UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
GRADUATE PROGRAM IN COMPUTER SCIENCE

MICHAEL LOPES BASTOS

DiagNose.AI: A Novel Explainable Artificial Intelligence Framework for the Identification of *Candida* spp. from Volatile Organic Compounds using Electronic Noses

Recife

2026

MICHAEL LOPES BASTOS

DiagNose.AI: A Novel Explainable Artificial Intelligence Framework for the Identification of *Candida* spp. from Volatile Organic Compounds using Electronic Noses

Work submitted to the Graduate Program in Computer Science at the Centro de Informática of the Universidade Federal de Pernambuco, as a partial requirement for obtaining the degree of Doctor in Computer Science.

Area of Concentration: Computational Intelligence

Supervisor: Leandro Maciel Almeida

Recife

2026

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Bastos, Michael Lopes.
DiagNose.AI: A Novel Explainable Artificial Intelligence Framework for the Identification of Candida spp. from Volatile Organic Compounds using Electronic Noses / Michael Lopes Bastos. - Recife, 2026.
177f.: il.

Tese (Doutorado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-Graduação em Ciências da Computação, 2025.
Orientação: Leandro Maciel Almeida.
Inclui referências, apêndices e anexos.

1. Explainable Artificial Intelligence; 2. Electronic Nose; 3. Candida Infections. I. Almeida, Leandro Maciel. II. Título.

UFPE-Biblioteca Central

Michael Lopes Bastos

“DiagNose.AI: A Novel Explainable Artificial Intelligence Framework for the Identification of Candida spp. from Volatile Organic Compounds using Electronic Noses”

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

Aprovada em: 25/11/2025.

_____

Orientador: Prof. Dr. Leandro Maciel Almeida

BANCA EXAMINADORA

_____

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

_____

Prof. Dr. Rodrigo Gabriel Ferreira Soares
Centro de Informática / UFPE

_____

Profa. Dra. Margaret Mary VanDeMark Powers-Fletcher
Department of Pathology and Laboratory Medicine
University of Cincinnat

_____

Prof. Dr. Diego Fernando Cuadros
Digital Edidemiology Laboratory/University of Cincinnat

_____

Profa. Dra. Rossana de Aguiar Cordeiro
Departamento de Microbiol / UFC

I dedicate this work to God, my family, friends, and to my wife and daughter, my greatest pillars throughout my journey so far. Everything will always be about you.

ACKNOWLEDGEMENTS

I am grateful to my parents, who gave me life, provided the foundation for my education, and did everything to make this possible. To my friends Severino Augusto, Romualdo Júnior, Wandersson Saraiva, and Irlândio Santana for supporting me throughout my journey. To my sister Michely and my uncle Valdemir, for their strength and inspiration that I have never lacked. To my parents-in-law Adalvina and Aílton, for embracing me as a son and bringing calm during tense moments. To my brothers-in-law Vanderlei, Ariel, Ayla, and José for their constant support, and especially to my girlfriend, fiancée, and now wife, Maria Andressa, for being the love of my life and never letting go of my hand.

Daughter, all of this was, is, and always will be for you. Thank you for being my greatest gift.

I am also grateful to my advisor, Prof. Dr. Leandro Almeida, for his patience, empathy, and invaluable knowledge shared throughout all these years of study. My gratitude equally extends to my co-advisor, Prof. Dr. Margaret Powers-Fletcher (Maggie), and to my colleague Christina Cox, from the University of Cincinnati, who made possible one of the greatest dreams of my life: the academic exchange experience. During this period, I am immensely grateful to Michael and Wendy Zappone for their friendship and fraternal support, which made my journey in the United States even more special. Their support, dedication, and collaboration were crucial to the success of this dream.

I would also like to extend my thanks to the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for their fundamental support in the development of this research. Finally, but not least, I sincerely thank the entire team of Medical Mycology at UFPE, especially my colleague Cícero and Prof. Rejane, as well as all members of the Regional Center for Nuclear Sciences of the Northeast/UFPE, for the valuable partnership established in support of this work.

"I PROPOSE to consider the question, ‘Can machines think?’"(TURING, 1950). Or smell? (Author himself)

RESUMO

As infecções fúngicas, especialmente as causadas por *Candida* spp., representam um desafio crítico em unidades de terapia intensiva, estando associadas a elevadas taxas de mortalidade (40–60%). Esse cenário é agravado pela lentidão e pela baixa sensibilidade ($\sim$50%) do atual padrão-ouro de diagnóstico, a hemocultura. Com o objetivo de superar essas limitações, esta tese propõe, desenvolve e valida um novo *Framework* para a identificação de microrganismos a partir da análise de Compostos Orgânicos Voláteis (VOCs). Essa abordagem estabelece um fluxo de trabalho sistemático que contempla: (i) o desenvolvimento de um protocolo para experimentação e aquisição de dados com Narizes Eletrônicos (E-nose); (ii) a construção e preparação de bases de dados de VOCs de *Candida*, tanto com isolados de cultura quanto em caldo de sangue; (iii) a aplicação e avaliação de modelos de classificação tradicionais e de séries temporais; e (iv) a concepção de uma arquitetura pioneira de explicabilidade (XAI) baseada em um ensemble de técnicas, voltada a assegurar a transparência das predições. A eficácia do Framework foi validada na diferenciação de espécies de *Candida* em diferentes contextos, incluindo cultura e caldo de sangue. Os resultados atestam a robustez da abordagem, com os modelos de classificação alcançando acurácias de 97,46% na abordagem com cultura e 98,18% no contexto com caldo de sangue. Nesse sentido, a principal contribuição desta tese é a criação de um framework computacional que integra uma arquitetura inédita de ensemble de explicabilidade, baseada na combinação de múltiplos métodos, a fim de fornecer interpretações consistentes e multifacetadas das decisões do modelo. A validação dessa abordagem, por meio de estudos de ablação e sensibilidade, confirma seu potencial para aumentar a confiança nos resultados e favorecer a adoção clínica da solução. Assim, o Framework consolida-se como uma contribuição metodológica significativa para a ciência da computação, com impacto direto e relevante na saúde.

Palavras-chaves: Compostos Orgânicos Voláteis, Inteligência Artificial Explicável, Análise de Séries Temporais, Biomarcadores, Nariz Eletrônico, *E-nose*, Infecções por *Candida* spp.

ABSTRACT

Fungal infections, especially those caused by *Candida* spp., represent a critical challenge in intensive care units, being associated with high mortality rates (40–60%). This scenario is aggravated by the slowness and low sensitivity ($\sim$50%) of the current gold standard for diagnosis, the blood culture. To overcome these limitations, this thesis proposes, develops, and validates a new *Framework* for the identification of microorganisms based on the analysis of Volatile Organic Compounds (VOCs). This approach establishes a systematic workflow that includes: (i) the development of a protocol for experimentation and data acquisition with Electronic Noses (E-noses); (ii) the construction and preparation of databases of *Candida* VOCs, from both culture isolates and in blood broth; (iii) the application and evaluation of traditional and time-series classification models; and (iv) the design of a pioneering explainability (XAI) architecture based on an ensemble of techniques, aimed at ensuring the transparency of predictions. The effectiveness of the Framework was validated in the differentiation of *Candida* species in different contexts, including culture and blood broth. The results attest to the robustness of the approach, with the classification models achieving accuracies of 97.46% in the culture-based approach and 98.18% in the blood broth context. In this sense, the main contribution of this thesis is the creation of a computational framework that integrates a novel explainability ensemble architecture, based on the combination of multiple methods, in order to provide consistent and multifaceted interpretations of the model's decisions. The validation of this approach, through ablation and sensitivity studies, confirms its potential to increase confidence in the results and favor the clinical adoption of the solution. Thus, the Framework is established as a significant methodological contribution to computer science, with a direct and relevant impact on healthcare.

Keywords: Volatile Organic Compounds, Explainable Artificial Intelligence, Time Series Analysis, Biomarkers, Electronic Nose, *E-nose*, *Candida* Infections.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Computational intelligence, driven by advances in machine learning and Artificial Intelligence (AI), has been consolidating itself as a transformative pillar for several areas of knowledge. In the health field, in particular, the application of computational techniques has opened new frontiers in the diagnosis, treatment, and monitoring of diseases. The ability to analyze large volumes of complex data and identify subtle patterns, often imperceptible to human analysis, offers unprecedented potential for the creation of clinical decision support systems that are faster, more accurate, and accessible (FAIYAZUDDIN et al., 2025; FURIZAL; MA'ARIF; RIFALDI, 2023).

One of the most critical challenges in the health area is the diagnosis of infectious diseases, where speed and accuracy are direct determinants of the patient's outcome. Traditional diagnostic methods, although established, often face limitations such as high cost, the need for complex laboratory infrastructure, and, crucially, the long time to obtain results. This delay represents a window of vulnerability, especially in intensive care environments, where therapeutic decisions need to be made in a matter of hours, not days (GILL et al., 2023).

To address this gap, a promising frontier in computing applied to health lies in the fusion of AI with advanced sensing technologies (CHEN et al., 2024). The analysis of volatile biomarkers, for example, emerges as a non-invasive approach for the detection of pathologies. In this context, the Electronic Noses (*E-Noses*) — a device that combines an array of chemical sensors with pattern recognition algorithms — represents a powerful computational tool, capable of generating multidimensional and temporal data from "olfactory signatures" of biological samples (FARRAIA et al., 2019). The computational challenge, therefore, lies in developing robust methodologies to extract significant knowledge from this raw data and translate it into interpretable and reliable diagnoses.

1.1 MOTIVATION

To validate and demonstrate the potential of a new computational approach, it is essential to apply it to a high-impact problem with well-defined limitations. Invasive fungal infections (IFIs), especially those caused by *Candida spp.*, represent precisely this scenario. These infections are a serious problem in Intensive Care Units (ICUs), with mortality rates ranging from

40% to 60% (LI et al., 2018).

One of the reasons contributing to high mortality rates is the clinical challenge in recognizing and diagnosing Invasive Fungal Infections (IFIs) during the initial stages of treatment (PAPPAS et al., 2018a). According to Barantsevich and Barantsevich (2022), traditional culture-based methods typically require 2 to 7 days to determine results, which is a critical limitation given the severity of the pathology (BARANTSEVICH; BARANTSEVICH, 2022). Research indicates that a delay of more than 12 hours in initiating appropriate therapy can significantly increase mortality chances, highlighting the urgent need for diagnostics-driven antifungal stewardship to manage these infections effectively (VERGIDIS et al., 2016; CHAKRABARTI et al., 2022).

Currently, blood culture is the gold-standard method for the laboratory diagnosis of candidemia, by isolating the etiological agent for identification (BEYDA; ALAM; GAREY, 2013). However, there are other methods based on the identification of Volatile Organic Compounds (VOCs) that can be used to identify these fungal agents, such as: Gas Chromatography-Mass Spectrometry (GC-MS), Solid-Phase Microextraction (SPME), Simultaneous Distillation-Extraction (SDE), and Selected Ion Flow Tube Mass Spectrometry (SIFT-MS) (MORATH; HUNG; BENNETT., 2012).

The need for a diagnostic method that is at the same time fast, accurate, and low-cost is, therefore, urgent and evident. This complex problem, which involves the analysis of subtle biological signals to differentiate multiple pathogen species, serves as the ideal use case for the development and validation of a new computational paradigm (WU et al., 2022).

One of the still little-explored methods, but with great growth potential, is the Electronic Noses, or *E-Noses*, as it is also called. This technology combines a set of gas sensors and Artificial Intelligence (AI) to recognize VOC patterns and classify the "olfactory fingerprints" released by these compounds. This technology is already used in several areas, such as food safety, agricultural applications, and in the field of disease diagnosis (MATYSIK; HERBARTH; MUELLER., 2009) apud (MORATH; HUNG; BENNETT., 2012).

Given the above, it is understood that there is a great challenge in the rapid identification of fungi in hospitalized patients and in clinical conditions that require extra care (FILHO, 2009), (PAPPAS et al., 2016). In response to this challenge, this thesis proposes, develops, and validates the *Framework DiagNose.AI*, an original and complete computational methodology for the identification of microorganisms from the analysis of Volatile Organic Compounds (VOCs). Instead of a one-off application, the Framework establishes a systematic workflow that integrates five essential pillars:

(I) Experimental Protocol and Data Acquisition: The definition of a standardized process for collecting data from culture and blood broth samples with Electronic Noses.

(II) Data Engineering: The methodology for building and making available robust databases on VOCs (Volatile Organic Compounds) collected by the electronic nose, aiming to foster research in this area.

(III) VOC Mapping: Creation of a database relating the VOCs emitted by different Candida species and identified by the gas sensors. of *Candida* and the VOCs detected by different types of gas sensors;

(IV) Predictive Modeling: The systematic application of classification models, using time series and traditional models, to ensure maximum accuracy.

(V) Explainable AI (XAI): The development of a pioneering Ensemble XAI architecture, designed to provide transparent and reliable explanations, crucial for adoption in high-risk environments such as medicine (WANG et al., 2020a).

Thus, the motivation for this work presents itself in two complementary dimensions: on one hand, the clinical urgency for faster and more effective diagnostic methods for fungal infections; on the other, the scientific opportunity to propose a complete and innovative computational methodology. The DiagNose.AI Framework emerges precisely at the intersection of these challenges, offering a solution that not only responds to a concrete demand in healthcare but also contributes in an original and systematic way to the advancement of applied Artificial Intelligence.

1.2 OBJECTIVES

In this section, the objectives related to the construction of this project are defined, divided into General Objective and Specific Objectives.

1.2.1 General Objective

The general objective of this project is to propose, develop, and validate the *Framework DiagNose.AI*, an innovative and complete computational methodology for the identification of microorganisms, using species of *Candida* spp. as the main focus. The Framework integrates

the analysis of volatile organic compounds (VOCs) through Electronic Noses with AI models and a novel Explainable Artificial Intelligence (XAI) architecture based on an Ensemble, aiming to ensure a fast, accurate, and interpretable diagnosis.

1.2.2 Specific Objectives

The specific objectives of this thesis, which support the pillars of the DiagNose.AI Framework, are:

- **To define and validate** an experimental protocol for the acquisition of Volatile Organic Compounds (VOCs) data with Electronic Noses, applicable to both culture media and blood broth samples;
- **To build and characterize** two new time series databases, detailing the VOC profiles for the studied *Candida* species;
- **To map VOCs** by means of a database of the VOCs emitted by candida and identified by the gas sensors;
- **To systematically develop and evaluate** the performance of classification models, with an emphasis on time series, for the accurate identification of microorganisms from the generated data;
- **To validate the end-to-end effectiveness** of the DiagNose.AI Framework through its application in the classification of different *Candida* species in different scenarios, demonstrating its potential for clinical impact.
- **To design and evaluate** a pioneering Explainable Artificial Intelligence (XAI) architecture, based on an ensemble of methods, to provide robust and multifaceted explanations for the model's predictions;

1.3 ORGANIZATION OF THIS THESIS

The structure of this work will be organized as follows: Chapter 2 will present the Theoretical Foundation, providing a brief explanation of the main topics that guide the project. Chapter 3 will detail the Related Works, with emphasis on initiatives aimed at the Identification

of Fungi, the use of E-noses in the detection of volatile compounds, and explainability methods focused on XAI Ensembles. Next, Chapter 4 will describe the Materials and Methods (THE DIAGNOSE.AI FRAMEWORK: DEVELOPMENT AND METHODOLOGY) used for the implementation of the proposed solution. The Results and discussions (VALIDATION OF THE DIAGNOSE.AI FRAMEWORK: RESULTS AND DISCUSSIONS) on the development will be addressed in Chapter 5. Chapter 6 will present the main Conclusions reached so far in this investigation. Finally, Chapter 7 (MAIN CONTRIBUTIONS AND FUTURE PERSPECTIVES) will discuss the main milestones of the project and some Future Perspectives for this work.

2 THEORETICAL FOUNDATION

This chapter provides a brief explanation of fungi, with an emphasis on *Candida* spp., and the main methods used for its identification. Additionally, it will cover information on the use of the Electronic Noses and the most currently used Artificial Intelligence and time series techniques in this context. Furthermore, the context of the main explainability methods will also be discussed.

2.1 THE KINGDOM OF FUNGI

From a historical and scientific point of view, it can be said that fungi have not had a very relevant role in the past. Initially studied by botanists, little was said about this category of microorganisms, being studied together with and classified within the plant kingdom. Fungi were classified into their own kingdom only a few decades ago, as they possess functional and morphological characteristics very distinct from their kingdom of origin and from any other existing one. Named the Fungi Kingdom, it is estimated that there are between 2.2 and 3.8 million species in all of nature, among them, only 144,000 have been properly described and classified (ARAúJO, 2021). Nowadays, fungi are considered the largest group of eukaryotes among all kingdoms. However, the scarcity of specialists and the late start of studies focused on this kingdom still make it largely unknown, with only an approximate knowledge of 8% of the total Funga supposed for the entire world being estimated, complicating studies aimed at understanding the group's evolution (ARAúJO, 2021).

To better understand the characteristics of a fungus, according to Araújo (2021), it is important to comprehend the phylogenetic scenario of the kingdom itself. Although the history of mycology is directly linked to botanists, the Fungi Kingdom is, in terms of evolution, closer to animals than to plants. In biology, there is a supergroup called Opisthokonta, which includes some species of the Animalia Kingdom, other eukaryotes, and the fungi themselves (ARAúJO, 2021). This supergroup is represented especially by two major categories, Holozoa and Holomycota, which differ according to their forms of digestion. Both lineages seek their food in the environment (they are heterotrophic), which eliminates any kinship with plants (ARAúJO, 2021).

Unlike the members of the Animalia Kingdom belonging to Holozoa, which perform internal

digestion by ingesting food and then carrying out the degradation and absorption of nutrients in the body, fungi belonging to Holomycota perform this process externally, subsequently absorbing only the nutrients. This action is called heterotrophy by absorption and is considered one of the most relevant characteristics of the Fungi Kingdom, considering that, to perform this process, fungi release digestive enzymes, which chemically modify the surrounding area. The result of this action ends up being a great ecological contribution of fungi to nature, as they modify the soil biomass, allowing a portion of lifeless organic matter to re-emerge in the future (ARAúJO, 2021).

Given the above, the existence of a great diversity of fungi with relevant importance for living beings and the environment is clear (AGUIRRE et al., 2016). However, amidst all this diversity, there are also fungi that can be a problem for humans, such as some yeasts found in the human body itself that, in case of low immunity, can spread and reach regions where they cannot be hosted (MIOTTO et al., 2004). One of these fungi is *Candida* spp., one of the subjects of this study, addressed in more detail in the next Section.

2.1.1 The genus *Candida*

According to Sarma e Upadhyay (2017), fungi have emerged as one of the main causes of diseases in humans, mainly affecting patients hospitalized for long periods or with a compromised immune system (PAPPAS et al., 2018b). The genus *Candida* is responsible for about 80% of the fungal infections already reported, representing a great challenge for health professionals in various clinical areas, given the diagnostic and therapeutic difficulties of infections caused by this agent (COLOMBO; GUIMARãES, 2003).

In this sense, *C. albicans*, as well as *C. glabrata*, *C. krusei*, *C. tropicalis*, and *C. parapsilosis* are responsible for a large part of hospital-acquired infections, primarily affecting the oral (superficial) or vaginal regions of patients. Depending on the conditions, the fungus can come into contact with the bloodstream, causing deep infections in the body (DADAR et al., 2018; WHALEY et al., 2017). *C. albicans*, for example, is part of the human microflora and is commonly found in the gastrointestinal, respiratory, and genitourinary tracts. In general, it is a harmless fungus, but it can become opportunistic in immunocompromised or immunodeficient individuals (SARMA; UPADHYAY, 2017; DADAR et al., 2018).

Among the non-*Candida albicans Candida* (NCAC) types, one of the species that also causes great concern is *C. parapsilosis*, which causes 17 to 50% of fungemia according to

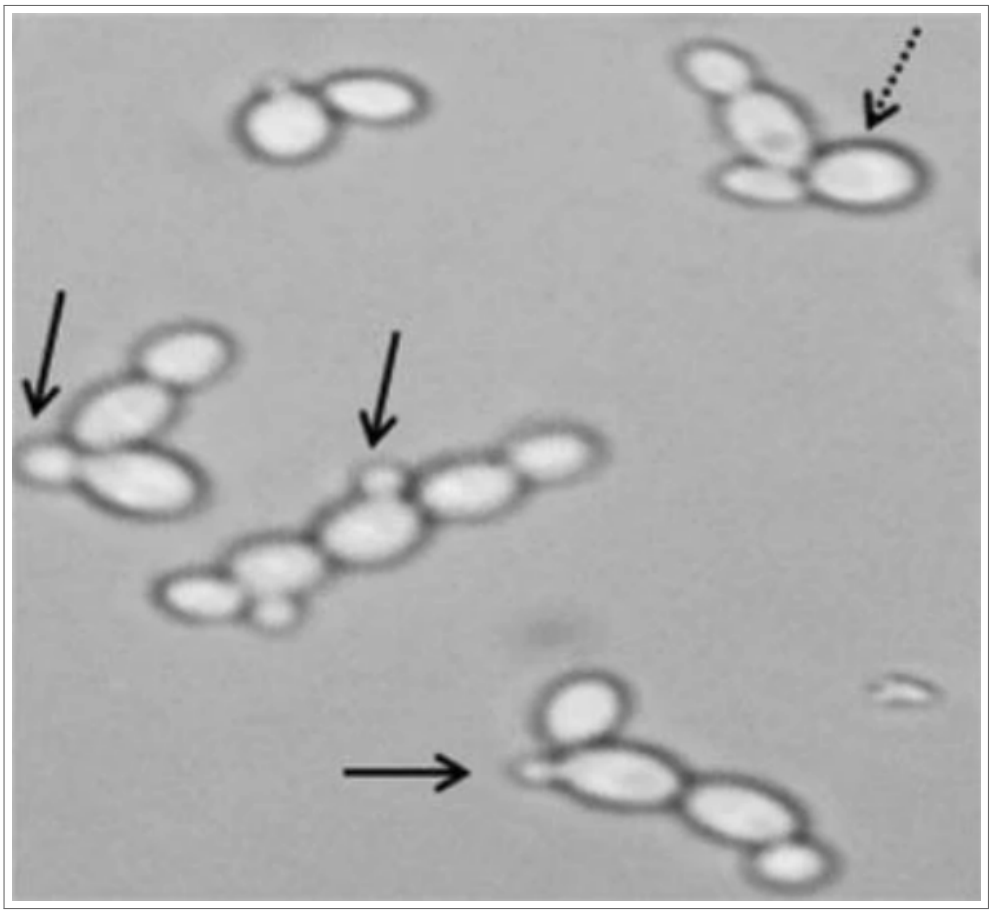
Whaley et al. (2017) and 46% according to another study Corona et al. (2011), especially in babies and newborns. This species can also create tenacious biofilms on central venous catheters and other clinically implanted equipment in patients undergoing invasive medical procedures (TóTH et al., 2019). Another NCAC that draws the attention of health professionals is *C. krusei*. This species shows high resistance to some drugs and causes a high mortality rate, ranging from 20% to 67% (SM et al., 2019; NAVARRO-ARIAS et al., 2019). Historically, the main cases of *C. krusei* are reported in neonates, people with pathogens in intra-abdominal abscesses, endocarditis, infectious arthritis, urethral obstruction, esophagitis, eye infections, cancer patients, and bone marrow transplant recipients (SAMARANAYAKE et al., 1994).

Regarding identification, it is usually possible to differentiate *Candida* isolates in cultures through methods that involve the germ tube test, chlamydospore formation, and the fermentation and digestion of sugars (Figure 1) (ALAM et al., 2014). According to Alam et al. (2014), the germ tube test has a rapid identification time for *C. albicans* (between 2 and 4 hours), but it requires well-trained professionals for identification and is not precise, given that about 5% of *C. albicans* isolates do not produce germ tubes, while on the other hand, some isolates of *C. tropicalis* also have the ability to produce them. The type of agar used and the temperature can also influence the identification and differentiation of some species, as some grow better with specific types of agar and at known temperatures (ALAM et al., 2014).

Given the complications these tests bring regarding precision in interpretation, new ways for the presumptive identification of yeasts were developed. Different chromogenic media aimed at the isolation and detection of *Candida* species have been created. These techniques are based on the analysis of different colored colonies with diverse shapes that imply the cleavage of chromogenic substrates by species-specific enzymes. Examples of commercially used chromogenic agars today include Tween 80 Corn Meal agar, cornmeal agar, CHROMagar Candida, Fluroplate, Candichrom, Pagano-Levin agar, Costa-de Lourdes Branco and albicans ID agar, CHROMagar, and BiGGY agar. Figure 2 demonstrates the growth morphology on Tween 80 Corn Meal for different *Candida* species (ALAM et al., 2014).

In this sense, among the main methods currently available for the identification of *Candida* spp., are polymerase chain reaction (PCR) assays, (1$\rightarrow$3)-β-D-glucan (BDG) (CLANCY; NGUYEN, 2013), and loop-mediated isothermal amplification (LAMP) (FALLAHI et al., 2020), with the first two being considered the gold standards for *Candida* identification today. However, as previously seen, most of these methods have sensitivity issues and are very expensive, making their use unfeasible in some regions.

Figure 1 – Demonstration of budding (marked with solid arrows) and yeast form (dashed arrow) in *Candida albicans*

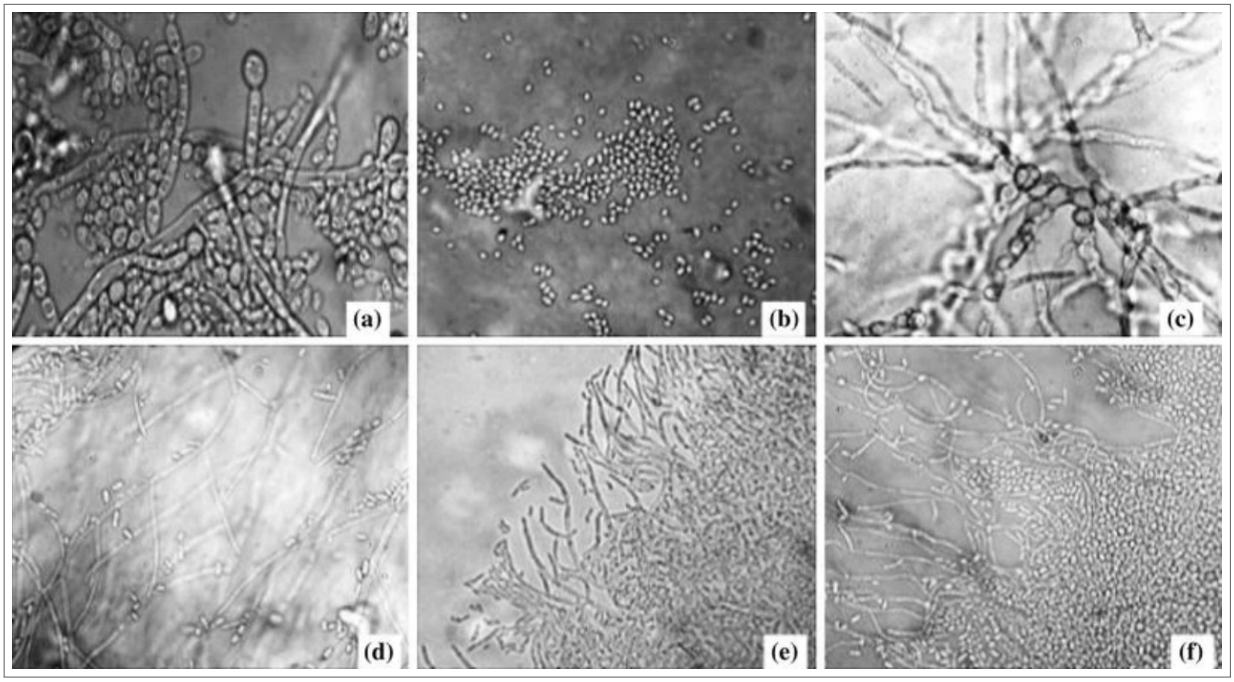


Source: ALAM et al. (2014)

Another major problem associated with these methods is the time required to obtain results (which can take from 2 to 7 days). This directly harms the treatment of patients and reinforces the need for the creation of new methods that can make this process more efficient (PINI et al., 2019). Another factor that raises an alarm about the damage caused by this fungus is the economic one. Studies show that in the United States alone, a total of US$ 4.6 billion was recently spent on direct medical costs related to hospitalizations for fungal diseases. *Candida* infections alone (26,735 hospitalizations) were responsible for a cost of US$ 1.4 billion to the American coffers, not including unrecognized fungal diseases and expenses with unnecessary exams, medical procedures, and improper treatments, until a correct diagnosis of the fungal infection was identified (TERRERO-SALCEDO; MARGARET, 2020).

