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

ELLEN CRISTINA XAVIER COELHO

Análise de desempenho de modelos de aprendizagem de máquina na classificação de séries temporais dos compostos voláteis fúngicos

Recife
2023

ELLEN CRISTINA XAVIER COELHO

Análise de desempenho de modelos de aprendizagem de máquina na classificação de séries temporais dos compostos voláteis fúngicos

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

Orientador: Prof. Leandro Maciel Almeida

Recife
2023

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

C672a Coelho, Ellen Cristina Xavier
Análise de desempenho de modelos de aprendizagem de máquina na classificação de séries temporais dos compostos voláteis fúngicos / Ellen Cristina Xavier Coelho – 2023.
137 f.: il., fig., tab.

Orientador: Leandro Maciel Almeida.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2023.
Inclui referências e apêndices.

1. Inteligência computacional. 2. Candida. 3. Nariz eletrônico. 4. Pré-processamento de séries temporais. I. Almeida, Leandro Maciel (orientador). II. Título

006.31 CDD (23. ed.) UFPE - CCEN 2024 – 010

Ellen Cristina Xavier Coelho

“Análise de desempenho de modelos de aprendizagem de máquina na classificação de séries temporais dos compostos voláteis fúngicos”

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: 27 de julho de 2023.

BANCA EXAMINADORA

_____

Prof. Dr. Cleber Zanchettin
Centro de Informática / UFPE

_____

Prof. Dr. Cicero Pinheiro Inácio
Departamento de Micologia / UFPE

_____

Prof. Dr. Leandro Maciel Almeida
Centro de Informática / UFPE
(**Orientador**)

AGRADECIMENTOS

Primeiramente, a Deus pelas oportunidades que foram inseridas em minha jornada até o presente momento. Eternamente grata a minha família por todo apoio durante minha formação, em especial ao meu pai, Erico Coelho, por ter sempre me incentivado a seguir em frente nos estudos.

Ao Prof. Dr. Leandro Maciel Almeida por todas as orientações ou contribuições prestadas ao desenvolver desta pesquisa, assim como aos conselhos pessoais e profissionais.

Aos professores da pós-graduação por todo o conhecimento compartilhado; a todas as pessoas que, de alguma forma, contribuíram para a minha formação acadêmica e que compartilham da minha alegria em concluir mais uma etapa.

Ao meu noivo Hugo Xavier, meus familiares e amigos, que em momentos em que precisei de ajuda estavam disponíveis para conversar e ajudar no que fosse necessário para alcançar os objetivos, em especial, Marina Bezerra e João Nunes.

À toda equipe do Laboratório de Micologia Médica Sylvio Campos pela disponibilização dos materiais necessários para a realização deste trabalho.

Por fim, à banca examinadora pelo tempo e disposição para ler e compreender o trabalho abordado e fornecer críticas a fim de melhorar o presente projeto.

RESUMO

Fungos são organismos essenciais para garantir a manutenção da vida no meio ambiente. Leveduras do gênero *Candida* fazem parte da microbiota do corpo humano, colonizando a pele e mucosas dos tratos digestivo e urinário, bucal e vaginal. Em condições normais, a maioria não causa danos ao indivíduo, entretanto, esses mesmos fungos podem apresentar características patogênicas. Atualmente, existem cerca de duzentas espécies do gênero *Candida*, no entanto, pouco mais de vinte são nocivas ao homem. O principal fator do desenvolvimento de infecções por essas leveduras, ocorre pelo rompimento do equilíbrio parasita-hospedeiro, desencadeado por alterações na barreira tecidual, alterações na microbiota normal e pelo comprometimento do sistema imunológico. Esse desequilíbrio pode causar desde doenças superficiais até invasivas, e, principalmente em hospedeiros debilitados, pode ser fatal. Dado a criticidade das infecções causadas por esses fungos, a identificação fúngica é fundamental para um tratamento eficaz. Métodos alternativos vêm sendo estudados e desenvolvidos com o objetivo de aliar a entrega de um resultado rápido e preciso a um baixo custo, e que não precise de um conhecimento especializado para operá-los. Uma tecnologia que se destaca entre esses métodos é a dos narizes eletrônicos (*e-nose*). Os dados produzidos por esses instrumentos, normalmente, possuem características complexas. Dessa forma, o objetivo deste trabalho é buscar por métodos e técnicas satisfatórios com menores custos computacionais, especificamente o tempo de processamento, para o processo de identificação das espécies de fungos *Candida*. Neste estudo, a metodologia envolveu a coleta de amostras de compostos orgânicos voláteis (COVs) emitidos por seis espécies do gênero *Candida* usando um nariz eletrônico. Os dados gerados foram submetidos a um processo de pré-processamento, incluindo técnicas de remoção de características de baixa variância e recursos altamente correlacionados. Em seguida, aplicou-se uma técnica de redução de dimensão chamada UMAP para melhorar a separação das classes. Cinco classificadores baseados em séries temporais foram treinados e avaliados usando validação cruzada. Os resultados foram analisados para avaliar o desempenho e a eficácia das técnicas utilizadas na identificação das espécies. O estudo abrangeu três experimentos, demonstrando a evolução das técnicas e o desempenho dos classificadores em cada estágio. O tempo de processamento e as métricas de desempenho, como acurácia, sensibilidade e precisão, foram usados para avaliar o sucesso da metodologia. No geral, o classificador TimeSeries Forest obteve os melhores resultados com a acurácia, sensibilidade de precisão de 100% e um tempo de processamento de 0.58 segundos. Como trabalhos futuros tem-se a expansão e a diversificação das espécies de leveduras clínicas do gênero *Candida* utilizadas para estudo, visando cobrir uma gama mais ampla de problemas e patologias, além disso, a criação e utilização de um ambiente controlado para realização de novos experimentos, a fim de minimizar erros causados por fatores externos.

Palavras-chave: *Candida*; nariz eletrônico; pré-processamento de séries temporais.

ABSTRACT

Fungi are essential organisms for maintaining life in the environment. Yeasts of the *Candida* genus are part of the human body's microbiota, colonizing the skin and mucous membranes of the digestive, urinary, oral, and vaginal tracts. Under normal conditions, most do not harm individuals, but these same fungi can exhibit pathogenic characteristics. Currently, there are approximately two hundred species of the *Candida* genus, but just over twenty are harmful to humans. The main factor in the development of infections by these yeasts is the disruption of the host-parasite balance, triggered by changes in tissue barriers, alterations in normal microbiota, and the compromise of the immune system. This imbalance can lead to diseases ranging from superficial to invasive, and can be fatal, especially in weakened hosts. Given the critical nature of infections caused by these fungi, fungal identification is essential for effective treatment. Alternative methods are being studied and developed to combine fast and accurate results with low cost, without the need for specialized knowledge to operate them. One technology that stands out among these methods is electronic noses (e-nose). Data produced by these instruments typically have complex characteristics. Therefore, the aim of this work is to seek satisfactory methods and techniques with lower computational costs, specifically processing time, for the identification of *Candida* fungal species. In this study, the methodology involved collecting samples of volatile organic compounds (VOCs) emitted by six *Candida* species using an electronic nose. The generated data underwent a pre-processing process, including techniques for removing low-variance features and highly correlated resources. Subsequently, a dimension reduction technique called UMAP was applied to improve class separation. Five time-series-based classifiers were trained and evaluated using cross-validation. The results were analyzed to assess the performance and effectiveness of the techniques used in species identification. The study encompassed three experiments, demonstrating the evolution of techniques and the performance of classifiers at each stage. Processing time and performance metrics, such as accuracy, sensitivity, and precision, were used to evaluate the success of the methodology. Overall, the TimeSeries Forest classifier achieved the best results with 100% accuracy, sensitivity, and precision, with a processing time of 0.58 seconds. As future work, there is the expansion and diversification of clinical yeast species of the *Candida* genus used for study, aiming to cover a broader range of problems and pathologies. Additionally, creating and utilizing a controlled environment for conducting new experiments is planned to minimize errors caused by external factors.

Keywords: *Candida*; electronic nose; pre-processing of time series.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizagem de Máquina
ANN	Artificial Neural Network
AP	Aprendizagem Profunda
CAWPE	Cross-validation Accuracy Weighted Probabilistic Ensemble
CG	Cromatografia a Gás
CG-EM	Cromatografia Gasosa acoplada a Espectrometria de Massas
CNN	rede neural convolucional
COVs	Compostos Orgânicos Voláteis
CV	Cross Validation
DCNN	Deep Convolutional Neural Network
DNN	Deep Neural Network
DP	Deep Learning
DrCIF	Diverse Representation Canonical Interval Forest
ECG	Eletrocardiograma
EEG	Eletroencefalograma
EM	Espectrometria de Massas
FFT	Fast Fourier Transform
FN	Falso Negativo
FP	Falso Positivo
FV	Fisher Vector
HIVE-COTE	Hierarchical Vote Collective of Transformation-based Ensemble
IA	Inteligência Artificial
kNN	k-Nearest Neighbor
LSTM	Long Short-Term Memory
MEG	Magnetoencefalografia
ML	Machine Learning
MMO	*Mixed Metal Oxides*
MOX	*Metal Oxide*
MTSC	Classificação de Séries Temporais multivariado
PCA	Principal Component Analysis

PCR	Reação em Cadeia da Polimerase
RBF	Radial-Basis Function
RDP	Random Dimension Permutation
RF	Random Forest
ROCKET	Random Convolutional Kernel Transform
SDRA	*Síndrome do Desconforto Respiratório Agudo*
SVM	Support Vector Machine
TapNet	Multivariate Time Series Classification with Attentional Prototypical Network
TDE	Temporal Dictionary Ensemble
TFS	Time Series Forest
TS	Times Series
TSC	Classificação de Séries Temporais
TVN	Taxa de Verdadeiros Negativo
TVP	Taxa de Verdadeiros Positivos
UFPE	Universidade Federal de Pernambuco
UMAP	Uniform Manifold Approximation and Projection
UTI	Unidade de Terapia Intensiva
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
WT	Wavelet Transform

SUMÁRIO

1 INTRODUÇÃO

As doenças infecciosas são conhecidas desde os tempos pré-históricos, porém passaram a ser registradas quando o homem começou a residir em comunidades. Durante séculos essas doenças foram um enigma para o homem, mas despertava grande interesse da comunidade científica (PETRUSEVSKI, 2013). Nesse contexto, segundo Pigatto *et al.*, as infecções fúngicas têm sido consideradas um problema de saúde pública, devido à alta prevalência e o aumento na incidência na última década (PIGATTO A; LOVISON, 2019).

Além disso, a recente pandemia global de COVID-19 predispôs um número relativamente alto de pacientes à *Síndrome do Desconforto Respiratório Agudo* (SDRA), que traz o risco de desenvolver superinfecções (ARASTEHFAR et al., 2020), evidenciando, assim, a continuidade dos estudos ligados a essas doenças. Essas infecções podem ser causadas por diferentes microrganismos, entretanto, os fungos do gênero *Candida* são considerados os agentes mais comum. Esses fungos podem provocar desde infecções superficiais até disseminadas, que acometem pessoas sadias e/ou imunocomprometidas, e se mostram de difícil terapia, com recorrências frequentes (PEIXOTO J.A; ROCHA1, 2014).

A candidíase consiste em uma extensa variedade de síndromes clínicas causadas por *Candida*, constituído de aproximadamente 200 espécies diferentes de leveduras, que vivem normalmente nos mais diversos nichos corporais (PEIXOTO J.A; ROCHA1, 2014). Espécies de leveduras pertencentes ao gênero *Candida*, incluindo *Candida albicans*, *Candida glabrata*, *Candida parapsilosis*, *Candida tropicalis* e *Candida krusei*, são as espécies fúngicas mais prevalentes que fazem parte da microbiota normal dos indivíduos saudáveis. Todavia, quando há uma ruptura no balanço normal da microbiota ou o sistema imune do hospedeiro encontra-se comprometido, as espécies do gênero *Candida* tendem a manifestações agressivas, tornando-se patogênicas (ARASTEHFAR et al., 2020).

As manifestações clínicas causadas por *Candida* spp são variadas, e podem ser classificadas como superficiais ou profundas, ocasionando desde alterações na qualidade de vida dos indivíduos até lesões graves, frequentemente fatais (Sá, 2017). Por isso, definir corretamente o diagnóstico da candidíase é um dos pontos de maior importância para o sucesso terapêutico (PEIXOTO J.A; ROCHA1, 2014).

Em laboratórios clínicos, infecções fúngicas podem ser detectadas com base em métodos convencionais, como cultura de células e após identificação por métodos fenotípicos, imunológicos e/ou genotípicos (CHERKAOUI et al., 2010). No entanto, o diagnóstico dessas condições pode ser problemático, tanto por ausência de sintomas clínicos específicos, quanto pela dificuldade em se obter resultado pelos métodos diagnósticos tradicionais (BORGES, 2009), pois, embora a cultura e a microscopia continuem sendo as técnicas padrão para diagnóstico, a sensibilidade e a especificidade desses métodos são limitadas (COSTA C. P., 2020). Atualmente, o método padrão-ouro para o diagnóstico de uma infec-

ção de corrente sanguínea por fungo é baseada na detecção direta do agente em cultivos sanguíneos (hemocultura). Entretanto, este método se caracteriza pela baixa sensibilidade e demora de, no mínimo, dois dias para detecção e uma semana para identificação (SIQUEIRA J. P. Z.; ALMEIDA, 2018).

Relatos recentes de patógenos fúngicos emergentes, *Candida auris* (ALLAW et al., 2021) e *Candida blankii* (JR JOãO NOBREGA DE ALMEIDA; SILVIA V, 2018), enfatizam essa dificuldade de identificação por diferentes métodos de diagnóstico, retratam a preocupação com a resistência dos agentes aos medicamentos, bem como problemas com a erradicação. Assim, o diagnóstico rápido e específico é fundamental para a adequada introdução terapêutica e para evitar a transmissão do agente.

Sabe-se que os microrganismos produzem uma ampla gama de produtos metabólicos, que são os produtos finais de processos bioquímicos e resultado de interações ambientais e genéticas (REES et al., 2018). Segundo Costa *et al.*, através desses metabólitos voláteis, tratados como impressões digitais únicas de cada espécie, é possível entender melhor sistemas biológicos complexos e dar suporte uma nova abordagem na pesquisa microbiana, chamada de diagnóstico microbiano (COSTA C. P., 2020).

Um instrumento, conhecido como nariz eletrônico (*e-nose*), vem desempenhando um papel cada vez maior na detecção de gases de propósito geral em muitas aplicações como (YAN XIUZHEN GUO; ZHANG, 2015): análise de odores, controle de qualidade da indústria de alimentos, proteção do meio ambiente, saúde pública, detecção de explosivos e aplicações em voos espaciais. Esse instrumento compreende um conjunto de sensores químicos eletrônicos com especificidade parcial e um sistema de reconhecimento de padrão apropriado, tornando-o capaz de captar e detectar Compostos Orgânicos Voláteis (COVs) (J.W., 1994). Dessa maneira, pode-se utilizá-lo como principal instrumento de estudo no levantamento de perfil de metabólitos voláteis de micro-organismos como a *Candida* spp.

1.1 OBJETIVOS

Este trabalho propõe a utilização de um protótipo de um nariz eletrônico para aquisição dos COVs das espécies de *Candida* mais comum: *C. Albicans, C. Glabrata, C. Haemulonii, C. Kodamaea ohmeri, C. Krusei, C. Parapsilosis* e *C. Tropicalis*. O objetivo principal deste trabalho é explorar dados gerados, através do mapeamento de metabólitos em conjuntos de observações, a fim de levantar características e elaborar estratégias que facilitem a identificação de *Candida* spp, e assim, propor um método auxiliar no processo de diagnóstico. Com base nisto, os objetivos específicos foram traçados:

- Criar a base de dados, através da coleta das leituras geradas pelos sensores do protótipo de um nariz eletrônico;
- Compreender o comportamento do dados coletados, analisando as características para cada espécie;

- Elaborar e explorar as séries temporais geradas pelos dados coletados;
- Aplicar técnicas e métodos de processamento digital de sinais, englobando as etapas de pré-processamento, extração de atributos, seleção de atributos e classificação;
- Avaliar o desempenho das técnicas e métodos utilizados, verificando quais combinações produzem os melhores resultados.

1.2 ORGANIZAÇÃO DO TRABALHO

O presente trabalho está divido nos seguintes capítulos:

1 INTRODUÇÃO: apresenta um introdutório de apresentação ao estudo e aos objetivos desta dissertação;

2 REFERENCIAL TEÓRICO: descreve a natureza dos dados que serão abordados e trabalhados, assim como apresenta o conceito base relacionado a Séries Temporais, a técnicas de pré-processamento, além de abordar os conceitos de Aprendizagem Profunda (AP) (em inglês, Deep Learning (DP)) que serão utilizados nos capítulos subsequentes.

3 REVISÃO DA LITERATURA: aborda trabalhos na literatura que estão relacionados ao diagnóstico fúngico, assim como estudo e análise dos COVs, e a outros dados de natureza semelhante.

4 MATERIAIS E MÉTODOS: explana sobre as estratégias e abordagens elaboradas e aplicadas nesta dissertação.

5 ANÁLISE EXPERIMENTAL: descreve e discute os resultados dos experimentos efetuados para análise e avaliação das técnicas trabalhadas.

6 CONCLUSÃO: apresenta as considerações finais sobre os principais tópicos abordados nesta dissertação, incluindo as contribuições alcançadas e as indicações de trabalhos futuros.

Por fim, anexos a este documento, têm os Apêndices A e B, nos quais são apresentados os registros da modificação da resistência dos sensores detectada pelo *e-nose* quando exposto a um conteúdo gasoso para, respectivamente, um único ciclo por espécie e todos os ciclos por espécie.

2 REFERENCIAL TEÓRICO

Neste capítulo, faz-se um breve estudo sobre os fungos do gênero *Candida*, abordando o interesse médico, as características e os principais métodos de identificação das diferentes espécies deste gênero, com descrições dos fungos cobertos pelo trabalho. Além disso, aborda uma introdução à Inteligência Artificial e uma breve discussão sobre séries temporais, além das métricas utilizadas para mensurar os resultados desta pesquisa.

2.1 CANDIDA: FUNGO DE INTERESSE MÉDICO

Segundo Sidrim e Rocha (SIDRIM J. J. C.; ROCHA, 2004), a primeira documentação de leveduras do gênero *Candida* spp. como patógeno é atribuída a Langenbeck, que em 1839 observou e isolou, da cavidade oral de um paciente com afta bucal, um micro-organismo, que atualmente é a mais importante levedura patogênica do homem, a *Candida albicans*. As infecções causadas por leveduras do gênero *Candida* dependendo do local da lesão, podem ser classificadas de duas formas distintas: candidíase superficial ou de mucosa e candidíase profunda ou sistêmica (BARBEDO L. S.; SGARBI, 2010).

A candidíase superficial é a infecção mais comum dentre as candidíases, acometendo a pele e mucosas, e normalmente é causada por *C. albicans*, que é a espécie comensal mais comum na boca, vagina e trato gastrointestinal de indivíduos saudáveis. Dentre as manifestação causadas por tal espécie destaca-se a candidíase pseudomembranosa, conhecida como sapinho, que acomete principalmente recém-nascidos e pacientes com imunidade muito baixa, até mesmo pessoas de maior idade (CREPALDI, 2021); e a candidíase vulvovaginal, que é considerada a segunda causa mais comum de infecção genital em mulheres em idade reprodutiva. Aliás, estudos mostram que 70-75% das mulheres, em algum momento de sua vida, vão apresentar pelo menos um caso desta infecção (CREPALDI, 2021).

Já a candidíase sistêmica acontece preferencialmente em pacientes com a imunidade comprometida, onde o microrganismo se dissemina através do sangue, e pode se instalar em órgãos vitais como cérebro, coração e rins (PEIXOTO J.A; ROCHA1, 2014). Em 2021, no Brasil, segundo o Ministério da Saúde, a taxa de incidência chegava a 2,49 casos de candidemia [1] por 1.000 admissões hospitalares, nos hospitais públicos terciários, que corresponde a uma taxa de 2 a 15 vezes maior que relatadas em países da Europa e o do EUA.

Segundo a Anvisa, alguns fatores são reconhecidos como risco para infecção invasiva por *Candida* (ANVISA, 2013), como:

- Permanência $>$ 4 dias em Unidade de Terapia Intensiva (UTI);

[1] Infecção de corrente sanguínea causada por leveduras do gênero Candida.

- Antibioticoterapia de largo espectro;
- Cirurgia abdominal;
- Cateterização venosa central;
- Nutrição parenteral total;
- Imunodepressão;
- Índice APACHE II > 10;
- Ventilação mecânica > 48h
- Neutropenia;
- E quimioterapia citotóxica.

Quando se trata de infecção generalizada grave do organismo, causada por microrganismos patogênicos (fungos, bactérias ou vírus), o gênero *Candida* é relatado como a terceira causa em geral no mundo (PAPPAS P. G., 2018). Além disso, num estudo realizado em 2016, (LI, 2016), que avaliava 190 pacientes com candidemia, a taxa de mortalidade desses pacientes hospitalizados foi de 27,9% num período de 30 dias após a coleta da amostra de sangue, sendo 16,7% com apenas 7 dias. Sendo assim, as manifestações clínicas causadas por *Candida* spp variam desde uma infecção leve até potencialmente fatal, e isso enfatiza o interesse médico neste micro-organismo.

2.2 ASPECTOS GERAIS DO GÊNERO CANDIDA

O reino Fungi abrange um grande grupo de organismos eucarióticos[2], que inclui diversas formas, como leveduras, bolores, mofos, cogumelos, fungos gelatinosos e orelha-de-pau (SIDRIM J. J. C.; ROCHA, 2004). Os fungos são seres dispersos no meio ambiente, em vegetais, ar atmosférico, solo e água e, embora sejam estimados em 250 mil espécies, menos de 150 foram descritos como patógenos aos seres humanos (ANVISA, 2013).

Os fungos de interesse médico são de dois tipos morfológicos: leveduras, que são unicelulares ou fungos filamentosos, que são multicelulares, a Figura 1, ilustra as estruturas microscópicas básicas desses fungos. As leveduras têm como estrutura primária, células que se reproduzem por brotamento, único ou múltiplo, em geral, de forma arredondada. Estas células são esporos de origem assexual e se denominam blastoconídios. Os fungos filamentosos possuem como elemento constituinte básico a hifa, que pode ser septada ou não septada.

[2] Os seres eucarióticos presentam uma membrana nuclear que envolve o material nuclear como os cromossomos e o nucléolo.

Figura 1 – Estruturas microscópicas básicas de fungos: filamentosos, possuem como elemento constituinte básico a hifa (a, b, c); e as leveduras, têm como estrutura primária, células que se reproduzem por brotamento, único ou múltiplo, em geral, de forma arredondada (d).

Fonte: (ANVISA, 2004).

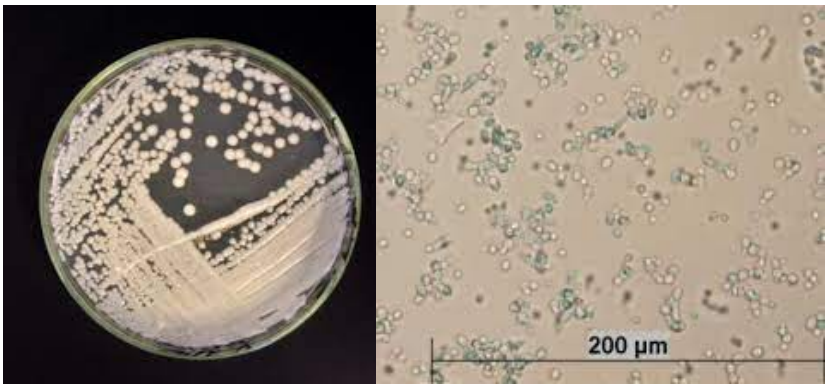
Esses conceitos fundamentais representam a base para a identificação de um fungo, pois a classificação de filamentosos é feita, em regra, pelas características morfológicas, tanto macroscópicas (cor, aspecto, textura da colônia, etc.), quanto microscópicas (forma e cor da hifa, presença ou não de septos, tipo e arranjo de esporos, etc.), além da velocidade de crescimento (lenta, moderada ou rápida). A identificação de leveduras, ao contrário, é feita, principalmente, por características fisiológicas, desde que, a morfologia destes fungos não é muito variada e não permite distinção entre espécies e, em regra, entre gêneros.

Os fungos pertencentes ao gênero *Candida* são fungos diploides e polimórficos, podendo apresentar estruturas leveduriformes[3] hialinas, com formação de blastoconídios, pseudo-hifas e em algumas circunstâncias pode apresentar hifas verdadeiras. As espécies do gênero *Candida*, geralmente, crescem bem em meios de composição relativamente simples como o Sabouraud dextrose, desenvolvendo colônias com um aspecto característico, como ilustrado na Figura 2. Na sua maioria, produzem colônias glabras (lisas), de coloração branca ou bege, úmida e cremosa ou, às vezes, rugosa e seca, com odor característico (BEZERRA, 2015).

Apesar da *Candida albicans* receber destaque nessa introdução, nas últimas décadas, outras espécies denominadas de não-albicans, tem mostrado aumento relevante em manifestações infectuosos fúngicas como: *C. tropicalis, C. parapsilosis, C. krusei, C. glabrata, C. rugosa, C. guilliermondi, C. lusitaniae, C. haemulonii, C. lipolytica, C. kodamaea ohmeri*, entre outras (GIOLO MURIEL PADOVANI; SVIDZINSKI, 2010).

[3] Colonia de funfos com aspecto pastoso ou cremoso

Figura 2 – Características macro e micromorfológicas de *Candida* spp. Na sua maioria, produzem colônias glabras (lisas), de coloração branca ou bege, úmida e cremosa ou, às vezes, rugosa e seca. (a) Colônias em meio Sabouraud dextrose; (b) Blastoconídios, preparação com lacto fenol azul algodão



(a) (b)

Fonte: (RODRIGUES, 2013).

2.3 IDENTIFICAÇÃO DE ESPÉCIES DE CANDIDA SPP.

O diagnóstico laboratorial das infecções fúngicas inclui a observação direta das amostras biológicas para a pesquisa de elementos fúngicos, a cultura celular, testes bioquímicos e eventualmente testes serológicos e/ou observações histopatológicas. A observação direta é um método rápido e de baixo custo mas, na maioria dos casos, não permite a identificação da espécie ou mesmo do gênero dos fungos que não apresentem estruturas morfológicas parasitárias típicas, como os pertencentes aos gêneros *Candida* (PEREIRA, 2010).

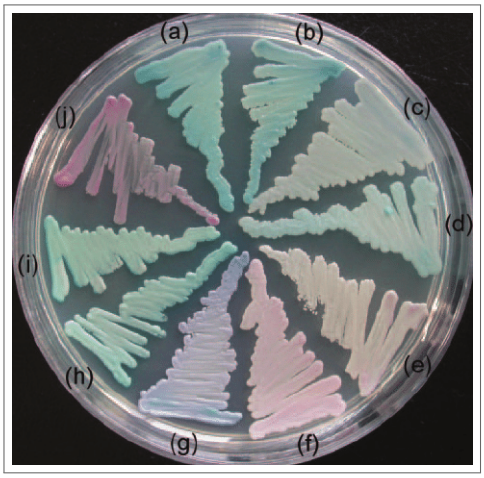
Os primeiros testes para elucidação das espécies (ou gênero) de leveduras eram baseados nas singularidades do crescimento de colônias quando em meio de cultivo, filamentação em cultivo em lâmina, diferenciação do crescimento do tubo germinativo, além de teste de assimilação de fontes de carbono e nitrogênio (auxonograma) e de fermentação de carboidratos (ziminograma). Atualmente, esses testes ainda são utilizados, contudo a maioria permite apenas a identificação do gênero e possui baixa sensibilidade, além de risco de contaminação das culturas (KOEHLER et al., 2016).

Para estes testes, normalmente os fungos são isolados por plaqueamento, ou seja, uma amostra previamente recolhida (por exemplo a partir de solo, líquidos, superfícies e do ar) é disposta numa placa de *Petri* com meio de cultura próprio para o seu crescimento. O plaqueamento pode ser realizado, por diluição da amostra em água ou em solução salina de baixa concentração e posterior espalhamento na placa de meio de cultura. Os fungos crescem onde existe matéria orgânica disponível, viva ou morta, geralmente adaptando-se ao calor e umidade. Assim, para os fungos cresçam na placa de *Petri*, é indispensável que os meios de cultura possuam nutrientes básicos que alimente-os, no entanto, há outros fatores externos que afetam seu crescimento, como o pH, a temperatura, a umidade e a luz (TROVÃO J.; PEREIRA, 2019).

2.3.1 Cultura e identificação no meio CHROMagar

Em 1994, Odds e Bernaerts, ao tentar encontrar um teste adequado para a detecção, e que possuísse ampla utilização, testaram e validaram o meio de cultivo CHROMagar Candida desenvolvido pela empresa *CHROMagar Company* (Paris, França). Com o CHROMagar Candida, um teste fenotípico tradicional, foi possível diferenciar, com base na coloração e morfologia das colônias, as espécies *C. albicans, C. krusei e C. tropicalis* provenientes de diferentes amostras biológicas, incluindo amostras de sangue (sem interferência do pigmento heme presente nas hemácias), após 48h de incubação a 37°C (JODDS, 1994). A figura a seguir ilustra colônias de diferentes gêneros de *Candida* cultivadas em CHROMagar Candida por 48 h a 30°C.

Figura 3 – Colônias cultivadas em CHROMagar Candida por 48 h a 30°C: (a) C. pseudorugosa XH 1026, (b) C. pseudorugosa XH 1164, (c) C. rugosa CBS 613, (d) C. rugosa AS 2.1498, (e) C. parapsilosis ATCC 22019, (f) C. krusei AS 2.3194, (g) C. tropicalis AS 2.3195, (h) C. dublinensis CBS 7988, (i) C. albicans ATCC 90028 e (j) C. glabrata ATCC 90030



Fonte: (LE; XU; BAI, 2007).

A realização da cultura fúngica envolve dois passos: coleta e semeadura. O primeiro passo é a coleta dos materiais biológicos, como fezes, sangue, urina, saliva, catarro ou até mesmo células de um órgão que esteja contaminado. O material coletado deve ser representativo do processo infeccioso investigado, evitando-se contaminação com as áreas adjacentes. O segundo passo é a semeadura em meios de cultura específico, procede-se com o repique em placa de *Petri* ou ágar, onde a amostra permanece incubada (ANVISA, 2013). A hemocultura é o padrão ouro para o diagnóstico de infecções da corrente sanguínea e depende de técnicas convencionais subsequentes, incluindo coloração de Gram, subcultura seguida por testes bioquímicos e teste de sensibilidade aos antifúngicos. Além disso, todo o procedimento leva em média de 48 a 72 horas para ser concluído e possui limitações, já que essas metodologias necessitam de processos metabólicos do micro-organismo (SOUZA, 2021).

Além disso, apesar de ser uma alternativa para diagnóstico, o método possui o ônus de necessitar de pessoal capacitado na interpretação dos resultados, além de não excluir

a utilização de testes complementares para identificação das espécies, incluindo outros meios contendo substratos cromogênicos (KOEHLER et al., 2016). Com a necessidade de rapidez no diagnóstico e reconhecimento das espécies de *Candida*, aliada com o surgimento de novas tecnologias de biologia molecular, pesquisadores desenvolveram um método de identificação de amplo espectro, aplicado diretamente em amostras clínicas: o método baseado na Reação em Cadeia da Polimerase (PCR), onde a DNA-polimerase sintetiza fragmentos de DNA após a hibridização de sequências específicas (*primers*) (KOEHLER et al., 2016).

2.3.2 Identificação de fungos por meio do DNA

A biologia molecular se desenvolveu durante o século XX, trazendo grandes avanços para a compreensão acerca do desenvolvimento celular, da replicação do DNA e funcionamento dos genes. A revolucionária técnica da PCR permitiu o estudo de sequências de ácidos nucléicos, proporcionando avanços suntuosos para áreas como a genética, medicina forense, sequenciamento do genoma humano e microbiano (GARCIA, 2018).

A técnica de Reação em Cadeia da Polimerase (PCR) é amplamente empregada na identificação de fungos, o processo baseia-se na síntese enzimática *in vitro* de cópias de fragmentos específicos de DNA, em que a partir de uma única molécula do ácido nucléico é possível gerar bilhões de moléculas similares em uma reação, imitando assim a replicação natural do DNA (MULLIS, 1990). Essas cópias podem então ser analisadas, frequentemente por meio de técnicas como sequenciamento para identificar a espécie de fungo.

As vantagens do PCR na identificação de fungos incluem a alta sensibilidade, que permite a detecção de fungos mesmo em pequenas quantidades de material genético, a rapidez do processo e a capacidade de automatização. No entanto, as desvantagens incluem a necessidade de equipamentos especializados e pessoal treinado, o risco de contaminação cruzada em laboratórios, e a possibilidade de não diferenciar entre espécies muito semelhantes, exigindo técnicas adicionais para confirmação(CASADO LARISSA DA SILVA ; GOMES, 2022).

Já o sequenciamento de DNA é uma técnica poderosa que envolve a determinação da ordem exata das bases nucleotídicas nas sequências de DNA de fungos, onde fragmentos são sequenciados usando técnicas como o sequenciamento de próxima geração (NGS) ou o sequenciamento de Sanger. As sequências resultantes são comparadas a bancos de dados genômicos para identificar as espécies de fungos presentes. Essa abordagem oferece alta precisão na identificação de fungos, mesmo quando características morfológicas não são distintivas o suficiente. No entanto, requer equipamentos especializados, recursos computacionais para análise de dados e expertise técnica (ZAMPARETTE, 2017).

A Tabela 1 resume as principais vantagens e desvantagens de cada método apresentado acima.

Tabela 1 – Comparação de metodologias convencionais para identificação fúngica

Metodologia	Técnicas baseadas na morfologia dos fungos (Meio CHROMagar)	Reação em Cadeia da Polimerase (PCR)	Sequenciamento de DNA
Vantagens	• Ser fácil e de rápida preparação; • Permite a identificação de acordo com a coloração que o fungo apresenta no meio semeado (NASCIMENTO, 2016).	• Altamente sensível; • Pequenas quantidade de sequências de DNA ou RNA específicas para possibilitar a identificação; • Resultados levam de 4 a 8 horas (CASADO LARISSA DA SILVA ; GOMES, 2022).	• Alta sensibilidade e precisão; • Precisa de uma quantidade mínima de DNA para possibilitar a identificação; • Permite identificar vários microrganismos em uma única amostra. • É automatizado, com resultados em 6 a 8 horas; (ZAMPARETTE, 2017)
Desvantagens	• Cultivo da amostra para isolamento do fungo em um meio de cultura específico; • Um crescimento adequado do fungo varia de 1 a 30 dias dependendo dos fatores de cultivos e intrínsecos do fungo; (NASCIMENTO, 2016); • Não automatizado, necessário especialista para operá-las.	• Baixa precisão e resolução; • Resultados não são expressos como números (CASADO LARISSA DA SILVA ; GOMES, 2022).	• Difícil implementação em laboratórios de microbiologia; • Alto custo dos consumíveis. (ZAMPARETTE, 2017)

Fonte: A autora (2023).

2.4 COMPOSTOS ORGÂNICOS VOLÁTEIS

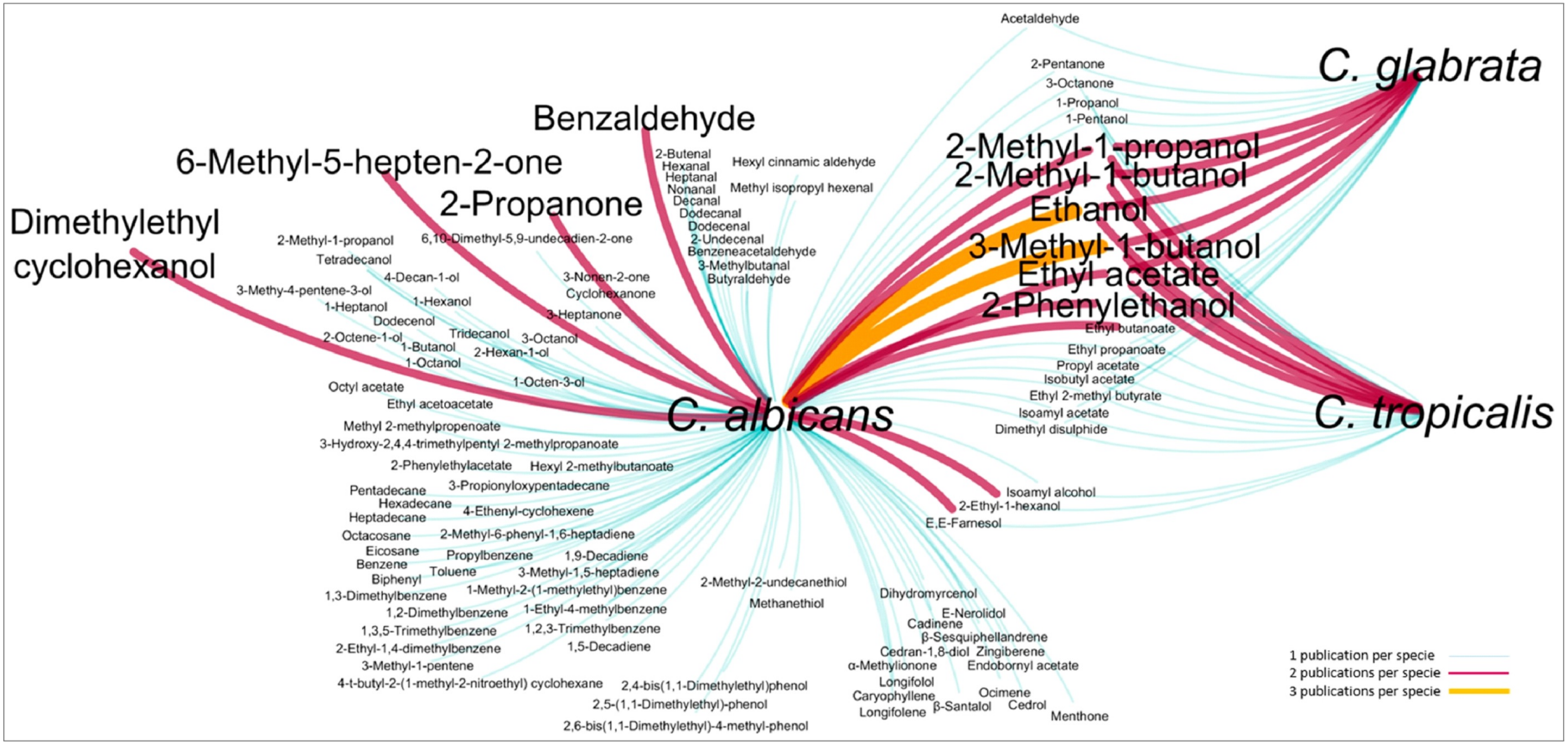
Os Compostos Orgânicos Voláteis (COVs) são substâncias orgânicas de baixa massa molecular que são facilmente vaporizadas à temperatura ambiente. Geralmente, tem de baixa a média solubilidade em água e apresentam odor característico (ANDREAS, 2010). De acordo com Moraes *et al.*, mais de 300 COVs já foram identificados em fungos e pertencem a diversas classes químicas tais como hidrocarbonetos simples, álcoois, aldeídos, cetonas, ésteres, terpenos, compostos aromáticos e outros derivados benzênicos (MORAES G. K. A.; FERRAZ, 2020).

A composição dos COVs produzido por fungos é dependente da espécie do microorganismo, do substrato, tempo de incubação, tipo de nutrientes, temperatura e outros parâmetros ambientais. Os fungos além de serem capazes de produzir uma grande variedade de compostos voláteis, também são capazes de metabolizá-los (KORPI; JäRNBERG; PASANEN, 2009). Os COVs fúngicos são derivados tanto das vias metabólicas primárias quanto das secundárias, principalmente da oxidação metabólica da glicose e são provenientes de diferentes precursores, como acetato, aminoácidos, ácidos graxos e cetoácidos (MORAES G. K. A.; FERRAZ, 2020).

Um estudo recente realizou um levantamento bibliográfico dos metabólicos celulares de C. *albicans* e não-C. *albicans* (C. *glabrata* e C. *tropicalis*)(COSTA C. P., 2020). Esse estudo foi capaz de explorar em profundidade os metabólitos voláteis dessas espécies e identificá-los em uma fração volátil que compreende 126 metabólitos, distribuídos, segundo ele, em várias famílias químicas: ácidos, álcoois, aldeídos, hidrocarbonetos, ésteres, cetonas, compostos monoterpênicos e sesquiterpênicos, norisoprenóides, fenóis e compostos sulfurados. A Figura 4 apresenta esses metabólitos, e a quantidade de publicações que os relatam.

Em 2015, Hertel *et al.*, seguindo a abordagem anterior, foi capaz distinguir *Candida albicans, glabrata, krusei e tropicalis*, através voláteis de assinatura (HERTEL et al., 2015). Os meios de crescimento foram analisados após 8 e 24 h usando cromatografia gasosa off-line e espectrometria de massa. A identificação de voláteis de assinatura foi assistida usando vários bancos de dados microbianos. Os padrões voláteis recuperados permitiram a discriminação de espécies de *Candida in vitro*. Para *C. albicans* 3-metil-2-butanona e estireno e para *C. krusei* uma combinação de p-xileno, 2-octanona, 2-heptanona e acetato de n-butila foram encontrados como específicos. O 1-hexanol foi encontrado em *C. tropicalis*, mas é emitido por uma variedade de outros microrganismos. *C. glabrata* foi caracterizada pela ausência desses voláteis.

Figura 4 – Representação visual dos metabólitos voláteis relatados na literatura para C. *albicans*, C. *glabrata* e C. *tropicalis*. Os nós representam espécies de *Candida*, e as linhas fazem a ligação entre espécies e metabólitos, alguns deles comuns a 2 ou 3 espécies. A cor e a espessura da linha representam o número de citações de cada metabólito por espécie: azul claro - 1 publicação, violeta - 2 publicações e laranja - 3 publicações publicações



Fonte: (COSTA C. P., 2020).

2.5 SENSORES DE GASES

Resistividade elétrica é uma das grandezas mais utilizadas em dispositivos de instrumentação. Ela representa a caraterística que um determinado material tem de dificultar a passagem de um fluxo de cargas elétricas através dele. Vários dispositivos eletrônicos utilizam a variação da magnitude da resistividade elétrica de um material como forma de medir alterações ou modificações de propriedades físicas, químicas ou biológicas (ROSSETTO, 2021).

Os sensores resistivos de gás auxiliam a identificar e quantificar seus níveis no ambiente, além de qualidade do ar e a umidade relativa (OLIVEIRA, 2021). Os mais comumente utilizados para sensoriamento de gases são, basicamente, os cromatógrafos e os sensores semicondutores de óxido metálico - *Metal Oxide* (MOX) ou óxido multimetal - *Mixed Metal Oxides* (MMO), que se baseiam em variações no parâmetro de resistência elétrica, através de transduções indireta e direta, respectivamente. Contudo, suas empregabilidades baseiam-se na aplicação de técnicas de aferição distintas, e são muito dispendiosas economicamente e/ou energeticamente (ROSSETTO, 2021).

A cromatografia gasosa é um método usado há mais tempo e que se baseia em função da variação de condutividade térmica existente entre as moléculas gasosas. Para realizar a medição, o dispositivo, que utiliza esta técnica, separa os diferentes gases entre suas colunas (duas ou mais) e mede o gradiente térmico existente entre cada uma delas. Como característica desses sensores, sua medição é feita de forma indireta, ou seja, nesse caso, se faz necessário uma transdução inicial de química para térmica e, por fim, de térmica para elétrica (MARCELLIS A.; FERRI, 2011).

Já os dispositivos MOX são sensores constituídos de uma camada de detecção baseada em materiais semicondutores. Esses sensores têm sido amplamente estudados devido à sua estrutura simples, ao baixo custo e abundância na crosta terrestre (OLIVEIRA, 2021). Estes materiais apresentam alta sensibilidade para diferentes analitos[4] (redutores e oxidantes), versatilidade, boa estabilidade e baixo custo de produção que facilita sua integração em dispositivos sensoriais (OLIVEIRA, 2021). A Figura 5 ilustra diferentes sensores de gás comerciais fabricados pela Figaro e FIS, com diferentes tamanhos e configuração de pinos.

Além disso, cada sensor MOX é projetado para detectar uma propriedade específica da substância odorante (ARAUJO; GAMBOA; SILVA, 2020). Em geral, os sensores de gás são dispositivos constituídos por duas partes principais, a primeira é um elemento ativo que altera suas propriedades físicas ou químicas na presença daquele que detecta e a segunda parte é um transdutor, que converte as alterações no propriedades do elemento ativo em um sinal elétrico. Esses sensores normalmente possuem uma membrana seletiva, impedindo a passagem de partículas ou materiais indesejados, atuando como um primeiro filtro de ruído (GAMBOA; ALBARRACIN-ESTRADA; DELGADO-TREJOS, 2011). A Figura 6

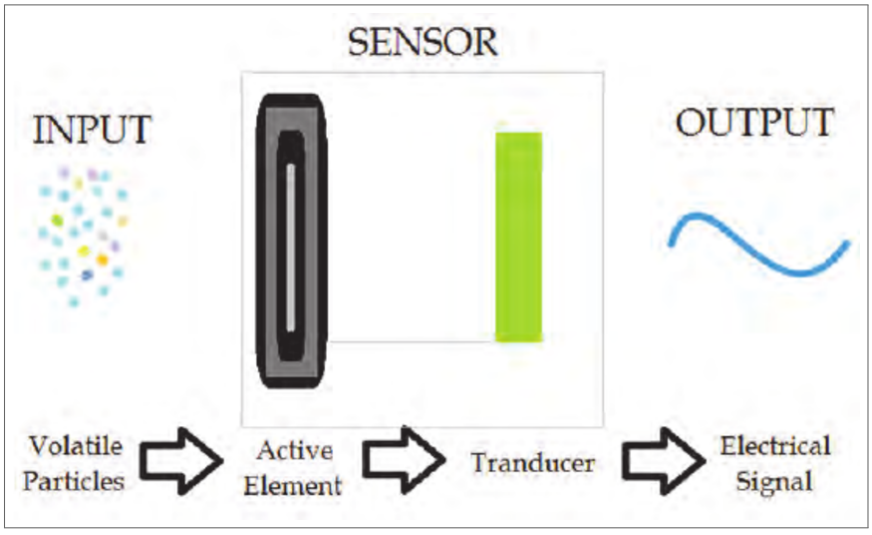
[4] Substância ou Componente químico de uma amostra que é alvo de análise ou tem interesse para uma análise.

