anomalias em contextos como a detecção de falhas em equipamentos industriais. Comumente, uma técnica de detecção gera um *score* de anomalia para cada instância, que é então usado para classificá-las entre anômalas ou normais. Esta classificação é baseada em um limiar de decisão (*threshold*) estabelecido de forma que se o *score* de uma determinada instância for superior ao *threshold*, esta instância é considerada anômala, caso contrário é classificada como normal. Neste trabalho foi utilizado um modelo *Sparse Autoencoder* (SAE) para a detecção de anomalias *online* que vem ganhando popularidade neste cenário e foi investigado o impacto de diferentes abordagens não supervisionadas para definição de *thresholds*. Para os experimentos foi utilizada uma base de dados pública referente a um problema de detecção de anomalias no metrô da cidade do Porto. A abordagem de cálculo do *threshold* impactou fortemente as métricas de avaliação da detecção. Por exemplo, a abordagem baseada em erro máximo garantiu a menor taxa de falsos positivos. Por sua vez, a abordagem baseada em intervalo interquartil obteve o maior número de verdadeiros positivos, e, consequentemente *recall*, enquanto que a abordagem baseada em 99-percentil garantiu o maior F1-Score. Foi avaliado ainda o uso de três tipos de filtros passa-baixa em duas abordagens distintas para a suavização do *score* de anomalia. De uma forma geral, a aplicação de filtros diretamente sobre o *score* de anomalia maximizou verdadeiros positivos, enquanto sua aplicação após uma classificação prévia das instâncias minimizou os falsos positivos. Além disso, foi verificado que a utilização do filtro foi essencial para detectar sequências de anomalias. Desta forma, a seleção de abordagens de definição de *thresholds* e de aplicação de filtros deve ser definida em função dos objetivos específicos do modelo.

Palavras-chave: detecção de anomalias; sparse autoencoder; *threshold*; filtro passa-baixa.

ABSTRACT

Anomaly detection techniques are widely used to identify instances with patterns differing from the general behavior of a data set. The development of new techniques, such as those based on deep learning, and the higher availability of data have increased anomaly detection use in contexts such as failure detection in industrial equipment. Frequently, a detection technique generates an anomaly score for each instance, later used to classify it as anomalous or normal. This classification is based on an established detection threshold such that if a given instance's score is higher than the established limit, it is considered anomalous. Otherwise, it is normal. In this work, the impact of different unsupervised approaches to define a threshold was investigated for anomaly detection by a Sparse Autoencoder (SAE) model. The experiments were based on a public database from an anomaly detection problem in Porto metro. The threshold calculation method strongly impacted detection evaluation metrics. For example, the maximum error approach guaranteed the lowest false positive ratio. On the other hand, the inter quantile range approach yielded the highest true positive numbers and, consequently, higher recall, and the 99-percentile-based approach had the highest F1-Score. We also evaluated using three low-pass filters in two different approaches to smooth anomaly scores. Generally, filter applications directly on the anomaly score maximized true positives, while their application after a previous instance classification minimized false positives. Besides this, filter usage was essential to detect anomalous sequences. Thus, the selection of threshold definition techniques and filter application must be defined in function of the model-specific goals.

Keywords: anomaly detection; sparse autoencoder; threshold; low pass-filter.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AD	Detecção de Anomalia (do inglês *Anomaly Detection*)
AE	*Autoencoder*
APU	Unidade de Produção de Ar (do inglês *Air Production Unit*)
BOX	Boxplot Extremo
DL	Aprendizagem Profunda (do inglês *Deep Learning*)
EMA	Média Exponencial Móvel (do inglês *Exponential Moving Average*)
FP	Falso Positivo (do inglês *False Positive*)
FPR	Taxa de Falso Positivo (do inglês *False Positive Rate*)
IQR	Intervalo Interquartil (do inglês *Inter Quantile Range*)
KL	Kullback-Leibler
LPF	Filtro Passa-Baixa (do inglês *Low-Pass Filter*)
MAE	Erro Absoluto Médio (do inglês *Mean Absolute Error*)
MAX	Erro Máximo
MSE	Erro Quadrático Médio (do inglês *Mean Squarred Error*)
N95	95-percentil
N97	97-percentil
N99	99-percentil
PCA	*Principal Component Analysis*
Pré-Classificação	Abordagem Pré-Classificação de Filtro
Pós-Classificação	Abordagem Pós-Classificação de Filtro
SAE	*Sparse Autoencoder*
SMA	Média Móvel Simples (do inglês *Simple Moving Average*)
SMM	Mediana Móvel Simples (do inglês *Simple Moving Median*)

TP Verdadeiro Positivo (do inglês *True Positive*)

LISTA DE SÍMBOLOS

α	Parâmetro de Suavização do Filtro EMA
β	Peso de Esparsidade
λ	Decaimento do Peso L2
μ	Média Aritmética do Atributo
ρ	Parâmetro de Esparsidade
$\hat{\rho}$	Parâmetro de Ativação Médio do Neurônio h
σ	Desvio Padrão do Atributo
$\|\|\mathbf{W}\|\|$	Peso dos Neurônios
b	Índice de Bloco
h	Índice das Unidades Escondidas
i	Índice do Ciclo
k	Índice de Instância
l	Índice de Parâmetros
m	Número de Características
n	Tamanho da Janela dos Filtros SMA e SMM
Q1	Primeiro Quartil do *score* de Anomalia
Q3	Terceiro Quartil do *score* de Anomalia
RE	Erro de Reconstrução
$s2$	Número de Neurônios na Camada Latente
t	Índice Temporal
$thres$	*Threshold*
$thres^w$	*Threshold* Calculado na Janela w

$thres^{exp}$	*Threshold* Experimental
W	Entrada do Filtro
w	Índice da Janela de Treinamento
$\mathbf{X}$	Entrada do Modelo SAE
$\hat{X}$	Saída do Modelo SAE
x_{ij}	Valor da j-ésima Característica do Ciclo i
Y	Classificação Detectada
y	Valor do Atributo
Z	Saída do Filtro
z	Variável Normalizada por *z-score*

SUMÁRIO

1 INTRODUÇÃO

Nesse capítulo, será apresentado um contexto geral relacionado a essa dissertação, os problemas de pesquisa abordados, os objetivos definidos e um breve resumo do trabalho desenvolvido.

1.1 CONTEXTO GERAL

Anomalias são padrões que diferem significativamente do comportamento geral observado em um conjunto de dados. Métodos de Detecção de Anomalia (do inglês *Anomaly Detection*) (AD) têm sido usado em contextos diversos de aplicação, tais como: identificação de fraudes bancárias, detecção de padrões anômalos em exames médicos de imagem e detecção de falhas em componentes industriais de forma incipiente (CHANDOLA; BANERJEE; KUMAR, 2009). Um fator comum entre estes contextos é o não balanceamento das classes, já que anomalias ocorrem com uma frequência muito inferior ao padrão normal (LIU; TING; ZHOU, 2008). Por serem problemas não balanceados de classificação, muitos modelos podem apresentar boas métricas de avaliação apesar de não detectarem corretamente as anomalias. Por exemplo, caso haja poucas instâncias anômalas, o modelo pode classificar todos os exemplos como normais e mesmo assim obter uma acurácia alta.

Há diferentes formas de detectar anomalias, e.g. modelos analíticos, conhecimento técnico e baseados em dados, sendo a última a mais explorada atualmente por exigir menor conhecimento prévio do cenário de aplicação, além de serem adaptáveis a diferentes casos (ZHANG; YANG; WANG, 2019). Três tipos de abordagens de AD baseadas em dados se destacam na literatura: detecção supervisionada, não supervisionada e semi-supervisionada (CHANDOLA; BANERJEE; KUMAR, 2009). O que distingue essas abordagens uma das outras é a presença ou não de dados rotulados, i.e., instâncias identificadas previamente como sendo anomalias ou não-anomalias. A detecção supervisionada necessita de conhecimento prévio da etiqueta de todas as instâncias do treinamento, enquanto a não supervisionada não necessita.

Com o avanço de metodologias não supervisionadas ou semi-supervisionadas que utilizam Aprendizagem Profunda (do inglês *Deep Learning*) (DL), a performance de métodos de detecção de anomalias e predição de falhas melhorou drasticamente e vêm sendo aplicados para diversos cenários (AUDIBERT et al., 2022). Um modelo que ganhou muita popularidade foi o

Autoencoder (AE) e suas variações como *Sparse Autoencoder* (SAE) para atividades de AD tanto na avaliação de imagens quanto séries temporais, devido a sua capacidade de aprender apenas os *features* mais importantes para serem utilizados como referência de normalidade dos dados (BLáZQUEZ-GARCÍA et al., 2021). Para aplicações temporais, também há um crescimento de técnicas de aprendizado *online*, já que o comportamento de um sistema pode ser modificado ao longo do tempo e esta forma de aprendizagem permite o modelo de AD se adaptar a novas condições que podem ser desconhecidas até então. A detecção *online* em particular é comumente usada para a manutenção preditiva.

1.2 PROBLEMAS DE PESQUISA

De uma forma geral, no contexto de AD, um *score* de anomalia de um SAE é obtido a partir do erro de reconstrução do modelo para cada padrão de entrada avaliado. O erro de reconstrução é definido a partir da diferença entre os dados de saída estimados pelo SAE e os dados de entrada fornecidos ao modelo. Quando o erro de reconstrução de uma instância ultrapassar um limiar de decisão (*threshold*), o padrão de entrada é classificado como anômalo. Porém, o erro de reconstrução pode muitas vezes apresentar picos espúrios não justificados que afetam diretamente as métricas de avaliação e a taxa de Falso Positivo (do inglês *False Positive*) (FP).

Nessa dissertação, dois aspectos práticos do uso de modelos de AD compartilhados pelos SAEs são explorados. O primeiro aspecto diz respeito à escolha dos *thresholds* para classificação de padrões de forma não supervisionada. Valores muito baixos para os *thresholds* podem resultar em aumento de FP, enquanto valores muito altos podem diminuir a taxa de FP e reduzir a taxa de Verdadeiro Positivo (do inglês *True Positive*) (TP). Essa decisão é crucial para a detecção de anomalias e pode ser fundamentada em diferentes princípios. Borghesi et al. (2019) e Tun, Nyaung e Phyu (2020) exploraram o cálculo de *threshold* baseado no princípio de N-percentil com diferentes valores de n, enquanto Perini, Bürkner e Klami (2023) utilizam diferentes abordagens de cálculo, porém não fazem uma avaliação extensa de impacto em diferentes métricas de detecção nem fazem um estudo de relevância estatística.

O segundo aspecto é a redução dos FP que são padrões que apresentam valores de *score* de anomalia altos, mas que na realidade são não-anômalos. Para suavizar o *score* de anomalia e diminuir a influência de picos espúrios produzidos pelo SAE, foram aplicados Filtro Passa-Baixa (do inglês *Low-Pass Filter*) (LPF) como sugerido por Ribeiro, Pereira e Gama (2016) e

Davari et al. (2021). Os autores sugeriram aplicar um LPF após uma classificação prévia das instâncias de forma a obter um novo *score* de anomalia entre 0 e 1 que foi posteriormente reclassificado. Este trabalho também propôs avaliar uma nova aplicação de LPF para suavizar diretamente o erro de reconstrução do SAE e reduzir o impacto dos picos encontrados.

1.3 OBJETIVOS

Esta seção apresenta o objetivo geral e os objetivos específicos deste trabalho.

1.3.1 Objetivo Geral

Investigar o impacto de diferentes configurações de detecção nas métricas de avaliação de um SAE no aprendizado *online*.

1.3.2 Objetivos Específicos

Como objetivos específicos para alcance da proposta geral, tem-se:

- Avaliar a influência da técnica de determinação do *threshold* nas métricas de avaliação de AD sem influência de filtros;
- Explorar a aplicação de LPFs distintos com duas abordagens de aplicação e seu respectivo impacto nas métricas de avaliação e na incidência de FP e TP;
- Avaliar o impacto dos parâmetros de entrada dos LPFs usados no comportamento de detecção e métricas de avaliação para cada configuração de detecção testada.

1.4 TRABALHO DESENVOLVIDO

Neste trabalho, foi investigada a influência da abordagem de cálculo de *threshold* adaptativo no aprendizado *online* de AD por meio quatro princípios distintos em uma configuração sem aplicação de LPF, contemplando um total de seis formas de cálculo de *threshold*. Além disso, foram avaliadas duas abordagens de aplicação de três LPFs distintos e seus impactos nas métricas de avaliação de forma estatística. Para isso, foram utilizados diferentes parâmetros de entrada e as configurações foram aplicadas para todos os *thresholds* avaliados. As 42

configurações de detecção avaliadas foram experimentadas na base de dados de uma aplicação de manutenção preditiva no metrô da cidade do Porto, Portugal, disponibilizada por Davari et al. (2021).

Foi observado que a abordagem de cálculo de *threshold* impactou estatisticamente as métricas de avaliação da detecção. Por exemplo, a abordagem baseada em erro máximo garantiu a menor taxa de FP, enquanto a abordagem baseada em intervalo interquartil apresentou o TP mais alto e 99-percentil obteve o melhor F1-Score. Já em relação à aplicação de LPF, a abordagem de aplicação diretamente sobre o *score* de anomalia maximizou o TP e deteriorou o FP, enquanto a aplicação após a classificação prévia das instâncias minimizou FP e TP. Apesar disso, ambas as abordagens permitiram a detecção de sequências de anomalias que não foi possível com as configurações sem filtro. Dos três filtros testados, foi observado que a mediana móvel simples foi consistente para a redução de FP em ambas as abordagens e que a média exponencial móvel foi a mais impactada pelo parâmetro de entrada.

1.5 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação foi estruturada em seis capítulos, sendo o primeiro responsável por uma introdução ao tema de detecção de anomalias e os problemas que foram explorados. O Capítulo 2 apresentam conceitos que são explorados neste trabalho como AD, a arquitetura de um SAE e conceitos de *thresholds* e LPFs que foram aplicados para a detecção. O Capítulo 3 apresenta a metodologia geral usada para o desenvolvimento e suas particularidades adotadas e o Capítulo 4 apresenta o estudo de caso analisado e os respectivos experimentos deste trabalho realizdos para alcançar os objetivos previamente estabelecidos. Já o Capítulo 5 apresenta os resultados obtidos e realiza uma discussão a respeito dos experimentos executados no estudo de caso. Por fim, o Capítulo 6 apresenta as principais conclusões alcançadas com o desenvolvimento deste trabalho bem como sugestões para trabalhos futuros.

2 DETECÇÃO DE ANOMALIAS

Nesta seção serão apresentados alguns conceitos que são explorados neste trabalho. Esta seção apresenta uma introdução ao tema de AD, com ênfase na arquitetura AE na subseção 2.1, além de apresentar alguns *thresholds* que são utilizados na literatura e que serão avaliados neste trabalho na seção 2.2. Por fim, a seção 2.3 apresenta o conceito de LPF que também serão avaliados.

Anomalias são padrões presentes em dados que não correspondem a um comportamento normal. Ao longo dos anos, vários métodos para a AD foram desenvolvidos para as mais diversas aplicações, tais como: identificar fraude em cartão de crédito, detectar anomalias em exames de imagem e detectar falhas em componentes industriais. Por isso, as técnicas de detecção de anomalias são desenvolvidas para casos específicos e dependem diretamente da natureza dos dados, disponibilidade de dados devidamente etiquetados e tipos de anomalias a serem detectadas (CHANDOLA; BANERJEE; KUMAR, 2009). Três abordagens de AD se destacam na literatura:

- **Detecção Supervisionada:** são conhecidas quais instâncias são normais e anômalas por meio de etiquetas ou tabelas verdade e utiliza classificadores para predizer a qual classe uma nova instância mais se relaciona. Essa abordagem tende a ser mais robusta, mas apresenta desafios como bases de treinamento desbalanceadas, reconhecimento de tipos de anomalias não representadas antes no treinamento e alto custo de obtenção de dados rotulados corretamente e com boa representatividade;

- **Detecção Não Supervisionada:** são técnicas que não utilizam rótulos ou conhecimento prévio para realizar a detecção. Para esta categoria de detecção geralmente são usadas estruturas de clusterização, isolamento e técnicas de DL, tais como os AEs;

- **Detecção Semi-Supervisionada:** são técnicas que utilizam uma parcela de instâncias com etiquetas conhecidas, tipicamente as instâncias anômalas são conhecidas, ou são treinadas apenas com dados normais, e classificam a anomalia a partir do preceito de que estas são diferentes das instâncias em que foi realizado o aprendizado.

Destas três abordagens, a detecção não supervisionada vem ganhando bastante destaque nos últimos anos pelo menor custo e ser adequada para cenários com pouca ou nenhuma disponibilidade de dados rotulados, além de ser capaz de identificar tipos de anomalias que

não foram necessariamente observados anteriormente. Porém, é importante ressaltar que a abordagem não supervisionada, em geral, pode gerar uma maior taxa de FP e tende a ser menos robusta do que abordagens supervisionadas (UMER et al., 2022).

A detecção não supervisionada pode ser categorizada ainda em tradicional ou baseada em DL. Dentre as abordagens tradicionais, as mais conhecidas são baseadas em: (1) estatística, quando a detecção de anomalias depende da relação da instância com a distribuição das amostras; (2) distância entre instâncias, sendo anomalias instâncias mais afastadas de seus vizinhos; (3) densidade da região, sendo regiões menos densas consideradas anômalas; (4) clusterização, em que clusters menores indicam anomalia; e (5) *ensemble* que consideram a combinação de modelos para reduzir a dependência de um modelo exclusivo, com destaque para a técnica paralela do Isolation Forest (WANG; BAH; HAMMAD, 2019).

Com o aumento do poder computacional ao longo dos anos e devido a acurácia destes modelos, métodos de detecção baseados em DL vêm sendo amplamente utilizados, em especial para cenários que utilizam um grande número de dados. Os modelos de DL podem extrair automaticamente características dos dados originais com alta dimensionalidade e identificar com acurácia a saúde de equipamentos e ganharam muito destaque nos últimos anos, em especial os SAEs (SUN et al., 2016; WEN; GAO; LI, 2019; DAVARI et al., 2021).

2.1 AUTOENCODER

Os AEs são modelos de DL que ganharam muito destaque para a detecção de anomalia em diversos cenários por serem capazes de aprender padrões complexos de dados que podem ser treinados de forma não supervisionada e serem capazes de indicar potenciais problemas ou modos de falha que ainda não ocorreram no equipamento sob análise para aplicações relacionadas a manutenção preditiva (CANCEMI et al., 2023). Os AEs são redes neurais não supervisionadas treinadas para reconstruir os valores de entradas na camada de saída a partir da minimização do erro de reconstrução por *backpropagation*, podendo estas serem imagens, séries temporais ou dados tabulares (LI et al., 2022; ALAMPAY; ABU, 2020). O objetivo destas redes neurais é aprender a função aproximada da identidade para a saída $\hat{x}_l$ que melhor representa a entrada x_l. Quando parte dos dados de entrada são correlacionados, o AE vai compreender uma parcela destas correlações (NG et al., 2011).

Os AEs são redes neurais simétricas constituídos de um *encoder*, uma camada latente e um *decoder*. O *encoder* tem como função condensar a entrada do AE para uma menor

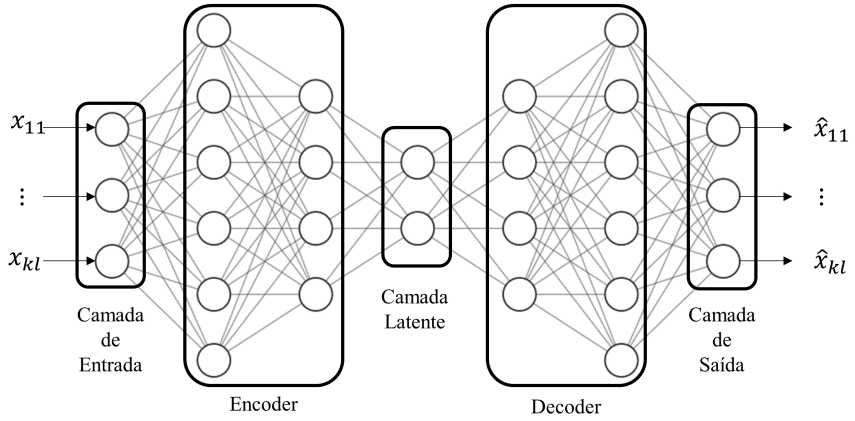
dimensão, a camada latente possui a informação codificada que representa os dados originais em sua menor dimensionalidade, e o *decoder* que reconstrói a informação codificada. Por isso, AEs também são aplicados para reduzir a dimensão de dados com alta dimensionalidade, de forma similar a aplicação do *Principal Component Analysis* (PCA), porém de forma não-linear (PURKAIT, 2019).

Para adaptar este modelo para a tarefa de AD, o erro de reconstrução da entrada de uma determinada instância é considerado um *score* de anomalia para essa instância. Para classificar as instâncias como normais ou anômalas é utilizado um valor *threshold*. Se o erro de reconstrução for abaixo do *threshold* informado, as instâncias são consideradas normais. Caso contrário, são consideradas anômalas. Por isso, a definição do valor do *threshold* de decisão é crítica, uma vez que valores baixos aumentam a incidência de FP e valores altos podem reduzir o TP.

Há quatro variações principais de AEs que são: Undercomplete Autoencoder, Variational Autoencoder, Denoising Autoencoder e Sparse Autoencoder (SAE). SAE são usados tipicamente para aprender características ligadas com classificação e por isso são muito usados para detectar anomalias em conjuntos com alta dimensionalidade (GOODFELLOW; BENGIO; COURVILLE, 2016; TUN; NYAUNG; PHYU, 2020; DAVARI et al., 2021).

SAE é uma derivação de AE em que um termo de penalização de esparsidade é adicionado ao AE original na camada latente. Com isso, o SAE possuí um número de neurônios na primeira camada escondida maior e apenas uma parcela de neurônios é ativada e treinada simultaneamente durante o codificação e decodificação (WEN; GAO; LI, 2019). A Figura 1 apresenta a arquitetura geral de um SAE.

Figura 1 – Arquitetura do SAE.



Fonte: A autora (2024).

A Equação 2.1 apresenta a função de custo proposta por (NG et al., 2011). A função de custo pode ser dividida em três partes: (1) o Erro Quadrático Médio (do inglês *Mean Squarred Error*) (MSE) de reconstrução; (2) a regularização dos pesos L2; e (3) a regularização de esparsidade por meio da divergência de Kullback-Leibler (KL).

$$J_{SAE} = \frac{1}{m}\sum_{l=1}^{m}|\hat{\mathbf{X}} - \mathbf{X}|^2 + \lambda||W||_2^2 + \beta\sum_{h=1}^{s2}KL(\rho||\hat{\rho}_h) \tag{2.1}$$

onde no termo de MSE: m é o número de entrada, $\hat{\mathbf{X}}$ é a reconstrução da entrada $\mathbf{X}$ que é composta pelos l parâmetros de uma instância k. No termo de regularização dos pesos: λ é o decaimento de peso usado para prevenir o sobreajuste e $||W||$ são os pesos dos neurônios do SAE. No terceiro termo, de regularização de esparsidade, β é o peso de esparsidade, h é o índice das unidades escondidas e $s2$ é o número de neurônios na camada latente, ρ e $\hat{\rho}_h$ são, respectivamente, o parâmetro de esparsidade e o parâmetro de ativação médio do neurônio h.

A regularização de esparsidade por meio da divergência de KL também pode ser escrita pela equação 2.2. Esta regularização adiciona uma penalidade adicional quando o parâmetro de ativação médio de um neurônio h desvia significativamente do parâmetro de esparsidade preestabelecido. ρ tipicamente é um valor baixo e próximo a zero que reduz a ativação média de cada unidade escondida h. Desta forma, a divergência de KL é uma função que mensura quão diferente duas distribuições de Bernoulli com médias ρ e $\hat{\rho}_h$ realmente são.

$$\sum_{h=1}^{s2}KL(\rho||\hat{\rho}_h) = \sum_{h=1}^{s2}\rho log\frac{\rho}{\hat{\rho}_h} + (1-\rho)log\frac{1-\rho}{1-\hat{\rho}_h} \tag{2.2}$$

2.2 THRESHOLD

A definição do *threshold* pode afetar consideravelmente o desempenho de um modelo de AD. Esta tarefa é especialmente desafiadora devido a ausência de etiquetas em conjunto de dados reais e em abordagens não supervisionadas de detecção. Perini, Bürkner e Klami (2023) reforçam que classificar um determinado score de anomalia com um fator de contaminação (proporção esperada de anomalias em um determinado conjunto de dados ou cenário) ou *threshold* definido incorretamente deteriora a performance do detector e, consequentemente, reduz a confiabilidade no sistema de detecção desenvolvido para melhorar a tomada de decisão.

Perini, Bürkner e Klami (2023) estudaram 21 abordagens para determinar um fator de contaminação que se refere a proporção de anomalias esperadas em um conjunto de dados

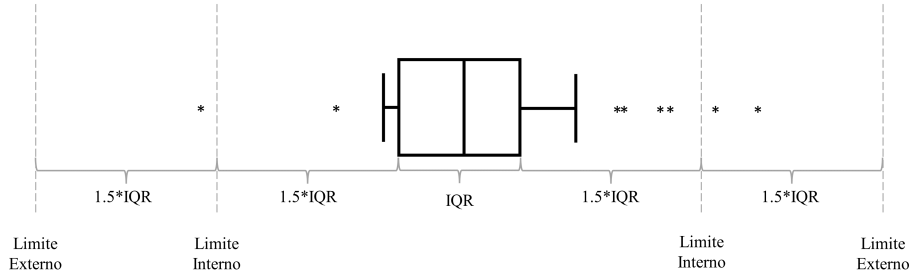
de forma não supervisionada para definição posterior do limiar de decisão. Dentre estas abordagens foram avaliados 9 categorias de estimadores de *threshold* baseados em kernel, curva, normalidade, regressão, filtro, teste estatístico, momento estatístico, quantil e transformação, além de um desenvolvimento próprio denominado γGMM. O algoritmo proposto pelos autores é executado em três passos e estima o fator de contaminação na perspectiva bayesiana a posteriori a partir de um conjunto de detectores de anomalia não supervisionados com um ajuste em um modelo de mistura Bayesiana-Gaussiana com processo prévio de Dirichlet seguido da estimativa da probabilidade dos pontos mais extremos em *score* serem anomalias. Para os experimentos foram utilizadas 22 bases de dados e 10 detectores de anomalia com princípios distintos como Isolation Forest e Autoencoder. A abordagem própria γGMM apresentou o menor Erro Absoluto Médio (do inglês *Mean Absolute Error*) (MAE) e maior taxa de deterioração F1, onde uma maior deterioração representa melhores valores.

2.2.1 Percentil

Uma abordagem intuitiva de determinar o *threshold* é baseado no conceito de N-percentil. A partir dos *scores* de anomalia obtidos durante o treinamento de um modelo de AD, é possível os ordenar de forma crescente e estimar o *threshold* como um percentil de N desta amostra. Dessa forma, quão maior for o N determinado, mais alto será o *threshold* e menor deverá ser o FP.

Após a definição de N, é calculado o N-percentil do score de anomalia gerado pelo modelo de AD ao longo do treinamento. Borghesi et al. (2019) aplicaram o conceito de N-percentil do erro de reconstrução para o cálculo de *threshold* em um Autoencoder semi-supervisionado testando os percentis 95, 97 e 99 e não encontrou uma relação positiva ao simplesmente aumentar o limiar de decisão devido a redução de TP. Tun, Nyaung e Phyu (2020) também adotaram o conceito de N-percentil em um SAE, com uma variação de n entre 1 e 100, mas determinaram o melhor *threshold* pelo conceito de *recall*, o que torna a abordagem semi-supervisionada. Apesar de ambos trabalhos explorarem o conceito de percentil, não foram exploradas abordagens com diferentes estratégias para a definição do limiar em modelo completamente não supervisionado.

Figura 2 – Parâmetros do *boxplot*.



Fonte: A autora (2024).