It is therefore evident that the diversity of fungal species and the limitations of conventional diagnostic methods represent a critical gap in the clinical environment. It is precisely in this gap that the intersection of biology and computer science offers a promising solution. Fungal

Figure 2 – Use of Tween 80 Corn Meal agar (CMA) to verify the growth morphology of different *Candida* species. **(a)** *C. albicans*, **(b)** *C. glabrata*, **(c)** *C. parapsilosis*, **(d)** *C. krusei*, **(e)** *C. Kefyr*, and **(f)** *C. tropicalis*



Source: ALAM et al. (2014)

metabolism, which results in the emission of Volatile Organic Compounds (VOCs), generates chemical "signatures" that can be captured by technologies like Electronic Noses, producing data in the form of time series, allowing their analysis through different Artificial Intelligence methods.

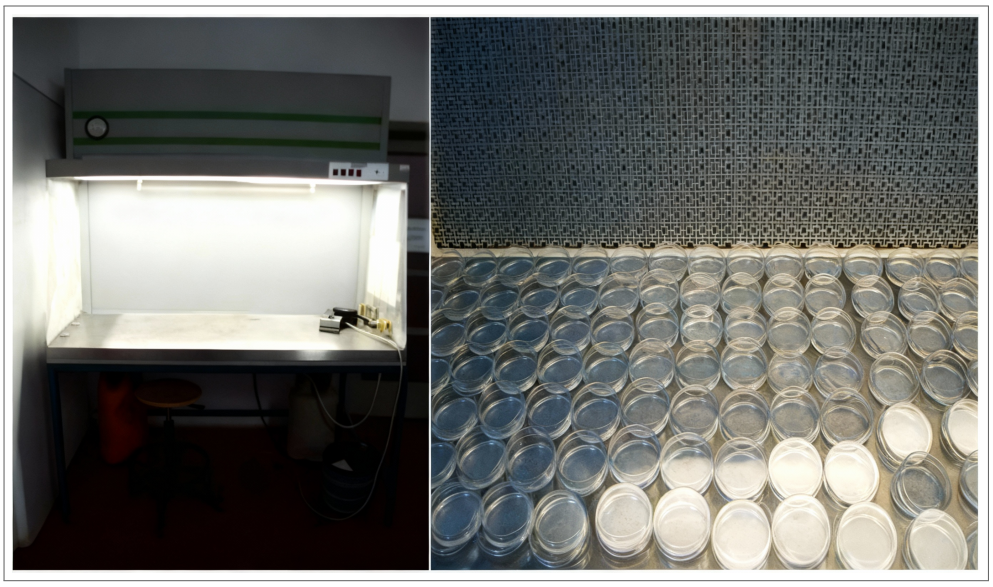
2.1.2 Process of obtaining colonies and culture media

To better understand the process of obtaining colonies and culture media, Trovão e Pereira. (2019) explain that fungi are generally separated by plating a previously collected sample. This sample is derived from the material where the presence of fungi is to be identified, such as soil, liquid, air, or a specific surface sample. This material is placed on a *Petri* dish with the appropriate culture medium for its growth. The plating process can be done in different ways, such as by diluting the sample in water, for example, or by using a low-concentration saline solution, which should then be distributed on the plate that will be used for the culture (TROVãO; PEREIRA., 2019). Thus, it can be said that the preparation of culture media can be done by adding its components to distilled water or by solubilizing commercial lyophilized media, followed by sterilization in an autoclave for about 15 minutes at a pressure of 1 ATM

and a temperature of 121 ºC (TROVãO; PEREIRA., 2019).

In the same context, after a rapid cooling of the medium, avoiding complete solidification, antibiotics can be used to prevent the growth of unwanted bacteria, which are meticulously spread over the *Petri* dish. This entire procedure needs to be performed in a flow hood (left part of Figure 3), already sterilized with ultraviolet light and alcohol, not allowing external contamination, under total aseptic conditions. After the complete cooling of the culture medium in the *Petri* dishes, a 5-day reservation period for the medium is indicated to ensure that there is no contamination of any kind. At the end of this period, the plates will be ready for use in the inoculation or propagation of the colonies (right part of Figure 3).

Figure 3 – Example of a laminar flow hood on the left and a set of culture media in the drying process on the right



Source: TROVãO; PEREIRA. (2019)

Among culture media, there is a usual classification used to describe them as "rich," "generic," and "poor," related to the nutrient concentrations of each. In general, poor media are manipulated to instigate the formation of sexual structures in media that do not easily sporulate. Such incitement is based on the concept that the accumulation of nitrogen and carbon sources and the cultivation temperature encourage the mode of reproduction, whether asexual or sexual. One of the possibilities of cultivation in poor media, aiming to influence sporulation, can be the cultivation in a medium whose surface has been placed over part of the organism it parasitizes. In this context, this methodology allows for *in situ* laboratory growth and the visualization of reproductive formations. In addition, external factors such as pH, temperature, humidity, and light must be considered, as they can affect the growth of

fungi, as well as the probable more energetic sporulation of some species, which can create a predominance of these over those with less sporulation or slower growth (TROVãO; PEREIRA., 2019).

2.1.3 Handling samples: preservation and propagation of cultures

In biology, the area responsible for studies related to the mycological diagnosis of fungal infections is called Medical Mycology. It is based on the correct execution of the processing and collection of clinical specimens. The entire context related to the preservation, transport, and handling of clinical material is of great relevance in obtaining efficient and reliable results (MOLINARO; CAPUTO; AMENDOEIRA., 2012).

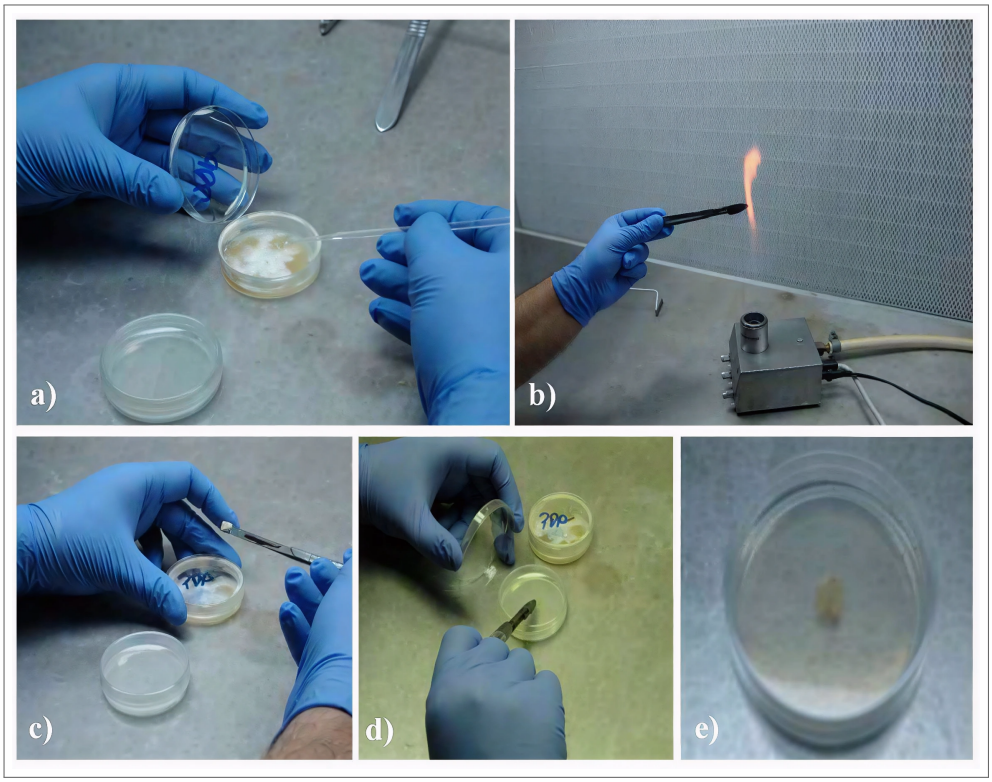
In the mycological context, there are some techniques used for handling and propagating cultures, the most common being:

- **Serial dilution** - A common technique that can be performed in different scenarios, such as the separation of two fungal strains mixed in a tube or plate, the counting of colonies in a sample, the separation of fungi from soil and liquid substrates, and the definition of inoculum quality in fermentation processes or liquids.
- **Fungal seeding (propagation) technique** - This method is used for the identification of filamentous fungi through their morphological characteristics, using inoculation on plates. It is a well-developed and widely used axenic culture (pure cultures) technique.
- **Microscopy** - In this case, a direct examination of a part of the colony is performed. This procedure is carried out using a sterilized needle or an L-shaped platinum loop, used to cut the colony and place it on the slide.

Based on the above, a greater emphasis will be given to the fungal seeding or propagation technique. In this procedure, after acquiring the varieties of cultures separated by plating, they must be kept in pure cultures by propagation, or as it is also called, "subculturing". This action refers to the inoculation of a minimal fraction of the fungus on *Petri* dishes using fresh culture medium. To perform the spreading, scalpels or inoculation loops, a *Bunsen* burner, and parafilm can be used, all in a flow hood under total aseptic conditions (TROVãO; PEREIRA., 2019). For subculturing, the scalpel or inoculation loop is first placed in the flame of the *Bunsen* burner until it changes color and becomes incandescent. After that, one must wait for the metal to

cool (to avoid deteriorating the structures due to heat), to transfer the mycelium or yeast to a new medium, already with the corresponding agar. If it is necessary to repeat the process, the scalpel or inoculation loop must again be exposed to the flame of the *Bunsen* burner, in order to always maintain correct sterilization of the sample handling material (TROVãO; PEREIRA., 2019). An example of this process can be seen in Figure 4.

Figure 4 – a) Seeding with inoculation loop; b) Sterilization of scalpel; c) Seeding with scalpel; d) Seeding in new culture medium; e) Subcultured *Petri* dish after propagation



Source: TROVãO; PEREIRA. (2019)

Regarding the preservation of cultures, it can be done by different types of mechanisms and for the desired length of time. For periods of many months, cooling the culture can be done at 4°C; however, for a more prolonged preservation period, another procedure must be used, such as lyophilization or the use of liquid nitrogen. In this sense, preserving cultures at 4°C tends to be done in two ways. The first corresponds to cooling colonies between 5 and 7 days at 4°C, while the second option is to subculture the culture in a solid medium tube, adding mineral oil until the surface of the colony is completely covered, followed by subsequent cooling at 4°C. For this culture to be used at another time, it is necessary to subculture the colony in a new culture medium and incubate it for a period of 5 to 7 days at approximately 25°C, and the procedure should be redone if the colony loses vitality (TROVãO; PEREIRA., 2019).

One of the widely used enriched media in microbiology is "blood broth" — a base broth

(e.g., BHI or peptone-based medium) to which sterile defibrinated blood is added in defined proportions (blood:broth ratio usually between 1:5 and 1:10, adjusted according to the experimental objective). The base medium must be prepared and sterilized by autoclaving (121 °C, 15 min) and, after cooling to approximately 40–45 °C, the sterile blood is added, avoiding violent shaking that could cause hemolysis. These precautions preserve the integrity of the red blood cells and plasma factors necessary for the growth of fastidious microorganisms, including species of the genus *Candida*. (FALCONER; HAMMOND; GILLESPIE, 2020; BUXTON, 2016)

The handling of blood broth must occur under strictly aseptic conditions (hood/laminar flow) and the final medium must be gently homogenized, refrigerated, and used within a short period due to its limited shelf life. In clinical practice and experimental protocols, the blood:broth ratio is a critical parameter: for example, ratios close to 1:10 are often used in adults, while in pediatric samples with reduced volumes, higher ratios (e.g., up to 1:100) are tested without substantial loss of sensitivity. Recent literature on the optimization of blood cultures and recovery methods reinforces the need for rigor in preparation and handling to ensure microbiological recovery and experimental reproducibility. (FALCONER; HAMMOND; GILLESPIE, 2020; BUXTON, 2016)

A detailed understanding of cultivation methodologies, both in solid (plating) and liquid (blood broth) media, is fundamental, as these protocols establish the basis for generating the distinct datasets analyzed in this project. Critical parameters such as the composition of the culture medium, the incubation temperature, or the blood:broth ratio in the case of blood broth, directly influence fungal metabolism and, consequently, the profile of emitted Volatile Organic Compounds (VOCs). Since these VOCs are the "signature" that the Electronic Noses aims to identify, rigor in the standardization of these procedures — regardless of the sample type — is an indispensable prerequisite to ensure the generation of consistent and high-fidelity data, essential for the training and validation of Artificial Intelligence models capable of performing an accurate and reliable classification.

2.1.4 Laboratory methodologies for fungal identification

In recent decades, remarkable progress has been noted in the field of fungal identification and detection. New techniques are emerging with the aim of improving the recognition time of these microorganisms, which have brought both social and financial problems (MARCOS; PINCUS, 2013). As briefly mentioned, the early diagnosis of fungal infections is extremely

important for the execution of an efficient treatment by the medical team. However, there are numerous factors that can hinder this process, such as the reduction in the number of clinical mycologists, diagnosis time, associated costs, and issues of sensitivity and specificity of the method to be used. Another important factor is that the diagnosis of these agents must meet the needs related to the constant emergence of new variants. These are generally found in cases related to patients with immune deficiency, common in countries that demand a high level of medical care and do not have the adequate resources for a correct diagnosis (KOZEL; WICKES., 2014).

Within this context, among the more traditional approaches to fungal identification are: direct microscopic examination of clinical samples, histopathology, culture, and serology. Among the emerging technologies, molecular diagnosis and the detection of antigens in clinical samples can be mentioned (KOZEL; WICKES., 2014). All these approaches require the use of people with a high level of specific training in mycological handling. Furthermore, the growth in the number of fungi identified by clinical mycologists in recent years increases the need for the development of new, more efficient methods than those traditionally used (KOZEL; WICKES., 2014). Below is a brief description of the most used and most promising methods in fungal identification today.

2.1.4.1 Culture-based fungal identification

Despite being a less sensitive and time-consuming method, the isolation of a fungal pathogen by culture remains the gold standard for the diagnosis of Invasive Fungal Infections (IFI) in most situations, playing a very important role in providing *in vitro* sensitivity data. Therefore, when discussing the future of fungal diagnostic media, this method still requires attention. In recent years, a series of studies have been combining ways to improve the performance and application of fungal culture through two special research categories: surveillance fungal culture and identification techniques using proteomics (MALDI-TOF MS) (TERRERO-SALCEDO; MARGARET, 2020).

Over the years, studies have proven the good utility of the surveillance culture-based approach for patients at risk of acquiring an IFI. As an example, the study reported in Terrero-Salcedo e Margaret (2020) *apud* (HONG et al., 2017) can be cited, in which the results indicate that, although 80% of mycology laboratories do not routinely perform mycology tests on samples from patients with cystic fibrosis, the inclusion of selective fungal culture media doubled

the yield of clinically important fungi, when compared to isolated routine bacterial culture conditions. This report may indicate that current procedures for evaluating the diagnosis of infectious diseases may fail in the identification in vulnerable patients (TERRERO-SALCEDO; MARGARET, 2020).

Regarding identification methods, proteomic detection with MALDI-TOF MS is still a widely used method. However, it can be time-consuming and laborious, heavily dependent on the experience of the clinical mycologists responsible for handling the samples. Even the most skilled experts can have difficulty in identification, as not all clinically important molds can be accurately identified using only phenotypic methods. For this reason, many clinical mycology laboratories are subject, to a certain extent, to DNA sequencing as the current gold standard for the identification of fungi.

Although the most reported use of sequencing for the identification of fungal pathogens has improved, there are still many limitations to this method. For example, the process of accurately identifying fungi to the species level using sequencing still has its limitations. The use of sequencing information from multiple genetic targets in a recognition algorithm needs to be used several times to enable accurate distinction between species. In addition, there are problems regarding the cost of the tests, the limited commercial availability of the technology in laboratories, and problems related to the databases used to compare the sequencing results (TERRERO-SALCEDO; MARGARET, 2020). Considering these limitations, the approach proposed in this project is even more promising, as it opens new perspectives for identification combined with reduced costs, ease of portability, greater transparency, and high levels of accuracy.

2.1.5 Volatile Organic Compounds (VOCs)

According to Morath, Hung e Bennett. (2012), Volatile Organic Compounds (VOCs) are carbon-based solids and liquids that enter a gaseous state at a temperature of approximately 20 °C, having low solubility in water. In this regard, about 250 VOCs have been identified based on fungi, through mixing with components such as simple hydrocarbons, heterocycles, aldehydes, ketones, alcohols, phenols, thioesters and derivatives, as well as benzene and cyclohexane derivatives. Fungal volatiles can be derived from both secondary and primary metabolism pathways.

Among the main methods used in the identification of fungal VOCs, some have gained prominence, they are:

- **Gas Chromatography-Mass Spectrometry (GC-MS)**: one of the most used methods due to its powerful separation of volatiles and highly sensitive detection, but it is not used for the identification of new compounds;
- **Solid-Phase Microextraction (SPME)**: used for reducing preparation time by compacting the extraction, concentration, and introduction steps into a unified stage, making it more sensitive than other methods;
- **Simultaneous Distillation-Extraction (SDE)**: a method that combines steam distillation and solvent extraction, commonly used to analyze the VOCs of *Penicillium roqueforti* in comparison with SPME. However, it has proven inadequate over the years as it could not define a complete volatile profile;
- **Selected Ion Flow Tube Mass Spectrometry (SIFT-MS)**: provides rapid and broad-spectrum detection of VOC characteristics in slightly complex gas compounds. It is generally used for studies of VOCs generated by *Aspergillus, Candida, Mucor, Fusarium*, and *Cryptococcus* species;
- **Proton-Transfer-Reaction Mass Spectrometry (PTR-MS)**: performs the ionization of organic molecules in their gaseous state through their reaction with H3O+ and can be used for the identification of fungal VOCs. Some of its main features include the ability for fine-scale time detection and response, near-real-time analysis without sample separation, derivatization, or concentration, with its sensitivity being comparable to the GC-MS method;
- **Electronic Noses - E-Noses**: A still little-explored method, but very promising for the identification of volatile fungal compounds. This type of technology combines a set of gas sensors and artificial intelligence for the recognition of VOC patterns and "smell fingerprints." It is already being used in different areas, such as food safety, agricultural applications, and in the field of disease diagnosis.

In the context of identifying VOCs of *Candida* spp., the work developed by Hertel et al. (2018b) sought to identify whether it is possible for Specific Volatile Organic Compounds (SVOCs) to be found in patients with oral candidiasis, through breath analysis, for the development of a new diagnostic tool. In the context of the work, samples were collected from

20 individuals, 10 with *Candida* spp. and 10 without the presence of the fungus. These samples were analyzed by gas chromatography and mass spectrometry. As a result, a total of 143 VOCs were identified in the collections made in both categories of individuals, but the volatiles found did not contain specific signatures, generally known to be emitted by *Candida* spp. *in vitro*. However, some patterns were identified containing nine volatile compounds (2-methyl-2-butanol, hexanal, longifolene, methyl acetate, 1-heptene, acetophenone, decane, 3-methyl-1-butanol, chlorobenzene), where characteristic changes were identified at the time after antifungal therapy.

Furthermore, in the research led by Perl et al. (2011), an innovative method for identifying volatile organic compounds was demonstrated, which uses a multi-capillary column ion mobility spectrometer (MCC-IMS) and was evaluated in the study for the identification of VOCs in the headspace of *A. fumigatus* and four *Candida* species, namely *Candida albicans*, *Candida parapsilosis*, *Candida glabrata*, and *Candida tropicalis*, being validated for *A. fumigatus* and *C. albicans* through the currently best-known method for this type of purpose, GC/MS. With the use of the GC/MS method on the samples, isoamyl alcohol, cyclohexanone, 3-octanone, phenylethyl alcohol, p_0642_1/p_683_1, and p_705_3 were identified as discriminating volatiles, while with the use of the method proposed by the authors, the MCC-IMS, 3-octanone and phenylethyl alcohol were identified. Some substances were not correctly detected by the MCC-IMS method, namely isoamyl alcohol and cyclohexanone. In a general context, the MCC-IMS method becomes an important and viable alternative according to the authors, due to its feasibility for rapid analysis and complex gas mixtures, eliminating the need for pre-concentration or sample preparation, in addition to not depending on the water vapor content and certain configurations. In this sense, it was possible to discriminate fungi at the genus level of the germs analyzed by the volatile metabolic profile, affirming its efficiency for VOC detection. However, discrimination at the species level for *Candida* species is not yet feasible by the method.

Given this, with the intention of conducting a small survey of the most commonly found VOCs in variations of *Candida* spp., a brief investigation was made in the literature on works involving the description of volatiles of this fungus. Table 1 describes all the volatiles found in the articles and also in the VOC database platform[1].

The reported information is of great importance for the construction of Electronic Noses that use sensors directly related to the VOCs identified for the reported *Candida* species. All

[1] https://bioinformatics.charite.de/mvoc/index.php?site=ergebnis

Table 1 – List of VOCs related to some *Candida* species found in the survey conducted in this study

VOCs	***Candida* Species**	**VOCs**	***Candida* Species**
Decane (DEC)	*C. glabrata*, *C. parapsilosis*, *C. tropicalis*, *C. albicans*	Tyrosol	*C. auris*
Methyl Acetate, 1HE: 1-Heptene (MET)	*C. glabrata*, *C. parapsilosis*, *C. tropicalis*, *C. albicans*	Palmitelaidic Acid	*C. auris*
2-Methyl-2-Butanol (2ME)	*C. glabrata*, *C. parapsilosis*, *C. tropicalis*, *C. albicans*	1-Dodecanol	*C. albicans*
Hexanal (HEX)	*C. glabrata*, *C. parapsilosis*, *C. tropicalis*, *C. albicans*	E-Nerolidol	*C. albicans*
Longifolene (LON)	*C. glabrata*, *C. parapsilosis*, *C. tropicalis*, *C. albicans*	E-Farnesol	*C. albicans*
1-Heptene, Acetophenone (ACE)	*C. glabrata*, *C. parapsilosis*, *C. tropicalis*, *C. albicans*	Ethyl Hexanoate	*C. sake 41E*
Chlorobenzene (CHL)	*C. glabrata*, *C. parapsilosis*, *C. tropicalis*, *C. albicans*	3-Methylbutyl Pentanoate	*C. sake 41E*
2-Phenylethanol	*C. glabrata*, *C. parapsilosis*, *C. tropicalis*, *C. albicans*	Methylpropyl 2-Hexanoate	*C. sake 41E*
Cyclohexanone	*C. glabrata*, *C. parapsilosis*, *C. tropicalis*, *C. albicans*	3,7-Dimethyl-6-Octen-1-Ol	*C. sake 41E*
3-Methyl-2-Butanone	*C. albicans*	Methylbutyl 3-Hexanoate	*C. sake 41E*
1-Hexanol	*C. tropicalis*	Phenylethyl 2-Acetate	*C. sake 41E*
P-Xylene	*C. krusei*	3-Methylbutyl Cyclopentanecarboxylate	*C. sake 41E*
2-Octanone	*C. krusei*	6-Octen-1-Ol, 3,7-Dimethyl-, Propionate	*C. sake 41E*
N-Butyl Acetate	*C. krusei*	3-Methylbutyl Octanoate	*C. sake 41E*
2-Heptanone	*C. krusei*, *C. kodamaea ohmeri*	3-Methyl-L-Butanol	*C. kodamaea ohmeri*
Benzyl Alcohol	C. auris	2-Methyl-L-Butanol	*C. kodamaea ohmeri*
3-Methyl-1-Butanol (3ME) - Isoamyl Alcohol	*C. glabrata*, *C. parapsilosis*, *C. tropicalis*, *C. albicans*, *C. auris*, *C. kodamaea ohmeri*	Phenylethyl Alcohol	*C. auris*, *C. sake 41E*

this information was filtered from a set of studies ((WONGCHOOSUK; LUTZ; KERDCHAROEN., 2009a), (ALVAREZ et al., 2019), (ARRARTE et al., 2017a), (SEMREEN et al., 2019), (HERTEL et

al., 2018b)) obtained through the mentioned literature investigation.

Understanding the VOC profiles emitted by each fungal species is, therefore, the fundamental premise on which the entire work is based, being crucial for the VOC mapping stage carried out by the XAI Ensemble and for the understanding of similarities between the different VOC profiles emitted by the *Candida* species. In this context, the data engineering stage of the *Framework DiagNose.AI* assumes the responsibility of structuring this information, both in the construction of databases that correlate the volatiles to the microorganisms and in its preparation for analysis by Artificial Intelligence models.

2.1.6 Electronic Noses: Digital Olfaction Technology

From ancient China to modern medicine, olfaction has been one of the mechanisms used to identify diseases. Some pathologies, such as those caused by fungi, have "olfactory signatures" that, although not normally identified by the human nose, can be detected by electronic gas sensors (GAS). These sensors are the basis for the construction of Electronic Noses (E-noses), devices developed to mimic the human olfactory capability, identifying and distinguishing complex odors through an array of semi-selective sensors (HAYASAKA et al., 2020; LIMA et al., 2019).

An E-nose analyzes gas mixtures through its array of sensors sensitive to Volatile Organic Compounds (VOCs), identifying patterns related to the components emitted into the air (SAIDI et al., 2020). With a relatively low cost and small size, this technology has been introduced in various fields, such as air quality analysis, monitoring of toxic gases, assistance in medical diagnoses, and quality testing in beverages and food (CHEN et al., 2019; ZHAN et al., 2020; LI et al., 2017). Currently, Electronic Noses are widely used due to their speed, convenience, and objectivity.