Figura 5 – Sensores de gás comerciais fabricados pela Figaro e FIS, com diferentes tamanhos e configuração de pinos

Fonte: (GAMBOA; ALBARRACIN-ESTRADA; DELGADO-TREJOS, 2011).

mostra um diagrama simplificado de um dispositivo deste tipo, no qual podem ser vistas as partes principais de um sensor de gás e a natureza das entradas e saídas.

Figura 6 – Diagrama esquemático simplificado de um sensor de gás. Existem diferentes tipos de sensores de gás. Seu princípio de funcionamento baseia-se na alteração da condutividade de um material sensível quando este absorve ou reage com os gases do ambiente

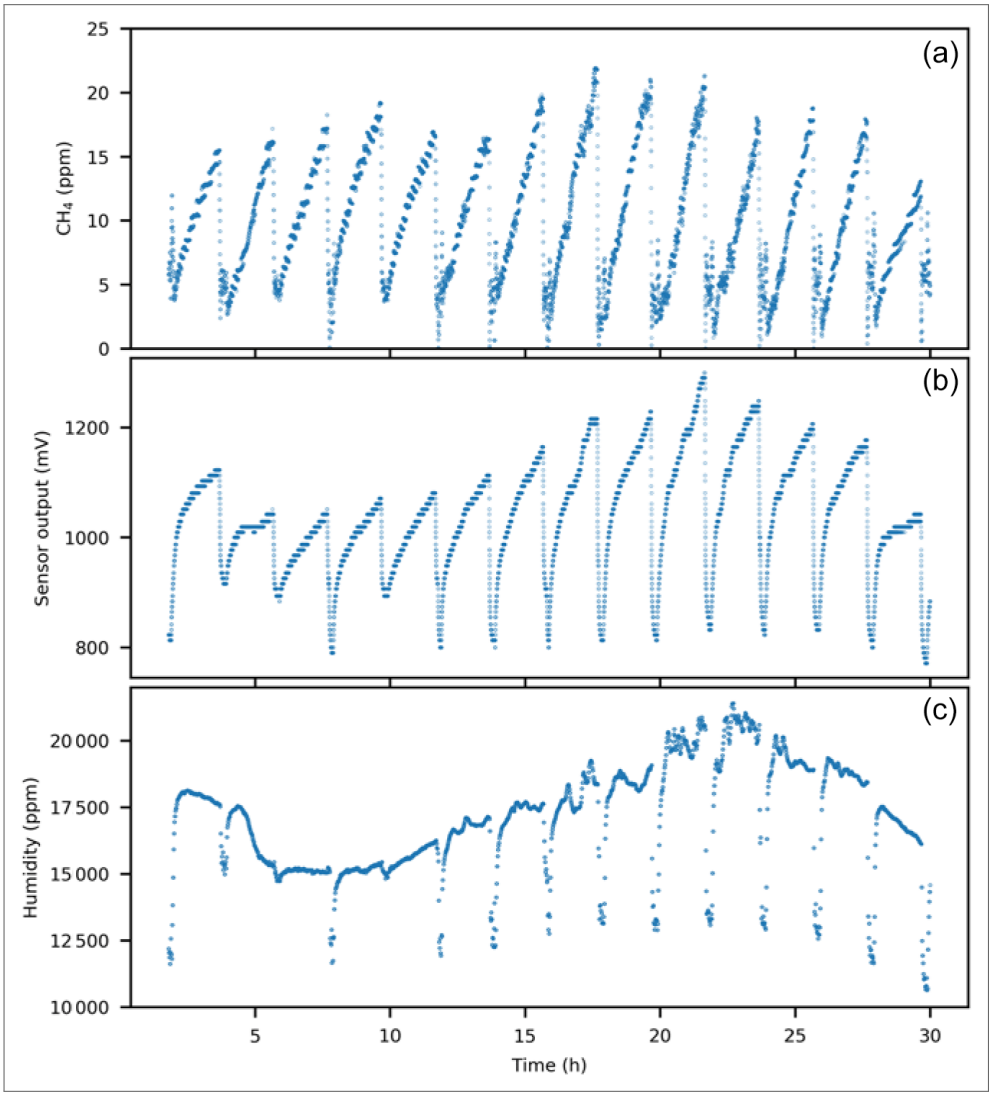


Fonte: (GAMBOA; ALBARRACIN-ESTRADA; DELGADO-TREJOS, 2011).

Assim, os diferentes valores de resistência que cada substância desencadeia nos sensores MOX formam séries temporais (ARAUJO; GAMBOA; SILVA, 2020). O sensor TGS2611, responsável por detectar o metano (CH4), é comumente utilizado em estudos para entender o comportamento gases de efeito estufa (BASTVIKEN et al., 2020). A Figura 7 ilustra os dados reais de uma câmara de fluxo em um lago no ano de 2019. Observa-se nesta

figura a formação da série temporal, dados não tratados, como resposta do sensor MOX.

Figura 7 – Dados reais de uma câmara de fluxo em um lago em junho de 2019 com 14 ciclos de abertura-fechamento automatizados da câmara ao longo de 30h. Os painéis mostram (a) frações molares de metano, (b) sinal de saída do sensor não tratado e (c) umidade absoluta



Fonte: (BASTVIKEN et al., 2020).

2.6 SÉRIES TEMPORAIS

Uma série temporal é um conjunto de observações ordenadas no tempo (não necessariamente igualmente espaçadas) e que apresentam dependência serial, isto é, dependência de instantes de tempo. A característica mais importante deste tipo de dados é que as observações vizinhas são dependentes, assim, a ordem das observações é crucial para a análise. Vale ressaltar que, além do tempo, uma série pode ser função de outra variável, como, por exemplo, espaço, profundidade, frequência, etc. (MORETTIN P. A.; TOLOI, 2006).

As observações de séries temporais aparecem em campos diversos do conhecimento como, por exemplo, a Economia (preços diários de ações, taxa mensal de desemprego,

produção industrial), Medicina (eletrocardiograma, eletro encefalograma), Epidemiologia (número mensal de casos de Covid-19), Meteorologia (precipitação pluviométrica, temperatura diária, velocidade do vento), entre outros. A análise de uma única sequência de dados é chamada de análise de série temporal univariada, enquanto que a análise de várias coleções de dados para a mesma sequência de períodos de tempo é chamada de análise multivariada (G (MADDALA, 2009)

Segundo Villa, o objetivo da análise em séries temporais é a construção de modelos com propósitos determinados, tais como (VILLA, 2019): investigar o mecanismo gerador da série temporal, fazer previsões de valores futuros, descrever o comportamento da série e procurar periodicidades relevantes nos dados. Os modelos fundamentais utilizados para descrever séries temporais são processos estocásticos, controlados por leis probabilísticas (MORETTIN P. A.; TOLOI, 2006).

Um processo estocástico é definido como sendo uma coleção de variáveis aleatórias sequenciadas no tempo e definidas em um conjunto de pontos T, que pode ser contínuo ou discreto. A variável aleatória no tempo t é denotada por Z(t), em que t = 0, $\pm$1, $\pm$2, ...,T. Para exemplificação, adota-se dois dispositivos biomédicos, que fundamentam-se nesse processo: o oxímetro, capaz de medir a oxigenação do sangue, e o monitor multiparâmetro de sinais vitais, é um equipamento que faz a leitura dos sinais vitais do paciente, como a pressão arterial não-invasiva, a frequência cardíaca, a temperatura corporal e glicose, e, assim como o oxímetro, a saturação de oxigênio.

No primeiro dispositivo, um único sensor é acoplado ao paciente, posicione-o na ponta do dedo indicador ou dedo médio de qualquer das mãos, como ilustrado na Figura 8. O sensor, por sua vez, vai traduzir a amplitude com que aquele pulso está sendo transmitido, gerando um formato de onda bem característico, conhecido como curva pletismográfica. Na Figura 9, ilustra-se esta curva, neste caso, uma série temporal univariada.

Já na Figura 10, podemos observar o comportamento gerado pelo monitor multiparamétrico, que trabalha com mais de um sensor acoplado ao paciente. Dessa maneira, em um mesmo instante de tempo várias magnitudes de sinais, de diferentes naturezas são traduzidas pelos seus respectivos sensores, formando, assim, uma série temporal multivariada.

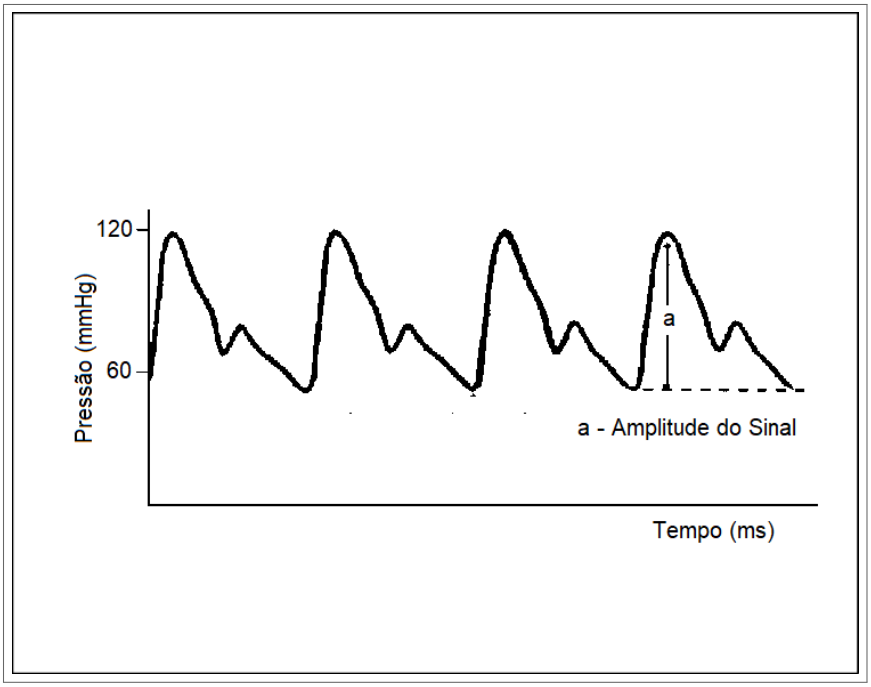
Assim, com as séries temporais geradas pelos sensores MOX, é possível criar uma assinatura específica para cada espécie de fungo processado e, dessa maneira, conseguir rotulá-los de forma rápida e precisa (FARRAIA et al., 2019).

Figura 8 – (a) Oxímetro de Pulso Portátil com Curva Sense 10 da Marca Alfamed, dispositivo biomédico que fundamenta no processo estocástico, medindo a oxigenação do sangue ao longo do tempo; (b) Posicionamento e localização do sensor (dedo da mão)

(a) (b)

Fonte: Disponível em <https://alfamed.com/produtos/oximetro-de-pulso-sense-10/>. Acesso em: 24 Mar 2023

Figura 9 – Pletismografia de pulso amplitude de pulso, exemplo de série temporal univariada. O sensor, traduz a amplitude com que o pulso está sendo transmitido, gerando um formato de onda bem característico, conhecida como curva pletismográfica. Imagem adaptada



Fonte: (LUIZ et al., 1992).

Figura 10 – (a) Monitor Multiparamétrico de sinais vitais 12"canais, dispositivo biomédico que fundamenta no processo estocástico, medindo múltiplos parâmetros ao longo do tempo como a pressão arterial não-invasiva, a frequência cardíaca, a temperatura corporal e glicose, a saturação de oxigênio, entre outros; (b) Display ilustrando as curvas geradas pelos diferentes sensores, um exemplo de série temporal multivariada.

(a) (b)

Fonte: Disponível em <https://www.bleymed.com.br/monitor-multiparametrico-pre-configurado-isaiah>. Acesso em: 24 Mar 2023

2.7 TÉCNICAS ANALÍTICAS DE IDENTIFICAÇÃO DA COMPOSIÇÃO QUÍMICA

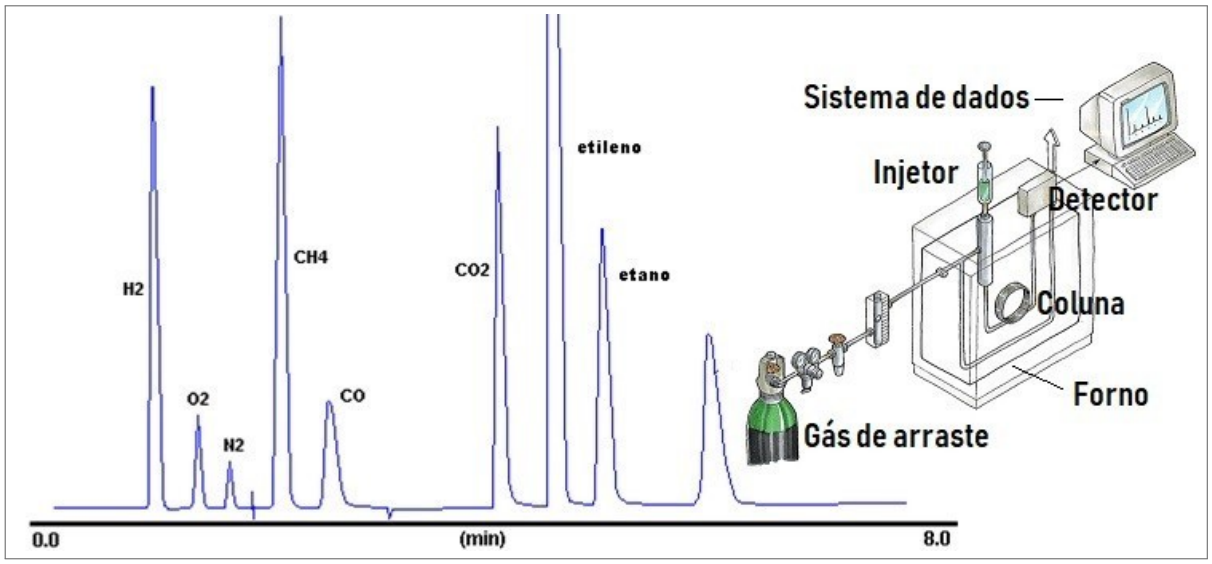
Sabe-se que os compostos orgânicos voláteis são constituídos por componentes químicos. Dessa forma, também há a possibilidade de utilizar as técnicas de análise química instrumental para a identificação fúngica. Essas técnicas podem ser agrupadas em três grandes áreas principais: Cromatografia, Eletroquímica e Espectroscopia, com cada uma delas caracterizando-se por suas particularidades e pelas espécies químicas de interesse (analitos) possíveis de detecção e/ou quantificação. O emprego de uma ou mais técnicas analíticas visa a separação e posterior identificação dos componentes químicos em uma amostra, uma técnica amplamente utilizada para tal finalidade é a Cromatografia Gasosa acoplada a Espectrometria de Massas (CG-EM) (VIEIRA, 2018).

2.7.1 Cromatografia gasosa

A Cromatografia em fase gasosa, ou Cromatografia a Gás (CG), é uma técnica analítica utilizada para a separação de compostos químicos em matrizes complexas onde a fase móvel é composta por um gás. A técnica consiste na introdução da amostra, através de um sistema de injeção aquecido e uma coluna cromatográfica para onde são “arrastadas” a amostra. Assim, a fase estacionária, presente na coluna é responsável pela separação dos compostos, viabilizando a saída desses compostos da coluna em tempos diferentes (tempo de retenção) para que possam ser identificados em um detector apropriado (COLLINS C. H.; BRAGA, 1997).

A Figura 11 exibe os princípios de uma cromatografia a gás: a fase móvel da cromatografia gasosa é um gás de arraste inerte em fluxo continuo na coluna. A amostra

Figura 11 – Princípios de Cromatografia Gasosa, uma técnica analítica utilizada para a separação de compostos químicos em matrizes complexas onde a fase móvel é composta por um gás. A técnica consiste na introdução da amostra, através de um sistema de injeção (onde amostras líquidas são vaporizadas). Em seguida, um gás inerte, "arrasta" a amostra pela coluna cromatográfica. A fase estacionária, presente na coluna é responsável pela separação dos compostos, fazendo com que cada um saia da coluna, a um tempo diferente. Por fim, um detector é responsável pela identificação dos sinais eletrônicos produzidos pelo sistema de CG e um software específico é utilizado para transformar estes sinais em picos cromatográficos



Fonte: (COLLINS C. H.; BRAGA, 1997).

é vaporizada ao ser introduzida rapidamente nessa corrente de gás e então é arrastada através da coluna. As substâncias, já separadas, chegam em tempos distintos ao detector, que gera um sinal para cada fração da amostra, registrando picos no formato de um cromatograma. Por fim, um detector é responsável pela identificação dos sinais eletrônicos produzidos pelo sistema de CG e um software específico é utilizado para transformar estes sinais em picos cromatográficos.

A temperatura da coluna pode sofrer uma variação, linear ou não, sendo chamada de cromatografia gasosa com temperatura programada. Durante a análise a temperatura da coluna sofre uma elevação, diminuindo a retenção de substâncias com maior ponto de ebulição. A programação da temperatura fornece condições para que haja uma maior simetria dos picos e uma melhor detectabilidade para aqueles picos com tempos de retenção excessivamente grandes (COLLINS C. H.; BRAGA, 1997).

2.7.2 Espectrometria de massas por impacto eletrônico

A técnica de Espectrometria de Massas (EM) por impacto de elétrons pode ser definida como uma técnica analítica de identificação de substâncias através da determinação de suas massas moleculares em forma iônica baseado na movimentação desses íons em um campo eletromagneticamente carregado. Esse movimento é determinado pela razão massa/carga (m/z) do analito após o bombardeamento de elétrons de alta energia, resultando

em fragmentos altamente energéticos. Esses fragmentos passam por um campo magnético que os separa baseados pelo m/z de cada íon, chegando ao detector que converte cada sinal em um espectro (SOUZA, 2008). Os espectros podem ser apresentados de forma gráfica ou tabelada sendo o gráfico mais vantajoso por expor as sequências de fragmentação que podem ser reconhecidas com mais facilidade conforme o uso da técnica. A abundância relativa dos picos em comparação ao pico base (o íon que representa 100%) também pode ser fornecida pelo espectro gráfico (MONTEIRO, 2008). Com a modernização dessas duas técnicas analíticas, diversas empresas passaram a investir em equipamentos de cromatografia gasosa acoplada à espectrometria de massas CG-EM. Esse artifício aumenta a concentração de amostra no gás de arraste, aproveitando a maior difusão do gás. As velocidades de varredura são capazes de processar cada espectro de massa por pico eluído do cromatógrafo (MONTEIRO, 2008). A Tabela 2 mostra as principais vantagens e desvantagens dos dois métodos apresentados nas subseções acima.

Tabela 2 – Comparação de metodologias para identificação da composição química

Metodologia	Cromatografia gasosa (CG)	Cromatografia gasosa acoplada à espectrometria de massas
Vantagens	• Alto poder de resolução (análise de muitos componentes de uma única amostra); • Alta sensibilidade; • Análise quantitativa (NASCIMENTO RONALDO FERREIRA DO;LIMA, 2018).	• Altamente sensível; • Ser um detector universal, com análise qualitativa inequívoca; • Possibilidade de realizar a deconvolução de picos não separados (utilizando as massas/cargas) (Cromvallab, 2022).
Desvantagens	• Substâncias voláteis e estáveis termicamente (ou formar um derivado com estas características); • Requer preparo da amostra, necessário especialista para operá-la; • Tempo e custo elevado; • Eficiência qualitativa limitada (NASCIMENTO RONALDO FERREIRA DO;LIMA, 2018).	• Custo do equipamento e custo operacional mais alto; • Necessidade de condições adicionais especiais (gases ou solventes mais puros, colunas mais eficientes e resistentes, melhoria no preparo de amostra, redução nos volumes de injeção, restrição de uso de alguns aditivos e solventes típicos de cromatografia); • Maior manutenção e treinamento, necessário especialista para operá-la (Cromvallab, 2022).

Fonte: A autora (2023).

2.8 PROCESSAMENTO DE SINAIS

O processamento de sinais consiste na análise e/ou modificação de sinais utilizando teoria fundamental, aplicações e algoritmos, de forma a extrair informações dos mesmos e/ou torná-los mais apropriados para alguma aplicação específica. Este processo pode ser feito de forma analógica ou digital, e pode utiliza matemática, estatística, computação, heurística e representação, modelagem, análise, síntese, descoberta, recuperação,detecção, aquisição, extração, aprendizagem, segurança e forense (MOURA, 2009).

2.8.1 Pré-processamento

A fase de pré-processamento visa essencialmente estudar os dados de forma a prepará-los para a fase seguinte. Muitas aplicações exigem um pré-processamento complexo, pois o

conjunto de dados original pode não ser completamente confiável, completo e consistente à partida, provocando uma dificuldade acrescida para muitas arquiteturas de *Deep learning* trabalharem sobre o mesmo.

O pré-processamento de dados é composto pelo seguinte conjunto de etapas (CAMILO CáSSIO OLIVEIRA; SILVA, 2009):

2.8.2 Limpeza dos dados

A limpeza dos dados refere-se ao processo que aumenta a qualidade dos dados de entrada, removendo dados ruidosos, completando dados incompletos e corrigindo inconsistências nos mesmos. Na eventualidade desta etapa não ser aplicada, torna-se complicado considerar que os dados são fiáveis, o que consequentemente leva a uma desconfiança nos resultados de qualquer processo de aprendizagem de um algoritmo (CAMILO CáSSIO OLIVEIRA; SILVA, 2009).

2.8.3 Integração dos dados

A integração dos dados trata-se do procedimento para reunir dados provenientes de diversas fontes num armazenamento de dados coerente, como uma base de dados, por exemplo. No entanto, como os dados provém de múltiplas fontes, a probabilidade de haver inconsistências e redundâncias nos mesmos é bastante alta. Por esta razão, geralmente as etapas limpeza dos dados e integração dos dados são executadas como uma só etapa de pré-processamento na preparação dos dados para um armazenamento (CAMILO CáSSIO OLIVEIRA; SILVA, 2009).

2.8.4 Transformação dos dados

Nesta etapa, os dados são transformados ou consolidados de modo que o processo resultante seja mais eficiente e os padrões encontrados sejam mais fáceis de compreender. Estas transformações ou consolidações de dados resultam de operações de agregação, generalização, normalização e discretização dos dados (CAMILO CáSSIO OLIVEIRA; SILVA, 2009). É importante mencionar que grande parte dos erros são corrigidos durante esta etapa de transformação de dados, nomeadamente erros que têm como base erro humano, como por exemplo os erros originados por um processamento de dados incorreto.

As técnicas de *Dataset Augmentation* e de Data Balance fazem parte desta etapa de pré-processamento. A técnica de *Dataset Augmentation* pode ser vista como uma forma de pré-processar os dados de treino, podendo ser uma forma de reduzir o erro de generalização da maioria dos modelos. Sucintamente, trata-se de uma técnica com o objetivo de criar novos "dados"com diferentes orientações, através dos dados existentes. Escalas, translações, rotações, inversões ou mesmo corte das imagens em locais ligeiramente diferentes são alguns exemplos de aumento de dados (GOODFELLOW, 2016). Importante

referir que esta técnica está associada ao aumento do conjunto de dados para o treino do modelo, podendo ser bastante útil no caso de haver uma escassa quantidade de dados.

As técnicas de *Data Balance* têm como objetivo, garantir que todas as classes de saída de um determinado algoritmo de aprendizagem estejam balanceadas, ou seja, no conjunto de dados não deverá haver uma diferença significativa no número de exemplos de diferentes classes. Algumas estratégias que podem ser seguidas para isto, podem passar pelo redimensionamento do conjunto de dados, ignorando exemplos das classes que possuem um tamanho maior de exemplos, ou usar custos diferentes para classificar cada uma das classes.

Segundo Dorian Pyle (PYLE, 1999), o pré-processamento é a etapa que requer um maior esforço envolvido ao longo de todo o processo, estimando ainda que cerca de 80% do tempo despendido no processo seja utilizado nesta etapa.

2.8.5 Redução dos dados

O conjunto de dados pode conter um grande número de entradas, podendo tornar o processo demasiadamente lento e pouco eficiente quando diretamente utilizados nos algoritmos. Desta forma, frequentemente são realizadas técnicas de redução do tamanho dos dados, para a obter um conjunto de dados com dimensão inferior, mas, ao mesmo tempo, mantendo a sua qualidade e integridade (SILVA, 2021). Assim, esta etapa é responsável por reduzir o tamanho dos dados, por exemplo, agregando, eliminando recursos redundantes/irrelevantes ou aplicando a técnica de *clustering*. A aplicação de técnicas de redução de dados permite que os dados de entrada tenham menos volume, mantendo sempre a sua integridade e produzindo os mesmos resultados analíticos. Algumas dessas técnicas incluem (SILVA, 2021):

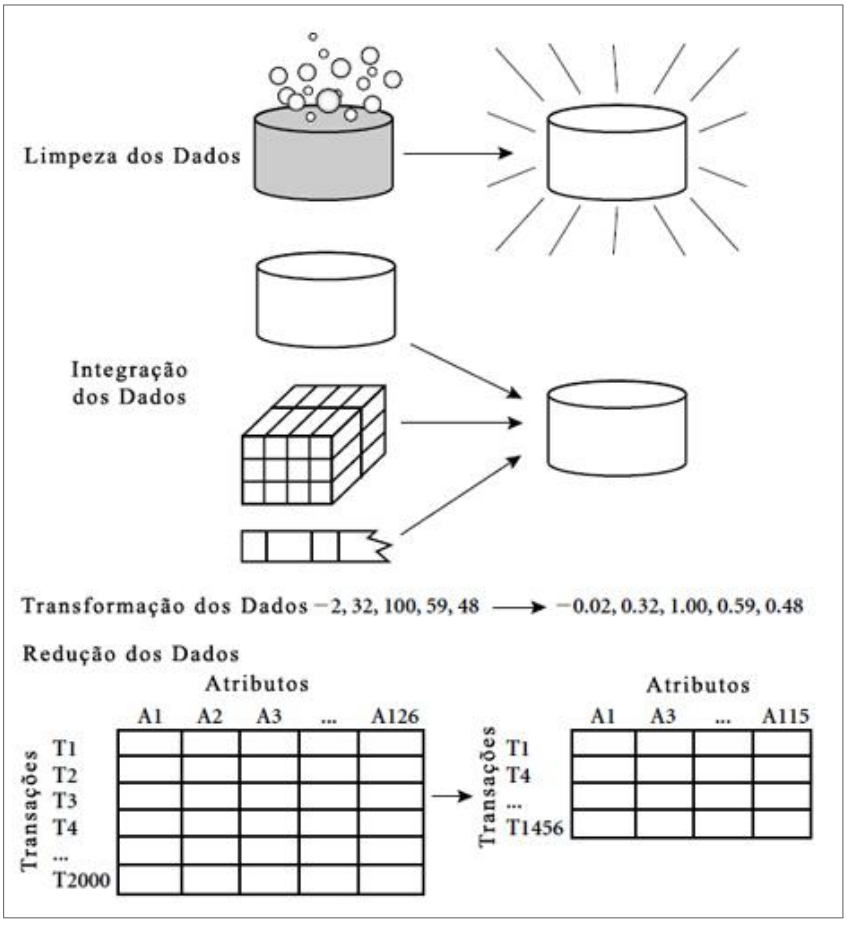
- Redução de dimensionalidade: Remove atributos irrelevantes do conjunto de dados, de forma a encontrar o conjunto mínimo de atributos, embora o resultado final deva ser idêntico ou melhor que o resultado original com todos os atributos;
- Redução de numerosidade: Os dados são substituídos por representações alternativas menores.

A Figura 12 resume as quatro etapas de pré-processamento de dados descritas anteriormente.

2.8.6 Extração de Atributos

A segunda etapa no processo é a extração de atributos, aplicada após o pré-processamento dos dados, tratando-se de uma etapa onde serão identificados os atributos presentes no conjunto de dados. Mais concretamente, a extração de atributos é extrair e converter a informação relativa aos dados de entrada num conjunto de características chamado de

Figura 12 – Etapas do Pré-Processamento de Dados.. Frequentemente, os dados são encontrados com diversas inconsistências: registros incompletos, valores errados e dados inconsistentes. A etapa de limpeza dos dados visa eliminar estes problemas de modo que eles não influam no resultado dos algoritmos usados. Além disso, é comum obter-se os dados a serem minerados de diversas fontes: banco de dados, arquivos textos, planilhas, vídeos, imagens, entre outras. Surge então, a necessidade da integração destes dados de forma a termos um repositório único e consistente. Para isto, é necessária uma análise aprofundada dos dados observando redundâncias, dependências entre as variáveis e valores conflitantes. Alguns algoritmos trabalham apenas com valores numéricos e outros apenas com valores categóricos. Nestes casos, é necessário transformar os valores numéricos em categóricos ou os categóricos em valores numéricos. Não existe um critério único para transformação dos dados e diversas técnicas podem ser usadas de acordo com os objetivos pretendidos. Em alguns casos, o volume de dados é tão grande que torna o processo de análise dos dados impraticável. Nestes casos, as técnicas de redução de dados podem ser aplicadas para que a massa de dados original seja convertida em uma massa de dados menor, porém, sem perder a representatividade dos dados originais



Fonte: (CAMILO CáSSIO OLIVEIRA; SILVA, 2009).

vetor de características, reduzindo o padrão de representação de dados. Esta etapa de seleção deve ser bem realizada, onde uma boa seleção dos atributos contribui para um menor custo computacional dos cálculos das medidas, armazenamento e agrupamento dos dados. O conjunto de características selecionará as informações relevantes dos dados de entrada de forma a executar depois a tarefa de classificação (MARQUES, 2018).

Existem bastantes metodologias de extração de atributos, sendo que segundo o estudo efetuado no trabalho de Choran *et al.* existem algumas que são frequentemente utilizadas em análise de sinal, tais como o Fast Fourier Transform (FFT) e o Wavelet Transform (WT). O FFT é uma ferramenta computacional que facilita a análise do sinal, através de ferramentas matemáticas, como análise de espectro de potência e simulação de filtro por meio de computadores digitais, já WT é uma transformação linear muito parecida com a transformada de Fourier, com uma importante diferença: ela permite a localização no tempo de diferentes componentes de frequência de um dado sinal (COCHRAN WILLIAM W. COOLEY, 1967).

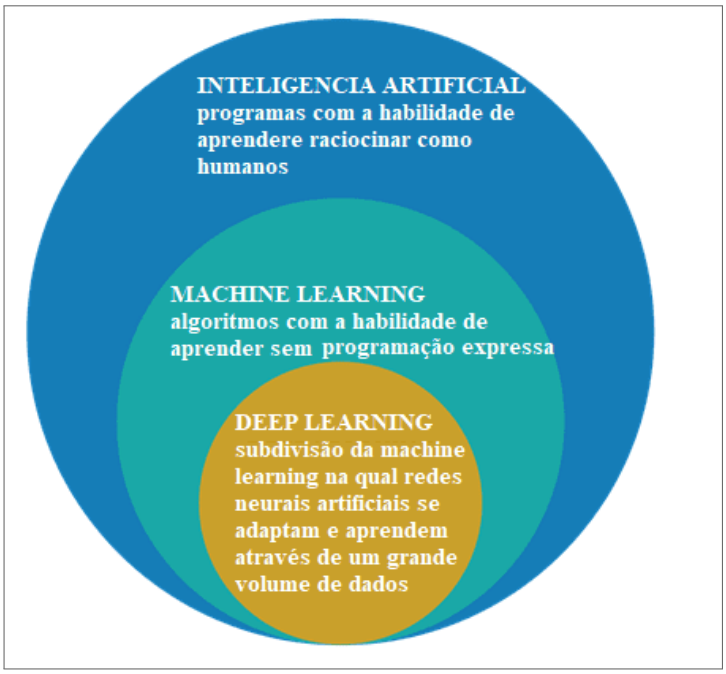
2.8.7 Aprendizagem de máquina

O principal componente de *hardware* de um nariz eletrônico é uma matriz de sensores de gás não específicos, ou seja, sensores que interagem com uma ampla gama de produtos químicos com intensidades variadas. Correspondentemente, o analito estimula os sensores na matriz, o que gera uma resposta característica chamada "impressão digital". O principal componente de *software* de um *e-nose* é sua extração de características e algoritmos de reconhecimento de padrões, que processam a resposta característica do sensor, extraem e selecionam informações úteis e realizam o reconhecimento de padrões (YAN XIUZHEN GUO; ZHANG, 2015)

Inteligência Artificial (IA) é uma área na Ciência da Computação responsável por simular a inteligência e o comportamento humano usando apenas máquinas. Aprendizagem de Máquina (AM), em inglês Machine Learning (ML), é um subcampo do desenvolvimento de IA. Essa vertente surgiu a partir do desenvolvimento da ciência da computação no que diz respeito a reconhecimento de padrões e da própria IA. Um aspecto importante a ser destacado é a diferenciação entre IA, aprendizagem de máquina e Aprendizagem Profunda (AP), em inglês Deep Learning (DP). Em que, todas essas definições são vinculadas, porém são conceitos distintos.

Na Figura 13, ilustra-se a diferenciação entre as três definições supracitadas. A IA é a capacidade de uma máquina simular uma inteligência humana. Já ML é um subconjunto da IA que permite os computadores realizarem um processo de aprendizagem sem serem programados claramente para este objetivo, podendo se adaptar a novos cenários e dados. E *Deep Learning* é um subconjunto do *Machine Learning* que diz respeito a redes neurais complexas.

Figura 13 – Hierarquia das áreas de estudo. A área de estudo da Inteligência Artificial é o estudo e projeto de agentes inteligentes, ou seja, qualquer sistema que consiga tomar uma decisão baseado em uma heurística pode ser considerado inteligente. Dentro da inteligência artificial existem diversas técnicas diferentes que modelam essa inteligência. Algumas técnicas podem ser classificadas na área de *Machine Learning*, que de forma geral, aprendem a tomar uma decisão baseadas em exemplos de um problema, e não de uma programação específica. Um subgrupo específico de técnicas de ML são chamadas de *Deep Learning*, geralmente utilizam redes neurais profundas e dependem de muitos dados para o treinamento



Fonte: (SANTANA, 2018).

2.8.8 Tipos de aprendizagem de máquina

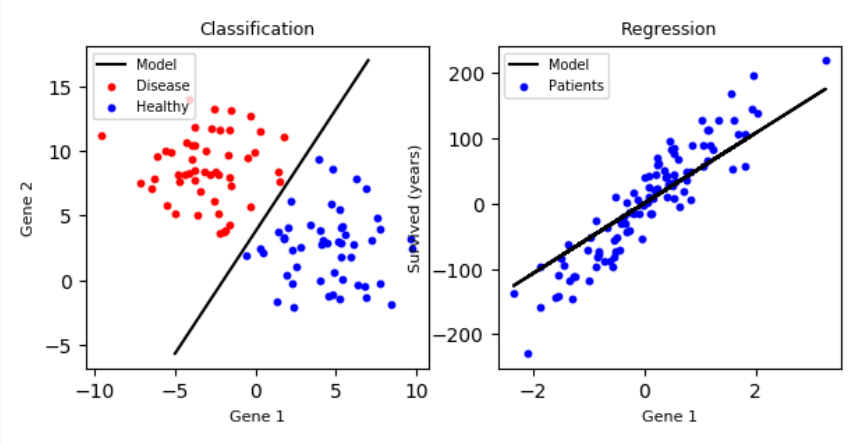
Ao tratar de aprendizagem de máquina, é de suma importância salientar os três tipos existentes: Aprendizagem Supervisionada, (em inglês, *Supervised Learning*), Aprendizagem Não-Supervisionada (em inglês, *Unsupervised Learning*) e Aprendizagem por Reforço (em inglês, *Reinforcement Learning*).

O principal objetivo da aprendizagem supervisionada é conceber um modelo, a partir de uma base de dados rotulada, que permita realizar previsões a respeito de dados não vistos ou futuros. Logo, o termo supervisionado compreende as amostras em que é conhecida a saída. As duas principais subcategorias de *Supervised Learning*, são: Classificação e a Regressão (RASCHKA S.; MIRJALILI, 2017). A principal diferença entre os algoritmos de regressão e classificação que os algoritmos de regressão são usados para prever os valores contínuos, como preço, salário, idade, etc. e os algoritmos de classificação são usados para prever/classificar os valores discretos, como masculino ou feminino, verdadeiro ou falso, etc.

Imagine-se realizando um estudo sobre um grupo de pacientes e analisando as informações genéticas deles. Na Figura 14, ilustra-se exemplos paras os respectivos métodos dependendo do gene analisado. No processo de classificação cada exemplo de aprendizagem

está associado a um valor alvo qualitativo, que corresponde a uma classe (por exemplo, doente, saudável). Pode haver duas classes (classificação binária) ou mais (classificação multiclasse). Enquanto que no processo de regressão, cada exemplo de aprendizado está associado a um valor alvo quantitativo (por exemplo, tempo de sobrevivência). O objetivo do modelo é estimar a saída correta, dado um vetor de características (DROUIN, 2017).

Figura 14 – Exemplos de Aprendizagem Supervisionada.No aprendizado de máquina supervisionado, os conjuntos de dados são coleções de exemplos de aprendizado no formato (x, y), onde x é um vetor de recursos e y é seu valor alvo correspondente. Características são variáveis observadas que descrevem cada exemplo. Na classificação, busca-se um modelo que seja capaz de diferenciar entre pacientes doentes e saudáveis, enquanto que para regressão busca-se estimar a quantidade de pacientes com esse gene

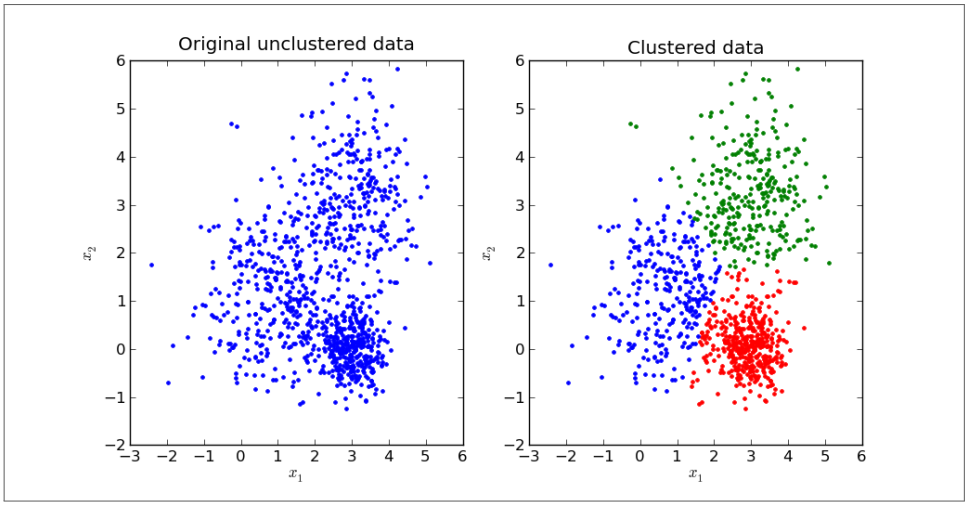


Fonte: (DROUIN, 2017).

Já na aprendizagem não-supervisionada, o programa irá lidar com uma base de dados desconhecida, sem rótulos ou estrutura conhecida. Dessa forma, esse tipo é capaz de extrair informações a partir da estrutura da base de dados, sem a orientação de rótulos ou uma função recompensa. A principal técnica para a aprendizagem não supervisionada é o *Clustering*, que consiste em uma técnica de divisão do banco de dados em subgrupos, ou *clusters*, a partir da similaridade dos dados. Essa técnica começa com a atribuição do número de *clusters* a serem encontrados (TOWARDS DATA SCIENCE, 2017). Na Figura 15, observa-se o agrupamento com base em 3 *clusters*.

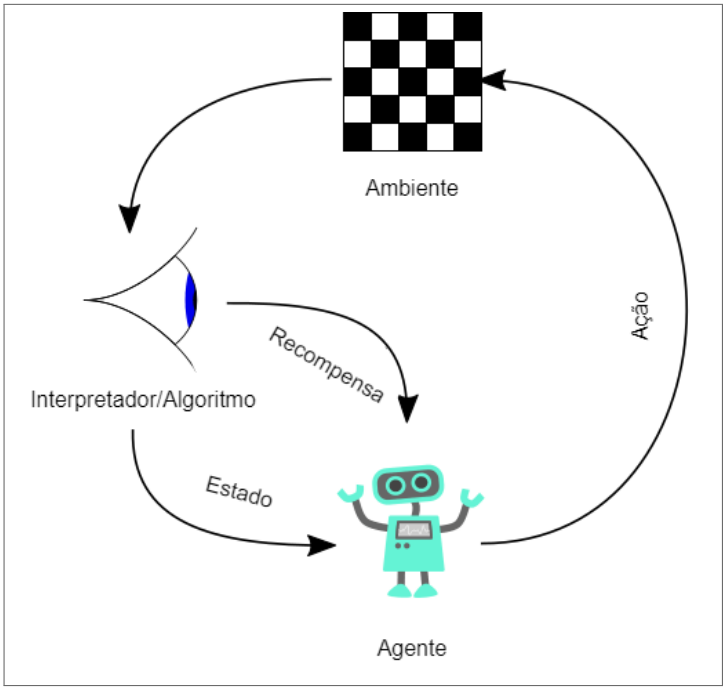
Por fim, na Aprendizagem por Reforço, o algoritmo irá melhorar seu desempenho com base em suas interações com o meio. Para isso um sistema de recompensas, ou um sinal de prêmio, irá sinalizar quando uma ação do programa é a correta ou não. Vale ressaltar que o *feedback* fornecido pela recompensa pode determinar como certa uma ação que talvez não fosse a ótima para o caso em questão. Na Figura 16 é pode ser visualizado o conceito desse tipo de aprendizagem. Logo, uma abordagem em que uma função de recompensa seja capaz de majorar o prêmio, com base nas ações, é bastante promissor. Um exemplo clássico para esse tipo de aprendizagem é um modelo para jogar xadrez, em que a partir da ampla gama de movimentos possíveis no tabuleiro a recompensa é definida pela vitória ou derrota no fim do jogo (RASCHKA S.; MIRJALILI, 2017).

Figura 15 – Exemplo de Aprendizagem Não Supervisionada. O *Clustering* é uma técnica simples e fácil usada para classificar ou agrupar um conjunto de dados em um certo número de *clusters*. Essa técnica começa com a atribuição do número de clusters a serem encontrados



Fonte: (TOWARDS DATA SCIENCE, 2017).

Figura 16 – Exemplo de Aprendizagem por Reforço. O conceito desse tipo de aprendizagem para um jogo de xadrez. Com base nas ações e observações dos movimentos no tabuleiro como a recompensa é possível melhorar o desempenho do algoritmo (agente)



Fonte: (RASCHKA S.; MIRJALILI, 2017).

A Classificação de Séries Temporais (TSC) é o estudo através de reconhecimento de padrões e da teoria do aprendizado computacional em inteligência artificial, em que os recursos do vetor de entrada são valorizados e ordenados de forma real (RUIZ ALEJANDRO PASOS; FLYNN, 2021). Esse cenário adiciona uma camada de complexidade ao problema, pois características importantes dos dados podem ser perdidas pelos algoritmos tradicionais. Nos últimos anos, novos conjunto de algoritmos TSC vem sendo desenvolvido, e trazendo melhorias significativas em relação ao estado da arte anterior (BAGNALL A., 2017).

O foco principal tem sido no TSC univariado, ou seja, o problema onde cada caso tem uma única série e um rótulo de classe. Na realidade, é mais comum encontrar problemas de Classificação de Séries Temporais multivariado (MTSC) em que a série temporal para um único caso tem várias dimensões. Reconhecimento de atividade humana, diagnóstico baseado em Eletrocardiograma (ECG), Eletroencefalograma (EEG) e Magnetoencefalografia (MEG) e problemas de monitoramento de sistemas são todos inerentemente multivariados (RUIZ ALEJANDRO PASOS; FLYNN, 2021).

2.9 ALGORITMOS DE CLASSIFICAÇÃO

Como visto na Seção 2.5, e na Seção 2.6, o problema dessa dissertação é identificar um conjunto de múltiplos COVs, que são emitidos por culturas de fungos. A este conjunto característicos de cada espécie de fungo é denominado assinatura de odor. Cada assinatura é um conjunto de séries temporais lida pelo nariz eletrônico e apresenta um perfil característico do fungo. Conforme comentado na Seção 1.1, esse trabalho contempla 6 gêneros de fungos *Candida*: *C. Albicans, C. Glabrata, C. Haemulonii, C. Kodamaea ohmeri, C. Krusei, C. Parapsilosis e C. Tropicalis.*. Portanto, para solucionar esse problema de classificação, é necessário encontrar modelos de aprendizagem de máquina, especializados em classificar séries temporais, que apresentarem os melhores resultados. A *Sktime* é uma biblioteca Python especializada em dados de séries temporais. Nesta seção, será abordado uma descrição de todos os algoritmos que foram utilizados nesse domínio de identificação de diferentes espécies de *Candida*.

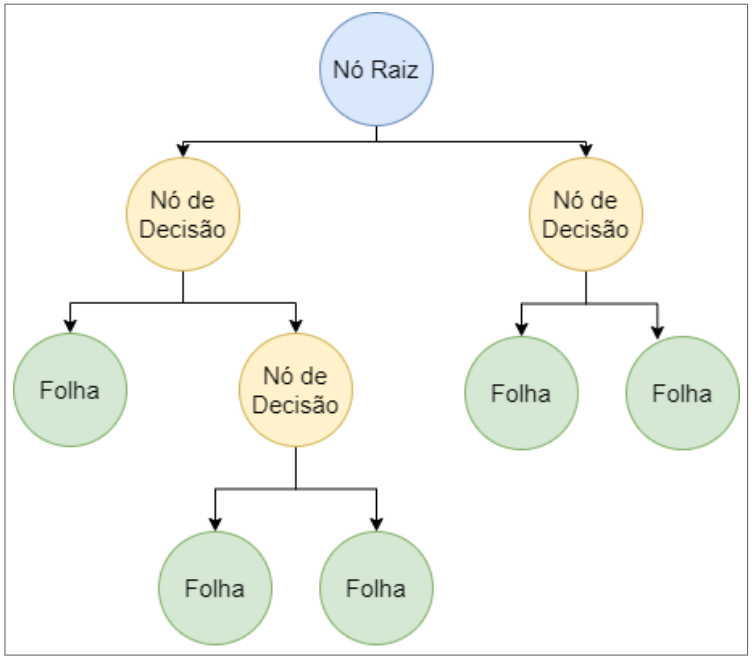
2.9.1 Time Series Forest

Árvores de decisão são modelos em estrutura de árvore onde cada nodo interno representa um teste de uma determinada variável que define um caminho para chegar da raiz à folha da árvore (nó terminal). A folha da árvore representa a decisão a ser tomada, ou seja, o resultado da classificação. Uma das principais vantagens da utilização desse tipo de estrutura gráfica é a fácil visualização do modelo gerado. Além disso, uma árvore de decisão pode facilmente ser transformada em um conjunto de regras (o caminho da raiz

até cada folha é único e pode ser transformado em regra) que é outra forma conveniente de modelar uma determinada situação (XU M.; WATANACHATURAPORN, 2005).