2.2.2 Erro Máximo

Esta abordagem é baseada no trabalho desenvolvido por Li et al. (2022), onde o *threshold* foi computado a cada ciclo do aprendizado online de um AE e considera como limite de anomalia o maior valor de score encontrado na janela de treinamento. De forma intuitiva, essa abordagem de *threshold* tende a gerar uma taxa de FP inferior a primeira abordagem, porém dependendo do comportamento das anomalias pode reduzir o TP. Apesar do trabalho comparar o modelo proposto com diferentes detectores da literatura, não testou diferentes abordagens de cálculo de *threshold*.

2.2.3 Intervalo Interquartil

A abordagem baseada em Intervalo Interquartil (do inglês *Inter Quantile Range*) (IQR) é amplamente utilizada e foi aplicada por Antwarg et al. (2021) para o cálculo do limiar de decisão para detecção de anomalias baseado no erro de reconstrução de um Autoencoder. Além disso, esta abordagem de *threshold* atingiu o quarto menor MAE e a sétima maior deterioração de F1-Score em Perini, Bürkner e Klami (2023) dentre as 21 abordagens testadas.

O IQR é uma forma de avaliar o grau de espalhamento de uma distribuição em torno da mediana. Para o cálculo desta abordagem, apresentado na equação 2.3, os *scores* de anomalia das instâncias i são ordenados de forma crescente e é considerado o limite interno superior da distribuição por representar aproximadamente 3 desvios padrões apresentado na Figura 2.

$$thres = Q3 + 1.5 * (Q3 - Q1) \tag{2.3}$$

onde $Q1$ e $Q3$ são, respectivamente, o primeiro e terceiro quartil do *score* de anomalia.

2.2.4 Boxplot Extremo

Esta abordagem foi aplicada por Davari et al. (2021) e também é baseada na análise de *boxplot* do *score* de anomalia. Sendo assim, se o *score* do conjunto de validação ou teste for superior ao limite superior do *score* de treinamento, apresentado na Figura 2 como limite externo superior, uma determinada instância é considerada anomalia. A abordagem é bem similar a abordagem IQR, porém considera como anomalia os valores mais extremos da distribuição e tende a obter um menor FP por isso. Então, o *threshold* é computado de acordo com a equação 2.4.

$$thres = Q3 + 3 * (Q3 - Q1) \tag{2.4}$$

onde $Q1$ e $Q3$ são, respectivamente, o primeiro e terceiro quartil do *score* de treinamento.

2.3 FILTROS

Um grande desafio para AD não supervisionado é a alta taxa de FP durante a detecção. É comum que os algoritmos indiquem que instâncias normais sejam anomalias. De acordo com Ribeiro, Pereira e Gama (2016), um LPF suaviza mudanças abruptas de um sinal a ser filtrado, o que reduz sua amplitude e a taxa de FP. Para AD, o sinal a ser filtrado é a sequência de *scores* de anomalia de conjunto de instâncias ordenadas no tempo. O LPF, uma vez aplicado, poderia por exemplo atenuar picos no *score* de anomalia, dentro de uma sequência de dados bem comportados, o que possivelmente seriam casos de FP.

Há diversos tipos de LPF, sendo os mais simples baseados em janelas deslizantes, tais como Média Móvel Simples (do inglês *Simple Moving Average*) (SMA), Mediana Móvel Simples (do inglês *Simple Moving Median*) (SMM) e Média Exponencial Móvel (do inglês *Exponential Moving Average*) (EMA). Estes filtros podem ser utilizados para o processamento de sinais digitais para suprimir ruídos brancos sem necessitar de uma frequência de corte específica (PASTELL, 2016).

2.3.1 Média Móvel Simples

Consiste em gerar um sinal de saída que equivale a média aritmética de sinais de entrada durante uma janela de n instâncias. Este filtro pode ser centrado, considerando valores passados e futuros em relação a uma instância, ou lateral que considera apenas n instâncias passadas para realizar a suavização.

$$Z_t = \frac{W_{t-n} + W_{t-(n+1)} + ... + W_{i-1} + W_t}{n} \tag{2.5}$$

onde Z e W são, respectivamente, saída e entrada do filtro, t representa o índice temporal e n o tamanho da janela. Vale ressaltar que n menores provocam uma reação mais brusca do filtro e que este é o único parâmetro de entrada.

2.3.2 Mediana Móvel Simples

Similar ao SMA, porém utiliza o conceito de mediana para o cálculo da saída do filtro. Por isso, o SMM é menos susceptível a valores extremos e apresenta uma suavização mais rápida. Assim como SMA, o único parâmetro de entrada necessário é o tamanho da janela a ser utilizada n.

2.3.3 Média Móvel Exponencial

Diferente dos demais, este filtro utiliza uma relação exponencial para efetuar a suavização. Matematicamente, EMA pode ser representada pela equação 2.6.

$$Z_t = Z_{t-1} + \alpha(W_t - Z_{t-1}) \tag{2.6}$$

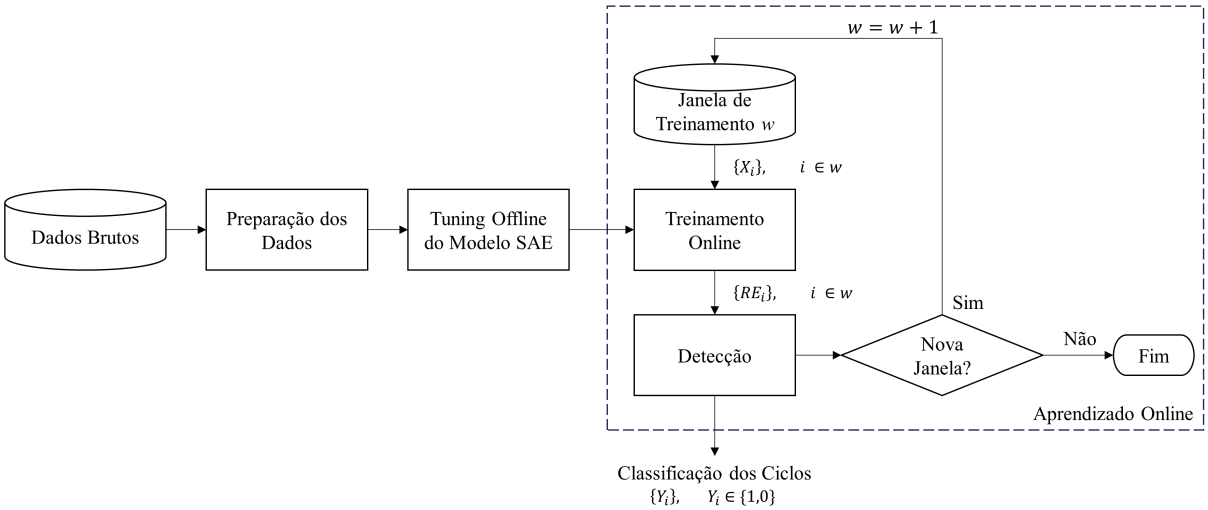
onde Z, W são, respectivamente a saída e entrada do filtro, t representa o índice temporal e α é o parâmetro de suavização.

Esta classe de filtros foi aplicada por Ribeiro, Pereira e Gama (2016) e Davari et al. (2021) para atenuar sinais com alta variabilidade em detecção de anomalia após a classificação de anomalias do modelo. É importante salientar que a escolha do parâmetro de suavização é muito relevante e que valores pequenos tornam a resposta do filtro mais lenta devido a maior inércia da entrada.

3 METODOLOGIA E IMPLEMENTAÇÃO

A presença de anomalias rotuladas em aplicações no mundo real não é comum, em grande parte devido ao alto custo do auxílio de especialistas. Por isso, ferramentas de AD não supervisionadas ou semi-supervisionadas vem sendo cada vez mais utilizadas em aplicações reais, com destaque para técnicas de aprendizagem de máquina profunda como os AEs. Para AD, dada uma instância de entrada, o modelo AE realiza inicialmente a sua reconstrução. O erro de reconstrução da entrada do modelo é usado então para definir um *score* de anomalia. Assim, um *threshold* é usado para classificar instâncias como normais ou anômalas. Se o erro de reconstrução for abaixo do limiar, as instâncias são consideradas normais. Caso contrário, são consideradas anômalas. Por isso, a definição do valor do *threshold* é crítica, uma vez que valores baixos aumentam a incidência de FP. Esse é problema comum em sistemas de AD, que apesar de serem sensíveis para detectar anomalias reais, retornam um grande número de falsos alarmes.

Figura 3 – Diagrama da metodologia adotada.



Fonte: A autora (2024).

O presente trabalho avaliou a influência da seleção do *threshold* para modelos de SAE para AD não supervisionada. Outro fator que também foi avaliado neste trabalho foi a aplicação de LPF que almejam reduzir a taxa de FP detectados (RIBEIRO; PEREIRA; GAMA, 2016). A Figura 3 ilustra brevemente a metodologia adotada por este trabalho. Para este fim, foi realizado uma preparação dos dados brutos conforme descrito na seção 3.1 e uma etapa de *tuning offline*, descrita na seção 3.2, foi realizada para encontrar os hiperparâmetros a serem utilizados no

modelo SAE para a etapa de aprendizado *online*. O aprendizado *online* foi dado até que todas as janelas de treinamento fossem percorridas e nele foram realizadas duas etapas de detecção, sendo uma no conjunto de validação e outra no conjunto de teste. Para isso, o modelo SAE recebeu um conjunto de dados X_i de uma janela w e calculou um erro de reconstrução RE_i do modelo treinado. Na etapa de detecção, as instâncias avaliadas foram classificadas e armazenadas em uma variável Y_i em que 1 representa anomalia e 0 uma instância normal. A seção 3.3 apresenta as configurações de detecção utilizadas considerando as abordagens de aplicação de LPF testadas, bem como os tipos de LPF e abordagens de *threshold* avaliados. Após o término do aprendizado *online*, a classificação de ciclos foi avaliada de acordo com as métricas e ferramentas descritas na seção 3.5.

3.1 PREPARAÇÃO DOS DADOS

A metodologia desenvolvida pode ser usada para diversos cenários de detecção de anomalia envolvendo séries temporais e dados tabulares. Um aspecto importante da implementação foi o processo de normalização dos dados, que para este estudo foi adotado por *z-score*, conforme a equação 3.1, para garantir que todos os atributos estivessem na mesma escala.

$$z = \frac{y - \mu}{\sigma} \tag{3.1}$$

onde z é uma variável normalizada, y é o valor de um determinado atributo, μ e σ são, respectivamente, a média aritmética e o desvio padrão dos valores do atributo y.

3.2 TUNING OFFLINE DO MODELO SAE

Para realizar a definição dos hiperparâmetros da arquitetura SAE a ser utilizada, foi utilizado o *framework* KerasTuner (O'MALLEY et al., 2019). A busca aleatória por hiperparâmetros foi dada de forma não supervisionada com objetivo de obter uma arquitetura que atendesse uma redução do erro de reconstrução no conjunto de validação em um conjunto de dados reduzido. Os hiperparâmetros que foram otimizados e seus respectivos intervalos são apresentados na Tabela 1.

Vale ressaltar que foi considerada uma arquitetura com 3 camadas escondidas com *decoder* simétrico ao *encoder*. A proporção adotada entre número de neurônios das camadas foi

Tabela 1 – Espaço de busca de parâmetros para a arquitetura SAE.

Hiperparâmetros	**Intervalo**
# Neurônios na 1ª camada escondida	[120, 240, 480]
β	[3, 5, 7, 9]
λ	[1e-7, 1e-5, 1e-3]
ρ	[0.01, 0.05, 0.1, 0.2]
Otimizador	Adam
Taxa de Aprendizado	[1e-3, 1e-2, 1e-1]
Tamanho de Mini-Lote	[32, 64, 128]

Fonte: A autora (2024).

conforme descrita a seguir:

- # Neurônios na 2ª camada escondida: # Neurônios na 1ª camada escondida/2
- # Neurônios na 3ª camada escondida: # Neurônios na 1ª camada escondida/4
- # Neurônios na camada latente: # Neurônios na 1ª camada escondida/10

A técnica *Random Search* foi utilizada para gerar de forma aleatória um total de 50 combinações de hiperparâmetros que foram treinados 10 vezes devido a variações estocásticas inerentes ao treinamento de arquiteturas de DL. Além disso, os modelos foram treinados por 250 épocas com parada antecipada com paciência de 50 épocas para evitar o sobreajuste dos modelos gerados.

3.3 DETECÇÃO

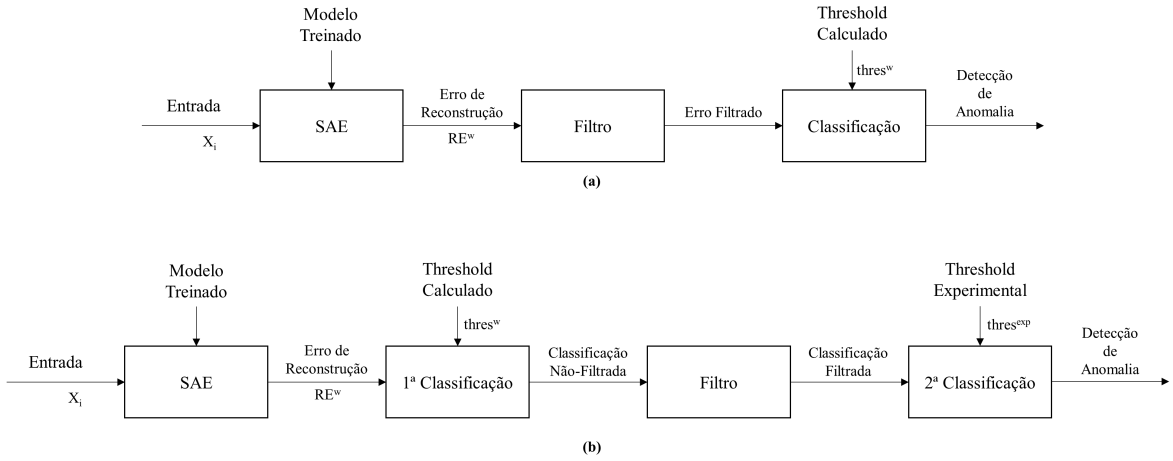
Neste trabalho foram testadas um total de 36 opções de configuração com variação de: técnica de cálculo de *threshold* adaptativo baseado nos fundamentos apresentados em 2.2; abordagem para aplicação de filtro que são explicadas na subseção 3.3.1; e, por fim, tipo de LPF, dentre as três opções apresentadas na subseção 3.3.2. Além disso, para utilizar como referência, foram experimentadas 6 configurações com alteração apenas da técnica de cálculo de *threshold* para desconsiderar efeitos da aplicação do LPF na avaliação de AD.

3.3.1 Seleção da Abordagem do Filtro

Após o cálculo do *threshold* de cada janela w no conjunto de treinamento, foi realizada a detecção de anomalia no conjunto de validação e de teste por duas abordagens distintas. As

abordagens apresentadas na Figura 4 indicam como foi aplicado o LPF nos experimentos.

Figura 4 – **(a)** Abordagem Pré-Classificação: Filtro sob o erro de reconstrução, **(b)** Abordagem Pós-Classificação: Filtro sob a classificação preliminar.



Fonte: A autora (2024).

A Abordagem Pré-Classificação de Filtro (Pré-Classificação) (Figura 4.(a)) foi utilizada para suavizar o erro de reconstrução RE_i para reduzir o número de FP e possibilitar a detecção de sequências de falhas de forma totalmente não supervisionada, já que uma das características do conjunto de dados usados é a presença de ciclos subsequentes com anomalias. Para isso, o sinal de erro de reconstrução RE^w da janela w é dado como entrada de um LPF e resulta em um sinal filtrado de erro de reconstrução. Em seguida, é executada a classificação de cada ciclo avaliado sob o sinal filtrado considerando o *threshold* $thres^w$ calculado durante a respectiva janela de treinamento. Se o sinal filtrado de um ciclo i da janela w de validação ou teste for menor do que o *threshold*, o ciclo é classificado como normal (0). Caso contrário, anômalo (1).

A Abordagem Pós-Classificação de Filtro (Pós-Classificação) (Figura 4.(b)) foi baseada nos trabalhos de Ribeiro, Pereira e Gama (2016) e Davari et al. (2021) e a entrada do LPF é uma classificação inicial. Nesse caso, o LPF é aplicado sob a sequência binária gerada na primeira classificação baseada apenas no erro de reconstrução RE^w da janela w e o *threshold* computado na janela w de treinamento $thres^w$. Para realizar a detecção final, é necessário utilizar um segundo *threshold* constante pré-definido $thres^{exp}$. Neste trabalho, foi adotado $0,5$ como segundo *threshold* assim como realizado por Ribeiro, Pereira e Gama (2016). Então, os valores de classificação filtrada acima de $thres^{exp}$ são etiquetadas como anomalia e abaixo são considerados ciclos normais.

3.3.2 Seleção de Filtro

Com objetivo de avaliar o efeito da aplicação de LPF na otimização de FP e suas consequências em TP e F1-Score, foram avaliados três tipos de LPF simples baseados em janelas deslizantes SMA, SMM e EMA apresentados em 2.3. Os parâmetros de entrada apresentados na Tabela 2 foram experimentados para estudar o efeito de cada filtro na suavização do erro de reconstrução. Além disso, foram considerados o mesmo tamanho de janela em SMA e SMM para permitir a comparação direta entre os filtros e para o EMA foram considerados valores pequenos (entre 0.01 e 0.09) por terem uma suavização mais lenta e α maiores (entre 0.1 e 0.9) para uma suavização mais rápida.

Tabela 2 – Parâmetros de entrada avaliados

Filtro	**Parâmetro de Entrada**	**Intervalo**
SMA	Tamanho da Janela (n)	[3, 10]
SMM	Tamanho da Janela (n)	[3, 10]
EMA	Parâmetro de Suavização (α)	0,01; 0,03; 0,05; 0,07; 0,09; 0,1; 0,3; 0,5; 0,7; 0,9

Fonte: A autora (2024).

3.4 APRENDIZADO ONLINE

Neste trabalho foi desenvolvida uma arquitetura de aprendizagem *online*, de forma a aproximar de uma aplicação de manutenção preditiva convencional avaliada nos dados de estudo de caso e permitir períodos de aprendizagem mais curtos utilizando a técnica de janelas deslizantes. Devido as variações estocásticas de modelos de DL, o processo de aprendizado *online* foi repetido 10 vezes para cada configuração de detecção testada para realização avaliação estatística das métricas e melhor avaliação do comportamento destas.

3.4.1 Treinamento do Modelo

Na etapa de aprendizado o SAE foi treinado de forma não supervisionada por meio da minimização do erro de reconstrução calculado através do MSE. A etapa de aprendizado utilizou dois conjuntos de dados: um conjunto de dados sem anomalias de treinamento e um conjunto de validação com uma semana de informações com $1/3$ do tamanho. O treinamento

utilizou validação cruzada para evitar o enviesamento da rede com a parada antecipada com um máximo de 500 épocas e uma paciência de 100 épocas.

3.4.2 Cálculo do Threshold

A partir do erro da reconstrução obtido no conjunto de treinamento, foi possível calcular o *threshold* adaptativo $thres^w$ da janela w de acordo com a opção do *threshold* selecionada no módulo de detecção. Sendo assim, se o erro de reconstrução do conjunto de validação ou teste de uma instância i da janela w for superior ao $thres^w$, esta é considerada anomalia. Essa abordagem foi adotada já que o comportamento de cada uma das janelas pode variar ao longo do tempo.

3.4.3 Detecção no Conjunto de Validação

Após o cálculo do *threshold* $thres^w$ de cada janela w, foi realizada a detecção de anomalia no conjunto de validação. A AD no conjunto de validação depende da seleção do LPF e também da abordagem de filtro selecionada e segue o direcionamento descrito na subseção 3.3.1. A partir da classificação realizada, as anomalias detectadas no conjunto de validação de cada janela deslizante w foram removidas do conjunto de treinamento da janela $w + 1$, o que possibilitou um tratamento não supervisionado de dados julgados como anômalos. A ausência de anomalias no treinamento é essencial para otimizar a detecção de anomalias por SAE.

3.4.4 Detecção no Conjunto de Teste

Após a remoção de anomalias verificadas no conjunto de validação, foi realizada a detecção de anomalias no conjunto de teste com o mesmo procedimento e parâmetros adotados inicialmente. Após o aprendizado e teste do AD na janela w, todas as janelas foram deslizadas em uma semana e o processo foi repetido até que não houvesse nenhuma entrada a ser testada.

3.5 AVALIAÇÃO DA DETECÇÃO

As anomalias que foram detectadas em cada janela de teste foram avaliadas e comparadas com as anomalias conhecidas e rotuladas em (DAVARI et al., 2021). Para isto, foram computados

algumas métricas tradicionais de estudos de detecção de anomalia e classificação, como *Recall*, Precisão e F1-Score, além da Taxa de Falso Positivo (do inglês *False Positive Rate*) (FPR). As equações utilizadas estão descritas nas equações 3.2, 3.3, 3.4 e 3.5 respectivamente. Nessa avaliação foi considerado como positivo uma anomalia e negativo um ciclo normal.

$$Recall = \frac{TP}{TP + FN} \tag{3.2}$$

$$Precisão = \frac{TP}{TP + FP} \tag{3.3}$$

$$F1 - Score = 2 * \frac{Precisão * Recall}{Precisão + Recall} \tag{3.4}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.5}$$

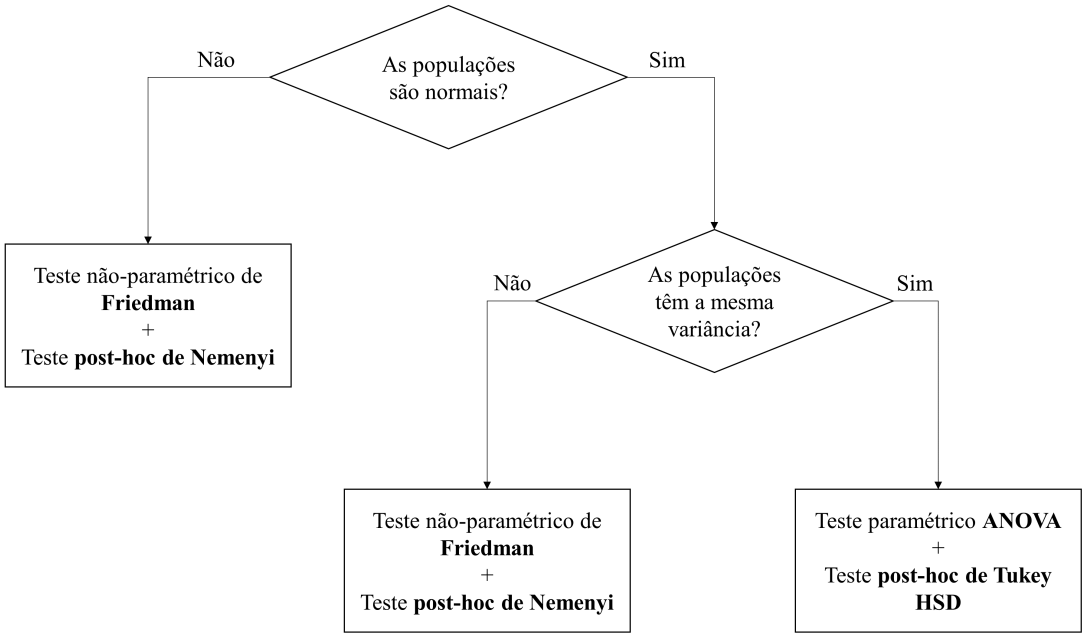
Além destas métricas, também foi avaliada a capacidade dos modelos de aprendizagem profunda de predizer com uma antecedência de duas horas a presença da anomalia. Por fim, foi avaliada a influência dos filtros em ambas abordagens apresentadas nesta seção e da seleção do *threshold* utilizado o conjunto de dados pré-processado em ciclos considerando a minimização de FPR, a maximização de TP pela maximização do *recall* e a maximização de F1-Score.

Por fim, foi realizada uma avaliação estatística das métricas avaliadas utilizando o pacote Autorank desenvolvido por Herbold (2020) considerando as diretrizes propostas por Demšar (2006). A Figura 5 apresenta a sequência de testes estatísticos realizadas para a avaliação de mais de duas populações.

Inicialmente, é avaliado se as populações observadas podem ser consideradas normais por meio do teste estatístico de Shapiro-Wilk com correção de Bonferoni. Caso o $p - valor$ deste teste seja inferior a $0,0083$, considerando um $\alpha = 0,05$ e 6 testes de hipóteses, por exemplo, a hipótese nula de que não diferença entre a distribuição avaliada e a distribuição normal é rejeitada com 95% de confiança. Se não, as populações são consideradas normais por não ser possível rejeitar a hipótese nula. Caso as populações sejam consideradas normais, é necessário avaliar a homogeneidade das variâncias com a aplicação do teste estatístico de Barlett. Para este teste a hipótese nula é de que as variâncias das amostras são homogêneas e se o $p-valor$ for inferior a $0,05$ as amostras são consideradas heterogêneas quanto a variância.

Se as populações não forem normais ou se forem normais mas forem heterogêneas quanto a variância, é realizado o teste estatístico não-paramétrico de Friedman para determinar se há

Figura 5 – Fluxograma do pacote Autorank.



Fonte: A autora (2024).

diferenças significativas entre as médias ou medianas respectivamente. Se $p-valor$ foi inferior a $0,05$, a hipótese nula de que não há diferenças significativas na medida de tendência central é rejeitada com 95% de confiança. Quando há diferenças, é aplicado o teste post-hoc de Nemenyi para verificar se as distâncias entre as médias ou medianas são superiores a distância crítica calculada. Se a distância for superior, as populações são consideradas estatisticamente diferentes.

Se as populações são normais e são homogêneas quanto a variância, o teste estatístico paramétrico ANOVA é aplicado para verificar se houve ou não uma diferença estatisticamente relevante entre as médias das populações avaliadas. Se $p-valor$ for inferior a $0,05$, a hipótese nula de que a média de todas as populações são iguais é rejeitada com 95% de confiança. Se diferenças forem constatadas, é aplicado o teste post-hoc de Tukey HSD para verificar se há diferenças mínimas significativas entre os pares de amostras.

Com esta avaliação, é possível assumir se há diferenças significativas entre as métricas avaliadas nos cenários estudados. A avaliação foi realizada inicialmente entre os *thresholds* para a aplicação sem filtro e depois foi realizada considerando cada uma das 36 opções testadas para avaliar também o impacto do parâmetro de entrada do filtro e a comparação com o respectivo modelo sem filtro.

4 ESTUDO DE CASO

Nesta seção será apresentado particularidades referentes ao estudo de caso desenvolvido nesta dissertação, a exemplo da preparação da base de dados e *thresholds* adotados.