However, the raw data generated by these sensors is a complex, high-dimensional temporal signal. The transformation of these signals into a meaningful classification or diagnosis is directly dependent on advanced computational methods for analysis and pattern identification (WANG et al., 2020a).

2.1.7 Artificial Intelligence for Olfactory Pattern Recognition

The interpretation of complex data from an E-nose is fundamentally a pattern recognition challenge. The objective is to identify a unique “olfactory signature” in the sensor signals

that corresponds to a specific condition, such as the presence of a microorganism. Artificial Intelligence (AI), and more specifically Machine Learning, provides the necessary toolkit to automate this task, allowing algorithms to learn how to map VOC signals to a diagnostic result from previously labeled data (GANCARZ et al., 2019).

Different algorithms, such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Ensemble methods (e.g., Random Forest), have been successfully applied to classify E-nose data in various scenarios, such as identifying food quality and predicting diseases (CAYA et al., 2020; KUSBANDHINI; WIJAYA; HIDAYAT., 2021; TURPPA et al., 2019).

Despite their high predictive accuracy, many machine learning models, especially the more complex ones like deep neural networks, operate as "black boxes". Their internal decision-making processes are not inherently transparent, making it difficult for human experts to understand why a particular prediction was made (ADADI; BERRADA, 2018; RUDIN, 2019). This lack of interpretability is a significant barrier to adoption in critical domains like medicine, where reliability, safety, and verifiability are non-negotiable.

This context presents two fundamental challenges that this thesis addresses. First, the need to select machine learning models that are intrinsically suited to the temporal nature of the E-nose sensor data, which leads to the exploration of time-series classifiers. Second, the critical need to overcome the "black box" problem by developing methods for eXplainable Artificial Intelligence (XAI) to ensure the model's transparency and reliability, a topic that will be further explored in subsection 2.1.9.

2.1.8 Time Series

In the world of data, it is often possible, when studying an event, to find datasets where its instances are constructed according to the order of time. This type of ordered sequence of observations over time is described as a Time Series (TS). Examples of this type of data collection include daily closing stock prices, monthly unemployment numbers, quarterly crime rates in a given region, and annual birth rates (all have time as a determining variable). In this sense, it can be said that the essential attribute of a TS is the correlation of its observations. A large part of conventional statistical methods, based on random samples, requires different techniques and are not applicable (WEI, 2013).

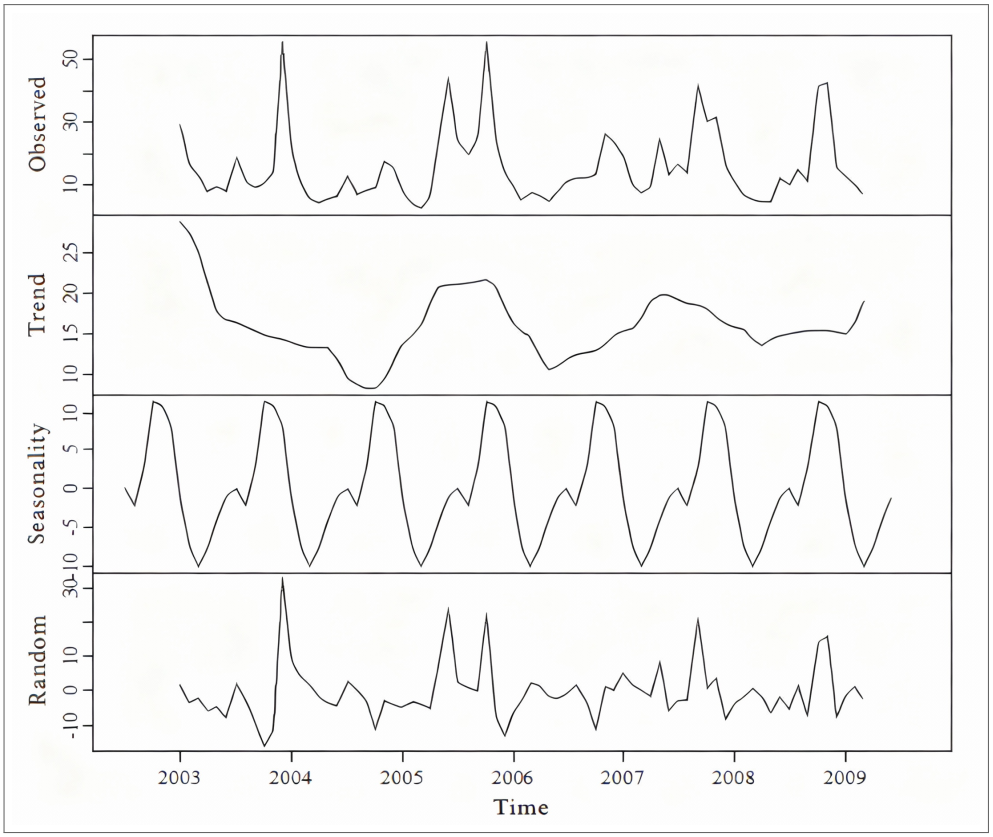
According to Vasconcelos (2022), there are three basic patterns in relation to Time Series: trend, seasonality, and cycle. In the case of a trend, it occurs when it is identified that the

data increases or decreases over time, and there may be scenarios where the trend changes its direction, such as going from a growing trend to a decreasing one. In the case of series that remain constant over time, it is said that there is no trend.

Regarding seasonality, it occurs in cases where the series is influenced by seasonal aspects, such as semester, month, week, or even events with a known period of occurrence. In this way, cyclical changes occur at unknown times. That is, seasonal changes are equivalent to cyclical movements completed within a year, while cyclical variations are complementary cyclical movements in intervals longer than one year (VASCONCELOS, 2022).

In addition to the variations already mentioned, there are also irregular elements, characterized by movements that cannot be explained by trends or cycles. These unregulated variations occur occasionally and influence the increase or decrease of the series' values. The factors responsible for their occurrence can range from occasional customs surcharges to unexpected wars (VASCONCELOS, 2022). To illustrate these behaviors, Figure 5 presents a composition of a time series, highlighting the observed, trend, seasonality, and random values.

Figure 5 – Example of a Time Series decomposition into components of observed values, seasonality, trend, and randomness



Source: WEI (2013)

In this sense, fields such as engineering, sciences, sociology, and statistics have encountered

time series and have been conducting analyses for problem mitigation. In this process, after choosing a suitable group of models, it is possible to estimate parameters, analyze their relevance to the data, and enable the use of the most aligned model to improve the understanding of the procedure that creates the series. With an efficient model implemented and selected, it can be used in different ways, according to the application for which it is being developed (VASCONCELOS, 2022).

These developed models can be used for different purposes, such as, for example, a summary description of the data. In some cases, it is important to identify the presence of seasonal components and remove them, to avoid confusing them with long-term trends, which is usually identified as seasonal adjustment. Still in this same context, other uses of time series models can include the separation or filtering of noise from signals, forecasting future values of a series (predicting sales or population data, for example), and managing future values of a given series by changing parameters (VASCONCELOS, 2022).

Among the main attributes of a time series, besides being stochastic (dependent on or resulting from a random variable), it must also be stationary, which means that the entire process needs to be in harmony with respect to a certain mean and with constant variance. When a series is not stationary, it only allows for the analysis of its behavior at a specific time, not allowing its use for other time fractions, not being very useful for making predictions. In this sense, a time series will rarely be stationary, and it should be converted to this format when necessary (VASCONCELOS, 2022).

Regarding the analysis of time series to verify how future observations can be influenced by the past, the coefficients of the sample autocorrelation function (ACF) and the sample partial autocorrelation function (PACF) are calculated with their respective lags. These coefficients need to alternate between the confidence interval (CI) of the ACF and PACF statistics, except for the first lag (MERELLES et al., 2019). In this sense, the CI can be calculated as follows:

$$CI = \frac{t_y}{\sqrt{N}} \tag{2.1}$$

Where:

CI $=$ confidence interval;

$t_y =$ value of the Student's t-statistic with N-1 degrees of freedom;

N $=$ sample size.

In this context, the ACF provides a basis for linear dependence in the series, that is, the way one observation can influence future ones. On the other hand, the PACF brings the level of direct linear association between instances divided by K periods. Through the ACF, it is possible to perceive the following characteristics of Time Series: seasonality, randomness, correlation, and stationarity. As for the trend, it can be identified through a series graph (VASCONCELOS, 2022).

In the scenario of time series, the order of the data is paramount, since, for this type of data, neighboring values are dependent, and the focus of the analyses is related to these dependencies. In most studies, TS are used to make predictions of the series in real time. This is done by using a past time period to predict a future occurrence. In this work, the time series models will be used for the classification of TS and not for prediction. In other words, Machine Learning (ML) models will be used to learn, using as a basis Time Series already with their respective labels, so that they can be classified when new instances are put to the test.

The sequential nature of the data generated by the E-nose makes the Time Series approach the most suitable for analysis, allowing the capture of dynamic patterns that could be lost in a static analysis. For this reason, the systematic application and evaluation of TS classification models form the analytical core of the *Framework DiagNose.AI*, seeking better accuracy in the distinction between species.

2.1.9 Explainability (XAI) Methods in Machine Learning

The high performance of Machine Learning models has made great strides in recent years. However, there are still some challenges regarding the clarity of their learning process. The lack of interpretability of the results of these algorithms can be a problem in fields where it is essential to understand the results for decision-making, such as healthcare and autonomous systems (RUDIN, 2019). For these reasons, the need arose for the implementation of techniques that, in some way, would allow a better understanding and interpretability of these models known as "black-boxes." This field of research was titled eXplainable AI (XAI), which has as its main objective to solve the interpretability problems of Machine Learning models, extracting insights through their behavior.

In this field of research, a large part of the studies are developed to improve the interpretation of computer vision and Natural Language Processing (NLP) models (RUDIN, 2019;

GUIDOTTI et al., 2019). In the field of images, for example, the methods have shown relevant results, being able to highlight important regions in the images that had greater weight in the decision-making process, without it being necessary to be a domain expert to understand the method's explanation for the result returned by the model. However, the same cannot be said for the explainability methods used for time series classification, where they generally require a greater level of technical understanding of what is being explained(RUDIN, 2019).

Visually, it is simpler for a human to understand an image or text than the signals that are generated by time series. In general, a certain expertise or additional method is needed to explain this data (even before they have gone through any kind of model). A study by Rudin (2019) highlights that more research focused on the explainability of Time Series should be done, especially when it comes to technical systems, the medical domain, and business applications. The same study emphasizes that, since 2019, the number of studies focused on XAI for Time Series has been growing, which strengthens the development of new methods that can fill the existing gaps in this field today.

According to Rudin (2019), Adadi e Berrada (2018), it is possible to categorize XAI methods based on two main criteria, Ante-Hoc and Post-Hoc models. Ante-Hoc groups models that are natively interpretable, as their training structure presents a certain transparency in their decision-making process, such as a decision tree. Despite this, this interpretability does not necessarily make the model explainable, as in some cases, it still requires a certain expertise to understand the report of the training process.

The Post-Hoc method, on the other hand, is not directly coupled to the model, being a separate process from its training flow. It aims to demonstrate and promote new perceptions of how a given model performed its learning after its training phase, without altering its internal structure, as is the case with LIME(RIBEIRO; SINGH; GUESTRIN, 2016). This type of technique is applied when the model in question is a black-box model, that is, when the classification algorithm does not provide a reason for its decision-making or it is not accessible or clearly understandable. Examples of these models are Artificial Neural Networks (ANN), Support Vector Machine (SVM), or a Random Forest (RF). Thus, a Post-Hoc method can be explained as a function g that receives the input of a classifier f, which will be trained with a dataset D.

Furthermore, these models can be subdivided into two more categories, global and local. Global models are characterized by seeking to demonstrate the logic used by them for any input, returning a general explanation of their decision-making process, which is valid for the

entire dataset. Local explanation models, on the other hand, detail the explanation behavior of the model for a particular instance, in which case, the explanations will be given only for a specific instance (RUDIN, 2019).

Within the Post-hoc and Ante-Hoc model techniques, there may still be another subdivision, based on the types of classifiers they can interpret. In these cases, we can name these models as Agnostic and Specific. Agnostic Models, such as LIME (RIBEIRO; SINGH; GUESTRIN, 2016), have the ability to provide explanations for any type of classifier, be it an ANN or a Random Forest. Specific models, on the other hand, are created to bring interpretability to a specific type or a specific family of classifiers. This is the case with Grad-CAM(SELVARAJU et al., 2017), used as an explainability technique for models such as CNNs. By default, Ante-Hoc models are considered all specific, as they are used only to explain their own decision-making process (RUDIN, 2019).

Regarding exclusively the XAI methods created and or adapted for time series, the current state of the art brings a limited series of options, both in quantitative and qualitative terms. Some literature investigations, such as those by Rudin (2019), Guidotti et al. (2019), bring some approaches and their classifications, also highlighting their negative points and some suggestions of spaces not yet filled by the current tools. Among the most prominent XAI techniques for time series in the literature are LIME(RIBEIRO; SINGH; GUESTRIN, 2016), Grad-CAM (SELVARAJU et al., 2017), SHAP (LUNDBERG; LEE, 2017), DeepLIFT (SHRIKUMAR; GREENSIDE; KUNDAJE, 2017), Occlusion Sensitivity (ZEILER; FERGUS, 2014), mWDN (WANG et al., 2018), and SAX-VSM (SENIN; MALINCHIK, 2013). Each of them has particular characteristics and seeks different forms of explainability.

2.1.9.1 The main XAI methods for Time Series

One of the most referenced methods in the literature today is LIME (RIBEIRO; SINGH; GUESTRIN, 2016), or *Local Interpretable Model-agnostic Explanations*, which is a technique that clarifies the predictions of machine learning models, offering local and understandable explanations. It allows users to understand the reasons behind the decisions of complex classifiers, promoting transparency and trust. However, the technique has its limitations. Because it is based on simplified local models, LIME may not fully capture the complexity of the original model, which can lead to explanations that are not fully representative of the model's global behavior. In addition, the selection of representative instances and the generation of

explanations can be challenging in cases with high data dimensionality or when the model is extremely complex. Despite its advantages in making AI models more transparent and reliable, LIME requires caution in the interpretation of its explanations. Users should be aware that the generated explanations are approximate and may not reflect the entirety of the model's decision process. This is particularly important in critical applications, where decisions based on incorrect or misinterpreted predictions can have significant consequences(RIBEIRO; SINGH; GUESTRIN, 2016).

Grad-CAM is a technique that provides visual explanations for the decisions of convolutional neural networks, highlighting the regions of an image that are important for classification. It generates heat maps that indicate the critical areas, increasing the transparency of deep learning models and facilitating the identification of biases in the dataset. Like LIME, this model was also used in the study by Schlegel et al. (2019a) in problems related to Time Series. Despite its advantages, Grad-CAM has limitations, such as the production of coarse heat maps that may not capture fine details and the dependence on gradients, which can be challenging to interpret. These restrictions can affect the accuracy and interpretability of the generated visualizations (SELVARAJU et al., 2017).

Regarding the SHAP method, or *SHapley Additive exPlanations*, it is a unified approach to interpreting the predictions of machine learning models. Based on Shapley values from game theory, SHAP calculates the contribution of each feature to a specific prediction, providing an additive importance measure for each feature. This technique is particularly useful for unraveling the contribution of individual features in complex models, such as decision trees and deep neural networks. However, SHAP has significant disadvantages. One of the main ones is its computational complexity, especially in datasets with many features, as it requires the calculation of all possible groupings of properties. In addition, Shapley values can be misinterpreted, and access to the data is necessary to calculate the values for new data. This can be a challenge in terms of time and computational resources, limiting the applicability of SHAP in scenarios with time constraints or processing capacity(LUNDBERG; LEE, 2017).

Regarding the *Deep Learning Important FeaTures* (DeepLIFT) method, it is a technique designed to decompose the predictions of a neural network and identify the important features that contribute to the final decision. DeepLIFT compares the activation of each neuron with a 'reference activation' and assigns contribution scores based on the difference. This approach allows DeepLIFT to highlight dependencies that may be missed by other methods, offering a more detailed view of how the inputs affect the network's outputs. One of the main advantages

of DeepLIFT is its ability to reveal these dependencies efficiently, in a single backpropagation, after making a prediction. This makes it a valuable tool for the interpretation of deep learning models, especially in fields where interpretability is essential. However, a disadvantage of DeepLIFT is that it can be computationally intensive, especially in networks with a large number of features, which can limit its use in situations with time or computational resource constraints(SHRIKUMAR; GREENSIDE; KUNDAJE, 2017).

Continuing the analysis of these explainability methods, we also have *Occlusion Sensitivity*, which is a visualization technique that helps to understand and interpret convolutional neural networks (CNNs). It consists of systematically hiding different parts of an input and observing how the model's output changes. This allows for the identification of which parts of the input are most important for the model's decision (ZEILER; FERGUS, 2014). An advantage of this method is that it is intuitive and easy to understand, as it directly correlates the importance of different regions of the input with the model's output. However, a disadvantage is that it can be computationally expensive, as it requires multiple passes through the model for each possible occlusion (ZEILER; FERGUS, 2014).

Another prominent method is mWDN, or Multilevel Wavelet Decomposition Network. It is a wavelet-based neural structure proposed for the interpretable analysis of time series. mWDN integrates the advantage of discrete multilevel wavelet decomposition in frequency learning, allowing the fine-tuning of all parameters within a deep neural network framework. This method is notable for its ability to incorporate wavelet-based frequency analysis into deep learning models, offering a new approach to modeling important frequency information that is often overlooked. One of the main contributions of mWDN is the proposal of two deep learning models: the Residual Classification Flow (RCF) and the multi-frequency LSTM (mLSTM), for time series classification and prediction, respectively (WANG et al., 2018). These models use sub-series decomposed by mWDN at different frequencies as input, learning all parameters globally through the backpropagation algorithm. In addition, mWDN has demonstrated excellent performance in extensive experiments with both academic and industrial time series datasets (WANG et al., 2018).

However, despite its advantages, mWDN can present computational challenges, especially when dealing with large volumes of data or complex time series, due to the need to process multiple decompositions at different frequencies. This can require significant computational resources and make the process slower compared to simpler methods(WANG et al., 2018).

Finally, we can mention SAX-VSM, another method widely referenced in the literature in

recent years. This method combines *Symbolic Aggregate approXimation* (SAX) and the *Vector Space Model* (VSM) for the interpretable classification of time series. SAX transforms time series into symbolic sequences, reducing the dimensionality of the data and allowing for a more abstract and manageable representation. VSM is used to classify these symbolic sequences, treating them as documents in a vector space where the frequency of the symbols is analyzed to determine their importance for a given class (SENIN; MALINCHIK, 2013).

This method is particularly useful because it not only provides accurate and fast classification but also offers an interpretable generalization of the class, identifying and classifying time series patterns by their relevance. However, SAX-VSM can be computationally costly during the learning phase, despite its efficiency in classification. In addition, although it offers superior interpretability compared to other algorithms, there are still challenges in symbolic representation that can limit the capture of fine nuances in the time series(SENIN; MALINCHIK, 2013).

In conclusion, the limitations of the discussed methods demonstrate that the search for explainability solutions in time series must transcend the use of heatmaps, which, although useful, may not be the most appropriate form of visual representation for all contexts. It is essential to explore other forms of representation that complement the visual representation, such as detailed textual explanations that aggregate and summarize the crucial information of the series, providing a deeper and more accessible understanding. In addition, new explainability methods must be developed with clear human guidance, aligned with the specific domain of the problem, ensuring that the interpretations are relevant and intuitive for the experts in the area. Finally, the validation of these methods with real, not synthetic, data is fundamental to ensure the applicability and reliability of the techniques in the real world, where data-driven decisions can affect lives or bring potential risks.

Based on this survey, it is possible to extract some important information about the main characteristics of these methods and to understand how new approaches can be developed, taking into account the main limitations of these techniques.

In the context of this research, the emphasis will be on an explainability method aimed at the clinical-medical area, which brings a clear context about the decision-making process of the model used and which can be easily interpreted by a medical or laboratory professional. This new technique will be applied to the model resulting from the AI techniques developed for the data of the *Candida* volatiles obtained by the E-noses of this study. The main idea is to create an explanatory model based on the needs of the clinical context, seeking, from

feedback from potential users, to improve the understanding of the results obtained, bringing both graphical and textual explanations that can clearly demonstrate the correlation of the volatiles identified by the E-nose, the volatiles existing in each *Candida* species of the study, and the final result obtained by the model.

3 RELATED WORK

This chapter provides a brief review of the literature regarding subjects with the greatest similarity to this proposal. In this sense, some works related to the process of classifying and identifying microorganisms, such as fungi and bacteria, using an Electronic Nose and Artificial Intelligence techniques will be highlighted, with an emphasis on techniques related to time series, as well as studies that address the use of Ensemble techniques for explainability.

3.1 SIMILAR WORKS ON CULTURE IDENTIFICATION USING AI AND ELECTRONIC NOSES

In the study conducted by Castro et al. (2022), an Electronic Nose and standard machine learning techniques are used for the clinical identification of *Candida* species cultures. In the work, the authors emphasize the difficulty that currently exists in standard methods for identifying fungi of the *Candida* type and propose a new technique that uses the identification of Volatile Organic Compounds from cultures by means of an Electronic Nose and Artificial Intelligence techniques, a methodology similar to the one used in this project. According to the authors, this type of technique should contribute to assisting the treatment of patients affected by the fungus, allowing for appropriate intervention, reducing complications related to infections, and consequently, the percentage of deaths (CASTRO et al., 2022).

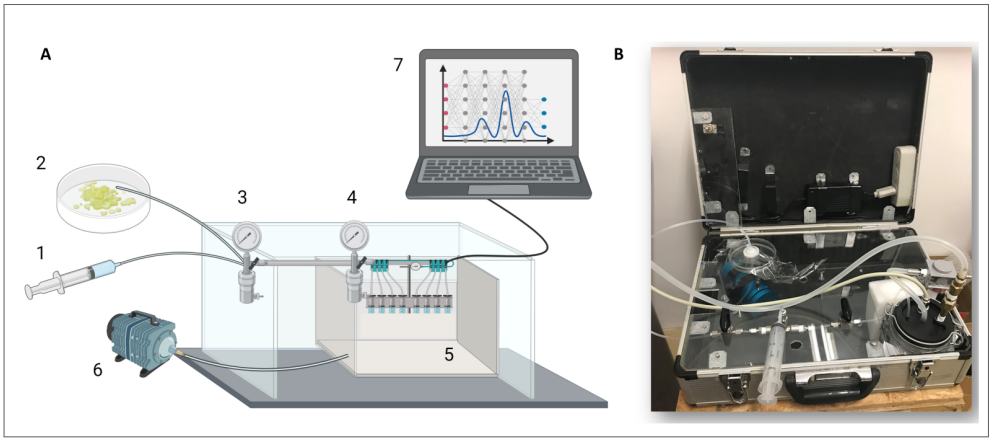
In this context, the use of AI techniques, more specifically Machine Learning, combined with E-noses for fungal identification is still an emerging but very promising method. This type of system analyzes the VOCs that are released by microorganisms after a period of culture in a *Petri* dish, through the use of physicochemical sensors (CASTRO et al., 2022).

In the proposal in question, the E-nose was applied to assist in the identification of samples of three *Candida* species: *C. albicans* (90028 readings), *C. parapsilosis* (22019 readings), and *C. krusei* (6258 readings); all samples were obtained from the ATCC company [1] (CASTRO et al., 2022). In this scenario, the entire reading, analysis, and processing process proposed by the authors is demonstrated in Figure 6.

After collecting the information from the VOC readings generated by the Electronic Nose, the data were prepared for processing by the Machine Learning models. The authors used k-fold cross-validation, with 10-fold repetition. Automated Machine Learning (AutoML) and 28 other

[1] https://www.atcc.org/

Figure 6 – **A) Components of the Electronic Nose used in the study**. The E-nose has a manual injection and suction system (item 1) that collects a certain volume of air from the sample in the *Petri* dish (item 2). Items 3 and 4 represent control valves for air injection and reception through item 1. The sample reading stage begins with the opening of valve 3 and aspiration of the air contained in item 2. After that, valve 2.1 is closed and 4 is opened for the insertion of the sample air for VOC analysis by the chamber (item 5). Once the air has been injected into the analysis chamber, the existing sensors perform the reading and generate data through the reaction that occurs at the moment of interaction of the volatiles with the sensor surface, converting them into digital signals that are sent to the computer system (item 7). Finally, item 6 is an activator responsible for cleaning the chamber, removing accumulated air and injecting filtered air back into the system. **B) Image representing the real Electronic Nose device used in the study**



Source: CASTRO et al. (2022)

models were executed, separated by categories: Naive Bayes, Gaussian Processes, K-Nearest Neighbors, Semi-Supervised Learning, Linear Models, Probability Calibration, Support Vector Machine, Neural Networks, Discriminant Analyses, Decision Trees, and Ensemble Methods (CASTRO et al., 2022). As results, the work indicates that AutoML obtained the best result among all strategies, with an average accuracy of 93%, using a class-balanced model with a weighting strategy. However, the study indicates that further research is still needed applying the Electronic Nose to larger databases, with more varied samples of fungal species. This can be done by improving the aforementioned data acquisition process, which should also imply the training of more robust models, such as deep learning (CASTRO et al., 2022).

Another study that brings great contributions to this research is that of Vasconcelos (2022), which seeks to correctly classify the volatiles emitted by colonies of anemophilous fungi using an Electronic Nose combined with Time Series techniques. For the research, the author built two main datasets, one called "Plate," which corresponds to data collected from *Petri* dishes cultivated with colonies of anemophilous fungi, and another called "Open," which receives data from the ambient air with the propagation of anemophilous fungi colonies in open *Petri* dishes.

In the study, several subcultures of anemophilous fungi species were used, namely *As-*

pergillus spp. (7,424 instances on plate), *Cladosporium spp.* (13,000 instances on plate - 563 open), *Fusarium spp.* (8,452 instances on plate - 1,213 open), *Penicillium spp.* (5,109 instances on plate), and *Rhizomucor spp.* (1,189 instances on plate - 1,189 open), all marked according to their collection method. For the training, these data were organized into three different bases: the first base was called "Plate," with all the data from the Plate-type readings; the second was called Plate_TR_Open_TS or Pl_TR_Ab_TS, as it uses the Plate base as training data and defines the "Open" readings as test values; the third base was called Plate_Open or Pl_Ab, as it gathers values from the open plate readings for both training and testing (VASCONCELOS, 2022).

For the training and execution of the models, the author used the k-fold cross-validation technique with 10 repetitions for most of the bases, collecting the metrics of Accuracy, Sensitivity, Specificity, and Time elapsed. The resulting values were calculated based on the means and standard deviation obtained from the 10 iterations for each model. The chosen models were MrSEQL, ROCKET, Arsenal, HIVE-COTE V2, TSF, cBOSS, kNN, RISE, and WEASEL. Of these, the one that obtained the best result for the "Plate" base was MrSEQL, with values of 94.5% average accuracy and a standard deviation of 5.2 between iterations. The Plate_TR_Open_TS base did not obtain relevant values in relation to the results of the models, where the one with the highest accuracy was Arsenal, with only 58.8% and a standard deviation of 5.3.