O procedimento padrão para criar uma árvore de decisão utiliza uma estratégia de "divisão e conquista" onde, a cada passo, uma determinada variável é selecionada (a que separa melhor os dados com relação à variável alvo) e os dados são particionados de acordo com os valores da variável (cada valor define um ramo, ou seja, caminho a ser seguido na árvore). O processo é repetido recursivamente para os ramos inferiores utilizando somente o subconjunto de dados de cada ramo e as variáveis que sobraram. O processo continua até não haver mais como dividir os subconjuntos de dados ou acabarem os dados. A figura a seguir ilustra a estrutura de uma árvore de decisão:

Figura 17 – Estrutura de uma árvore de decisão. A folha da árvore representa a decisão a ser tomada. Cada caminho da raiz até cada folha é único e pode ser transformado em regra.



Fonte: A autora (2023).

O Time Series Forest (TFS) é um método de conjunto de árvores, conhecido como floresta de séries temporais, onde é proposto para classificação de séries temporais. O TFS emprega uma combinação de ganho de entropia e uma medida de distância, conhecida como ganho de entrada (entropia e distância), para avaliar as divisões.

Estudos experimentais mostram que o ganho de entrada melhora a precisão do TFS. O TFS amostra aleatoriamente recursos em cada nó da árvore e tem complexidade computacional linear no comprimento da série temporal e pode ser construído usando técnicas de computação paralela. A curva de importância temporal é proposta para capturar as características temporais úteis para a classificação.

Estudos experimentais mostram que o TFS usando recursos simples, como média, desvio padrão e inclinação, é computacionalmente eficiente e supera concorrentes fortes,

como classificadores de um vizinho mais próximo com distorção de tempo dinâmica.

2.9.2 KNeighbors

O algoritmo k-Nearest Neighbor (kNN) é um algoritmo de aprendizado supervisionado do tipo *lazy*, introduzido por (AHA D. W., 1991). A ideia geral desse algoritmo consiste em encontrar os k exemplos rotulados mais próximos do exemplo não classificado e, com base no rótulo desses exemplos mais próximos, é tomada a decisão relativa à classe do exemplo não rotulado. Os algoritmos da família kNN requerem pouco esforço durante a etapa de treinamento. Em contrapartida, o custo computacional para rotular um novo exemplo é relativamente alto, pois, no pior dos casos, esse exemplo deverá ser comparado com todos os exemplos contidos no conjunto de exemplos de treinamento.

Na Figura 18 é ilustrada essa ideia para um problema de classificação, com um conjunto de exemplos de treinamento descrito por dois atributos, no qual, exemplos com rótulo positivo (+) referem-se a pacientes doentes e exemplos com rótulo negativo (-) a não doentes. Considerando o algoritmo kNN para classificação, com k = 1, o novo exemplo (?) seria classificado de acordo com o único vizinho mais próximo, que é da classe positiva (+). Três parâmetros importantes devem ser determinados para a execução de kNN:

- quais exemplos rotulados, i.e., exemplos de treinamento, devem ser lembrados;
- qual a medida que quantifica a similaridade entre o exemplo não classificado e os exemplos de treinamento;
- quantos/quais vizinhos mais próximos devem ser considerados.

Assim o KNeighborsClassifier é uma versão adaptada do scikit-learn kNN para dados de séries temporais. Este classificador suporta medidas de distância de séries temporais.

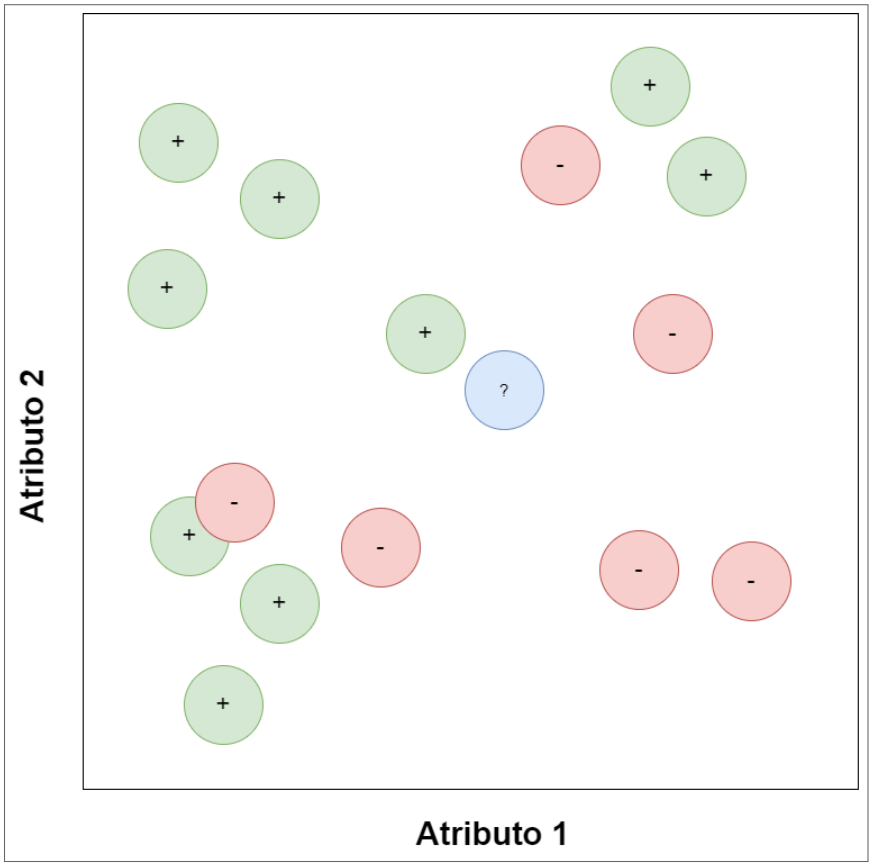
2.9.3 Random Convolutional Kernel Transform (ROCKET)

Na Figura 20 acima, nota-se que existem duas classes de observações: os pontos azuis e os pontos roxos. Existem várias maneiras de separar essas duas classes, conforme mostrado no gráfico à esquerda. No entanto, queremos encontrar o “melhor” hiperplano que possa maximizar a margem entre essas duas classes, o que significa que a distância entre o hiperplano e os pontos de dados mais próximos de cada lado é a maior. Dependendo de qual lado do hiperplano um novo ponto de dados localiza, podemos atribuir uma classe à nova observação.

Parece simples no exemplo acima. No entanto, nem todos os dados são linearmente separáveis. De fato, no mundo real, quase todos os dados são distribuídos aleatoriamente, o que dificulta a separação linear de diferentes classes.

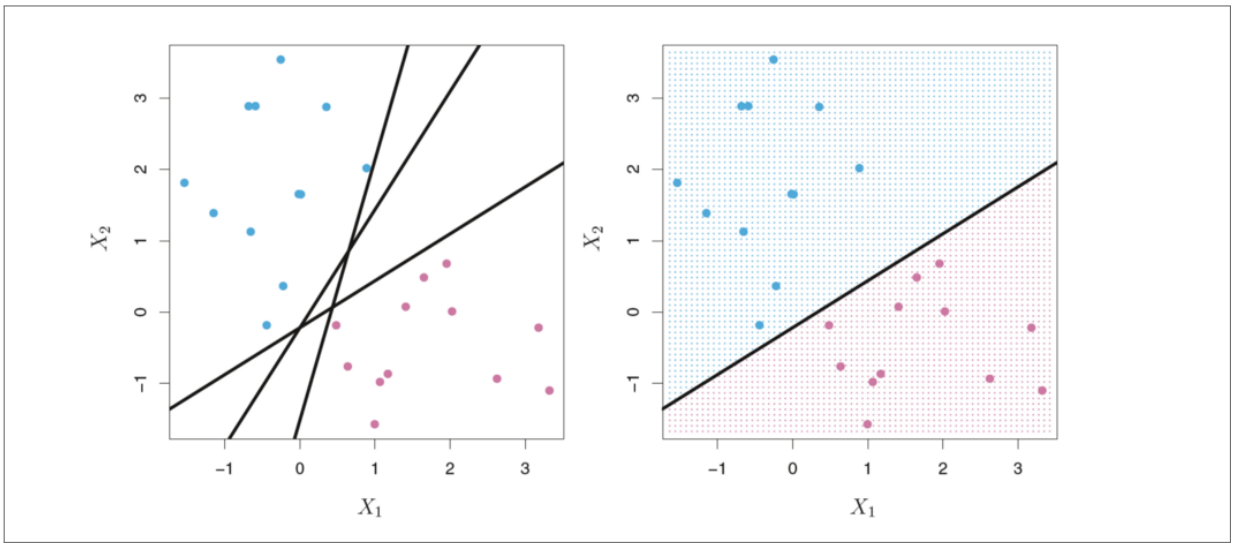
Dessa forma, na Figura 20, encontra-se uma maneira de mapear os dados do espaço bidimensional para o espaço tridimensional, onde possibilita encontrar uma superfície

Figura 18 – Exemplo de classificação do método *k-Nearest Neighbor*. O rótulo positivo (+) referem-se a pacientes doentes e o rótulo negativo (-) a não doentes. Com k = 1, o novo exemplo (?) seria classificado de acordo com o único vizinho mais próximo, que é da classe positiva (+)



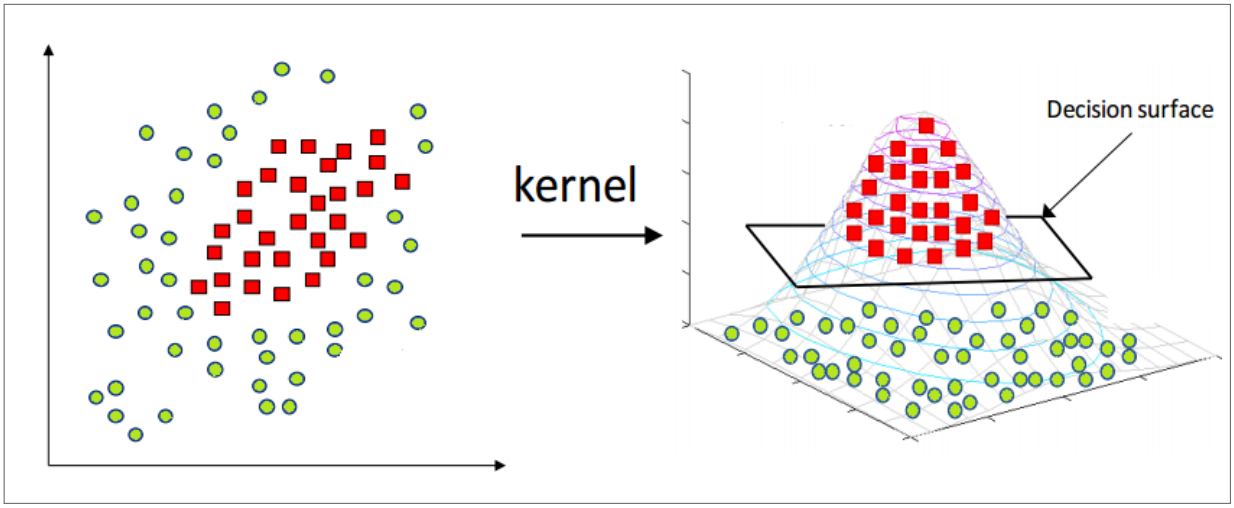
Fonte: A autora (2023).

Figura 19 – Exemplo de separação de grupos. Os pontos azuis e os pontos roxos são classes de observação. Existem vários hiperplanos capazes de separar essas duas classes, conforme mostrado no gráfico à esquerda. O melhor hiperplano é aquele que a distância entre o hiperplano e os pontos de dados mais próximos de cada lado é a maior. Dependendo de qual lado do hiperplano um novo ponto de dados localiza, podemos atribuir uma classe à nova observação



Fonte: Disponível em <https://lamfo-unb.github.io/2020/07/04/SVM/>. Acesso em: 10 Abr 2023

Figura 20 – Exemplo de separação de grupos utilizando Kernel, uma maneira de mapear os dados do espaço bidimensional para o espaço tridimensional, onde possibilita encontrar uma superfície de decisão que se divide claramente entre as diferentes classes



Fonte: Disponível em <https://medium.com/@srilakshmit3512/a-brief-introduction-to-svm-support-vector-machine-778090b09933>. Acesso em: 10 Abr 2023

de decisão que se divide claramente entre as diferentes classes. No entanto, quando há mais e mais dimensões, os cálculos dentro desse espaço tornam-se cada vez mais caros computacionalmente.

Assim, em essência, o que o *kernel* faz é oferecer uma maneira mais eficiente e menos dispendiosa de transformar dados em dimensões superiores. Existem diferentes núcleos. Os mais populares são o *kernel* polinomial e o *kernel* da função de base radial (em inglês, Radial-Basis Function (RBF)).

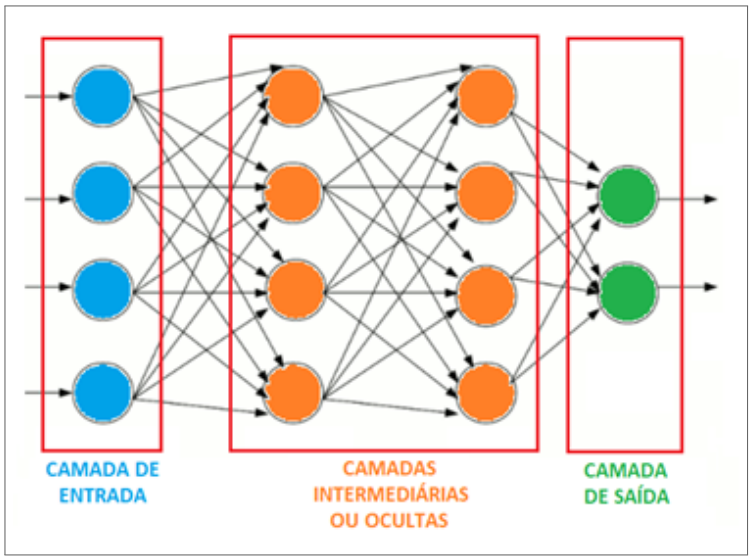
De acordo com Dempster *et al.*, os *kernels* convolucionais são um instrumento único e poderoso que pode capturar muitos dos recursos usados pelos métodos existentes para classificação de séries temporais (DEMPSTER; PETITJEAN; WEBB, 2020). E mostram que, em vez de aprender os pesos do kernel, um grande número de kernels aleatórios em combinação são extremamente eficazes para capturar padrões discriminativos em séries temporais. Além disso, os kernels aleatórios têm requisitos computacionais muito baixos, tornando o aprendizado e a classificação extremamente rápidos.

O ROCKET faz uso importante da proporção de valores positivos (ou ppv) para resumir a saída dos mapas de recursos, permitindo que um classificador pondere a prevalência de um padrão em uma determinada série temporal. Assim, diferente dos outros métodos para classificação de séries temporais, que podem ter alta complexidade computacional e se concentram em um único tipo de recurso, como forma ou frequência, o Rocket baseá-se em redes neurais convolucionais para classificação de séries temporais, mostramos que classificadores lineares simples usando kernels convolucionais aleatórios.

2.9.4 Multivariate Time Series Classification with Attentional Prototypical Network (TapNet)

Uma rede neural de múltiplas camadas, mais especificamente uma rede neural convolucional (CNN), recebe este nome porque possui camadas de convolução que, juntamente com outros tipos de camadas, determinam uma arquitetura específica para esta finalidade, como veremos adiante. Uma rede neural profunda possui camadas de neurônios: uma camada de entrada, algumas camadas intermediárias (ou ocultas), e uma camada de saída, como na Figura 21.

Figura 21 – Exemplo de camadas de neurônios de um CNN. Camada de Entrada é onde os padrões são apresentados à rede; Camadas Intermediárias ou Escondidas é onde é feita a maior parte do processamento, através das conexões ponderadas, e podem ser consideradas como extratoras de características; Camada de Saída é onde o resultado final é concluído e apresentado.



Fonte: Disponível em http://www2.decom.ufop.br/imobilis/fundamentos-de-redes-neurais/. Acesso em: 10 Abr 2023

Uma imagem 300×300, por exemplo, hipoteticamente, em uma rede completamente conectada, teríamos cada pixel da imagem ligado a cada um dos neurônios na entrada. Para isso, seria preciso de 90 mil neurônios, resultando em uma quantidade extremamente alta de parâmetros, e uma descorrelação espacial entre os pixels da imagem, ou seja, um enorme desperdício de recurso.

Por isso, a aplicação de camadas convolucionais sobre os pixels da imagem, reduz muito a quantidade de parâmetros e facilita a descoberta de padrões. Geralmente utiliza-se filtros espaciais lineares como visto no seção anterior. Uma camada convolucional realiza o aprendizado de múltiplos filtros, onde cada filtro extrai uma informação da imagem. Como dito anteriormente, outras camadas são utilizadas em conjunto com a convolucional.

Existem diversos métodos de classificação que utilizam a ideia de mapear dados de série temporal original em uma matriz de imagem análoga e usar um modelo clássico existente da CNN para extrair recursos e classificar as mistura de gases. No entanto, sabe-

se que na classificação das imagens a posição e a ordenação das informações são recursos importantes, por isso, nesse método a direção de operação de convolução é limitada a uma única e determinada ordem. Dessa maneira, uma mudança na direção do desenvolvimento de dados da série temporal pode levar a grandes mudanças nas tendências e localizações das curvas. Assim, o resultado da classificação pode ser afetado diretamente caso haja perda de informações na mudança de direção, por exemplo.

Pensando nestas limitações de mapeamento, Zhang *et al.* propôs um classificador chamado TapNet (ZHANG YIFENG GAO, 2020). A Figura 22 mostra a arquitetura geral do modelo TapNet, que compreende três componentes principais: permutação aleatório de dimensão (em inglês, *Random Dimension Permutation*), codificação de série temporal multivariada (em inglês, *Multivariate Time Series Encoding*) e aprendizagem de protótipo de atenção (em inglês, *Attentional Prototype Learning*).

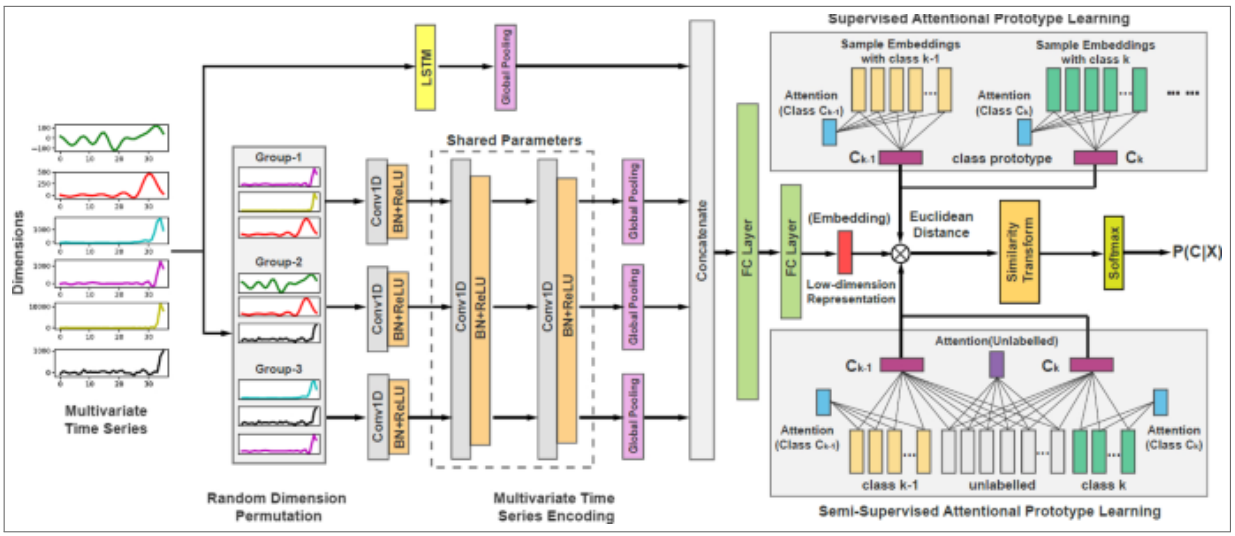
A entrada desse modelo é um conjunto de séries temporais multivariadas com dimensões múltiplas. Na figura, adotou-se como exemplo uma série temporal de 6 dimensões. Para cada dimensão, as séries temporais compartilham a mesma duração. Este modelo propõe um método de permutação de dimensão aleatória (em inglês, Random Dimension Permutation (RDP)) para combinar aleatoriamente as dimensões em grupos diferentes, porém com um tamanho de grupo fixo. Dessa maneira, para este exemplo, as seis dimensões foram divididas em três grupos fixos com diferentes permutações de dimensão.

Após a permutação da dimensão, é aplicado a memória de longo e curto prazo (em inglês, Long Short-Term Memory (LSTM))e uma convolução de 1-camada dimensional para modelar as informações sequenciais de séries temporais e as relações entre o tempo. Depois disso, uma representação de *feature* (*embedding)* de baixa dimensão é aprendido através da codificação de série temporal multivariada (*Multivariate Time Series Encoding*) e assim, são utilizados como entrada na aprendizagem dos protótipos de atenção (*Attentional Prototype Learning*) para cada classe. Dessa maneira, o protótipo da classe é uma combinação ponderada das amostras de treinamento na mesma classe, onde o pesos das amostras de treinamento são treinados por uma camada de atenção. A intuição por trás disso é aprender um protótipo de classe para cada classe. Portanto, amostras de dados na mesma classe possuem distâncias menores entre si, enquanto amostras de dados em classes diferentes possuem distâncias maiores.

2.9.5 Hierarchical Vote Collective of Transformation-based Ensemble (HIVE-COTE)

O HIVE-COTE é um meta-ensemble heterogêneo para classificação de séries temporais, ele forma seu conjunto a partir de classificadores de múltiplos domínios, incluindo *shapelets* independentes de fase, dicionários baseados em bag-of-words e intervalos dependentes de fase (LINES; TAYLOR; BAGNALL, 2016). Desde que foi proposto pela primeira vez em 2016, o algoritmo permaneceu no estado da arte para precisão no arquivo de classificação de séries temporais UCR (MIDDLEHURST M., 2021). Ao longo do tempo foi sendo atuali-

Figura 22 – Arquitetura geral do modelo TapNet, que compreende três componentes principais: permutação aleatório de dimensão, codificação de série temporal multivariada e aprendizagem de protótipo de atenção



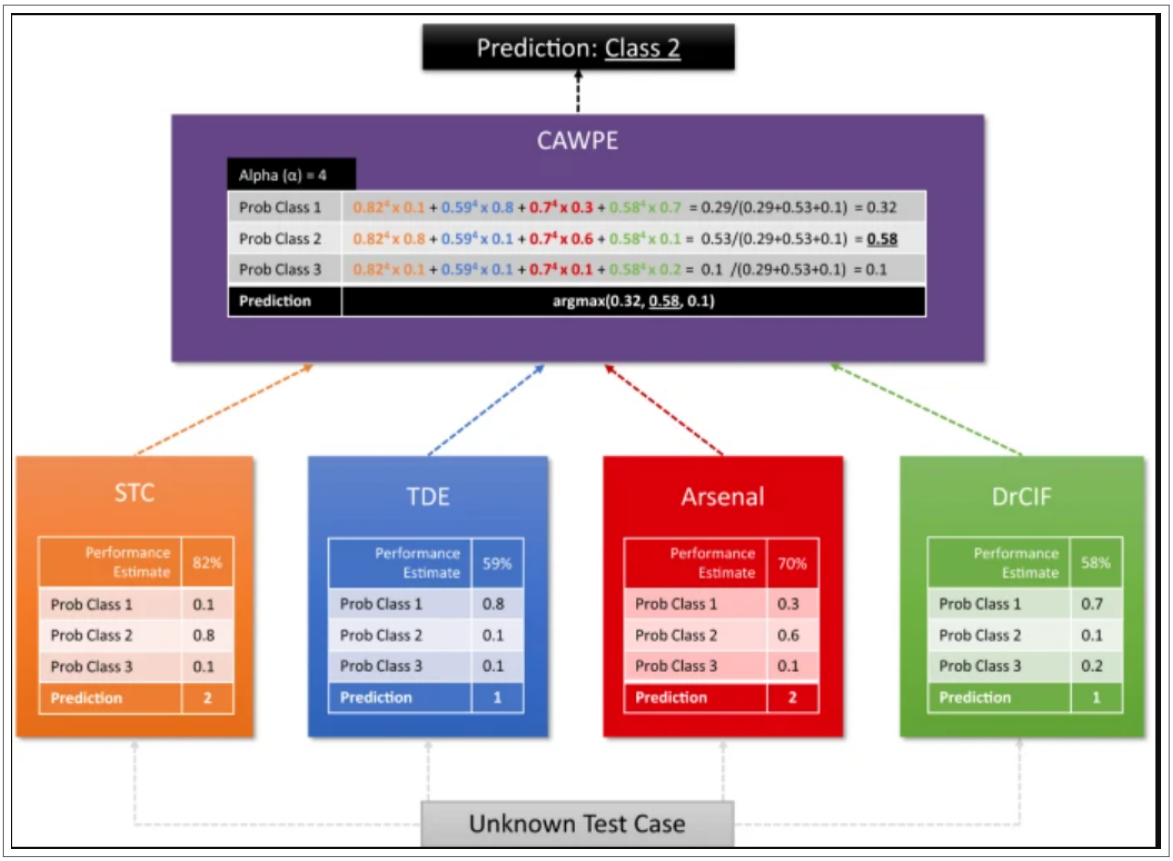
Fonte: (ZHANG YIFENG GAO, 2020).

zado de forma incremental, culminando no seu estado atual, HIVE-COTE 1.0. Durante esse tempo, vários algoritmos foram propostos que correspondem à precisão do HIVE-COTE. O HIVE-COTE V2 é um modelo híbrido, especializado em classificação de séries temporais. No artigo (MIDDLEHURST M., 2021) mostra que o HIVE-COTE V2 substituiu três dos quatro classificadores que compunham o HIVE-COTE V1 . Os novos módulos de componentes são: a transformação Shapelet baseada em Shapelet Transform Classifier; o conjunto baseado em convolução dos classificadores ROCKET, chamado de Arsenal; a representação baseada em dicionário Temporal Dictionary Ensemble (TDE); e o Diverse Representation Canonical Interval Forest (DrCIF) baseado em intervalo.

Uma visão geral da estrutura do conjunto de HIVE-COTE V2 para um problema de três classes é exemplificada na Figura 23. Cada módulo é treinado de forma independente e produz uma estimativa da probabilidade de associação de cada classe para dados não vistos. A unidade de controle Cross-validation Accuracy Weighted Probabilistic Ensemble (CAWPE) combina estas probabilidades, ponderadas por uma estimativa da qualidade do módulo encontrada nos dados do treinamento.

Assim, para novos dados, cada módulo produz uma estimativa probabilística para cada classe. O controlador constrói uma distribuição inclinada através exponenciação (usando $= 4$ por padrão) para atenuar as diferenças nos classificadores e ponderado com a estimativa de precisão. Cada módulo do HIVE-COTE V2 contém novos recursos e melhorias em relação às versões anteriores. Isso inclui um novo modelo com melhorias, extensões multivariadas e melhorias de contratação.

Figura 23 – Visão geral da estrutura atualizada do HIVE-COTE V2 para um problema de três classes. Cada módulo é treinado de forma independente e produz uma estimativa da probabilidade de associação de cada classe para dados não vistos. A unidade de controle CAWPE combina estas probabilidades, ponderadas por uma estimativa da qualidade do módulo encontrada nos dados do treinamento.



Fonte: (MIDDLEHURST M., 2021).

2.9.6 Medidas de Avaliação

Nesta secção são apresentadas algumas das medidas de avaliação mais comuns. Para calcular as fórmulas referentes às medidas de avaliação é necessário a quantificação dos seguintes termos, tendo em conta o que o algoritmo classificador previu e o que realmente são as classificações corretas (HOSSIN M ; SULAIMAN, 2015).

- Verdadeiro Positivo (VP): representa a quantidade de casos referentes à classe positiva que foram classificados corretamente. Relativamente ao tema do projeto, corresponde à quantidade de arritmias detetadas corretamente;
- Falso Positivo (FP) - representa a quantidade de casos referentes à classe negativa, mas que o algoritmo classificou como sendo da classe positiva, ou seja, corresponde à quantidade de arritmias que não foram detetadas pelo mesmo;
- Verdadeiro Negativo (VN) - representa a quantidade de casos referentes à classe negativa que foram classificados corretamente. Assim sendo, corresponde à quantidade de casos que foram detetados como não sendo arritmias corretamente;

- Falso Negativo (FN) - representa a quantidade de casos referentes à classe positiva, mas que o algoritmo classificou como sendo da classe negativa, ou seja, corresponde à quantidade de arritmias que foram detetadas pelo algoritmo incorretamente;

2.9.7 Matriz de Confusão

A matriz de confusão é utilizada em problemas de classificação binária (problema com duas classes possíveis de classificar), para encontrar a exatidão e precisão do modelo. Esta é apresentada sob a forma de uma tabela com duas entradas , uma constituída pelas classes desejadas e a outra pelas classes previstas pelo modelo. Já as células são preenchidas pelos valores que correspondem ao cruzamento das entradas.

Tabela 3 – Matriz de Confusão: uma tabela que mostra as frequências de classificação para cada classe do modelo (HOSSIN M ; SULAIMAN, 2015)

	Positivos previstos	**Negativos previstos**
Positivos originais	VP	FN
Negativos originais	FP	VN

Fonte: (HOSSIN M ; SULAIMAN, 2015).

A matriz de Confusão em si não é uma medida de desempenho, no entanto grande parte das métricas de desempenho são baseadas na matriz de confusão e nos valores que esta contém (HOSSIN M ; SULAIMAN, 2015).

2.9.8 Sensibilidade

Sensibilidade, também conhecido por Recall ou Taxa de Verdadeiros Positivos (TVP), corresponde à taxa de acerto na classe positiva (HOSSIN M ; SULAIMAN, 2015). Esta medida de avaliação, relativamente ao tema do trabalho, corresponde à proporção de pacientes que realmente possuem arritmias cardíacas, tal como o algoritmo teve como previsão. A fórmula aplicada para esta medida é a seguinte:

$$Sensibilidade = \frac{VP}{VP + FN}$$

2.9.9 Especificidade

A medida de avaliação Especificidade, Taxa de Verdadeiros Negativo (TVN), corresponde à taxa de acerto na classe negativa. Esta medida quantifica a proporção de casos negativos que foram corretamente classificados, ou seja, indica-nos a proporção de pacientes que

não possuem arritmias cardíacas, tal como o algoritmo previu (HOSSIN M ; SULAIMAN, 2015). Assim sendo, pode-se determinar que a especificidade corresponde ao oposto da sensibilidade. Para calcular esta medida aplica-se a seguinte fórmula:

$$Especificidade = \frac{VN}{VN + FP}$$

2.9.10 Precisão

A precisão é uma medida de avaliação que nos indica qual a proporção de pacientes que foram diagnosticados como tendo arritmias, e que na verdade tiveram. Os pacientes que foram diagnosticados como tendo arritmias são os VP e FP, já os que sofrem realmente desse problema são os VP (HOSSIN M ; SULAIMAN, 2015). Para avaliar o algoritmo com esta métrica, é utilizada a seguinte formula:

$$Precisão = \frac{VP}{VP + FP}$$

2.10 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foi apresentado um referencial teórico sobre o problema de reconhecimento fúngico. Foram discutidos os conhecimentos básicos acerca das leveduras estudadas neste trabalho, abordando técnicas para identificação de fungos: convencionais (CHROMagar, PCR e Sequenciamento de DNA) e alternativas (por composição química - CG, CG-EM - ou através de um nariz eletrônico). Ainda foram abordados os componentes de um nariz eletrônico: os sensores e a natureza de entrada (COVs) e saída (séries temporais); e conceitos de Inteligência Artificial, ligados ao principal componente de *software*. Para cada uma dessas metodologias, foram mostrados vantagens e desvantagens. Além disso, foram apresentados modelos de IA que serão testados para solucionar o problema desta pesquisa e também as métricas que serão utilizadas para mensurar os resultados. Dessa maneira, a partir dessa apresentação teórica conclui-se que foi possível conhecer a natureza dos dados que serão levantados na pesquisa, bem como os fundamentos que regem os métodos de identificação fúngica e as modificações e avanços nesse processo de reconhecimento. Além disso, infere-se que a utilização de tecnologias alterativas como os narizes eletrônicos e a IA podem trazer benefícios no processo de identificação de diferentes espécies de *Candida*, melhorando, por exemplo, o desempenho e a produtividade. O próximo passo será revisar a literatura para situar este trabalho com outros em áreas correlatas.

3 REVISÃO DA LITERATURA

Neste capítulo serão introduzidos os trabalhos que influenciaram os métodos propostos nessa dissertação, contendo uma breve revisão da literatura sobre o tema de interesse, abordando trabalhos relacionados nas áreas de detecção e identificação de fungos e classificação de séries temporais.

3.1 TRABALHOS RELACIONADOS A IDENTIFICAÇÃO CONVENCIONAL DE FUNGOS DO GÊNERO *CANDIDA*

Na prática clínica existem diferentes métodos que podem ser utilizados, mas que diferem quanto à sensibilidade e especificidade diagnóstica e estes aspectos estão diretamente relacionados à confiabilidade nos resultados e diretrizes de tratamento. O estudo, realizado por Valente *et al.*, fez uma revisão integrativa da literatura, evidenciando a metodologia mais eficiente para identificação fúngica (VALENTE; LOPES; REIS, 2021). Para isso, foram utilizados periódicos nacionais e internacionais indexados em diferentes bases de dados, sendo selecionados 20 artigos publicados entre os anos 2005 a 2020. Como resultados, constatou-se que o método CHROMagar possui, segundo a literatura, sensibilidade de 97% a 83,3% e especificidade de 97,9% para a identificação de *C. albicans*, porém apresenta limitações para identificação de espécies não-albicans como *C. parapsilosis*. A sensibilidade e especificidade de CHROMagar Candida, explanados no estudo, foram determinadas como 96,9% e 97,9% para *C. albicans* , 100% e 98,5% para *C. tropicalis* , 100% e 100% para *C. glabrata* e *C. krusei* e 100% e 98,3% para *C. parapsilosis*, respectivamente. Assim, identificou corretamente a maioria de *C. albicans , C. tropicalis , C. glabrata, e C. krusei* com altas sensibilidades e especificidades comparáveis. No entanto, para *C. kefyr e C. haemulonii* não puderam ser diferenciados de *C. parapsilosis*.

Já o estudo realizado por Mulet *et al.*, fez a mesma comparação, de sensibilidade e especificidade, para diferentes espécies de *Candida*, em dois meios de referência (CHROMagar™ Candida e CHROMagar™ Candida Plus), mais também em diferentes tempos de incubação (24h, 36h e 48h) (MULET et al., 2020). Apesar de encontrar resultados semelhares ao anterior, este estudo observou que embora algumas espécies possam ser presumivelmente identificadas em uma incubação de 24 ou 36 h, os melhores resultados em sensibilidade e especificidade foram obtidos em 48 h de incubação. Com isso, os resultados foram de sensibilidade e especificidade de 100% para *C. albicans, C. tropicalis e C. krusei*, para ambos os meios; 100% e 86,9% para em meio CHROMagar™ Candida e 100% e 100% em meio CHROMagar™ Candida Plus para *C. glabrata*; já as espécies *C. auris, C. lusitaniae e C. parapsilosis*, a sensibilidade e a especificidade não foram calculadas em meio CHROMagar™ Candida, pois este meio não as diferenciam, porém em meio

CHROMagar™ Candida Plus, foram encontradas, 100%, 80% e 100% de sensibilidade, respectivamente e 100% de especificidade para estas espécies.

Contudo, de acordo com Zhai *et al.*, os testes fenotípicos, como o CHROMagar, geralmente levam mais de 48 horas para completar a identificação da levedura podendo ainda apresentar resultados equivocados (ZHAI et al., 2018). Isso gera atrasos na introdução de medidas terapêuticas eficazes para o tratamento da infecção, aumentando os custos e os riscos de vida que essas infecções apresentam quando agravadas. Sendo assim, são necessários métodos diagnósticos que permitam a identificação da levedura de forma rápida, precisa e eficaz, como os métodos moleculares (SIQUEIRA J. P. Z.; ALMEIDA, 2018). Segundo Zhai *et al.*, várias abordagens estão sendo desenvolvidas para permitir uma identificação mais rápida e segura de espécies fúngicas para substituir os problemas de imprecisão e demora dos testes fenotípicos tradicionais (ZHAI et al., 2018). As técnicas moleculares realizam amplificação nas regiões altamente conservadas do DNA e que estão presentes em fungos patogênicos e assim é possível identificar as espécies de *Candida* pelos diferentes formatos de PCR.

Nos últimos 10 anos, vários métodos não baseados em cultura tentaram superar as limitações dos diagnósticos convencionais baseados em cultura. Existem vários testes de diagnóstico comercialmente disponíveis, estabelecidos e reconhecidos para a detecção de *Candida* (KOC et al., 2022). Nas pesquisa realizadas por Zhang *et al.*, encontra-se um levantamento bibliográfico e um comparativo de diferentes métodos de PCR, aplicados para diversas espécies de *Candida* (J HUNG GC, 2016). No geral, a tabela a seguir ilustra as tecnologias adotada nesse artigo.

Tabela 4 – Sensibilidade e Especificidade de técnicas associadas para o diagnóstico de Candidemia

Tecnologia	**Sensibilidade**	**Especificidade**
Light cycle	95%	97%
Multiplex PCR	98%	88%
Nested PCR	86%	54%
PCR Elisa	83%	92%
PCR sequencing	72%	91%
TaqMan	100%	97%

Fonte: (J HUNG GC, 2016).

Em 2021, o estudo realizado por Valente *et al.*, observou o avanço da técnica Nested-PCR e a melhorias dos resultados encontrados, que obteve uma sensibilidade e especificidade de 100% e 97%, respectivamente, para a identificação de *Candida* spp (VALENTE; LOPES; REIS, 2021). Além disso, em 2019, um estudo realizou um comparativo de desempenho de ensaio de PCR em tempo real multiplex específico para *Candida* (Fungiplex® Candida IVD Real-Time PCR Kit) e outro ensaio de PCR em tempo real para diagnós-

tico estabelecido (LightCycler SeptiFast Test) em relação à detecção de *Candida* a partir de amostras de sangue total (FUCHS; LASS-FLöRL; POSCH, 2019). Amostras clínicas de 58 pacientes foram analisadas por hemocultura padrão e testadas simultaneamente com o Fungiplex Candida PCR e o teste SeptiFast para detecção molecular de *Candida* spp. Na tabela a seguir encontra-se os resultados explanados nesse artigo.

Tabela 5 – Comparativo de desempenho de técnicas PCR para ensaios de amostras clínicas de 58 pacientes por hemocultura padrão

Tecnologia	**Sensibilidade**	**Especificidade**
Fungiplex Candida PCR	100%	94.1%
LightCycler SeptiFast Test	60%	96.1%

Fonte: (FUCHS; LASS-FLöRL; POSCH, 2019).

Apesar disso, estes métodos possuem algumas limitações como o alto custo e a possibilidade de resultados falso-positivos por contaminação cruzada (*carryover*), além de necessitar de pessoal capacitado na interpretação dos resultados. Com o avanço tecnológico, a inteligência artificial passou a fornecer soluções para vários problemas e a ter um desempenho ainda melhor em comparação com os humanos em diferentes domínios de aplicação, como diagnóstico médico (SRI; KARUNA, 2021).

3.2 TRABALHOS RELACIONADOS A IDENTIFICAÇÃO FÚNGICA BASEADO EM APRENDIZAGEM DE MÁQUINA

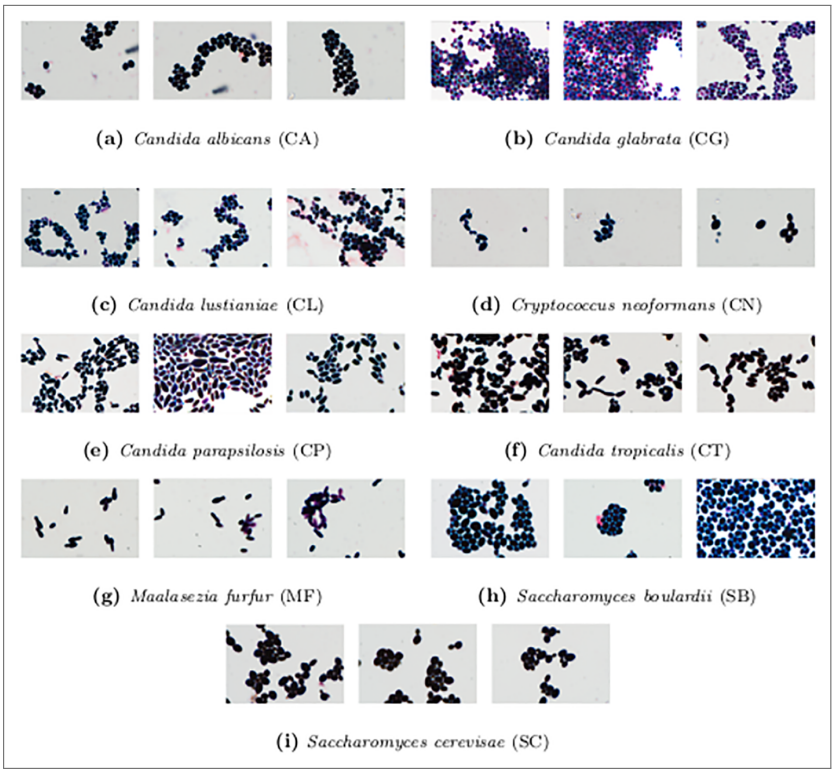
A capacidade dos modelos de inteligência artificial de aprender os recursos hierárquicos de imagens, tornou o diagnóstico microbiano muito poderoso e capaz de extrair características complexas de alto nível como representações de dados. Embora muito trabalho esteja sendo feito no campo da análise de imagens médicas, a classificação de imagens microscópicas de fungos é uma área relativamente nova (SRI; KARUNA, 2021).

Em 2020, Zieliński *et al.* propôs um método com uma abordagem de aprendizado de máquina baseada em redes neurais profundas para classificar imagens microscópicas de espécies de fungos (ZIELIńSKI et al., 2020). Dessa maneira, este método, baseados na coloração microbiológica básica (coloração de Gram), utiliza uma câmera acoplado ao microscópio simples para identificação de nove espécies fúngicas (*Candida albicans, Candida glabrata, Candida lustianiae, Cryptococcus neoformans, Candida parapsilosis, Candida tropicalis, Candida krusei, Saccharomyces phylum, Saccharomyces boulardii E Maalasezia furfur*).

Nesta pesquisa, as cepas foram cultivadas em ágar Sabouraud a 37°C por 48h. Após esse tempo, foram feitas preparações microscópicas e coradas pelo método de Gram. As imagens, ilustradas na Figura, foram tiradas usando um microscópio (*Olympus BX43*),

uma uma câmera fotográfica (*Olympus BP74*) e *software CellSense*. Como método de linha de base, o estudo adotou o bloco do classificador do arquiteturas de rede bem conhecidas para de imagens como *AlexNet, InceptionV3, ResNet* e variações estas redes utilizando Support Vector Machine (SVM) e Random Forest (RF). Esta abordagem realizou diversos testes baseados em varredura por agregamento em *patch*, e os melhores resultados encontrados, apresentados na tabela 6, foi a partir da utilização da rede AlexNet, com mecanismo de agregação de *patches* baseado no Fisher Vector (FV) e o algoritmo SVM.

Figura 24 – Exemplos de imagens tiradas de uma câmera acoplado ao microscópio simples para identificação de espécies fúngicas



Fonte: (ZIELIńSKI et al., 2020).

Tabela 6 – Resultados de Testes pela AlexNet FV SVM: rede escolhida por uma abordagem para descrever e classificar imagens microscópicas de fungos

Espécies	**Acurácia**
Candida albicans, Candida lustianiae, Candida parapsilosis, Candida tropicalis, Candida krusei, Saccharomyces phylum	$100\% \pm 0\%$
Maalasezia furfur	$95\% \pm 5\%$
Cryptococcus neoformans	$75\% \pm 15\%$
Candida glabrata	$75\% \pm 25\%$

Fonte: (ZIELIńSKI et al., 2020).

Em 2021, Bettauer *et al.* apresentou uma abordagem, também utilizando *Deep Learning* para explorar a complexa variedade de morfologias exibidas pela *C. albicans* (BETTAUER et al., 2021). O sistema desenvolvido, chamado de *Candescence* tem como objetivo principal detectar automaticamente células da microscopia utilizando contraste de imagem diferencial e rotular cada célula detectada com uma das nove morfologias vegetativas. Assim, o foco desse estudo é capturar a essência apenas da *C. albicans*, sua morfologia, desenvolvendo modelos usando redes adversarias generativas e identificando subcomponentes do espaço latente que controlam variáveis técnicas, trajetórias de desenvolvimento ou mudanças morfológicas. O desempenho (acurácia) encontrada nessa abordagem foi de aproximadamente 82%.

Sabe-se que os microrganismos produzem uma ampla gama de produtos metabólicos, que são produtos finais de processos bioquímicos e resultado de interações ambientais e genéticas, como metabólitos voláteis que podem ser usados como impressões digitais metabólicas voláteis únicas de cada espécie. A pesquisa realizada por Costa *et al.* explorou uma estratégia metabolômica[5] baseada no uso de uma cromatografia gasosa multidimensional avançada para o mapeamento abrangente de metabólitos celulares de *C. albicans e não-C. albicans (C. glabrata e C. tropicalis)* (COSTA C. P., 2020). Esta pesquisa representa um estudo detalhado sobre o exometaboloma[6] das espécies de *Candida*, e não de classificação. Os resultados apresentados nesta pesquisa indicam um aumento de 70% de metabólitos não relatados anteriormente para *C. albicans*, 91% para *C. glabrata* e 90% para *C. tropicalis*. Contribuindo, assim, para o desenvolvimento de uma plataforma abrangente para o manejo da detecção de fungos e para a distinção de espécies.

Segundo Mota *et al.*, sistemas de nariz eletrônico estão sendo propostos como alternativas adequadas às técnicas de identificação de fungos atualmente disponíveis. Através de uma revisão sistemática da literatura de 16 artigos, os estudos revisados demonstraram que a detecção eficaz de fungos foi possível por meio de sistemas de nariz eletrônico baseados em sensores, que podem realmente funcionar como uma ferramenta de triagem de Compostos Orgânicos Voláteis (COVs) para várias aplicações (MOTA; TEIXEIRA-SANTOS; RUFO, 2021).