4.1 PREPARAÇÃO DOS DADOS

Para este trabalho, foi selecionado um conjunto de dados público para detecção de falhas (VELOSO et al., 2022), disponibilizado no repositório UCI por Davari et al. (2021). A base de dados apresenta sinais analógicos e digitais relacionados a Unidade de Produção de Ar (do inglês *Air Production Unit*) (APU) de um trem do metrô da cidade do Porto, Portugal. Os dados foram coletados entre fevereiro e agosto de 2020 e disponibilizados de forma tabular com uma frequência de amostragem de 0,1 Hz. Em cada tempo, são registradas as seguintes variáveis relacionadas à APU:

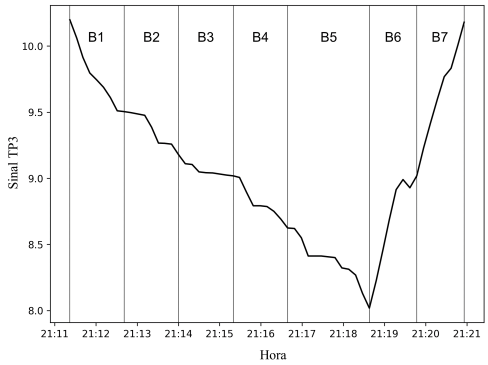
- **TP2**: sinal analógico de pressão no compressor (bar);
- **TP3**: sinal analógico que indica a pressão gerada no painel pneumático (bar);
- **H1**: sinal analógico ativado quando a pressão no pressostato de comando é superior a pressão de operação (10,2 bar);
- **DV_pressure**: sinal analógico referente a pressão resultante de uma queda de pressão que ocorre quando há descarga de água nas torres de secagem de ar (bar). Esta pressão é zero quando o compressor está atuando em sobrecarga;
- **Reservoirs**: sinal analógico que indica a pressão dos tanques de ar instalados nos trens (bar);
- **Oil_Temperature**: sinal analógico da temperatura do óleo contido no compressor (°C);
- **Motor_current**: sinal analógico que indica a corrente de uma fase do motor trifásico utilizado (A). Este sinal deve ser próximo de 0 A quando o compressor for desligado ou próximo a 4 A quando o compressor estiver operando sem carga ou próximo a 7 A quando o compressor estiver trabalhando em sobrecarga;

- **COMP**: sinal digital da unidade de controle eletrônico que indica o sinal elétrico na válvula de entrada de ar do compressor. Quando este sinal está ativo, não há admissão de ar no compressor devido ao desligamento do equipamento ou operando sem carga;
- **DV_Electric**: sinal digital da unidade de controle eletrônico que comanda a válvula solenoide de saída. Quando este sinal está ativo, o compressor está trabalhando em sobrecarga, caso contrário o compressor está desligado ou sem carga;
- **TOWERS**: sinal digital da unidade de controle eletrônico que define qual torre está responsável por secar o ar e qual torre está drenando a umidade. Quando este sinal está ativo, a torre dois está trabalhando, caso contrário, a torre um está trabalhando;
- **MPG**: sinal digital da unidade de controle eletrônico que indica a ativação da válvula de entrada de ar quando a pressão do APU é inferior a 8,2 bar e consequentemente ativa COMP;
- **LPS**: sinal digital da unidade de controle eletrônico que indica quando a pressão é inferior a 7 bar;
- **Pressure_switch**: sinal digital da unidade de controle eletrônico que indica pressão detectada na válvula de controle piloto;
- **Oil_Level**: sinal digital da unidade de controle eletrônico referente ao nível de óleo no compressor. Quando este sinal está ativo, o nível de óleo está abaixo do esperado;
- **Caudal_impulses**: sinal digital da unidade de controle eletrônico que é produzido pelo fluxômetro quando há fluxo de ar em um segundo.

A detecção de anomalias foi realizada por ciclos como proposto por (RIBEIRO; PEREIRA; GAMA, 2016) e (DAVARI et al., 2021), que são intervalos de dados definidos através do comportamento da variável analógica **TP3**, como apresentado na Figura 6. O fluxo para preparação dos dados é apresentado na Figura 7 e pode ser dividido em duas etapas principais: (1) Identificação do Ciclo e (2) Caracterização do Ciclo.

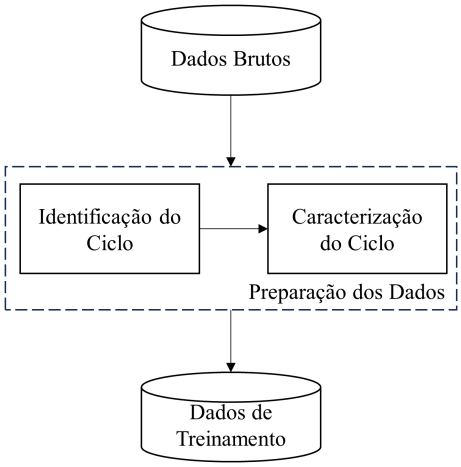
Inicialmente, para a identificação do ciclo, foi realizada uma avaliação de tendência da variável **TP3** e uma classificação das instâncias. Para a classificação foi considerado que uma tendência positiva indica um aumento de pressão gerada no painel pneumático e é relacionado a um período ativo do compressor (Run), enquanto uma tendência negativa representa um

Figura 6 – Exemplo de ciclo.



Fonte: A autora (2024).

Figura 7 – Fluxo para Preparação dos Dados.



Fonte: A autora (2024).

período em que o compressor está ocioso (Idle). Então, a definição de ciclo foi realizada considerando que o ciclo sempre se iniciaria em um período ocioso. Nesta etapa, foi verificado que apesar da base de dados original não possuir nenhum valor numérico indefinido, a série temporal não é contínua. Então, poderia gerar ciclos com períodos de tempo muito elevados devido a ausência de informação dentro deste ciclo. Por isso, foi definido que um ciclo só poderia ser considerado válido se não houvesse a ausência de dados por mais de 20 minutos ou se o período fosse uma anomalia, caso contrário o ciclo seria descartado. Também foram identificados ciclos em que havia uma repetição de sinais analógicos durante longos períodos de tempo o que pode ter ocorrido devido a problema de sensoriamento que também foram descartados.

Após a identificação dos ciclos, cada ciclo é então caracterizado através da segmentação dos sinais digitais e analógicos ao longo do ciclo, por meio de blocos. O período ocioso de

Tabela 3 – Entradas para os modelos SAE por cada ciclo.

Entrada	**Descrição da Entrada**
B_b^{TP2}, b=1,...,7	7 blocos para o parâmetro analógico **TP2**
B_b^{TP3}, b=1,...,7	7 blocos para o parâmetro analógico **TP3**
B_b^{H1}, b=1,...,7	7 blocos para o parâmetro analógico **H1**
B_b^{DVP}, b=1,...,7	7 blocos para o parâmetro analógico **DV_pressure**
B_b^{Res}, b=1,..,7	7 blocos para o parâmetro analógico **Reservoirs**
B_b^{OT}, b=1,...,7	7 blocos para o parâmetro analógico **Oil_Temperature**
B_b^{MC}, b=1,...,7	7 blocos para o parâmetro analógico **Motor_current**
COMP	Contagem de uns para o parâmetro digital **COMP**
DV_electric	Contagem de uns para o parâmetro digital **DV_electric**
TOWERS	Contagem de uns para o parâmetro digital **TOWERS**
MPG	Contagem de uns para o parâmetro digital **MPG**
LPS	Contagem de uns para o parâmetro digital **LPS**
Pressure_switch	Contagem de uns para o parâmetro digital **Pressure_switch**
Oil_Level	Contagem de uns para o parâmetro digital **Oil_Leve**l
Caudal_impulses	Contagem de uns para o parâmetro digital **Caudal_impulses**
T_idle	Período ocioso de um ciclo
T_run	Período ativo de um ciclo

Fonte: A autora (2024).

parâmetros analógicos do ciclo foi dividido em 5 blocos iguais, enquanto o período ativo foi dividido em 2 blocos equivalentes. Para os sinais analógicos foi realizada a média do sinal ao longo do bloco, enquanto para os sinais digitais foi realizada uma contagem de instâncias com valor 1. Dessa forma, foi gerado um conjunto de atributos para detecção de anomalias nos ciclos. A Tabela 3 apresenta uma caracterização dos atributos que foram utilizados para a detecção de anomalia. Após a definição deste conjunto, foi realizado o processo de normalização dos dados por *z-score* para garantir que todos os atributos estivessem na mesma escala.

4.2 TUNING OFFLINE DO MODELO SAE

Para realizar o *tuning offline* do modelo SAE foi seguida a implementação definida na Seção 3.2. O conjunto de dados utilizados nesta etapa correspondeu a um período de seis semanas nos meses de fevereiro e março da base original MetroPT, em que quatro semanas foram dedicadas ao treinamento e duas semanas à validação. Os hiperparâmetros que foram otimizados e seus respectivos intervalos foram apresentados na Tabela 1.

A Tabela 4 apresenta os hiperparâmetros selecionados pelos cinco melhores modelos gerados randomicamente, ou seja, os cinco modelos que geraram menor erro durante o período de

Tabela 4 – Melhores hiperparâmetros encontrados.

Hiperparâmetros	1° Modelo	2° Modelo	3ª Modelo	4º Modelo	5º Modelo
# Neurônios na 1ª camada escondida	**480**	480	480	240	480
β	**3**	5	5	5	7
λ	**1e-7**	1e-5	1e-3	1e-3	1e-3
ρ	**0.2**	0.2	0.2	0.2	0.2
Taxa de Aprendizado	**1e-3**	1e-3	1e-3	1e-3	1e-3
Tamanho de Mini-Lote	**32**	64	32	64	128

Fonte: A autora (2024).

validação. Foi possível perceber que houve uma preferência dos modelos por uma maior quantidade de neurônios na primeira camada escondida e que os parâmetros que mais oscilaram foi β, λ e tamanho de mini-lote. Já os hiperparâmetros ρ e taxa de aprendizado estabilizaram nos limites superiores do espaço de busca de parâmetros apresentado na Tabela 1.

A partir do *tuning*, foi adotada a arquitetura do modelo que apresentou melhores resultados apresentada na Tabela 5.

Tabela 5 – Arquitetura final do modelo SAE.

Hiperparâmetros	Escolha
Nós da camada de Entrada	59
# Neurônios na 1ª camada escondida	480
# Neurônios na 2ª camada escondida	240
# Neurônios na 3ª camada escondida	120
# Neurônios na camada latente	48
β	3
λ	1e-7
ρ	0.2
Otimizador	Adam
Taxa de Aprendizado	1e-3
Tamanho de Mini-Lote	32

Fonte: A autora (2024).

4.3 SELEÇÃO DE THRESHOLD

Como mencionado na subseção 3.3, neste trabalho foram testadas um total de 36 opções de configuração com variação de: cálculo de *threshold* adaptativo; abordagem para aplicação de filtro; e, por fim, tipo de LPF. Na presente seção, serão apresentadas apenas a seleção de *threshold*, já que os detalhes a respeito das abordagens de aplicação de filtro e tipo de LPF foram apresentadas respectivamente nas subseções 3.3.1 e 3.3.2.

Este trabalho propôs investigar o impacto da seleção da abordagem de cálculo do *threshold* adaptativo na classificação das anomalias no estudo de caso referente a base MetroPT. Para isso, foram avaliadas as quatro formas distintas descritas nesta subseção com a nomenclatura de ciclos introduzidas na subseção 4.1.

4.3.1 N-Percentil

A primeira abordagem avaliada para determinar o *threshold* foi baseada no conceito de N-percentil. Para isso, é avaliado o erro de reconstrução RE ao longo de uma janela de treinamento w, que é o *score* de anomalia da janela w, baseado que é calculado baseado no MSE expressa na equação 4.1.

$$RE_i = \frac{1}{m}\sum_{j=1}^{m}(\hat{x}_{ij} - x_{ij})^2 \tag{4.1}$$

onde RE_i representa o erro de reconstrução do ciclo i, m é o número de características e x_{ij} é a j-ésima característica do ciclo i. Sendo assim, cada ciclo i da janela de treinamento w gera um valor de erro RE_i e o conjunto de todos os erros de uma janela é representado como RE^w.

Para este trabalho, optou-se por testar três valores distintos de N, sendo eles 95, 97 e 99 tal como Borghesi et al. (2019). Estes *thresholds* serão abreviados da seguinte forma 95-percentil (N95), 97-percentil (N97) e 99-percentil (N99) para melhor compreensão no texto.

4.3.2 Erro Máximo

Esta abordagem é baseada no trabalho desenvolvido por (LI et al., 2022), onde o *threshold* foi computado a cada ciclo do aprendizado *online* de acordo com o maior erro de reconstrução encontrado na janela de treinamento RE^w, conforme a equação 4.2. Como mencionado na subseção 2.3, é esperado que essa abordagem gere um menor FP, porém também deve gerar um menor TP.

$$thres^w = max_{i\in\{1:n\}}RE_i \tag{4.2}$$

onde n é o número de ciclos i na janela de treinamento w. Para facilitar a compreensão Erro Máximo (MAX).

4.3.3 Intervalo Interquartil

A abordagem baseada em IQR é amplamente utilizada na literatura e para o cálculo desta abordagem, apresentado na equação 4.3, os RE_i dos ciclos i de uma janela w são ordenados de forma crescente.

$$thres^w = Q3_w + 1.5 * (Q3_w - Q1_w) \tag{4.3}$$

onde $Q1_w$ e $Q3_w$ são, respectivamente, o primeiro e terceiro quartil do erro de reconstrução do treinamento da janela w, RE^w. Essa abordagem será simplificada para IQR.

4.3.4 Boxplot Extremo

Esta abordagem foi aplicada por Davari et al. (2021) e também é baseada na análise de *boxplot* de RE^w. Sendo assim, se o erro de reconstrução RE_i de um ciclo i do conjunto de validação ou teste de uma janela w for superior a $thres^w$, o ciclo i é considerado anomalia. Então, a cada janela de treinamento w, o *threshold* é computado de acordo com a equação 4.4.

$$thres^w = Q3_w + 3 * (Q3_w - Q1_w) \tag{4.4}$$

onde $Q1_w$ e $Q3_w$ são, respectivamente, o primeiro e terceiro quartil do erro de reconstrução do treinamento da janela w, RE^w. Essa abordagem será simplificada para Boxplot Extremo (BOX).

4.4 APRENDIZADO ONLINE NO ESTUDO DE CASO

Como mencionado na seção 3.4, este trabalho optou por desenvolver uma arquitetura de aprendizagem *online* de forma a aproximar de uma aplicação de manutenção preditiva convencional relacionada ao estudo de caso avaliado. Para isso, foi adotada a técnica de janelas deslizantes considerando um conjunto de aprendizado (três semanas de treinamento e uma semana de validação) e teste (uma semana).

Além disso, o cálculo do *threshold* foi dado de forma adaptativa em cada janela w de acordo com a opção do *threshold* selecionada no módulo de detecção como apresentado na

subseção 4.3. Sendo assim, se o erro de reconstrução do conjunto de validação ou teste de um ciclo i da janela w for superior ao $thres^w$, o ciclo i é considerado anomalia.

Para o estudo de caso avaliado, também foi verificada a capacidade das configurações de detecção que maximizaram o F1-Score antecipar a presença de anomalias. Ou seja, a métrica **Antecipação** criada representa quantos ciclos de antecedência uma anomalia foi detectada considerando um limite de 10 ciclos, já que este valor representa uma antecipação de aproximadamente 3 horas de operação ao considerar uma média de 20 minutos por ciclo. É importante salientar que se uma anomalia não for detectada pelo modelo ou se for detectada apenas no ciclo em que esta ocorre, é atribuído o valor 0 para a métrica de Antecipação.

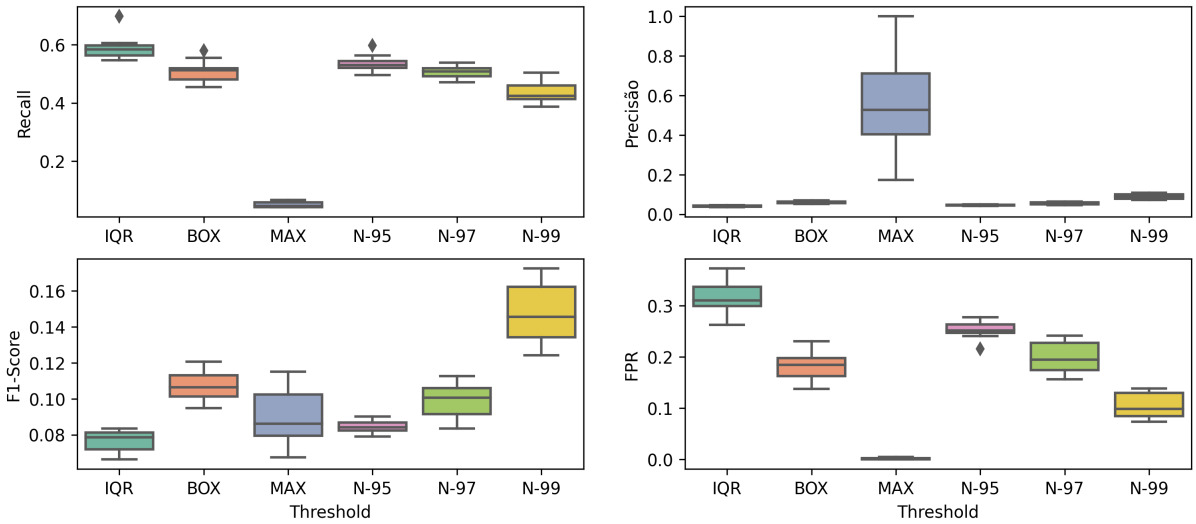
5 RESULTADOS E DISCUSSÕES

Nesta seção da dissertação, é realizada uma avaliação da influência da seleção do *threshold* na detecção de anomalias bem como a aplicação de cada abordagem de aplicação de filtro e tipo de filtro.

5.1 AVALIAÇÃO DO IMPACTO DA SELEÇÃO DO THRESHOLD

Para realizar uma avaliação do impacto da seleção do *threshold* nas métricas de AD, cada uma das seis abordagens de cálculo de *threshold* foi testada 10 vezes ao longo da base de dados considerando a arquitetura apresentada na Tabela 5 sem a aplicação de filtros para evitar influência externa. A distribuição das métricas é apresentada por meio de *boxplots* na Figura 8 e numericamente na Tabela 6. Já a Figura 9 apresenta um resumo visual do ranqueamento dos modelos com diferentes abordagens de cálculo de *threshold* com base na análise post-hoc de Nemenyi pelo pacote Autorank utilizado em *python*.

Figura 8 – Comportamento das métricas de avaliação por *threshold*.



Fonte: A autora (2024).

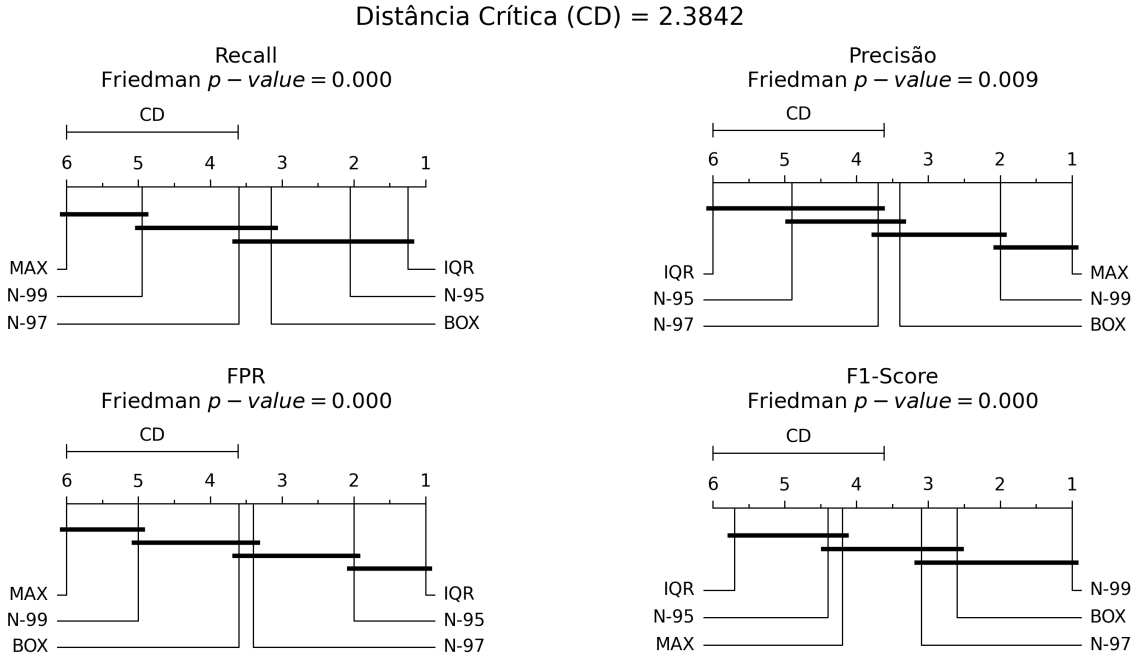
O *recall* dos modelos apresentou uma amplitude de variação da média elevada, tendo a abordagem MAX alcançado $5,04\%$ enquanto as demais alcançaram um *recall* entre $43,70$ e $58,91\%$. Essa discrepância se deu pelo baixo TP obtido pela abordagem MAX devido aos valores elevados de *threshold*. Já o elevado *recall* da abordagem IQR ocorreu devido aos menores valores de *threshold* calculados, o que consequentemente eleva o TP. Do ponto de

Tabela 6 – Métricas de Avaliação de AD para diferentes *thresholds*.

Métrica	**Threshold**					
	IQR	BOX	MAX	N95	N97	N99
Recall (%)	**58,91 ± 4,29**	50,67 ± 4,04	5,04 ± 0,97	53,53 ± 2,86	50,50 ± 2,29	43,70 ± 3,50
Precisão (%)	4,09 ± 0,34	6,00 ± 0,54	**56,37 ± 24,88**	4,60 ± 0,20	5,49 ± 0,63	8,92 ± 1,38
F1-Score (%)	7,65 ± 0,60	10,70 ± 0,81	9,05 ± 1,56	8,46 ± 0,34	9,89 ± 0,99	**14,72 ± 1,74**
FPR (%)	31,51 ± 3,24	18,25 ± 2,85	**0,14 ± 0,14**	25,28 ± 1,78	20,04 ± 3,11	10,48 ± 2,49

Fonte: A autora (2024).

Figura 9 – Ranqueamento dos modelos com diferentes abordagens de *thresholds* com base na análise post-hoc de Nemenyi.



Fonte: A autora (2024).

vista estatístico, foi observado que foi identificada uma diferença do *recall* médio e que os seguintes grupos são estatisticamente similares: MAX e N99; N99, N97 e BOX; N97, N95, BOX e IQR.

A precisão dos modelos foi de uma forma geral baixa (variando entre $4,09$ a $8,92\%$), com exceção da abordagem MAX que obteve uma precisão média de $56,37\%$. A precisão foi uma métrica fortemente influenciada pela elevada taxa de FP obtida pelas demais métricas. Para MAX, essa métrica chegou a atingir de 100% em uma das repetições realizadas devido ao elevado *threshold* calculado no treinamento, mas é importante salientar que nesta repetição de MAX mencionada a taxa de TP também foi muito baixa, o que também resultou em um *recall* de $4,20\%$, além do elevado desvio padrão provocado por esta abordagem ($24,88\%$) quando

comparado aos demais. Isso mostrou que apesar da abordagem MAX apresentar resultados interessantes no ponto de vista de falsos alarmes, esta abordagem também tem dificuldade de detectar corretamente anomalias. Estatisticamente, a aplicação do pacote Autorank acusou que houve uma diferença estatisticamente significativa entre as seis técnicas de *threshold* aplicadas, mas foi possível agrupá-las nos seguintes grupos: IQR, N95 e N97; N95, N97 e BOX; N97, BOX e N99; N99 e MAX.

O F1-Score é uma métrica muito relevante na avaliação de AD por ser uma média harmônica entre *recall* e precisão. Por isso, é muito interessante maximizar essa métrica, já que ela representa um balanço ideal entre TP e FP. Nos experimentos avaliados sem aplicação de filtro, foi observado uma amplitude de F1-Score entre $7,65\%$, quando a abordagem foi IQR, a $14,72\%$ quando a abordagem foi N99, que é bem inferior a faixa de variação encontrada em outras métricas. Do ponto de vista estatístico, foi verificado que ocorreu diferenças significativas entre os *thresholds*, mas que foi possível agrupar em: IQR, N95 e MAX; N95, MAX, N97 e BOX; N97, BOX e N99.

Por fim, a métrica FPR apresentou, como esperado, seu menor valor com a abordagem MAX ($0,14\%$) e seu maior valor médio ocorreu com a aplicação da abordagem IQR ($31,51\%$) devido aos *threshold* serem respectivamente mais altos e mais baixos em relação às demais abordagens. Esta métrica foi avaliada devido ao impacto de falsos alarmes em AD não supervisionado e é uma das métricas que tipicamente são minimizadas na criação de modelos de AD. Estatisticamente foi verificado que também ocorreu diferença entre abordagens, mas que foi possível agrupar de acordo com semelhança estatística os seguintes grupos: MAX e N99; N99, BOX e N97; BOX, N97 e N95; N95 e IQR.

Dessa forma, foi observada uma alta influência da seleção do *threshold* nas métricas de avaliação nos experimentos. Com a aplicação do pacote Autorank e avaliação do teste post-hoc de Nemenyi foi observada uma semelhança estatística para todas as métricas para [N97, N99 e BOX]. Além disso, foi observado que para precisão, F1-Score e FPR os modelos baseados em [IQR, N95 e N97] também foram estatisticamente semelhantes, enquanto *recall*, precisão e FPR de [N99 e MAX] foram similares. A Tabela 7 apresenta os valores de *threshold* calculados em cada uma das janelas de treinamento durante o treinamento *online*.

Essas semelhanças são justificáveis pelos valores de *threshold* que foram encontrados durante os experimentos, que de fato foram muito próximos, com exceção da abordagem MAX que apresentou *thresholds* muito elevados, apesar de não gerar métricas de *recall*, precisão e F1-Score estatisticamente diferentes da abordagem N99. É importante destacar que além

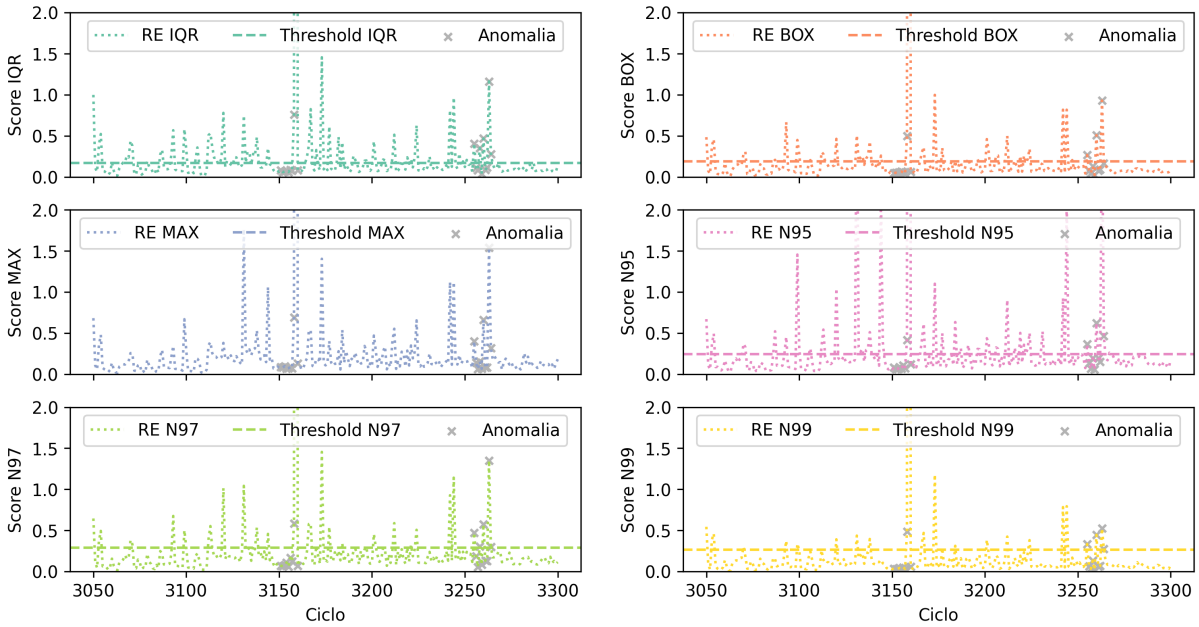
Tabela 7 – *Thresholds* calculados nas janelas de treinamento por tipo de abordagem.