Finally, for the Plate_Open base, the results were slightly similar to the "Open" base, with values of 94.9% accuracy for the best-placed model, Arsenal. However, the author recommends the use of the model that came in second place, ROCKET, with 94.3%, because, according to his analysis, the loss of performance in the metrics is outweighed by the gain in execution time, of almost 8x. Thus, based on the results obtained, the author considers the results for the Plate_Open and "Plate" databases to be satisfactory, with very positive values for both. However, future works were pointed out regarding the expansion and diversification of the database, the investigation of new models, and the creation of a low-cost device to measure the presence of fungi in environments (VASCONCELOS, 2022).

Another very interesting use of Electronic Noses and Time Series is in the work of Nascimento (2022), which uses the combination of techniques to identify bacteria of the type *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Enterococcus faecalis*, and *Escherichia coli* in infected wounds. The study used as a basis 14 bacterial cultures, divided as four for the species *Staphylococcus aureus* (5,340 readings), another four for *Pseudomonas aeruginosa*

(5,359 readings), four more for *Escherichia coli* (3,341 readings), and two for *Enterococcus faecalis* (981 readings), which resulted in a final set of 15,021 readings, derived from the volatiles emitted by each culture and collected by the *E-nose*.

These collections were carried out after four distinct periods, at 6h, 9h, 24h, and 78h, in order to evaluate whether there are relevant improvements in the results at shorter or longer time intervals. In the organization of the base, the author defined two more variables for the base: one referring to the label of the species itself and another to the culture time used for collecting the information by the *E-nose*. In addition, due to inconsistencies related to the number of readings per cycle, it was also necessary to level the data, aiming to maintain homogeneity in the data (NASCIMENTO, 2022).

For the process of classifying the information, the author defined four Time Series models: K-Nearest Neighbors (KNN), specifically the distance-based version using Dynamic Time Warping (DTW), Time Series Forest, HIVE-COTE 1.0, and InceptionTime. The entire training, validation, and testing phase, as in the previous works, also underwent k-fold cross-validation, with $k = 10$, applying the metrics of Accuracy, F1-Score, Precision, Recall, and Specificity (in addition to the standard deviation of the iterations).(NASCIMENTO, 2022; LIN et al., 2019).

Among all the evaluated models, *InceptionTime* performed best in the validation and testing phases, with 98.99% accuracy for the former and 94.65% for the latter. In addition, *InceptionTime* also obtained better sensitivity and specificity in both stages, with values of 94.65% and 97.01% (respectively) for the validation stage and 93.68% and 96.71% (respectively) for the testing stage (NASCIMENTO, 2022).

Thus, based on the results obtained, the author describes the identification of bacteria through the use of Electronic Noses and time series techniques as possible, based on the satisfactory results of the presented models. In this scenario, emphasis is also given to the possibility of making the identification of bacteria in wounds faster, more accurate, and automatic using this methodology compared to existing ones. As future contributions, the importance of obtaining a more robust and diversified database is highlighted, in addition to the possibility of using new sensors, more appropriate for identifying the VOCs emitted by the studied species. This would allow for a more specific collection by the E-nose and a possible improved classification of the chosen model, also in real environments (hospital institutions) (NASCIMENTO, 2022).

With a more comprehensive approach, Mota, Teixeira-Santos e Rufo (2021) conducted a Systematic Literature Review regarding the detection and identification of fungi through

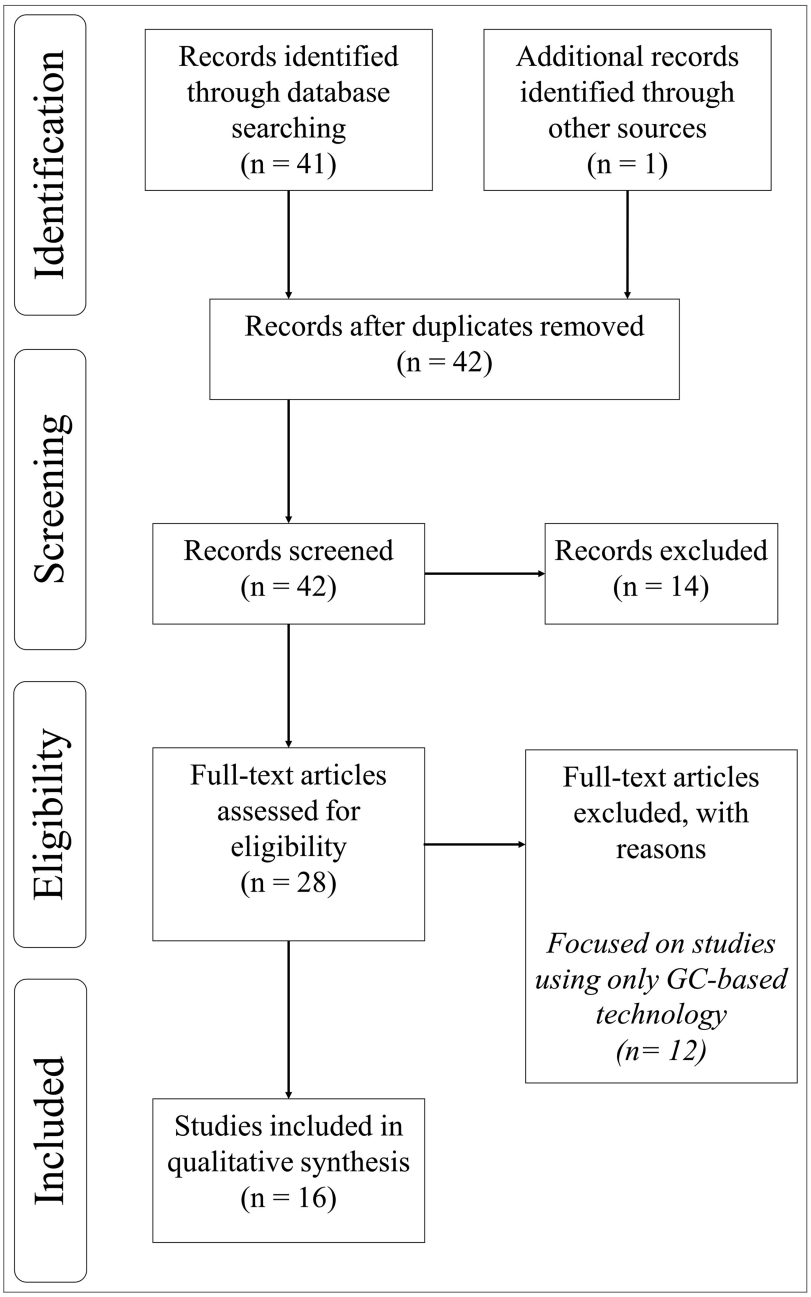
Electronic Noses technology. At the beginning of the discussion, the authors highlight the importance of researching new methods for identifying fungi, given that the most used ones today, such as MALDI-ToF, have high associated costs, requiring a laboratory and specialized human resources to perform the correct identification and are rarely used outside the clinical environment. In contrast, new technologies in biosensor engineering are enabling the creation of portable, faster, and cheaper devices (E-Noses), using the concept of identifying Volatile Organic Compounds. Consequently, these advantages have expanded the possibilities of applications and research, benefiting various areas of knowledge, including the clinical field. Several challenges were faced, which brought the first good results for the identification of respiratory diseases.

Regarding the applied methodology, the study was built based on the PRISMA methodology, using PubMed as the main search engine, for publications made up to January 6, 2020. The search *string* was assembled with the combination of the keywords *nose* or *E-nose*, combined with any term originating from the word *fungus*, resulting in (("electronic nose*" OR eNose) AND fung*). As inclusion criteria, the following points were used: 1) Study originating in English; 2) The study sought the identification of fungi, and; 3) The study performs the identification by means of Electronic Nose. In this context, as exclusion criteria, articles that did not have as a premise the use of sensors or pattern recognition were discarded (MOTA; TEIXEIRA-SANTOS; RUFO, 2021). Figure 7 demonstrates the search flow carried out by the authors, using the PRISMA methodology as a structural basis.

As shown, a total of 16 studies resulted from the filtering phases. Of these, the vast majority focused on the food industry and the fungus of the genus *Penicillium*. In general, nine different E-Nose devices were used, the two most common being the PEN3 (*Airsense Analytics Inc., Germany*) and the BH114 (*Bloodhound Sensors Ltd, UK*), with three studies each. Among them, the main technical differences are related to the sensors used for VOC identification. Different types were used, such as Conducting Polymer (CP), Metal Oxide Semiconductor (MOS), Quartz Crystal Microbalance (QCM), Surface Acoustic Wave (SAW) sensors, or a mix of different technologies. This discrepancy can imply differences in the identification results of each approach. Another factor that can interfere is that the vast majority of solutions use ambient air for baseline measurements, which can cause some kind of contamination by external VOCs at the time of analysis in different environments (MOTA; TEIXEIRA-SANTOS; RUFO, 2021).

As conclusions of the study, the authors reported that the identification of fungi through

Figure 7 – Flow of the bibliographic search conducted by Mota, Teixeira-Santos e Rufo (2021) using the PRISMA methodology as a basis



Source: MOTA; TEIXEIRA-SANTOS; RUFO (2021)

the use of Electronic Noses devices proved to be very positive, highlighting its use in bakery products, cherry tomatoes, dry wheat grain, rice grains, peaches, and even in urban trees. In addition, this type of technology also proved to be efficient in the early identification of ochratoxigenic species in grape samples and in the classification of microorganisms causing root canal infections. However, the study emphasizes the need for optimization of experimental conditions and standardization of the detection method (MOTA; TEIXEIRA-SANTOS; RUFO, 2021).

Another study with great synergy with the present work is that carried out by Aksebzeci et al. (2010), on the classification of microorganisms in the root canal using Electronic Nose. According to the authors, an efficient and rapid identification can help health professionals make more assertive decisions about treatment forms, such as the use of different types of irrigants, antibiotics, and intracanal medications. The difficulties associated with the cultivation process and the complexity in isolating the prevalent anaerobic pathogens end up forcing professionals to apply experimental treatments to patients. Thus, the study sought to identify seven types of pathogens commonly found in root canal infections, through the information of the volatiles collected by an E-Nose.

In this research, the dataset was formed by 5 repeated samples of 7 different types of species *(Candida albicans, Candida glabrata, Fusobacterium nucleatum, Porphyromonas gingivalis, Pseudoramibacter alactolyticus, Streptococcus sanguinis, Enterococcus faecalis)* in 4 repetitions. At each concentration, a set of 35 examples was classified with 3 different methods of discriminant analysis (there was no use of AI, only the PCA method.). With the aim of specifying an ideal profile for the use of the Electronic Noses in the application, the authors used 3 different approaches to test the sensor responses. Three different sensor baseline values were also used to obtain the normalization of the sensor responses. Considering that the number of sensors (32 carbon-black polymer composite sensors) is comparatively larger than the number of collected samples, the impact of two different dimensionality reduction methods on the classification performance was also investigated (AKSEBZECI et al., 2010).

As main results, the study highlights that the quadratic type discriminant analysis surpasses the other varieties of this same method. It was also observed that the classification performance is reduced whenever the concentration drops and that the models in which the minimum sensor reading values in the sample were accepted as baseline bring a better performance in the classification process. The results showed that the Electronic Nose was able to accurately distinguish between different types of bacteria and fungi present in root canals, with an accuracy rate of more than 90%. The authors conclude that this technique can be used as a complementary tool to conventional microbiological diagnosis in endodontics, allowing for a faster and more accurate identification of the microorganisms present in infected root canals (AKSEBZECI et al., 2010).

With a greater emphasis on the state of the art of Electronic Noses, the article by Chen et al. (2019) reports a literature review on GAS sensor *arrays* (E-noses), highlighting the main technologies used for their design, the AI models most used in the literature for applications

related to these sensors, and the main areas where they are being applied. The study also makes a correlation between the different types of gas sensors and the different detection methods used by them. Among the main ones are thermal (catalytic), mass, electrochemical, optical, semiconductor, and surface acoustic wave sensors. Another highlight of the work is the detail about intelligent gas sensor *arrays*, demonstrating how the detection and discrimination of volatile organic compounds and other compounds that can be found in the air is done.

The author explains that this interaction between the gas sensor and the material is done through the capture and storage of changes in resistance, current, or frequency, according to the different types of measurement metrics. One of them is sensitivity, a parameter of the gas sensor that describes the device's response in relation to the target gas molecules at a certain concentration. One of the most referenced *arrays* in the study is the metal oxide (MOX), where its composition, positive and negative points, and use are reported (CHEN et al., 2019).

Another interesting related work in the context of E-Noses is that of Peng et al. (2018). The study provides a general overview of the mechanisms and definitions involved in the construction of an Electronic Noses. The authors highlight that this type of application has already been used in areas such as medicine and diagnosis, food production, and environmental monitoring and can be separated into three main parts: planning and construction of the gas sensor *array*, feature extraction from the signal generated by the "smell fingerprint" of the gas, and pattern recognition regarding the olfactory characteristics. The authors' proposal aims at the construction of an efficient solution for the last topic.

In this sense, the work uses different classifiers on the data derived from gas sensor *arrays* that collected Carbon Monoxide, Methane, Hydrogen, and Ethylene. In this process, through comparisons and results, the authors highlight the use of a set of techniques related to *Deep Convolutional Neural Networks* (DCNNs), called GasNet, as the most successful in the process of classifying gas types, despite being more commonly applied to image classification. They maintain that this was the first approach to use this type of network in gas classification, demonstrating better results in relation to widely used classifiers in the literature, such as SVM and MLP (PENG et al., 2018).

Still in this same context, the study by Wongchoosuk, Lutz e Kerdcharoen. (2009b) reports very concisely the use of an Electronic Nose for the identification of human odors, specifically from the armpits. The study used metal oxide sensors from the TGS line (TGS813, TGS825, TGS2602, TGS880, and TGS822) for the capture and identification of VOCs released by the armpits of the study participants. Adjustments were made on the maximum and minimum

resistance in relation to the average values of their 10 neighboring data points and also an adjustment of the hardware and software parameters related to humidity.

For the analysis of these volatiles, the authors used statistical models such as the paired T-test and a principal component analysis (PCA) algorithm, which was implemented with the intention of identifying patterns related to the volatiles of each study participant. As final results, the work indicated that it is possible to identify and distinguish the volatiles of different people through the odor of the armpits, even if the individual has used deodorant. The sensors that had the most response to the volatiles were the TGS2602 and the TGS822 (WONGCHOOSUK; LUTZ; KERDCHAROEN., 2009b).

In addition, according to the authors, it is also possible to explore the use of E-noses as a new biometric identification marker, given that the PCA results showed very distinct patterns for the two individuals who participated in the study. However, it is understood that, for a better understanding of the effectiveness of this process, it is necessary to carry out the study with more people, creating a more precise statistical validation (WONGCHOOSUK; LUTZ; KERDCHAROEN., 2009b).

In the context of this study, Table 2 presents a brief summary and comparison of the main similar works and the present project, highlighting the classification method (whether it is time series or not), the highest accuracy achieved, the type of E-nose (sensors used), whether or not statistical analysis was performed, and the volatiles analyzed in each study. Based on the analysis of the main results of each study, it is possible to observe that there are currently no studies directed towards the analysis and identification of the *Candida* species used in this study (*C. albicans, C. parapsilosis, C. krusei, C. haemulonii, C. kodamaea ohmeri, and C. glabrata*) using the most updated Time Series techniques. In addition, the present project obtained a success rate superior to all other studies, which indicates a great potential for its implementation in an operational environment.

Table 2 – Comparison of the characteristics of similar works on VOC identification with E-nose with this project

Work	Uses Time Series?	Best Accuracy	E-nose Type	Performs statistical analysis?	Analysis of which VOC type?
(CASTRO et al., 2022)	No	93%	10 sensors, with manual injection and suction of volatiles	No	Fungi - *C. albicans, C. parapsilosis, and C. krusei*
(VASCONCELOS, 2022)	Yes	94.9%	10 sensors, with automatic injection of volatiles	No	Fungi - *Aspergillus spp., Cladosporium spp., Fusarium spp.*
(NASCIMENTO, 2022)	Yes	94.65%	10 sensors, with automatic injection of volatiles	Yes	Bacteria - *Staphylococcus aureus, Pseudomonas aeruginosa, Enterococcus faecalis, and Escherichia coli*
(MOTA; TEIXEIRA-SANTOS; RUFO, 2021)	N/A	–	Compares various versions	No	Systematic Review
(AKSEBZECI et al., 2010)	No	75%	32 gas sensors	Yes	Fungi: *Candida albicans, Candida glabrata, Fusobacterium nucleatum, Porphyromonas gingivalis, Pseudoramibacter alactolyticus, Streptococcus sanguinis, Enterococcus faecalis*
(CHEN et al., 2019)	N/A	–	Compares various versions	No	Systematic Review
(PENG et al., 2018)	No	>90%	8 gas sensors	No	Carbon Monoxide, Methane, Hydrogen, and Ethylene
(WONGCHOOSUK; LUTZ; KERD-CHAROEN., 2009b)	No	–	5 gas sensors	Yes	Armpit odor
This project	**Yes**	**97%**	**10 sensors, with automatic injection of volatiles**	**Yes**	**Fungi - *C. albicans, C. parapsilosis, C. krusei, C. haemulonii, C. kodamaea ohmeri, and C. glabrata***

Comparative analysis between the studies

3.2 RELATED WORK ON XAI ENSEMBLES IN TIME SERIES AND BIOMEDICAL APPLICATIONS

Several recent studies have proposed the use of XAI ensembles in different domains. For instance, Rezk et al. (REZK; EL-GHAFAR; HASSAN, 2024) developed a voting ensemble of machine learning models for heart disease prediction and applied SHAP and LIME to interpret its outputs. The combination allowed for greater trust by identifying the most influential clinical features. However, their approach focused on explaining an ensemble of predictive models using tabular data, rather than creating an ensemble of the XAI methods themselves to produce a single, more robust explanation for temporal data.

Ganguly and Singh (GANGULY; SINGH, 2023) developed an explainable ensemble learning framework for diabetes management prediction. Their approach combined multiple machine learning models with post-hoc interpretability techniques to enhance transparency and clinical reliability. While the ensemble improved predictive accuracy and provided useful feature-level explanations, the study did not incorporate user-centered interpretive interfaces or semantic contextualization of the explanations, limiting its direct applicability in patient-oriented healthcare settings.

Shtayat et al. (SHTAYAT et al., 2023) proposed an explainable ensemble deep learning framework for intrusion detection in industrial Internet of Things environments. Their approach integrated multiple neural models and interpretability tools to enhance both detection accuracy and model transparency. However, despite its strong analytical performance, the system lacked user-centered interpretive interfaces, limiting accessibility for non-technical operators and domain experts.

Huang et al. (HUANG et al., 2022) developed a deep ensemble learning framework for human activity recognition using wearable sensors. The proposed approach combined multiple neural network architectures to enhance classification accuracy and robustness across diverse activity patterns. While the ensemble demonstrated improved generalization and feature learning capabilities, the study did not include integrated interpretability mechanisms or user-oriented visualization interfaces, which may hinder broader practical adoption.

Esser-Skala and Fortelny (ESSER-SKALA; FORTELNY, 2023) investigated the interpretability of biologically inspired deep neural networks, emphasizing the need for reliable explanation techniques in bioinformatics. Their analysis provided valuable insights into the transparency of complex models but did not incorporate ensemble-based interpretability or human-centered

validation components.

Hou et al. (HOU et al., 2024) conducted a comprehensive survey on self-explainable AI methods for medical image analysis, highlighting architectures with built-in interpretability. While these models promote intrinsic transparency, the study underscored that most existing solutions remain architecture-dependent and lack integration with complementary XAI techniques or user-oriented explanation frameworks.

Zhang et al. (ZHANG et al., 2023) applied interpretable machine learning approaches to metabolomics data, successfully identifying biomarkers associated with Parkinson's disease. Despite demonstrating high predictive performance and meaningful biological insights, their framework did not employ ensemble-based explainability or provide semantic, domain-aligned explanations for clinical end-users.

Theissler et al. (THEISSLER et al., 2022) reviewed explainable AI methods for time series analysis, emphasizing the scarcity of ensemble strategies and the importance of incorporating human-in-the-loop paradigms in safety-critical domains. Their findings suggest that future XAI systems should better integrate interpretability, reliability, and usability considerations for real-world adoption.

Although XAI methods such as LIME, SHAP, and Grad-CAM are widely adopted, their isolated application to time series data remains challenging due to the temporal dependencies and high dimensionality inherent to such signals (THEISSLER et al., 2022; ROJAT et al., 2021; SCHLEGEL et al., 2019b). Recent studies have begun to explore ensemble-based XAI frameworks as a means of leveraging the complementarity among interpretability techniques (REZK; ELGHAFAR; HASSAN, 2024; HUANG et al., 2022; SHTAYAT et al., 2023). However, few works have specifically investigated these strategies for complex multivariate and temporal sensory data, such as those generated by electronic noses. Table 3 summarizes how the main related studies compare to the objectives of the present work.

Our work positions itself as an original contribution by proposing, for the first time in this domain, an XAI ensemble focused on interpreting chemo-temporal data associated with the identification of *Candida* species in yeast or blood broth. In addition to being a pioneer in the joint application of these techniques to E-nose data, the proposed approach stands out for its integration with a textual explanation module based on a scientific VOCs database, providing end users with a deeper understanding of the model's outputs.

Table 3 – Comparison between XAI ensemble approaches and the proposed method. The table summarizes the main characteristics of representative studies addressing explainability across different domains. Prior works have applied XAI ensembles in areas such as image analysis (ZOU et al., 2022), industrial intrusion detection (SHTAYAT et al., 2023), wearable sensor data (HUANG et al., 2022), cardiovascular health (REZK; EL-GHAFAR; HASSAN, 2024), and diabetes management (GANGULY; SINGH, 2023), yet without focusing on multivariate time-series classification with human-centered explanations. Other approaches explored individual XAI techniques in biological data analysis (ESSER-SKALA; FORTELNY, 2023; ZHANG et al., 2023) or developed self-explainable models (HOU et al., 2024), but lacked ensemble strategies or semantic interpretability. Theissler et al. (THEISSLER et al., 2022) provided a comprehensive review, identifying ensemble-based XAI for time-series data as an underexplored area. The proposed method differs by integrating complementary XAI techniques into an ensemble designed for chemo-temporal data, enriched by a VOC semantic base, textual explanations, and usability-oriented design, addressing interpretability, robustness, and accessibility gaps not fully covered in previous studies.

Work	**Data Type**	**Domain**	**XAI Ensemble**	**Textual Explanation**	**VOC Semantic Base**	**Time-Series Focus**	**Human-Centered Design**
L. Zou et al. (ZOU et al., 2022)	Image	Respiratory Infections	Yes	No	No	No	No
Shtayat et al. (SHTAYAT et al., 2023)	Network traffic	Industrial IoT	Yes	No	No	Partial	No
Huang et al. (HUANG et al., 2022)	Multisensor	Human Activity Recognition	Partial	No	No	Yes	No
Ganguly & Singh (GANGULY; SINGH, 2023)	Clinical tabular	Diabetes Management	Yes	Partial	No	No	Partial
Rezk et al. (REZK; EL-GHAFAR; HASSAN, 2024)	Clinical tabular	Cardiovascular Health	Yes	Partial	No	No	Partial
Esser-Skala & Fortelny (ESSER-SKALA; FORTELNY, 2023)	Biological sequences	Bioinformatics	No	No	No	No	No
Hou et al. (HOU et al., 2024)	Medical images	Healthcare	No	Yes	No	No	Partial
Zhang et al. (ZHANG et al., 2023)	Metabolomic signals	Biomedical / Metabolomics	No	No	No	Yes	No
Theissler et al. (THEISSLER et al., 2022)	Mixed (survey)	Time-series / General	No	No	No	Yes	Yes (review)
XAI Ensemble for VOCs (This Work)	**Multivariate chemo-temporal**	**Health / Biolabs**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Comparative analysis between related works on XAI ensembles in different domains and the proposal of this study.

This approach can also be generalized to other applications involving densely multivariate time series and complex sensory data, such as in artificial olfaction systems, environmental sensor networks, and metabolic biosensors.

3.3 POSITIONING THE RESEARCH IN RELATION TO EXISTING LITERATURE

The in-depth analysis of related works reveals a scenario of great potential, but also of important limitations. On one hand, the state of the art consistently demonstrates that the combination of Electronic Noses (E-noses) with Artificial Intelligence techniques is a viable and promising approach. Studies such as those by (CASTRO et al., 2022), (NASCIMENTO, 2022), and (AKSEBZECI et al., 2010) have achieved high accuracy rates in the classification of various microorganisms, validating the fundamental premise that the profiles of Volatile Organic Compounds (VOCs) contain distinct "signatures" that can be learned by machine models.

However, when confronting this technological promise with clinical reality, the limitations of current diagnostic methods become even more evident. The reliance on blood culture, with its response time of days and sensitivity of approximately 50%, not only delays the start of targeted antifungal therapy but also contributes to the empirical and often inappropriate use of broad-spectrum drugs. This scenario negatively impacts clinical outcomes, increases the risk of antimicrobial resistance, and raises hospital costs. The urgency for a method that breaks with this paradigm — being at the same time fast, sensitive, and accessible — is not only a technological demand but an imminent need of modern medical practice.

Despite the advances in classification, a critical analysis reveals the first major gap in the literature: *methodological fragmentation*. Most studies present themselves as one-off and isolated applications. There is a lack of a *complete and replicable methodological Framework* that covers the end-to-end workflow: from the definition of an experimental protocol, through data engineering, to modeling and, crucially, the interpretation of the results. This absence of a systematic approach makes it difficult to compare studies and generalize findings.

The second gap is the *scarcity of public and well-documented databases*. Most works use private datasets, which prevents the reproducibility of experiments and the benchmarking of new algorithms by the scientific community. In the specific context of identifying *Candida* species through VOCs, this scarcity is particularly notable, representing a significant barrier to the advancement of the area.

Finally, the third and most significant gap lies in the *superficiality of Explainable Artificial*

Intelligence (XAI) in this domain. The application of XAI to data derived from VOCs is incipient and, when present, limited to individual techniques whose results are difficult for the target audience to interpret. Common outputs, such as *heatmaps*, although useful in other contexts, are not intuitive for health professionals who need to understand *why* a certain VOC profile led to a specific diagnosis. There is a lack of *textual and semantically rich explanations* that connects the model's prediction to existing biochemical knowledge. In addition, many of these methods were not designed for the dense, multivariate, and temporal nature of the signals from an E-nose, being adaptations of techniques created for computer vision or tabular data, which limits their effectiveness and reliability. A proposal for an *XAI ensemble approach* designed to overcome these barriers, increasing the robustness and clarity of the interpretations, was not identified in the literature.