Em 2020, Loulier *et al.* testou com o sistema *e-nose* culturas alguns microrganismos como *Armillaria gallica, Armillaria ostoyae, Fusarium avenaceum, Fusarium culmorum, Fusarium oxysporum, Fusarium poae, Rhizoctonia solani, Trichoderma asperellum, Phytophthora cactorum, P. cinnamomi, P. plurivora, P. ramorum.* Os resultados sugeriram que as principais diferenças entre as respectivas faixas de emissão de COVs das espécies testadas residem na produção de um componente volátil: o sesquiterpeno. Nesse sistema, o nariz eletrônico pode discriminar entre os odores emitidos por *P. ramorum, F. poae, T. asperellum e R. solani*, que representaram mais de 88% da variação do Principal Compo-

[5] Estudo do conjunto de metabólitos (produto do metabolismo de uma determinada molécula ou substância) de um determinado sistema biológico

[6] Conjunto de metabolitos excretados para o meio que envolve a célula (ou fluídos extracelulares).

nent Analysis (PCA).

3.3 TRABALHOS RELACIONADOS A IDENTIFICAÇÃO DE GASES BASEADO EM APRENDIZAGEM DE MÁQUINA

Sabe-se que o principal componente de *hardware* de um nariz eletrônico é um conjunto de sensores de gás capazes de interagir com uma ampla gama de produtos químicos de vários compostos. As partículas de gás estimulam os sensores na matriz, o que gera uma resposta característica de série temporal. O sistema de reconhecimento de padrões passa então a processar as formas de onda geradas, extraindo, filtrando e selecionando informações úteis, além de realizar a classificação dos dados desse gás (YAN XIUZHEN GUO; ZHANG, 2015).

O sistema de reconhecimento faz uso de modelos de Machine Learning (ML). Muitas abordagens usaram algoritmos clássicos de ML, como k-Nearest Neighbor (kNN),Artificial Neural Network (ANN) e SVM para classificar a série temporal multivariada (WANG et al., 2021) (ZHENG et al., 2014). Para que um nariz eletrônico seja confiável, precisamos de alta precisão na classificação e um modelo robusto capaz de lidar com dados de ruído e com mínimo ou nenhum resultado falso-negativo.

Muitas abordagens começaram a converter a série temporal (em inglês, Times Series (TS)) do gás em imagens e usá-lo com algoritmos de aprendizado profundo para classificar dados de gás. A ideia principal por trás da lógica de conversão de alguns estudos é criar uma matriz com informações de relacionamento entre os pontos do TS e como eles estão correlacionados, usando algoritmos como Gramian Angular Field (WANG; OATES, 2015) e Markov Transition Field (GAMBOA, 2017) .

Em 2018, o modelo chamado como *GasNet* foi o primeiro modelo proposto a utilizar a ideia de rede neural convolucional profunda (em inglês, Deep Convolutional Neural Network (DCNN)) para classificação de gás (PENG P.; ZHAO, 2018). Este modelo trabalha com a ideia de mapear dados de série temporal do gás original em uma matriz de imagem análoga e usar um modelo clássico existente da rede neural convolucional (CNN) para extrair recursos e classificar as mistura de gases. A rede proposta neste artigo foi composta por 38 camadas divididas entre camadas convolucionais, uma camada *pooling* e uma camada totalmente conectada. O método DCNN é usado para classificar dados de imagem, portanto, em seu trabalho, eles precisaram adaptar a entrada de uma CNN para receber dados de matriz de gás. Eles obtiveram um ganho de 15,3% na precisão em relação aos resultados anteriores (SVM), seus resultados foram de 95,2% na precisão.

Seguindo a mesma abordagem, o estudo de Han *et al.* foi capaz de classificar e comparar cinco tipos de gases misturados mapeando dados de séries temporais em dados de matriz de imagens analógicas e usando cinco tipos de redes neurais convolucionais (HAN et al., 2019). Depois disso, este mesmo estudo ajustaram os parâmetros das CNN e obtiveram uma taxa final de reconhecimento de gás de 96,67%.

Wang *et al.* propôs diferentes abordagens para classificar dados de gás usando não apenas algoritmos ML convencionais, mas também métodos Deep Neural Network (DNN), incluindo uma abordagem híbrida (WANG et al., 2021). Em seu trabalho, eles usaram a rede GasNet para comparar com resultados anteriores e propuseram uma nova arquitetura CNN chamada SimResNet-9, que é baseada na ResNet (HE et al., 2015), uma arquitetura criada para reconhecimento de imagem. Neste estudo, realizou-se o pré-processamento dos dados, convertendo o TS em imagens e inserindo-os em sua arquitetura profunda. Os autores exploraram três métodos de pré-processamento e também o impacto do ruído externo nos dados. O método SimResNet-9 proposto obteve resultados de acurácia superiores aos trabalhos anteriores alcançando uma acurácia de classificação de 95% para a base de dados utilizada.

No entanto, na classificação das imagens a posição e a ordenação das informações são recursos importantes, por isso, nesse método a direção de operação de convolução é limitada a uma única e determinada ordem. Dessa maneira, uma mudança na direção do desenvolvimento de dados da série temporal pode levar a grandes mudanças nas tendências e localizações das curvas. Por isso, o resultado da classificação pode ser afetado diretamente caso haja perda de informações na mudança de direção, por exemplo. Portanto, mais recentemente, para superar algumas limitações na transformação de TS em imagem, novos modelos foram criados especialmente para abordar problemas de classificação para natureza desses dados.

Em 2022, Castro *et al.* apresentou um estudo para identificação de espécies de *Candida* por meio da análise de séries temporais de compostos orgânicos voláteis de culturas adquiridas e interpretadas por métodos de *e-nose* e Inteligência Artificial (IA). Segundo este estudo, a abordagem proposta contribui para estabelecer um tratamento ágil e adequado, reduzindo as complicações da doença e o número de óbitos. No entanto, este estudo trabalhou apenas com três espécies de *Candida* (*C. albicans, C. parapsilosis e C. krusei)*. Por isso, incentiva que outros trabalhos lidem com mais tipos de fungos, pois quanto mais tipos de fungos forem explorados, mais ampla será a gama de problemas e patologias cobertas, e, dessa forma, os novos estudos podem auxiliar na criação de uma nova metodologia de identificação mais rápida e confiável usando métodos de inteligência artificial.

3.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram apresentados trabalhos que utilizam os conceitos e princípios, explanados nesta dissertação, para detecção e identificação de fungos. Foram abordados trabalhos que usaram técnicas convencionais e por modelos de inteligência artificial para identificação fúngica. Além de apresentar estudos que utilizaram *e-nose* para identificação de gases, englobando modelos de classificação específicos para séries temporais. No entanto, apesar de apresentarem bons resultados, ainda apresentam limitações, como, para nos métodos convencionais, o alto custo, a necessidade de profissional capacitado

para manipulação e o tempo para completar a identificação; já para os métodos baseados em Aprendizagem de Máquina (AM), a possibilidade de resultados falso-positivos devido a baixa a cobertura de espécies estudadas e as restrições nos métodos de IA, devido a natureza dos dados analisados (imagens ou séries temporais). Dessa maneira, infere-se a necessidade de aprofundar o estudo de métodos que ampliem a cobertura de espécies de *Candida* com um baixo custo, e que sejam especializados na natureza de dados de saída de um nariz eletrônico (as séries temporais). Os próximos passos serão apresentar a metodologia utilizada nesta dissertação, apresentando o nariz eletrônico (responsável por coletar os COVs emitidos pelas colônias de fungos), a construção da base de dados e os procedimentos de processamento aplicados.

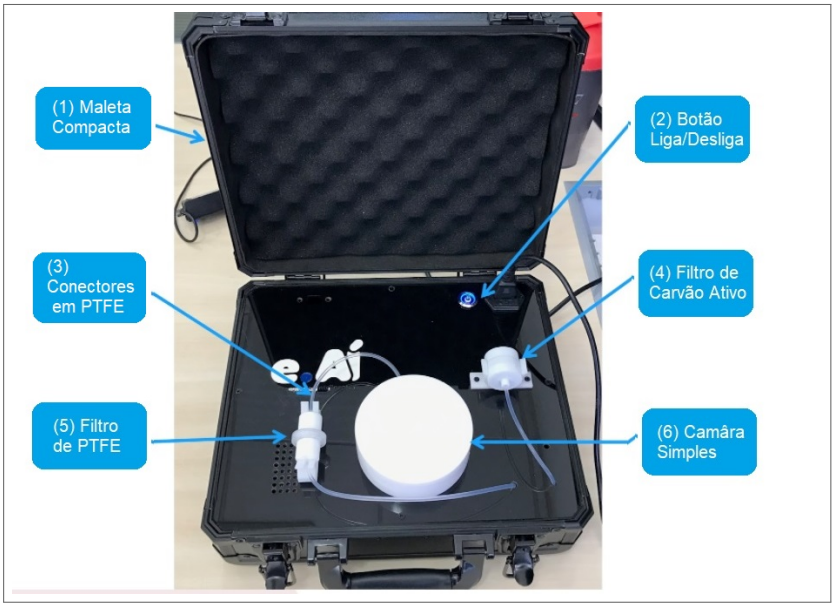
4 MATERIAIS E MÉTODOS

Neste capítulo será discutido a metodologia executada neste trabalho que englobam o funcionamento do nariz eletrônico, a construção da base de dados e os procedimentos de pré-processamentos aplicados sobre os dados para obter os resultados da classificação de séries temporais.

4.1 VISÃO GERAL

O protótipo de nariz eletrônico foi construído por pesquisadores da UFPE, IFPE, CRCN-NE e com financiamento da *e.Aí Tecnologias Inteligentes LTDA*, e foi utilizado nesse estudo para a captação dos sinais eletrônicos gerados pelas leituras dos COVs produzidos pelos fungos.Conforme pode ser observado na Figura 25, o *eAi* é um sistema acoplado dentro de uma maleta compacta, composto por dispositivos físicos (*hardwares*), sistemas computacionais (*softwares*) e um protocolo de coleta de dados.

Figura 25 – Protótipo de nariz eletrônico, um sistema acoplado dentro de uma maleta compacta, composto por dispositivos físicos (*hardwares*), sistemas computacionais (*softwares*) e um protocolo de coleta de dados.



Fonte: *eAi* (2022).

Os principais componentes de hardwares, são visíveis na figura anterior:

1. O nariz é acoplado a uma maleta compacta e portátil;
2. O sistema é ativado ou desativado facilmente por um botão de liga/desliga;
3. Todas as conexões e compartimentos são feitos de PTFE;

4. Possui um filtro de carvão ativado;

5. Possui um filtro de PTFE;

6. Câmara de Ar, com sensores acoplados a ela, é o local onde será realizado a coleta;

Além desses, o dispositivo contém outros componentes físicos dispostos internamente, como:

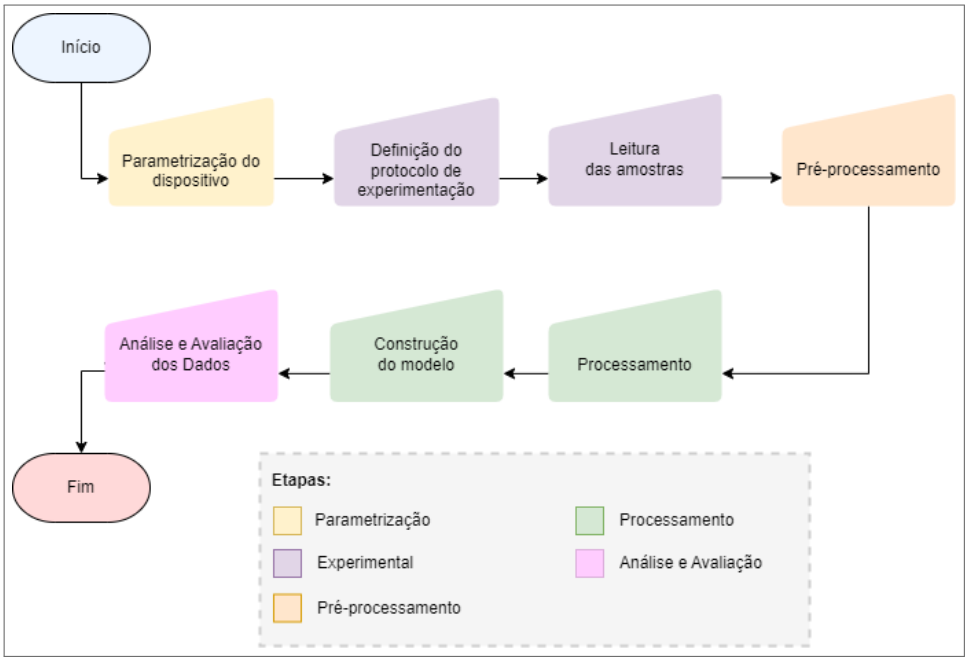
1. Sensores Eletrônicos (*E-sensores*), responsáveis pelas leituras ou coleta de informações dos compostos voláteis liberados pelas amostras ao longo do tempo;

2. Válvulas de Controle e

3. uma Bomba de ar, responsável pela aspiração ou injeção de gases ou ar na câmara.

A metodologia executada se divide basicamente em cinco etapas: parametrização do equipamento, estabelecimento do protocolo experimental (onde inclui o processo de definição do protocolo e a leitura das amostras), pré-processamento (onde inclui o aplicação das técnicas), processamento (onde inclui a aplicação dos métodos de IA e construção do modelo) e análise e avaliação dos dados. A Figura 26 ilustra o fluxograma referente a metodologia utilizada, que será descrita com mais detalhes nas seções subsequentes.

Figura 26 – Fluxograma ilustrativo da metodologia adotada. Esta metodologia divide-se em cinco etapas: parametrização do equipamento, estabelecimento do protocolo experimental, pré-processamento, processamento e análise e avaliação dos dados



Fonte: A autora (2023).

4.2 PARAMETRIZAÇÃO DO EQUIPAMENTO

Esta etapa pode ser subdividida em seleção dos sensores a serem utilizados na matriz de sensores e parametrização do *hardware* do dispositivo. Este dispositivo possui dez *e-sensores*: sete detectores de gases específicos, e os outros três detectam a temperatura (Cº), a pressão (kPa), a umidade (%), respectivamente. Os sensores de gás específicos, feitos de semicondutores de óxido metálico *Metal Oxide* (MOX), pertencem a família TGS, fabricados pela *FIGARO Engineering Inc*, sendo eles: TGS826, TGS2611-E00,TGS2603, TGS813, TGS822, TGS2602 e TGS823.

O contato das moléculas gasosas com a superfície dos sensores provocam mudanças no sinal de saída de acordo com a composição e intensidade da massa de ar a ser analisada. Mais especificamente, a resistividade dos sensores aumenta na presença de ar e diminui na presença dos gases sensíveis. A Tabela 7 mostra a faixa de detecção de partículas, conforme os *datasheets* dos Figaro®.

Tabela 7 – Escopo de Detecção de Concentração de partículas para os sete sensores MOX utilizados

Sensor	**Faixa de Detecção de Concentração**
TGS826	10-1000 ppm iso-butano, hidrogênio, amônia e etanol
TGS2611-E00	100-100000 ppm etanol, hidrogênio e metano
TGS2603	0,1-3 ppm trimetilamina; 0,3-3 ppm H2S e metil mercaptano; 1-30 ppm etanol e H2
TGS813	500-100000 ppm CO, metano, etanol, propano, iso-butano e hidrogênio
TGS822	50-5000 ppm metano, CO, n-Hexano, benzeno, etanol e acetona
TGS2602	0,1-3 ppm H2S; 1-30 ppm H2, NH3, etanol e tolueno
TGS823	50-5000 ppm metano, CO, n-Hexano, benzeno, etanol e acetona

Fonte: A autora (2023).

Esses sensores de gás foram escolhidos por serem os mais comumente usados para a construção de narizes eletrônicos, devido ao menor custo, confiabilidade e a fácil aquisição em comparação as demais. Além disso, escolha de diferentes sensores foi afim de possibilitar a criação de uma impressão digital de cada amostra, uma vez que cada sensor tem uma resposta diferente para uma determinada amostra e o conjunto das respostas gera uma saída para cada tipo de gás, aliado ao fato de que cada sensor trabalha com um espectro e focos de detecção para diferentes gases. A Tabela 8 resume o foco principal e as principais funcionalidades de cada sensor empregado neste estudo. Isso demonstra que a seleção de cada sensor foi cuidadosamente considerada para maximizar a obtenção de características da impressão digital, permitindo a identificação dos sensores mais relevantes e, quando apropriado, a eliminação de sensores redundantes.

Tabela 8 – Foco de Detecção ou Principal Funcionalidade dos sensores MOX utilizados

Sensor	**Principal Funcionalidade**
TGS826	Detecção de Amônia
TGS2611-E00	Detecção de Metano
TGS2603	Detecção de odores e contaminantes do ar (Alta sensibilidade à série de aminas e gases com odor sulfuroso e alta sensibilidade a odores de alimentos)
TGS813	Detecção de gases combustíveis (Alta sensibilidade ao metano, propano e butano)
TGS822	Detecção de Vapores de Solvente (Alta sensibilidade ao álcool e solvente orgânico)
TGS2602	Detecção de contaminantes do ar (Alta sensibilidade a contaminantes gasosos do ar)
TGS823	Detecção de Vapores de Solventes Orgânicos (Alta sensibilidade a vapores de solventes orgânicos, como etanol)

Fonte: A autora (2023).

4.3 PROTOCOLO DE EXPERIMENTAÇÃO

4.3.1 Amostras de Leveduras do gênero *Candida* spp

O sucesso na visualização e isolamento do agente fúngico depende, além da coleta, transporte, conservação e armazenamento adequados e volume suficiente da amostra, de seu processamento correto antes do exame micológico (ANVISA, 2013). Por isso, todo processo de coleta, isolamento e incubação de amostras foi realizado pelo Laboratório de Micologia Médica/UFPE.

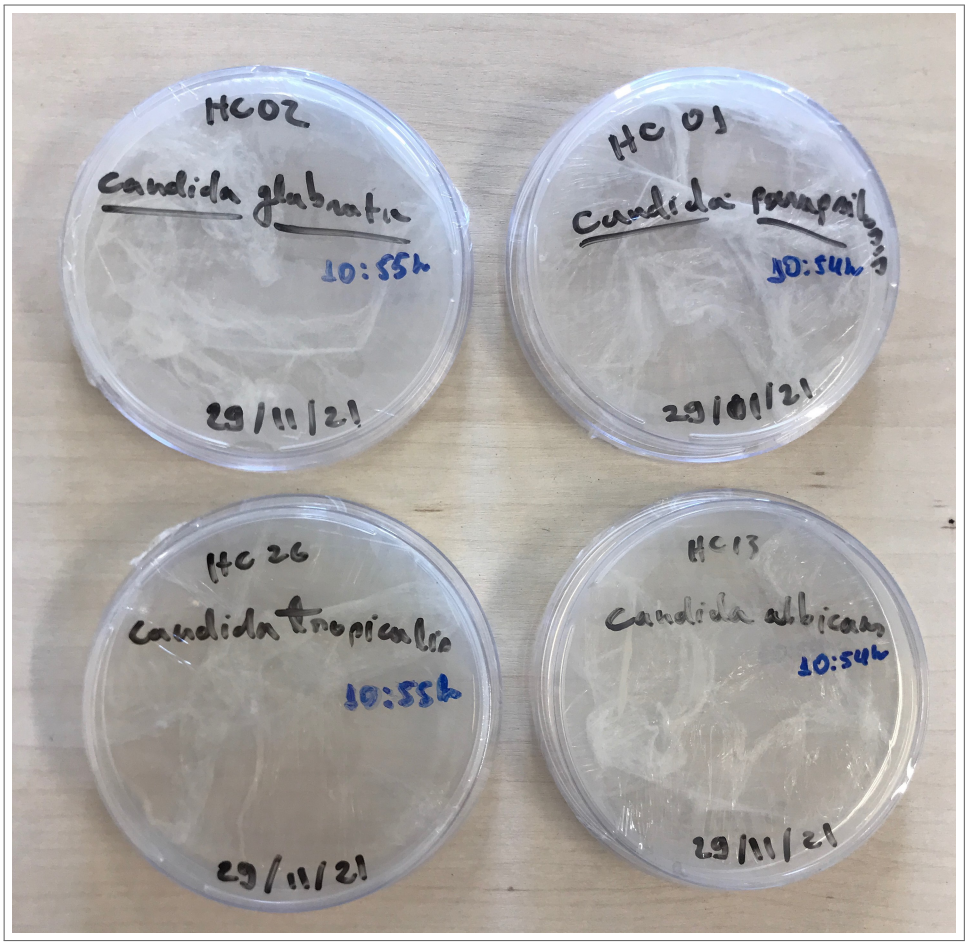
As coletas das amostras clínicas foram realizadas em pacientes internados no Hospital das Clínicas/UFPE, preparadas em lâminas sem adição de corante ou clarificante e, quando necessário, coradas com Giemsa. Concomitantemente, as amostras foram semeadas em duplicata na superfície do meio Sabouraud Dextrose Ágar chamado de DIFCO adicionado de 50 mg/L de cloranfenicol contido em placas de *Petri*, mantidas à temperatura de 30 °C e 37 °C por até 15 dias. Além disso, essas placas foram rotuladas por espécie e data, para serem encaminhadas pro Centro de Informática. A Figura 27 apresenta algumas destas placas.

Para este trabalho, iremos tratar com seis diferentes espécies: *Candida Albicans, C. Glabrata, C. Haemulonii, C. Kodamaea ohmeri, C. Krusei, C. Parapsilosis e C. Tropicalis*

4.3.2 Coleta de dados

O nariz eletrônico utilizado trabalha em 3 macro-etapas: Aspiração dos gases, Espera de estabilização do sinal dos sensores e Restauração do domínio de Coleta. Para cada uma dessas etapas foi adotado um espaço de tempo de execução de respectivamente, 20

Figura 27 – Isolamento de *Candida* spp em placas de *Petri*, semeadas por matériais de pacientes internados no Hospital das Clínicas/UFPE, rotuladas por espécie e data. Estas placas foram preparadas em lâminas sem adição de corante ou clarificante e, quando necessário, coradas com Giemsa. Concomitantemente, em duplicata na superfície do meio Sabouraud Dextrose Ágar chamado de DIFCO adicionado de 50 mg/L de cloranfenicol contido em placas de *Petri*, mantidas à temperatura de 30 °C e 37 °C por até 15 dias.

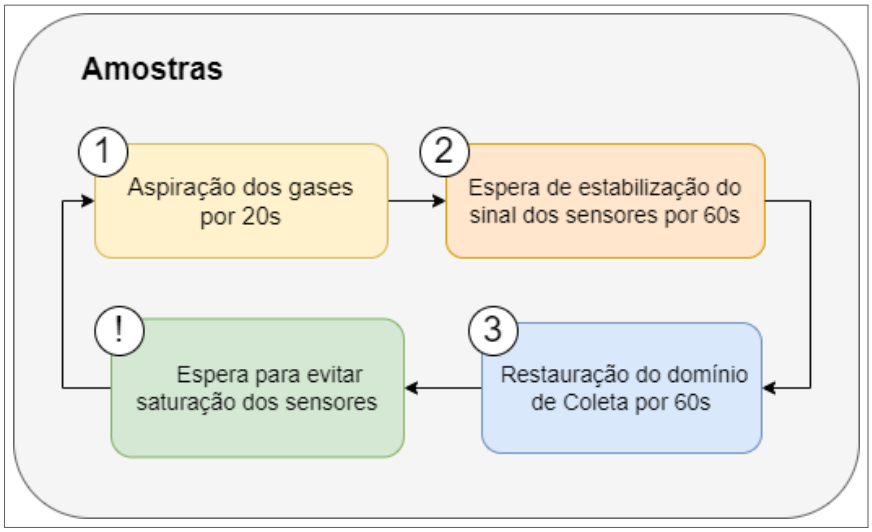


Fonte: A autora (2023).

segundos, 60 segundos e 60 segundos. A primeira etapa é a aspiração dos COVs da câmara de amostra (projetada para encaixar uma placa de *Petri* aberta) para a câmara de sensores por um espaço de tempo. A segunda etapa é a estabilização dos COVs com os sensores na câmara de sensores, também por um espaço de tempo. A terceira e última etapa é a purga/limpeza da câmara de sensores através da injeção de ar filtrado por carvão ativado, mais uma vez por um espaço de tempo.

A junção dessas três etapas, para este trabalho, é chamado de ciclo. Portanto, um ciclo dos experimentos totaliza 140 segundos. Além disso, neste equipamento, ainda existe duas variáveis programável, a quantidade de ciclos (que define a quantos ciclos serão realizados por coleta estudada) e o espaço de tempo entre ciclos (afim evitar a saturação dos sensores pela emissão de COVs). Todo esse conjunto de ciclos formam o que chamamos de *amostra*. De maneira geral, a Figura 28 ilustra uma visão macro do fluxo do funcionamento dos ciclos do nariz eletrônico utilizado nos experimentos.

Figura 28 – Macro-etapas do funcionamento dos ciclos do nariz eletrônico utilizado nos experimentos: Aspiração dos gases, Espera de estabilização do sinal dos sensores e Restauração do domínio de Coleta. A junção dessas três etapas, para este trabalho, é chamado de ciclo. Portanto, um ciclo dos experimentos totaliza 140 segundos. Além disso, há uma espaço de tempo entre ciclos afim de evitar saturação dos sensores.



Fonte: A autora (2023).

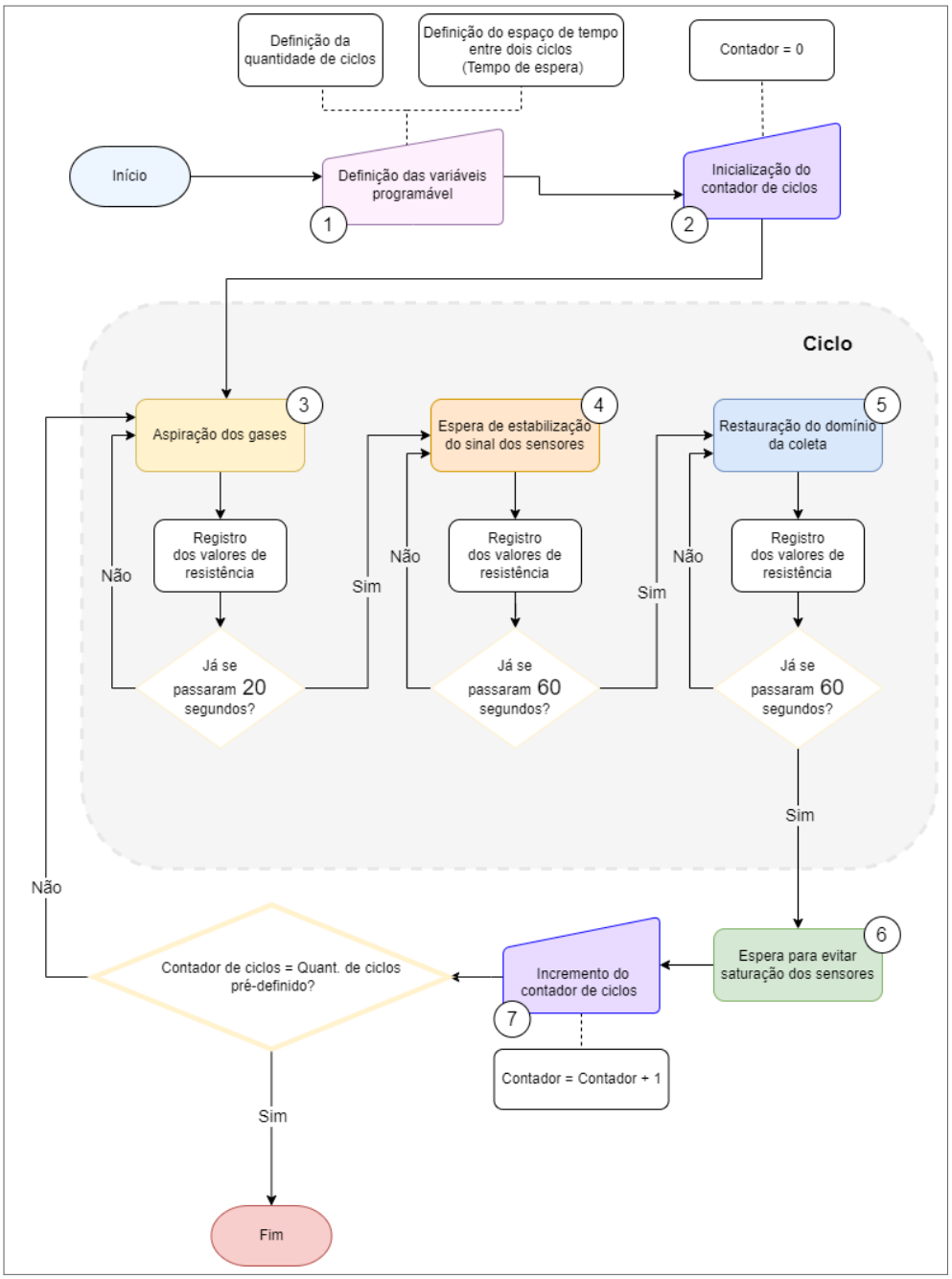
4.3.3 Roteiro de coleta de uma amostra

Cada amostra é constituída pelo registro da modificação da resistência do sensor semicondutores de óxido metálico (MOX) quando exposto a um conteúdo gasoso. O roteiro começa pela definição manual da quantidade de ciclos que serão coletados. Depois disso, o processo automatizado inicia-se. Como citado na subseção 4.3.2, o ciclo é dividido em 3 etapas principais (Aspiração, Estabilização e Restauração/Purga), para cada uma dessas etapas são registrados os valores de resistência. Logo após, o equipamento fica em repouso, onde alguns registros são realizados. A cada finalização de ciclo um contador é incrementado. O roteiro encerra-se quando valor desse contador se iguala a quantidade de ciclos pré-estabelecida. A Figura 29 ilustra o fluxo do funcionamento da coleta de dados pelo *e-nose* utilizado.

A Figura 30 ilustra a resposta de curva de um sensor MOX, o TGS826, para três ciclos. Além disso, nessa figura cada uma das etapas do ciclo foram ilustradas por cores distintas: Verde, Azul e Amarelo (Principais Etapas), Preto (Em repouso). A partir dessa figura, pode-se observar a forma de onda de um ciclo e evidenciar a representação gráfica da forma com que uma onda evolui ao longo do tempo. Este tipo de observação será essencial na comparação de amostras com diferente conteúdos, seja por espécies distintas ou por alteração de parâmetros, como quantidade de ciclos.

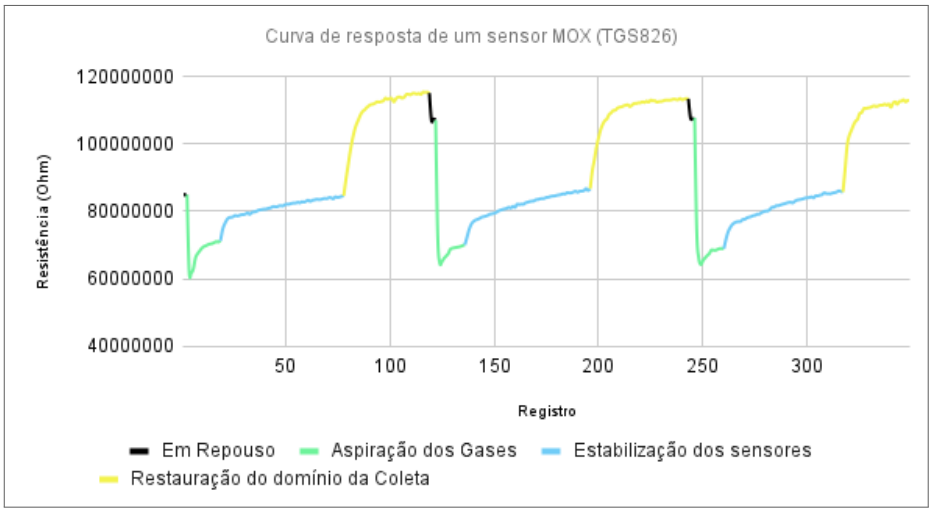
Além disso, afim de validar comportamento dos COVs exalados após um longo período de tempo, outro parâmetro adotado neste trabalho foi o tempo de coleta após o cultivo, ou melhor, o tempo de incubação. Este tempo, nada mais é, do que o intervalo entre a hora da coleta da amostra e a hora de semeação da levedura, registrado na Placa de *Petri*.

Figura 29 – Visão mais detalhadas das etapas do funcionamento dos ciclos do nariz eletrônico utilizado nos experimentos. O roteiro começa pela definição manual da quantidade de ciclos que serão coletados. Depois disso, o processo automatizado inicia-se. Durante 20 segundos são registrados os valores de resistência para a primeira etapa, a Aspiração, logo após inicia-se a etapa de Estabilização, registrando as resistências durante 60 segundos. E por último, inicia-se o processo de Restauração/Purga, que também dura 60 segundos. O equipamento fica em repouso, onde alguns registros são realizados. A cada finalização de ciclo um contador é incrementado. O roteiro encerra-se quando valor desse contador se iguala a quantidade de ciclos pré-estabelecida.



Fonte: A autora (2023).

Figura 30 – Registro da modificação da resistência do sensor TGS826 quando exposto a um conteúdo gasoso, ilustração de uma amostra contendo três ciclos com todas as suas etapas.



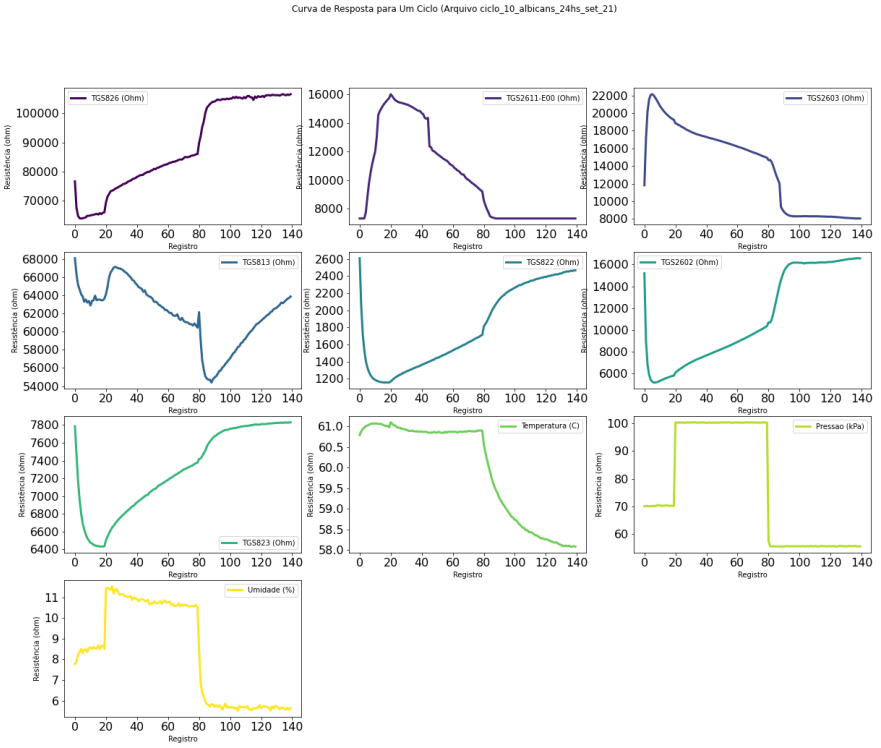
Fonte: A autora (2023).

No geral, foram coletadas amostras com 1h, 3h, 24h ou 1 semana (168h) após o semeio. Assim, foram gerados 69 amostras, com diferentes valores de ciclo e tempo de incubação.

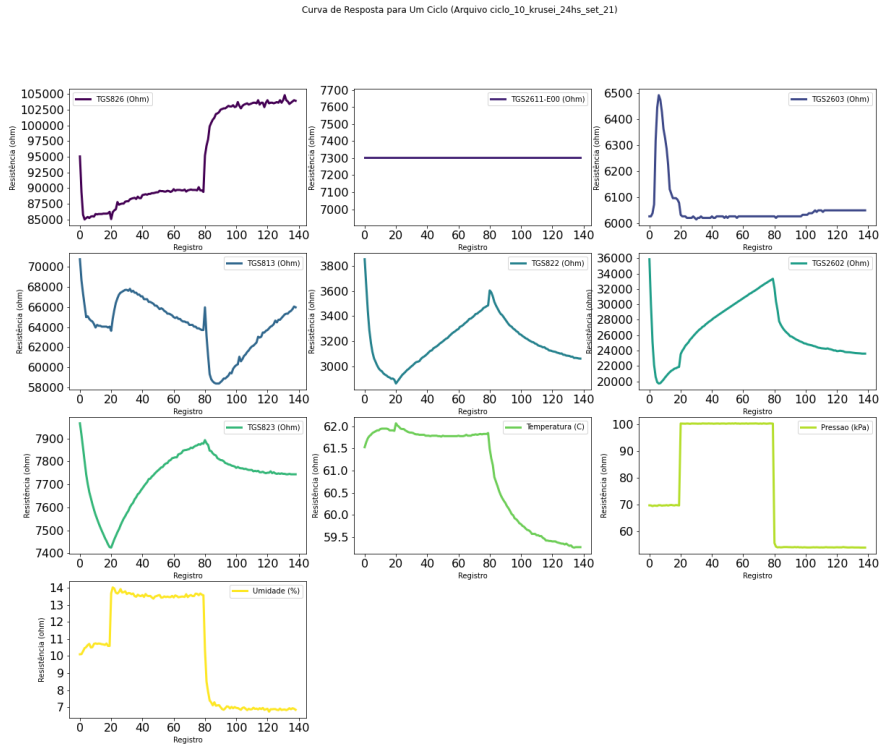
4.3.4 Curvas de Resposta por Gênero de Candida *spp*

Como citado na seção 2.5, cada sensor MOX, possui uma sensibilidade de resposta focada a uma ou mais substância. Dessa maneira, o conteúdo gasoso analisado, produzido por um fungo, pode ter, ou não, suas características típicas representados por um dado sensor. Na Figura a seguir, ilustra-se curva resposta de um único ciclo para duas espécies distintas: *C. Albicans e C. Krusei*, para os 10 sensores utilizados neste estudo. Observa-se que o sensor TGS2611-E00 é constante para *C. Krusei* enquanto que para *C. Albicans* possui uma forma de curva característica. No Apêndice A, observa-se esta mesma representação gráfica, para as outras cinco espécies. Já no Apêndice B, apresenta a curva resposta de todos os ciclos, coletados neste estudo, por categoria (sete espécies). Em suma, a *C. Parapsilosis*, apresenta o mesmo comportamento, de constâcia da resistência, da *C. Krusei*.

Figura 31 – Registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso para único ciclo por espécie: (a) *C. Albicans* e (b) *C. Krusei*



(a)



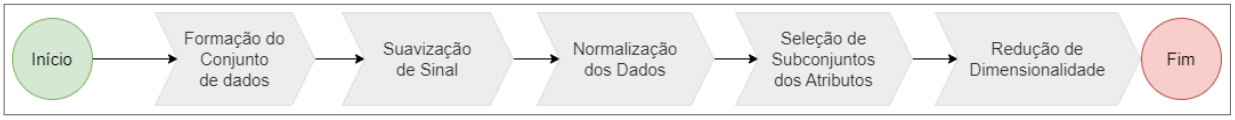
(b)

Fonte: A autora (2023).

4.4 PROCEDIMENTOS DE PRÉ-PROCESSAMENTO

As saídas digitais geradas pelos sensores de narizes eletrônicos devem ser analisadas e interpretadas a fim de fornecer o padrão olfativo da amostra. Esse procedimento envolve a extração de certas características significativas das curvas de resposta do sensor, para produzir um conjunto de dados numéricos que podem ser processados pelo sistema de reconhecimento e classificação do nariz eletrônico.

Figura 32 – Fluxograma ilustrativo das etapas de pré-processamento dos dados. O pré-processamento inicia-se com a formação do conjunto de dados, passa pelo suavização do sinal, normalização dos dados, seleção de subconjuntos dos atributos e, por último, pelo processo de redução de dimensionalidade.



Fonte: A autora (2023).

4.4.1 Formação do conjunto de dados

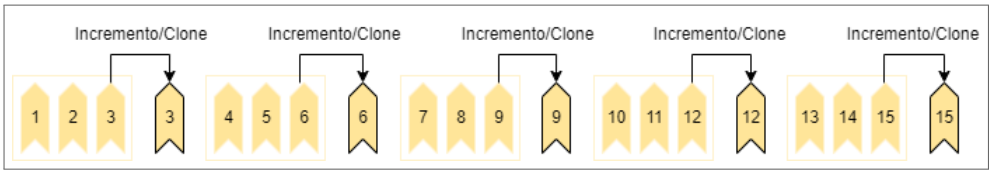
Como citado na seção 4.3.3, as leituras dos diferentes fungos seguem um protocolo ou roteiro de coleta, gerando, assim, arquivos que foram utilizados para criar as bases de dados. No total, foram gerados 28 arquivos, com diferente números de ciclos, de no mínimo 5 ciclos. Para este estudo, foram construídos quatro conjunto de dados base: uma é composta pelas leituras de todas as três fases do ciclo (Aspiração, Estabilização e Restauração); as outras três bases são compostas pelas leituras isoladas de cada uma dessas fases do ciclo. Além disso, adotou-se que uma instância é um ciclo.

Além disso, afim de manter a consistência no número de pontos coletados em cada etapa de cada ciclo, adotou-se uma estratégia para garantir que o conjunto de dados possua a mesma quantidade de dados em cada caso, independentemente das flutuações na coleta de dados. Isso pode ser útil em situações onde a consistência no tamanho do conjunto de dados é importante para análises subsequentes ou para manter o desempenho de algoritmos que requerem entradas de tamanho fixo. Esta estratégia seguiu-se as seguintes critérios:

- Se o número de coletas da etapa é menor do que o número desejado;
  - Calcule a diferença de pontos entre o número desejado e o número real de coletas;
  - Crie uma janela, que é uma região de tempo ou espaço para coletas adicionais;
  - Repita a última amostra dentro dessa janela para preencher a diferença e alcançar o número desejado de coletas.

Para exemplificação, a figura a seguir ilustra esta estratégia, considerando que a janela encontrada foi 3 coletas. Assim, a cada 3 coletas, o último valor dessa janela é incrementado, mantendo a continuidade da série temporal coletada. Lembra-se que esse processo é repetido para todas as etapas do ciclo.

Figura 33 – Formação do conjunto de dados: Incrementação de dados para amostras com menor número de coletas desejadas, calculá-se a diferença de pontos, cria-se uma janela e repete-se a última amostra.

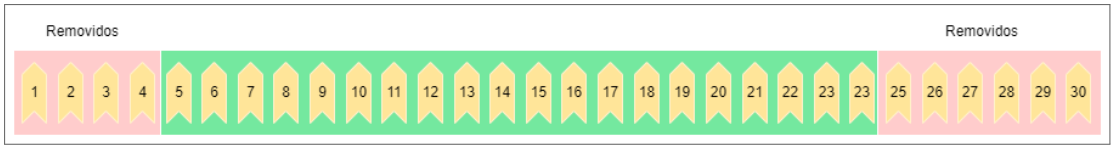


Fonte: A autora (2023).

- Se o número de coletas da etapa é maior do que o número desejado.
  - Calcule a diferença de pontos entre o número real de coletas e o número desejado;
  - Remova dados da janela inicial e final para reduzir o número de coletas à quantidade desejada.

  A figura a seguir ilustra uma exemplificação onde a diferença considerada foi de 10 coletas. Dessa forma, remove-se as coletas iniciais e finais e mantém-se as coletas centrais. Lembra-se que esse processo é repetido para todas as etapas do ciclo.

Figura 34 – Formação do conjunto de dados: Remoção de dados para amostras com maior número de coletas desejadas, calculá-se a diferença de pontos, remove-se dados da janela inicial e final.



Fonte: A autora (2023).

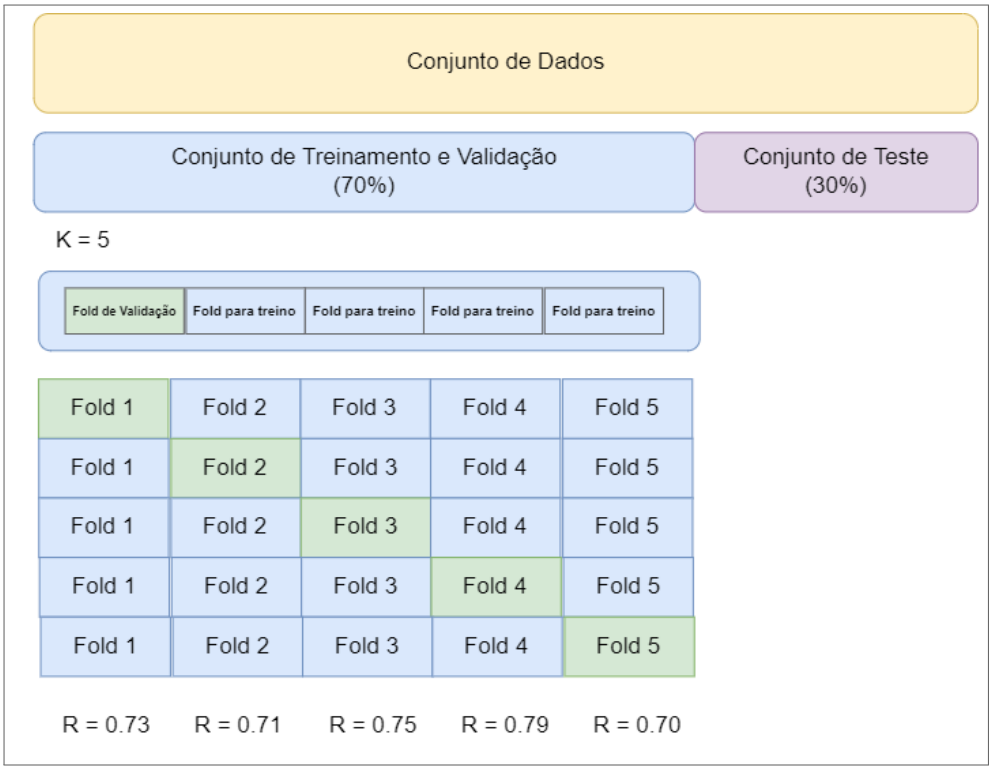
4.4.1.1 Aleatorização de Processamento dos Modelos

A Cross Validation (CV) é uma técnica muito utilizada para avaliação de desempenho de modelos de aprendizado de máquina. Esta técnica consiste em particionar os dados em conjuntos, onde um conjunto é utilizado para treino e validação do modelo e outro conjunto é utilizado para teste e avaliação do desempenho do modelo. A utilização do CV tem altas chances de detectar se o seu modelo está sobreajustado aos seus dados de treinamento, pois, ao particionar os dados disponíveis em três conjuntos, reduz-se drasticamente o número de amostras que podem ser usadas para aprender o modelo, e os

resultados podem depender de uma escolha aleatória específica para o par de conjuntos (treinamento, validação).