Janela	Threshold calculado no treinamento					
	IQR	BOX	MAX	N95	N97	N99
1	0,099 ± 0,017	0,155 ± 0,030	5,430 ± 4,824	0,118 ± 0,018	0,131 ± 0,029	0,240 ± 0,083
2	0,053 ± 0,013	0,077 ± 0,017	17,944 ± 19,308	0,062 ± 0,012	0,095 ± 0,018	0,107 ± 0,026
3	0,073 ± 0,021	0,101 ± 0,028	101,773 ± 60,188	0,083 ± 0,022	0,128 ± 0,053	0,251 ± 0,113
4	0,061 ± 0,011	0,142 ± 0,188	17,847 ± 16,179	0,064 ± 0,011	0,103 ± 0,090	0,132 ± 0,033
5	0,149 ± 0,072	0,180 ± 0,072	331,973 ± 301,866	0,236 ± 0,152	0,271 ± 0,188	0,351 ± 0,201
6	0,072 ± 0,048	0,152 ± 0,084	109,838 ± 246,148	0,088 ± 0,044	0,120 ± 0,085	0,275 ± 0,149
7	0,082 ± 0,040	0,133 ± 0,053	12,729 ± 15,439	0,081 ± 0,024	0,096 ± 0,031	0,177 ± 0,044
8	0,189 ± 0,076	0,277 ± 0,078	61,180 ± 56,240	0,283 ± 0,143	0,235 ± 0,082	0,378 ± 0,167
9	0,075 ± 0,121	0,098 ± 0,052	21,901 ± 18,645	0,149 ± 0,244	0,168 ± 0,303	0,138 ± 0,019
10	0,072 ± 0,037	0,109 ± 0,104	663,980 ± 1622,547	0,105 ± 0,102	0,208 ± 0,292	0,294 ± 0,395
11	0,095 ± 0,025	0,157 ± 0,038	509,973 ± 721,542	0,088 ± 0,037	0,126 ± 0,055	0,350 ± 0,116
12	0,050 ± 0,011	0,086 ± 0,014	20,700 ± 24,872	0,054 ± 0,011	0,057 ± 0,010	0,092 ± 0,026

Fonte: A autora (2024).

de ser mais susceptível a erros espúrios ao longo do treinamento, a remoção de anomalias da janela de treinamento seguinte neste trabalho foi feito de forma não supervisionada na AD na janela de validação. Dessa forma, os valores de *threshold* $thres^w$ computados pela abordagem MAX são mais afetados caso haja a presença de alguma anomalia em uma janela de treinamento.

Figura 10 – Exemplo de comportamento da AD em uma janela de teste.



Fonte: A autora (2024).

A Figura 10 apresenta o comportamento da AD na janela de teste 4 das repetições que apresentaram melhor métrica de F1-Score em todo o aprendizado *online*. Vale ressaltar que o *threshold* de MAX não está visível na Figura 10 por possuir um valor muito mais elevado

que prejudicaria a visibilidade do *score* de anomalia, além disso, devido a estocasticidade inerente deste modelos, o erro de reconstrução pode variar entre experimentos. Assim como apresentado na avaliação das métricas e pelos valores de *threshold* apresentado na Tabela 7, a abordagem MAX tende a gerar um FP muito baixo devido ao elevado *threshold* calculado, enquanto a abordagem IQR tem um TP superior às demais abordagens por gerar um *threshold* mais baixo. Sendo assim, é necessário escolher a abordagem de *threshold* mais adequada ao contexto de aplicação. Por exemplo, se falsos alarmes possuem um alto custo ou anomalias forem muito escassas, a abordagem MAX pode ser interessante, enquanto a abordagem BOX ou N99 podem ser interessantes caso TP seja crítico. Além disso, é muito importante salientar que independente de qual abordagem foi selecionada para o cálculo do *threshold* adaptativo, o modelo SAE não foi capaz de detectar sequências de anomalias presentes na base de dados avaliada.

Portanto, essa seção apresentou a diferença de comportamento das métricas de avaliação de acordo com a abordagem de *threshold* utilizada para o cálculo adaptativo não supervisionado em uma aplicação de AD de um modelo SAE não supervisionado. Além disso, foi ressaltada a alta influência desta decisão e necessidade de avaliação prévia de criticidade das métricas para seleção de técnica de cálculo de *threshold* que mais condiz com o cenário a ser adotado.

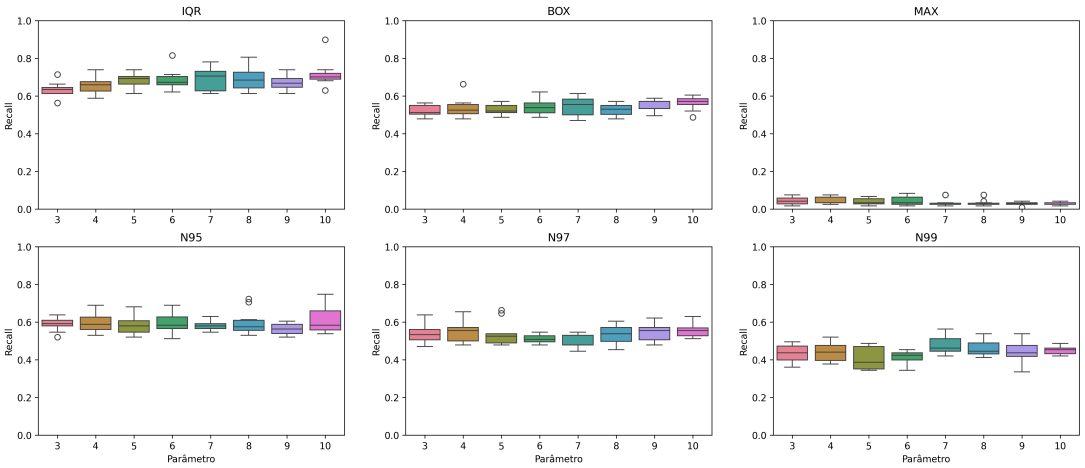
5.2 AVALIAÇÃO DO IMPACTO DA ABORDAGEM PRÉ-CLASSIFICAÇÃO DE DETECÇÃO

Nesta seção, será apresentado o impacto da aplicação dos LPF com a abordagem Pré-Classificação nas métricas avaliadas, bem como a influência do parâmetro de entrada nos resultados obtidos. Além disso, será realizado um breve comparativo entre a detecção com aplicação de LPF e sem a utilização de filtros.

5.2.1 Média Móvel Simples

O primeiro filtro avaliado foi o SMA e para evitar influência da técnica de *threshold* adotada, a avaliação será realizada para cada um dos *thresholds* utilizados. O *recall* é apresentado na Figura 11 de acordo com as janelas de entrada avaliadas (3 a 10) e foi observado que a configuração Pré-Classificação + SMA aumentou os valores médios da métrica para todas as técnicas de *threshold* em relação a abordagem sem filtro, com a exceção de MAX. Isso ocorreu

Figura 11 – Comportamento do *recall* de acordo com o parâmetro de entrada do filtro SMA e abordagem Pré-Classificação.

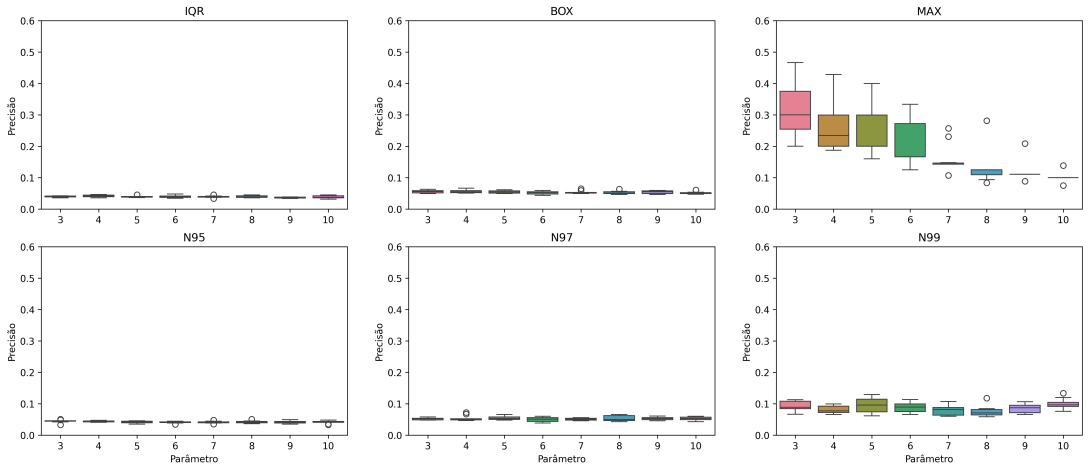


Fonte: A autora (2024).

porque a suavização por SMA manteve o sinal filtrado mais elevado momentaneamente após uma instância com erro de reconstrução mais elevado. Essa influência gerou um aumento de TP dos demais *thresholds*, o que indicou uma melhoria real na detecção, além da possibilidade de detectar sequências de anomalias. O mesmo não ocorreu com MAX porque o aumento do tamanho da janela levou a uma diluição do pico a partir de uma janela de tamanho 7 para valores abaixo do que o alto *threshold* calculado e passou a reduzir o TP. Do ponto de vista estatístico, foi observado que os modelos sem filtro não se diferenciaram dos modelos com aplicação de Pré-Classificação + SMA. Além disso, houve diferença estatística referente aos parâmetros de entrada apenas para N99, enquanto IQR, BOX, MAX, N95 e N97 apresentaram os respectivos p-valores: Friedman de $0,069$, ANOVA de $0,259$, Friedman de $0,319$, Friedman de $0,702$ e ANOVA de $0,093$. Apesar da diferença estatística para N99, não foi identificada uma relação estatística clara de queda de *recall* em relação aos parâmetros de entrada do SMA. Isso indica que foi observada uma melhoria do *recall* quando comparado à detecção sem filtro, mas que a escolha do parâmetro de entrada não apresentou influência clara nesta melhoria.

Para a precisão, confirmou-se que houve uma redução da média da métrica com a configuração Pré-Classificação + SMA quando comparada a abordagem sem filtro, apresentada na Figura 12. Essa mudança ocorreu devido ao aumento maior de FP, quando comparado ao aumento de TP, provocado pela suavização realizada pelo filtro que aumentou o *score* de anomalia após um valor mais elevado de erro de reconstrução pelo número de janelas utilizado.

Figura 12 – Comportamento da precisão de acordo com o parâmetro de entrada do filtro SMA e abordagem Pré-Classificação.

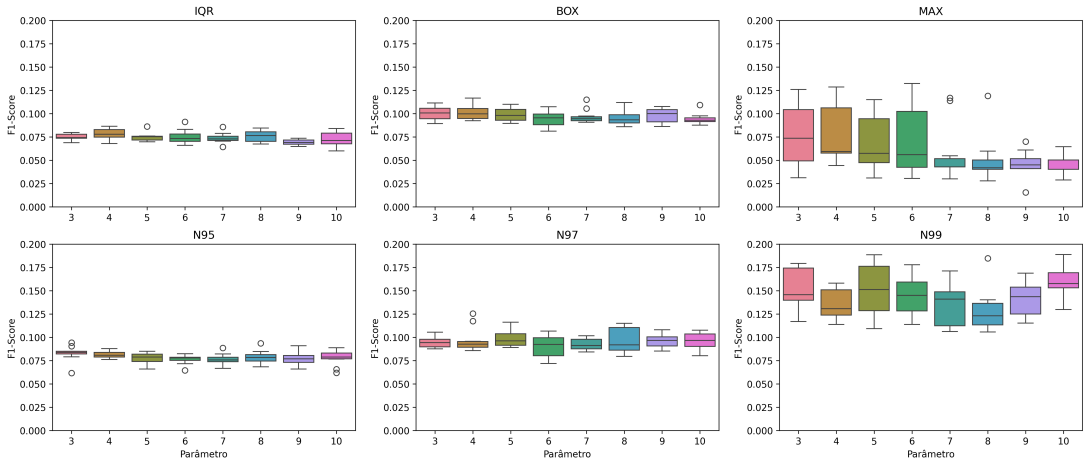


Fonte: A autora (2024).

Porém, do ponto de vista estatístico, a precisão para a abordagem sem filtro foi estatisticamente semelhante a detecção com aplicação de filtro em todas as abordagens de *threshold*. Além disso, só houve preferência estatística relacionada a escolha do parâmetro de entrada do filtro menores para a abordagem de *threshold* MAX, já que a configuração Pré-Classificação + SMA com maiores janelas gerou um aumento de FP diretamente proporcional ao aumento da janela para esta abordagem e não modificou substancialmente o TP. Também foi observado que não houve diferenças referente a escolha da entrada do filtro do ponto de vista estatístico para: IQR, p-valor ANOVA de $0,084$; BOX, p-valor Friedman de $0,291$; N95, p-valor ANOVA de $0,255$; e N97, p-valor Friedman de $0,768$. O comportamento da precisão foi mais irregular do que o comportamento do *recall*, mesmo que estatisticamente o parâmetro de entrada influencie a precisão para N99.

Do ponto de vista de F1-Score apresentado na Figura 13, também foi verificado que, estatisticamente, apenas o modelo com a configuração Pré-Classificação + SMA com técnica BOX foi estatisticamente inferior à abordagem sem filtro. Para as demais técnicas de *threshold*, a abordagem sem filtro apresentou valores de F1-Score estatisticamente semelhantes aos modelos de detecção Pré-Classificação + SMA, com um aumento da métrica média apenas para N99 para as janelas 3, 5 e 10. Dessa forma, mesmo com as melhorias no *recall* e redução de precisão, não foi detectado impacto estatisticamente significativo na métrica F1-Score. Também foi observado que a escolha do parâmetro de entrada dos filtros não exerceu diferença estatisticamente relevante para: IQR, p-valor ANOVA de $0,108$; BOX, p-valor Friedman de

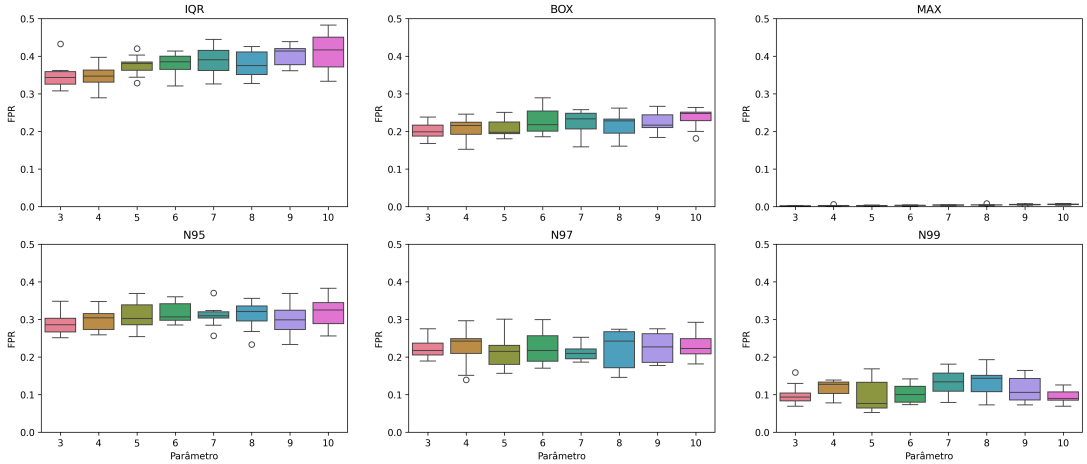
Figura 13 – Comportamento do F1-Score de acordo com o parâmetro de entrada do filtro SMA e abordagem Pré-Classificação.



Fonte: A autora (2024).

$0,464$; MAX, p-valor Friedman de $0,055$; N95, p-valor ANOVA de $0,257$;e N97, p-valor Friedman de $0,652$. Além disso, o padrão do F1-Score não apresentou relação clara com a escolha de parâmetro para N99, apesar de apresentarem diferença estatisticamente significativa entre alguns grupos de forma similar a precisão. Esse fator mostrou que o impacto da precisão foi superior ao do *recall* no cálculo do F1-Score.

Figura 14 – Comportamento do FPR de acordo com o parâmetro de entrada do filtro SMA e abordagem Pré-Classificação.

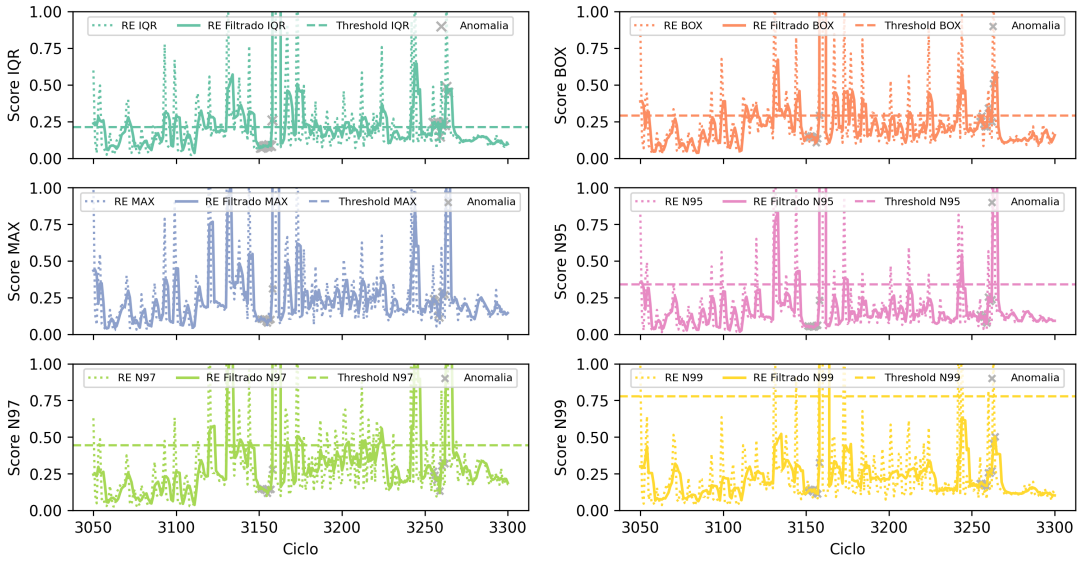


Fonte: A autora (2024).

A Figura 14 apresenta o comportamento do FPR de acordo com os parâmetros de entrada avaliados para a configuração Pré-Classificação + SMA. Foi verificado estatisticamente que a aplicação do filtro reduziu o FPR apenas para o *threshold* IQR quando comparada a abordagem

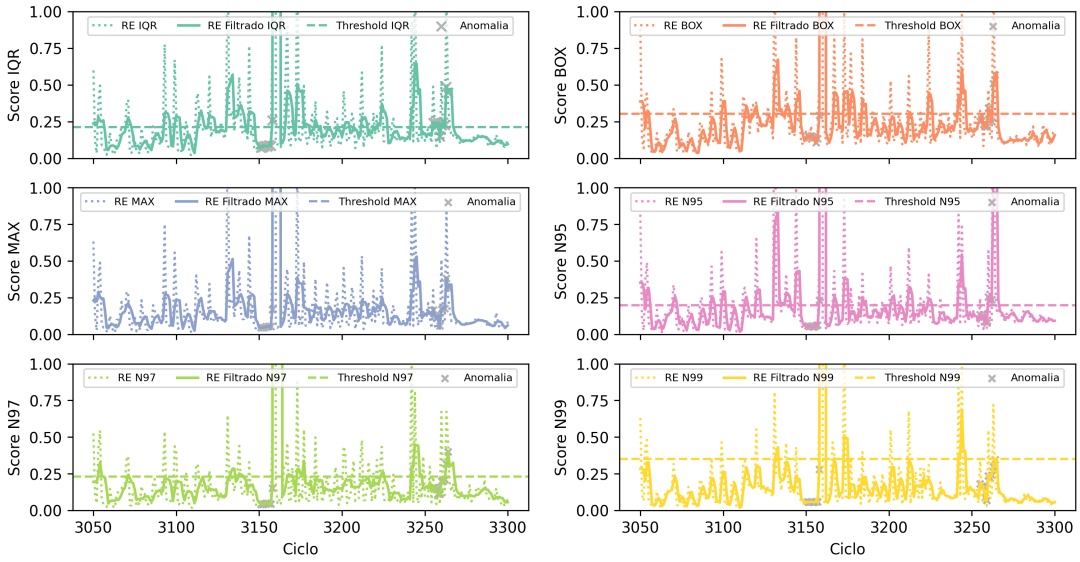
sem filtro, aumentou o FPR significativamente para N95, enquanto não gerou diferença para as demais abordagens. Do ponto de vista de parâmetro de entrada, não houve diferença estatística para: BOX, p-valor ANOVA de $0,196$; N95, p-valor ANOVA de $0,394$; e N97, p-valor ANOVA de $0,976$. Além disso, IQR e MAX apresentaram um aumento claro de FPR com o aumento do parâmetro de entrada. Essa tendência ocorreu devido ao aumento de FP com aumento da janela devido ao comportamento do filtro. Isso mostra que com a configuração Pré-Classificação + SMA, não houve uma redução clara de FPR, já que um pico de erro de reconstrução passou a ser propagado por mais instâncias ao invés de serem minimizados conforme ilustrado na Figura 15. Os modelos que foram capazes de minimizar o FPR utilizaram janelas de tamanho n = 4; 3; 3; 3; 8; 10 para os *thresholds* IQR, BOX, MAX, N95, N97 e N99 respectivamente, enquanto os modelos de detecção que maximizaram o F1-Score utilizaram as janelas n = 4; 3; 4; 3; 5; 3. Um fator interessante que é possível observar nas Figuras 15 e 16 é que a aplicação de Pré-Classificação + SMA possibilitou a detecção de sequências de anomalias, enquanto a aplicação sem filtro detectou apenas anomalias pontuais. Vale ressaltar que assim como na Figura 10, os *thresholds* de MAX não estão apresentados devido ao seu valor extremamente elevado que prejudicaria a visualização da imagem.

Figura 15 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro SMA com abordagem Pré-Classificação com minimização de FPR.



Fonte: A autora (2024).

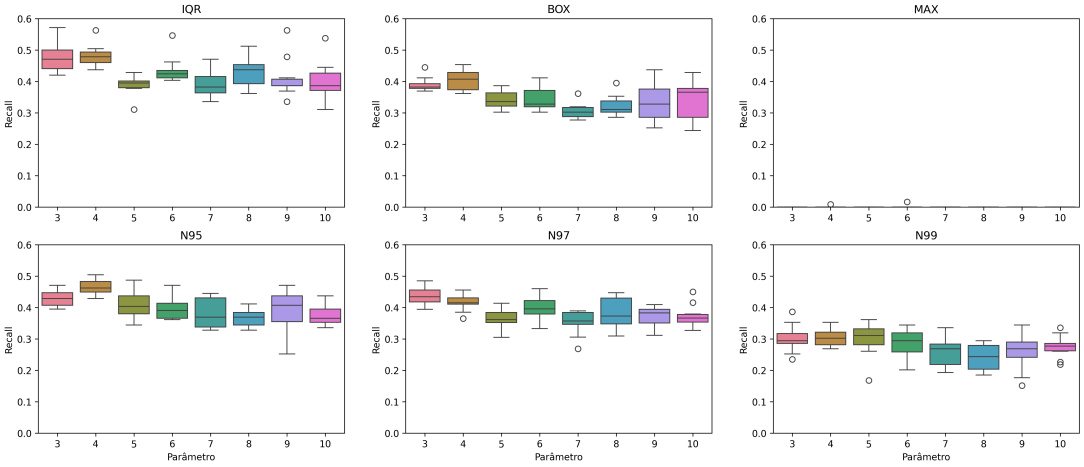
Figura 16 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro SMA com abordagem Pré-Classificação com maximização de F1-Score.



Fonte: A autora (2024).

5.2.2 Mediana Móvel Simples

Figura 17 – Comportamento do *recall* de acordo com o parâmetro de entrada do filtro SMM e abordagem Pré-Classificação.

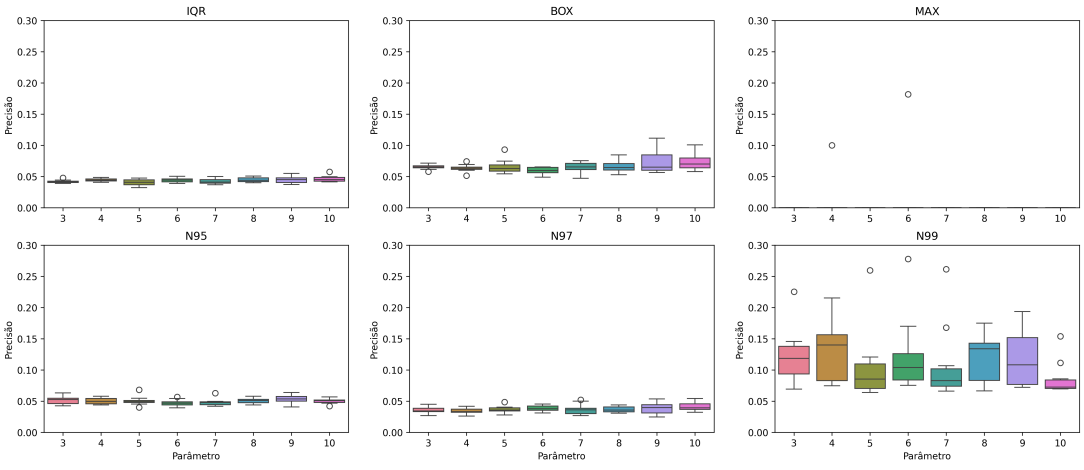


Fonte: A autora (2024).

Para o filtro SMM considerando a mesma abordagem de aplicação de filtro (Pré-Classificação + SMM), foi observado um impacto diferente nas métricas quando comparado a configuração Pré-Classificação + SMA como será descrito a seguir. A Figura 17 apresenta o comportamento do *recall* para os diferentes parâmetros de entrada testados para o filtro SMM. Para esta configuração, foi observada uma redução da média *recall* quando comparado a abordagem sem

filtro e a Pré-Classificação + SMA, sendo a técnica MAX a mais impactada. Essa redução está relacionada a redução de TP, já que o filtro SMM se baseia na mediana que é menos impactado por erros de reconstrução elevados isolados, ao contrário do SMA que pode ser muito influenciado por poucos ou até um único valor elevado. Estatisticamente foi constatada que a aplicação do filtro gerou redução de *recall* para as técnicas de *threshold* MAX, N97 e N99, enquanto para os demais *thresholds* foi observada uma semelhança estatística com a aplicação de filtro com janelas 3 e 4. Também foi observada uma influência dos parâmetros de entrada para todos os *thresholds*, com exceção de MAX, p-valor Friedman de $0,540$, avaliados para a detecção com uma tendência estatisticamente comprovada de redução de *recall* com o aumento da janela de entrada do filtro para N97. Isso ocorreu devido a redução de TP com o aumento do parâmetro de entrada do SMM, já que a mediana de uma janela maior tende a resultar valores menores se houver uma predominância de instâncias com erro de reconstrução inferiores independente de quão alto seja o valor de um determinado pico encontrado. A respeito de MAX, foi observada um *recall* praticamente nulo para esta configuração, já que a aplicação de SMM gerou um TP nulo em quase todas as repetições independente do tamanho da janela devido aos altos *thresholds* calculados por esta técnica.

Figura 18 – Comportamento da precisão de acordo com o parâmetro de entrada do filtro SMM e abordagem Pré-Classificação.

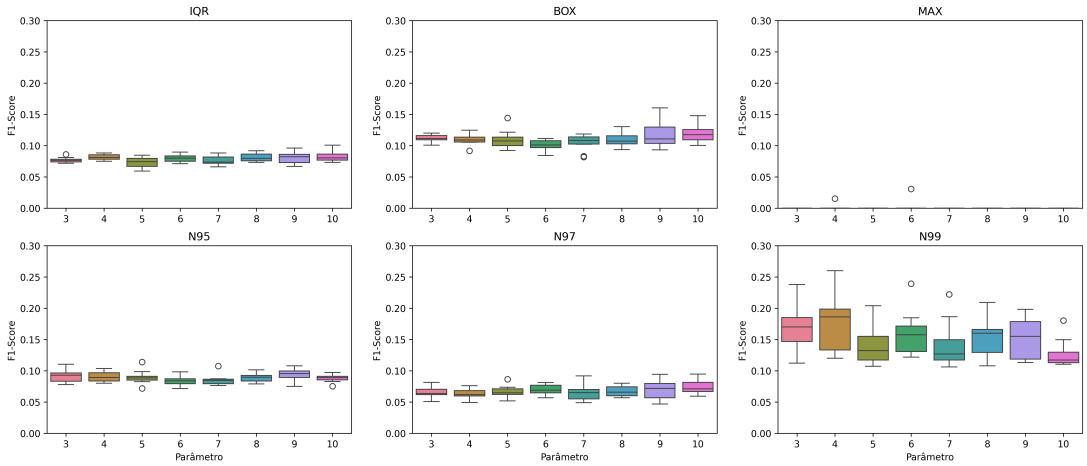


Fonte: A autora (2024).