Given these identified gaps — clinical, methodological, data, and explainability —, the *Framework DiagNose.AI* proposed in this thesis was conceived to directly address each of these deficiencies. It responds to fragmentation by proposing a complete workflow; it contributes to the lack of data by detailing the construction of two new databases; and, most importantly, it attacks the main computational gap by developing a novel *Ensemble XAI* architecture. Thus, this thesis is positioned not only as an application of AI but as a methodological contribution that aims to strengthen and mature the area as a whole.

4 THE *DIAGNOSE.AI FRAMEWORK*: DEVELOPMENT AND METHODOLOGY

This chapter details the architecture and validation of the DiagNose.AI Framework, the main methodological contribution of this thesis. The methodology presented here is the result of a rigorous iterative research process, which consolidated a systematic and complete solution for the identification of microorganisms from the analysis of Volatile Organic Compounds (VOCs).

The work follows on from initial explorations on the use of the Electronic Noses and AI in the identification of *Candida* spp. (CASTRO et al., 2022), but significantly expands the approach by proposing a complete Framework. A fundamental step for this expansion and for the robust validation of the methodology was the international collaboration established with the College of Medicine at the University of Cincinnati, in the United States. This partnership was crucial to enable one of the most important phases of the research: experimentation with patient blood broth samples. Conducting these experiments at an international reference center, using a versatile and portable reading device developed as part of this thesis, allowed not only the construction of a novel clinical database but also the consolidation of the Framework in a highly complex and relevant scenario.

The methodology was implemented and validated through an iterative process, culminating in a functional proof of concept. The developed software artifacts, as well as the details of the computational configuration used for the experiments, are documented in Appendix A of this work.

The content of this chapter is structured to detail each of the components that were designed and integrated to form the DiagNose.AI Framework. As illustrated in Figure 8, which presents an overview of the workflow, the methodology covers everything from sample collection to the issuance of a final explainable report. The following sections will describe in detail each of the steps consolidated in this Framework:

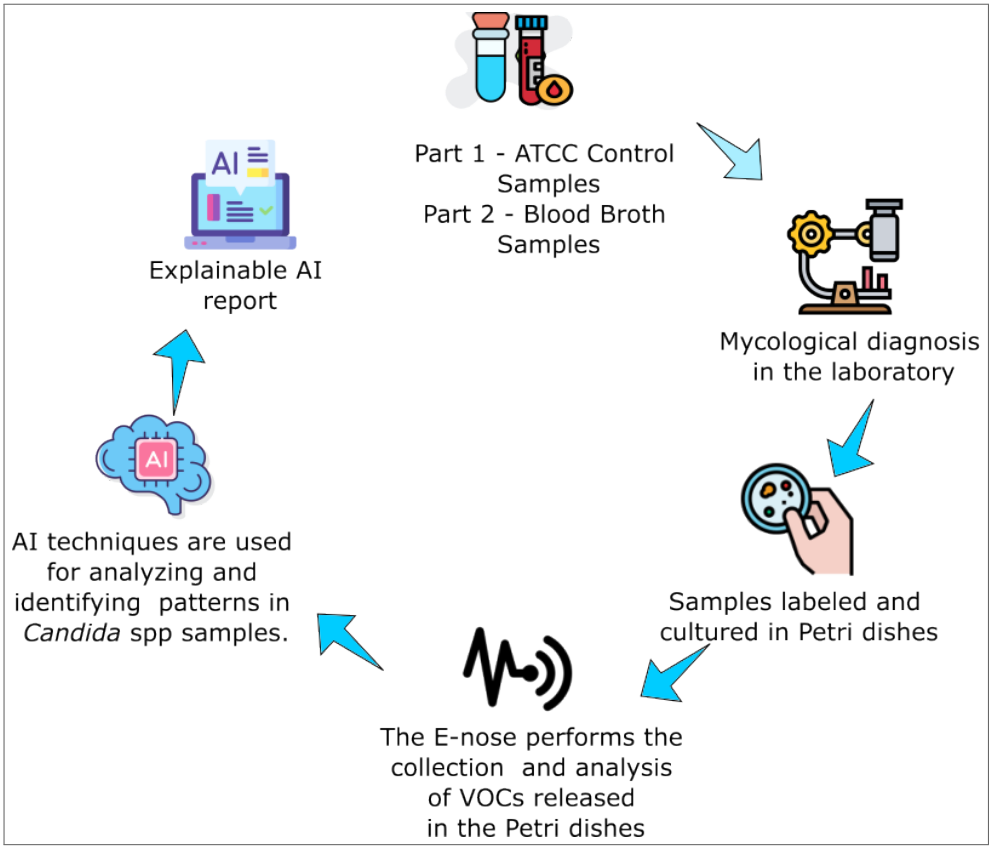
(I) The structure of the reading devices and the data acquisition protocol;

(II) The methodology for data engineering and preparation of the databases;

(III) The predictive modeling approach with an emphasis on time series;

(IV) And, finally, the design and validation of the *XAI Ensemble* explainability architecture.

Figure 8 – Overall workflow of the DiagNose.AI Framework. The methodology ranges from the preparation and collection of data from samples (ATCC and blood broth) with the Electronic Noses, through preprocessing and analysis by AI models, to the generation of a final explainable report on the species identification.



Source: Created by the author

4.1 COMPONENT I: THE DATA ACQUISITION PROTOCOL

The validity of any Artificial Intelligence system is intrinsically dependent on the quality and consistency of the data with which it is trained. In the domain of diagnosis through sensors, where the variability of samples and environmental conditions can introduce significant noise, the absence of a standardized collection process represents one of the main barriers to the reproducibility and reliability of the results. Recognizing this fundamental gap, the first component of the DiagNose.AI Framework is the design and consolidation of a rigorous Data Acquisition Protocol. This section details the proposed methodology, which was designed to ensure the generation of high-fidelity and consistent volatile compound data, serving as the fundamental foundation for the subsequent stages of modeling and explainability. The protocol was validated in multiple scenarios, using different reading devices and sample types to ensure its robustness and versatility.

Table 4 – Description of the functions of each sensor

Sensors	Main Function
TGS826	Ammonia Detection
TGS2611-E00	Methane Detection
TGS2603	Detection of odors and air contaminants (High sensitivity to the amine series and sulfurous odor gases and high sensitivity to food odors)
TGS813	Detection of combustible gases (High sensitivity to methane, propane, and butane)
TGS822	Detection of Solvent Vapors (High sensitivity to alcohol and organic solvent)
TGS2602	Detection of air contaminants (High sensitivity to gaseous air contaminants)
TGS823	Detection of Organic Solvent Vapors (High sensitivity to organic solvent vapors, such as ethanol)

Sensors used in the Electronic Nose to identify the volatiles emitted by the gases generated by the *Candida* spp. species.

4.1.1 The Reading Device for ATCC Culture Samples (Suitcase)

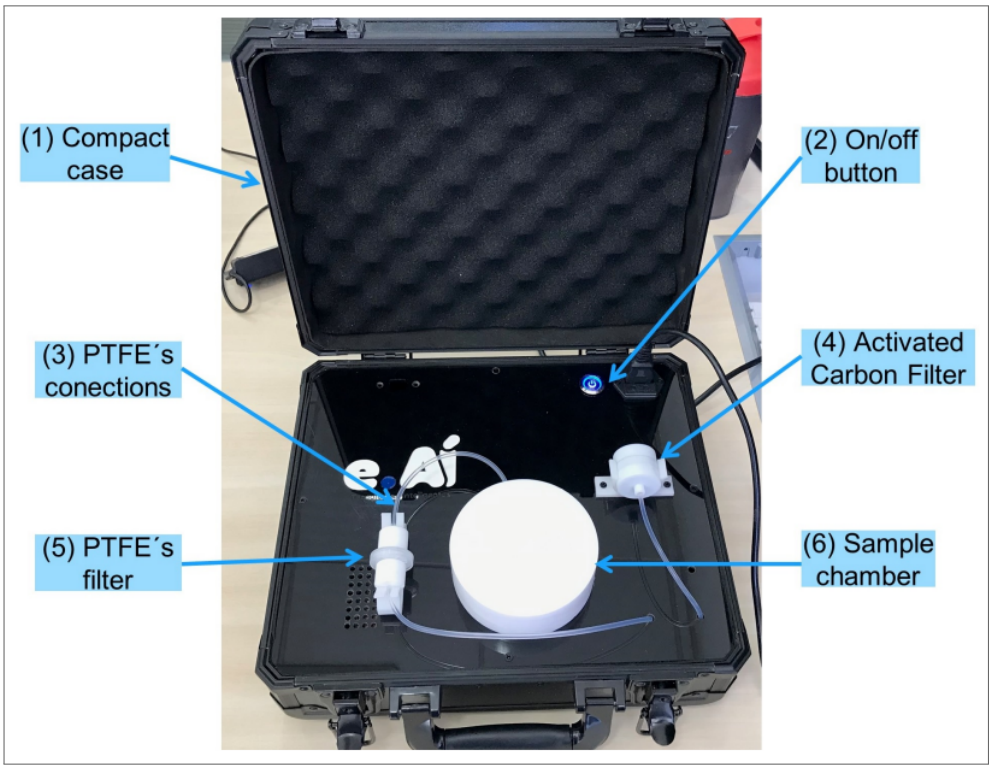
In parallel with the construction of the theoretical basis and the structuring of the problem, the first steps for the construction of the solution were carried out. The first version of the database was built from ATCC control samples[1]. These samples were used as reference strains by the Medical Mycology Laboratory/UFPE, and were then marked and cultivated on Petri dishes for analysis by the Electronic Nose, developed in partnership with the Northeast Regional Center for Nuclear Sciences/UFPE. The *E-Noses* performs the identification of the "olfactory fingerprints" released by the fungi through Volatile Organic Compounds. In this process, as mentioned earlier, the *E-Nose* uses 10 different categories of sensors, seven of them from the manufacturer Figaro Engineering Inc. (TGS826 (Ohm), TGS2611 (Ohm), TGS2603 (Ohm), TGS813 (Ohm), TGS822 (Ohm), TGS2602 (Ohm), TGS823 (Ohm)), and the other three are temperature (°C), pressure (kPa), and humidity (%) sensors used to analyze possible interferences of these parameters on the behavior of the samples. a summary of the main functions of the sensors used in the device is in Table 4

To provide greater agility in transporting the device, it was built and adapted inside a compact case, with adequate sealing and structure to support all the necessary elements for the operation of the Electronic Nose. In the box, in addition to the sensors coupled to an air chamber on the inside and an on/off button, there is a pump responsible for the

[1] https://www.atcc.org/about-us

suction/injection of gases or air into the chamber, a control valve, an air filter with activated carbon, and finally, a simple chamber to insert the Petri dish and collect the volatiles emitted by the reactions of the microorganisms. All connections between the components and the chamber surfaces are made of polytetrafluoroethylene (PTFE) due to its non-stick characteristic and low coefficient of friction, facilitating cleaning and preventing the permanence of volatiles between the suction and purge cycles. Figure 9 presents the Electronic Nose device used in the experiments with the ATCC culture samples.

Figure 9 – Electronic Nose device (Suitcase) used in the experiments with ATCC samples: (1) The Electronic Nose is packaged in a compact box; (2) It is activated by the on/off button; (3) All connections are made of PTFE; (4) It has an activated carbon filter and (5) a PTFE filter; (6) The sample chamber is also made of PTFE. For collection, the Petri dish is placed in the sample chamber (6), the chamber is closed and the E-nose is turned on (2). With the air already filtered (5), the device performs the aspiration in the Chamber for 20s, the air passes through the PTFE connections (3) and goes to the sensors that are on the inside of the case (1). After that, a stabilization phase occurs for 60s, followed by a purge phase, which performs the cleaning for another 60s (using the activated carbon filter - (4)). Three readings per second are made during this process.



Source: Author

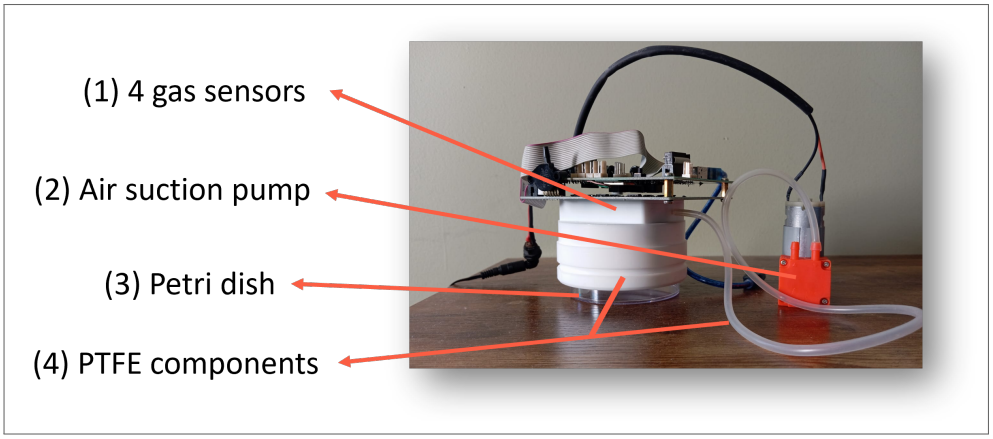
4.1.2 The Portable Reading Device for Validation in Blood Broth (Prototype)

The second version of the Electronic Nose was developed to facilitate the transport of the equipment between countries, allowing the creation of more heterogeneous databases, with

samples from different regions. In this context, this more compact version of the E-nose was used for the collection and reading of blood broth samples from infected patients, managed by the University of Cincinnati Medical Center. The prototype has seven distinct sensors: four gas sensors of the MQ line (MQ-135, MQ-3, MQ-7, and MQ-138) and three environmental sensors for monitoring temperature (°C), pressure (kPa), and humidity (%), used with the same objective as the previous version (*Suitcase*), that is, to check for possible environmental interferences in the readings.

Following the same logic as the *Suitcase* version, the prototype was built entirely with PTFE materials, avoiding the adhesion of compounds to the surfaces and preventing cross-contamination between readings. In addition, the portable version also has an air pump, responsible for transporting the VOCs from the *Petri* dish to the internal gas sensors of the E-nose. On the other hand, due to its portability, the purge phase must be performed manually, alternating between the *Petri* dish with the blood sample and another plate containing activated carbon. Figure 10 illustrates the new version of the device.

Figure 10 – Electronic Nose device (Prototype) used in the experiments with blood broth: (1) region where the gas sensors are embedded; (2) pump used for VOC aspiration; (3) region intended for placing the Petri dish; (4) PTFE components to avoid cross-contamination. For collection, the *Petri* dish with the sample is positioned below the E-nose (3), the safety cabinet is closed (where the experiments are performed) and the E-nose is activated by the system. The device performs the aspiration (2) for 60 s, the air passes through the PTFE connections (4) to the internal sensors of the prototype (1). Then, a stabilization phase of 120 s occurs, followed by a purge phase, in which the plate with the sample is replaced by a plate containing activated carbon, performing the cleaning for another 60 s.



Source: Author

4.2 COMPONENT II: THE DATA ENGINEERING METHODOLOGY

After the acquisition of raw data by the experimental protocol, the subsequent and equally critical step of the DiagNose.AI Framework is its transformation into a structured, high-quality knowledge asset. Sensor data, in its original form, is unsuitable for the direct training of machine learning models, requiring a robust engineering process to extract significant signals and format them appropriately. In addition, the literature in the area suffers from a notorious scarcity of public datasets, which hinders reproducibility and the benchmarking of new approaches. To address both gaps, the Data Engineering Methodology described in this section was conceived. This contribution covers not only the development of a pre-processing pipeline for microorganism VOC data but also the construction and characterization of two new databases, which represent a valuable resource for the scientific community.

4.2.1 Construction of the Culture Database (ATCC Samples)

In the first stage, the material cultivated by means of ATCC samples is labeled with its respective species (*C. albicans*, *C. glabrata*, *C. haemulonii*, *C. kodamaea ohmeri*, *C. krusei*, *C. parapsilosis*), cultivated on *Petri* dishes containing Sabouraud Dextrose Agar medium (see Figure 11) and taken for reading by the Electronic Nose (using the protocol explained in the subsection 4.2.2), resulting in the generation of the database. The culture of fungi in the laboratory is a fundamental technique in biological and industrial research, allowing for the in-depth investigation of the biochemical and physiological properties of different fungal species, as well as the production of biomolecules with wide application. Recent studies highlight that fungal cultures have been used for decades for the production of food, enzymes, and other biochemical compounds, and that their use is growing with technological advances in fungal biotechnology (ROTH; WESTRICK; BALDWIN, 2023).

However, it also involves several associated costs, from the acquisition of materials and equipment to the maintenance of ideal growth conditions for the fungi. The effective management of these costs is fundamental to ensure the financial viability of fungal research in the laboratory. Among the main costs associated with the culture of fungi in the laboratory, the following stand out:

- **Culture media**: culture media are essential for the growth of fungi in the laboratory,

providing the necessary nutrients for the metabolism and reproduction of fungal cells. Culture media can be purchased ready-made or prepared from specific ingredients, such as malt extract, peptone, agar, and sugars. The costs vary depending on the type and quality of the culture media.

- **Laboratory equipment**: fungal culture requires specific laboratory equipment, such as incubators, ovens, magnetic stirrers, Petri dishes, and pipettes. This equipment can be expensive and requires regular maintenance to ensure proper functioning.
- **Fungal species**: the acquisition of fungal species involves costs, depending on the source and rarity of the species. Some fungal species are protected by laws and require special licenses for acquisition.
- **Electricity**: many laboratory equipment, such as incubators and ovens, consume significant electrical energy, which can increase operational costs.
- **Time and labor**: the process of culturing fungi in the laboratory is time-consuming and requires specialized labor to maintain ideal growth conditions. This includes salary, scholarship, and training costs for laboratory technicians.
- **Waste disposal**: the proper disposal of waste generated during the fungal culture process, such as used culture media and contaminated cultures, involves additional costs, such as the acquisition of equipment and chemicals for treatment and proper disposal.

In summary, the culture of fungi in the laboratory is an important technique in biological and industrial research, but it also involves several associated costs that must be properly managed to ensure the financial viability of the research. Careful planning and consideration of the costs involved are fundamental to maximize the return on investment in fungal culture in the laboratory.

4.2.2 VOC Collection from ATCC samples

After laboratory culture, the VOCs are aspirated by the Electronic Nose with different cultivation times (24h, 48h, and 72h), in order to increase the heterogeneity of the data and allow for better generalization by models in the future. This aspiration at different times also

Figure 11 – Examples of a *C. albicans* sample (URM8368) used to create the database (isolate tested using the culture in an *in vivo* and *ex vivo* model). Cultivation performed on a *Petri* dish using Sabouraud Dextrose Agar culture medium.

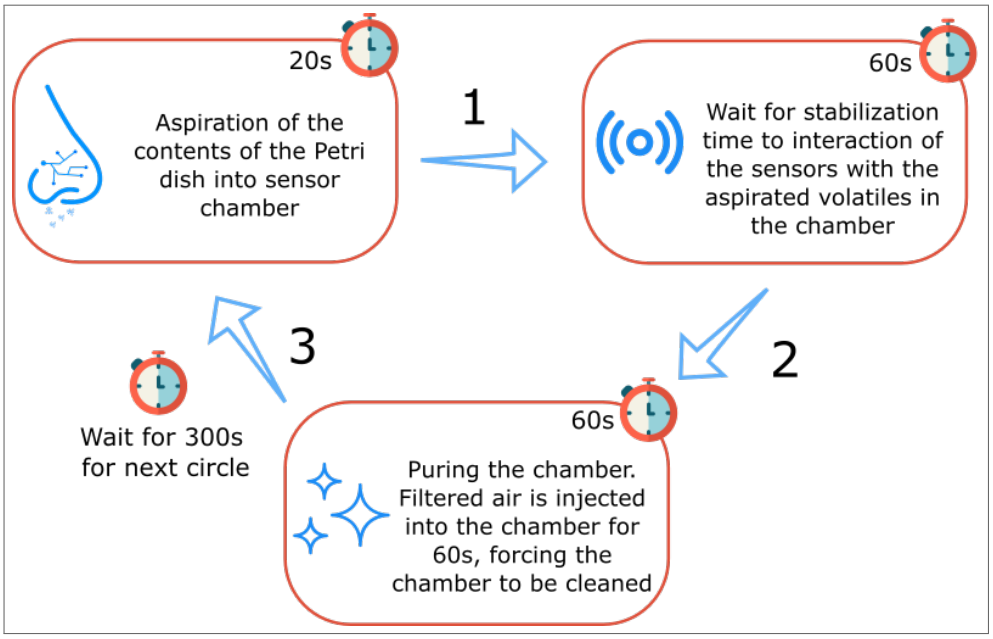
Source: Author

aims to identify whether it is possible to obtain accurate results more quickly, which is of great importance to assist health professionals in decision-making.

For each collected sample, the *E-Nose* performs a collection protocol based on three categories of actions: aspiration, stabilization, and purge (cleaning step) (as seen in Figure 12), where the completion of all three characterizes the conclusion of a cycle. For each sample, a volume of three readings per second is collected for 20 seconds in the aspiration phase, for 60 seconds in stabilization, and another 60 seconds in the cleaning phase, totaling an average of 420 readings per cycle on each sensor (for each sample, a predefined number of cycles is executed). Considering that numerous samples of the same species are necessary to obtain diversity in the data (so that the AI models can satisfactorily learn the patterns of each species), a relevant amount of data was collected in this first stage, with 20,189 instances of *C. albicans*, 19,068 of *C. glabrata*, 6,989 of *C. haemulonii*, 7,067 of *C. kodamaea ohmeri*, 17,255 of *C.*

krusei, and 20,234 of *C. parapsilosis*, totaling 90,802 samples collected in approximately 514 cycles with cultures on different days. It is common to have cycles of different sizes, due to a reading inconsistency in the *E-Nose*. To solve this, it was necessary to combine the sizes of the cycles, explained in more detail in the Sample Classification Process section.

Figure 12 – *E-Nose* collection cycle. (1) Chamber suction step (2) Sensor stabilization step (3) Chamber cleaning (purge)



Source: Author

After the construction of the first version of the database, the need was identified to perform a data analysis, seeking to observe the existence of behaviors or indications of patterns for the different sensors related to each of the species. In addition, this initial verification was important to identify cleaning and/or restructuring strategies for the base, to enable its use by the learning models.

It is important to highlight that this first stage of experimentation aims to create a conceptual basis for new experiments with blood, seeking to validate the hypothesis that the E-nose, combined with AI techniques, is capable of identifying the patterns emitted by *Candida*. Additionally, in an operational setting, the analyzes should also be conducted in a controlled environment (the collected material should be taken to a sterile environment). This is intended to avoid potential noise in the data.

4.2.3 Analysis and processing of ATCC samples

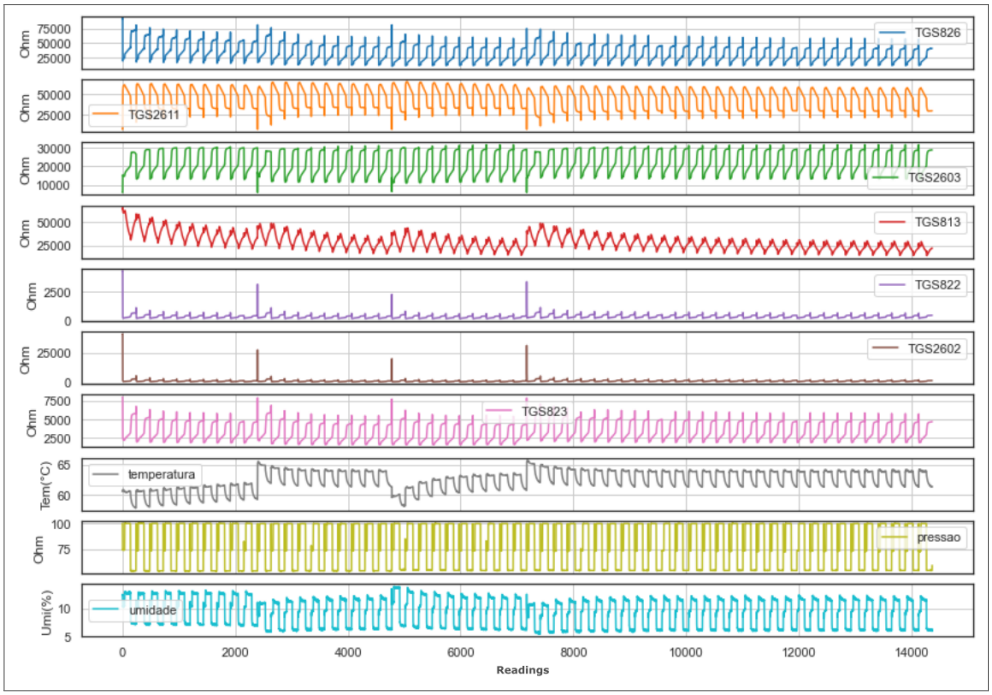
After data generation, a descriptive analysis was performed to better understand its behavior and which AI models might be more suitable to identify the patterns generated by the samples. For this, it was first necessary to perform an analysis and visualization of the data to get an idea of how they relate in relation to each sensor for each *Candida* spp. collection. From there, a new database was built with the dataset of all collected species, only with the sensors considered significant and with the addition of new columns to label the samples in relation to their species and cultivation time. Another important point in this information visualization stage was the use of UMAP (*Uniform Manifold Approximation and Projection*) and PCA (*Principal Component Analysis*), dimensionality reducers, which helped to better understand the clustering of the data. In this sense, as initial steps for the pre-processing and visualization of the information, four relevant points were verified in relation to the data:

- Whether all the sensor data for the same species have a similar behavior;
- Whether there are differences in information between the same species at different collection times;
- Whether there is a predominance of a sensor per species;
- Whether there is a clear division between the data and how they are grouped.

For the analysis of the first point, graphs were generated with data from all sensors related to the collections of each *Candida* species to be analyzed. In these, the wave patterns of each collection were observed, following the chronological order of reading, visualized in Figure 13 for the *C. albicans* data.

As can be seen, each of the sensors has a specific wave pattern, varying in well-defined intervals. There are some reading peaks in some regions that may signal detection errors by the sensors, indicating the presence of possible *outliers*. The pressure and humidity sensors have an almost constant reading cycle, not interfering at any time with the reading pattern of the other gas sensors. The temperature sensor, despite oscillating at some points, also does not interfere with the reading of the other devices. This may be an indication that the alteration of these parameters does not cause, in this case, any interference in the captures of the other sensors, and they can be removed from the analysis.