Portanto, antes de treinar cada um dos modelos adotados neste estudo, procede-se com a separação aleatória dos dados em conjuntos de treinamento, validação e teste. É uma prática comum dividir a base de dados em uma proporção de 70% para treinamento e 30% para teste. O treinamento inicial ocorre no conjunto de treinamento, seguido pela avaliação no conjunto de validação. Quando o experimento demonstra ser bem-sucedido nessa fase, a avaliação final é conduzida no conjunto de teste, permitindo assim a avaliação do desempenho do modelo com dados que ele nunca viu anteriormente.

Figura 35 – Representação gráfica da técnica de validação cruzada kFold, que consiste em dividir uma base qualquer em k partes (*folders*) e rodar o modelo k vezes. É muito comum dividir a base em treinamento/teste a 70%/30%. Em cada rodada k−1 *folders* são o conjunto de treinamento, esse processo garante que cada subconjunto será utilizado para teste em algum momento da avaliação do modelo.



Fonte: A autora (2023).

Um método de aplicação do *Cross Validation* é o kFold que consiste em dividir a base de dados de treinamento de forma aleatória em K subconjuntos (em que K é definido previamente) com aproximadamente a mesma quantidade de amostras em cada um deles. A cada iteração, treino e validação, um conjunto formado por K−1 subconjuntos são utilizados para treinamento e o subconjunto restante será utilizado para validação gerando um resultado de métrica para avaliação como a acurácia. Ou seja, encontra-se resultados diferentes de acordo com cada iteração K do nosso modelo. Esse processo garante que cada subconjunto será algum momento da avaliação do modelo. A Figura 35 faz uma

representação gráfica da técnica de validação cruzada kFold e uma abstração de resultados (R) para cada subdivisão.

Dessa maneira, o resultado final é a média de todos os resultados:

$$\bar{R} = \frac{\sum_{i=1}^{k} Ri}{K} = \frac{0.73 + 0.71 + 0.75 + 0.79 + 0.70}{5} = 0.736 \tag{4.1}$$

Os experimentos foram processados através de códigos escritos em *Python* e o formato das bases utilizado foi por ciclo. Além disso para garantir a representatividade dos conjuntos foi adotada a validação cruzada kFold por 10 repetições aleatórias (a partir da técnica da biblioteca *Python StratifiedShuffleSplit*) e dividindo a base em treinamento/teste a 70%/30%,(a partir da técnica da biblioteca *Python scikit-learn, train_test_split*).

4.4.2 Suavização do Sinal

Uma das etapas de pré-processamento dos sinais dos sensores foi a aplicação do filtro da média móvel modificado, a fim de reduzir os possíveis ruídos gerados pelos sensores. Essa alteração proporciona maior suavização no conjunto de dados, removendo possíveis picos de sinais devido a ruídos provenientes do circuito, pois desconsidera os valores discrepantes do conjunto de dados. Esse filtro foi usado, pois é o mais simples, rápido, robusto e de fácil implementação (ALESSIO et al., 2002). Uma média móvel simples sobre n elementos é constituída das médias não ponderadas dos subconjuntos de n elementos em um conjunto de dados. Cada ciclo possui em média 140 registros. O valor adotado de n é 10. A diferença entre o filtro modificado e o original é que o primeiro só leva em consideração os valores que estejam em torno da média com diferença máxima de três vezes o desvio padrão, ou seja, $|pi - \mu| \leq 3\sigma$, em que pi é o elemento i do conjunto de dados p, μ é a média do conjunto e σ o desvio padrão.

4.4.3 Representação do domínio frequência

Sabe-se que os sensores de gás possuem um comportamento que decompõe uma função temporal em frequências. Dessa forma, o sinal gerado pode ser representado outro domínio, como o de frequência. Comumente, operação matemática que associa a representação domínio frequência a uma função temporal é a transformada de Fourier. Esta transformada não é limitada a funções temporais, contudo para fins de convenção, o domínio original é comumente referido como domínio do tempo.

Além disso, outra alternativa da transformada de Fourier é a transformada de *wavelet* que permite analisar periodicidade de eventos em diferentes escalas da variabilidade temporal e não necessita de uma série estacionária. Assim, a ferramenta é apropriada para analisar eventos irregularmente distribuídos.

4.4.4 Normalização de Dados

Nesta etapa, os dados de atributo são dimensionados de modo a se enquadrarem em um intervalo pequeno, de 0 a 1, pelo método mais comum de normalização de dados é a máximo-mínimo. Esta normalização é particularmente útil para algoritmos de classificação envolvendo redes neurais ou medições de distância, como classificação e agrupamento de vizinhos mais próximos. Já se estiver usando o algoritmo de *backpropagation*, a normalização ajuda a evitar que atributos com intervalos inicialmente grandes superem atributos com intervalos inicialmente menores.

4.4.5 Seleção de Atributos

4.4.5.1 Sensores

A matriz de correlação linear das leituras dos sensores do *e-nose* indica se existe correlação linear entre os dados dos sensores. Para isso, foi adotado uma escala de -1 a 1 sobre uma única cor, a verde. Assim, quanto mais escuro é o verde, mais próximo a 1 e consequentemente mais forte é a correlação linear direta, já quanto mais claro é o verde, mais próximo a -1, e assim mais forte é a correlação linear inversa.

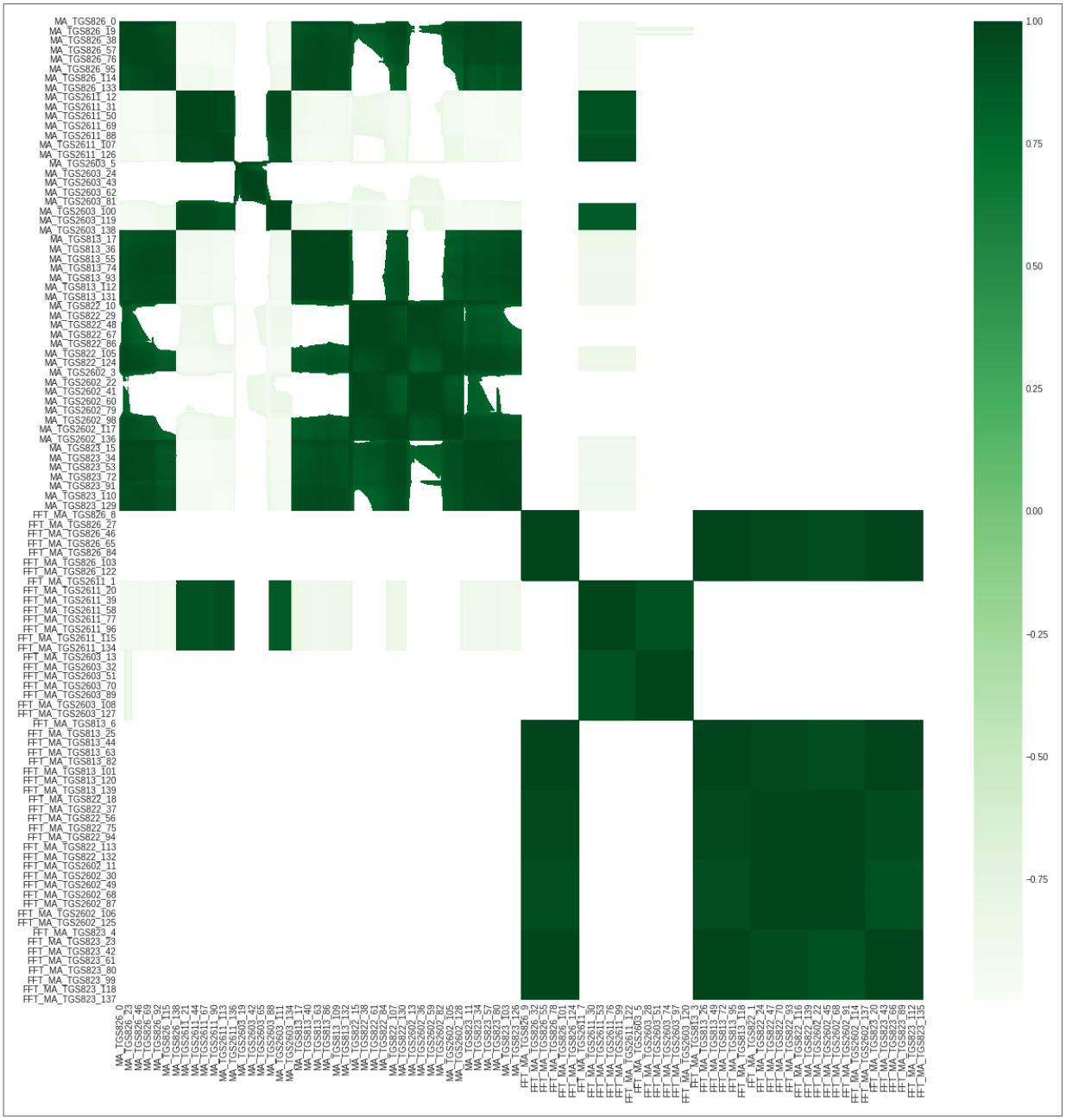
O ideal é não haver correlação linear entre os sensores, indicando independência entre sensores. Portanto, é desejável que a correlação esteja no escalar 0, isto é, ausência de correlação linear, melhor para o processo de classificação. Por outro lado, as características com alta correlação são mais linearmente dependentes e, portanto, têm quase o mesmo efeito na variável dependente. Dessa forma, quando dois recursos têm alta correlação, podemos descartar um dos dois recursos. As matrizes de correlações dos sensores do *e-nose* estão mostradas na Figura 36.

Os critérios específicos para definir "baixa variância"ou "alta correlação"em um conjunto de dados de séries temporais podem variar com base na natureza dos dados. Neste estudo, adotou-se os seguintes critérios::

1. Baixa Variância em Séries Temporais: A baixa variância em séries temporais geralmente significa que uma característica tem pouca variação nos valores ao longo do tempo. A variância de uma série temporal pode ser calculada em várias janelas de tempo e comparada. Critérios para baixa variância podem ser definidos com base em uma variação mínima que é considerada informativa para a previsão ou identificação de padrões. Para isso, foi utilizado um método conhecido como *Variance Threshold*, onde é encontrado um valor de limiar para a variância. Se a variância de uma característica for inferior a este limiar, essa característica é classificada como tendo baixa variância.

2. Alta Correlação em Séries Temporais: A alta correlação entre características em séries temporais implica que elas estão linearmente relacionadas ao longo do tempo.

Figura 36 – Matriz de correlação linear dos atributos derivado de dados coletados por sensores do *e-nose*. Quanto mais escuro é o verde, mais próximo a 1 e consequentemente mais forte é a correlação linear direta, já quanto mais claro é o verde, mais próximo a -1, e assim mais forte é a correlação linear inversa. O ideal é não haver correlação linear entre os sensores, indicando independência entre sensores. Portanto, é desejável que a correlação esteja no escalar 0, isto é, ausência de correlação linear, melhor para o processo de classificação.



Fonte: A autora (2023).

Critérios para alta correlação podem ser definidos com base em um valor mínimo para o coeficiente de correlação, que indica a força e a direção dessa relação linear. Se o coeficiente de correlação absoluto entre duas características for maior que 0.95, essas características podem ser consideradas altamente correlacionadas.

4.4.5.2 Cálculo da área da curva

Na Geometria, o cálculo de áreas sob curvas é uma maneira de calcular o comprimento de arcos e volumes; enquanto que na Física, para é calculado o trabalho realizado por uma força, momento, centros de massa e momento de inércia, além de várias outras aplicações. Dessa maneira, a magnitude da área da curva gerado pelos sensores, pode ser considerado como outro parâmetro de estudo.

4.4.6 Redução de dimensionalidade

A redução da dimensão é definida como uma forma de converter um conjunto de dados de alta dimensão num conjunto de dados de dimensão inferior, assegurando que a informação que fornece é semelhante em ambos os casos. Esta técnica é frequentemente utilizada na aprendizagem de máquinas para obter um modelo de previsão mais apertado, resolvendo ao mesmo tempo os problemas de regressão e classificação apresentados pelos algoritmos. Assim, a redução da dimensionalidade é responsável pela identificação e remoção de características que diminuem o desempenho do modelo de aprendizagem da máquina.

Além disso, há várias técnicas de dimensionalidade, dentre elas a Uniform Manifold Approximation and Projection (UMAP), uma técnica de redução de dimensão que pode ser usada para visualização semelhante ao *t-SNE*, mas também para redução de dimensão não linear. Os gráficos apresentados na Figura 37 ilustram as formas de redução da quantidade de atributos de uma base para 2 e 3 atributos, respectivamente, para a primeira base de dados formadas, contendo todos as etapas e todos os atributos.

4.4.7 Escolha dos classificadores

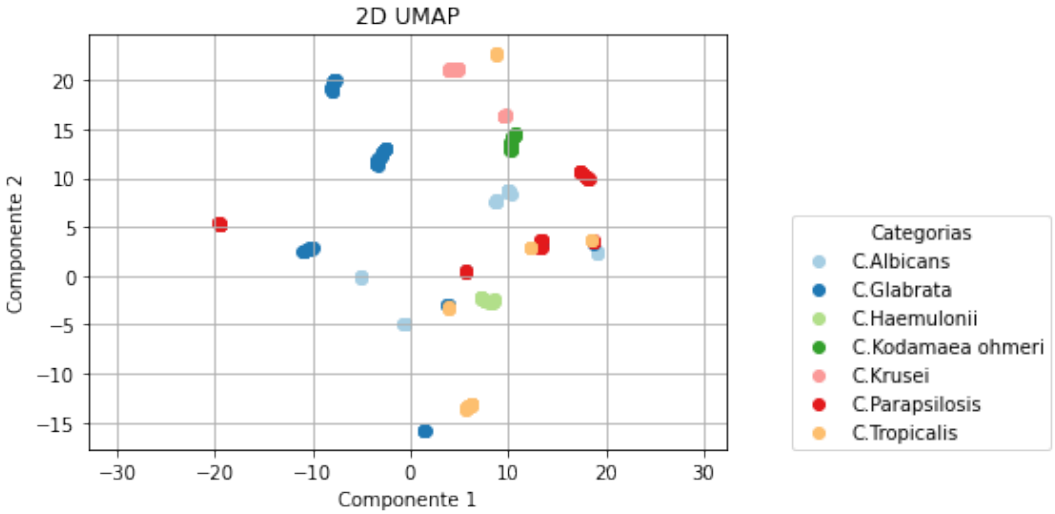
A escolha de utilizar classificadores como o TimeSeries Forest, KNeighbors TimeSeries, Rocket, HIVE-COTE e TapNet foi baseada na necessidade de explorar diferentes abordagens para a tarefa de identificação de espécies de *Candida* com base em dados de compostos orgânicos voláteis. Cada um desses classificadores possui características únicas que podem se adequar a diferentes aspectos do problema, como a complexidade dos dados e a necessidade de processamento eficiente. O TimeSeries Forest foi escolhido devido ao seu desempenho notável em problemas de séries temporais, enquanto o KNeighbors TimeSeries oferece uma abordagem baseada em vizinhos próximos. O Rocket foi incluído por sua capacidade de transformação de dados em recursos adequados para classificação. O HIVE-COTE é uma escolha robusta que combina vários classificadores, explorando a

diversidade de técnicas. Por fim, o TapNet é uma adição valiosa devido à sua capacidade de aprendizado profundo, que pode revelar padrões complexos nos dados. A combinação desses classificadores proporciona uma abordagem abrangente e diversificada.

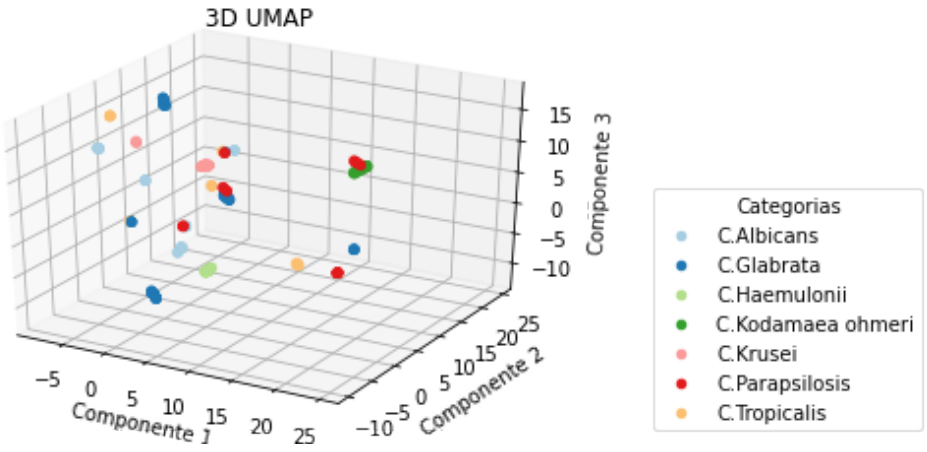
4.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram apresentados os materiais e métodos que foram utilizados para solucionar problema de reconhecimento fúngico. Foram apresentados o fluxo da metodologia adotada, explanando os sensores e a funcionalidade disposto no protótipo do *e-nose* construído, demonstrando o protocolo ou roteiro de experimentação utilizado e ilustrando o passo-a-passo da coleta de dados, apresentando desde como e onde os materiais a serem analisados foram semeados até a curva resposta do registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso. Além disso, foram explanados os procedimentos de pré-processamento aplicados e o motivo para a qual foram adotadas. Dessa maneira, este capítulo apresenta descrição dos procedimentos experimentais utilizados para a obtenção e tratamentos dos dados. E conclui-se que para os experimentos sejam reproduzidos é essencial que os protocolos e roteiros sejam seguidos. O próximos passo será a apresentação dos experimentos realizados neste estudo.

Figura 37 – Ilustração dos dados após a aplicação da UMAP, uma nova técnica de aprendizado múltiplo para redução de dimensão. A redução da dimensão é uma forma de converter um conjunto de dados de alta dimensão num conjunto de dados de dimensão inferior, assegurando que a informação que fornece é semelhante em ambos os casos (a) 2D e (b) 3D da base de dados construída a partir das leituras dos COVs exalados por sete gêneros de *Candida* spp.



(a)



(b)

Fonte: A autora (2023).

5 ANÁLISE EXPERIMENTAL

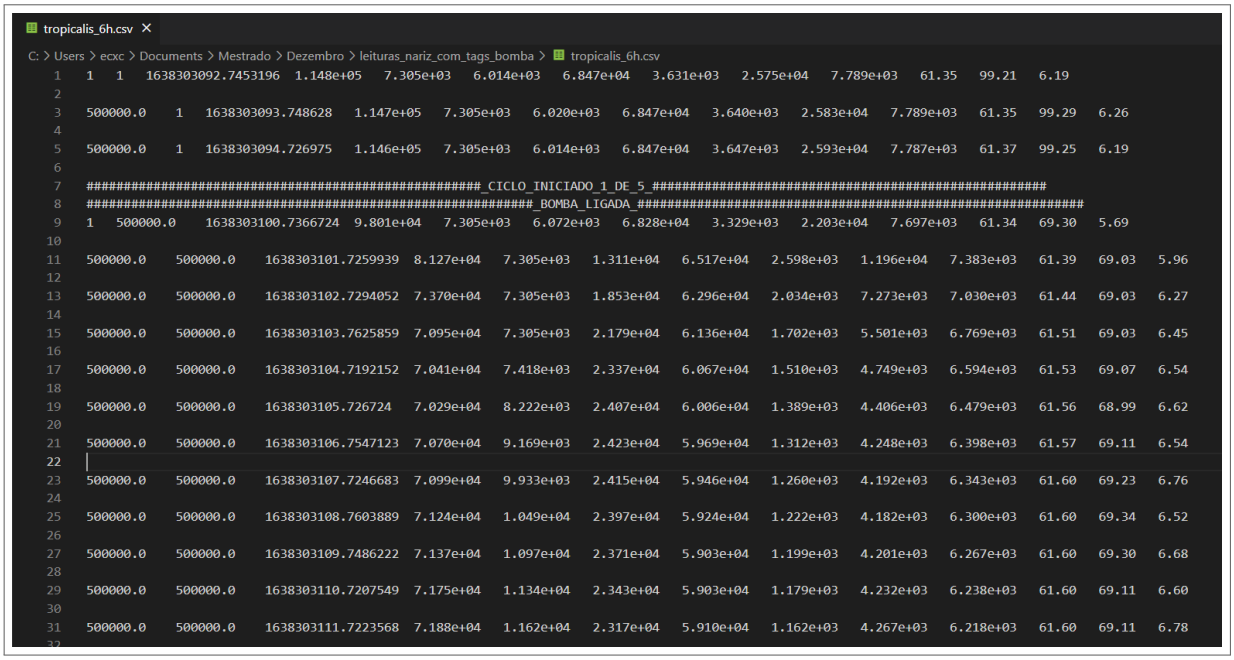
Neste capítulo serão apresentados os experimentos realizados e os resultados obtidos com os métodos de classificação, tanto para validação, quanto para teste, dos para as bases de dados criadas, além das métricas e matrizes de confusão. Por último, serão apresentadas as comparações dos resultados deste trabalho com outros estudos presentes na literatura.

5.1 EXPERIMENTOS

5.1.1 Dados

Como citado na subseção 4.3.1, os repiques de diferentes espécies de *Candida*, coletados pelo Laboratório de Micologia Médica Sylvio Campos da Universidade Federal de Pernambuco (UFPE), foram inseridos na câmara simples (Figura 25). Os sensores acoplados a esta câmera são responsáveis pela leitura dos COVs exalados pela amostra. A Figura 38 mostra um trecho de uma saída de leitura do nariz eletrônico.

Figura 38 – Trecho de saída de uma leitura do nariz eletrônico. Pode-se observar que o arquivo construído possui algumas marcações como 'CICLO_INICIADO', 'BOMBA_LIGADA', 'CICLO_FINALIZADO' e 'BOMBA_DESLIGADA'. Estas marcações são fundamentadas no protocolo de coleta, separando, assim, os ciclos e as etapas do ciclo. Além disso, observa-se também as leituras que não fazem parte do ciclo, chamadas de leituras de repouso, como nas linhas 1 a 3.


tropicalis_6h.csv

C: > Users > ecxc > Documents > Mestrado > Dezembro > leituras_nariz_com_tags_bomba > tropicalis_6h.csv

```
1   1   1638303092.7453196  1.148e+05   7.305e+03   6.014e+03   6.847e+04   3.631e+03   2.575e+04   7.789e+03   61.35   99.21   6.19
2
3   500000.0    1   1638303093.748628   1.147e+05   7.305e+03   6.020e+03   6.847e+04   3.640e+03   2.583e+04   7.789e+03   61.35   99.29   6.26
4
5   500000.0    1   1638303094.726975   1.146e+05   7.305e+03   6.014e+03   6.847e+04   3.647e+03   2.593e+04   7.787e+03   61.37   99.25   6.19
6
7   ##########################################################_CICLO_INICIADO_1_DE_5_#########################################################
8   ##########################################################_BOMBA_LIGADA_#################################################################
9   1   500000.0    1638303100.7366724  9.801e+04   7.305e+03   6.072e+03   6.828e+04   3.329e+03   2.203e+04   7.697e+03   61.34   69.30   5.69
10
11  500000.0    500000.0    1638303101.7259939  8.127e+04   7.305e+03   1.311e+04   6.517e+04   2.598e+03   1.196e+04   7.383e+03   61.39   69.03   5.96
12
13  500000.0    500000.0    1638303102.7294052  7.370e+04   7.305e+03   1.853e+04   6.296e+04   2.034e+03   7.273e+03   7.030e+03   61.44   69.03   6.27
14
15  500000.0    500000.0    1638303103.7625859  7.095e+04   7.305e+03   2.179e+04   6.136e+04   1.702e+03   5.501e+03   6.769e+03   61.51   69.03   6.45
16
17  500000.0    500000.0    1638303104.7192152  7.041e+04   7.418e+03   2.337e+04   6.067e+04   1.510e+03   4.749e+03   6.594e+03   61.53   69.07   6.54
18
19  500000.0    500000.0    1638303105.726724   7.029e+04   8.222e+03   2.407e+04   6.006e+04   1.389e+03   4.406e+03   6.479e+03   61.56   68.99   6.62
20
21  500000.0    500000.0    1638303106.7547123  7.070e+04   9.169e+03   2.423e+04   5.969e+04   1.312e+03   4.248e+03   6.398e+03   61.57   69.11   6.54
22
23  500000.0    500000.0    1638303107.7246683  7.099e+04   9.933e+03   2.415e+04   5.946e+04   1.260e+03   4.192e+03   6.343e+03   61.60   69.23   6.76
24
25  500000.0    500000.0    1638303108.7603889  7.124e+04   1.049e+04   2.397e+04   5.924e+04   1.222e+03   4.182e+03   6.300e+03   61.60   69.34   6.52
26
27  500000.0    500000.0    1638303109.7486222  7.137e+04   1.097e+04   2.371e+04   5.903e+04   1.199e+03   4.201e+03   6.267e+03   61.60   69.30   6.68
28
29  500000.0    500000.0    1638303110.7207549  7.175e+04   1.134e+04   2.343e+04   5.903e+04   1.179e+03   4.232e+03   6.238e+03   61.60   69.11   6.60
30
31  500000.0    500000.0    1638303111.7223568  7.188e+04   1.162e+04   2.317e+04   5.910e+04   1.162e+03   4.267e+03   6.218e+03   61.60   69.11   6.78
32
```

Fonte: A autora (2023).

Pode-se observar que o arquivo construído possui marcações como "CICLO_INICIADO"e "BOMBA_LIGADA"; mas também existe "CICLO_FINALIZADO"e "BOMBA_DESLIGADA".

Estas marcações são fundamentadas no protocolo de coleta, separando, assim, os ciclos e as etapas do ciclo. Além disso, observa-se também as leituras que não fazem parte do ciclo, chamadas de leituras de repouso, como nas linhas 1 a 3.

5.1.2 Remoção das marcações

Sabe-se que para este trabalho a base da instância é o ciclo, sendo assim, fez-se necessário separar os 28 arquivos por ciclo, removendo as marcações, as leituras de repouso ou qualquer outro dado que não pertencem ao ciclo. No final, foram encontrados 330 instâncias, ou melhor, 330 ciclos de diferentes espécies. A Figura 39 ilustra um trecho dos arquivos formados, enquanto que na Figura 40 ilustra o conteúdo de um desses arquivos.

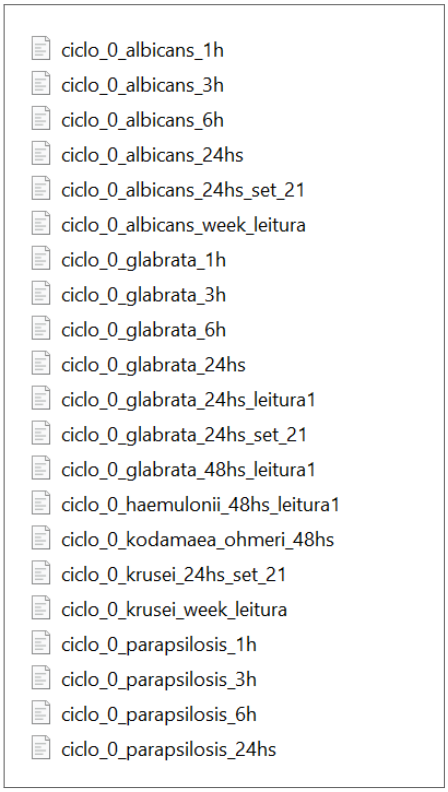
A Tabela 9 apresenta distribuição de instâncias para cada gênero de fungo (categoria). As bases não são balanceadas, sendo a categoria que mais apresenta instâncias a *C. Glabrata* com 100 casos e as que apresentam menos instâncias as *C. Kodamaea ohmeri, C. Krusei e C. Haemulonii* apenas 30 casos cada.

Tabela 9 – Distribuição de instâncias para cada gênero de fungo (categoria)

Gênero - Categoria	**Quantidade de instâncias**
C. Glabrata	100 (30.30%)
C. Albicans	50 (15.15%)
C. Parapsilosis	50 (15.15%)
C. Tropicalis	40 (12.121%)
C. Kodamaea ohmeri	30 (9.09%)
C. Krusei	30 (9.09%)
C. Haemulonii	30 (9.09%)
Total	**330 (100%)**

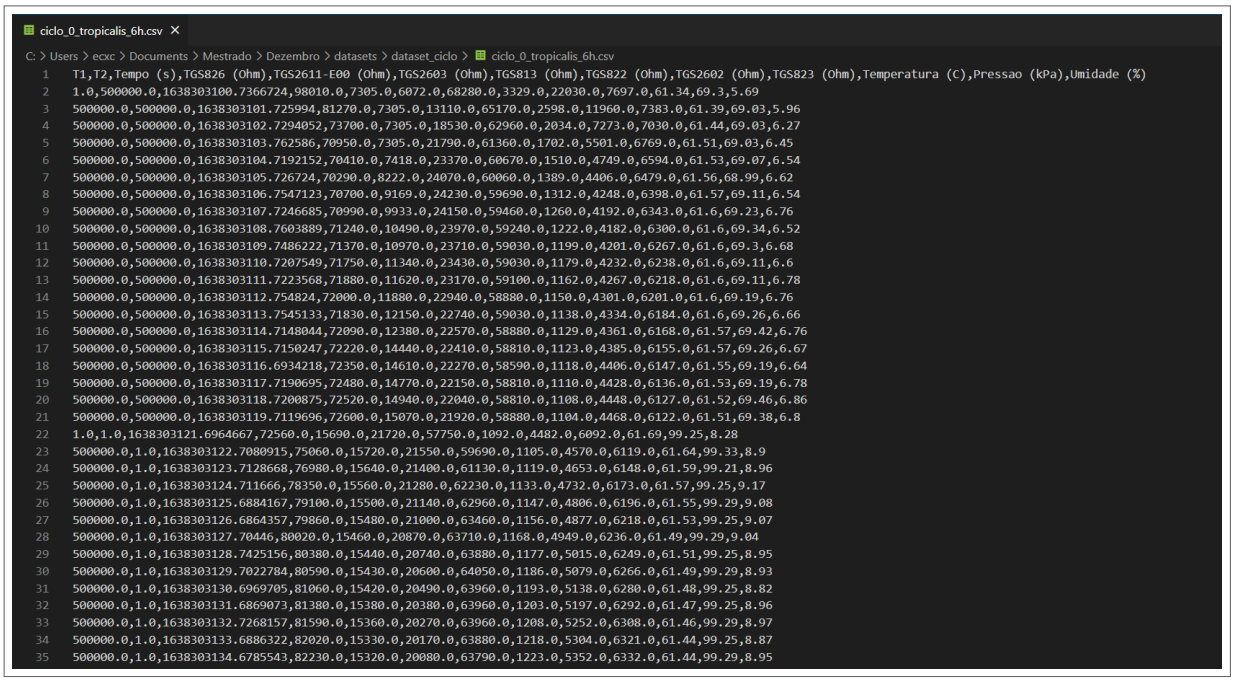
Fonte: A autora (2023).

Figura 39 – Trecho da saída da divisão dos arquivos de leitura em arquivos por ciclo temporal. No final, os 28 arquivos foram separados por ciclo, gerando, assim, novos arquivos.



Fonte: A autora (2023).

Figura 40 – Trecho de saída de uma leitura do nariz eletrônico por ciclo. No final, nos novos arquivos houve a remoção das marcações, das leituras de repouso ou qualquer outro dado que não pertenciam ao ciclo.



ciclo_0_tropicalis_6h.csv

C: > Users > ecxc > Documents > Mestrado > Dezembro > datasets > dataset_ciclo > ciclo_0_tropicalis_6h.csv

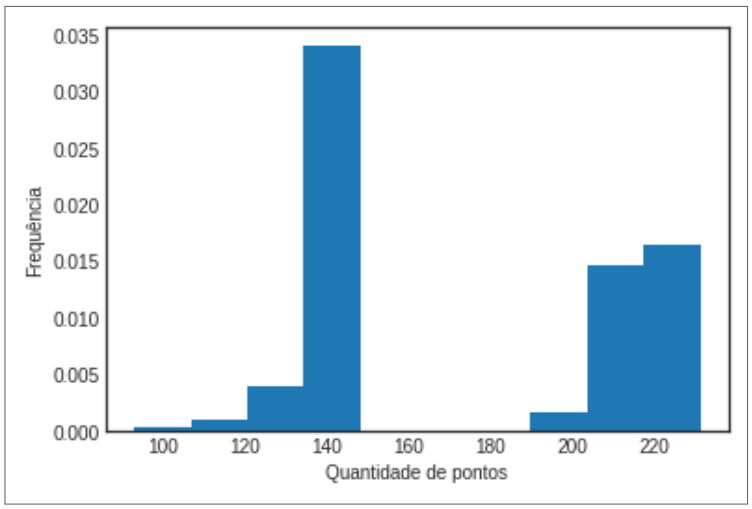
```
T1,T2,Tempo (s),TGS826 (Ohm),TGS2611-E00 (Ohm),TGS2603 (Ohm),TGS813 (Ohm),TGS822 (Ohm),TGS2602 (Ohm),TGS823 (Ohm),Temperatura (C),Pressao (kPa),Umidade (%)
1.0,500000.0,1638303100.7366724,98010.0,7305.0,6072.0,68280.0,3329.0,22030.0,7697.0,61.34,69.3,5.69
500000.0,500000.0,1638303101.725994,81270.0,7305.0,13110.0,65170.0,2598.0,11960.0,7383.0,61.39,69.03,5.96
500000.0,500000.0,1638303102.7294052,73700.0,7305.0,18530.0,62960.0,2034.0,7273.0,7030.0,61.44,69.03,6.27
500000.0,500000.0,1638303103.762586,70950.0,7305.0,21790.0,61360.0,1702.0,5501.0,6769.0,61.51,69.03,6.45
500000.0,500000.0,1638303104.7192152,70410.0,7418.0,23370.0,60670.0,1510.0,4749.0,6594.0,61.53,69.07,6.54
500000.0,500000.0,1638303105.726724,70290.0,8222.0,24070.0,60060.0,1389.0,4406.0,6479.0,61.56,68.99,6.62
500000.0,500000.0,1638303106.7547123,70700.0,9169.0,24230.0,59690.0,1312.0,4248.0,6398.0,61.57,69.11,6.54
500000.0,500000.0,1638303107.7246685,70990.0,9933.0,24150.0,59460.0,1260.0,4192.0,6343.0,61.6,69.23,6.76
500000.0,500000.0,1638303108.7603889,71240.0,10490.0,23970.0,59240.0,1222.0,4182.0,6300.0,61.6,69.34,6.52
500000.0,500000.0,1638303109.7486222,71370.0,10970.0,23710.0,59030.0,1199.0,4201.0,6267.0,61.6,69.3,6.68
500000.0,500000.0,1638303110.7207549,71750.0,11340.0,23430.0,59030.0,1179.0,4232.0,6238.0,61.6,69.11,6.6
500000.0,500000.0,1638303111.7223568,71880.0,11620.0,23170.0,59100.0,1162.0,4267.0,6218.0,61.6,69.11,6.78
500000.0,500000.0,1638303112.754824,72000.0,11880.0,22940.0,58880.0,1150.0,4301.0,6201.0,61.6,69.19,6.76
500000.0,500000.0,1638303113.7545133,71830.0,12150.0,22740.0,59030.0,1138.0,4334.0,6184.0,61.6,69.26,6.66
500000.0,500000.0,1638303114.7148044,72090.0,12380.0,22570.0,58880.0,1129.0,4361.0,6168.0,61.57,69.42,6.76
500000.0,500000.0,1638303115.7150247,72220.0,14440.0,22410.0,58810.0,1123.0,4385.0,6155.0,61.57,69.26,6.67
500000.0,500000.0,1638303116.6934218,72350.0,14610.0,22270.0,58590.0,1118.0,4406.0,6147.0,61.55,69.19,6.64
500000.0,500000.0,1638303117.7190695,72480.0,14770.0,22150.0,58810.0,1110.0,4428.0,6136.0,61.53,69.19,6.78
500000.0,500000.0,1638303118.7200875,72520.0,14940.0,22040.0,58810.0,1108.0,4448.0,6127.0,61.52,69.46,6.86
500000.0,500000.0,1638303119.7119696,72600.0,15070.0,21920.0,58880.0,1104.0,4468.0,6122.0,61.51,69.38,6.8
1.0,1.0,1638303121.6964667,72560.0,15690.0,21720.0,57750.0,1092.0,4482.0,6092.0,61.69,99.25,8.28
500000.0,1.0,1638303122.7080915,75060.0,15720.0,21550.0,59690.0,1105.0,4570.0,6119.0,61.64,99.33,8.9
500000.0,1.0,1638303123.7128668,76980.0,15640.0,21400.0,61130.0,1119.0,4653.0,6148.0,61.59,99.21,8.96
500000.0,1.0,1638303124.711666,78350.0,15560.0,21280.0,62230.0,1133.0,4732.0,6173.0,61.57,99.25,9.17
500000.0,1.0,1638303125.6884167,79100.0,15500.0,21140.0,62960.0,1147.0,4806.0,6196.0,61.55,99.29,9.08
500000.0,1.0,1638303126.6864357,79860.0,15480.0,21000.0,63460.0,1156.0,4877.0,6218.0,61.53,99.25,9.07
500000.0,1.0,1638303127.70446,80020.0,15460.0,20870.0,63710.0,1168.0,4949.0,6236.0,61.49,99.29,9.04
500000.0,1.0,1638303128.7425156,80380.0,15440.0,20740.0,63880.0,1177.0,5015.0,6249.0,61.51,99.25,8.95
500000.0,1.0,1638303129.7022784,80590.0,15430.0,20600.0,64050.0,1186.0,5079.0,6266.0,61.49,99.29,8.93
500000.0,1.0,1638303130.6969705,81060.0,15420.0,20490.0,63960.0,1193.0,5138.0,6280.0,61.48,99.25,8.82
500000.0,1.0,1638303131.6869073,81380.0,15380.0,20380.0,63960.0,1203.0,5197.0,6292.0,61.47,99.25,8.96
500000.0,1.0,1638303132.7268157,81590.0,15360.0,20270.0,63960.0,1208.0,5252.0,6308.0,61.46,99.29,8.97
500000.0,1.0,1638303133.6886322,82020.0,15330.0,20170.0,63880.0,1218.0,5304.0,6321.0,61.44,99.25,8.87
500000.0,1.0,1638303134.6785543,82230.0,15320.0,20080.0,63790.0,1223.0,5352.0,6332.0,61.44,99.29,8.95
```

Fonte: A autora (2023).

5.1.3 Remoção de outliers

Os *outliers* são dados que fogem da normalidade e que pode (e provavelmente irá) causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise. Sabe-se que idealmente a quantidade de coletas por ciclo deveria ser de 140 pontos (20 para aspiração - 14.28%, 60 para estabilização - 42.86% e 60 para purga/purificação - 42.86%). No entanto, segundo o histograma ilustrado na Figura 41, observa-se existe uma variabilidade do número de coletas realizadas por ciclo.

Figura 41 – Histograma da quantidade de pontos coletados por ciclo



Fonte: A autora (2023).

Levando em consideração a distribuição percentual dos pontos coletas por etapa (14.28%, 42.86% e 42.86%), todos os ciclos foi analisado individualmente, o que proporcionou a observação do comportamento de distribuição dos pontos, na tabela a seguir:

Tabela 10 – Distribuição percentual dos pontos coletados por ciclo

	Aspiração	**Estabilização**	**Purificação/Purga**
Média	13.67%	44.72%	41.61%
Moda	14.29%	42.86%	42.86%
Mediana	13.90%	44.05%	42.22%
Desvio padrão	4.91%	3.73%	3.74%
Valor mínimo	3.23%	41.94%	31.07%
Valor máximo	16.13%	55.34%	54.84%

Fonte: A autora (2023).

Dessa maneira, baseando-se no valor de percentual médio e no desvio padrão encontrado por etapa, cada ciclo foi avaliado afim de procurar por observações de ciclos fora do comum. Portanto, se o distribuição de pontos coletados por etapa no ciclo estão dentro

do intervalo adotado, o ciclo é considerado válido, se não, invalida-se o ciclo. No final, 22 ciclos foram considerados como *outliers*, resultando 308 ciclos válidos.

5.1.4 Formação do conjunto de dados

Apesar do processo anterior remover os ciclos que se diferenciavam drasticamente de todos os outros, a variabilidade da quantidade de pontos coletados se manteve, necessitando da adoção de uma estratégia, já explanada na seção 4.4.1: manter o número fixo de pontos coletados,

Para ilustrar o primeiro cenário, quando o número de coletas não alcança o desejado, considera-se o quarto ciclo do arquivo *glabrata_3h*. Neste ciclo, o número total de coletas foi de 116, com 15 coletas na etapa de aspiração, 59 na etapa de estabilização e 42 na etapa de purificação/purga. Isso resultou em diferenças de -5, -1 e -18 em relação à quantidade desejada de coletas para cada etapa. Para resolver essa discrepância, é necessário calcular uma janela de coleta adicional para cada etapa, a fim de atingir o número desejado de dados.

No segundo cenário, quando o número de coletas excede o desejado, considera-se como exemplo o primeiro ciclo do arquivo *haemulonii_48hs_leitura*. Neste ciclo, o número total de coletas foi de 218, com 30 coletas na etapa de aspiração, 99 na etapa de estabilização e 89 na etapa de purificação/purga. Isso resultou em diferenças de +10, +39 e +29 em relação à quantidade desejada de coletas para cada etapa. Para ajustar isso, é necessário reduzir o número de coletas, eliminando dados das etapas iniciais e centrais, de modo a atingir a quantidade desejada.

Por fim, esta estratégia permitiu a construção de um único conjunto de dados. A Figura 42 apresenta um trecho de saída desse conjunto, que no final obteve um de tamanho (308, 2801). Além disso, a Tabela 11 apresenta distribuição de instâncias para cada gênero do fungo para o conjunto de dados final.

Tabela 11 – Distribuição de instâncias para cada gênero de fungo (categoria) para o conjunto de dados final construído

Gênero - Categoria	**Quantidade de instâncias**
C. Glabrata	88 (28.571%)
C. Parapsilosis	50 (16.233%)
C. Albicans	42 (13.636%)
C. Tropicalis	39 (12.662%)
C. Kodamaea ohmeri	30 (9.740%)
C. Krusei	30 (9.740%)
C. Haemulonii	29 (9.415%)
Total	**308 (100%)**

Fonte: A autora (2023).

Figura 42 – Trecho de saída do conjunto de dados final, resultando um conjunto de tamanho (308, 2801).