Do ponto de vista da precisão, apresentada na Figura 18, foi observado um aumento dos valores médios da métrica, com exceção do *threshold* MAX e N97 que foram deterioradas em relação à detecção sem filtro de forma estatisticamente significativa. Para MAX foi observada uma redução de TP, como descrita anteriormente, e para N97 foi observado um aumento de

FP, ao mesmo tempo que foi observada uma queda de TP. Também foi observado que, do ponto de vista estatístico, a escolha do parâmetro de entrada não gerou influência significativa em nenhum dos *thresholds* apresentando as seguintes estatísticas: IQR, p-valor ANOVA de $0,056$; BOX, p-valor Friedman de $0,165$; MAX, p-valor Friedman de $0,540$; N95, p-valor ANOVA de $0,307$; N97, p-valor ANOVA de $0,405$; e N99, p-valor Friedman de $0,126$. Como não há influência estatística comprovada, não é possível inferir que haja uma relação entre parâmetro de entrada e comportamento da precisão.

Figura 19 – Comportamento do F1-Score de acordo com o parâmetro de entrada do filtro SMM e abordagem Pré-Classificação.

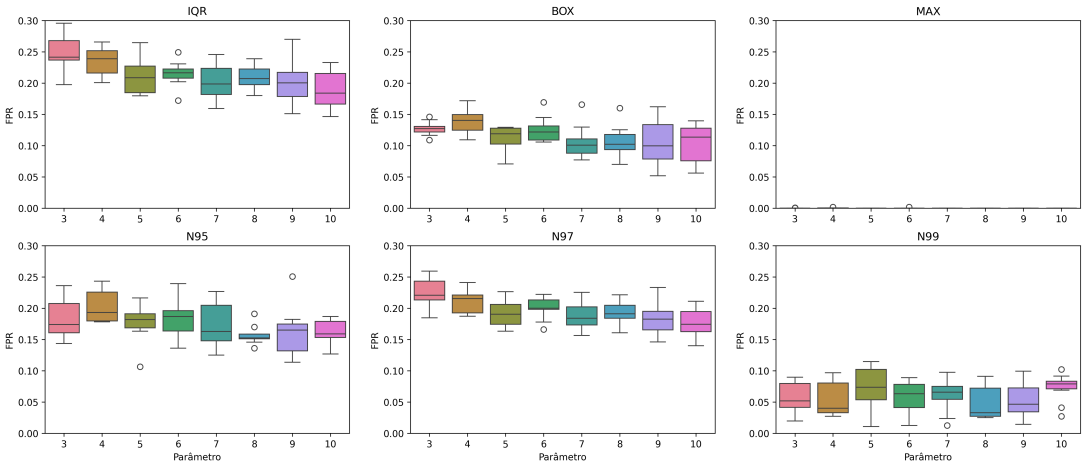


Fonte: A autora (2024).

Para o F1-Score, apresentado na Figura 19, foi observado que apenas MAX e N97 foram estatisticamente inferiores quando comparado a respectiva abordagem sem aplicação de filtro devido ao comportamento do *recall* e da precisão. Os demais *thresholds* se comportaram de forma estatisticamente similar a detecção sem aplicação de filtro apesar do aumento da média de F1-Score observada para IQR e N99. Além disso, foi observado que a escolha do parâmetro de entrada não gerou influência em F1-Score para: MAX, p-valor Friedman de $0,540$; N95, p-valor Friedman de $0,376$; e N97, p-valor ANOVA de $0,511$. Vale reforçar que toda a avaliação estatística desse trabalho considera uma confiabilidade de 95%. Apesar da seleção do parâmetro de entrada influenciar F1-Score para IQR, BOX e N99, nenhum *threshold* apresentou uma relação estatisticamente evidente entre a métrica e a janela de entrada.

A Figura 20 apresenta o comportamento da métrica FPR na configuração de detecção Pré-Classificação $+$ SMM. Estatisticamente foi confirmado que houve redução do FPR entre a aplicação de filtro e sem aplicação de filtro na detecção com os *thresholds* IQR, BOX, N95 e

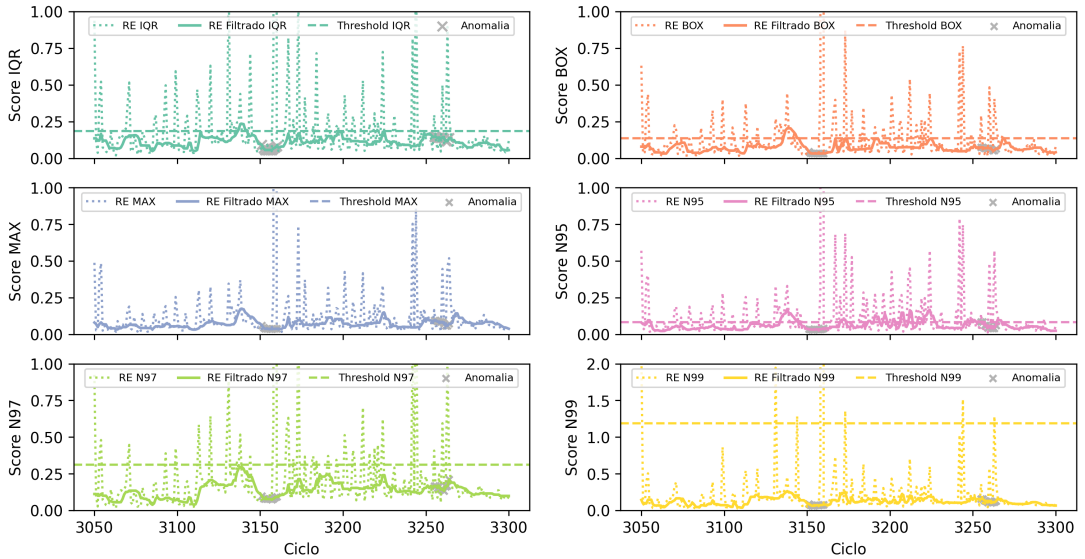
Figura 20 – Comportamento do FPR de acordo com o parâmetro de entrada do filtro SMM e abordagem Pré-Classificação.



Fonte: A autora (2024).

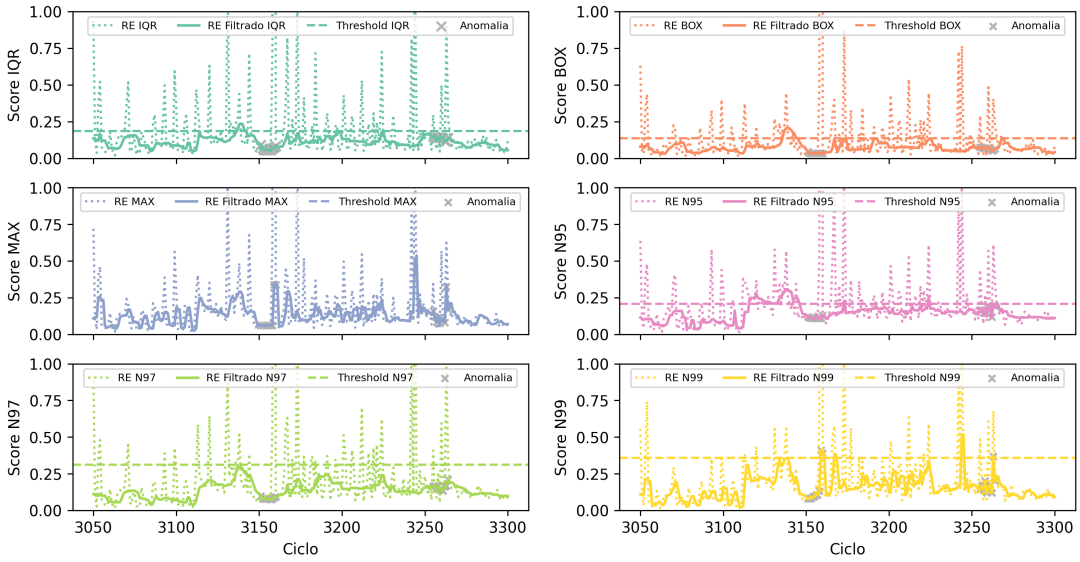
N99. Para MAX e N97, houve semelhança entre a aplicação de pequenas janelas de SMM para os modelos sem filtro. Além disso, foi observada que a seleção do tamanho da janela de entrada do SMM gerou diferenças estatisticamente significativas para todos os *thresholds* estudados, com exceção de MAX (p-valor Friedman de $0,013$, mas não atestada estatisticamente pelo teste de Nemenyi post-hoc) e N99 (p-valor ANOVA de $0,326$). Para IQR e N95 foi observado que uma janela menor tendeu a provocar uma maior redução de FPR, enquanto não foi observada uma tendência explícita para BOX e N97.

Figura 21 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro SMM com abordagem Pré-Classificação com minimização de FPR.



Fonte: A autora (2024).

Figura 22 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro SMM com abordagem Pré-Classificação com maximização de F1-Score.



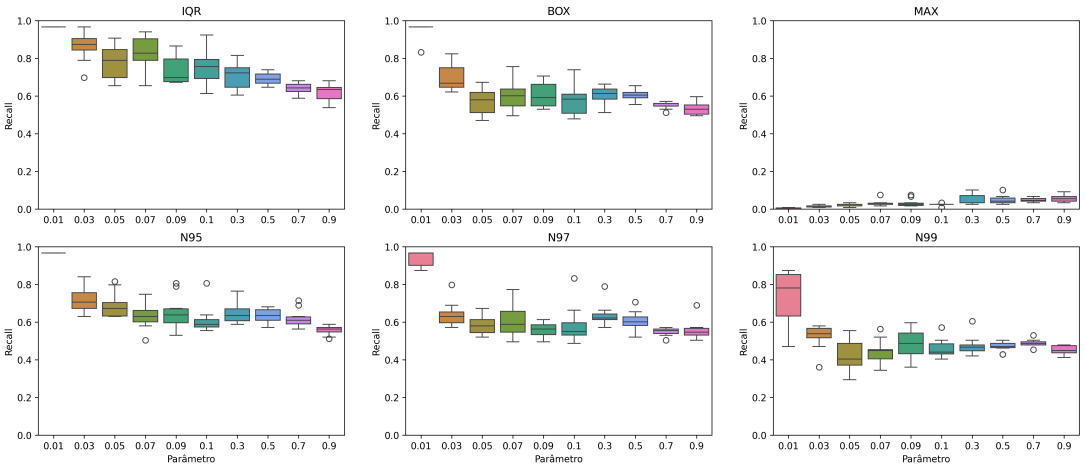
Fonte: A autora (2024).

Nas Figuras 21 e 22 é possível observar o comportamento do SMM sobre o erro de reconstrução para cada um dos *thresholds* considerando a minimização de FPR e maximização de F1-Score. Para a minimização de FPR com o filtro SMM, foram utilizadas as janelas de entrada n = 10; 10; 10; 8; 10, 8 enquanto para a maximização de F1-Score foram utilizadas n = 10; 1; 4; 9; 10; 4 para os *thresholds* IQR, BOX, MAX, N95, N97 e N99 respectivamente. Como foi discutido anteriormente, o SMM foi menos influenciado do que o SMA por picos de erro de reconstrução e suavizou de forma mais rápida, o que provocou uma redução de FPR. Essa mesma suavização também reduziu a detecção real do modelo, mas a maximização de F1-Score ocorreu em cenários com aumento na precisão ao invés de aumento do *recall*.

5.2.3 Média Móvel Exponencial

O último filtro avaliado para a abordagem Pré-Classificação foi o filtro EMA (Pré-Classificação + EMA) que apresentou uma variação de métricas muito grande a depender da seleção do parâmetro de entrada α selecionado. O *recall*, apresentado na Figura 23, apresentou um comportamento estatisticamente dependente do parâmetro de entrada para todos os *thresholds* avaliados. Foi observado que os modelos sem filtro foram estatisticamente semelhantes às abordagens com parâmetro de entrada mais elevado que pode ser explicado pela rápida suavização apresentada. Além disso, foi observado um aumento do *recall* com a redução do

Figura 23 – Comportamento do *recall* de acordo com o parâmetro de entrada do filtro EMA e abordagem Pré-Classificação.



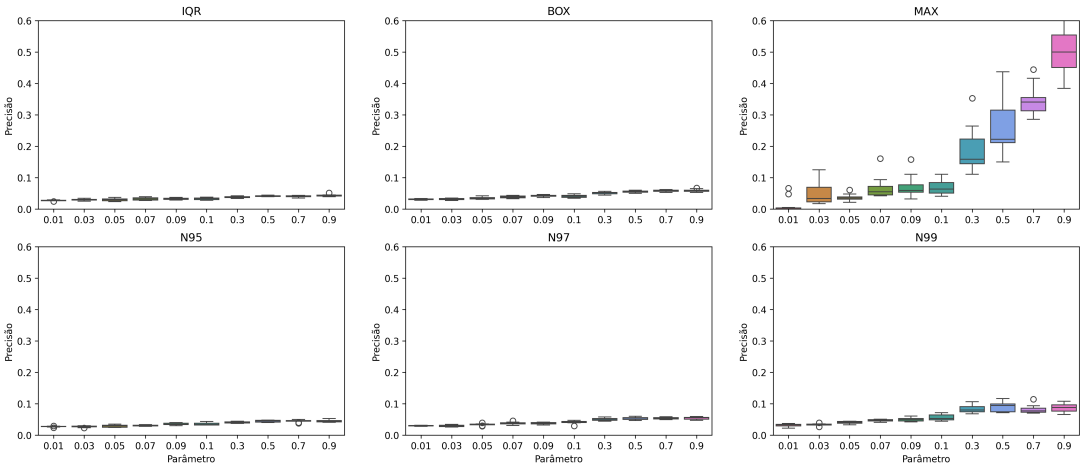
Fonte: A autora (2024).

parâmetro de entrada para IQR, BOX, N95, N97 e N99, enquanto o comportamento de MAX foi oposto. O aumento do *recall* foi dado pelo aumento de TP quando a suavização se deu de forma mais lenta já que o erro de reconstrução filtrado foi impactado por valores anteriores durante um número maior de instâncias que provocou um maior atraso e um consequente aumento de FP. O mesmo não ocorreu com MAX devido ao elevado *threshold* computado que mostrou melhores métricas de *recall* com α mais elevado devido a resposta mais imediata do filtro.

A Figura 24 apresenta o comportamento da precisão dos modelos que utilizaram a configuração Pré-Classificação + EMA e foi observada uma redução de precisão a depender do parâmetro de entrada selecionado. Quando comparado ao modelo sem filtro não foi identificada diferença estatisticamente significativa para aplicação de filtro EMA com parâmetros mais elevados e é interessante salientar que foi identificada uma tendência de aumento de precisão com o aumento do parâmetro de entrada para todos os *thresholds* avaliados. Essa tendência pode ser explicada pela redução significativa de FP com o aumento do α já que a suavização passa a ser mais ágil e impacta valores seguintes de forma mais rápida, o que torna menos propícia a falsos alarmes.

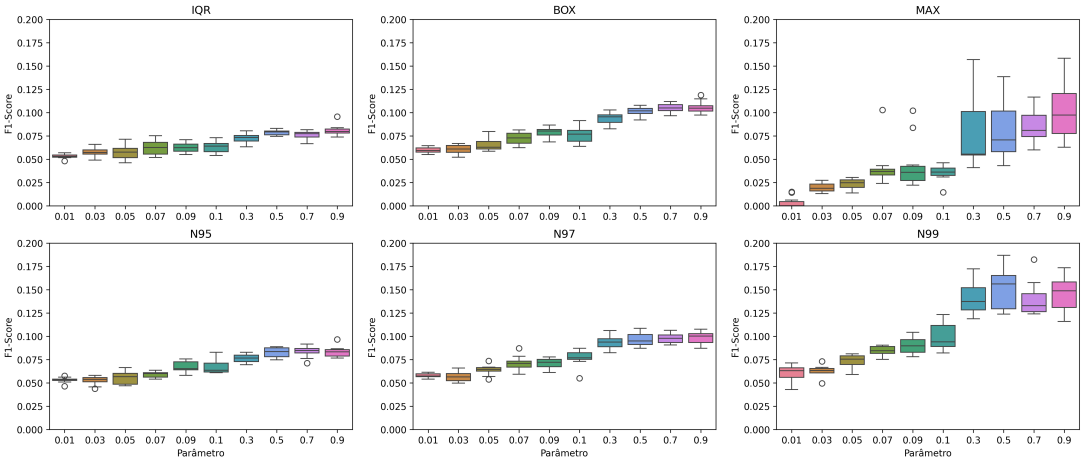
Já o F1-Score, apresentado na Figura 25, foi estatisticamente similar a abordagem sem-filtro para valores altos do parâmetro de entrada do EMA. Porém, para pequenos valores de α foi observada uma redução de F1-Score estatisticamente significativa. Isso se deve primariamente à redução de precisão observada nestas condições que impactou mais do que a melhoria

Figura 24 – Comportamento da precisão de acordo com o parâmetro de entrada do filtro EMA e abordagem Pré-Classificação.



Fonte: A autora (2024).

Figura 25 – Comportamento do F1-Score de acordo com o parâmetro de entrada do filtro EMA e abordagem Pré-Classificação.

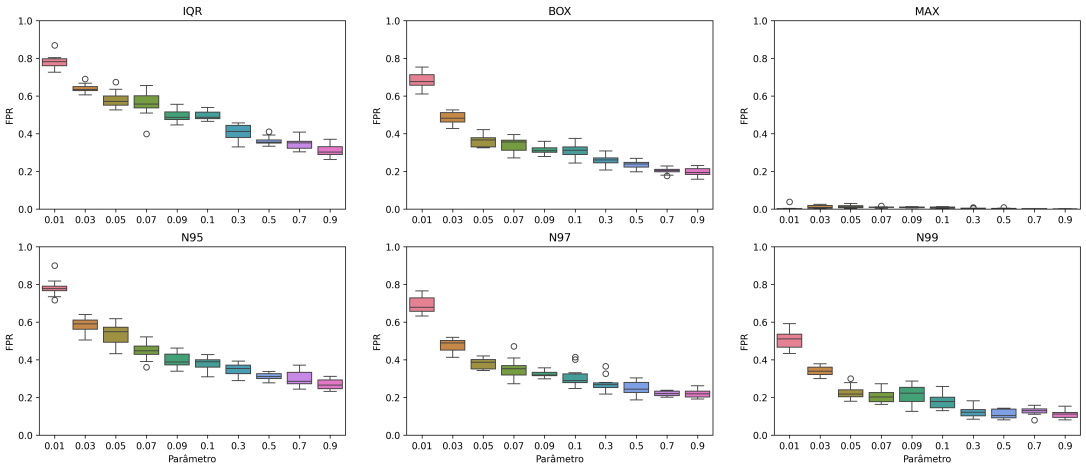


Fonte: A autora (2024).

do *recall*. Com isso, foi observada uma tendência de aumento de F1-Score com o aumento do parâmetro de entrada para todos os *thresholds* avaliados, assim como observado para a precisão.

A Figura 26 apresenta o comportamento da métrica FPR que foi estatisticamente similar com valores maiores de parâmetros de entrada ao modelo sem filtro. Porém, assim como para as outras métricas, foi observado um aumento estatisticamente significativo do FPR para pequenos valores de α devido ao aumento de FP provocado pela suavização mais lenta que pode ser observado na Figura 27 para todas as abordagens de *threshold* com exceção de MAX. Isso ocorreu porque a detecção com a configuração Pré-Classificação + EMA não gerou

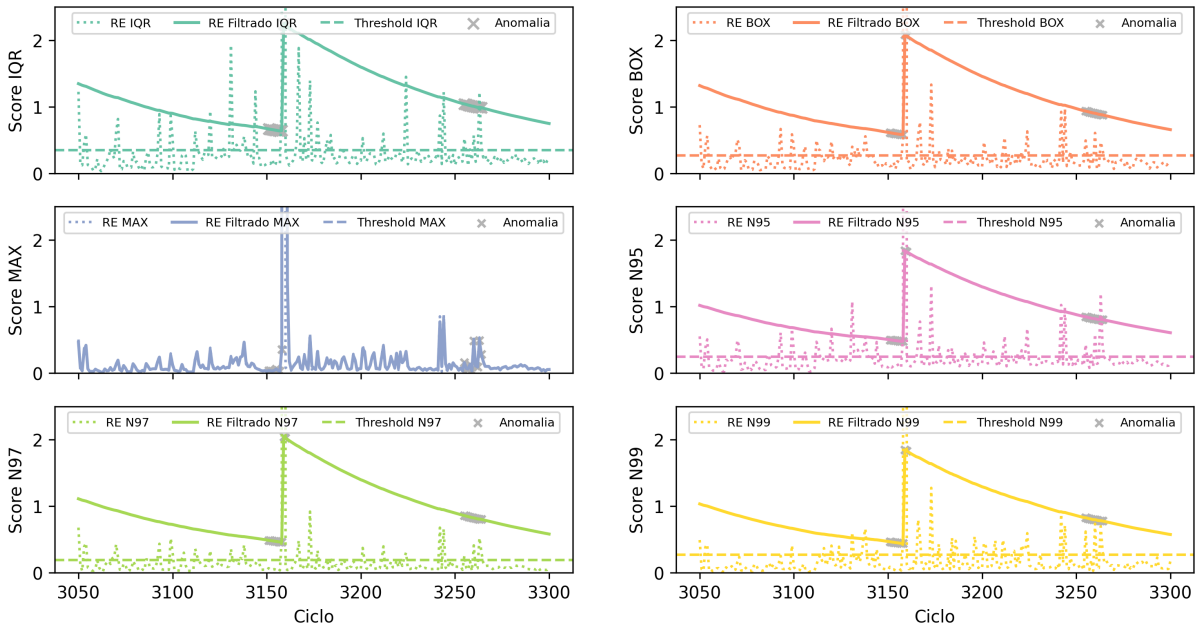
Figura 26 – Comportamento do FPR de acordo com o parâmetro de entrada do filtro EMA e abordagem Pré-Classificação.



Fonte: A autora (2024).

efeitos expressivos sobre o FPR para a técnica MAX devido a seu baixo FP atrelado ao alto *threshold* do modelo sem filtro. Para a maximização do *recall* foram utilizados os parâmetros de suavização de entrada de EMA α = 0,01; 0,01; 0,9; 0,01; 0,01; 0,01 para os *thresholds* IQR, BOX, MAX, N95, N97 e N99 respectivamente.

Figura 27 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro EMA com abordagem Pré-Classificação com maximização do *Recall*.

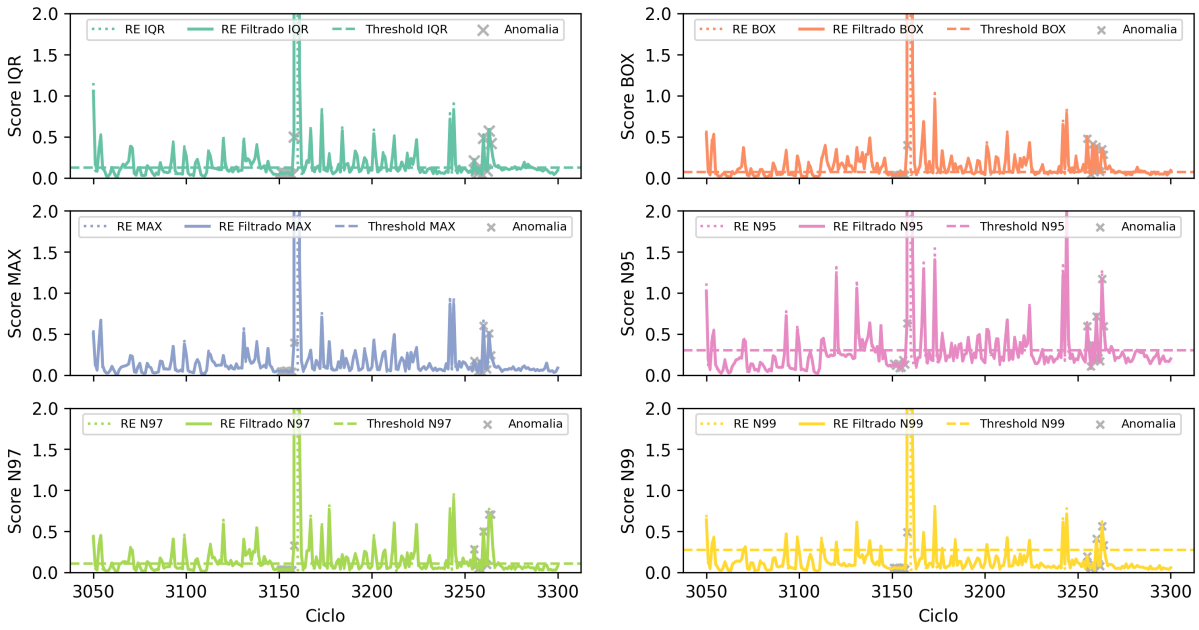


Fonte: A autora (2024).

Já a Figura 28 apresenta o comportamento da detecção com a minimização de FPR obtida

com parâmetros de suavização $\alpha = 0{,}9$ para todos os *thresholds*. Também foi possível observar que o filtro EMA, quando comparado a Pré-Classificação + SMA e Pré-Classificação + SMM, possuiu picos mais altos e foi mais influenciado por erros de reconstrução mais elevados devido a natureza do filtro. Assim a suavização mais lenta levou a uma propagação mais longa de picos encontrados no erro de reconstrução quando o α foi mais baixo.

Figura 28 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro EMA com abordagem Pré-Classificação com minimização de FPR.



Fonte: A autora (2024).

5.3 AVALIAÇÃO DO IMPACTO DA ABORDAGEM PÓS-CLASSIFICAÇÃO DE DETECÇÃO

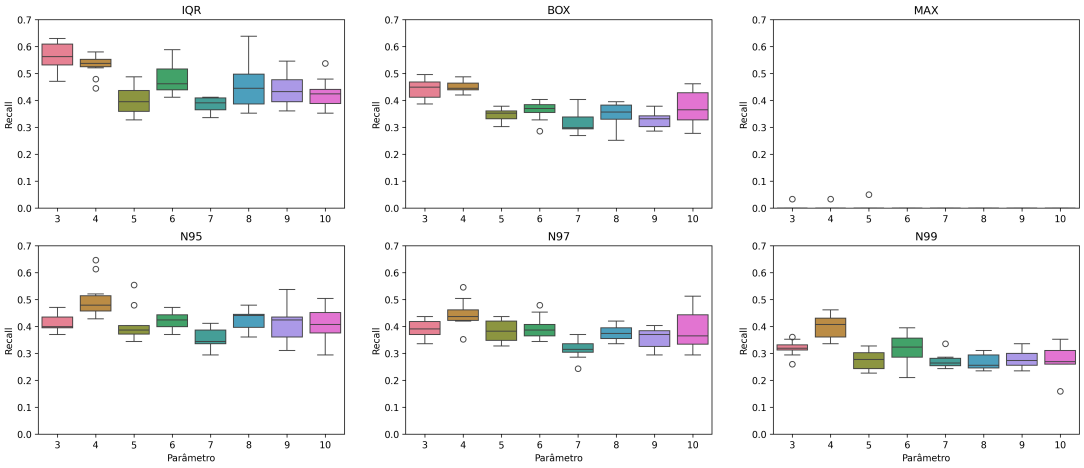
Nesta seção será apresentado o impacto da aplicação dos LPF com a abordagem Pós-Classificação nas métricas avaliadas, bem como a influência do parâmetro de entrada nos resultados obtidos. Além disso, será realizado um breve comparativo entre a detecção com aplicação de LPF nesta abordagem e sem a utilização de filtros.