Figure 13 – Readings data of each sensor over time for the *C. albicans* samples



Source: Author

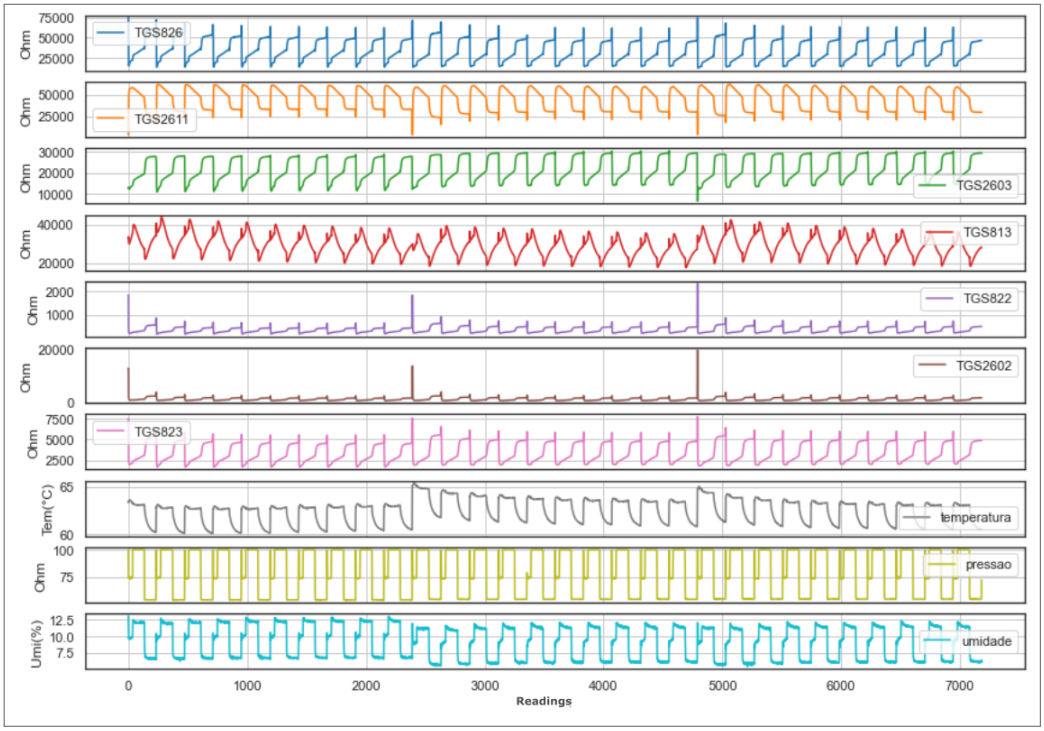
Another important point for this initial study is the identification of differences between data of the same species, but for different collection times. This point helps to visually point out if there are significant differences between the readings performed with cultures of different days, because the sooner the reading patterns are identified, the better the decision-making will be. Figure 14 shows reading data of one and two days for the species *C. krusei*.

As can be seen in Figure 14a and Figure 14b, there is a small distinction in the amplitude of the waves in relation to some sensors from one day to the next. This demonstrates that these devices have a difference in resistance between the volatiles of day 1 and the volatiles of day 2. One hypothesis is that the concentration of gases released by this species changes over time, decreasing in some cases and increasing in others, contributing to widening the differences in the patterns between the different days.

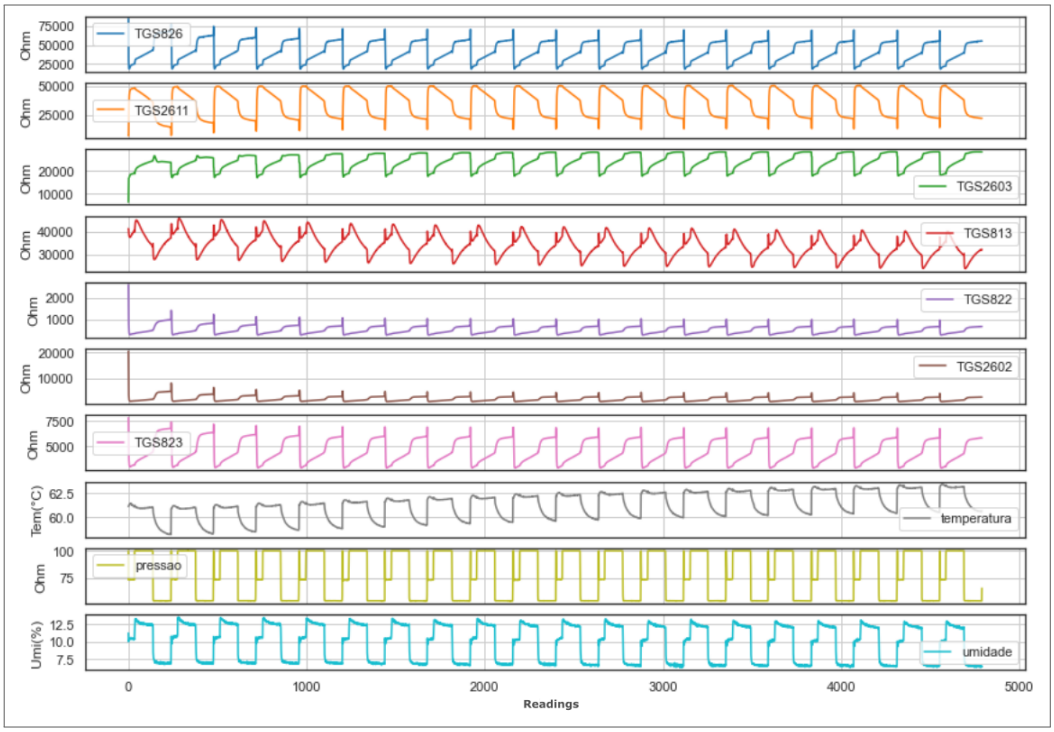
The third point is the possibility of the predominance of a certain sensor per species. This can indicate which sensor is able to differentiate itself more in relation to each species of *Candida*, contributing to the distinction of patterns and the selection of the characteristics used in the database consumed by the classification models. Figure 15 represents this information.

Each of the graphs highlighted in Figure 15 portrays the behavior of the sensors based on the resistance caused by the gases emitted by each species at the time of reading by the Electronic Nose. Seeking to identify a predominance of a sensor over the species, it is noted

Figure 14 – Reading data of *C. krusei* on different days.



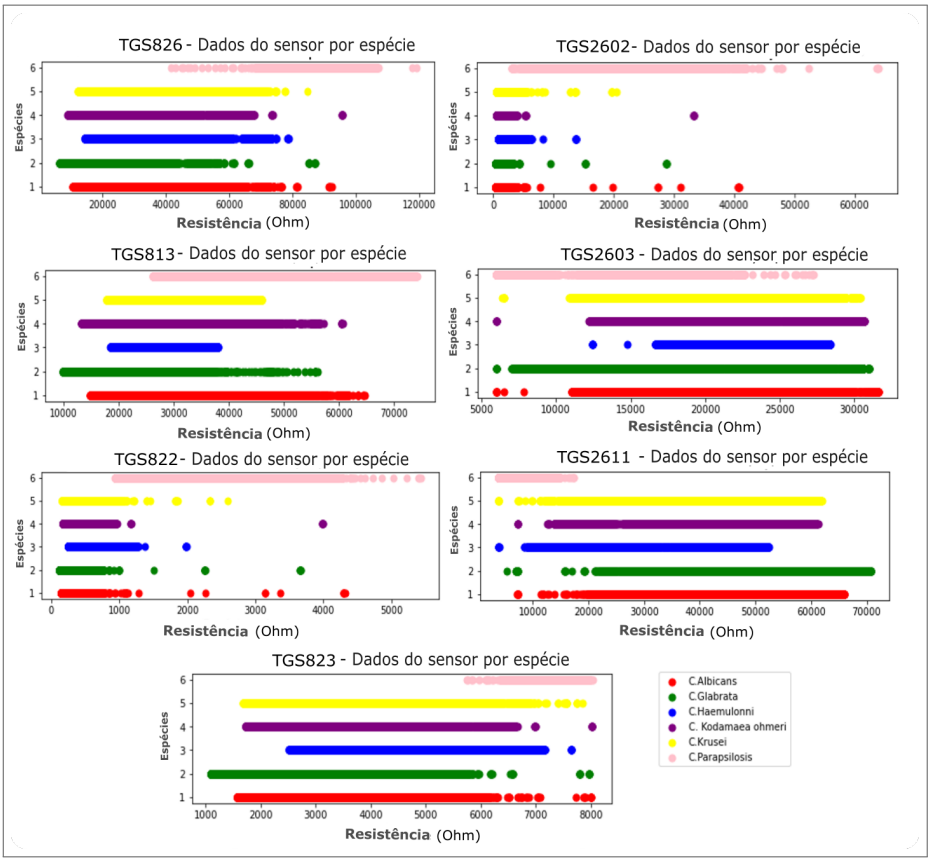
(a) Reading data of *C. krusei* after 1 day of cultivation.



(b) Reading data of *C. krusei* after 2 days of cultivation.

Source: Author

Figure 15 – Resistance data for each sensor by *Candida* species. In this case, it is possible to identify the sensitivity of each device with the readings of each species. For example, the TGS2602 sensor (Detection of air contaminants - upper right corner of the figure) has a higher resistance to *Candida parapsilosis* than to the other species. Thus, it is possible to say that this sensor is more sensitive to the volatiles of this species than to the volatiles of the other *Candida*.



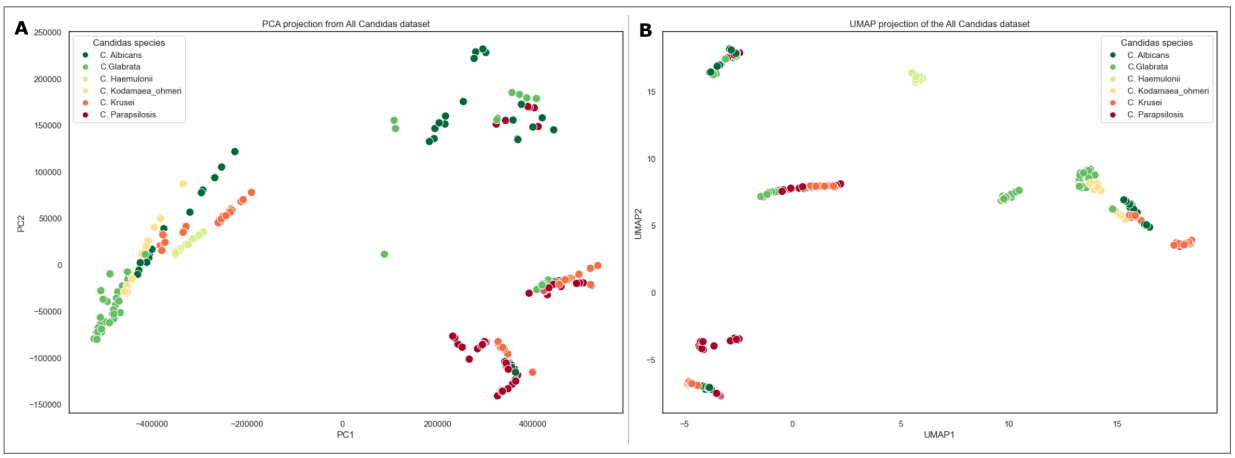
Source: Author

that the TGS2602 and TGS822 sensors have a greater number of readings spread over different levels of resistance in *C. parapsilosis*, with the values of the other types of *Candida* in very similar regions to each other, but very different from *C. parapsilosis*. The opposite occurs with the TGS2611 and TGS823 sensors, where the other candidates have more distributed resistances and *C. parapsilosis* is more focused in one region. This all shows that, in fact, some sensors have a predominance in relation to some species. However, for each of them to identify different levels of resistance in relation to each of the others, all the reading values end up being relevant, because together they become important characteristics for the identification of patterns by the models.

After analyzing the data of each species and sensor separately, the need was identified to understand how the entire dataset was grouped. For this, two dimensionality reduction techniques were applied: PCA (Principal Component Analysis) and UMAP (Uniform Manifold Approximation and Projection). In the case of PCA, according to Abdi e Williams. (2010), its

main objective is to extract relevant information from a tabulated dataset and convert it into a new set of orthogonal variables, called Principal Components. In this sense, it is possible to display patterns of similarity in the instances and variables as components of a graphic map. On the other hand, UMAP, according to McInnes, Healy e Melville (2018), is an innovative dimensionality reduction technique that is based on a theoretical framework of Riemannian geometry and algebraic topology, which makes the results derived from its reduction scalable and easily used in real data situations. Unlike PCA, it performs dimension reduction in a non-linear way, trying to keep similar cases close and different cases separate. In this study, for both techniques, a two-dimensional reduction was applied, which can be analyzed in Figure 16.

Figure 16 – Two-dimensional representation of the Principal Component Analysis (a) and the Uniform Manifold Approximation and Projection (b)



Source: Author

Analyzing the two projections, we can see small groups formed by each species. In the case of PCA, the difference in patterns of *C. parapsilosis* from the other *Candida* is evident, as its points are well dispersed from other data groups, with only a few samples separated from the main group. This demonstrates that this species has very particular characteristics and can probably be distinguished by machine learning models. Regarding the other species, it is noted that there is not much visible overlap between them. An evident problem is the existence of the same group in different parts of the PCA image. Some models may encounter barriers to distinguish this behavior. In the graph generated by UMAP, it is already possible to see a separation of the data, with groupings of species being made in different regions of the graph, without the existence of much overlap. This is explained by the way UMAP treats the reduction, through algebraic topology and similarity measures. It is important to highlight

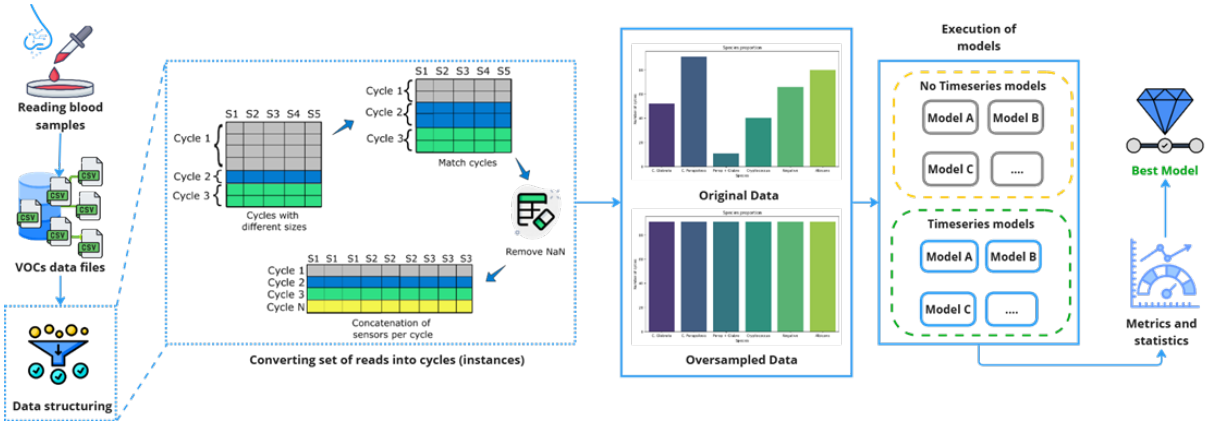


Figure 17 – Workflow of the *Candida* identification process using Electronic Nose and AI models. Blood aliquots are collected and analyzed by the device, which captures the volatile organic compounds (VOCs). The data are stored and preprocessed, including restructuring of the cycles and balancing by oversampling. The sets (original and balanced) are used to train traditional classification and time series models. The validation is done by Repeated K-Fold cross-validation with 10 repetitions. The best model is selected based on metrics such as accuracy, F1-score, and specificity, and is then deployed for real-time identification.

that, despite not being grouped in the same region of the graph, the species with similar characteristics end up being close to each other and, because they have very different reading averages within the same species – due to the differences in sensor reading – the same species may contain data that are not very close, since this method does not seek its resizing based on the principal components, but rather on similarity measures.

Finally, knowing how the data are arranged and grouped, the base was prepared for use by Time Series models, modified to 2 dimensions, one of the few supported by most of the models in this segment. From there, the experiments with the models began, which will be detailed in the next sections.

4.2.4 Construction of the Blood Broth Database

All stages of sample collection and data generation were carried out under the IRB protocol 2020-0313, in compliance with all necessary safety standards and protocols. For the blood culture experiments, Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) mass spectrometry was employed to rigorously evaluate and confirm the species identification of the samples prior to their utilization in the study. To better illustrate the two methodologies presented, Figure 18 shows the data collection and generation phase, while Figure 17 delineates the solution development phase.

The first methodological stage (Fig. 18) involved the preparation, storage, and analysis of

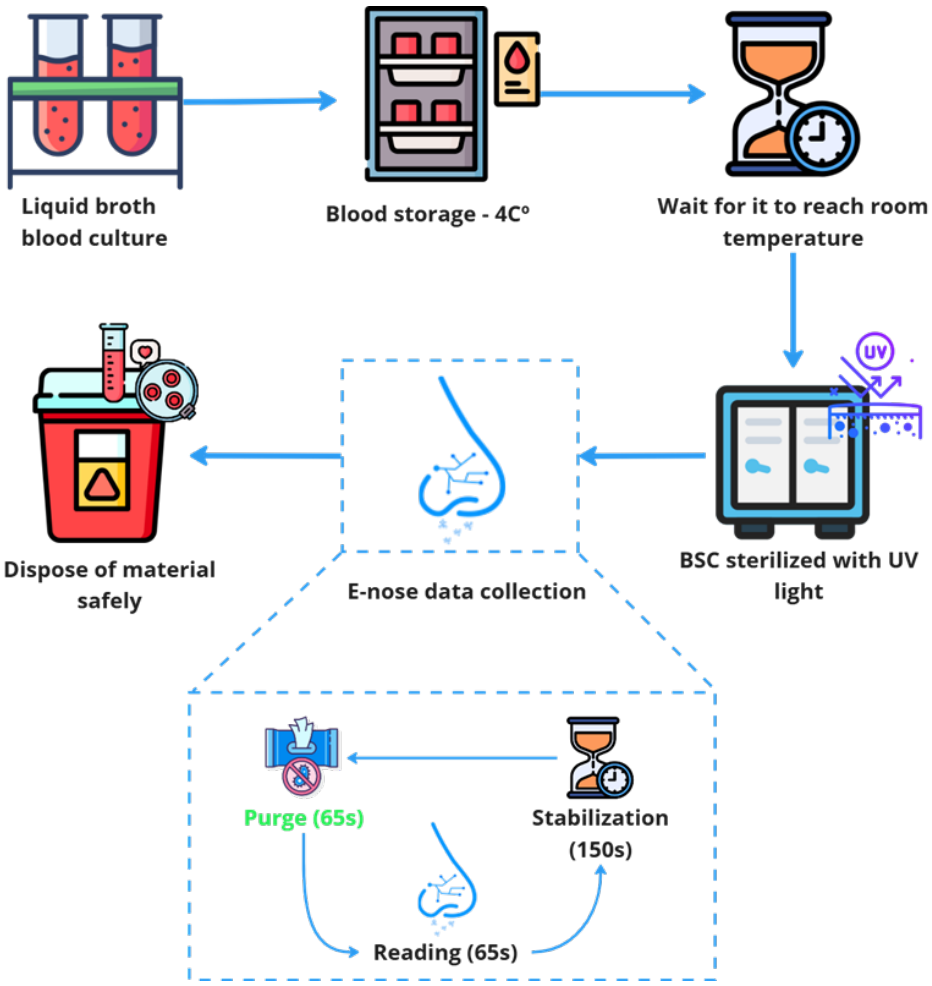


Figure 18 – Experimental setup for in situ analysis of blood samples. Step 1: Collection of the blood culture broth. Step 2: Storage of the sample at 4°C. Step 3: Wait for the sample to reach room temperature (25°C). Step 4: Sterilization of the collection environment. Step 5: Execution of the collection cycle (reading, stabilization, and purge). Step 6: Disposal of the used material.

the samples in a controlled environment, ensuring the standardization of the procedures and the reliability of the obtained results. Aliquots of residual broth were collected from clinical blood cultures with previously known subculture results, tested at the [clinical laboratory of the University of Cincinnati Medical Center]. The aliquots were stored in sterile tubes under refrigeration at 4°C, with a maximum period of 48 hours to preserve the integrity of the emitted VOCs. In total, 14 samples were collected, which generated 45,674 readings, whose subculture results were used to label the data analyzed by the Electronic Nose, defining the target variable of the dataset. This information was essential for the training of the AI models, allowing the identification of which VOC profiles corresponded to each type of microorganism. The samples included both fungal organisms isolated in subculture (*C. glabrata*, *C. albicans*, *C. parapsilosis*, or *Cryptococcus neoformans*), as well as mixtures of *C. glabrata* and *C. parapsilosis*, in addition to negative samples (without isolated fungi).

4.2.5 Preparation for reading and conducting the experiments

After storage and before the readings, the samples were kept for one hour at room temperature, allowing for thermal stabilization and minimizing the influence of temperature variations on the volatilization of the compounds. To ensure ideal conditions at the time of measurement, the temperature was checked (around 25°C) before each reading. The non-*Candida*, negative, and mixed samples were used to test whether the models could identify potential mixed infections, uninfected samples, and other types of infection besides those caused by *Candida*.

Additionally, to ensure optimal sensor performance, the Electronic Nose was turned on approximately one hour before the start of the readings. This pre-heating phase allowed the sensors to reach a stable operating temperature, reducing fluctuations that could impact the detection of VOCs. During the experiments, the minimum temperature recorded by the sensors was approximately 45°C. Environmental factors such as humidity (between 60% and 75%) and pressure (between 78 mBar and 82 mBar) were also monitored but were not included in the final dataset, being only considered for the control of possible environmental interferences. No additional calibration of the sensors was performed, and only the raw values were converted to the measurement units used in the analysis.

The experiments were conducted in a biological safety cabinet (BSC), ensuring a controlled environment free of external contamination. The BSC was sterilized with UV light for about 30 minutes before each experimental cycle, eliminating microbiological contaminants. As the sensors of the Electronic Nose are highly sensitive to external volatile compounds, the use of 70% alcohol was avoided, as its vapors could interfere with the detection of VOCs and compromise the accuracy of the measurements. To avoid cross-contamination, each sample was individually placed in a disposable Petri dish and positioned in the reading chamber.

The collection process included the stages of Purge, Reading, and Stabilization. In the Purge, the Electronic Nose was placed over a Petri dish containing activated carbon and activated for 65 seconds, allowing the device to be cleaned and residual VOCs to be removed. Then, the Reading took place, in which the plate with carbon was replaced by the plate containing the blood sample, and the reading was performed for another 65 seconds. After the reading, the device entered Stabilization, remaining inactive for 150 seconds to allow for the stabilization of the VOCs. This process was repeated between 10 and 40 times, with variations to evaluate interferences in the amount of VOCs after the exposure of the sample to the air.

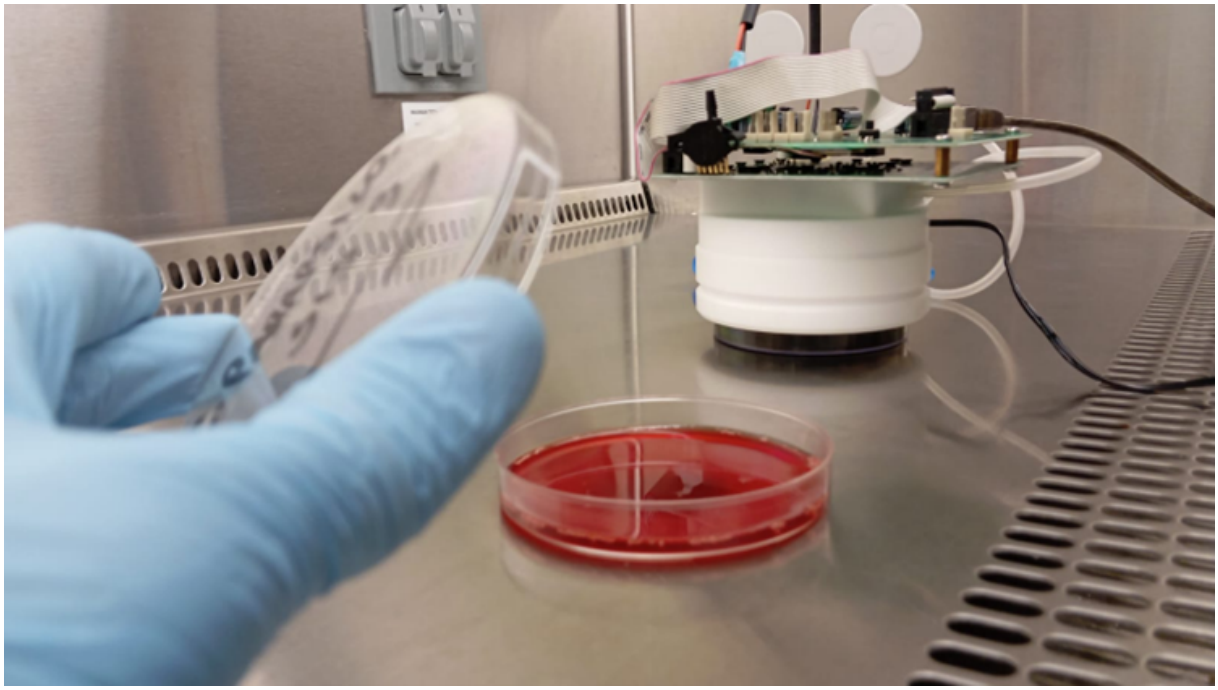


Figure 19 – Experimental setup: Petri dish with a blood sample and the Electronic Nose positioned for the purge stage, over the plate with activated carbon.

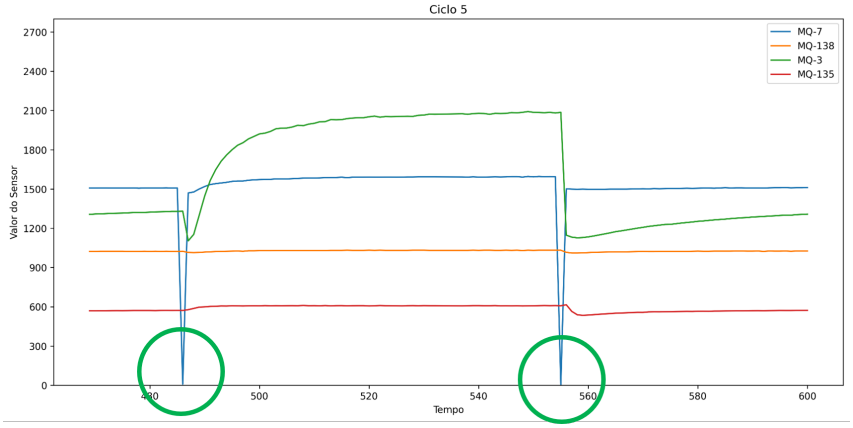
All the used material was properly discarded according to the current biosafety regulations, including Petri dishes, gloves, and contaminated waste, which were deposited in biosafety containers for the disposal of infectious waste. This rigorous methodological approach ensured the collection of reliable and reproducible data, minimizing environmental interferences and ensuring the validity of the analyses. The Electronic Nose used was specially developed for the experiments, containing 4 gas sensors (MQ-7, MQ-138, MQ-3, MQ-135) and 3 environmental sensors (temperature, pressure, and humidity). Figure 19 illustrates part of the collection process and the device used.