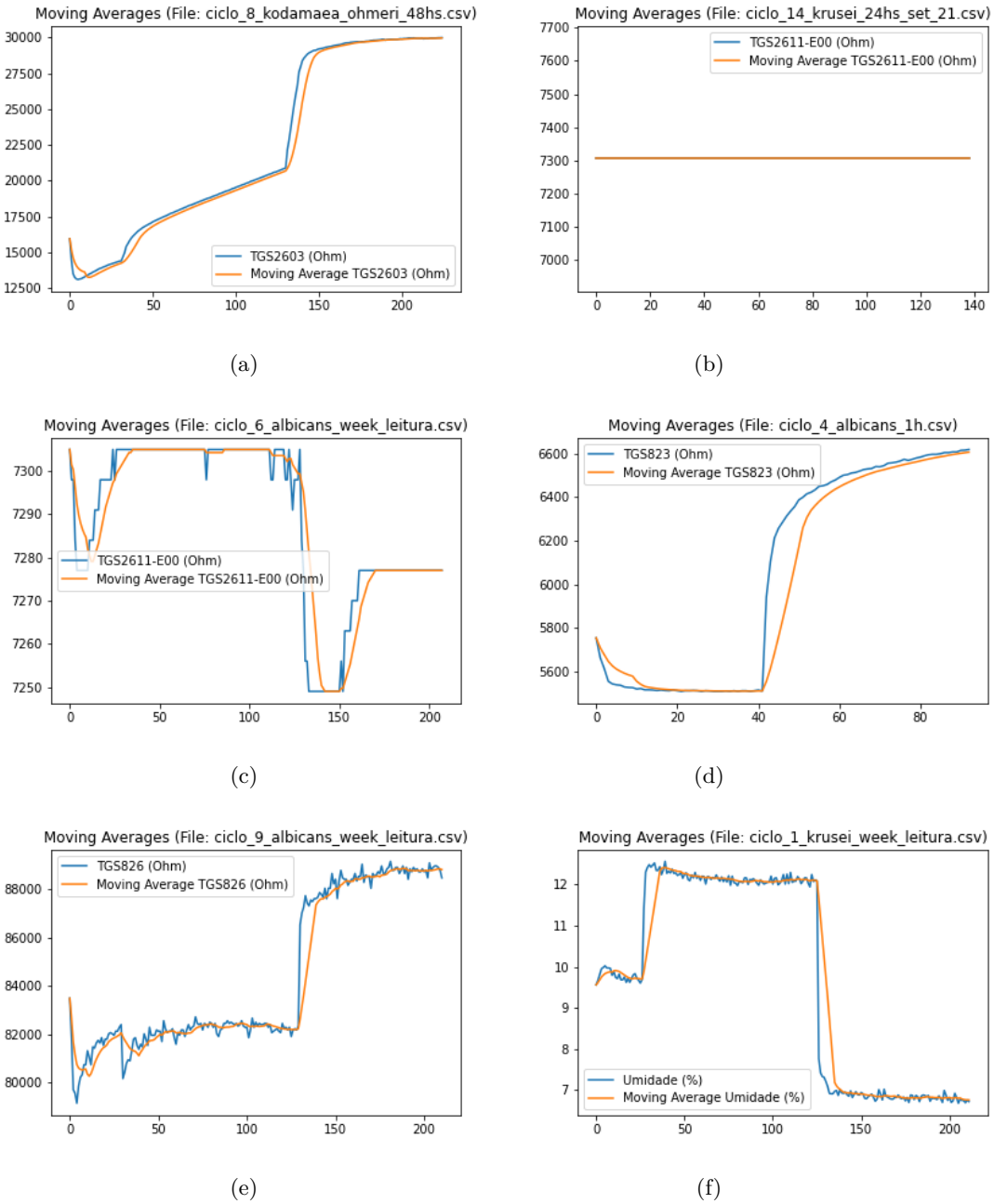
TARGET	TEMPO_APOS_	MA_TGS826_0	MA_TGS826_1	MA_TGS826_2	MA_TGS826_3	MA_TGS826_4	MA_TGS826_5	MA_TGS826_6	MA_TGS826_7	MA_TGS826_8	MA_TGS826_9	MA_TGS826_1(	MA_TGS826_11	MA_TGS826_12	MA_TGS826_13	MA_TGS826_14	MA_TGS826_
1	1	59080.0	55335.0	54.723.333.333.	54922.5	55418.0	56145.0	5.678.857.142.8	57383.75	57990.0	58509.0	58953.0	60173.0	61244.0	62151.0	62895.0	63405.0
1	1	95297.0	91142.0	87309.0	83582.0	79914.0	76292.0	72822.0	69326.0	65882.0	62484.0	63231.0	64337.0	65186.0	65881.0	66455.0	66964.0
1	1	104816.0	100123.0	95564.0	91093.0	86722.0	82425.0	78187.0	73988.0	69821.0	65640.0	66033.0	66685.0	67201.0	67597.0	67936.0	68219.0
1	24	91190.0	88890.0	8.754.333.333.3	86685.0	86238.0	8.594.833.333.3	8.577.428.571.4	85665.0	8.558.666.666.6	85541.0	84958.0	84830.0	84784.0	84921.0	85018.0	85128.0
1	24	100136.0	98408.0	96580.0	94783.0	92811.0	90858.0	88901.0	87091.0	85197.0	83217.0	82615.0	82442.0	82420.0	82512.0	82616.0	82742.0
1	24	99899.0	98314.0	96496.0	94580.0	92713.0	90751.0	88771.0	86851.0	84937.0	83030.0	82208.0	81912.0	81858.0	81896.0	81987.0	82073.0
1	24	99779.0	98079.0	96207.0	94158.0	92103.0	90114.0	88120.0	86132.0	84180.0	82234.0	81614.0	81389.0	81335.0	81410.0	81511.0	81575.0
1	24	99605.0	98171.0	96346.0	94392.0	92303.0	90215.0	88248.0	86226.0	84205.0	82194.0	81140.0	80532.0	80382.0	80361.0	80486.0	80600.0
1	24	99501.0	97917.0	96029.0	94048.0	92077.0	90075.0	88095.0	86089.0	84077.0	82087.0	81197.0	80863.0	80820.0	80862.0	80915.0	81047.0
1	24	99828.0	98557.0	96956.0	95080.0	93079.0	91061.0	89110.0	87082.0	85060.0	83069.0	81668.0	81037.0	80787.0	80729.0	80782.0	80935.0
1	24	99223.0	97898.0	96224.0	94285.0	92278.0	90250.0	88227.0	86199.0	84171.0	82557.0	81263.0	80601.0	80370.0	80292.0	80328.0	80375.0
1	24	99336.0	97880.0	96079.0	94092.0	92043.0	90042.0	87907.0	85802.0	83733.0	81697.0	80467.0	79899.0	79689.0	79648.0	79679.0	79761.0
1	24	98660.0	96864.0	94938.0	92883.0	90842.0	88816.0	86686.0	84571.0	82436.0	80386.0	79735.0	79452.0	79369.0	79410.0	79430.0	79491.0
1	24	99482.0	98211.0	96523.0	94651.0	92669.0	90640.0	88660.0	86659.0	84625.0	82612.0	81184.0	80437.0	80060.0	79982.0	79972.0	80013.0
1	24	99007.0	97529.0	95841.0	93933.0	91956.0	89922.0	87918.0	85894.0	83865.0	81857.0	80772.0	80270.0	80090.0	80059.0	80054.0	80121.0
1	24	99200.0	97732.0	95905.0	93982.0	92027.0	89973.0	87982.0	85948.0	83909.0	81886.0	80728.0	80183.0	79982.0	79930.0	79930.0	79956.0
1	24	98508.0	96771.0	94851.0	92820.0	90725.0	88641.0	86595.0	84465.0	82328.0	80219.0	79450.0	79110.0	78986.0	78955.0	78965.0	79000.0
1	24	98589.0	97616.0	96050.0	94204.0	92239.0	90301.0	88257.0	86203.0	84152.0	82434.0	80646.0	79696.0	79261.0	79096.0	79045.0	79010.0
1	24	98471.0	97215.0	95580.0	93742.0	91708.0	89679.0	87671.0	85680.0	83703.0	82075.0	80673.0	79933.0	79595.0	79450.0	79394.0	79420.0
1	24	98746.0	97527.0	95877.0	93981.0	92015.0	89951.0	87890.0	85962.0	83934.0	81883.0	80366.0	79511.0	79088.0	78910.0	78880.0	78905.0
1	24	98007.0	96503.0	94654.0	92704.0	90688.0	88700.0	86677.0	84666.0	82648.0	80735.0	79798.0	79351.0	79171.0	79125.0	79155.0	79160.0
1	24	98341.0	96791.0	94970.0	92941.0	90883.0	88797.0	86653.0	84535.0	82422.0	80319.0	79204.0	78695.0	78450.0	78370.0	78345.0	78365.0
1	24	98040.0	96675.0	94950.0	93038.0	91057.0	89029.0	86946.0	84928.0	82940.0	80882.0	79717.0	79149.0	78901.0	78845.0	78840.0	78865.0
5	3	87475.0	8.181.666.666.6	78302.5	76116.0	7.472.333.333.3	7.377.714.285.7	73141.25	7.272.111.111.11	72405.0	69737.0	68882.0	68841.0	69078.0	69399.0	69685.0	69925.0
5	3	113080.0	109787.0	105335.0	100518.0	95609.0	90754.0	85913.0	81158.0	76464.0	71809.0	68169.0	66876.0	66777.0	67083.0	67461.0	67820.0
5	3	110331.0	106354.0	101730.0	96866.0	92018.0	87254.0	82497.0	77815.0	73167.0	68793.0	66609.0	66070.0	66206.0	66537.0	66891.0	67193.0
5	3	109875.0	105683.0	100918.0	96062.0	91226.0	86512.0	81789.0	77081.0	72436.0	67841.0	66012.0	65645.0	65826.0	66128.0	66451.0	66748.0
5	3	109955.0	105860.0	101207.0	96429.0	91580.0	86778.0	82021.0	77276.0	72651.0	68033.0	66193.0	65759.0	65874.0	66166.0	66488.0	66760.0
6	1	91650.0	82965.0	7.764.333.333.3	74685.0	72984.0	7.193.666.666.6	7.130.714.285.7	70921.25	7.069.888.888.8	70581.0	68396.0	67985.0	68306.0	68791.0	69239.0	69652.0
6	1	108360.0	103808.0	99190.0	94679.0	90211.0	85848.0	81492.0	77165.0	72894.0	68716.0	67190.0	67584.0	68062.0	68469.0	68839.0	69108.0
6	1	107082.0	102777.0	98237.0	93748.0	89236.0	84823.0	80511.0	76219.0	71912.0	67712.0	66898.0	67006.0	67312.0	67630.0	67961.0	68155.0
6	1	108310.0	104472.0	99999.0	95386.0	90810.0	86377.0	82036.0	77648.0	73300.0	69036.0	66941.0	66555.0	66758.0	67099.0	67441.0	67692.0
6	1	108175.0	104127.0	99606.0	95031.0	90377.0	85901.0	81471.0	77034.0	72704.0	68244.0	66502.0	66321.0	66578.0	66943.0	67334.0	67586.0
6	3	83660.0	78155.0	7.484.333.333.3	72890.0	71718.0	7.100.666.666.6	7.055.285.714.2	70271.25	7.010.888.888.8	70014.0	68628.0	68359.0	68554.0	68892.0	69226.0	69531.0
6	3	113294.0	109518.0	104884.0	100004.0	95083.0	90177.0	85328.0	80524.0	75724.0	70989.0	68331.0	67492.0	67437.0	67651.0	67930.0	68190.0

Fonte: A autora (2023).

5.1.5 Suavização do sinal

A presença de ruído em um conjunto de dados pode aumentar a complexidade do modelo e o tempo de aprendizado, o que degrada o desempenho dos algoritmos de aprendizado, evidenciando, assim, a importância desta etapa. Na Figura 43, observa-se essa eliminação o ruído para diferentes sensores.

Figura 43 – Ilustração do registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso com e sem a etapa de suavização de sinal.



Fonte: A autora (2023).

5.1.6 Análise experimental das coletas de dados de temperatura (em ºC), pressão (em kPa) e umidade (em %)

Sabe-se que fatores externos podem afetar o crescimento dos fungos, como a temperatura e a umidade. Um gráfico de *violinplot* é uma representação gráfica que combina elementos de um gráfico de caixa (*boxplot*) com um gráfico de densidade de probabilidade. Ele é útil para visualizar a distribuição de dados em relação a diferentes categorias ou parâmetros.

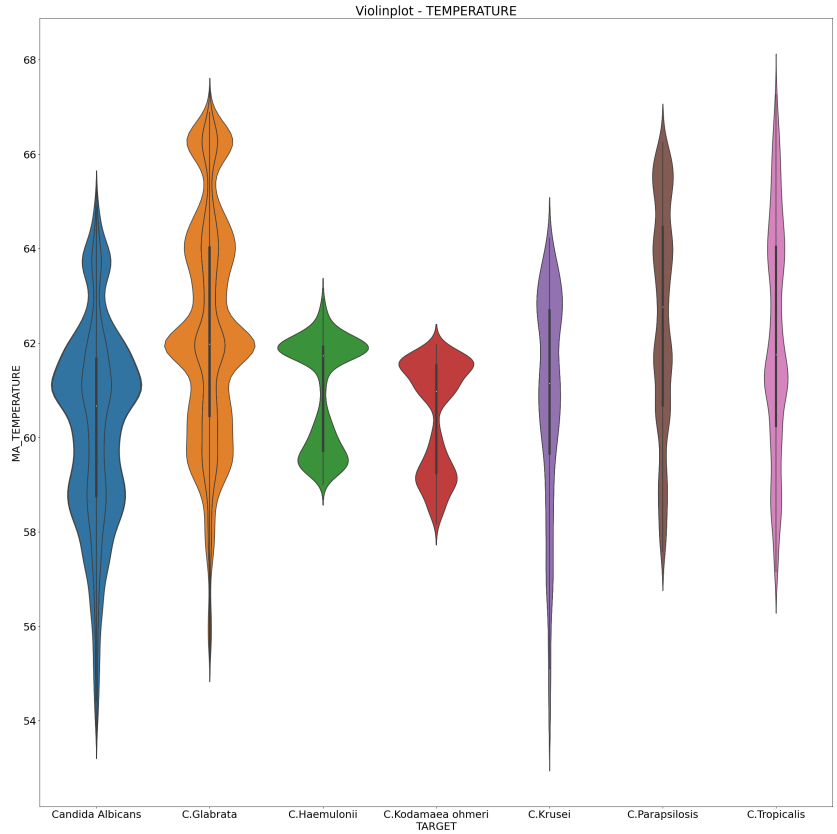
O violino é simétrico em relação ao eixo vertical e mostra a distribuição de dados para essa categoria. A parte mais larga do violino indica onde os valores são mais frequentes, enquanto as partes mais estreitas indicam áreas com menor frequência. A linha no centro do violino representa a mediana da distribuição para cada categoria e dentro de cada violino, é possível encontrar uma caixa que se assemelha a um gráfico de caixa (boxplot).

A Figura 44 apresenta os *violinplots* das leituras dos sensores de temperatura, pressão atmosférica e umidade dispostos no *e-nose* para cada uma das espécies fúngicas abordadas neste trabalho.

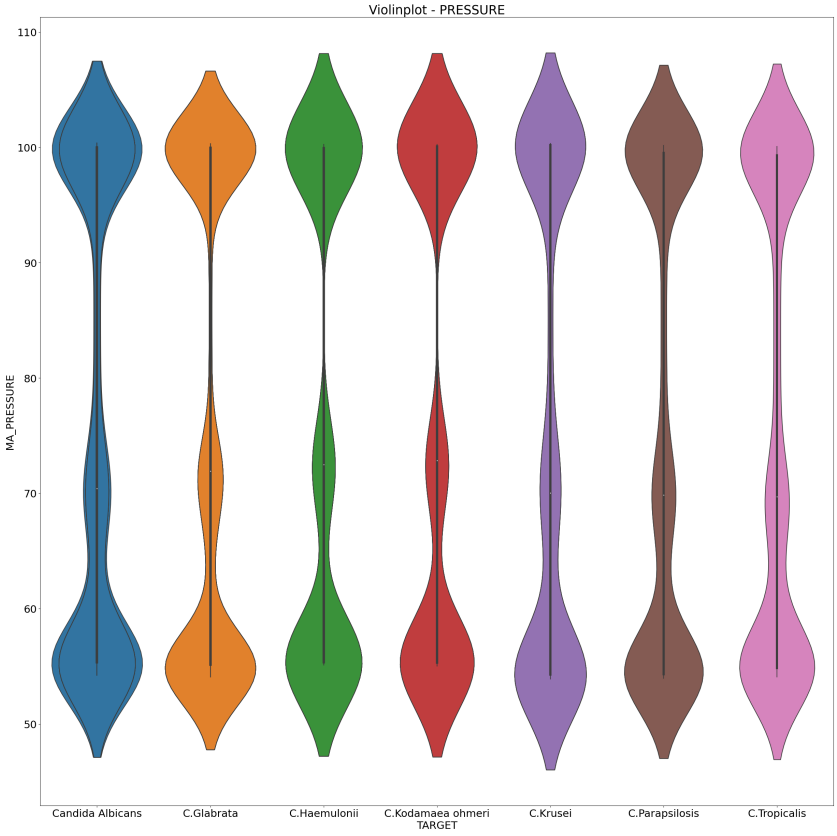
Ao analisar a forma dos violinos, pode-se avaliar se as variações das distribuições são comparáveis entre si. Quando os violinos exibem uma configuração similar, com larguras e medianas próximas, isso sugere que as distribuições têm uma dispersão e centralidade semelhantes. Isso implica que esses parâmetros não apresentam diferenças substanciais em relação à influência sobre a distribuição dos dados analisados, razão pela qual podem ser excluídos. Essa situação se aplica especificamente à pressão atmosférica.

No entanto, no caso da temperatura e umidade, é evidente que existe variabilidade nas distribuições. É importante ressaltar que, neste estudo, não foi possível estabelecer um ambiente controlado metrologicamente. Como resultado, as flutuações nas temperaturas e umidade nos ambientes, podem exercer um impacto significativo sobre os resultados das medições realizadas. Para evitar erros decorrentes dessa influência, optou-se por excluir a temperatura e pressão atmosférica dos parâmetros utilizados na identificação dos fungos.

Figura 44 – *Violinplots* das leituras dos sensores de (a) temperatura (em ºC), (b) pressão (em kPa) e (c) umidade (em %) das espécies fúngicas: *Candida Albicans*,*C. Krusei* , *C. Glabrata*,*C. Parapsilosis*, *C. haemulonii* ,*C. Tropicalis* e *C. Kodamaea ohmeri*.



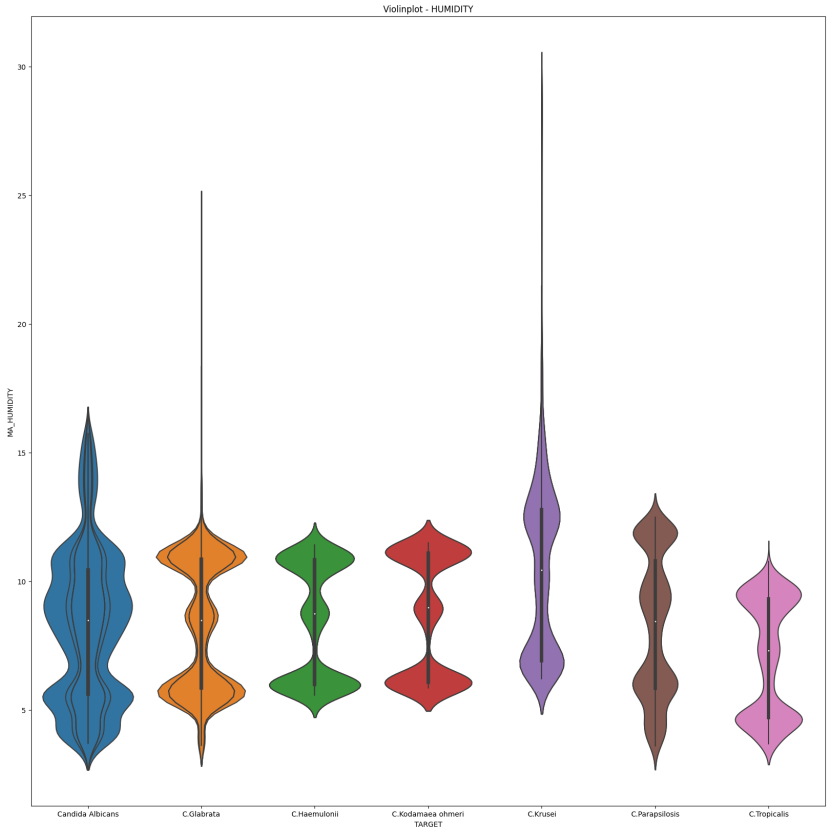
(a)



(b)

Fonte: A autora (2023).

Figura 44 – *Violinplots* das leituras dos sensores de (a) temperatura (em ºC), (b) pressão (em kPa) e (c) umidade (em %) das espécies fúngicas: *Candida Albicans*,*C. Krusei* , *C. Glabrata*,*C. Parapsilosis*, *C. haemulonii* ,*C. Tropicalis* e *C. Kodamaea ohmeri*.



(c)

Fonte: A autora (2023).

5.2 RESULTADOS ALCANÇADOS

Nesta seção serão apresentados os resultados alcançados, apresentando os pré-processamentos realizados sobre a base de dados:

5.2.1 Experimento I: Conjunto de dados final

Neste experimento, o conjunto de dados final foi submetido a um processo de suavização e normalização de dados. É importante recordar que, para o treinamento, foi empregada a técnica de validação cruzada, citado na seção 4.4.1.1. Nesse método, em cada repetição, um fold (ou subconjunto) dos dados é selecionado aleatoriamente como conjunto de validação, enquanto os K-1 folds restantes são utilizados como conjunto de treinamento. Consequentemente, os resultados obtidos neste experimento serão apresentados na curva de aprendizagem para métricas específicas para cada fold das 10 repetições da validação cruzada. Esse procedimento ajuda a fornecer uma visão geral mais robusta e confiável do desempenho do modelo, considerando a variabilidade inerente aos diferentes conjuntos de validação utilizados em cada repetição do processo de validação cruzada. As figuras 46 e

45 ilustram a curva de aprendizagem tanto para treinamento quanto para validação.

Além disso, o tempo de processamento é um indicador importante para avaliar o custo computacional e, consequentemente, os recursos necessários para implementar o modelo em ambientes de produção. Por isso, a Tabela 12 apresenta o tempo de processamento (em seg.) do conjunto de treinamento para os classificadores analisados.

Tabela 12 – Tempo de processamento (seg.) no conjunto de dados de treinamento para o Experimento I

Classificador	**Tempo (seg)**
TimeSeries Forest	323.85
KNeighbors TimeSeries	**6.41**
Rocket	809.39
HIVE-COTE V2	1567.3
TapNet	614.22

Fonte: A autora (2023).

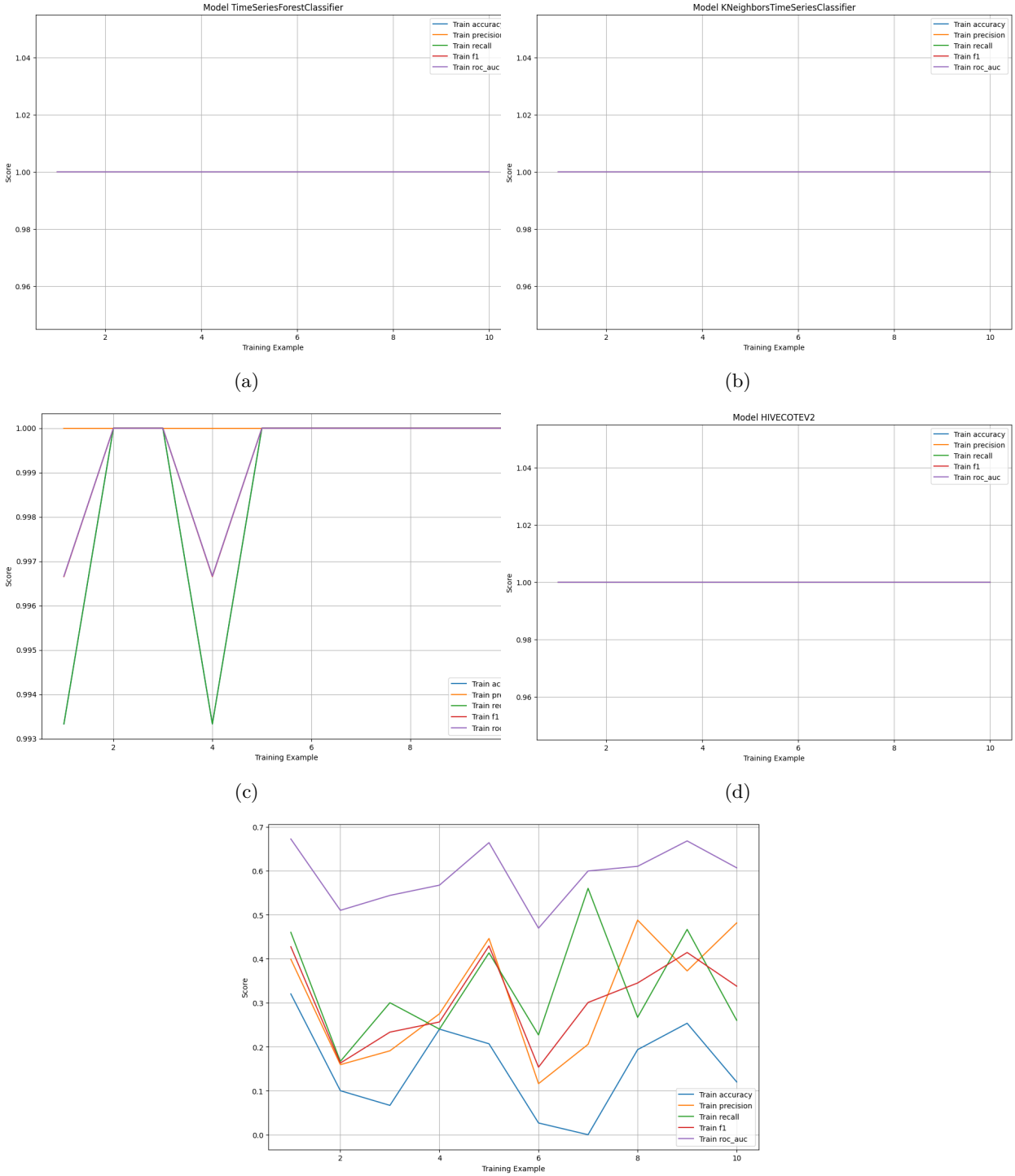
Analisando os resultados no processo de validação, com exceção do classificador TapNet, os resultados foram satisfatórios, com acurácias variando de 84% a 100%, com precisão variando de 87% a 100%, com especificidades variando de 86% a 100% e F1-Score variando de 82% a 100%. De maneira geral, os desvios padrões foram baixos para as acurácias, sensibilidades e especificidades dos modelos. Além disso, observa-se que os tempos de processamento foram bem distintos, variando entre 6.41 segundos a 1567.3 segundos (mais de 26 minutos).

Diferenças substanciais nessas métricas entre os diferentes classificadores podem ter um impacto significativo na aplicação prática do modelo. Por exemplo, a preferência por um classificador com alta sensibilidade pode ser justificável em cenários onde a detecção de infecções fúngicas é crítica, como em pacientes imunocomprometidos. No entanto, se o mesmo classificador sacrifica a especificidade em detrimento da sensibilidade, ele pode gerar um aumento de falsos positivos, levando a tratamentos desnecessários. Portanto, a escolha do classificador deve ser criteriosamente ponderada, levando em consideração as implicações clínicas e práticas das métricas de sensibilidade e especificidade dentro do contexto específico de aplicação.

Ao examinar a matriz de confusão da validação cruzada, é possível identificar padrões de erros que podem não ser evidentes na métrica de desempenho global, como a acurácia. Estas matrizes estão apresentadas na Figura 47. Isso ajuda a entender melhor como o modelo lida com diferentes classes, devido ao desequilíbrios de classe que precisam ser abordados. Além disso, a análise da matriz de confusão contribui para a identificação de oportunidades de ajuste de hiperparâmetros e melhorias no modelo, resultando em uma avaliação mais robusta e informada do seu desempenho.

A presença de uma classe com um menor percentual na matriz de confusão é um fenômeno relativamente comum, especialmente em problemas de classificação desequilibrados, nos quais as classes têm tamanhos muito diferentes. Isso ocorre porque o modelo de classificação tende a se sair melhor na previsão das classes majoritárias, uma vez que há mais exemplos para aprender. No entanto, para este primeiro experimento, apesar da *C. Parapsilosis* ser uma classe majoritária, independente do classificador, no processo de validação obtive o menor percentual de acerto. Este reconhecimento será importante para adotar estratégias, nas próximas experimentações, para lidar com o desequilíbrio, a fim de melhorar o desempenho de classificação nessas classes.

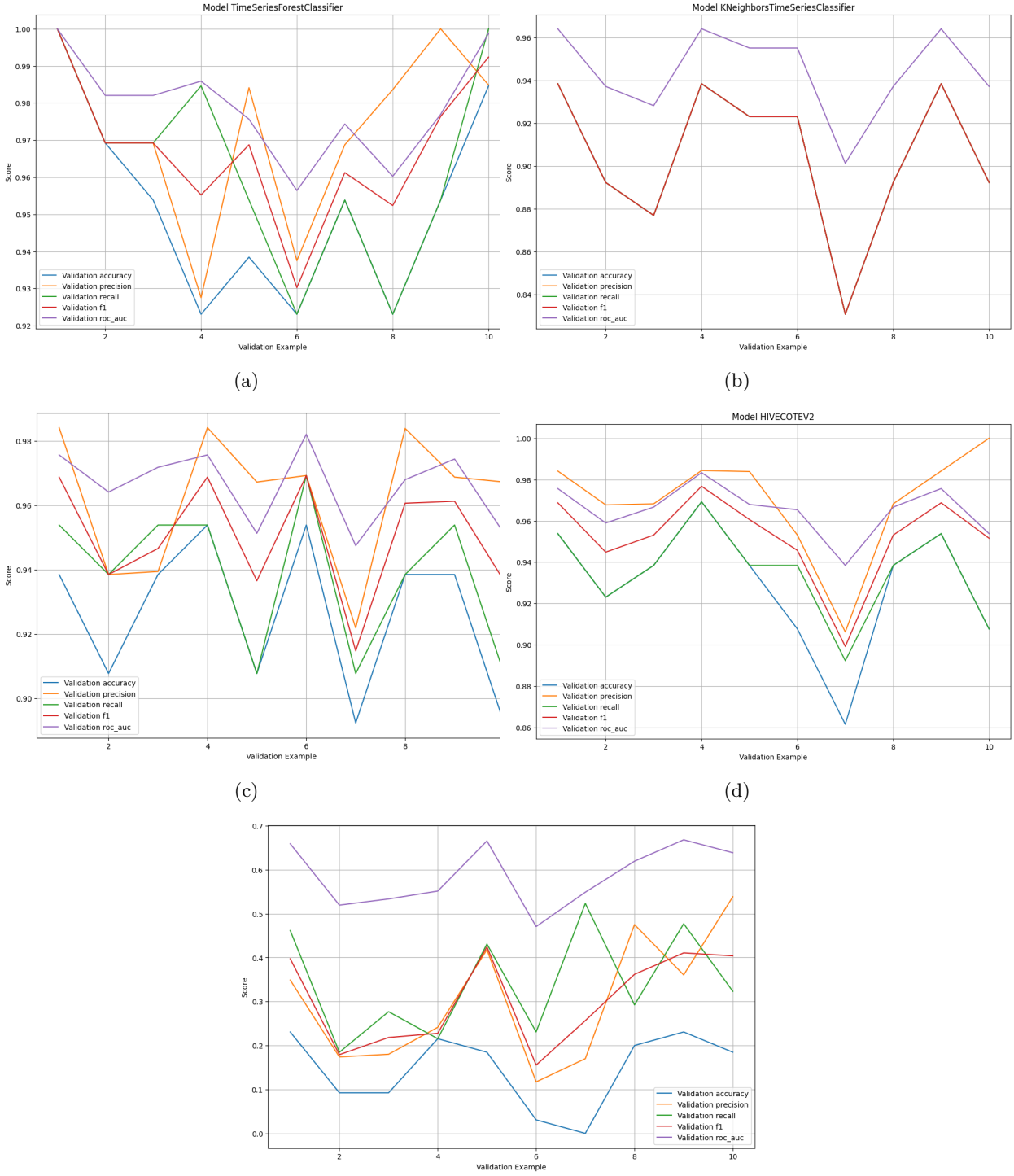
Figura 45 – Experimento I - Curva de aprendizagem das 10 repetições da validação cruzada k-Fold das métricas acurácia, sensibilidade, especificidade, precisão e recall dos 5 modelos de IA para o conjunto de treinamento. (a) Classificador KNeighborsTime, (b) Classificador Rocket, (c) Classificador TimesSeriesForest, (d) Classificador HIVE-COTE V2, (e) Classificador TapNeT



(a) (b) (c) (d) (e)

Fonte: A autora (2023).

Figura 46 – Experimento I - Curva de aprendizagem das 10 repetições da validação cruzada k-Fold das métricas acurácia, sensibilidade, especificidade, precisão e recall dos 5 modelos de IA para o conjunto de validação. (a) Classificador KNeighborsTime, (b) Classificador Rocket, (c) Classificador TimesSeriesForest, (d) Classificador HIVE-COTE V2, (e) Classificador TapNeT



(a)

(b)

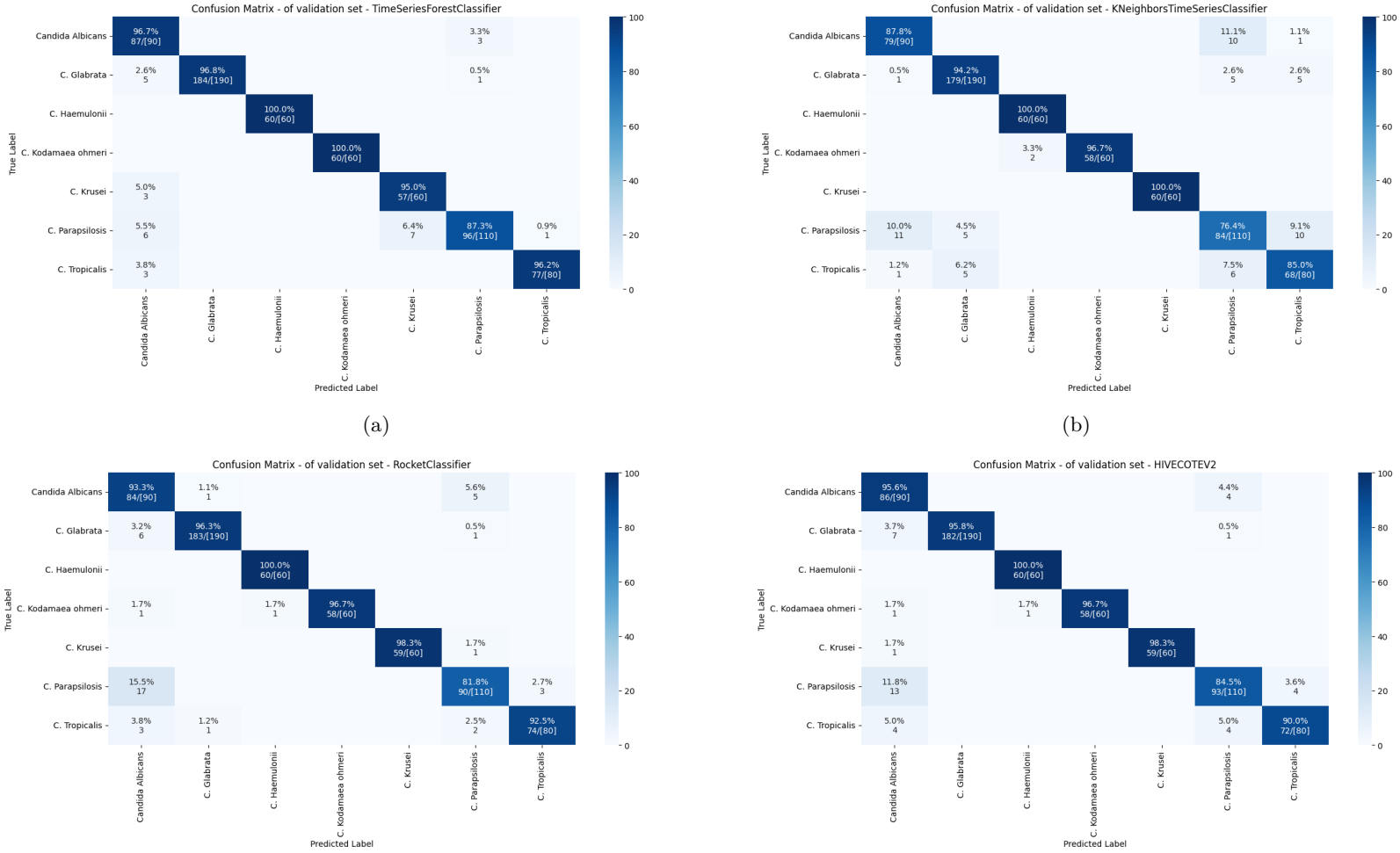
(c)

(d)

(e)

Fonte: A autora (2023).

Figura 47 – Experimento I - Matriz de confusão por classificador no conjunto de validação. (a) Classificador KNeighborsTime, (b) Classificador Rocket, (c) Classificador TimesSeriesForest, (d) Classificador HIVE-COTE V2, (e) Classificador TapNeT

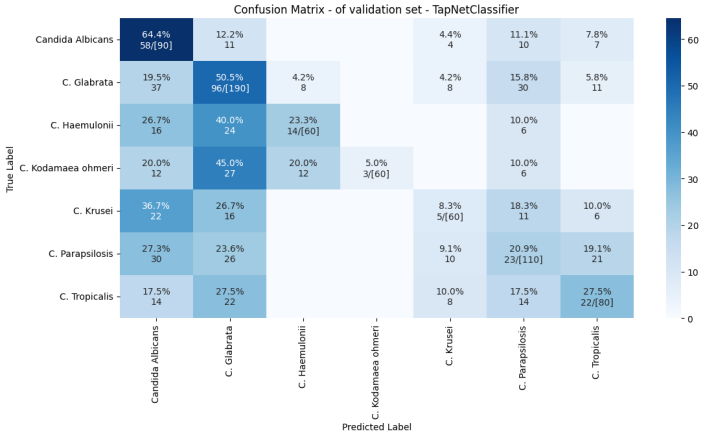


(a) (b)

(c) (d)

Fonte: A autora (2023).

Figura 47 – Experimento I - Matriz de confusão por classificador no conjunto de validação. (a) Classificador KNeighborsTime, (b) Classificador Rocket, (c) Classificador TimesSeriesForest, (d) Classificador HIVE-COTE V2, (e) Classificador TapNeT



(e)

Fonte: A autora (2023).

Como citado na seção 4.4.1.1, 30% do conjunto de dados foi utilizado como conjunto de testes. A tabela seguir ilustra os resultados encontrados para cada classificador neste conjunto:

Tabela 13 – Resultados Experimento I no conjunto de dados de teste: conjunto de dados final passou pela etapa de suavização e normalização de dados.

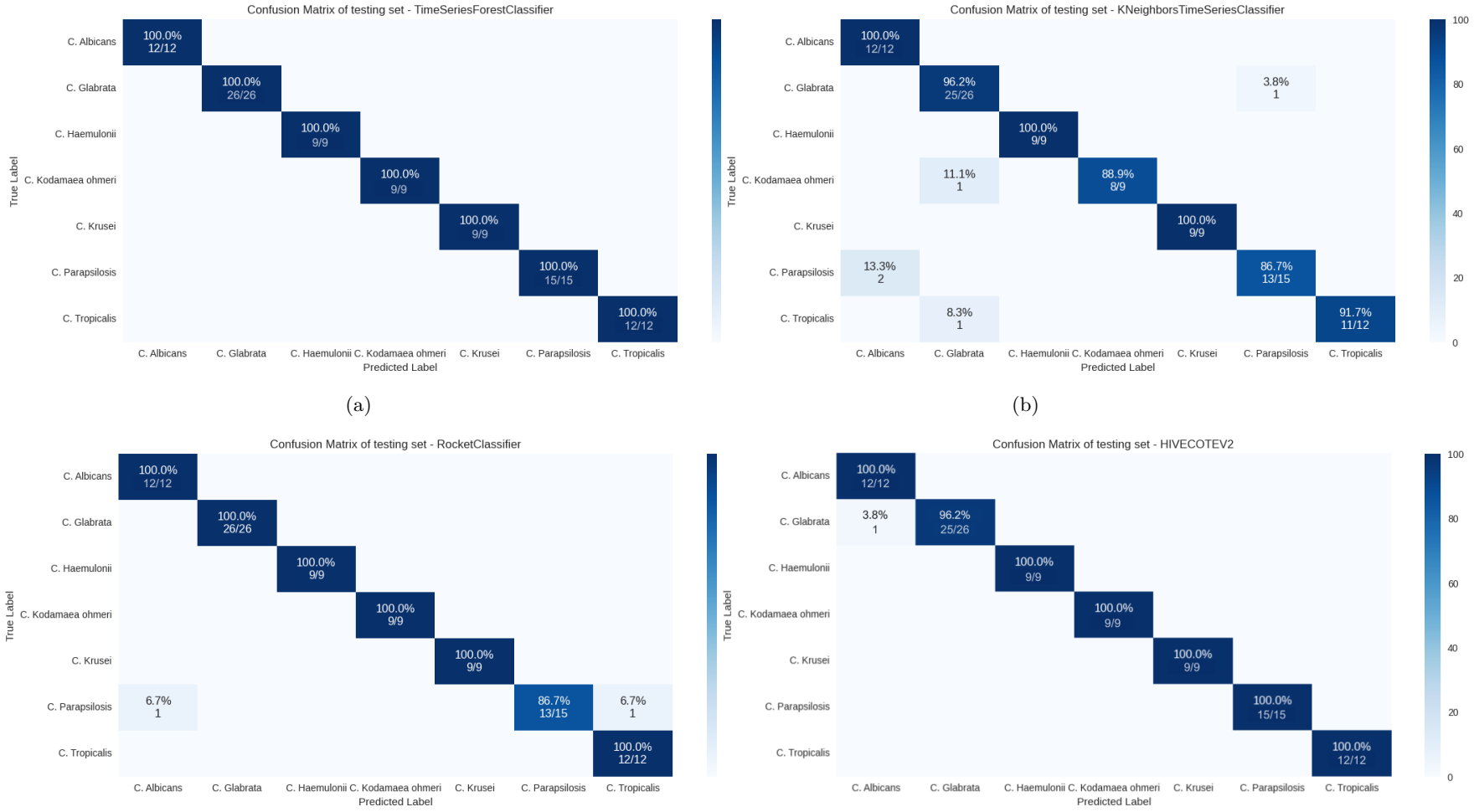
Classificador	**Acurácia**	**Precisão**	**Especificidade**	**F1-Score**
TimeSeries Forest	**100%**	**100%**	**100%**	**100%**
KNeighbors TimeSeries	94.6%	96.0%	94.7%	96.3%
Rocket	97.8%	97.3%	98.0%	97.6%
HIVE-COTE V2	99%	99%	99%	99%
TapNet	40%	41%	38%	51.3%

Fonte: A autora (2023).

Assim, o classificador TimeSeries Forest ficou em primeiro lugar nas 4 métricas, seguidos do HIVECOTEV2 e RocketClassifier. Já o estudo das matrizes de confusão exibe a distribuição dos registros em termos de suas classes atuais e de suas classes previstas. Neste trabalho, serão ilustradas as matriz de confusão encontradas sobre conjunto de teste. Mostradas na Figura 48, as matrizes apresentam semelhanças sutis dos perfis dos compostos voláteis dos sensores. Com relação ao desempenho insatisfatório do TapNet na análise de séries temporais do COVs pode ser atribuído, em grande parte, à complexidade das séries temporais multivariadas. O TapNet considera as dependências temporais entre os pontos de dados e, por isso, enfrentou dificuldades em capturar as relações complexas entre as variáveis e em lidar com as interações dinâmicas presentes nesse tipo de dado. Isso pode ter resultando em perda de informações relevantes e, consequentemente, na incapacidade do modelo em realizar tarefas de análise de séries temporais de forma eficaz.

Além disso, a possibilidade de adotar o classificador KNeighbors TimeSeries como uma escolha inteligente deve ser cuidadosamente considerada. Embora possa haver uma perda de desempenho nas métricas, é importante reconhecer o valor substancial de ganho no tempo de processamento. Em certos contextos, a velocidade de processamento pode ser de maior importância, particularmente quando se lida com sistemas em tempo real ou aplicações que requerem respostas rápidas. Uma justificativa sólida para essa escolha reside na necessidade de otimizar recursos computacionais, manter a eficiência e garantir uma resposta oportuna, superando as limitações de desempenho que podem ser aceitáveis em favor da agilidade e da capacidade de resposta em determinados cenários. Portanto, a decisão de priorizar o tempo de execução pode ser estratégica e benéfica, servindo como base para a introdução de um segundo experimento.

Figura 48 – Experimento I - Matriz de confusão por classificador no conjunto de testes. (a) Classificador KNeighborsTime, (b) Classificador Rocket, (c) Classificador TimesSeriesForest, (d) Classificador HIVE-COTE V2,

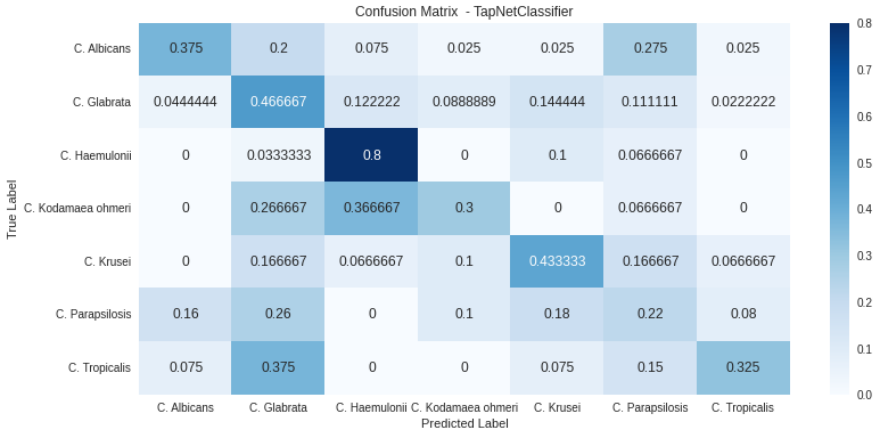


(a) (b)

(c) (d)

Fonte: A autora (2023).

Figura 48 – Experimento I - Matriz de confusão por classificador no conjunto de testes. (e) Classificador TapNeT



(e)

Fonte: A autora (2023).

5.2.2 Experimento II: Seleção de Recursos

Para este segundo experimento, o enfoque principal do processo de pré-processamento é a redução do custo computacional, especificamente o tempo de processamento. No entanto, em diferentes contextos, esse custo pode englobar outros fatores, como o uso de memória, capacidade de armazenamento e poder de processamento (CPU), embora estes não sejam abordados neste estudo. Com o intuito de alcançar esse objetivo, foram identificados e descartados recursos constantes, ou seja, aqueles com valores únicos ou muito semelhantes em todas as observações do conjunto de dados. Esses recursos não contribuem de forma significativa para a previsão do resultado. Para realizar essa remoção, seguindo os critérios citados em subseção 4.4.5, empregou-se o método para elimina características de baixa e alta correlacionalidade. Esse processo resultou na formação de um novo conjunto de dados contendo 308 registros com 8 características. A Figura 49 ilustra a matriz de correlação desse novo conjunto de dados.

A configuração de treinamento dos classificadores é um aspecto crucial em experimentos de aprendizado de máquina. Ela inclui hiperparâmetros e opções que afetam diretamente o desempenho do modelo. A Tabela 14 apresentão as melhores configurações consideradas ao treinar classificadores neste estudo.

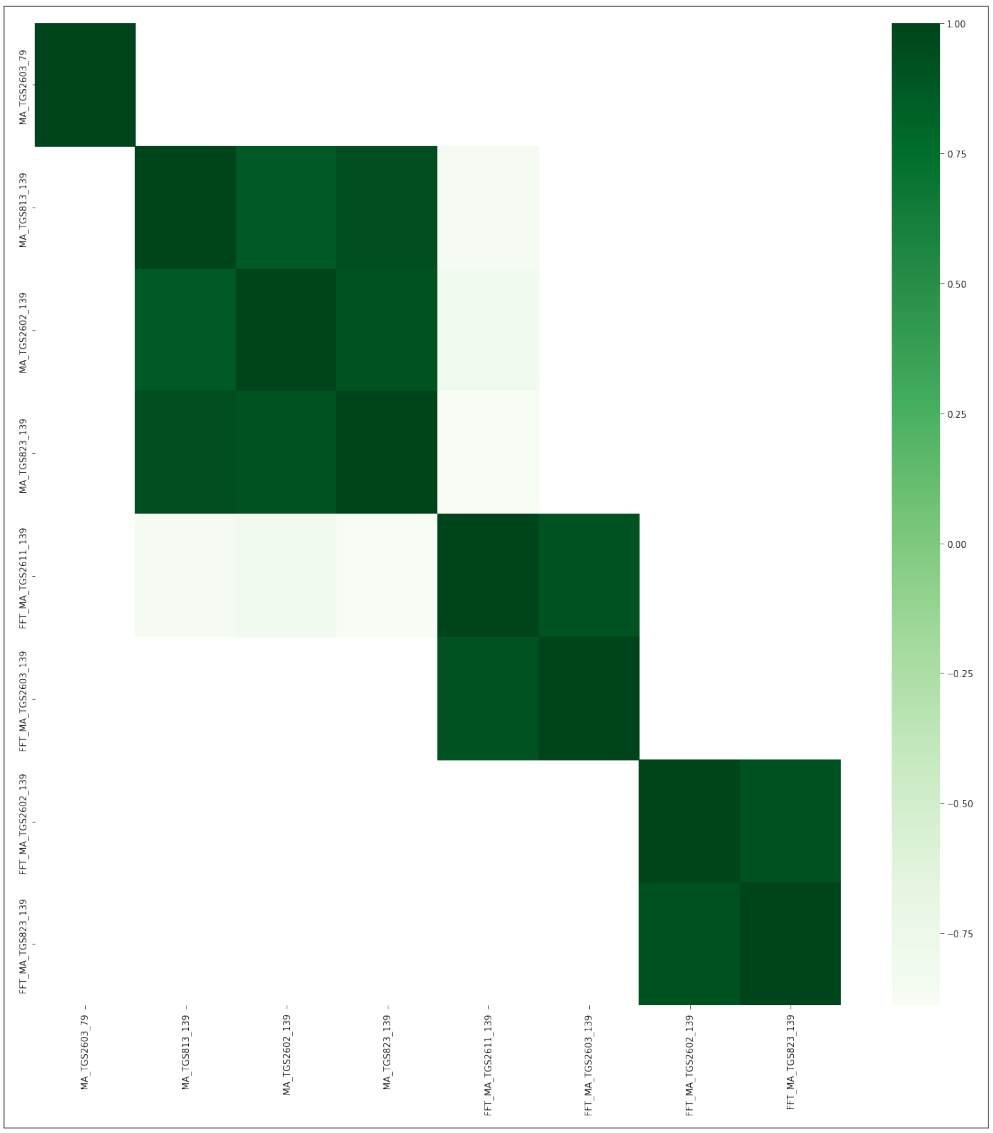
Tabela 14 – Configuração de treinamento dos classificadores

Classificador	**Parâmetros**
TimeSeries Forest	`n_estimators=80, random_state=42`
KNeighbors TimeSeries	`distance='euclidean', n_jobs=100`
Rocket	`num_kernels=300, random_state=42`
HIVE-COTE V2	`time_limit_in_minutes=0.25`
TapNet	`batch_size=20, n_epochs=120`

Fonte: A autora (2023).