5.3.1 Média Móvel Simples

Ao contrário da abordagem Pré-Classificação + SMA, o *recall*, apresentado na Figura 29, apresentou uma redução, porém só foi estatisticamente relevante para os *thresholds* MAX e N99 quando comparado a detecção sem filtro. Também foi observada estatisticamente uma

influência da escolha dos parâmetros de entrada para todos os *thresholds*, com exceção de MAX (p-valor Friedman de $0,660$) que apresentou métrica nula em praticamente todos as execuções devido a baixa detecção na primeira classificação relacionada ao elevado *threshold* que foi reduzida com a aplicação do SMA para a segunda classificação. É interessante comentar que apesar da relevância da escolha do parâmetro só foi verificada uma tendência clara de aumento do *recall* com redução da janela do filtro n para os *thresholds* IQR e N99. Apesar disso, foi verificado que para IQR, BOX, N95, N97 e N99 as janelas de entrada 3 e 4 apresentaram um melhor *recall* por aumentar a influência de instâncias com passado mais recente quando comparada a janelas maiores.

Figura 29 – Comportamento do *recall* de acordo com o parâmetro de entrada do filtro SMA e abordagem Pós-Classificação.

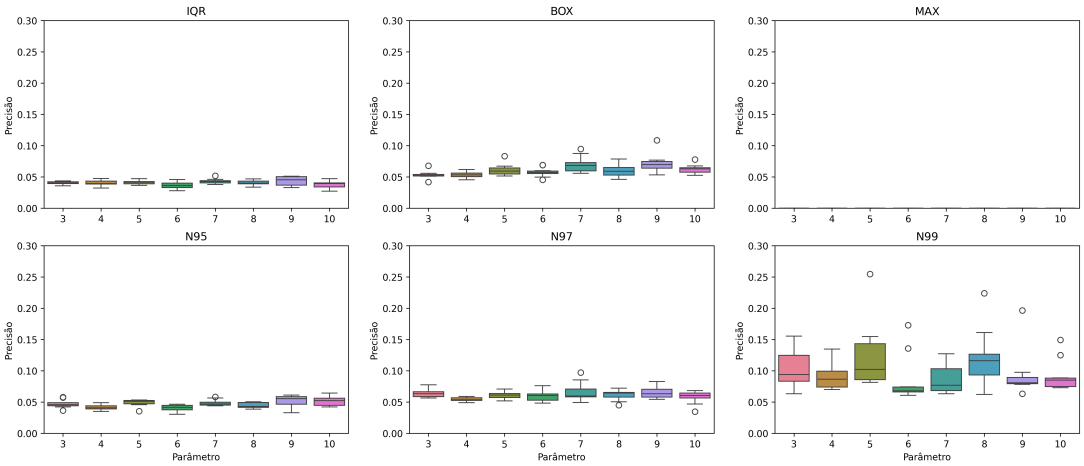


Fonte: A autora (2024).

A Figura 30 apresenta o comportamento da precisão nos experimentos realizados com a configuração Pós-Classificação + SMA. Foi possível constatar que, estatisticamente, a detecção sem filtro apresentou precisão superior à detecção por meio da técnica MAX de *threshold* com Pós-Classificação + SMA devido ao comportamento de TP observado para esta configuração de detecção para MAX. Ou seja, apesar da redução de FP verificada, a mudança na precisão não foi estatisticamente significativa para as outras abordagens de *threshold*. Além disso, verificou-se que a seleção da janela influenciou significativamente a precisão para BOX e N95, mas não foi verificada uma relação direta entre janela do filtro e comportamento da métrica. Para IQR, MAX, N97 e N99 não foi identificada influência na escolha do parâmetro de entrada do filtro, com p-valor Friedman de respectivamente $0,217$, $0,660$, $0,660$ e $0,127$.

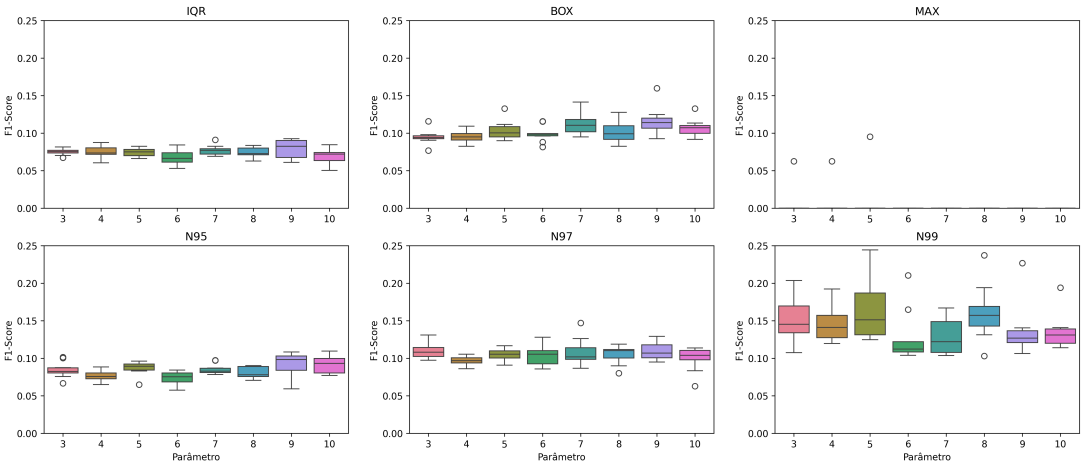
O F1-Score, apresentado na Figura 31, apresentou um aumento quando comparado à

Figura 30 – Comportamento da precisão de acordo com o parâmetro de entrada do filtro SMA e abordagem Pós-Classificação.



Fonte: A autora (2024).

Figura 31 – Comportamento do F1-Score de acordo com o parâmetro de entrada do filtro SMA e abordagem Pós-Classificação.

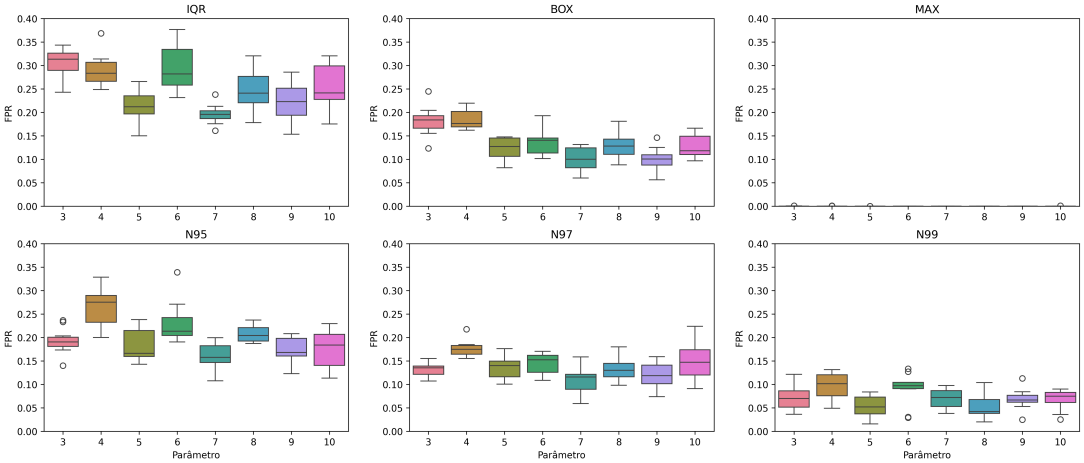


Fonte: A autora (2024).

detecção sem filtro devido ao aumento da precisão que mais impactou esta métrica, porém essa mudança só foi estatisticamente significativa para a técnica de *threshold* MAX que apresentou uma deterioração. Para esta métrica, não foi observada influência estatisticamente significativa para: IQR, p-valor Friedman de $0,369$; N97, p-valor ANOVA de $0,168$; e, N99, p-valor Friedman de $0,058$. Do ponto de vista de impacto do parâmetro de entrada, foi verificado que esta escolha não influenciou o F1-Score de IQR, MAX, N97 e N99 com os respectivos p-valores: Friedman de $0,275$; Friedman de $0,660$; ANOVA de $0,230$; Friedman de $0,039$ que apesar de ser inferior a $0,05$ não acusou diferenças no teste post-hoc de Nemenyi. Esse comportamento também foi diretamente ligada a precisão destas configurações. Apesar disto, foi observada

uma tendência de aumento de F1-Score com o aumento do tamanho da janela do filtro SMA para BOX.

Figura 32 – Comportamento do FPR de acordo com o parâmetro de entrada do filtro SMA e abordagem Pós-Classificação.



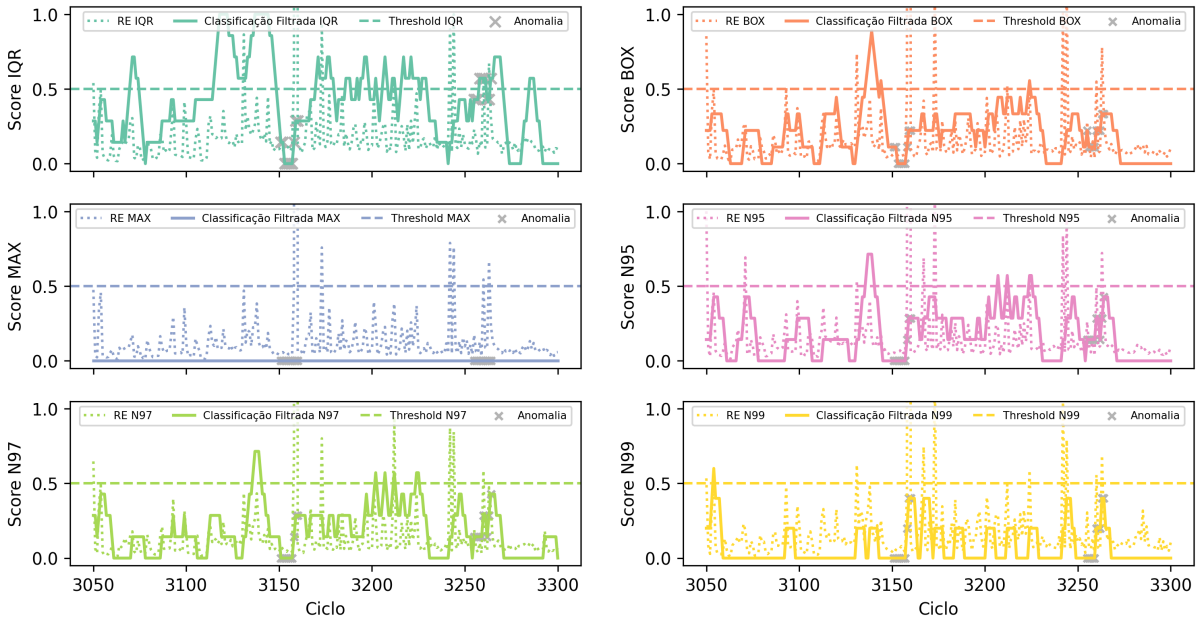
Fonte: A autora (2024).

A aplicação Pós-Classificação + SMA gerou reduções de FP e FPR e está apresentada na Figura 32. Porém, essa redução só foi considerada estatisticamente significativa, quando comparado ao modelo sem filtro, para a detecção que utilizou o *threshold* N97. Para os demais *thresholds* adotados, a abordagem sem filtro foi estatisticamente similar a utilização de filtros SMA com parâmetros de entrada pequenos. Além disso, foi observada uma influência na seleção de entrada para todos os *thresholds* estudados, com exceção de MAX (p-valor Friedman de $0,161$), porém só foi identificada uma tendência de redução de FPR com o aumento da janela para a abordagem BOX.

A Figura 33 apresenta o comportamento da classificação filtrada quando foram selecionadas as configurações de cada *threshold* que minimizaram a métrica FPR. Para a minimização do FPR foram selecionados os parâmetros n $=$ 7; 9; 6; 7; 7; 5 para os *thresholds* IQR, BOX, MAX, N95, N97 e N99 respectivamente de acordo com a relevância estatística e foi possível observar que o filtro reagiu de forma rápida, porém apresentou dificuldades em detectar as anomalias reais (redução de TP). É interessante salientar a preferência geral por janelas maiores na redução de FPR, devido a maior suavização da média por considerar um passado menos recente.

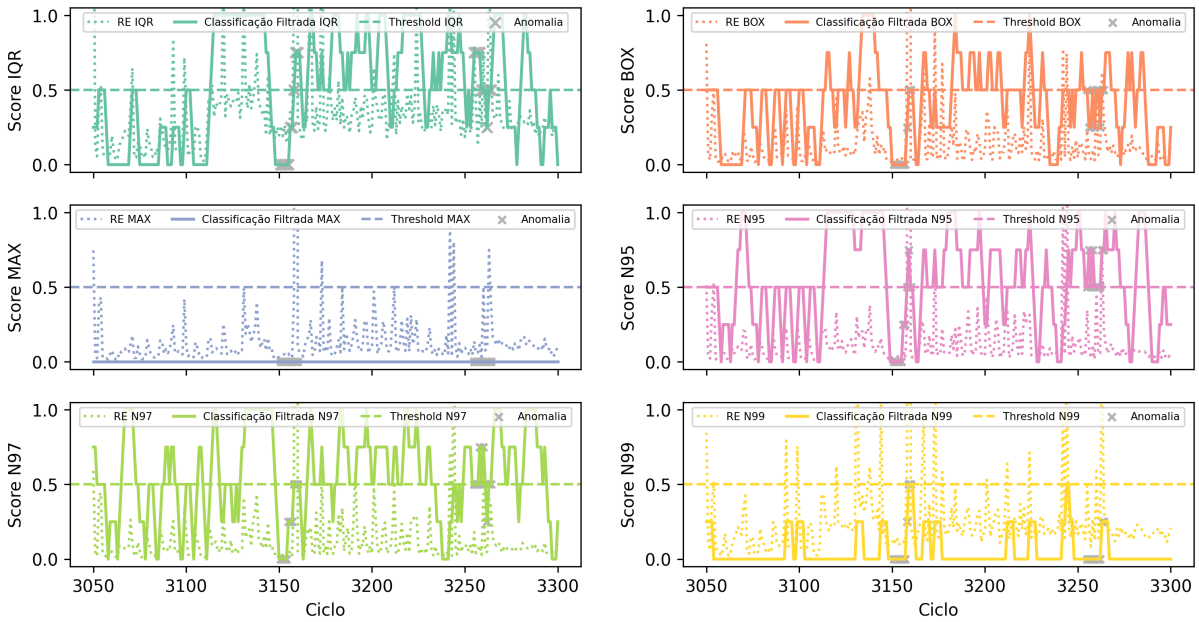
Já a Figura 34 apresenta o comportamento da detecção quando foram selecionadas as configurações que maximizaram o *recall*. Nesse cenário foram selecionados os parâmetros n

Figura 33 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro SMA com abordagem Pós-Classificação com minimização de FPR.



Fonte: A autora (2024).

Figura 34 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro SMA com abordagem Pós-Classificação com maximização do *recall*.

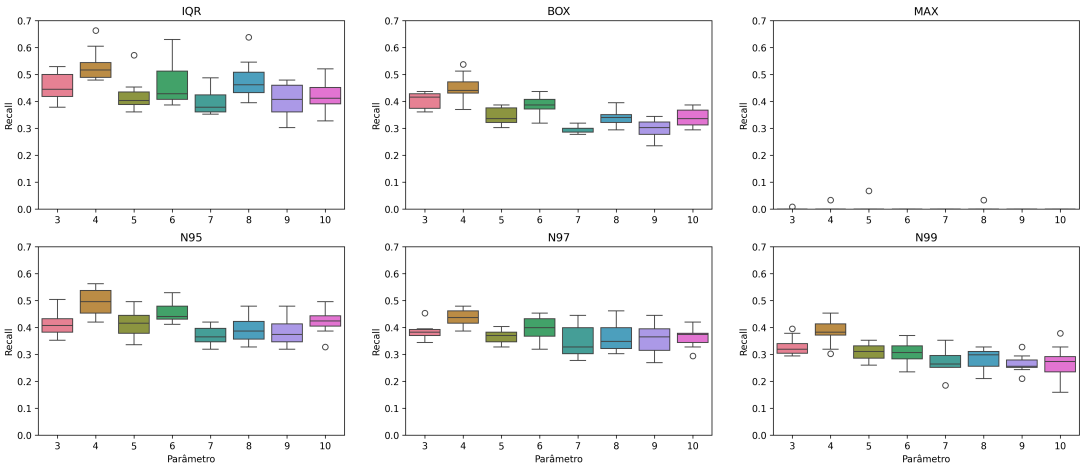


Fonte: A autora (2024).

$= 3; 4; 5; 4; 4; 4$ para os *thresholds* IQR, BOX, MAX, N95, N97 e N99 respectivamente de acordo com a relevância estatística e foi mais ruidoso devido ao tamanho de janela mais baixo. Além disso, a maximização do *recall* também provocou um aumento de FP. É importante salientar que a utilização da configuração Pós-Classificação + SMA para MAX apresentou classificação filtrada nula em muitos momentos devido ao elevado *threshold* calculado durante o treinamento das janelas.

5.3.2 Mediana Móvel Simples

Figura 35 – Comportamento do *recall* de acordo com o parâmetro de entrada do filtro SMM e abordagem Pós-Classificação.

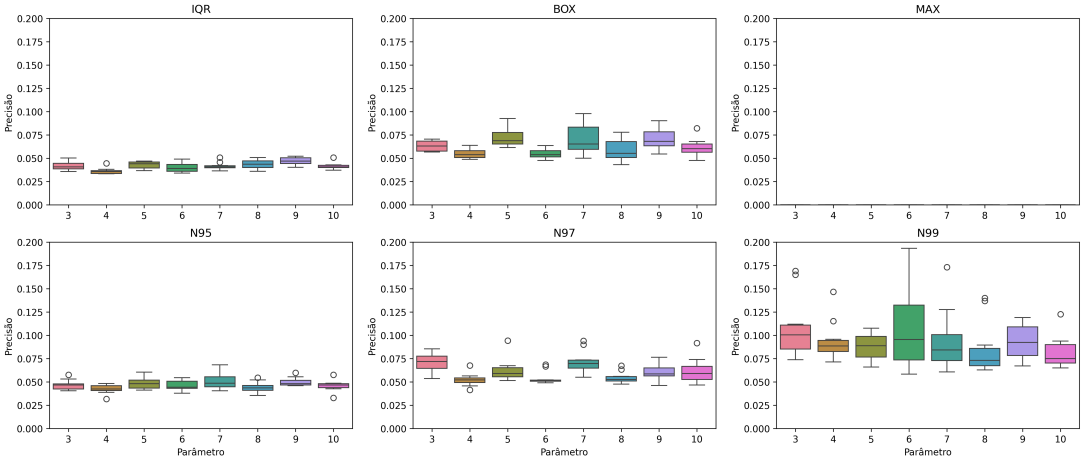


Fonte: A autora (2024).

Para a aplicação de SMM com a abordagem Pós-Classificação de aplicação de filtro na detecção (Pós-Classificação + SMM), foi observada uma redução do *recall* devido a redução de TP observada em geral para a abordagem Pré-Classificação + SMM apresentado na Figura 35 por causa do efeito de diluição da primeira classificação dos ciclos que foi mais impactado com o uso da mediana ao invés da média. A redução do *recall* foi observada estatisticamente para todos os *thresholds* com exceção de N97. Quanto a escolha do parâmetro de entrada, todos os modelos possuíram influência significativa no *recall*, com exceção de MAX (p-valor Friedman de $0,741$) que assim como à detecção com Pós-Classificação + SMA teve o TP extremamente prejudicado. Apesar da influência estatística do parâmetro de entrada, apenas N99 apresentou uma tendência estatística de aumento da métrica com a redução da janela de entrada n. Além disso, é interessante destacar que a janela com tamanho 4 apresentou um

recall estatisticamente superior para IQR, N95, N97 e N99, e que em geral as janelas pares apresentaram resultados superiores para essa métrica.

Figura 36 – Comportamento da precisão de acordo com o parâmetro de entrada do filtro SMM e abordagem Pós-Classificação.

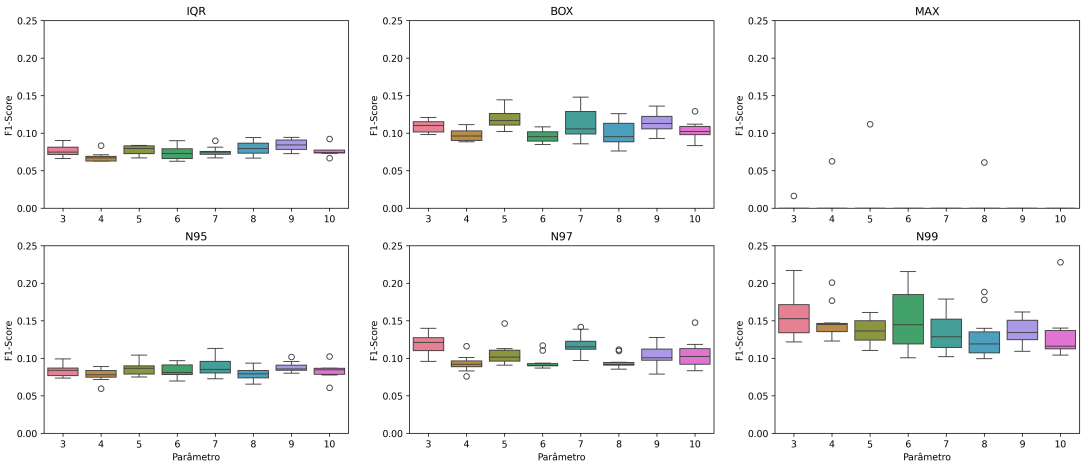


Fonte: A autora (2024).

A Figura 36 apresenta o comportamento da precisão de acordo com a escolha do parâmetro de entrada do filtro SMM para os *thresholds* avaliados. Foi observado um aumento geral da precisão, com exceção de MAX que foi deteriorada pela redução de TP. Esse aumento foi provocado pela redução de FP promovido por esta configuração já que o SMM é ainda menos susceptível do que o SMA à instâncias classificadas isoladamente como anomalia devido ao princípio de cálculo da mediana. Porém, do ponto de vista estatístico, o aumento da precisão não foi considerável quando realizada a comparação entre detecção com filtro e sem filtro. A respeito da influência do parâmetro de entrada do SMM, foi observado que para todos os *thresholds*, com exceção de MAX (p-valor Friedman de $0,741$) e N99 (p-valor Friedman $= 0,333$), houve uma alteração no comportamento da precisão, apesar de não haver tendência explícita para nenhuma técnica.

O F1-Score, apresentado na Figura 37, sofreu uma redução com a aplicação de Pós-Classificação + SMM para todos os *thresholds*, apesar de não haver uma diferença estatística significativa entre a detecção com filtro e sem filtro com exceção de MAX. Esta redução, mesmo que não significativa, foi provocada pela redução do *recall* observada nesta configuração. Além disso, foi observado que apesar de haver estatisticamente uma alteração do comportamento de F1-Score com a seleção do parâmetro de entrada do SMM para IQR, BOX e N97, os *thresholds* MAX (p-valor Friedman de $0,741$), N95 (p-valor ANOVA de $0,112$) e

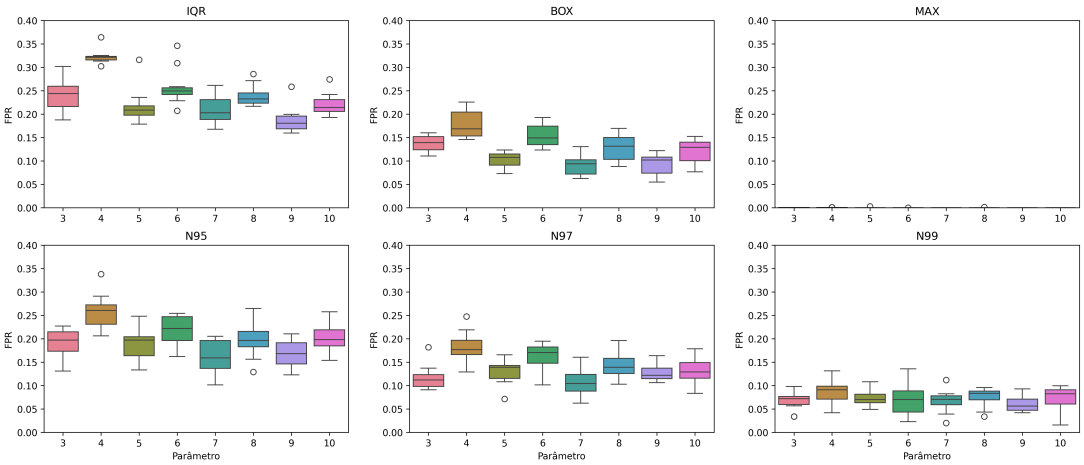
Figura 37 – Comportamento do F1-Score de acordo com o parâmetro de entrada do filtro SMM e abordagem Pós-Classificação.



Fonte: A autora (2024).

N99 (p-valor Friedman de $0,139$) não apresentaram diferença estatística. Além disso, não foi identificada uma relação clara entre tamanho da janela n e melhoria do F1-Score.

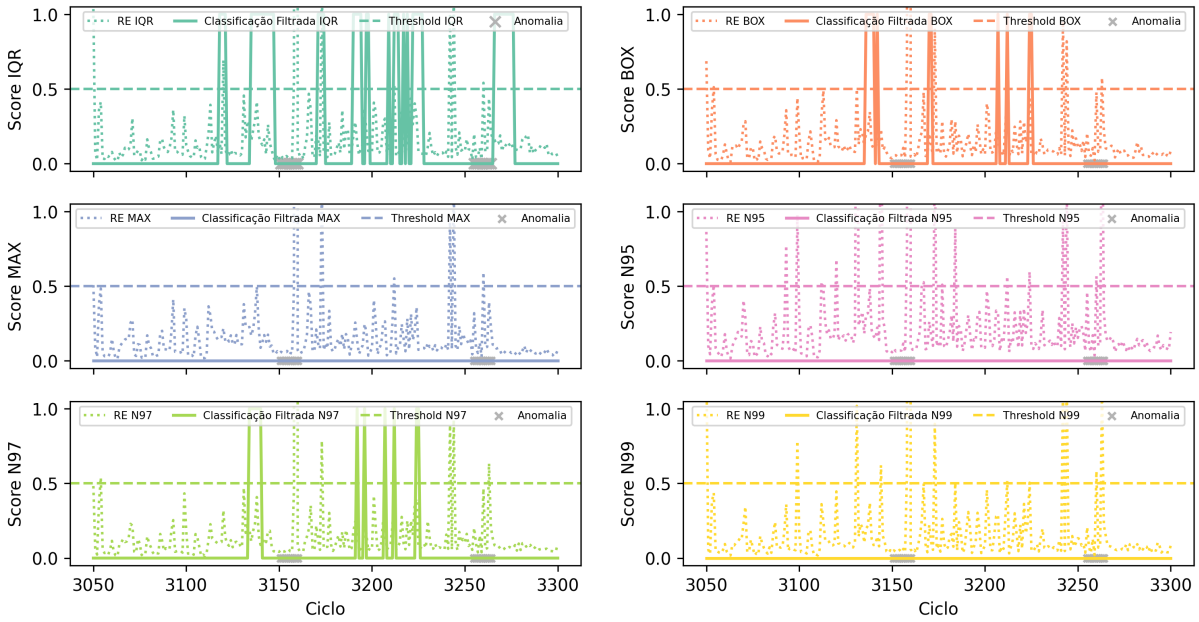
Figura 38 – Comportamento do FPR de acordo com o parâmetro de entrada do filtro SMM e abordagem Pós-Classificação.



Fonte: A autora (2024).

Já para o FPR, apresentado na Figura 38, também foi apresentada uma redução assim como na configuração Pós-Classificação + SMA. Apesar da mudança na métrica, a diferença não foi estatisticamente significativa para nenhuma abordagem de *threshold* para valores pequenos de janela de entrada, em particular para n $= 4$. Mesmo com a influência estatística entre FPR e escolha da janela n para todos os *thresholds* com exceção de MAX (p-valor Friedman de $0,003$, mas não teve diferença confirmada pelo teste post-hoc de Nemenyi), não ocorreu nenhuma

Figura 39 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro SMM com abordagem Pós-Classificação com minimização de FPR.