4.2.6 Pre-processing and Structuring of Data for Time Series Analysis

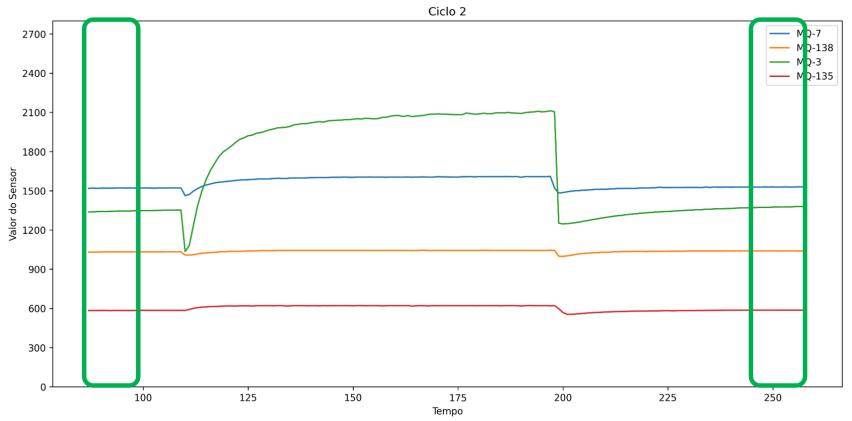
The second methodological stage (Fig. 17) involved the development of the pre-processing of the data and the implementation of the predictive models. After the collection of the samples and analysis by the Electronic Nose, each measurement generated a CSV file with records of the different phases of the experiment: reading, purge, and stabilization. As an initial step, these files were unified into a single structured dataset.

The pre-processing of the data involved the removal of outliers and missing values, the standardization of the length of the measurement cycles, and the consolidation of each cycle

(reading, purge, and stabilization) into a single instance, ensuring a direct correspondence in the training set (Figure 20). In addition, oversampling was applied to evaluate the impact of class balancing. The experiments were performed with both the original data and the balanced set, enabling a comparative analysis.



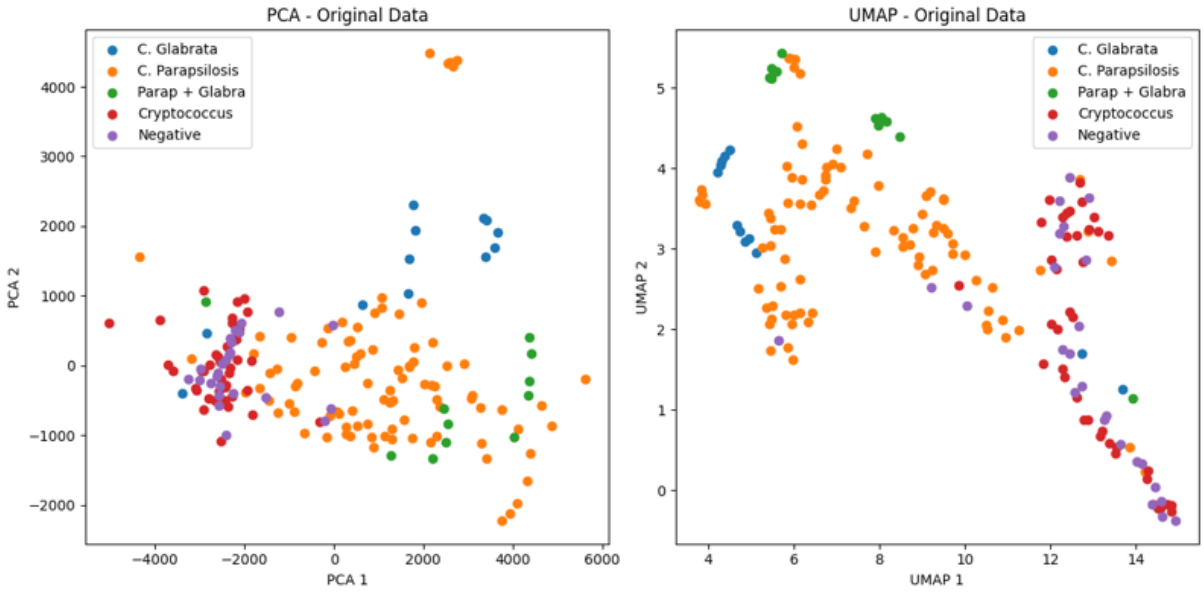
(a) Outlier removal step.



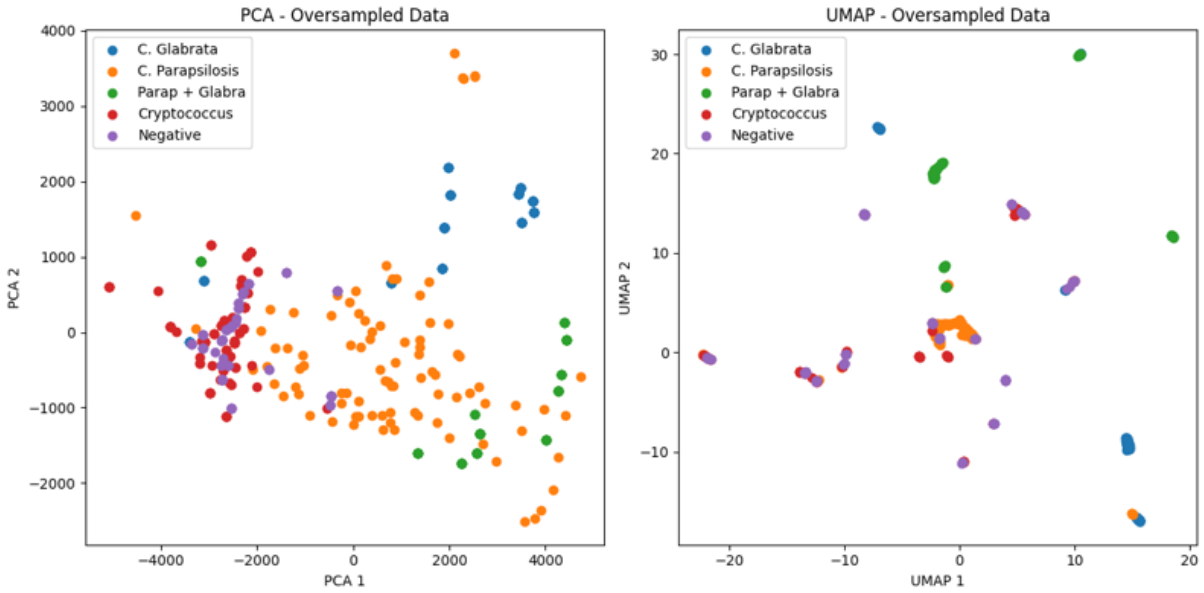
(b) Standardization of the length of the cycles.

Figure 20 – Pre-processing steps using a *C. glabrata* sample as a basis

To better understand the distribution of the data, the UMAP and PCA techniques were used for dimensionality reduction, which allowed for a clearer visualization of the data clustering, both in its original format and after the application of oversampling. Figure 21 presents the resulting dispersion.



(a) UMAP and PCA applied to the original data



(b) UMAP and PCA applied to the data with Oversampling

Figure 21 – Data visualization step using the UMAP and PCA dimensionality reduction techniques for the approaches with and without Oversampling

The visualization of the distribution of the samples by means of PCA and UMAP shows a good separation between the different classes in all the evaluated scenarios, with low overlap between them. Only a slight proximity is observed between the *Cryptococcus* samples and the negative samples, which is expected, since both belong to the group of non-*Candida* samples. Thus, this approximation does not compromise the identification process, which remains effective for the objectives of the study.

In the modeling phase, traditional models and time series models were selected. The choice of temporal models was based on the models used in the culture stage, prioritizing those with

high performance in similar patterns. The traditional models were chosen to ensure diversity of approaches and to evaluate whether less complex solutions could present competitive results.

Model performance was evaluated using a repeated stratified 5-fold cross-validation approach (2 repetitions, 5 folds), totaling 10 iterations. Stratification was applied to ensure the original class distribution within each fold (see Algorithm 1). Although each instance in the dataset consists of a temporal signal (a scent signature cycle), these cycles are treated as independent and identically distributed (i.i.d.) patterns rather than a continuous chronological sequence. Unlike forecasting tasks where temporal dependencies exist between samples, the classification of scent signatures focuses on the internal dynamics of each isolated cycle. Therefore, a standard stratified cross-validation was preferred over TimeSeriesSplit, as the latter is designed for forecasting models where the goal is to predict future values based on past observations, whereas our objective is robust pattern recognition across discrete events.

Algorithm 1 Repeated Stratified Cross-Validation Procedure

Input: Dataset $\mathcal{D}$, Repetitions $R = 2$, Folds $K = 5$
Initialize: Metric list $\mathcal{M} \leftarrow []$
for $r = 1$ to R **do**
    $\mathcal{D}_{shuffled} \leftarrow \text{Shuffle}(\mathcal{D}, \text{seed}_r)$
    $\{F_1, \ldots, F_K\} \leftarrow \text{StratifiedPartition}(\mathcal{D}_{shuffled}, K)$
    for $k = 1$ to K **do**
        $\mathcal{D}_{val} \leftarrow F_k$
        $\mathcal{D}_{train} \leftarrow \mathcal{D} \setminus F_k$
        Model $\leftarrow \text{Train}(\mathcal{D}_{train})$
        score $\leftarrow \text{Evaluate}(\text{Model}, \mathcal{D}_{val})$
        Append score to $\mathcal{M}$
    end for
end for
Output: $\text{Mean}(\mathcal{M})$, $\text{StandardDeviation}(\mathcal{M})$

Final results are reported as the mean performance accompanied by the standard deviation ($\mu \pm \sigma$). The evaluation metrics included accuracy, precision, F1-score, sensitivity, specificity, and processing time, following the same pattern as the tests with culture. Statistical tests were applied to check for significant differences, analyzing the impact of oversampling on the performance of the models. To check for normality, the Shapiro-Wilk test was used, which is suitable for small samples (< 50). Depending on the result, the Wilcoxon test (non-normal) or the paired t-test (normal) was applied, according to the recommendations of the literature (further details on this implementation can be found in the links provided in Appendix A).

4.3 COMPONENT III: THE PREDICTIVE MODELING APPROACH

With the data treated and structured, the models were selected based on the results of the data visualization phase and the study on *Inception Time* (FAWAZ et al., 2020), which compares it with other state-of-the-art models, including its predecessors, the Hierarchical Vote Collective of Transformation-based Ensembles 1 (LINES; TAYLOR; BAGNALL., 2016) and 2 (MIDDLEHURST et al., 2021) (HIVE-COTE 1 (HC1) and HIVE-COTE 2 (HC2)). The visualization of the information showed that the data do not have a large overlap and have a single division between them, so there are no restrictions on which categories of models to use. Thus, in addition to the techniques already mentioned, the K-Neighbors Time Series Classifier (KNN), which implements the K-nearest neighbors for time series (TAVENARD et al., 2020), the Time Series Forest Classifier (TSFC), an implementation of a Time Series Forest using intervals (BABAYEV; WIESE., 2021), the Shapelet Transform Classifier (STC), which uses transformed discriminatory subseries as a classifier (BAGNALL et al., 2019), the Random Interval Spectral Ensemble (RISE), built on the basis of trees and different sets of partial and automatic correlations of features (FLYNN; LARGE; BAGNALL, 2019), the ROCKET Classifier (ROCKET) (DEMPSTER; PETITJEAN; WEBB, 2020), and the BOSS Ensemble (BOSS) (SCHäFER, 2015) were also introduced in the experiments, all time series models that will be used due to the temporal characteristic of the data, translated through the *culture_day* parameter of the base.

As mentioned earlier, a total of 90,802 readings of the six *Candida* species were collected in about 514 cycles. However, to obtain an "olfactory fingerprint" of the data, it was necessary to concatenate all the readings of all the sensors of a cycle into a single row of the dataset, resulting in a new set of 397 instances with 652 columns (now, each sample is related to a cycle, see Figure 22 as an example of a dataframe). Therefore, the base was divided into a training, validation, and test set, with 60% for the first (238 cycles) and approximately 20% for the others (79 and 80 cycles), as can be seen in Figure 23.

Stratified cross-validation serves to maintain a homogenized proportion of the data sampling, seeking to ensure that the training set can represent the entire population, avoiding sampling bias (GéRON, 2021). For each subset used in the training, a result was obtained referring to 5 different metrics: accuracy, recall (sensitivity), F1-Score, precision, and specificity (MORTAZ, 2020). Accuracy (Equation 1) measures the proportion of the correct predictions of the model over the total number of evaluated examples. Recall (sensitivity) (Equation 2) is applied to measure the portion of patterns correctly identified by the classification model.

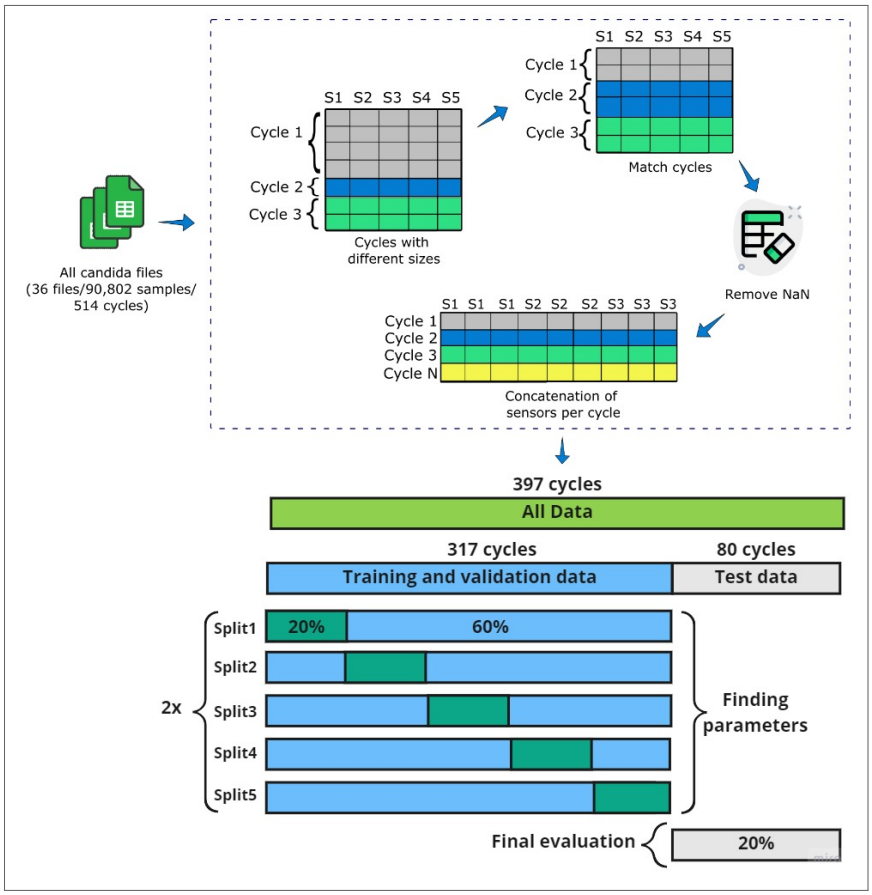
Figure 22 – Final structure of the dataset used - Example for the training set

	0	1	2	3	4	5	6	7	8	9	...	646	647	648	649	650	651
0	0.097089	0.888026	0.409823	0.211322	0.011470	0.002984	0.094389	0.097508	0.888311	0.412353	...	0.942633	0.147668	0.017130	0.006142	0.211215	0.6
1	0.834456	0.001447	0.038346	0.578831	0.709699	0.613095	0.976286	0.836283	0.001452	0.037597	...	0.354621	0.506213	0.470051	0.301036	0.879555	0.6
2	0.846607	0.001447	0.001259	0.637470	0.852780	0.833073	0.995040	0.842061	0.001452	0.001000	...	0.046484	0.561678	0.766249	0.546414	0.997163	1.0
3	0.032964	0.824803	0.370739	0.066015	0.018003	0.005947	0.090825	0.033225	0.825896	0.373669	...	0.853630	0.041059	0.037784	0.016002	0.224307	0.8
4	0.137392	0.833933	0.523603	0.256131	0.024064	0.007370	0.158710	0.138076	0.834368	0.526231	...	0.955287	0.184945	0.041651	0.017419	0.305259	0.6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
233	0.053176	0.819635	0.450645	0.101789	0.017146	0.006038	0.123063	0.053329	0.820363	0.453210	...	0.940946	0.085359	0.033602	0.013576	0.266856	0.8
234	0.724615	0.151579	0.552265	0.878849	0.306767	0.195811	0.787973	0.731672	0.149882	0.547964	...	0.105496	0.873582	0.580337	0.467981	0.985817	0.6
235	0.170837	0.682854	0.685152	0.276231	0.052296	0.022295	0.293707	0.172024	0.683777	0.688790	...	0.915215	0.206555	0.097539	0.050178	0.506437	0.6
236	0.247714	0.663216	0.725540	0.340955	0.061548	0.027736	0.336175	0.248706	0.663376	0.728344	...	0.909309	0.253377	0.090435	0.051028	0.507310	0.8
237	0.673845	0.192923	0.654753	0.688733	0.250894	0.123837	0.661965	0.709281	0.203306	0.636198	...	0.281647	0.727355	0.485233	0.312376	0.807331	0.2

238 rows × 652 columns

Source: Author

Figure 23 – Experiment development flow: the database of the readings of all species is united into a single base, creating a label to associate each row of the base with a type of *Candida*. The data are then normalized and separated by cycles. In this case, all the sensor data are concatenated into a single row, referring to the cycle in which they were generated. Only after this, these values are divided into training, validation, and test bases, following the guidelines of cross-validation with 10 k-folds.



Source: Author

Specificity (Equation 3) is used to test the ability to correctly determine the negative cases. Precision (Equation 4), on the other hand, is applied to measure the number of positive pat-

terns correctly predicted, based on the total number of patterns predicted in a positive class. Finally, the F1-Score (Equation 5) or F1-measure, portrays the harmonic mean between the values of precision and recall (HOSSIN; SULAIMAN, 2015). All these metrics are calculated based on the values of true positive (TP), false positive (FP), false negative (FN), and true negative (TN), obtained after crossing the predicted values with the current values of each class. All the equations related to each of these metrics can be seen highlighted.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

$$Recall(Sensitivity) = \frac{TP}{TP + FN} \tag{4.2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.4}$$

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \tag{4.5}$$

Seeking to improve the validity of the project, at the end of the experimentation process, a statistical analysis was applied using the Shapiro-Wilk normality test, the non-parametric Kruskal-Wallis test, and the Nemenyi post-hoc test to understand the statistical significance between the means of the results and to highlight the difference between the tested models, detailed in Chapter 5, Results and Discussions.

4.4 COMPONENT IV: THE XAI ENSEMBLE EXPLAINABILITY ARCHITECTURE

The methodology proposed here is based on the findings and strategies established earlier, in the identification of culture and blood broth samples. The main innovation of this methodology lies in the integration of a novel Explainable AI Ensemble approach with the best-performing model, trained on a new dataset collected from blood broth. This integration was implemented in the *DiagNose.AI* system through the development of four main artifacts, with the aim of enabling interpretable and reliable predictions. The ensemble mechanism represents the main methodological contribution, offering robust and consistent explanations for classification tasks in time series.

The initial development consisted of the construction of a database containing Volatile Organic Compounds (VOCs) associated with species of the genus *Candida*. To identify these VOCs, a review of the scientific literature was carried out, which resulted in a set of eight relevant publications (CAILLEUX et al., 1992; HERTEL et al., 2018a; JENKINS et al., 2019; SILVA et al., 2020; SANTOS et al., 2019; HERTEL et al., 2016; ARRARTE et al., 2017b; BENDA et al., 2008), in addition to the mVOC database (LEMFACK et al., 2024). The identified compounds were inserted into the database and classified according to their chemical classes, in order to facilitate the association with *Candida* species. In addition, an analysis of the technical manuals of the sensors present in the different versions of the E-nose was carried out, to identify which groups of VOCs are detectable by each sensor.

With the database built, the development of an ensemble-based explainability library was initiated. This approach integrates Grad-CAM, LIME, and SHAP to identify the most relevant sensors used by the final AI model in the prediction with E-nose time series data, using a majority voting strategy. Where each method generates a weight for each of the characteristics (sensors) it considers most important and returns the average of the three that obtained the highest weight (most relevant) during its execution. This choice results from a trade-off analysis between robustness, interpretability, and computational feasibility. Limiting the selection to three features per technique reduces the cognitive load of the generated explanation, making it more suitable for non-technical users, as suggested in (HOLZINGER et al., 2019; LIPTON, 2018).

Although there are other strategies for aggregating XAI outputs — such as weighted average of importances or the use of explanatory meta-models (ARRIETA et al., 2020) — majority voting was chosen due to its low computational complexity and greater transparency, which facilitates manual audits and integration with the tool's textual interface. This decision is also aligned with human-centered XAI design principles (MA, 2024).

Another reason for adopting majority voting based on the top three features is related to the specificities of the E-noses used in this study. Between the two versions of the device analyzed, one includes seven sensors, while the other has only four. Thus, increasing the number of selected features would not be appropriate, as it would include all the available sensors in the device with fewer channels. In this context, majority voting proved to be the most appropriate selection technique, since alternative approaches, such as weighted averages, intersection, or union, would not bring significant differences in the selection process.

The relevance is homogenized as follows: Grad-CAM calculates gradient-based activation

scores for the sensors' time series; LIME fits a local linear model to rank the importance of the sensors; and SHAP calculates Shapley values for the contributions of the sensors. These scores are normalized in the range [-1,1], ensuring comparability. This approach favors a controlled redundancy, helping to mitigate the biases of each individual technique (for example, the oversimplification of LIME or the computational variance of SHAP), promoting a more robust selection.

The library is also responsible for querying the database and executing the AI model. In addition, it allows for the consultation of three random instances of the database used in the training of the model for comparative purposes.

For the development of the third milestone of the project, the communication structure between the Ensemble XAI library and the interface of the DiagNose.AI system was built. This module was programmed to be executed automatically whenever the main application is started, allowing the retrieval of all relevant information from the system, including predictions, characteristics of the Ensemble XAI, mapping of VOCs, and data of training instances.

Finally, all modules were integrated into the graphical interface of the DiagNose.AI system. In addition to communicating with the E-nose for information collection, the system also began to perform the prediction and explainability of the results. In this version, in addition to the probability of the sample belonging to a certain species, a textual explanation is presented highlighting which sensors were most relevant. Additionally, the associations between the identified VOCs and their respective categories are described. The interface also offers visualization and comparison between the current sample and the training data, as well as the mapping between VOCs and relevant sensors. Figure 24 represents the complete flow of the system.

This process begins with the reading of the samples by the E-nose in a controlled environment. The generated data (time series of the variation of the electrical resistance of the sensors) undergo preprocessing before being sent to the AI model (see this process in Figure 25). In this stage, the reading data are organized into cycles, in which the information of all the sensors recorded at the same instant is aggregated into a single instance of the dataset — encompassing both the reading and purge phases. This structuring follows the same pattern adopted in previous studies, ensuring compatibility with the predictive model in use. However, this data format introduces an additional challenge in the context of explainability, as the main components under analysis — the sensors — are distributed in each instance, making the direct interpretation of their individual contributions difficult.

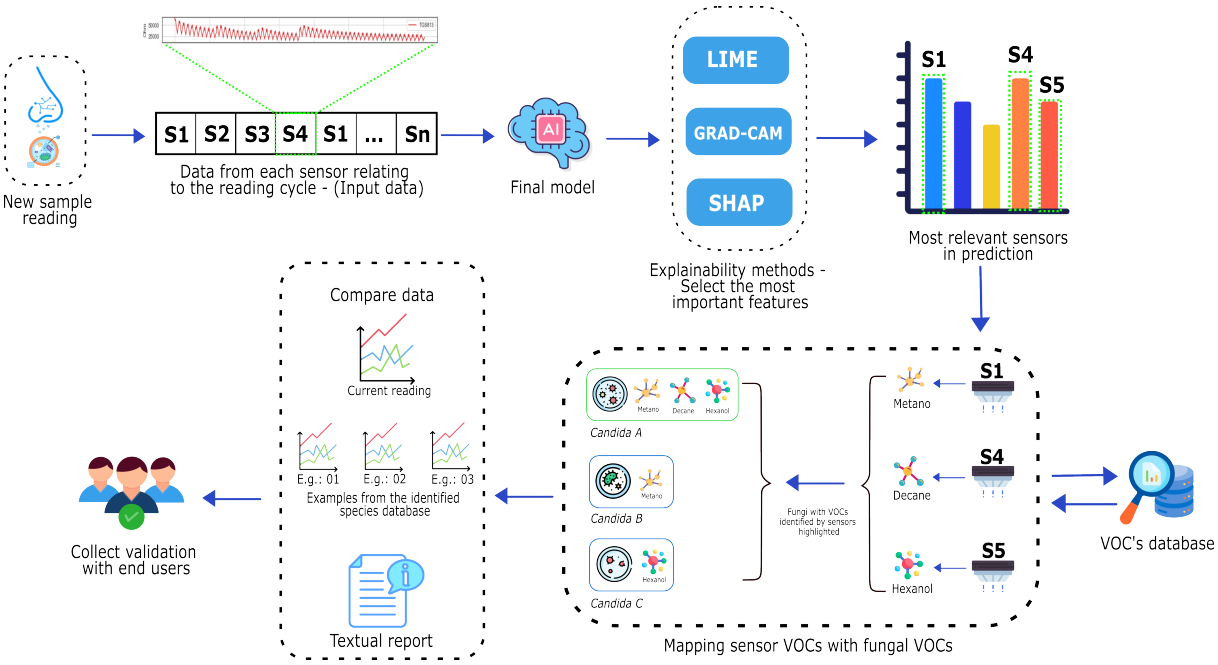


Figure 24 – Execution flow of *DiagNose.AI*. The process begins with the reading of the samples by the E-nose, followed by preprocessing and sending of the data to the AI model. Explainability techniques (LIME, SHAP, and Grad-CAM) are combined (Ensemble XAI) to identify the most relevant sensors and VOCs. Based on majority voting, only the most influential features (sensors) are selected. The system generates graphical and textual reports (s) comparing the predicted data with the actual data of the bank, validated by experts.

To overcome this issue, after the application of each technique of the VOCs Ensemble eXplainability Framework (Grad-CAM, LIME, and SHAP, with the average of three repetitions to ensure greater accuracy), the data are restructured in their original format, with readings organized by sensor. From these restructured data, the sensors that most frequently appear among the features of greatest relevance are identified based on their frequency and the scores attributed by each method. Then, the majority voting scheme is applied, in which each occurrence of a sensor among the highlighted features of a method counts as a vote. The sensors with the highest number of votes among all the methods are ultimately selected as the most relevant by the ensemble (see Algorithm 2).

The extracted information is then cross-referenced with the VOC database, allowing for the mapping of the compounds identified by the sensors with those described in the literature. The species whose VOC signature has the greatest correspondence with the highlighted sensors is considered the most probable. This cross-referencing allows for the comparison between the results of the model and the chemical data, offering an interpretable explanation. At the end, the user receives a detailed textual explanation accompanied by graphical visualizations. To ensure that this process was appropriate for the domain in question, a prototype of the application was validated in tests with professionals in the area.