Assim como no experimento anterior, antes do processo de classificação, o conjunto de dados passou pela etapa de suavização e normalização de dados. As figuras 50 e 51 ilustram, respectivamente, a curva dos resultados para treinamento e validação dos classificadores analisados. Enquanto que a Tabela 15 mostram o tempo de processamento (em seg.) no caso do conjunto de treinamento.

Figura 49 – Matriz de correlação linear dos sensores do *e-nose* após descarte de recursos. O verde mais escuro indica que mais forte é a correlação linear direta, o verde claro indica que mais forte é a correlação linear inversa.

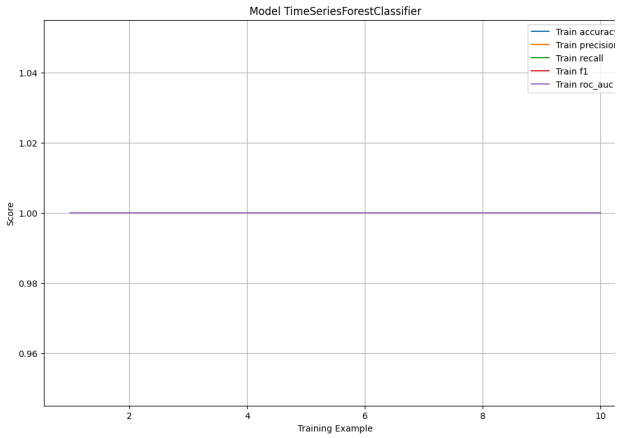


Fonte: A autora (2023).

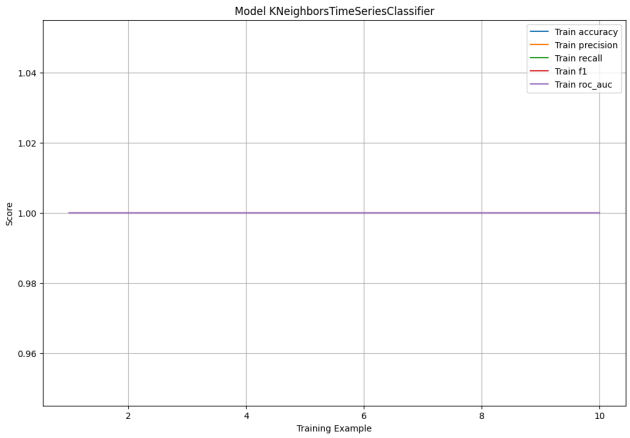
Assim como nos resultados encontrados no experimento anterior, os resultados encontrados foram satisfatórios, exceto para o classificador TapNet, porém neste caso, com um pequeno ganho. Neste experimento, possuem acurácias variando de 85% a 100%, com precisão variando de 90% a 100%, com especificidades variando de 87% a 100% e F1-Score variando de 91% a 100%. Além disso observa-se que os tempos de processamento foram reduziu de modo expressivo, variando de 0.58 a 326 segundos. Da mesma maneira, os desvios padrões foram baixos para as acurácias, sensibilidades e especificidades dos modelos.

O classificador TimeSeries Forest, KNeighbors TimeSeries e HIVECOTEV2 obtiveram o mesmo desempenho e ocuparam juntos o primeiro lugar nas 4 métricas. No experimento anterior, apresentou a possibilidade de adotar como solução o classificador KNeighbors

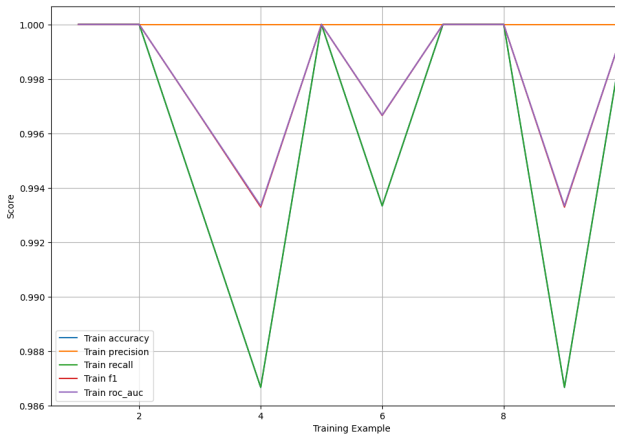
Figura 50 – Experimento II - Curva de aprendizagem das 10 repetições da validação cruzada k-Fold das métricas acurácia, sensibilidade, especificidade, precisão e recall dos 5 modelos de IA para o conjunto de treinamento. (a) Classificador KNeighborsTime, (b) Classificador Rocket, (c) Classificador TimesSeriesForest, (d) Classificador HIVE-COTE V2, (e) Classificador TapNeT



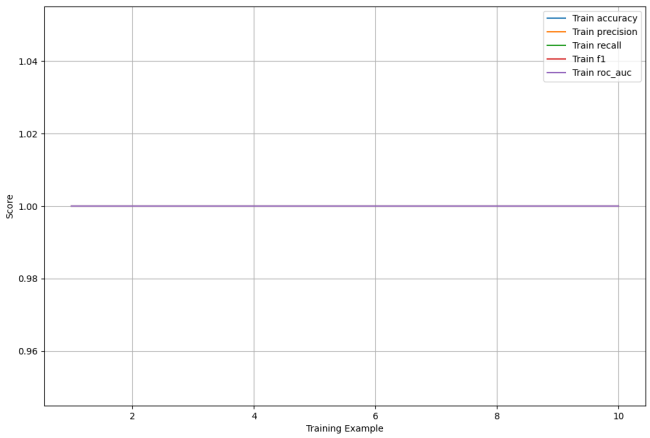
(a)



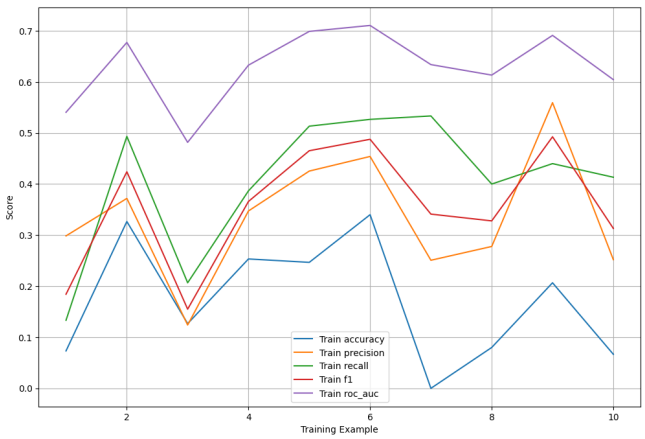
(b)



(c)



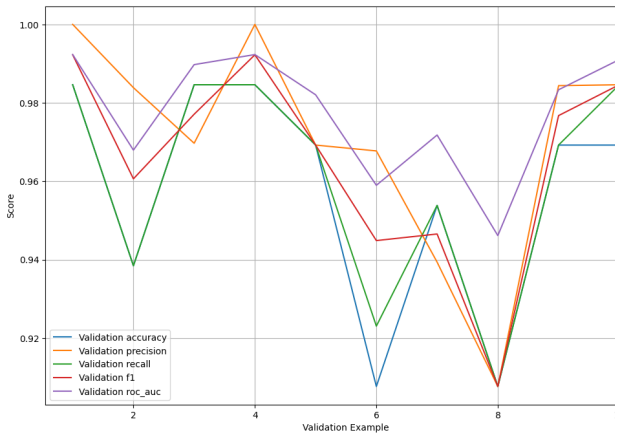
(d)



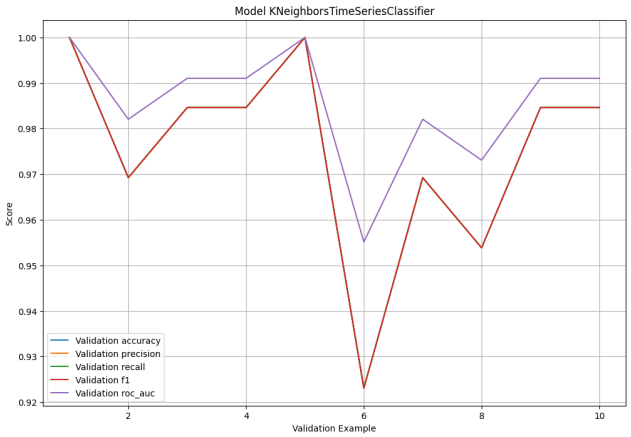
(e)

Fonte: A autora (2023).

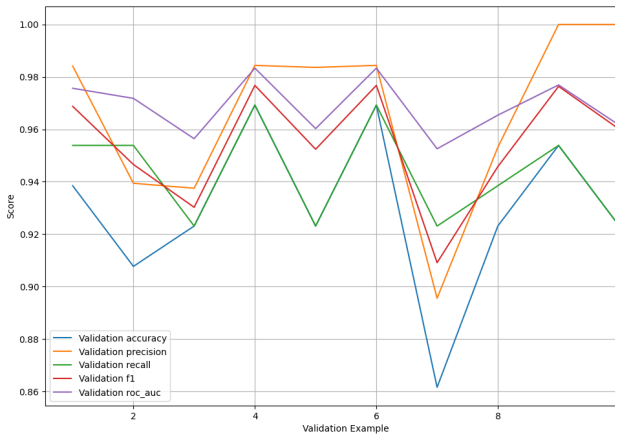
Figura 51 – Experimento II - Curva de aprendizagem das 10 repetições da validação cruzada k-Fold das métricas acurácia, sensibilidade, especificidade, precisão e recall dos 5 modelos de IA para o conjunto de validação. (a) Classificador KNeighborsTime, (b) Classificador Rocket, (c) Classificador TimesSeriesForest, (d) Classificador HIVE-COTE V2, (e) Classificador TapNeT



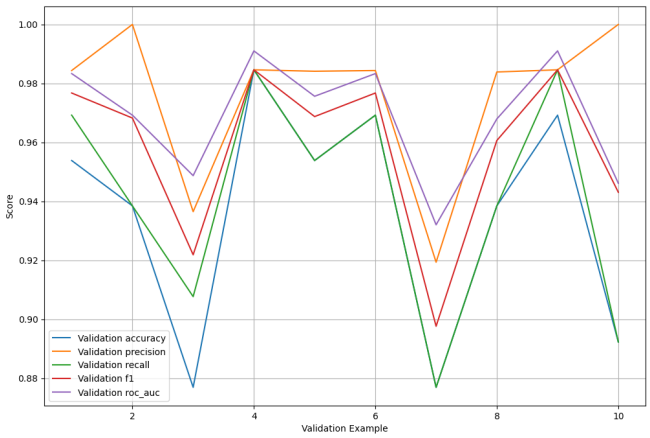
(a)



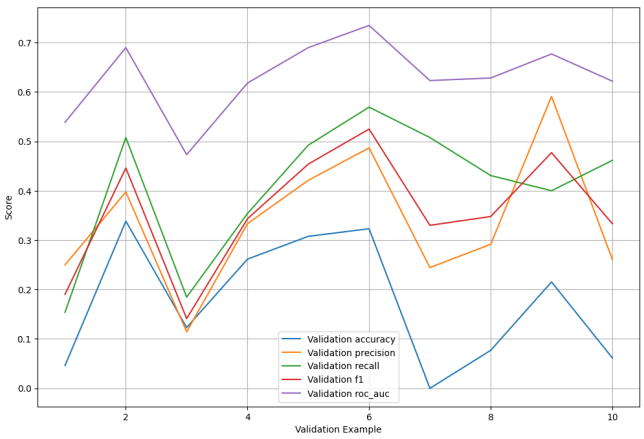
(b)



(c)



(d)



(e)

Fonte: A autora (2023).

Tabela 15 – Resultados Experimento II no conjunto de dados de treinamento: conjunto de dados final passou pela etapa de suavização e normalização de dados e aplicou-se o *Variance Threshold*, um seletor de recursos. Média e desvio padrão de 10 iterações da validação cruzada kFold das métricas acurácia, sensibilidade e especificidade, e do tempo de processamento ().

Classificador	**Tempo (seg)**
TimeSeries Forest	**0.58**
KNeighbors TimeSeries	0.94
Rocket	26.63
HIVE-COTE V2	326
TapNet	198.04

Fonte: A autora (2023).

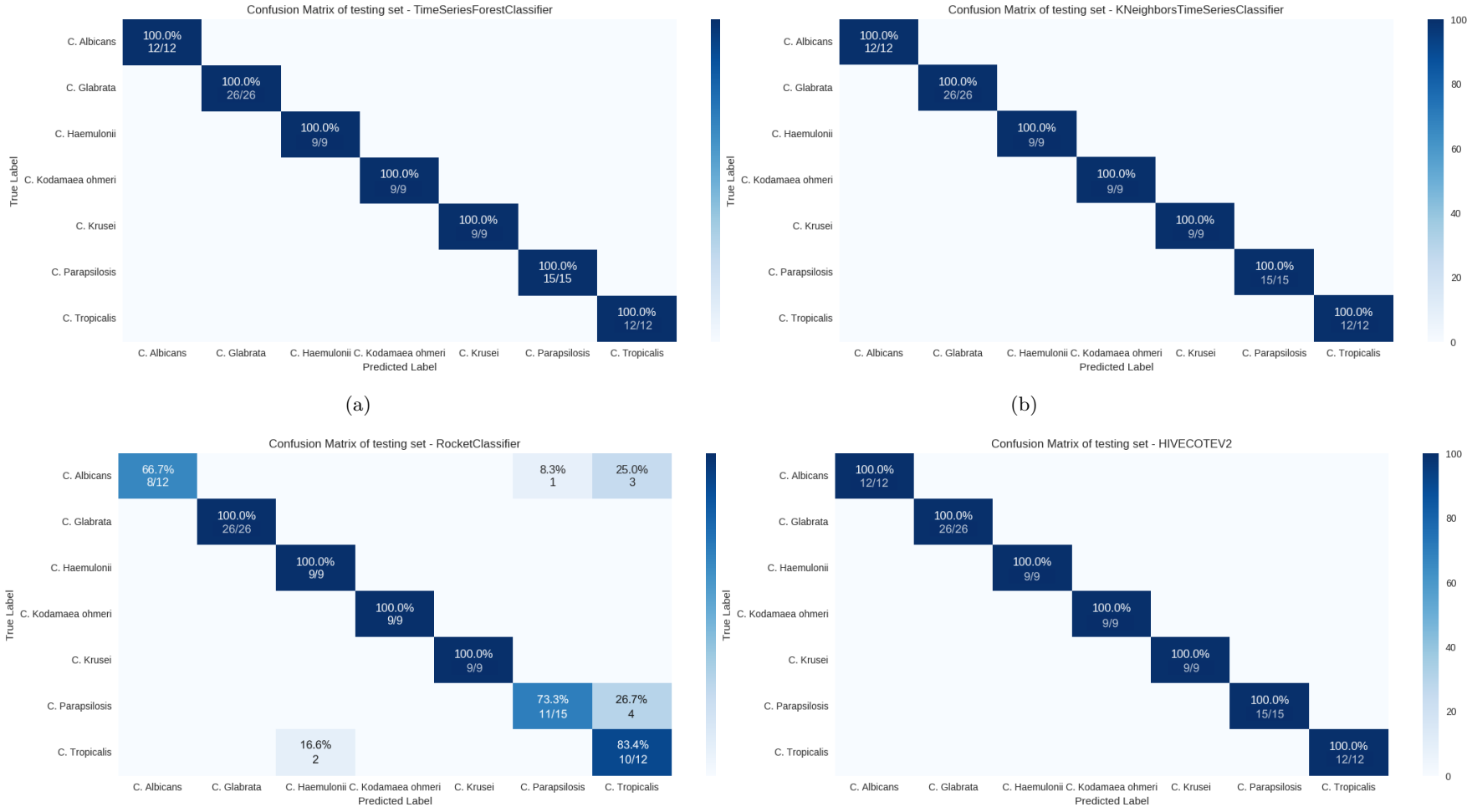
Tabela 16 – Resultados Experimento II no conjunto de dados de teste: conjunto de dados final passou pela etapa de suavização e normalização de dados e aplicou-se o *Variance Threshold*, um seletor de recursos.

Classificador	**Acurácia**	**Precisão**	**Especificidade**	**F1-Score**
TimeSeries Forest	**100.0%**	**100.0%**	**100.0%**	**100.0%**
KNeighbors TimeSeries	**100.0%**	**100.0%**	**100.0%**	**100.0%**
Rocket	87.0%	90.0%	88.7%	89.0%
HIVE-COTE V2	**100.0%**	**100.0%**	**100.0%**	**100.0%**
TapNet	41.9%	36.9%	44.1%	32.8%

Fonte: A autora (2023).

TimeSeries, o que foi consolidado neste experimento. O estudo das matrizes de confusão mostradas na Figura 52, que apresentam semelhanças dos perfis dos compostos voláteis captados dos sensores entre os classificadores desse experimento, porém com grande ganho no tempo de execução comparado ao experimento anterior.

Figura 52 – Experimento II - Matriz de confusão por classificador no conjunto de testes. (a) Classificador KNeighborsTime, (b) Classificador Rocket, (c) Classificador TimesSeriesForest, (d) Classificador HIVE-COTE V2, (e) Classificador TapNeT



(a)

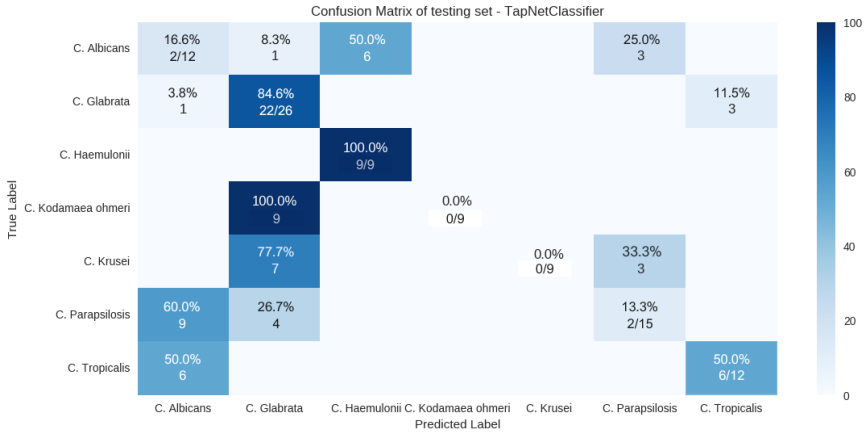
(b)

(c)

(d)

Fonte: A autora (2023).

Figura 52 – Experimento II - Matriz de confusão por classificador no conjunto de testes. (e) Classificador TapNeT



(e)

Fonte: A autora (2023).

5.2.3 Experimento III: Projeção *Uniform Manifold Approximation and Projection* (UMAP) das leituras dos sensores do *e-nose*

Como mencionado na subseção 4.4.6, o UMAP é uma coleção de técnicas que englobam a aproximação, redução de dimensão e projeção (visualização) de dados. Neste estudo, o UMAP foi empregado para examinar as interdependências entre classes em conjuntos de dados extensos e complexos. Especificamente, realizou-se a projeção UMAP das leituras dos sensores do dispositivo *e-nose* com base no gênero fúngico. Essa projeção proporciona uma representação visual que revela a sobreposição das classes nos conjuntos de dados, como ilustrado na Figura 37. Portanto, neste experimento, o UMAP foi aplicado ao conjunto de dados resultante do Experimento II, com o objetivo de mitigar essas sobreposições. A Figura 53 exibe as projeções obtidas.

Embora a projeção UMAP das leituras dos sensores do *e-nose* por categoria ainda revele alguma sobreposição entre as classes, é notável que essa sobreposição seja relativamente reduzida. Consequentemente, é provável que os modelos de classificação alcancem um desempenho satisfatório ao lidar com esses conjuntos de dados. As figuras 54 e 55 ilustram a curva de aprendizagem para treinamento e validação para este terceiro experimento, enquanto que a Tabela 17 mostra os resultados do tempo de processamento (em seg.) para o conjunto de validação.

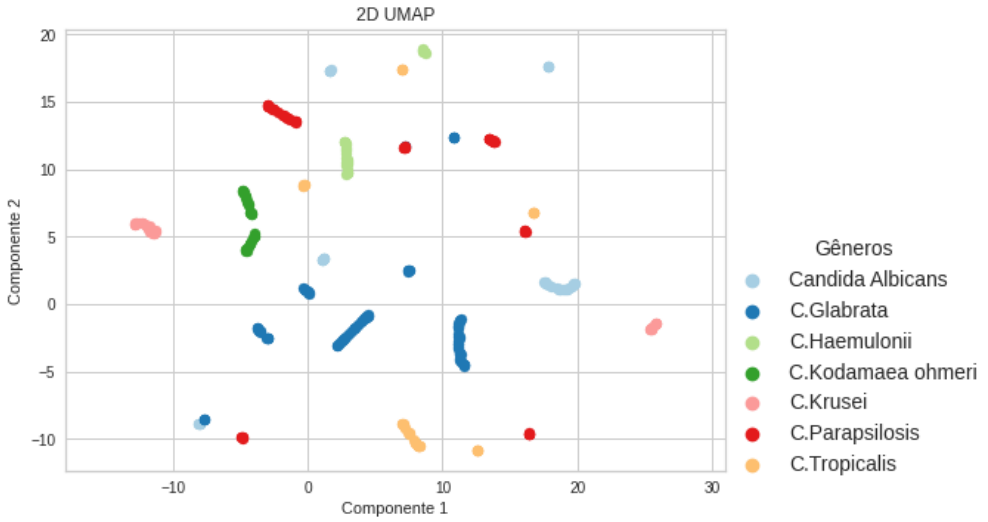
Neste experimento, diferentemente dos experimentos anteriores, é notável a variação do desempenho dos classificadores, com o TapNet apresentando um desempenho consideravelmente superior em comparação com experimentos anteriores, ao passo que o Rocket teve uma queda acentuada, exige uma análise aprofundada. Essa inconsistência pode ser atribuída a diversos fatores, como a sensibilidade dos modelos a técnicas de pré-processamento, incluindo a redução de dimensão com o UMAP, bem como a natureza intrínseca e os requisitos específicos de cada modelo.

Tabela 17 – Resultados Experimento III: aplicação do UMAP no conjunto de dados formado no Experimento II. Tempo de processamento para o conjunto de validação

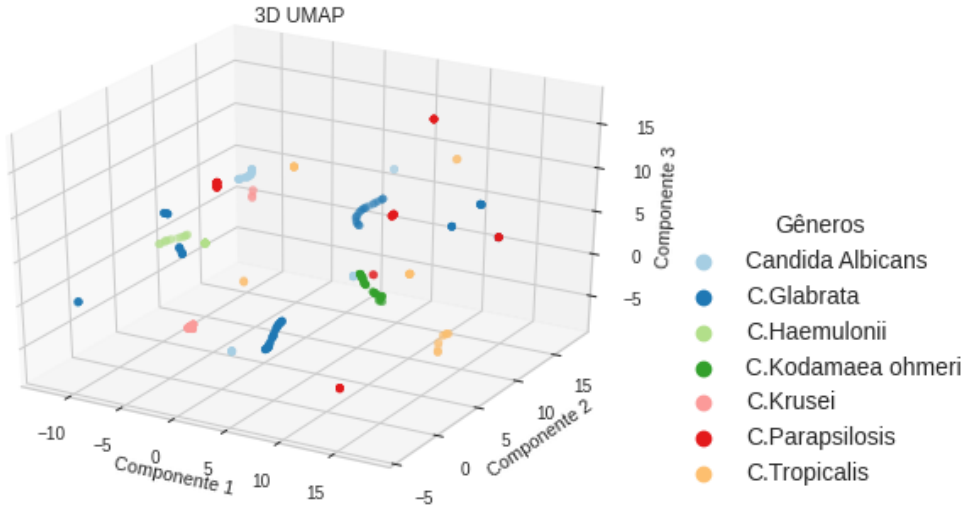
Classificador	**Tempo (seg)**
TimeSeries Forest	**0.17**
KNeighbors TimeSeries	0.70
Rocket	33.64
HIVE-COTE V2	252.7
TapNet	1602.78

Fonte: A autora (2023).

Figura 53 – Ilustração dos dados após a aplicação da UMAP, técnica de aprendizado múltiplo para redução de dimensão para (a) 2D da base de dados formado no Experimento II e (b) 3D da base de dados formado no Experimento II



(a)



(b)

Fonte: A autora (2023).

Para entender melhor essa variação, uma simples avaliação do impacto do pré-processamento é visualizar as projeções UMAP das leituras dos sensores do *e-nose* antes e depois do pré-processamento para identificar mudanças na separação entre classes, que já foram apresentadas na Figura 37 e 53.

Sabe-se que a abordagem do classificador Rocket é baseada em características, enquanto o TapNet utiliza camadas de convolução na arquitetura da rede neural para aprender automaticamente as representações das séries temporais. Dessa maneira, a redução de dimensão com o UMAP afetou adversamente a capacidade do Rocket de extrair características relevantes, enquanto que para o TapNet melhorou a qualidade das representações das séries temporais, tornando-as mais informativas e separáveis. E da mesma forma que nos experimentos anteriores, o conjunto de testes foi empregado para avaliar o desempenho dos classificadores. Os resultados correspondentes a cada classificador neste conjunto estão apresentados na Tabela 18.

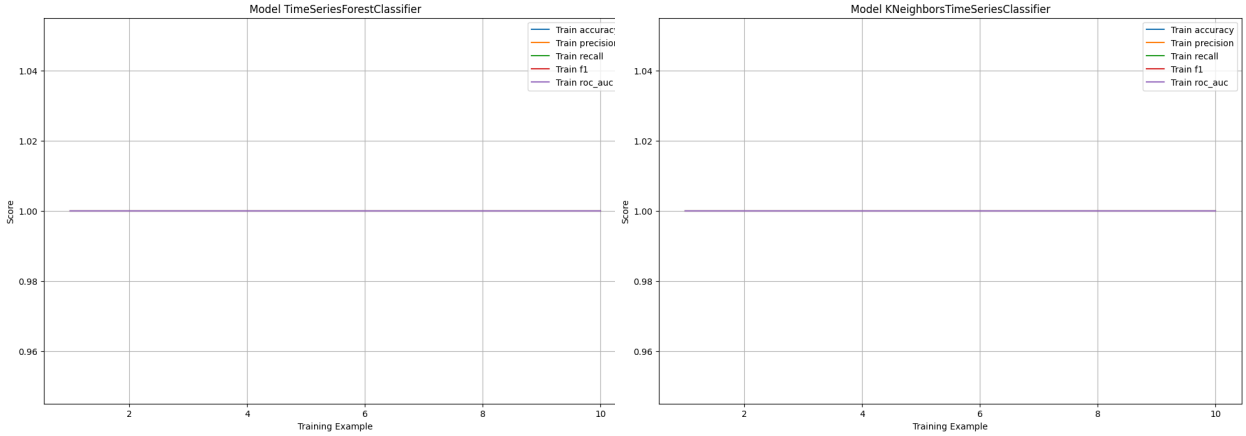
É importante ressaltar que, em certos casos, a redução significativa no tempo de processamento pode acarretar em uma possível degradação do desempenho do modelo. Portanto, a busca por um equilíbrio entre eficiência computacional e capacidade de previsão é crucial. Neste contexto, é notável que houve uma redução nos tempos de processamento para os classificadores KNeighbors TimeSeries, TimeSeries Forest e HIVE-COTE V2. No entanto, é preciso chamar a atenção para o fato de que o classificador TapNet apresentou o tempo de processamento mais elevado, atingindo 1602,78 segundos, equivalente a quase 26 minutos. Esse resultado suscita preocupações em termos de aplicabilidade prática e operacional. Importante mencionar que os classificadores foram executados no ambiente do Google Colab em um sistema com especificações de hardware que incluem um processador Intel(R) Xeon(R) CPU 2.20GHz e 12GB de RAM. É relevante destacar que modelos de aprendizado de máquina que envolvem cálculos intensivos, como redes neurais profundas ou operações complexas em grandes conjuntos de dados, tendem a demandar recursos de hardware mais substanciais, justificando, em parte, os tempos de processamento mais prolongados.

Tabela 18 – Resultados Experimento III no conjunto de dados de teste: conjunto de dados final passou pela etapa de suavização e normalização de dados.

Classificador	**Acurácia**	**Precisão**	**Especificidade**	**F1-Score**
TimeSeries Forest	**100%**	**100%**	**100%**	**100%**
KNeighbors TimeSeries	**100%**	**100%**	**100%**	**100%**
Rocket	56.2%	50.9%	31.08%	32.11%
HIVE-COTE V2	**100%**	**100%**	**100%**	**100%**
TapNet	98%	97%	96%	98%

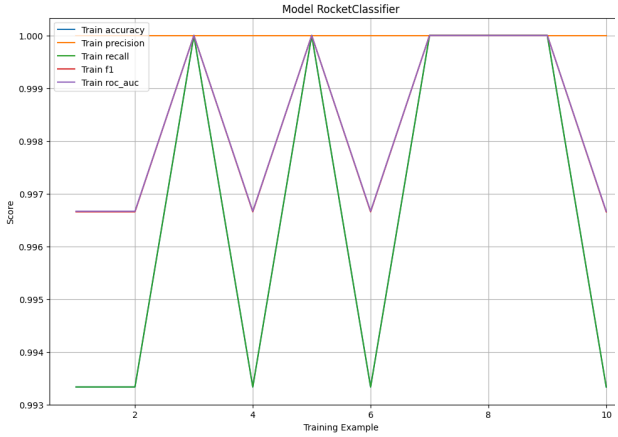
Fonte: A autora (2023).

Figura 54 – Experimento III - Curva de aprendizagem das 10 repetições da validação cruzada k-Fold das métricas acurácia, sensibilidade, especificidade, precisão e recall dos 5 modelos de IA para o conjunto de treinamento. (a) Classificador KNeighborsTime, (b) Classificador Rocket, (c) Classificador TimesSeriesForest, (d) Classificador HIVE-COTE V2, (e) Classificador TapNeT

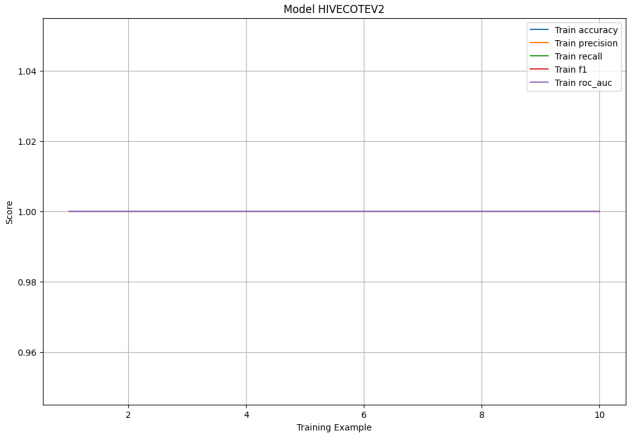


(a)

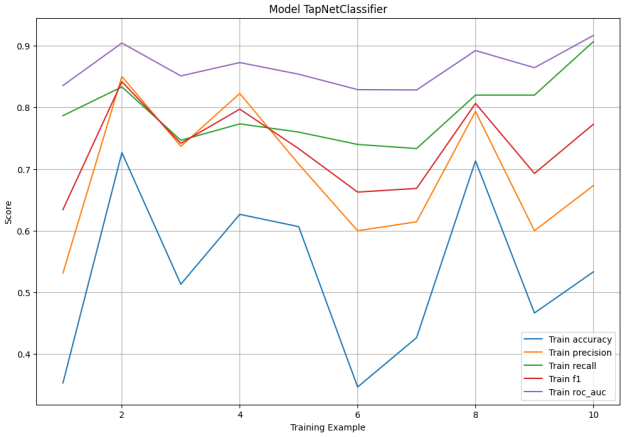
(b)



(c)



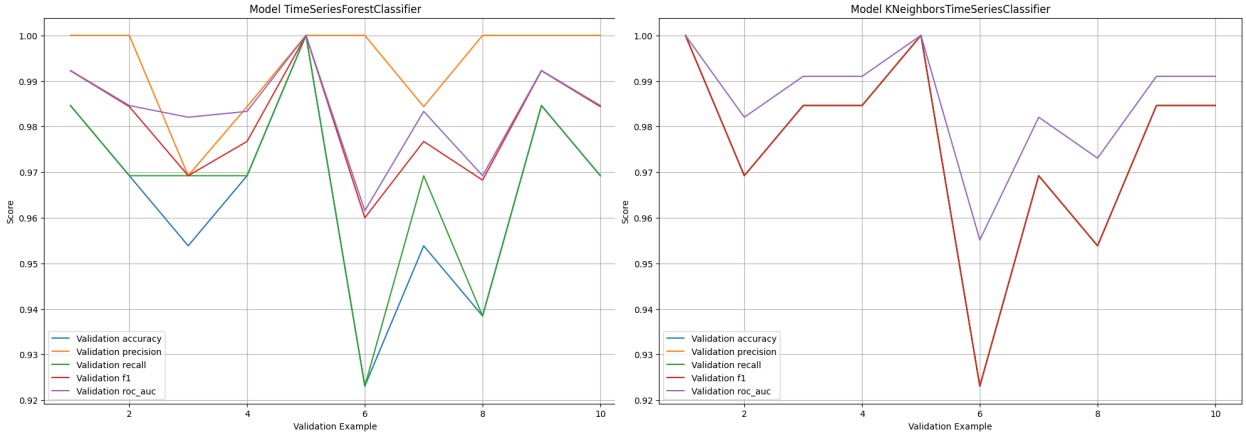
(d)



(e)

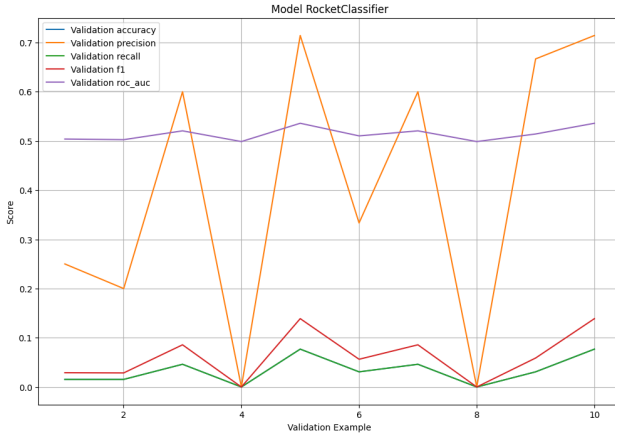
Fonte: A autora (2023).

Figura 55 – Experimento III - Curva de aprendizagem das 10 repetições da validação cruzada k-Fold das métricas acurácia, sensibilidade, especificidade, precisão e recall dos 5 modelos de IA para o conjunto de validação. (a) Classificador KNeighborsTime, (b) Classificador Rocket, (c) Classificador TimesSeriesForest, (d) Classificador HIVE-COTE V2, (e) Classificador TapNeT

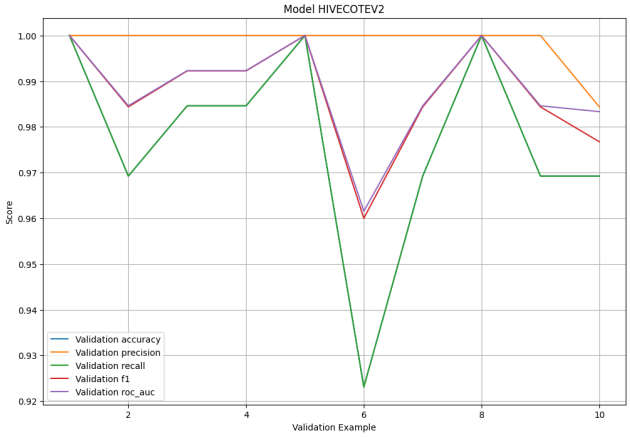


(a)

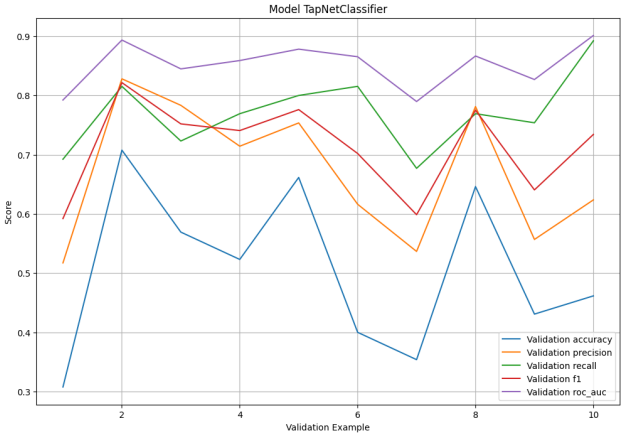
(b)



(c)



(d)



(e)

Fonte: A autora (2023).

5.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste estudo foram realizados três experimentos. No primeiro, o conjunto de dados formado passou pela etapa de suavização e normalização de dados. No segundo, um seletor de recursos que remove todos os recursos de baixa variância do conjunto de dados foi utilizado. E no terceiro, aplicou-se o UMAP, um conjunto de técnicas de aproximação, redução de dimensão e projeção (visualização). Apesar do grande conjunto de dados formado, em todos os experimentos, foram observados resultados satisfatórios, o qual foi possível encontrar classificadores que obtiveram como métrica de acurácia, sensibilidade e especificidade o valor de 100%, com diferentes tempo de processamento. Assim, infere-se que equipamento desenvolvido mostrou-se promissor, demonstrando ter sensibilidade para a detecção e identificação de leveduras do gênero *Candida* através dos COVs emitidos.

6 CONCLUSÃO

Neste estudo, explorou-se a aplicação da análise de compostos orgânicos voláteis (COVs), emitidos por meios biológicos, como uma ferramenta diagnóstica não invasiva, com foco nas leveduras do gênero *Candida*. Ancorados na Inteligência Artificial, buscou-se contribuir para esse processo, explorando o potencial dos sensores eletrônicos (*e-nose*) na rápida aquisição de assinaturas de COVs.

No decorrer deste projeto, foi investigado o desempenho de cinco classificadores em um cenário de classificação de dados de séries temporais. Cada classificador apresentou suas próprias vantagens e desafios, fornecendo *insights* valiosos para esta pesquisa. O Classificador TimeSeries Forest se destacou, obtendo consistentemente 100% de acurácia, sensibilidade e especificidade nos experimentos realizados. Sua habilidade em lidar com séries temporais complexas sugere que é um modelo promissor para problemas semelhantes no futuro. O KNeighbors TimeSeries também apresentou um bom desempenho, especialmente no Experimento III, destacando a importância da eficiência computacional em problemas de classificação de séries temporais. No entanto, o classificador Rocket não conseguiu atingir o mesmo nível de sucesso, evidenciando limitações na sua capacidade de lidar com as características específicas das séries temporais.

Surpreendentemente, o HIVE-COTE V2, um modelo de alto desempenho em experimentos anteriores, não conseguiu replicar seus resultados excepcionais neste estudo. Isso leva a inferir que as peculiaridades inerentes aos dados deste estudo desempenharam um papel crucial na escolha do modelo apropriado. Por fim, o desempenho aprimorado do classificador TapNet, devido à técnica UMAP para redução de dimensão, sugere que estratégias de pré-processamento podem ser fundamentais para melhorar a eficácia dos modelos.

Em resumo, a escolha do classificador adequado depende da natureza dos dados e do contexto do problema. Cada um dos modelos analisados contribuiu com informações valiosas sobre o desempenho, a eficiência computacional e a escalabilidade. A utilização desses modelos nos experimentos deste projeto aprofundou a compreensão da aplicação de classificadores de COVs, ou melhor, de séries temporais complexas.

No entanto, é crucial reconhecer as limitações inerentes a esta pesquisa. Uma das principais limitações reside na disponibilidade de dados e recursos. O estudo dependeu de um conjunto de dados específico e da capacidade dos sensores eletrônicos em uso, o que pode não refletir completamente a complexidade e a variabilidade do ambiente clínico real. Além disso, a interpretabilidade dos modelos propostos ainda é um desafio, uma vez que compreender os motivos subjacentes às decisões dos algoritmos de aprendizado de máquina é crucial para a aceitação clínica. Outra limitação se relaciona à validação clínica, que requer testes rigorosos e a adaptação do modelo a uma variedade de cenários

clínicos.

Portanto, este estudo oferece uma base para futuras pesquisas, acredita-se que o objetivo desta dissertação foi plenamente resolvido. A abordagem proposta neste trabalho demonstrou ser promissora e oferece um caminho válido para futuras investigações no campo da detecção e identificação de microrganismos, demonstrando ter sensibilidade para a detecção e identificação de leveduras do gênero *Candida*. Isso sugere um potencial de contribuir significativamente para a prática clínica.

6.1 TRABALHOS FUTUROS

Alternativas de trabalhos futuros foram identificadas, com o propósito de aprimorar o dispositivo e os resultados obtidos neste trabalho, nas quais se destacam:

1. Ampliar e diversificar a testagem de fungos *Candida*, podendo cobrir uma gama maior de problemas e patologias associadas a mais espécies, como também a ampliação das bases de dados. Isto é, mais e melhores dados para minimizar os erros implicam em modelos mais bem treinados, que generalizam adequadamente para casos novos, que ainda não são conhecidos pelos modelos

2. Projetar e utilizar um ambiente metrologicamente controlável para armazenamento das placas e realização dos testes, ou seja, a criação de um ambiente controlado a fim de garantir que as placas de *Petri* sejam afetadas pelos mesmos fatores externos (temperatura, umidade, luz, entre outros), e assim, minimizar erros causados por estes fatores;

3. Aprofundar o estudo dos modelos de classificadores especializados em séries temporais. O objetivo é encontrar modelos com resultados mais satisfatórios e com menores custos computacionais;

4. Realizar testes de robustez abrangentes nos modelos analisados, submetendo-os a uma variedade de condições e cenários clínicos simulados. Esses testes ajudarão a avaliar o desempenho do sistema em situações do mundo real e garantir as suas confiabilidades;

5. Estudar o impacto de partes das séries temporais lidas pelo nariz eletrônico nos classificadores. Isto é, testar modelos e possibilidades de classificação a partir de diferentes momentos da série temporal de um ciclo de leitura. Exemplos seriam as partes vistas na Figura 30, da Seção 4.3.4.

REFERÊNCIAS

AHA D. W., K. D. e. A. M. K. *Instance-based learning algorithms. Machine Learning.* 1991. 37–66 p.

ALESSIO, E.; CARBONE, A.; CASTELLI, G.; FRAPPIETRO, V. Second-order moving average and scaling of stochastic time series. *Physics of Condensed Matter*, v. 27, p. 197–200, 01 2002.

ALLAW, F.; ZAHREDDINE, N. K.; IBRAHIM, A.; TANNOUS, J.; TALEB, H.; BIZRI, A. R.; DBAIBO, G.; KANJ, S. S. *First* Candida auris *Outbreak during a COVID-19 Pandemic in a Tertiary-Care Center in Lebanon.* 2021. Disponível em: <https://www.mdpi.com/2076-0817/10/2/157>.

ANDREAS, H. The chemistry and biology of volatiles, wiley: Chichester. *John Wiley Sons Ltd*, 2010.

ANVISA. Detecção e identificação dos fungos de importância médica. *Agência Nacional de Vigilância Sanitária,Brasil*, VII, 01 2004.

ANVISA. Microbiologia clínica para o controle de infecção relacionada à assistência à saúde. módulo 3 : Principais síndromes infecciosas/agência nacional de vigilância sanitária. *Agência Nacional de Vigilância Sanitária,Brasil*, v. 9, 2013.

ARASTEHFAR, A.; CARVALHO, A.; NGUYEN, M. H.; HEDAYATI, M. T.; NETEA, M. G.; PERLIN, D. S.; HOENIGL, M. *COVID-19-Associated Candidiasis (CAC): An Underestimated Complication in the Absence of Immunological Predispositions?* 2020. Disponível em: <https://www.mdpi.com/2309-608X/6/4/211>.

ARAUJO, I.; GAMBOA, J. R.; SILVA, A. da. Modelos de deep learning para classificação de gases detectados por matrizes de sensores nariz artificial. p. 844–855, 01 2020.

BAGNALL A., L. J. B. A. e. a. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, v. 31, p. 606–660, 2017. Disponível em: <https://doi.org/10.1007/s10618-016-0483-9>.

BARBEDO L. S.; SGARBI, D. B. G. Candidíase. *DST - J bras Doenças Sex Transm*, v. 22, p. 22–38, 2010.

BASTVIKEN, D.; NYGREN, J.; SCHENK, J.; MASSANA, R. P.; DUC, N. T. Technical note: Facilitating the use of low-cost methane (CH4) sensors in flux chambers – calibration, data processing, and an open-source make-it-yourself logger. *Biogeosciences*, v. 17, n. 13, p. 3659–3667, 2020. Disponível em: <https://bg.copernicus.org/articles/17/3659/2020/>.

BETTAUER, V.; COSTA, A.; OMRAN, R.; MASSAHI, S.; KIRBIZAKIS, E.; SIMPSON, S.; DUMEAUX, V.; LAW, C.; WHITEWAY, M.; HALLETT, M. A deep learning approach to capture the essence of *Candida albicans* morphologies. *bioRxiv*, Cold Spring Harbor Laboratory, 2021. Disponível em: <https://www.biorxiv.org/content/early/2021/06/10/2021.06.10.445299>.

BEZERRA, K. K. S. Leveduras vaginais e ação antifúngica do extrato de própolis vermelha. *Programa de Pós-graduação em Sistemas Agroindustriais, Centro de Ciências e Tecnologia Agroalimentar, Universidade Federal de Campina Grande – Pombal – Paraíba – Brasil*, 2015.

BORGES, R. M. e. a. *Fatores de risco associados à colonização por Candida spp em neonatos internados em uma unidade de terapia intensiva neonatal brasileira.* 2009. 431-435 p.