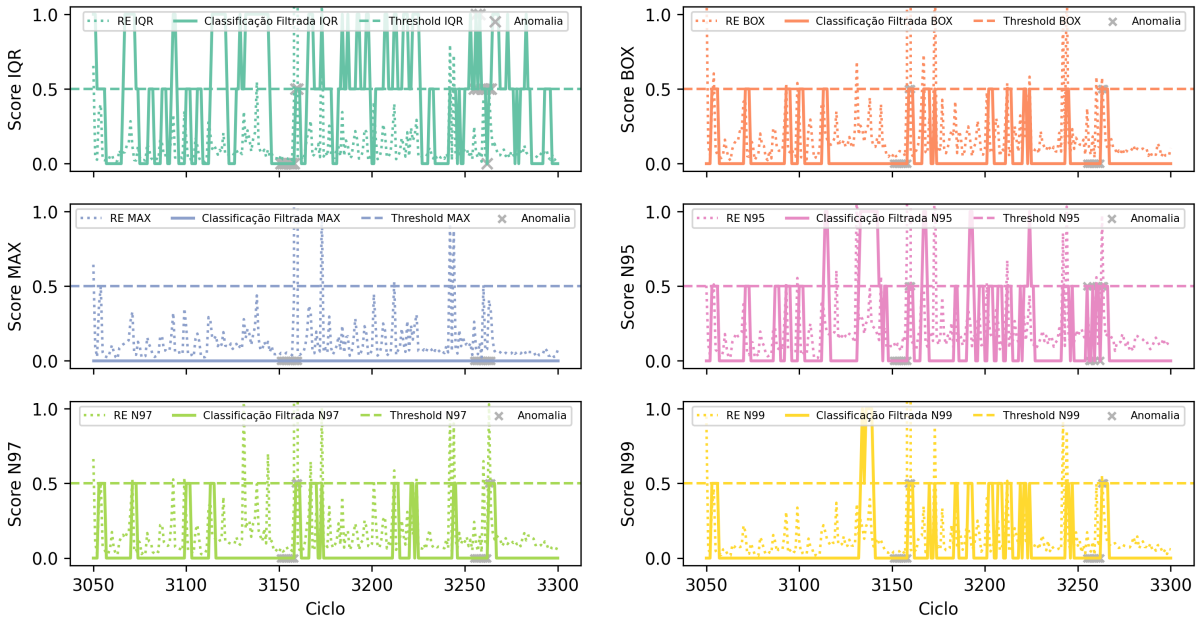
Fonte: A autora (2024).

relação entre aumento e redução de janela com a métrica em si. É interessante salientar que, assim como destacado na avaliação do *recall*, as janelas pares apresentaram FPR superior, sendo assim preferida a adoção de janelas ímpares para a minimização do FPR.

Quando foi avaliada a minimização do FPR para a configuração Pós-Classificação + SMM, foram selecionadas as janelas n = 9; 7; 7; 7; 7; 9 para os *thresholds* IQR, BOX, MAX, N95, N97 e N99 respectivamente, que se aproximaram dos parâmetros utilizados para a minimização de FPR para o filtro SMA. Isso ocorreu porque, de uma forma geral, a aplicação de janelas maiores diminuiu a relevância de classificações equivocadas de instâncias isoladas para ambos os filtros. Porém ao comparar as Figuras 33 e 39 é possível visualizar que a suavização de SMM para a abordagem Pós-Classificação apresentou picos devido ao cálculo da mediana enquanto a suavização de SMA foi contínua, o que prejudicou a detecção de sequências pelo filtro SMM nesse cenário apesar de ter gerado uma otimização na redução de FPR.

A Figura 40 apresenta o comportamento da detecção quando foi realizada uma maximização do *recall* utilizando as janelas n = 4; 4; 5; 4; 4; 4 para os *thresholds* IQR, BOX, MAX, N95, N97 e N99 respectivamente. Mesmo considerando a melhoria no *recall* foi observado que o filtro SMM, com essa seleção de janela de entrada, não captou bem sequências de anomalias, além de apresentar um atraso em relação as anomalias reais que aumenta o FP e reduz o TP. Um ponto que vale a pena salientar é que em trabalhos futuros é interessante testar

Figura 40 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro SMM com abordagem Pós-Classificação com maximização do *recall*.

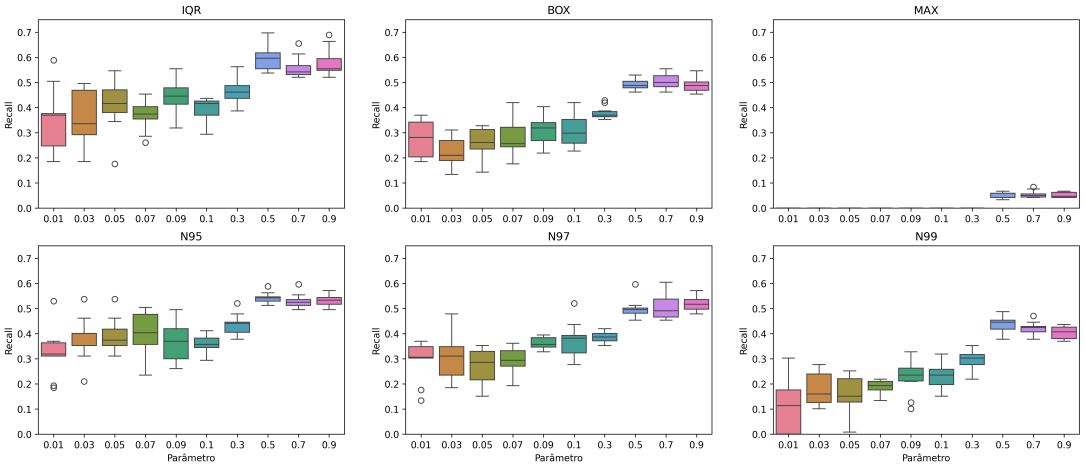


Fonte: A autora (2024).

diferentes *thresholds* experimentais além de $0,50$ para essa abordagem que pode modificar o comportamento das métricas e otimizar a detecção.

5.3.3 Média Móvel Exponencial

Figura 41 – Comportamento do *recall* de acordo com o parâmetro de entrada do filtro EMA e abordagem Pós-Classificação.

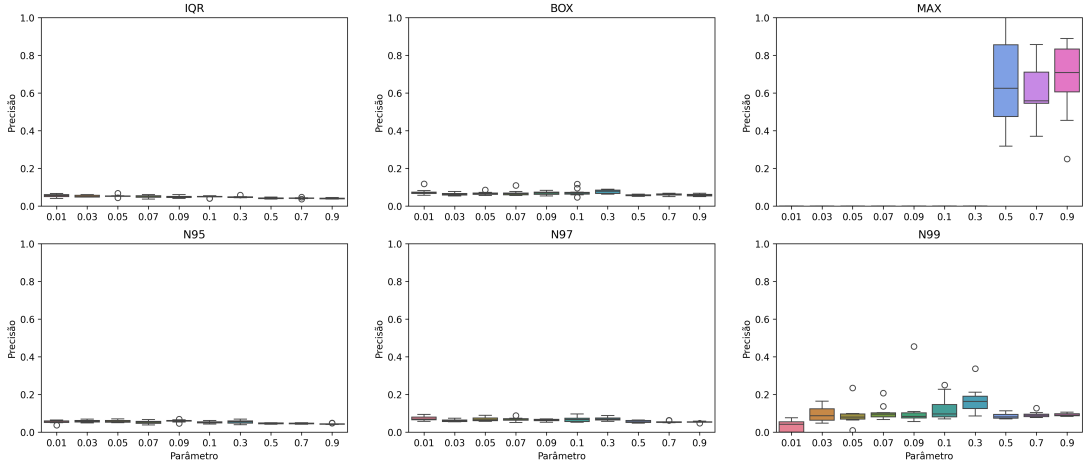


Fonte: A autora (2024).

Por fim, com a aplicação de EMA com abordagem Pós-Classificação de filtro (Pós-Classificação

+ EMA), foi observado uma grande variação no *recall* a depender do parâmetro de entrada assim como em Pré-Classificação + EMA. A Figura 41 apresenta o comportamento da métrica com diferentes parâmetros de entrada para essa configuração de detecção. Do ponto de vista estatístico, foi verificado que configurações que não utilizaram filtro para a detecção obtiveram *recall* similares aos que utilizaram filtros com α mais elevado. Além disso, foi verificado que houve uma influência estatística significativa na métrica para todos os *thresholds* com o parâmetro de entrada, mas a relação direta entre aumento do *recall* e aumento de α ocorreu para IQR, BOX, N95, N97 e N99. É interessante discutir que apenas a partir do parâmetro de entrada 0,5 a configuração com *threshold* MAX passou a detectar algumas instâncias anômalas de forma correta, apesar de ainda apresentar um baixo TP. Isso ocorreu porque com α mais elevado, o filtro apresentou uma suavização mais ágil e menos influenciada pelo sinal filtrado anterior.

Figura 42 – Comportamento da precisão de acordo com o parâmetro de entrada do filtro EMA e abordagem Pós-Classificação.

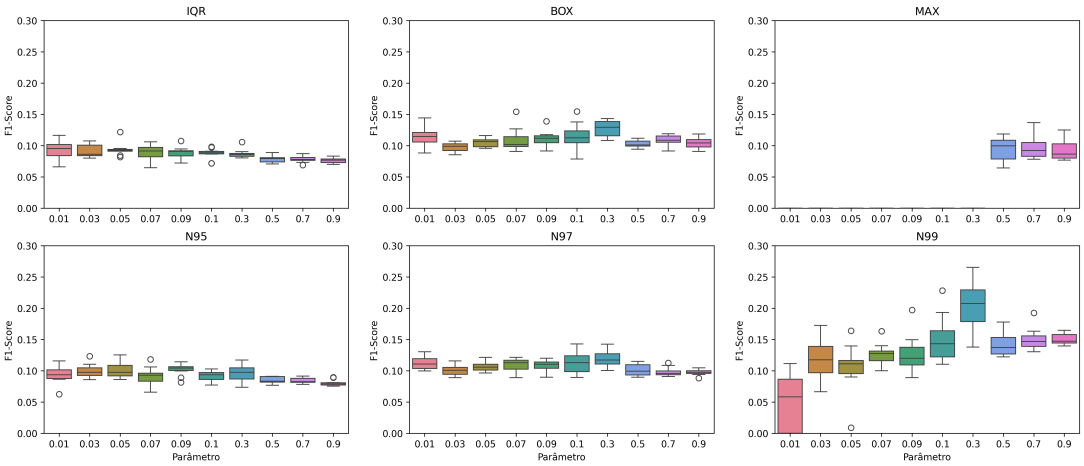


Fonte: A autora (2024).

O comportamento da precisão, apresentado na Figura 42, para esta configuração de detecção foi superior à detecção sem filtro devido a redução de FP alcançada pela abordagem Pós-Classificação de filtro, mas não foi identificada diferença estatística significativa entre o modelo sem filtro e modelos com α mais altos. Também foi verificado que a precisão dependeu do parâmetro de entrada para todos os *thresholds* avaliados, mas IQR, N95 e N97 apresentaram uma relação clara inversamente proporcional entre a métrica e a entrada do filtro, já que o aumento do parâmetro de entrada aumentou o FP a partir de 0,5 para todos os *thresholds*.

Assim como a detecção com Pós-Classificação + SMA, a utilização de EMA também gerou

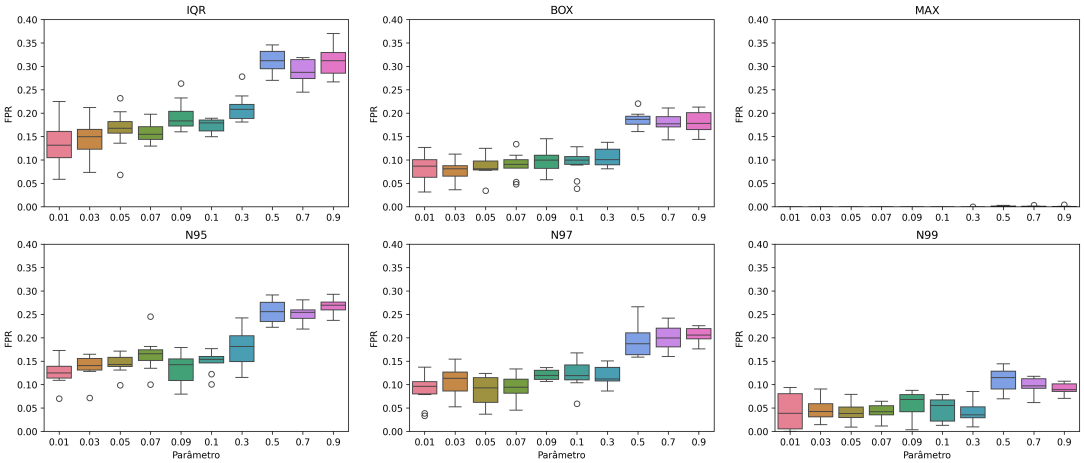
Figura 43 – Comportamento do F1-Score de acordo com o parâmetro de entrada do filtro EMA e abordagem Pós-Classificação.



Fonte: A autora (2024).

um aumento de F1-Score quando comparada a detecção sem filtro e não houve diferença estatística significativa para nenhum dos *thresholds* avaliados quando foram utilizados valores mais elevados de α. Para o F1-Score, apresentado na Figura 43, também foi identificada uma influência do parâmetro de entrada para todos os *thresholds*, mas só N99 apresentou uma relação direta entre aumento da métrica e aumento do parâmetro de entrada, já que o *recall* e a precisão também apresentaram o mesmo comportamento, enquanto para os outros *thresholds* foi verificada uma tendência contrária, mesmo que não estatisticamente significativa.

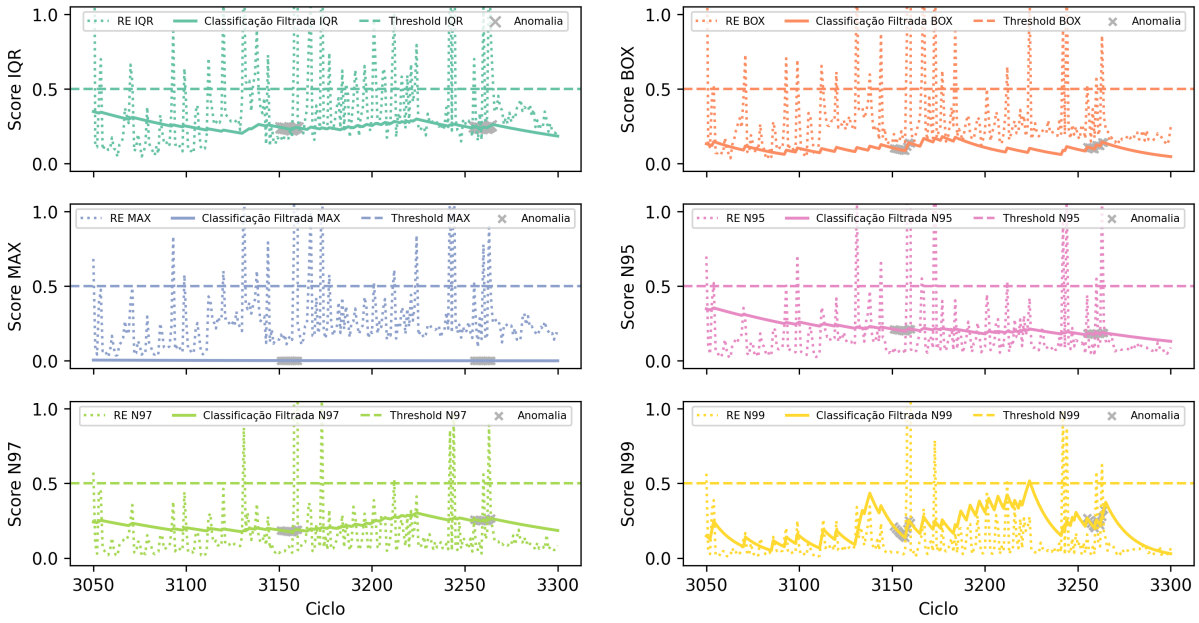
Figura 44 – Comportamento do FPR de acordo com o parâmetro de entrada do filtro EMA e abordagem Pós-Classificação.



Fonte: A autora (2024).

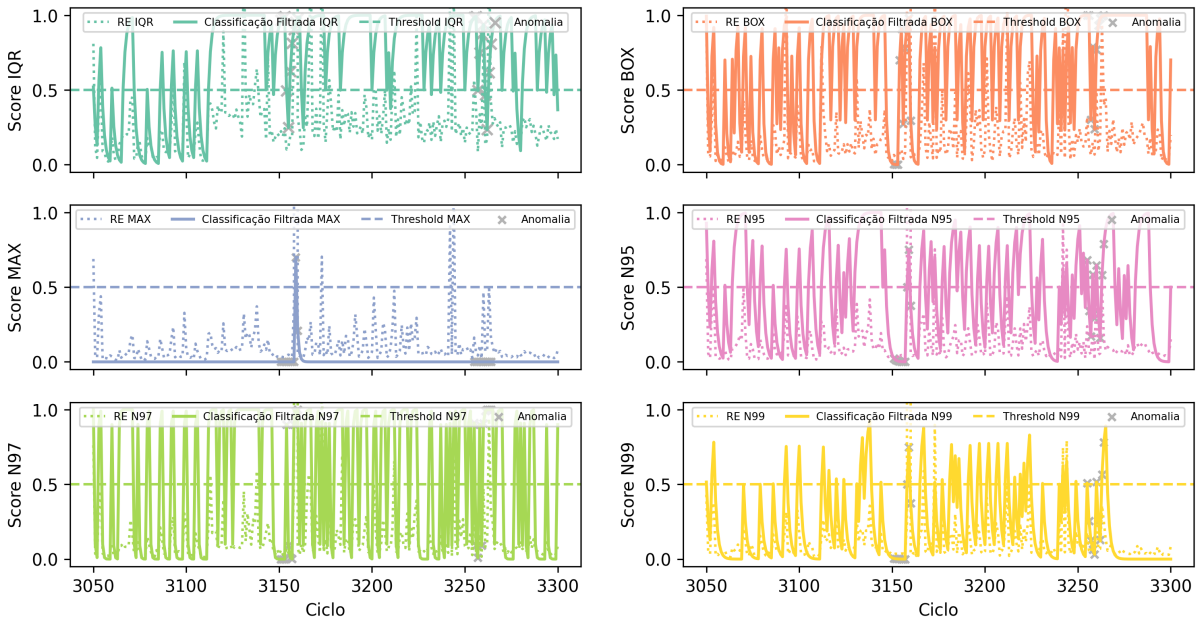
A utilização de Pós-Classificação + EMA também gerou redução de FPR quando compa-

Figura 45 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro EMA com abordagem Pós-Classificação com minimização de FPR.



Fonte: A autora (2024).

Figura 46 – Exemplo de comportamento da AD em uma janela de teste com a aplicação do filtro EMA com abordagem Pós-Classificação com maximização do *recall*.



Fonte: A autora (2024).

rado à detecção sem filtro, porém esta não foi estatisticamente significativa para parâmetros de entrada mais elevados. A influência entre a seleção de α e a métrica FPR foi estatisticamente relevante para todos os *thresholds* e pode ser visualizada na Figura 44. Além disso, foi identificada uma relação diretamente proporcional entre α e FPR para todos os *thresholds*, com exceção de MAX, e vale destacar o aumento considerável da métrica quando foram utilizados α acima de 0,5. A Figura 45 apresenta o comportamento da detecção desta configuração considerando uma minimização de FPR que utilizou os parâmetros de entrada de EMA α = 0,01; 0,03; 0,01; 0,01; 0,01; 0,07 para os *thresholds* IQR, BOX, MAX, N95, N97 e N99 respectivamente. Foi possível ver que com a redução de FPR a redução de TP foi maior para EMA do que para os demais filtros avaliados com a mesma abordagem de detecção.

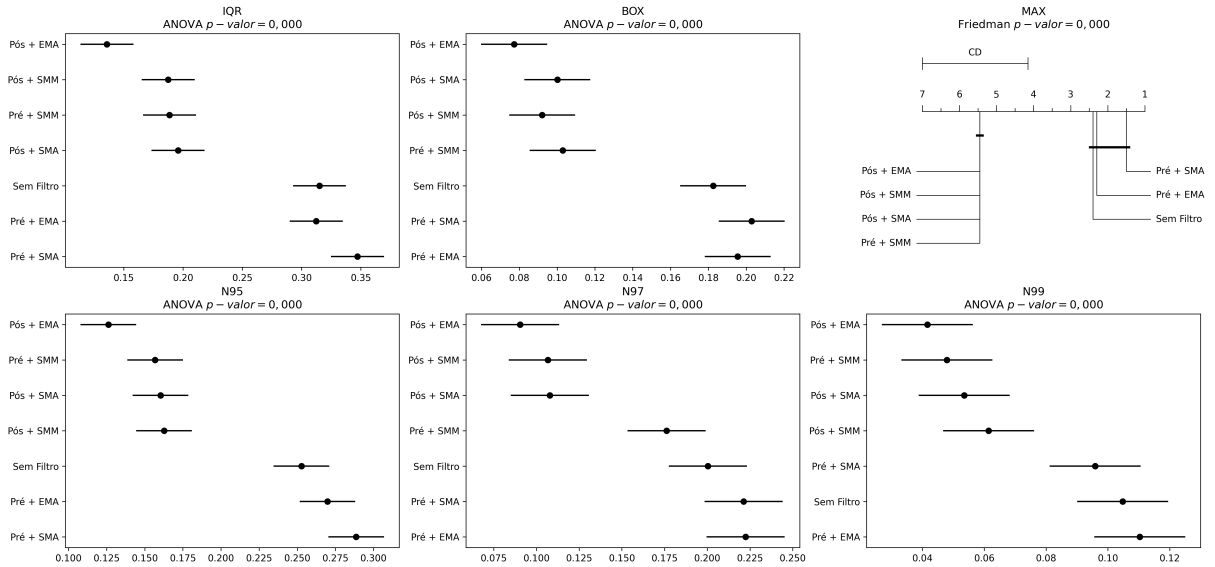
Já a Figura 46 apresenta o comportamento da detecção com configuração Pós-Classificação + EMA considerando a maximização do *recall*. Para esta avaliação foram utilizadas os parâmetros de entrada que estatisticamente apresentaram uma melhoria no *recall* α = 0,5; 0,7; 0,7; 0,5; 0,9; 0,5 para os *thresholds* IQR, BOX, MAX, N95, N97 e N99 respectivamente. É possível perceber que o aumento do *recall* também gerou um aumento significativo de FP e pela suavização ser mais ágil também diminui a detecção de sequências de anomalia.

5.4 AVALIAÇÃO DO IMPACTO DO USO DE FILTRO PARA DETECÇÃO

Para realizar a avaliação do impacto do filtro e da abordagem de utilização adotada de acordo com a técnica de *threshold* utilizada, serão apresentados os cenários que alcançaram uma minimização de FPR, uma maximização de *recall* e uma maximização de F1-Score juntamente com uma avaliação estatística destas métricas.

A Figura 47 apresenta a síntese dos testes estatísticos realizados a respeito do comportamento do FPR quando foi realizada a minimização do FPR. Foi observado que a abordagem Pós-Classificação de filtro provocou uma maior minimização de FPR quando comparado a abordagem Pré-Classificação de uma forma geral. Para os *thresholds* IQR, BOX, MAX, N95 e N99 foi obtida uma redução de FPR nas configurações de detecção Pós-Classificação + EMA, Pós-Classificação + SMM, Pós-Classificação + SMA e Pré-Classificação + SMM, sendo o primeiro responsável pela maior redução. Já para N97 a configuração Pré-Classificação + SMM não gerou redução do FPR significativa, porém Pós-Classificação + EMA, Pós-Classificação + SMM e Pós-Classificação + SMA geraram. É interessante ressaltar que para a abordagem Pré-Classificação + EMA foi verificada uma preferência por parâmetros de entrada mais ele-

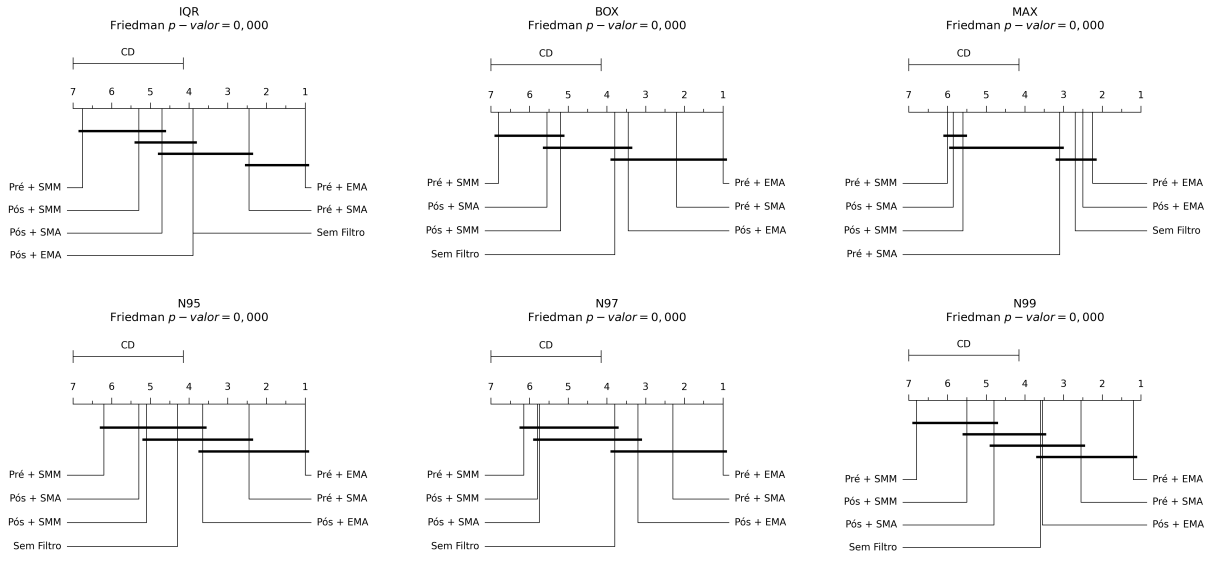
Figura 47 – Avaliação estatística dos melhores modelos selecionados de acordo com a minimização de FPR.



Fonte: A autora (2024).

vados, enquanto para Pós-Classificação + EMA houve uma preferência por valores menores. Quanto a detecção alcançada por MAX, apesar do FPR ser reduzido ao máximo, houve um impacto muito grande na detecção real TP que chegou até a zerar o *recall* e F1-Score. Dessa forma, para a minimização do FPR, a configuração Pós-Classificação + EMA com pequenos α apresentou os melhores resultados para todos os *thresholds* avaliados.

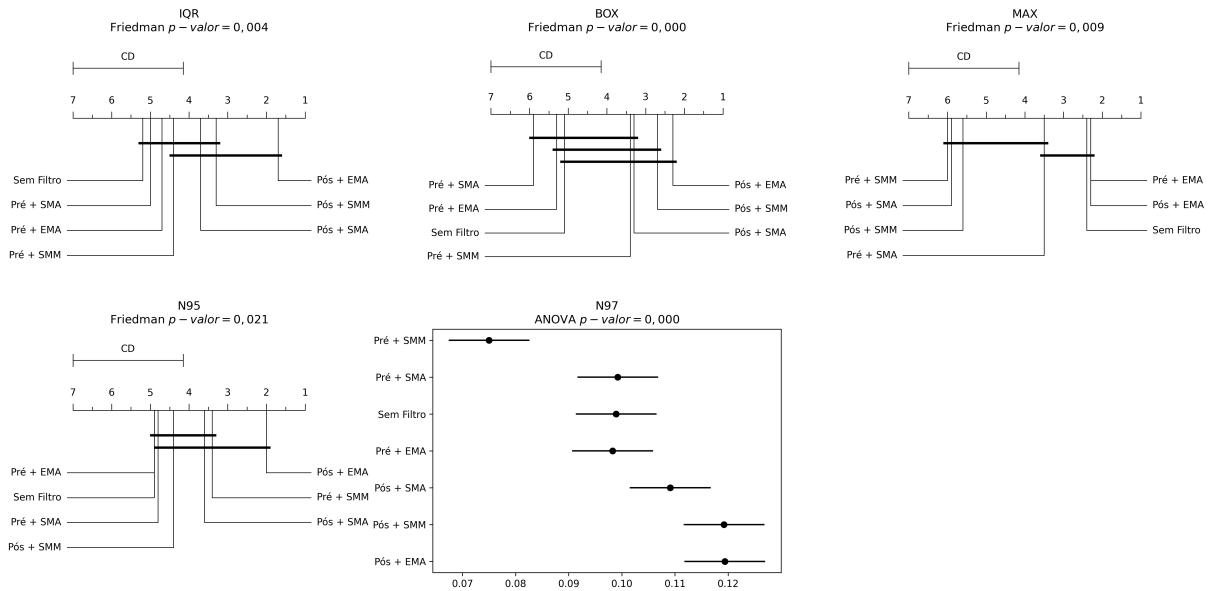
Figura 48 – Avaliação estatística dos melhores modelos selecionados de acordo com a maximização do *recall*.



Fonte: A autora (2024).

Quanto ao *recall*, a Figura 48 apresenta uma síntese dos testes estatísticos da métrica para as configurações que alcançaram uma maximização do *recall*. Foi verificado um aumento significativo do *recall* para o *threshold* IQR quando foi utilizada a configuração Pré-Classificação + EMA e Pré-Classificação + SMA, sendo que o aumento alcançado por Pré-Classificação + EMA também provocou um aumento expressivo de FPR. Para N95, além de aumento da métrica com Pré-Classificação + EMA e Pré-Classificação + SMA, houve um aumento com a configuração Pós-Classificação + EMA que não provocou o mesmo aumento de FPR da abordagem Pré-Classificação devido à lenta suavização para parâmetros de entrada muito baixos. BOX, N97 e N99 não tiveram o *recall* alterado estatisticamente, apesar de haver um aumento da métrica com a aplicação das configurações Pré-Classificação + EMA, Pré-Classificação + SMA e Pós-Classificação + EMA. Enquanto isso, MAX apresentou uma deterioração significativa do *recall* com a aplicação de filtro, com exceção das configurações Pré-Classificação + EMA e Pós-Classificação + EMA. Com isso, é possível inferir que o filtro SMM não auxiliou a maximização de *recall* em nenhuma das abordagens testadas e que a configuração Pré-Classificação + EMA e Pré-Classificação + SMA apresentaram resultados satisfatórios.

Figura 49 – Avaliação estatística dos melhores modelos selecionados de acordo com a maximização de F1-Score.



Fonte: A autora (2024).

Por fim, a maximização do F1-Score é apresentado na Tabela 8 e a síntese estatística é apresentada na Figura 49. Na tabela, as configurações destacadas em negrito representam as configurações de um determinado *threshold* e abordagem de filtro que alcançaram melhor

Tabela 8 – Métricas obtidas na maximização de F1-Score para todas as configurações.

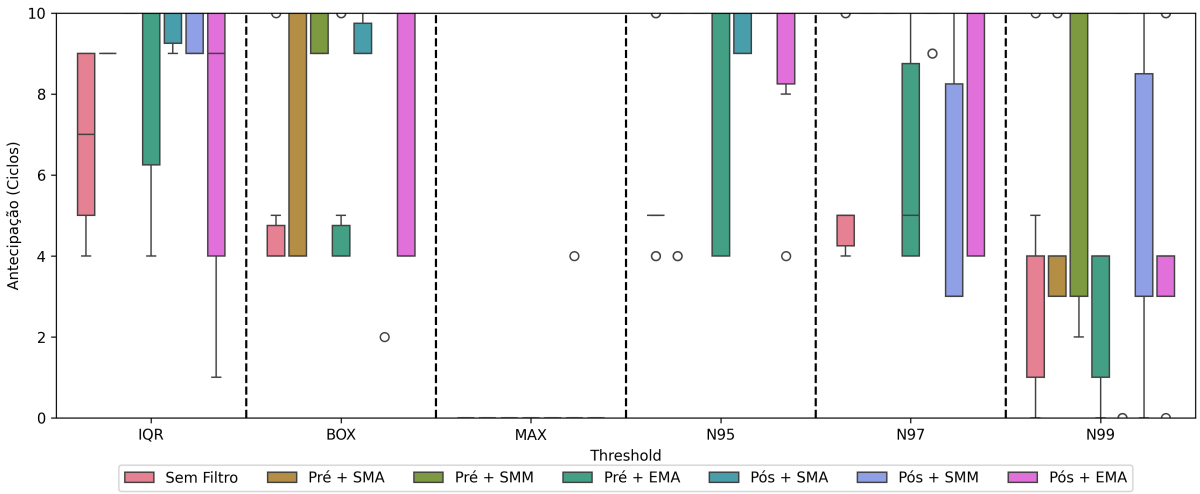
Threshold	Detecção	Parâmetro	Recall (%)	Precisão (%)	F1-Score (%)	FPR (%)	Antecipação
IQR	*Sem Filtro*	-	*58,91 ± 4,29*	*4,09 ± 0,34*	*7,65 ± 0,60*	*31,51 ± 3,24*	*2,53 ± 1,20*
	Pré + SMA	n = 4	65,55 ± 4,69	4,13 ± 0,35	7,77 ± 0,63	34,72 ± 3,25	3,35 ± 1,03
	Pré + SMM	**n = 10**	**40,08 ± 6,25**	**4,63 ± 0,50**	**8,28 ± 0,83**	**18,86 ± 2,95**	**2,90 ± 1,14**
	Pré + EMA	$\alpha = 0{,}9$	61,93 ± 4,48	4,33 ± 0,34	8,09 ± 0,61	31,24 ± 3,32	2,79 ± 0,64
	Pós + SMA	n = 9	43,95 ± 6,18	4,39 ± 0,75	7,95 ± 1,26	22,26 ± 4,38	3,22 ± 1,43
	Pós + SMM	n = 9	40,59 ± 6,09	4,70 ± 0,42	8,41 ± 0,76	18,75 ± 2,85	3,05 ± 1,02
	Pós + EMA	**$\alpha = 0{,}05$**	**40,67 ± 9,98**	**5,36 ± 0,60**	***9,39 ± 1,08***	**16,45 ± 4,31**	**3,53 ± 1,74**
BOX	*Sem Filtro*	-	*50,67 ± 4,04*	*6,00 ± 0,54*	*10,70 ± 0,81*	*18,25 ± 2,85*	*1,63 ± 0,35*
	Pré + SMA	n = 3	52,27 ± 2,99	5,56 ± 0,45	10,05 ± 0,72	20,29 ± 2,29	2,46 ± 0,34
	Pré + SMM	**n = 10**	**34,12 ± 6,01**	**7,31 ± 1,31**	**11,86 ± 1,35**	**10,30 ± 3,14**	**2,69 ± 1,13**
	Pré + EMA	$\alpha = 0{,}9$	53,36 ± 3,37	5,88 ± 0,44	10,58 ± 0,66	19,55 ± 2,44	1,83 ± 0,56
	Pós + SMA	n = 9	32,60 ± 3,14	7,16 ± 1,48	11,63 ± 1,78	10,01 ± 1,78	2,35 ± 0,63
	Pós + SMM	n = 5	34,45 ± 3,12	7,20 ± 0,97	11,86 ± 1,27	10,26 ± 1,70	2,27 ± 0,49
	Pós + EMA	**$\alpha = 0{,}3$**	**37,90 ± 2,61**	**7,69 ± 1,01**	**12,73 ± 1,31**	**10,57 ± 2,04**	**2,14 ± 0,43**
MAX	*Sem Filtro*	-	*5,04 ± 0,97*	*56,37 ± 24,88*	*9,05 ± 1,56*	*0,14 ± 0,14*	*0,00 ± 0,00*
	Pré + SMA	n = 4	4,62 ± 1,95	26,63 ± 9,20	7,79 ± 3,05	0,29 ± 0,12	0,00 ± 0,00
	Pré + SMM	n = 4	0,08 ± 0,03	1,00 ± 3,16	0,16 ± 0,49	0,02 ± 0,05	0,00 ± 0,00
	Pré + EMA	**$\alpha = 0{,}9$**	**5,80 ± 2,04**	**50,86 ± 8,00**	***10,28 ± 3,24***	**0,13 ± 0,06**	**0,01 ± 0,02**
	Pós + SMA	n = 3	0,34 ± 1,06	4,44 ± 14,05	6,25 ± 1,98	0,02 ± 0,03	0,00 ± 0,00
	Pós + SMM	n = 5	6,72 ± 2,13	3,33 ± 10,54	1,12 ± 3,54	0,03 ± 0,10	0,02 ± 0,08
	Pós + EMA	**$\alpha = 0{,}7$**	**5,46 ± 1,44**	**61,84 ± 15,36**	**9,78 ± 2,12**	**0,10 ± 0,09**	**0,01 ± 0,02**
N95	*Sem Filtro*	-	*53,53 ± 2,86*	*4,60 ± 0,20*	*8,46 ± 0,34*	*25,28 ± 1,78*	*1,91 ± 0,48*
	Pré + SMA	n = 3	58,74 ± 3,39	4,45 ± 0,49	8,28 ± 0,86	28,87 ± 2,96	2,72 ± 0,31
	Pré + SMM	**n = 9**	**39,41 ± 6,57**	**5,37 ± 0,65**	**9,35 ± 1,00**	**16,18 ± 4,05**	**3,01 ± 1,37**
	Pré + EMA	$\alpha = 0{,}7$	61,85 ± 4,95	4,50 ± 0,36	8,38 ± 0,62	30,10 ± 4,40	2,74 ± 0,77
	Pós + SMA	n = 9	41,26 ± 6,95	5,21 ± 0,90	9,22 ± 1,52	17,30 ± 2,71	3,34 ± 1,19
	Pós + SMM	n = 7	36,81 ± 3,36	5,03 ± 0,85	8,81 ± 1,21	16,27 ± 3,57	2,84 ± 0,55
	Pós + EMA	**$\alpha = 0{,}09$**	**36,97 ± 8,38**	**5,94 ± 0,62**	***10,17 ± 0,98***	**13,40 ± 3,14**	**2,95 ± 2,10**
N97	*Sem Filtro*	-	*50,50 ± 2,29*	*5,49 ± 0,63*	*9,89 ± 0,99*	*20,04 ± 3,11*	*1,78 ± 0,43*
	Pré + SMA	**n = 5**	**53,87 ± 6,54**	**5,49 ± 0,65**	**9,92 ± 0,98**	**21,55 ± 4,92**	**2,79 ± 0,63**
	Pré + SMM	n = 10	37,33 ± 3,61	4,19 ± 0,74	7,50 ± 1,18	17,62 ± 2,35	2,70 ± 0,67
	Pré + EMA	$\alpha = 0{,}9$	55,63 ± 5,16	5,39 ± 0,38	9,83 ± 0,65	22,25 ± 2,48	2,24 ± 1,11
	Pós + SMA	n = 9	35,71 ± 3,61	6,49 ± 0,92	*10,91 ± 1,16*	12,01 ± 2,62	2,88 ± 0,91
	Pós + SMM	n = 3	38,40 ± 2,91	7,09 ± 0,96	*11,92 ± 1,30*	11,73 ± 2,70	2,02 ± 0,38
	Pós + EMA	**$\alpha = 0{,}3$**	**38,57 ± 2,29**	**7,09 ± 1,02**	***11,94 ± 1,39***	**11,74 ± 2,18**	**2,18 ± 0,60**
N99	*Sem Filtro*	-	*43,70 ± 3,50*	*8,92 ± 1,38*	*14,72 ± 1,74*	*10,48 ± 2,49*	*1,44 ± 0,45*
	Pré + SMA	n = 3	43,36 ± 4,47	9,32 ± 1,62	15,24 ± 2,17	9,96 ± 2,69	1,44 ± 0,74
	Pré + SMM	**n = 4**	**30,42 ± 2,96**	**13,02 ± 4,83**	**17,65 ± 4,67**	**5,48 ± 2,77**	**1,13 ± 0,44**
	Pré + EMA	$\alpha = 0{,}9$	45,21 ± 2,27	8,72 ± 1,30	14,56 ± 1,76	11,03 ± 2,21	1,30 ± 0,30
	Pós + SMA	n = 8	26,81 ± 2,92	12,05 ± 4,54	16,02 ± 3,64	5,19 ± 2,50	1,49 ± 1,10
	Pós + SMM	n = 3	32,94 ± 3,36	10,84 ± 3,33	15,98 ± 3,31	6,69 ± 2,03	1,29 ± 0,64
	Pós + EMA	**$\alpha = 0{,}3$**	**29,66 ± 3,82**	**16,91 ± 7,20**	**20,31 ± 4,06**	**4,10 ± 2,27**	**1,11 ± 0,63**

Fonte: A autora (2024).

F1-Score. Os valores de F1-Score destacados em itálico representam os valores que foram estatisticamente superiores aos respectivos modelos sem filtro. Apesar de, em geral, a abordagem Pós-Classificação de aplicação de filtro resultar em aumento de F1-Score (relacionada a redução de FPR e aumento da precisão), a aplicação de filtros não gerou aumento da métrica de forma estatisticamente significativa para BOX e MAX quando comparada a detecção sem filtro. E que a aplicação de SMA e SMM em ambas as abordagens de aplicação de filtro

deteriorou significativamente o F1-Score para MAX. Para N99, não foi detectada nenhuma diferença estatística entre as configurações com filtro e sem filtro (p-valor Friedman $0,107$). Já a configuração Pós-Classificação gerou aumento significativo de F1-Score para IQR e N95 quando foi utilizado EMA e para todos os filtros para N97.

Figura 50 – Comportamento da antecipação em ciclos para uma sequência de duas anomalias.



Fonte: A autora (2024).

A métrica Antecipação apresentada na Tabela 8 representa quantos ciclos, em média, foram identificados como anomalia antes do evento anômalo de fato ocorrer de acordo com a tabela verdade para todas as anomalias do conjunto de teste. Vale ressaltar que esta métrica foi muito impactada por anomalias que não foram identificadas ou não foram antecipadas. A Figura 50 apresenta o comportamento dos modelos apresentados na Tabela 8 para uma sequência de anomalias de dois ciclos do conjunto de teste. Por essa Figura, é possível constatar que a aplicação de filtros foi de fato vantajosa para antecipar as anomalias e detectar sequências de anomalias com destaque para a configuração Pré-Classificação + EMA e a abordagem Pós-Classificação.

Em síntese, foi observado que a abordagem Pós-Classificação de filtro foi mais interessante para maximizar o F1-Score neste cenário devido a redução de FP, com destaque para o LPF EMA. Já para o aumento de TP, a abordagem Pré-Classificação se mostrou mais interessante com destaque para SMA e EMA. Além disso, cabe estudos futuros com ênfase na determinação do *threshold* experimental utilizado pela abordagem Pós-Classificação para buscar otimizações desta. Também foi observado que a seleção dos parâmetros de entrada do filtro são significativamente relevantes para o comportamento das métricas e deve ser avaliado previamente.

Quanto aos filtros que foram avaliados, verificou-se que a aplicação destes foi muito relevante para a detecção de sequências de anomalias, sendo SMM o menos afetado por picos e EMA o mais ágil quando são utilizados parâmetros de entrada mais altos e o mais lento quando são utilizados parâmetros mais baixos.

A respeito dos *thresholds* que foram aplicados para a detecção, observou-se que MAX apresentou resultados positivos para FP, porém sua detecção de TP foi muito afetada pelo *threshold* calculado durante o treinamento do modelo devido a picos de erro de reconstrução. Para a maximização de TP, os *thresholds* que apresentaram melhores resultados foram o IQR seguido por N95, BOX e N97. Já do ponto de vista de maximização de F1-Score, foi observado que N99 apresentou as melhores métricas, o que indica um melhor balanço entre *recall* e precisão. Portanto destaca-se a importância da seleção da técnica de *threshold* mais adequada de acordo com as métricas que devem ser priorizadas para a AD do cenário avaliado.

6 CONCLUSÃO

Na última seção deste trabalho são apresentadas as principais contribuições dessa dissertação, sugestões para trabalhos futuros e trabalhos que foram publicados.

6.1 CONCLUSÕES

Essa dissertação propôs investigar a influência da seleção da abordagem de cálculo de *threshold* adaptativo em um modelo *online* com arquitetura de SAE. Além disso, também foi avaliada a aplicação de três LPFs, SMA, SMM e EMA de duas formas: (Pré-Classificação) sob o erro de reconstrução da janela de teste, ou seja, imediatamente antes da classificação das instâncias; e (Pós-Classificação) após a classificação inicial das instâncias. Para ter uma referência da abordagem sem aplicação de filtro, também foram consideradas 6 aplicações com filtro utilizando a mesma base de dados e as abordagens de definição de *threshold* adaptativo IQR, BOX, MAX, N95, N97 e N99.

A respeito da avaliação do *threshold* apenas, não foi identificada diferença estatística entre [N97, N99 e BOX] para nenhuma das quatro métricas avaliadas. Dentre as abordagens estudadas, MAX apresentou a menor taxa de FP e TP, já que nesta abordagem foram computados *thresholds* muito elevados devido a erros pontuais demasiadamente altos durante o treinamento. Já IQR apresentou a maior taxa de TP, devido ao menor valor de *threshold* calculado nas janelas de treinamento pelo princípio de cálculo. Já do ponto de vista de F1-Score, foi observado que N99 apresentou a maior média, apesar de ser estatisticamente semelhante a BOX e N97. Sendo assim, visando maximizar o F1-Score, os *thresholds* N99, BOX e N97 foram mais interessantes por apresentar um melhor balanço entre precisão e *recall*.

De uma forma geral, a abordagem Pré-Classificação de filtro promoveu uma melhoria de TP e uma deterioração de FP, enquanto Pós-Classificação provocou o efeito oposto, uma melhoria de FP e deterioração de FP. Apesar disso, ambas as aplicações de filtro permitiram detectar sequências de anomalias, que não foi possível com a detecção sem filtro. Contudo, a configuração Pré-Classificação + SMM também reduziu ao máximo a métrica de FPR para todos os *thresholds* avaliados. Isso se deve a não pertubação do SMM diante a picos espúrios isolados de erro de reconstrução. Essa resposta do SMM também impactou TP e levou a deterioração do *recall* independente dos *thresholds*.

Para F1-Score, foi observado que a utilização de filtros não melhorou significativamente a detecção gerada por N99 e que as abordagens Pós-Classificação de aplicação geraram otimizações significativas para IQR, N95 e N97. Porém, foi observado que a abordagem Pós-Classificação foi mais promissora na redução de FP, sendo uma alternativa interessante para reduzir o número de falsos alarmes, como proposto por Ribeiro, Pereira e Gama (2016). Dentre as configurações testadas neste trabalho, foi observado que a configuração Pós-Classificação + EMA com o *threshold* N99 obteve o melhor F1-Score (20,31 $\pm$ 4,06).

6.2 SUGESTÕES PARA TRABALHOS FUTUROS

Para trabalhos futuros, é importante explorar outras formas de definição de *threshold* experimental para a abordagem Pós-Classificação, a exemplo da aplicação de algoritmos bio-inspirados para AD supervisionado. Além disso, outros valores determinísticos podem ser testados para gerar modelos otimizados quanto ao TP. Esses incrementos promovem mais autonomia e flexibilidade aos modelos de AD e a esta abordagem de filtro. Visando uma maior generalização dos resultados, também é interessante testar as abordagens avaliadas em outras bases de dados recorrentes na literatura, especialmente para o caso de detecção de sequências de anomalias, já que a aplicação de filtros apresentou efeito relevante para esta detecção. Também é interessante avaliar o efeito da aplicação de filtros em bases de dados sem sequências de anomalias para reforçar os resultados desta dissertação.

Também é interessante explorar a aplicação de teoria de resposta ao item com uma base de modelos não supervisionados com diferentes aplicações de filtro e *thresholds* e aproveitar o melhor que cada configuração tem a oferecer. A ideia seria criar um *ensemble* não supervisionado utilizando o conceito de teoria de resposta ao item já que técnicas convencionais de *ensemble* necessitam das etiquetas dos dados ou tabela verdade.

Por fim, é sugerido que sejam exploradas técnicas de explicabilidade agnósticas para tornar a detecção mais transparente para o usuário final dos modelos. Estas sugestões devem aumentar a credibilidade dos modelos de detecção de anomalias empregados. Além disso, a metodologia proposta pode ser aplicada a outros algoritmos não supervisionados de AD como *Isolation Forest* e *K-means* devido aos princípios de *score* de anomalia.

6.3 TRABALHOS PUBLICADOS

Os resultados parciais deste trabalho foram apresentados e publicados na forma de trabalho completo no congresso '2023: ANAIS DO XX ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC 2023)' que ocorreu entre os dias 25 e 29 de setembro de 2023 na cidade de Belo Horizonte. O artigo apresentado tem como título '*Decision Threshold Selection and Low-Pass Filter Application for Anomaly Detection with Sparse Autoencoder*' e foi apresentado no dia 29 de setembro de 2023 e publicado com o DOI: https://doi.org/10.5753/eniac.2023.234280 na base de trabalhos científicos SOL (SBC Open Lib).

Resultados finais de avaliação de filtro desenvolvidas nesta dissertação foram submetidas para a trilha do IJCNN (*The International Joint Conference on Neural Networks*) do IEEE WCCI 2024 (*IEEE World Congress on Computational Intelligence*) que ocorrerá na cidade de Yokohama (Japão) entre os dias 30 de junho e 5 de julho de 2024. O trabalho entitulado '*Low-Pass Filter Application for Anomaly Detection with Sparse Autoencoder*' foi aceito e será apresentado presencialmente na conferência.

REFERÊNCIAS

ALAMPAY, R. B.; ABU, P. Autocalibration of outlier threshold with autoencoder mean probability score. In: . New York, NY, USA: Association for Computing Machinery, 2020. (AICCC '19), p. 104–110. ISBN 9781450372633. Disponível em: <https://doi.org/10.1145/3375959.3375978>.

ANTWARG, L.; MILLER, R. M.; SHAPIRA, B.; ROKACH, L. Explaining anomalies detected by autoencoders using shapley additive explanations. *Expert Systems with Applications*, v. 186, p. 115736, 2021. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417421011155>.

AUDIBERT, J.; MICHIARDI, P.; GUYARD, F.; MARTI, S.; ZULUAGA, M. A. Do deep neural networks contribute to multivariate time series anomaly detection? *Pattern Recognition*, Elsevier, v. 132, p. 108945, 2022.

BLáZQUEZ-GARCÍA, A.; CONDE, A.; MORI, U.; LOZANO, J. A. A review on outlier/anomaly detection in time series data. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 54, n. 3, apr 2021. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3444690>.

BORGHESI, A.; BARTOLINI, A.; LOMBARDI, M.; MILANO, M.; BENINI, L. Anomaly detection using autoencoders in high performance computing systems. In: *The Thirty-First AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-19)*. AAAI, 2019. v. 33, p. 9428–9433. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/4993>.

CANCEMI, S.; Lo Frano, R.; SANTUS, C.; INOUE, T. Unsupervised anomaly detection in pressurized water reactor digital twins using autoencoder neural networks. *Nuclear Engineering and Design*, v. 413, p. 112502, 2023. ISSN 0029-5493. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0029549323003515>.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 41, n. 3, jul 2009. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/1541880.1541882>.

DAVARI, N.; VELOSO, B.; RIBEIRO, R. P.; PEREIRA, P. M.; GAMA, J. Predictive maintenance based on anomaly detection using deep learning for air production unit in the railway industry. In: *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2021. p. 1–10. ISBN 9781665420990. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9564181>.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, JMLR. org, v. 7, p. 1–30, 2006. Disponível em: <https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

HERBOLD, S. Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, The Open Journal, v. 5, n. 48, p. 2173, 2020. Disponível em: <https://doi.org/10.21105/joss.02173>.

LI, Z.; SUN, Y.; YANG, L.; ZHAO, Z.; CHEN, X. Unsupervised machine anomaly detection using autoencoder and temporal convolutional network. *IEEE Transactions on Instrumentation and Measurement*, v. 71, 2022. ISSN 15579662. Disponível em: <https://ieeexplore.ieee.org/document/9912415>.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. In: *2008 Eighth IEEE International Conference on Data Mining*. [S.l.: s.n.], 2008. p. 413–422.

NG, A. et al. Sparse autoencoder. *CS294A Lecture notes*, v. 72, n. 2011, p. 1–19, 2011. Disponível em: <https://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf>.

O'MALLEY, T.; BURSZTEIN, E.; LONG, J.; CHOLLET, F.; JIN, H.; INVERNIZZI, L. et al. *KerasTuner*. 2019. <https://github.com/keras-team/keras-tuner>.

PASTELL, M. *Measurements and Data Analysis for Agricultural Engineers using Python*. [S.l.]: Helsinki, Finland: Leanpub. com, 2016.

PERINI, L.; BüRKNER, P.-C.; KLAMI, A. Estimating the contamination factor's distribution in unsupervised anomaly detection. In: *Proceedings of the 40th International Conference on Machine Learning*. Honolulu, United States of America: Proceedings of Machine Learning Research, 2023. (PMLR 202), p. 27668–27679. Disponível em: <https://proceedings.mlr.press/v202/perini23a/perini23a.pdf>.

PURKAIT, N. *Hands-On Neural Networks with Keras: Design and create neural networks using deep learning and artificial intelligence principles*. [S.l.]: Packt Publishing Ltd, 2019.

RIBEIRO, R. P.; PEREIRA, P.; GAMA, J. Sequential anomalies: a study in the railway industry. *Machine Learning*, Springer New York LLC, v. 105, p. 127–153, 10 2016. ISSN 15730565. Disponível em: <https://link.springer.com/article/10.1007/s10994-016-5584-6>.

SHELLEY, M. *Frankenstein*. [S.l.]: Antofágica, 2023.

SUN, W.; SHAO, S.; ZHAO, R.; YAN, R.; ZHANG, X.; CHEN, X. A sparse auto-encoder-based deep neural network approach for induction motor faults classification. *Measurement*, v. 89, p. 171–178, 2016. ISSN 0263-2241. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0263224116300641>.

TUN, M. T.; NYAUNG, D. E.; PHYU, M. P. Network anomaly detection using threshold-based sparse. In: *IAIT2020: Proceedings of the 11th International Conference on Advances in Information Technology*. New York, NY, USA: Association for Computing Machinery, 2020. (IAIT2020). ISBN 9781450377591. Disponível em: <https://dl.acm.org/doi/10.1145/3406601.3406626>.

UMER, M. A.; JUNEJO, K. N.; JILANI, M. T.; MATHUR, A. P. Machine learning for intrusion detection in industrial control systems: Applications, challenges, and recommendations. *International Journal of Critical Infrastructure Protection*, v. 38, p. 100516, 2022. ISSN 1874-5482. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S1874548222000087>.

VELOSO, B.; RIBEIRO, R. P.; GAMA, J.; PEREIRA, P. M. The MetroPT dataset for predictive maintenance. *Scientific Data*, Nature Publishing Group UK London, v. 9, p. 764, 12 2022. ISSN 20524463. Disponível em: <https://www.nature.com/articles/s41597-022-01877-3>.

WANG, H.; BAH, M. J.; HAMMAD, M. Progress in outlier detection techniques: A survey. *IEEE Access*, v. 7, p. 107964–108000, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8786096>.

WEN, L.; GAO, L.; LI, X. A new deep transfer learning based on sparse auto-encoder for fault diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, v. 49, n. 1, p. 136–144, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8058000>.

ZHANG, W.; YANG, D.; WANG, H. Data-driven methods for predictive maintenance of industrial equipment: A survey. *IEEE Systems Journal*, v. 13, p. 2213–2227, 9 2019.