CAMILO CáSSIO OLIVEIRA; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Instituto de Informática Universidade Federal de Goiás*, 2009.

CASADO LARISSA DA SILVA ; GOMES, R. D. S. *ASPECTOS CLÍNICOS DA CERATITE FÚNGICA: Uma revisão de literatura.* 2022.

CHERKAOUI, A.; HIBBS, J.; EMONET, S.; TANGOMO, M.; GIRARD, M.; FRANCOIS, P.; SCHRENZEL, J. *Comparison of Two Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry Methods with Conventional Phenotypic Identification for Routine Identification of Bacteria to the Species Level.* 2010. 1169-1175 p. Disponível em: <https://journals.asm.org/doi/abs/10.1128/JCM.01881-09>.

COCHRAN WILLIAM W. COOLEY, J. . L. F. D. . D. H. H. . A. K. R. . W. L. W. . M. J. G. . E. N. D. . M. R. C. . D. W. P. T. *What is the fast Fourier transform?* 1967. 1664 - 1674 p.

COLLINS C. H.; BRAGA, G. L. B. P. S. Introdução a métodos cromatográficos. *Editora da UNICAMP*, p. 143–144, 1997.

COSTA C. P., B. A. R. A. A. . R. S. M. *Candida Species (Volatile) Metabotyping through Advanced Comprehensive Two-Dimensional Gas Chromatography.* [S.l.]: Microorganisms, 2020.

CREPALDI, C. da Rosa e Viviane Curi e Andrei Rosa e Antonio Carlos Filho e Cyra Maria Bianchi e Tahyná Deps e Marcus Crepaldi e Maria de Lourdes Crepaldi e M. CandidÍase bucal. *REVISTA FAIPE*, v. 11, n. 1, p. 155–163, 2021. ISSN 2179-9660. Disponível em: <https://www.revistafaipe.com.br/index.php/RFAIPE/article/view/239>.

Cromvallab. *Análises cromatográficas.* 2022. Disponível em: <https://cromvallab.com/2022/01/10/quais-as-vantagens-de-utilizar-o-detector-de-espectrometria-de-massas-acoplado-a-cromatografia-liquida-ou-gasosa/>.

DEMPSTER, A.; PETITJEAN, F.; WEBB, G. I. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, Springer Science and Business Media LLC, v. 34, n. 5, p. 1454–1495, jul 2020. Disponível em: <https://doi.org/10.1007%2Fs10618-020-00701-z>.

DROUIN, A. *Introduction to Machine Learning.* 2017. Disponível em: <https://aldro61.github.io/microbiome-summer-school-2017/sections/basics/>.

FARRAIA, M. V.; RUFO, J. C.; PACIêNCIA, I.; MENDES, F.; DELGADO, L.; MOREIRA, A. The electronic nose technology in clinical diagnosis: A systematic review. *Porto biomedical journal*, v. 4, n. 4, p. e42, 2019. ISSN 2444-8664. Disponível em: <https://europepmc.org/articles/PMC6924976>.

FUCHS, S.; LASS-FLöRL, C.; POSCH, W. Diagnostic performance of a novel multiplex pcr assay for candidemia among icu patients. *Journal of Fungi*, v. 5, n. 3, 2019. ISSN 2309-608X. Disponível em: <https://www.mdpi.com/2309-608X/5/3/86>.

GAMBOA, J. C. B. *Deep Learning for Time-Series Analysis*. 2017.

GAMBOA, J. R.; ALBARRACIN-ESTRADA, E.; DELGADO-TREJOS, E. Quality control through electronic nose system. In: _____. [S.l.: s.n.], 2011. ISBN 978-953-307-971-4.

GARCIA, A. C. H. Métodos de identificação molecular de doenças infectocontagiosas. 2018. trabalho de conclusão de curso (graduação em biomedicina. *Faculdade de Ciências da Educação e Saúde, Centro Universitário de Brasília*, 2018.

GIOLO MURIEL PADOVANI; SVIDZINSKI, T. I. E. Fisiopatogenia, epidemiologia e diagnóstico laboratorial da candidemia. *J. Bras. Patol. Med. Lab.*, 2010.

GOODFELLOW, I. Nips 2016 tutorial: Generative adversarial networks. arXiv, 2016. Disponível em: <https://arxiv.org/abs/1701.00160>.

HAN, L.; YU, C.; XIAO, K.; ZHAO, X. A new method of mixed gas identification based on a convolutional neural network for time series classification. *Sensors*, v. 19, p. 1960, 04 2019.

HE, K.; ZHANG, X.; REN, S.; SUN, J. *Deep Residual Learning for Image Recognition*. 2015.

HERTEL, M.; HARTWIG, S.; SCHüTTE, E.; GILLISSEN, B.; PREISSNER, R.; SCHMIDT-WESTHAUSEN, A.; PARIS, S.; KASTNER, I.; PREISSNER, S. Identification of signature volatiles to discriminate *Candida albicans*, glabrata, krusei and tropicalis using gas chromatography and mass spectrometry. *Mycoses*, v. 59, 12 2015.

HOSSIN M ; SULAIMAN, M. *International Journal of Data Mining Knowledge Management Process,*. 2015.

J HUNG GC, N. K. L. B. T. S. L. S. Z. Development of candida-specific real-time pcr assays for the detection and identification of eight medically important candida species. *Microbiol Insights*, v. 9, p. 21–28, 2016.

JODDS, F. C. B. R. Chromagar candida, a new differential isolation medium for presumptive identification of clinically important candida species. *Journal of Clinical Microbiology*, n. 8, 1994.

JR JOãO NOBREGA DE ALMEIDA; SILVIA V, C. D. Y. T. L. T. d. A. R. K. G. a. a. *Emerging microbes infection*. 2018.

J.W., B. P. G. A brief history of electronic noses. *Sens. Actuators B Chem*, 1994.

KOC, ; KESSLER, H. H.; HOENIGL, M.; WAGENER, J.; SUERBAUM, S.; SCHUBERT, S.; DICHTL, K. Performance of multiplex pcr and beta;-1,3-d-glucan testing for the diagnosis of candidemia. *Journal of Fungi*, v. 8, n. 9, 2022. ISSN 2309-608X. Disponível em: <https://www.mdpi.com/2309-608X/8/9/972>.

KOEHLER, A.; DALLEMOLE, D. R.; RIçA, L. B.; CORBELLINI, V. A.; RIEGER, A. Identificação de três espécies de candida por pcr em tempo real. *Revista Jovens Pesquisadores*, v. 6, n. 1, jun. 2016. Disponível em: <https://online.unisc.br/seer/index.php/jovenspesquisadores/article/view/7341>.

KORPI, A.; JäRNBERG, J.; PASANEN, A.-L. Microbial volatile organic compounds. *Critical Reviews in Toxicology*, Taylor Francis, v. 39, n. 2, p. 139–193, 2009. PMID: 19204852. Disponível em: <https://doi.org/10.1080/10408440802291497>.

LE, C.; XU, Y.-C.; BAI, F.-Y. *Candida pseudorugosa* sp. nov., a novel yeast species from sputum. *Journal of clinical microbiology*, v. 44, p. 4486–90, 01 2007.

LI, Y. e. a. Nosocomial bloodstream infection due to candida spp. in china: Species distribution, clinical features, and outcomes. *Mycopathologia*, 2016.

LINES, J.; TAYLOR, S.; BAGNALL, A. Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification. p. 1041–1046, 2016.

LUIZ, J.; AMARAL, G.; FERREIRA, A.; FEREZ, D.; GERETTO, P. Monitorização da respiração: Oximetria e capnografia. *Rev Bras Anestesiol*, v. 42, p. 52, 1992.

MADDALA, G. S. e. K. L. Introduction to econometrics. *John Wiley Sons Ltd*, 2009.

MARCELLIS A.; FERRI, G. D. Physical and chemical sensors" and "resistive, capacitive and temperature sensor interfacing overview" in analog circuits and systems for voltage-mode and current-mode sensor interfacing applications". *Netherlands: Springer*, p. 1–71, 2011.

MARQUES, E. C. M. *Redução de características baseadas em grupos semânticos aplicados à classificação de textos*. 2018.

MIDDLEHURST M., L. J. F. M. e. a. Hive-cote 2.0: a new meta ensemble for time series classification. 2021.

MONTEIRO, O. S. Caracterização do óleo essencial da pimenta dioica lindl. e sua aplicação como atrativo de abelhas euglossina. *Tese (Doutorado em Química). Universidade Federal da Paraíba, João Pessoa*, p. 148, 2008.

MORAES G. K. A.; FERRAZ, L. F. C. V. M. Compostos orgânicos voláteis de fungos endofíticos e suas aplicações biotecnológicas. *Revista Virtual Quimíca*, v. 12, n. 6, p. 1498–1510, 2020.

MORETTIN P. A.; TOLOI, C. M. Séries temporais. *São Paulo: Atual*, 2006.

MOTA, I.; TEIXEIRA-SANTOS, R.; RUFO, J. C. Detection and identification of fungal species by electronic nose technology: A systematic review. *Fungal Biology Reviews*, v. 37, 04 2021.

MOURA, J. "what is signal processing?, president's message". *IEEE Signal Processing Magazine*, v. 26, 2009.

MULET, J. V.; GARCíA, C.; TORMO, N.; GIMENO, C. Evaluation of a novel chromogenic medium for candida spp. identification and comparison with chromagar™ candida for the detection of *Candida auris* in surveillance samples. *Diagnostic Microbiology and Infectious Disease*, v. 98, p. 115168, 08 2020.

MULLIS, K. Target amplification for dna analysis by the polymerase chain reaction. *Annales de Biologie Clinique*, v. 48, n. 8, p. 579–582, 1990.

NASCIMENTO, L. J. D. *ESTRATÉGIAS DE IDENTIFICAÇÃO INTEGRADA DE FUNGOS MEDIANTE MÉTODOS MICROBIOLÓGICOS E NÃO-MICROBIOLÓGICOS*. 2016.

NASCIMENTO RONALDO FERREIRA DO;LIMA, A. C. A. d. P. G. A. V. P. A. d. *Cromatografia gasosa: aspectos teóricos e práticos*. 2018. 334 p. Disponível em: <http://www.repositorio.ufc.br/handle/riufc/39260>.

OLIVEIRA, F. K. F. d. Síntese e caracterização de molibdato de cério pelo método hidrotérmico assistido por micro-ondas: aplicação como sensor de gás ozônio. *Universidade Federal do Rio Grande do Norte. Dissertação (Mestrado em Ciência e Engenharia de Materiais*, 2021. Disponível em: <https://repositorio.ufrn.br/handle/123456789/45585>.

PAPPAS P. G., L. M. S. A. M. C. O.-Z. L. . K. B. J. Invasive candidiasis. *Nature Reviews Disease Primers*, v. 4, 2018.

PEIXOTO J.A; ROCHA1, M. G. N. R. V. V. K.-T. G. B. *CANDIDÍASE BUCAL REVISÃO DA LITERATURA*. 2014. 75-82 p.

PENG P.; ZHAO, X. P. X. Y. W. Gas classification using deep convolutional neural networks. *Sensors*, 2018.

PEREIRA, A. P. V. Identificação molecular de candidoses invasivas no centro hospitalar da cova da beira. *Instituto de Higiene e Medicina Tropical*, 2010.

PETRUSEVSKI, A. *History of infectious diseases development in the Old and the Middle Ages with the emphasis on the plague and leprosy*. 2013. 704-8 p.

PIGATTO A; LOVISON, O. V. A. C. F. *Prevalência de infecções fúngicas em um laboratório de análises clínicas da cidade de Veranópolis, Rio Grande do Sul*. 2019. 202-207 p.

PYLE, D. *Data Preparation for Data Mining*. [S.l.]: Morgan Kaufmann, 1999. ISBN 1-55860-529-0.

RASCHKA S.; MIRJALILI, V. *Python machine learning*. 2017.

REES, C. A.; BURKLUND, A.; STEFANUTO, P.-H.; SCHWARTZMAN, J. D.; HILL, J. E. Comprehensive volatile metabolic fingerprinting of bacterial and fungal pathogen groups. *Journal of Breath Research*, IOP Publishing, v. 12, n. 2, p. 026001, jan 2018. Disponível em: <https://doi.org/10.1088/1752-7163/aa8f7f>.

RODRIGUES, T. J. S. Cepas do complexo *Candida parapsilosis* de origem animal : classificação taxonômica, sensibilidade antifúngica e atributos de virulência in vitro. *Dissertação (Mestrado em Microbiliologia Médica) - Universidade Federal do Ceará. Faculdade de Medicina, Fortaleza*, 2013.

ROSSETTO, L. Dispositivo para análise e caracterização de materiais semicondutores utilizados como sensores de gás. *Universidade Federal do Rio Grande do Sul, Instituto de Física, Programa de Pós-Graduação em Ciência dos Materiais*, 2021.

RUIZ ALEJANDRO PASOS; FLYNN, M. L. J. M. M. B.-A. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, v. 35, p. 401–449, 2021. Disponível em: <https://doi.org/10.1007/s10618-020-00727-3>.

SANTANA, M. *Deep Learning: do Conceito às Aplicações*. 2018. Disponível em: <https://medium.com/data-hackers/deep-learning-do-conceito-%C3%A0s-aplica%C3%A7%C3%B5es-e8e91a7c7eaf>.

SIDRIM J. J. C.; ROCHA, M. F. G. Micologia médica à luz de autores contemporâneos. *Guanabara Koogan*, v. 04, p. 20–21, 2004.

SILVA, D. F. B. F. d. Pré-processamento de dados e comparação entre algoritmos de machine learning para a análise preditiva de falhas em linhas de produção para o controlo. *Mestrado em Engenharia Informática - Sistemas de Informação e Conhecimento. Instituto Superior de Engenharia de Porto*, 2021.

SIQUEIRA J. P. Z.; ALMEIDA, M. T. G. B. Biologia molecular como ferramenta de detecção fúngica no sangue: auxílio diagnóstico e redução de gastos. *Archives of Health Sciences- AHS*, v. 25, p. 41–45, 2018. Disponível em: <http://docs.bvsalud.org/biblioref/2019/12/1046416/artigo9.pdf>.

SOUZA, M. S. Aplicações da espectrometria de massas e da cromatografia líquida na caracterização estrutural de biomoléculas de baixa massa molecular. *Tese (Doutorado em Ciências – Bioquímica). Universidade Federal do Paraná, Curitiba*, p. 162, 2008.

SOUZA, C. d. IdentificaÇÃo e determinaÇÃo do perfil de sensibilidade de forma rÁpida de candida spp. isoladas de hemoculturas. *FACULDADE DE MEDICINA. UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL. PROGRAMA DE PÓS-GRADUAÇÃO EM MEDICINA: CIÊNCIAS MÉDICAS*, 2021. Disponível em: <https://lume.ufrgs.br/handle/10183/235029?locale-attribute=pt_BR>.

SRI, S. P. M. R.; KARUNA, T. Classification of fungi microscopic images – leveraging the use of ai. *Insights2Techinfo*, 2021. Disponível em: <https://insights2techinfo.com/classification-of-fungi-microscopic-images-leveraging-the-use-of-ai/>.

Sá, F. A. D. S. *COMPOSIÇÃO QUÍMICA E ATIVIDADE ANTI-CANDIDA DAS FOLHAS DE Myrcia tomentosa (Aubl.) DC. – MYRTACEAE*. 2017.

TOWARDS DATA SCIENCE. *K-Means Data Clustering*. 2017. Disponível em: <https://towardsdatascience.com/k-means-data-clustering-bce3335d2203>.

TROVÃO J.; PEREIRA, L. IntroduÇÃo ao estudo dos microfungos: Guia simples para a iniciação à identificação. 2019.

VALENTE, A.; LOPES, T.; REIS, M. ComparaÇÃo da sensibilidade e especificidade entre dois mÉtodos de identificaÇÃo de *Candida albicans*. v. 18, mar. 2021. Disponível em: <https://conhecer.org.br/ojs/index.php/biosfera/article/view/3823>.

VIEIRA, F. A. P. Estudo quÍmico e avaliaÇÃo da atividade microbiolÓgica do Óleo essencial das folhas de hyptis dilatata benth. *DEPARTAMENTO DE TECNOLOGIA QUÍMICA. CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA . UNIVERSIDADE FEDERAL DO MARANHÃO.*, 2018.

VILLA, Q. S. S. N. . E. B. F. . D. A. N. . A. A. G. . A. G. . T. E. de. Uso de modelagem univariada e multivariada com séries temporais como ferramenta de gestão do agronegócio na cultura de soja do brasil. *Revista Espacios*, 2019.

WANG, S.-H.; CHOU, T.-I.; CHIU, S.-W.; TANG, K.-T. Using a hybrid deep neural network for gas classification. *IEEE Sensors Journal*, v. 21, n. 5, p. 6401–6407, 2021.

WANG, Z.; OATES, T. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In: . [S.l.: s.n.], 2015.

XU M.; WATANACHATURAPORN, P. V. P. K. A. M. K. *Decision tree regression for soft classification of remote sensing data*. 2005. 322–336 p.

YAN XIUZHEN GUO, S. D. . P. J. L. W. C. P. J.; ZHANG, S. Electronic nose feature extraction methods: A review. *Sensors*, 2015.

ZAMPARETTE, C. P. *GUIA COMPLETO: Comparação de Metodologias de Identificação de Microrganismos*. 2017.

ZHAI, Y.; LIU, J.; ZHOU, L.; JI, T.; MENG, L.; GAO, Y.; LIU, R.; WANG, X. J.; LI, L.; LU, B.; CAO, Z. Detection of candida species in pregnant chinese women with a molecular beacon method. *Journal of Medical Microbiology*, v. 67, p. 783 – 789, 2018.

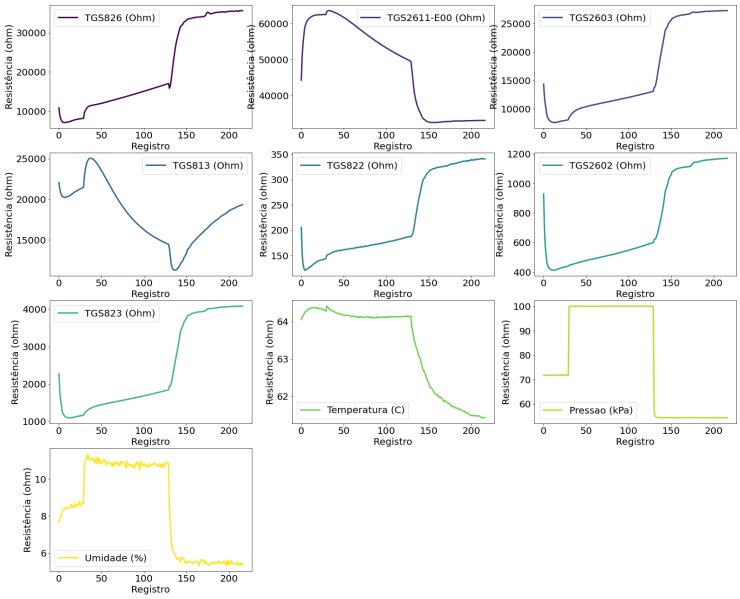
ZHANG YIFENG GAO, J. L. C.-T. L. X. Tapnet: Multivariate time series classification with attentional prototypical network. *Association for the Advancement of Artificial Intelligence*, 2020.

ZHENG, Y.; LIU, Q.; CHEN, E.; GE, Y.; ZHAO, J. Time series classification using multi-channels deep convolutional neural networks. *WAIM 2014. LNCS*, v. 8485, p. 298–310, 01 2014.

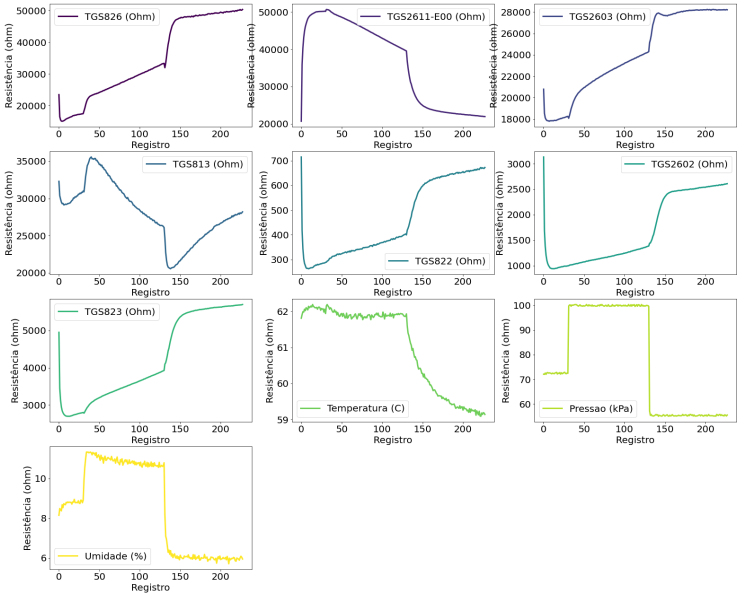
ZIELIńSKI, B.; SROKA-OLEKSIAK, A.; RYMARCZYK, D.; PIEKARCZYK, A.; BRZYCHCZY-WłOCH, M. Deep learning approach to describe and classify fungi microscopic images. *PLOS ONE*, Public Library of Science, v. 15, n. 6, p. 1–16, 06 2020. Disponível em: <https://doi.org/10.1371/journal.pone.0234806>.

APÊNDICE A – REGISTRO DA MODIFICAÇÃO DA RESISTÊNCIA DA MATRIZ DOS SENSORES QUANDO EXPOSTO A UM CONTEÚDO GASOSO PARA UM ÚNICO CICLO POR ESPÉCIE

Figura 56 – Registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso para único ciclo por espécie: (a) *Candida Glabrata*, (b) *C. haemulonii*, (c) *C. Kodamaea ohmeri*, (d) *C. Parapsilosis* e (e) *C. Tropicalis*



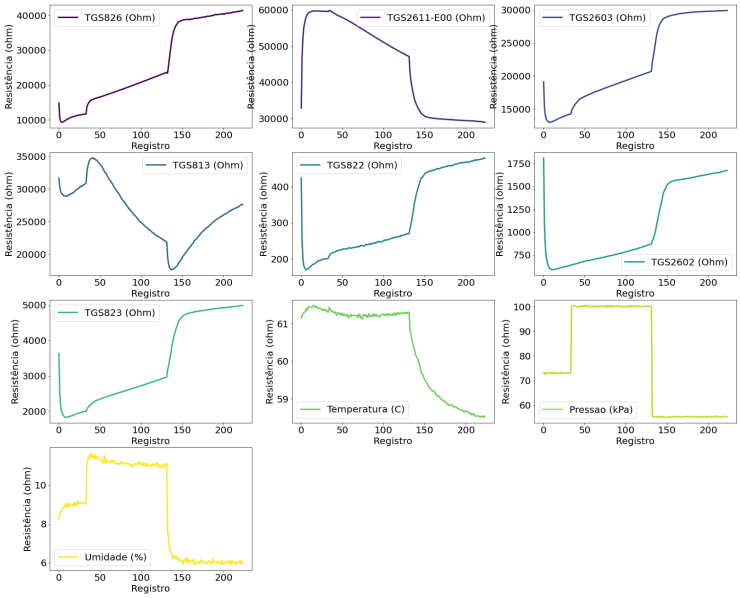
(a)



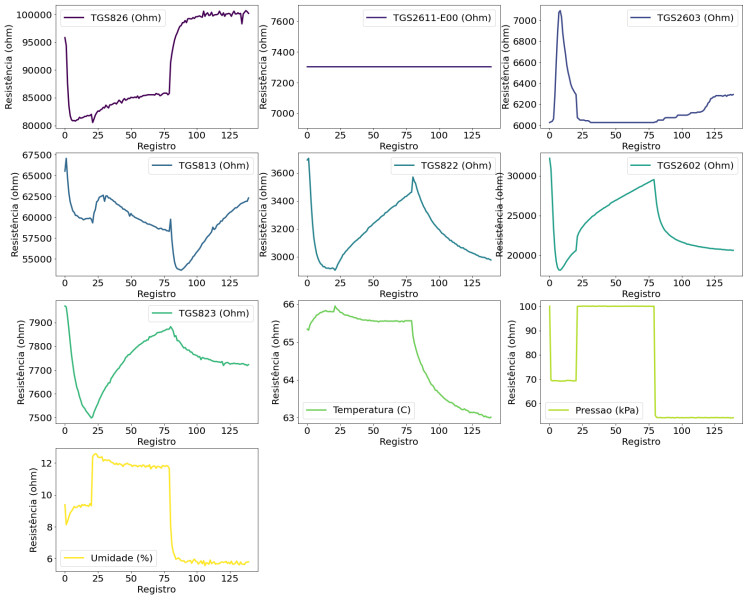
(b)

Fonte: A autora (2023).

Figura 56 – Registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso para único ciclo por espécie: (a)*Candida Glabrata*, (b) *C. haemulonii*, (c) *C. Kodamaea ohmeri*, (d) *C. Parapsilosis* e (e) *C. Tropicalis*



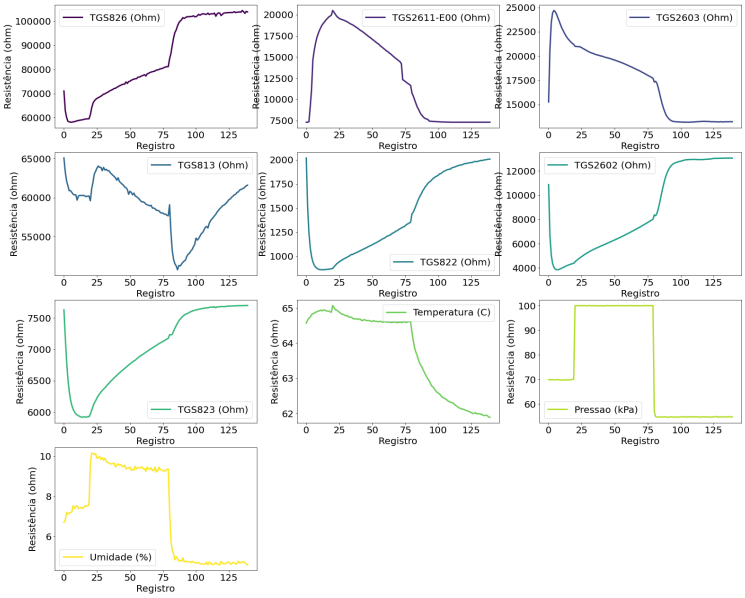
(c)



(d)

Fonte: A autora (2023).

Figura 56 – Registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso para único ciclo por espécie: (a) *Candida Glabrata*, (b) *C. haemulonii*, (c) *C. Kodamaea ohmeri*, (d) *C. Parapsilosis* e (e) *C. Tropicalis*

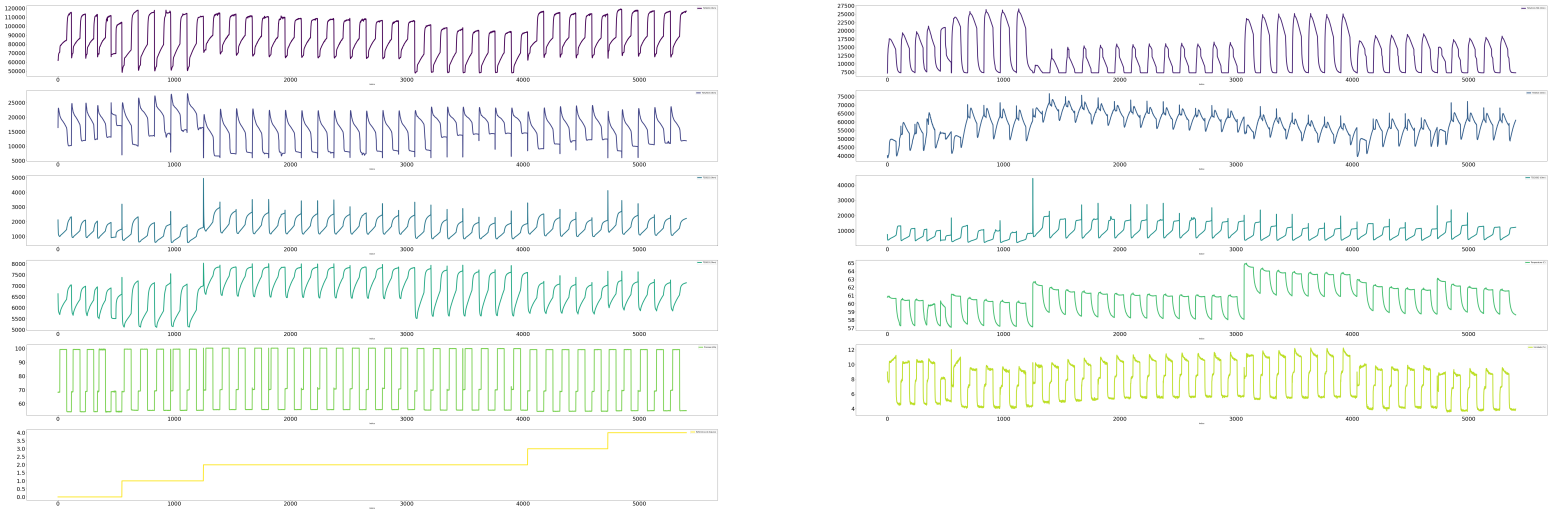


(e)

Fonte: A autora (2023).

APÊNDICE B – REGISTRO DA MODIFICAÇÃO DA RESISTÊNCIA DA MATRIZ DOS SENSORES QUANDO EXPOSTO A UM CONTEÚDO GASOSO PARA TODOS OS CICLO POR ESPÉCIE

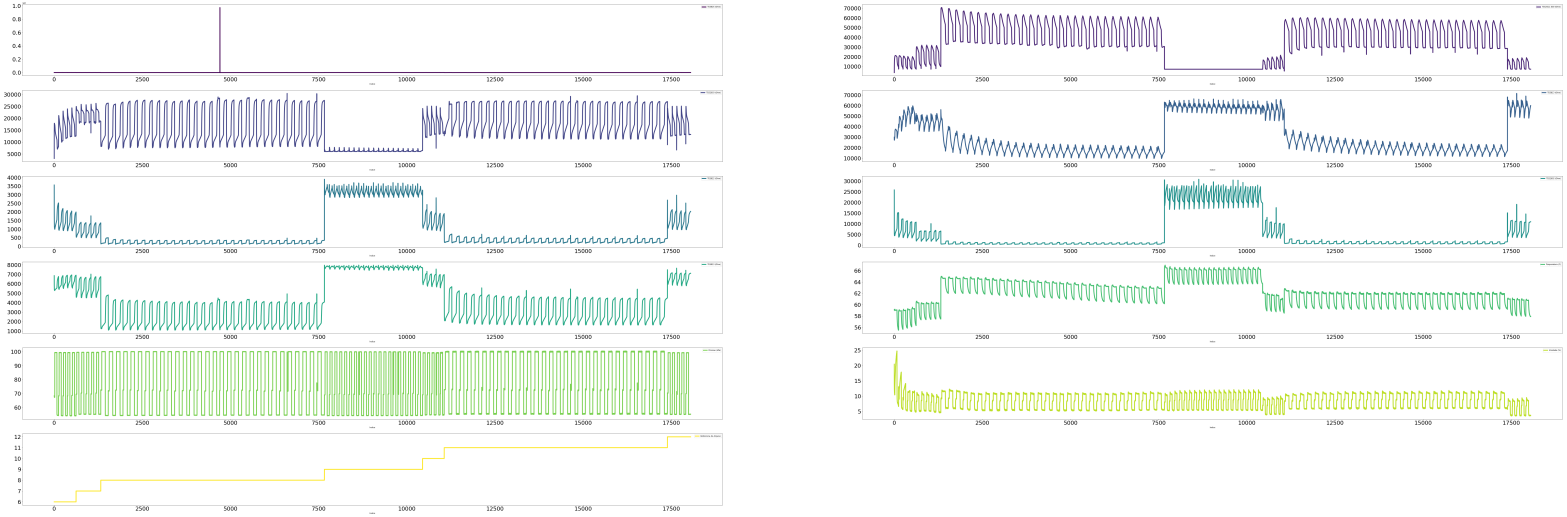
Figura 57 – Registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso para todos os ciclos por espécie: (a) *Candida Albicans*, (b) *C. Glabrata*, (c) *C. haemulonii*, (d) *C. Kodamaea ohmeri*, (e) *C. Krusei*, (f) *C. Parapsilosis* e (g) *C. Tropicalis*



(a)

Fonte: A autora (2023).

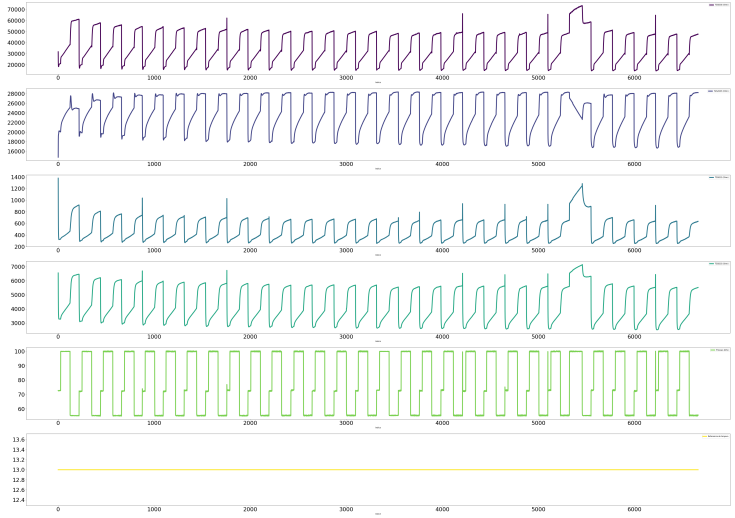
Figura 57 – Registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso para todos os ciclos por espécie: (a) *Candida Albicans*, (b) *C. Glabrata*, (c) *C. haemulonii*, (d) *C. Kodamaea ohmeri*, (e) *C. Krusei*, (f) *C. Parapsilosis* e (g) *C. Tropicalis*

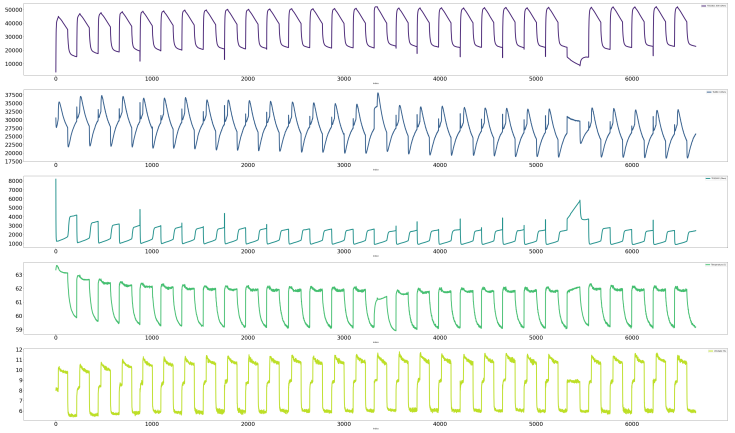


(b)

Fonte: A autora (2023).

Figura 57 – Registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso para todos os ciclos por espécie: (a) *Candida Albicans*, (b) *C. Glabrata*, (c) *C. haemulonii*, (d) *C. Kodamaea ohmeri*, (e) *C. Krusei*, (f) *C. Parapsilosis* e (g) *C. Tropicalis*

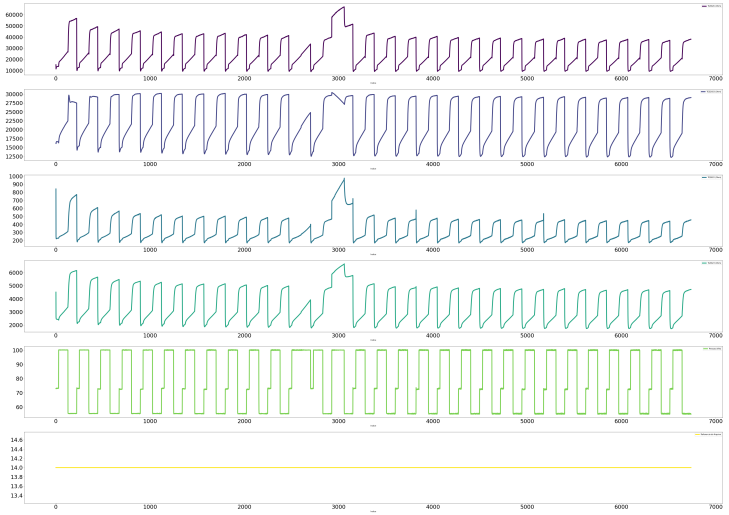


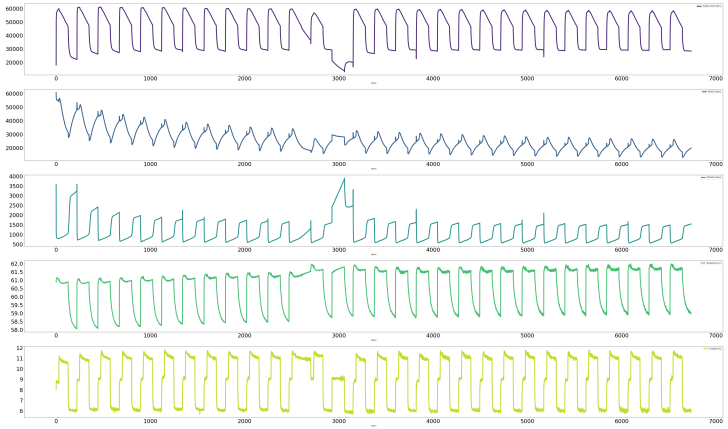


(c)

Fonte: A autora (2023).

Figura 57 – Registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso para todos os ciclos por espécie: (a) *Candida Albicans*, (b) *C. Glabrata*, (c) *C. haemulonii*, (d) *C. Kodamaea ohmeri*, (e) *C. Krusei*, (f) *C. Parapsilosis* e (g) *C. Tropicalis*

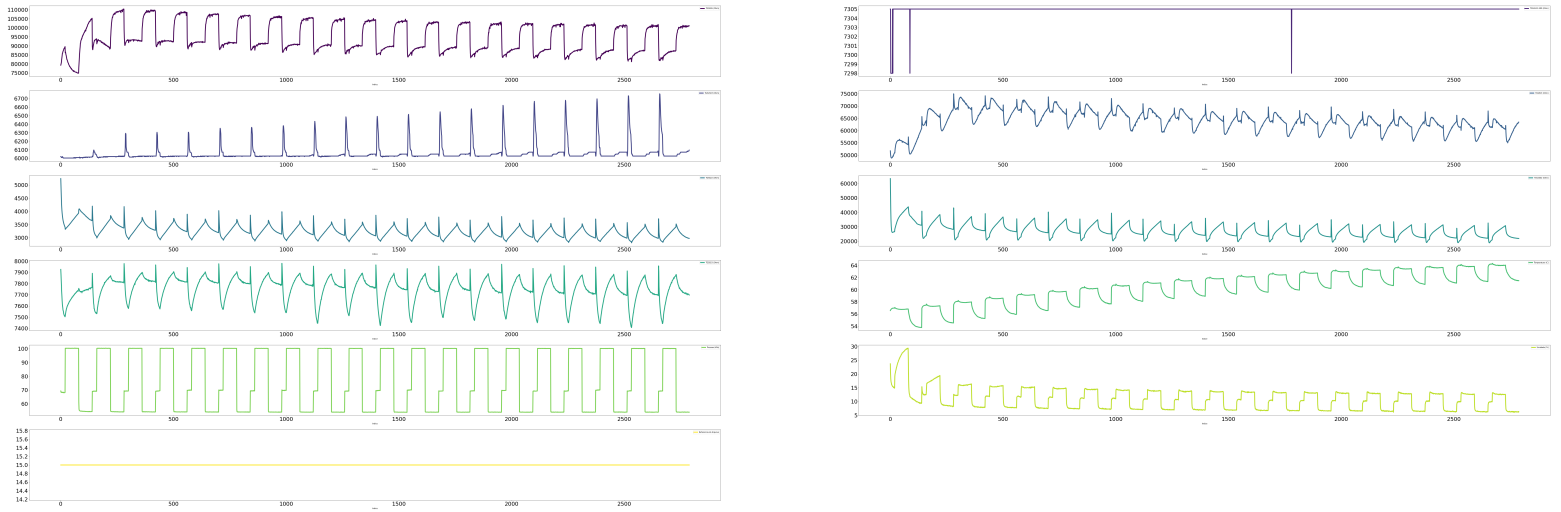




(d)

Fonte: A autora (2023).

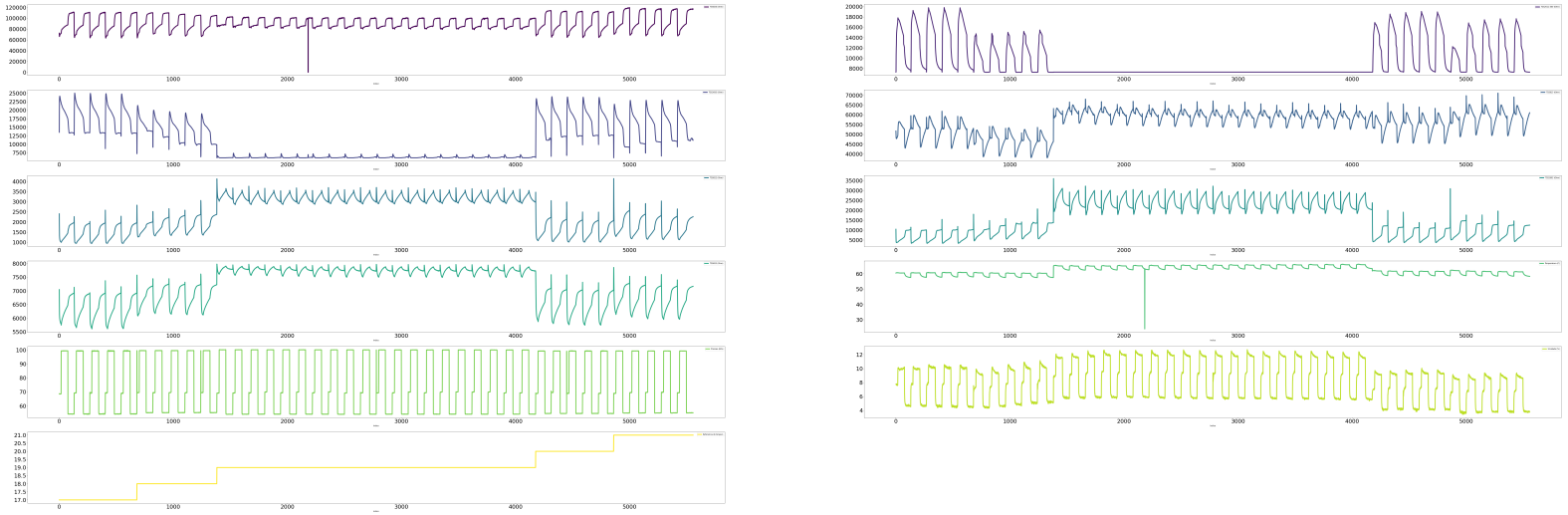
Figura 57 – Registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso para todos os ciclos por espécie: (a) *Candida Albicans*, (b) *C. Glabrata*, (c) *C. haemulonii*, (d) *C. Kodamaea ohmeri*, (e) *C. Krusei*, (f) *C. Parapsilosis* e (g) *C. Tropicalis*



(e)

Fonte: A autora (2023).

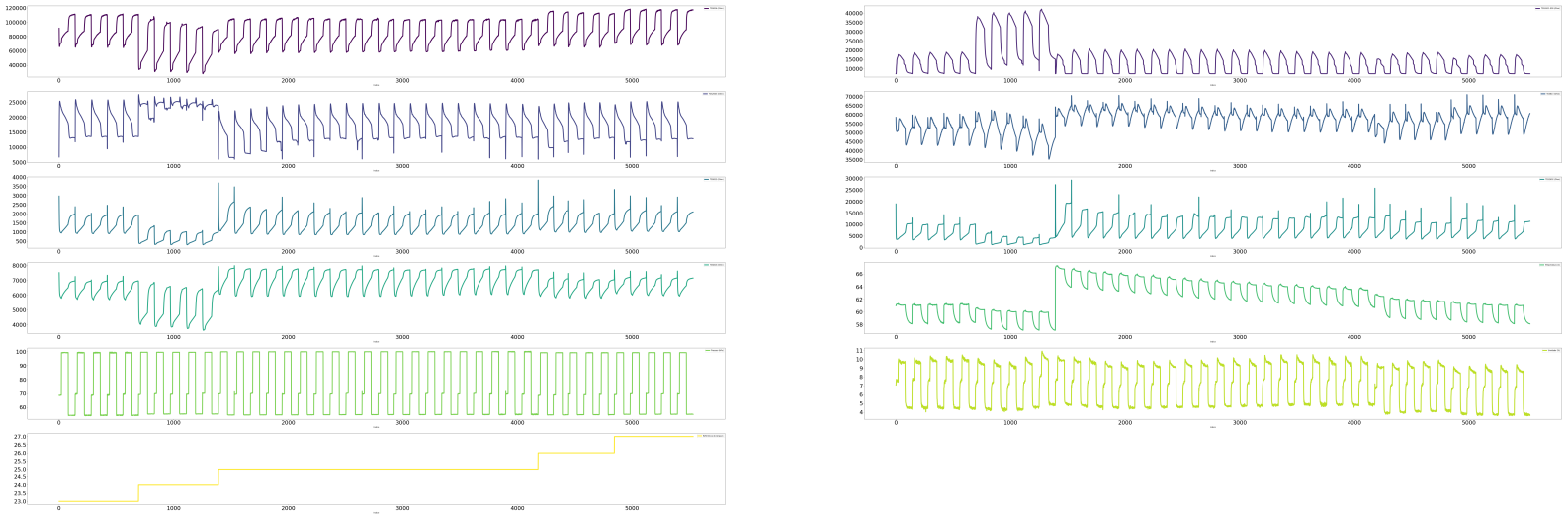
Figura 57 – Registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso para todos os ciclos por espécie: (a) *Candida Albicans*, (b) *C. Glabrata*, (c) *C. haemulonii*, (d) *C. Kodamaea ohmeri*, (e) *C. Krusei*, (f) *C. Parapsilosis* e (g) *C. Tropicalis*



(f)

Fonte: A autora (2023).

Figura 57 – Registro da modificação da resistência da matriz dos sensores quando exposto a um conteúdo gasoso para todos os ciclos por espécie: (a) *Candida Albicans*, (b) *C. Glabrata*, (c) *C. haemulonii*, (d) *C. Kodamaea ohmeri*, (e) *C. Krusei*, (f) *C. Parapsilosis* e (g) *C. Tropicalis*



(g)

Fonte: A autora (2023).