Algorithm 2 XAI Ensemble for Feature Relevance Identification

Require: Test instance x, Trained model M, Number of repetitions R
Ensure: Sensor ranking and biological mapping summary

function EXECUTEENSEMBLE(x, M, R)
    $EnsembleResults \leftarrow []$
    for $i \leftarrow 1$ **to** R **do**
        $W_{LIME} \leftarrow$ CalculateLIME(x, M)
        $W_{SHAP} \leftarrow$ CalculateSHAP(x, M)
        $W_{GRAD} \leftarrow$ CalculateGradCAM(x, M)
        $Top_{LIME} \leftarrow$ SelectTopFeatures$(W_{LIME}, 3)$
        $Top_{SHAP} \leftarrow$ SelectTopFeatures$(W_{SHAP}, 3)$
        $Top_{GRAD} \leftarrow$ SelectTopFeatures$(W_{GRAD}, 3)$
        $CommonFeatures \leftarrow$ Intersect$(Top_{LIME}, Top_{SHAP}, Top_{GRAD})$
        $EnsembleResults.append(CommonFeatures)$
    end for
    $Ranking \leftarrow$ ComputeFrequency$(EnsembleResults)$
    $BioMapping \leftarrow$ QuerySQLDatabase$(Ranking)$
    return $BioMapping$
end function

Initially, the pipeline processes the time-series data through the trained model M, where the learning process occurs via the optimization of a categorical cross-entropy loss function. In this stage, convolutional filters (if present) or dense connections learn to map specific biochemical signatures to the target fungal classes. To extract the "reasoning" behind these classifications, the function `ExecuteEnsemble` performs R iterations to ensure statistical consistency and mitigate the inherent stochasticity of individual XAI methods.

The first perspective, LIME, generates a localized linear approximation of the model's decision boundary by perturbing the input signal x and observing the response variance. This identifies features that are locally indispensable for the prediction. Simultaneously, SHAP provides global-local consistency by computing the *Shapley Value*, a solution concept from cooperative game theory that assigns a fair importance score to each sensor by evaluating its marginal contribution across all possible sensor combinations. Finally, Grad-CAM utilizes the gradient of the winning class flowing into the final convolutional layer to produce a localization map, highlighting the specific sensors and time intervals that most significantly activated the neural features.

The core innovation of the method lies in the `Intersect` and `ComputeFrequency` operations. By retaining only the sensors identified by the majority of the methods, the ensemble filters out explanatory noise. This refined ranking is then subjected to a *Biological Mapping* via SQL queries into the VOC's database. This final step transforms raw feature importance

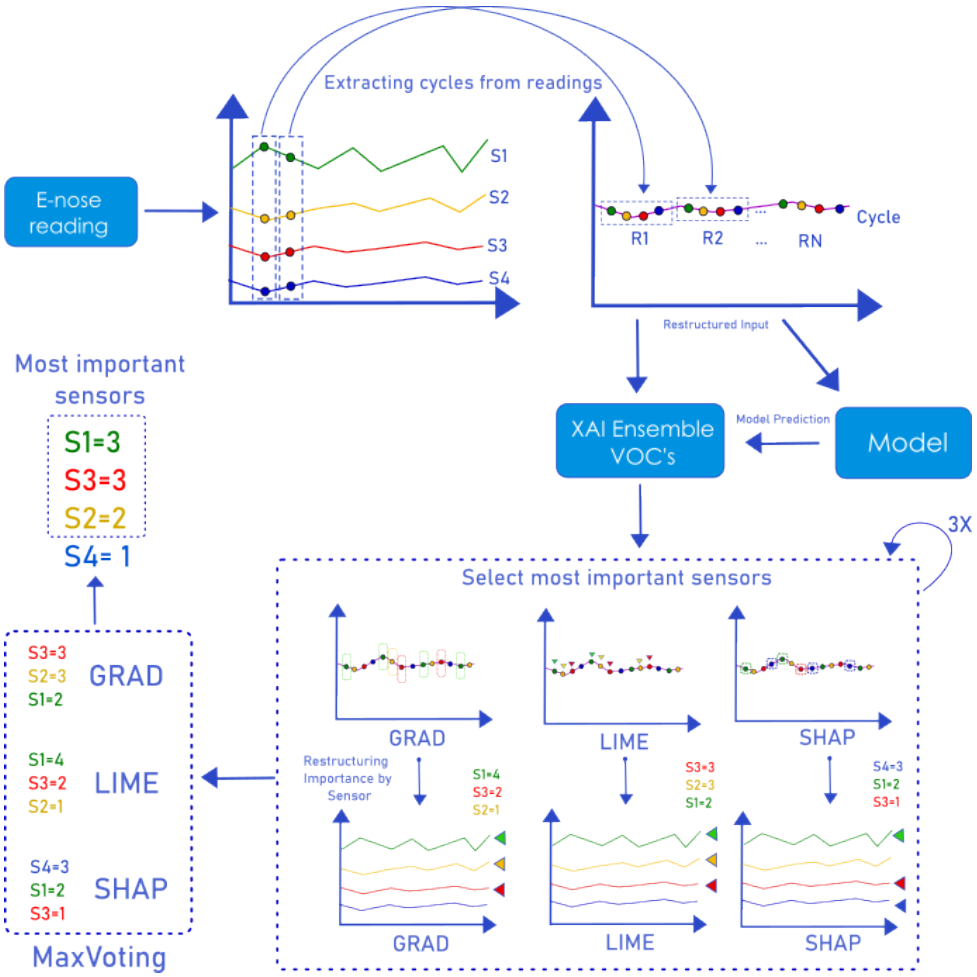


Figure 25 – The process begins with raw readings from the E-nose from multiple sensors (S1–S4), which are restructured into time cycles (R1–RN). These reformulated inputs are sent to the AI model for prediction and are processed simultaneously by the VOCs Ensemble XAI module. Each explainability technique (Grad-CAM, LIME, and SHAP) analyzes the input independently and selects the most important sensors. To ensure interpretability, the explanation goes through a restructuring stage, reorganizing the importance scores by sensor. After three iterations for greater robustness, a majority voting strategy aggregates the most frequently highlighted sensors among the methods. The final result identifies the most relevant features that contributed to the model's decision.

into mycological evidence by correlating the most active sensors with the VOC profiles of the identified fungus (e.g., *Candida albicans*), thus validating the model's decision against established clinical literature. To understand further details about the method's implementation, the repository links for their respective codes are available in Appendix A of this thesis.

4.4.1 Experimental Setup

One of the validation approaches of the XAI methodology was to use the model tested and evaluated on the blood broth VOC dataset. This first set is composed of 14 samples in blood broth, including 2 uninfected control samples, 1 sample of a non-*Candida* species

(*Cryptococcus neoformans*), 9 samples infected by 3 distinct *Candida* species (*C. albicans*, *C. glabrata*, *C. parapsilosis*), and 2 samples infected with a combination of two distinct *Candida* species (*C. parapsilosis* and *C. glabrata*). For each of these aliquots, between 10 and 40 reading cycles were performed. The data underwent a normalization process and were divided into training (60%), validation (20%), and test (20%) sets.

The E-nose used in this collection was specially developed for the experiments and includes 4 gas sensors (MQ-7, MQ-138, MQ-3, MQ-135) and 3 environmental sensors (temperature, pressure, and humidity), although only the gas sensors were used as input data. After the data restructuring stage described in the Methodology Section, the final set consisted of 546 rows (cycles) $\times$ 513 columns (sensor data).

Models such as *Time Series Forest*, *Inception Time*, and *Support Vector Classifier* (SVC) were trained and evaluated using a repeated cross-validation approach, dividing the data into ten distinct subsets at each repetition. This process was repeated several times to ensure that the models developed a strong generalization ability of their predictions, that is, that they maintained reliable performance when applied to new data, instead of just memorizing patterns of the training samples.

The performance evaluation was carried out using several metrics, including accuracy, precision, F1-score, recall (sensitivity), specificity, and processing time (for the training and testing phases). After the completion of the experiments, the *Inception Time* model obtained the best overall performance considering the evaluated metrics and, therefore, was adopted as the main model for the tests carried out in this study.

To broaden the validation of the methodology, the experiments were also conducted with the dataset from the experiments with culture, which was carried out using the Suitcase E-nose. As mentioned earlier, this second dataset is composed of 90,802 samples of five *Candida* species (*C. glabrata, C. haemulonii, C. kodamaea ohmeri, C. krusei*, and *C. parapsilosis*), collected over 514 cycles. The results of both experimental configurations are presented and discussed in the following sections.

5 VALIDATION OF THE DIAGNOSE.AI FRAMEWORK: RESULTS AND DISCUSSIONS

After the presentation of the DiagNose.AI Framework's architecture, this chapter is dedicated to its empirical validation, presenting the results that prove its effectiveness and robustness. It is fundamental to highlight that the foundation for all subsequent analysis lies in one of the central contributions of this thesis: the creation and characterization of two novel databases, an effort that addresses a critical gap in the literature and generates the necessary data assets for the rigorous validation of the proposed methodology.

Based on these new datasets, a series of experiments was conducted to test the Framework in scenarios with different levels of complexity. The first validated the methodology on previously laboratory-cultivated ATCC culture samples, while the second elevated the challenge to a clinical scenario closer to the Framework's real-world application environment, using infected blood broth directly. It is crucial to highlight that a real-world application, in this context, refers to the use of the Framework in a clinical setting directly utilizing blood samples. This, however, does not negate the necessity for the Framework to be executed in a controlled environment, isolated from the interference of external ambient VOCs. In both contexts, the performance of various classification models, with an emphasis on time series, was systematically evaluated. The evaluation metrics, described in Chapter 4, were employed to quantify the performance of the Framework's modeling and explainability components.

Additionally, this chapter will present the results that validate the pioneering Ensemble XAI architecture. Through comparative studies, including sensitivity and ablation analyses, the consistency and reliability of the explainability method were demonstrated. The consolidation of these results culminated in the integration of the methodology into the *DiagNose.AI* prototype system, whose outputs were evaluated by specialists to confirm the relevance and interpretability of the generated explanations.

5.1 THE GENERATED DATASETS

One of the fundamental contributions of this project, and a prerequisite for the validation of any machine learning approach, is the creation of robust and well-characterized datasets. The notorious scarcity of public databases in the area represents a barrier to scientific advancement and reproducibility. To address this gap, a significant effort was dedicated to the construction

and curation of two novel datasets, involving inter-institutional and international collaboration. Table 5 describes the main characteristics of each database generated within the scope of this work.

Table 5 – Characterization of the databases generated and used for the validation of the DiagNose.AI Framework. Each database represents a distinct validation scenario, with different sample origins, reading devices, and species profiles.

Database	Sample Origin	Reader (E-nose)	No. of Cycles*	Total Readings	No. of Samples	Species
Culture-UFPE	ATCC Culture	Suitcase (7 sensors)	514	90,802	41	*albicans, glabrata, haemulonii, kodamaea_ohmeri, krusei, parapsilosis*
BloodBroth-UC	Infected blood broth	Prototype (4 sensors)	546	46,574	14	*glabrata, parapsilosis, glabra + parapsi, cryptococcus, negative, albicans*

* Cycles are the readings after the restructuring process. in the case of the culture samples, many readings were disregarded to standardize the size of the cycles.

The generation of the described datasets constituted a fundamental methodological step, indispensable for the rigorous validation of the Framework. The process required significant interdisciplinary and inter-institutional mobilization, involving close collaboration of specialists from the Center for Informatics, the Medical Mycology team at UFPE, the Northeast Regional Center for Nuclear Sciences/UFPE, and, crucially, the international partnership with the College of Medicine at the University of Cincinnati. This synergy was essential to ensure the execution of a high-fidelity protocol, which ranged from the standardized cultivation of ATCC samples to the handling of clinical blood broth samples. The subsequent stage of data curation and structuring was a critical component to transform the raw signals from the sensors into a cohesive time-series format suitable for modeling. The public availability of these datasets is, therefore, a deliberate contribution of this thesis, aiming to promote reproducibility and catalyze new research at the intersection of VOC sensing and machine learning.

5.2 VALIDATION OF THE PREDICTIVE AND EXPLAINABILITY COMPONENTS

With the databases established, the next step consisted of validating the central analysis components of the DiagNose.AI Framework: the predictive modeling and the explainability architecture. The experiments were conducted in the two distinct scenarios to test the effectiveness and generalization of the methodology.

5.2.1 Scenario 1: Validation in a Laboratory Environment using ATCC Culture

The first experimental scenario aimed to evaluate the performance of the models in a controlled laboratory environment, using reference cultures from the ATCC collection. This context represents a fundamental validation step, as it allows for the verification of the algorithms' ability to recognize and differentiate the VOCs of *Candida* species at different stages of their growth (differentiating minutes from hours and hours from days). This controlled setting is essential to establish a high-confidence performance baseline, ensuring that the chemical signatures captured by the E-nose sensors are directly attributable to specific biological markers before progressing to complex clinical matrices. Thus, the goal is not only to assess the effectiveness of the models in the training process but also to understand how they respond to the introduction of unseen data in the validation and testing phases, reflecting their true generalization ability.

Table 6 presents a summary of the averages of all metrics collected in the training stage, as well as the standard deviation of accuracy for the 10 repetitions and the training time for each model in seconds. The recording of training time, alongside performance metrics, serves as a preliminary assessment of the computational efficiency of the proposed models, ensuring their suitability for future near real-time diagnostic applications.

In addition to the results regarding training, the averages obtained in the validation and testing phases were also recorded. The comparative analysis of these three stages shows a slight decrease in the models' performance between training and the subsequent phases. Such behavior is expected, as during training, models tend to capture *Candida* patterns with a high level of precision due to the marked differences between the analyzed species. However, in the validation and testing phases, which consist of data unseen by the models, it is natural for a generalization gap to emerge, reflecting the real-world capability of the algorithms.

This phenomenon, far from compromising the final evaluation, reinforces the robustness of the methodology, as it demonstrates that the models did not merely memorize the training data but were subjected to rigorous cross-validation across distinct temporal windows. Furthermore, the low standard deviation observed across multiple repetitions indicates the stability of the predictive process, a prerequisite for the subsequent implementation of reliable explanation modules. Tables 7 and 8 present, respectively, the detailed results of the validation and testing stages for each of the investigated models, allowing for a more comprehensive comparative analysis of their performance in different phases of the experimental process.

Table 6 – Result of the model training

Classifiers	Accuracy	F1-Measure	Recall (Sensitivity)	Precision	Standard Deviation (of accuracy)	Specificity	Training time (s)
Inception Time	0.97740	0.97679	0.97245	0.98192	0.00081	0.99667	548.83118
Random Interval Spectral Ensemble (RISE)	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	14.79765
Time Series Forest Classifier	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	19.22641
ROCKET Classifier	1.00000	1.00000	1.00000	0.99886	0.00000	1.00000	19.5395490
Shapelet Transform Classifier	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	36.58212
K-Neighbors Time Series Classifier	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	0.96999
HIVE COTE 1	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	7,287.21419
HIVE COTE 2	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	21.65166
BOSS Ensemble	1.00000	1.00000	1.00000	1.00000	0.00000	1.00000	81.42075

Average result of the training values for the Accuracy, F1-score, Recall, Precision, and Standard Deviation metrics calculated for each model during the 10 iterations of the stratified cross-validation

Table 7 – Result of the model validation

Classifiers	Accuracy	F1-Measure	Recall (Sensitivity)	Precision	Standard Deviation (of accuracy)	Specificity	Validation time (s)
Inception Time	0.95875	0.96700	0.96720	0.96811	**0.00915**	0.99205	544.09643
Random Interval Spectral Ensemble (RISE)	0.94965	0.94869	0.94566	0.96092	0.02496	0.99212	2.84949
Time Series Forest Classifier	**0.97895**	**0.98118**	**0.97931**	**0.98447**	0.03491	0.99194	3.97698
ROCKET Classifier	0.95117	0.95037	0.94604	0.96318	0.01776	0.99464	4.85449
Shapelet Transform Classifier	0.94737	0.94901	0.95462	0.95038	0.04077	0.99230	2.86250
K-Neighbors Time Series Classifier	0.95114	0.94788	0.94626	0.95929	0.01789	0.99100	15.35404
HIVE COTE 1	0.95789	0.95859	0.95837	0.96742	0.03158	**0.99479**	6.69850
HIVE COTE 2	0.93844	0.94232	0.94406	0.94758	0.02700	0.99470	2.67899
BOSS Ensemble	0.95429	0.95060	0.94543	0.96464	0.02050	0.99203	**0.83150**

Average result of the validation values for the Accuracy, F1-score, Recall, Precision, and Standard Deviation metrics calculated for each model during the 10 iterations of the stratified cross-validation

Table 8 – Result of the model testing stage

Classifiers	Accuracy	F1-Measure	Recall (Sensitivity)	Precision	Specificity	Test time (s)
Inception Time	**0.97468**	**0.97605**	**0.97817**	**0.97540**	**0.99513**	1.21489
Random Interval Spectral Ensemble (RISE)	0.65000	0.55758	0.57007	0.56251	0.94223	4.58585
Time Series Forest Classifier	0.67500	0.61261	0.58960	0.61261	0.91312	1.18719
ROCKET Classifier	0.78750	0.78105	0.85764	0.79171	0.93804	4.91326
Shapelet Transform Classifier	0.63750	0.58207	0.60258	0.59832	0.93028	0.61583
K-Neighbors Time Series Classifier	0.75000	0.73245	0.72192	0.81357	0.95635	52.86100
HIVE COTE 1	0.52500	0.40245	0.42669	0.39475	0.94009	11.93961
HIVE COTE 2	0.66250	0.58503	0.60360	0.62833	0.94342	2.57620
BOSS Ensemble	0.63750	0.50525	0.53555	0.49929	0.90165	**0.48432**

Result of the test values for the Accuracy, F1-score, Recall, and Precision metrics calculated for each model after the training and validation phase

As presented in the results, **the most outstanding model was Inception Time** (FAWAZ et al., 2020) (even with a lower performance in the training phase), executed with the default parameters, followed by the ROCKET Classifier (DEMPSTER; PETITJEAN; WEBB, 2020), K-Neighbors Time Series Classifier (TAVENARD et al., 2020), and the Time Series Forest Classifier (BABAYEV; WIESE., 2021), respectively. All metrics calculated for Inception Time were close to 100%, with a minimal standard deviation, demonstrating great consistency among the results in each subset of data used.

The comparative analysis of the results obtained in the validation (Table 7) and testing (Table 8) phases reveals a significant disparity in the performance of the classifiers, which justifies the emphasis on the highest-performing models. In validation, the Time Series Forest Classifier demonstrated the most robust average performance across the ten iterations, standing out with the highest Accuracy (0.97895), F1-Measure (0.98118), and Precision (0.98447). However, in the final testing phase (Table 8), the Inception Time classifier excelled, delivering the best overall performance, with an Accuracy of $\mathbf{0.97468}$ and a Specificity of $\mathbf{0.99513}$, while simultaneously showing a remarkably fast test time (1.21489 s), indicating excellent generalization capability.

The substantial difference in performance between these top-tier models and the others, such as the Random Interval Spectral Ensemble (RISE) and the HIVE COTE 2, is a crucial finding. While Inception Time and Time Series Forest are complex architectures designed to extract deep temporal features from series, the lower-performing classifiers failed to capture the complexity and inherent patterns in the VOC data with the same effectiveness. This substantial discrepancy validates the selection of more advanced models, which, despite potentially requiring longer validation times (such as Inception Time at 544.09643 s), offer superior generalization and prediction capabilities in the testing phase, making them essential for the reliability of the clinical diagnosis.

5.2.1.1 Cost and performance analysis of the culture experiments

Observing the executed models, both Inception Time and HIVE COTE 1 had relatively high training times when compared to the other models. This can be explained by the characteristic of the models, being derived from deep neural networks and ensembles with classifiers of different natures. Inception Time, for example, is executed based on 1500 epochs (default value in the original code of its repository). This implies a greater number of times the model

will seek to learn from the training data, requiring more time to learn about the instances. HIVE COTE 1, on the other hand, is an ensemble that combines the classification of a series of algorithms, where some of them may have had more difficulty in training the data, resulting in a peak in relation to the training time.

Despite being more computationally expensive in training and validation, in the testing stage, the Inception Time model had its return in less than a second, being one of the most efficient among all those evaluated. At this point, there is no great difference in relation to the return time in each model, reducing the problems related to the final use cost of the algorithms in an operational environment. However, this time may vary depending on the machine that will execute the application with the embedded model. Even so, the increase of a few seconds in the response time should not be a problem compared to the diagnostic methods currently used.

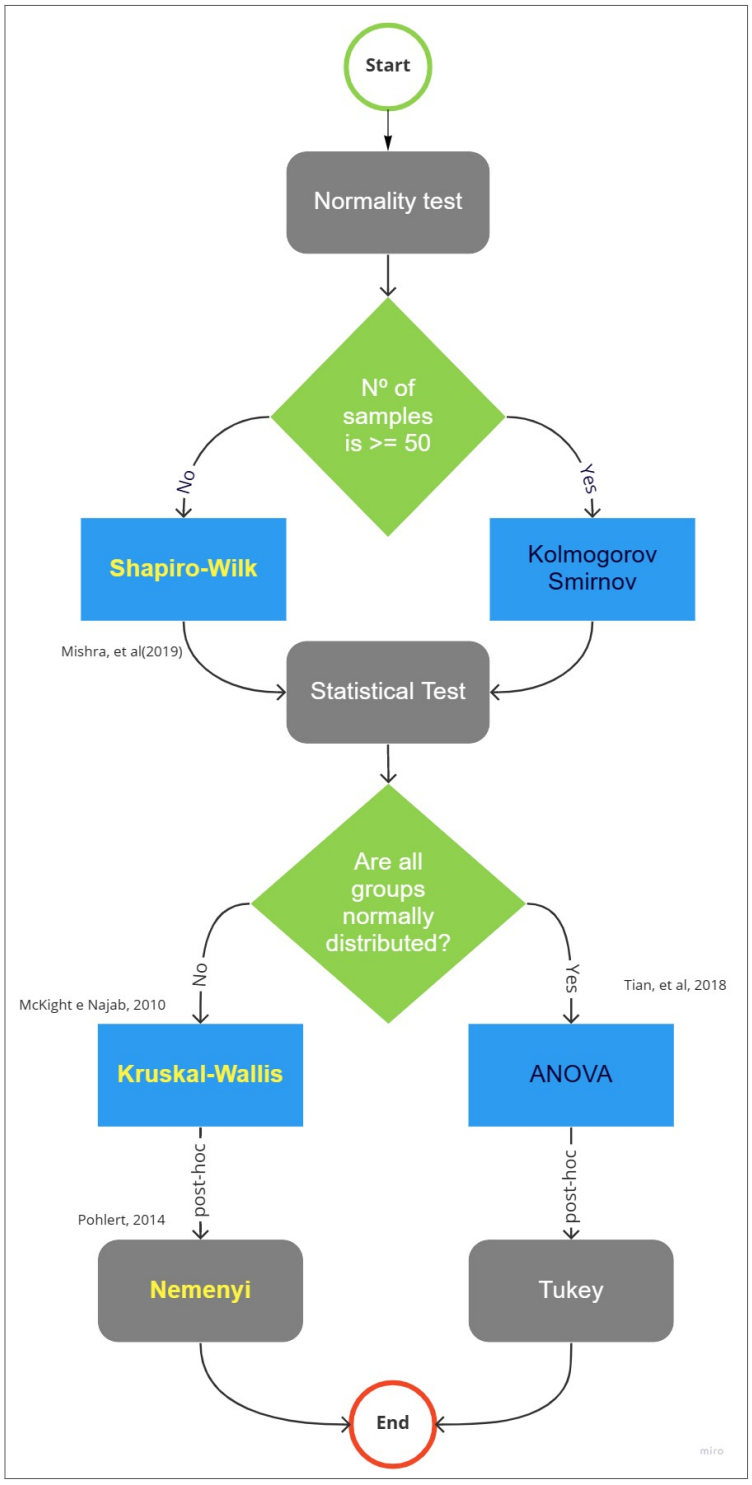
Still in the context of cost analysis, directing to a more general scope of the project, initially, there was no objective to value the total costs of construction and development of the application. Given the different fields and laboratories involved, it would be a rather arduous task for the purpose of the work. However, even knowing the costs related to each area involved, after completion, the expenditures will be focused on the large-scale production of the device and the development of the application that will embed the already trained AI model. For this reason, the proposal should stand out in relation to the other existing alternatives today, becoming a highly viable solution for regions poorly served by health technology.

5.2.1.2 Statistical analysis of the culture experiments

In addition to collecting the metrics, statistical tests were performed to verify the difference between the results of the different models. In this sense, a normality test was performed with the precision results obtained in the 10 repetitions for each model of the validation stage, followed by the significance test and the post-hoc test, to create a pairwise comparison between each of the selected algorithms. Figure 26 demonstrates the flow for choosing the most appropriate tests for use in this study.

Before performing the statistical tests, it was necessary to identify whether the data followed a normal distribution or not, in order to enable the choice of the most appropriate statistical test. In the literature, the most common ways to do this are through graphical evaluation and numerical evaluation, which include statistical normality tests. The verification

Figure 26 – Flow for selecting the most appropriate statistical tests for use in the accuracy groups of each model.

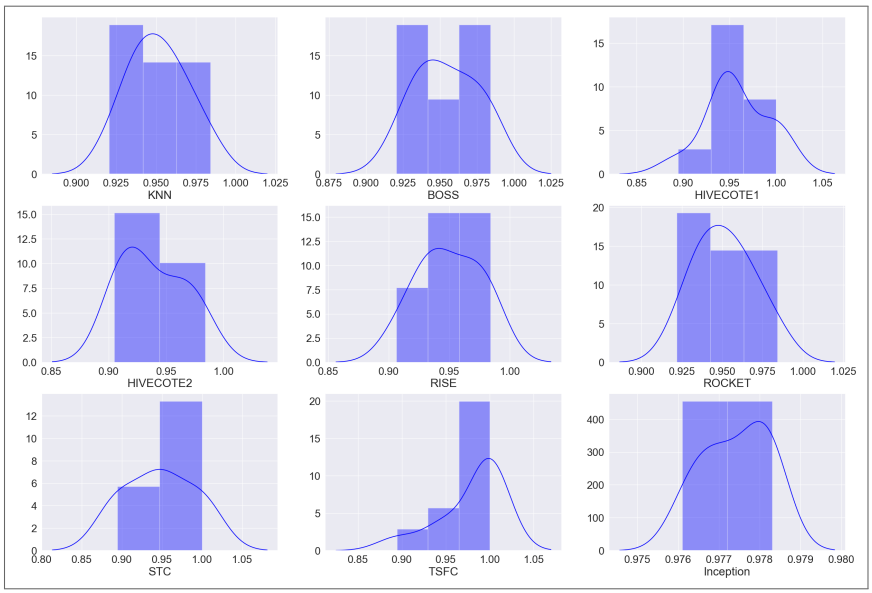


Source: Author

by means of graphs has the advantage of allowing a less sensitive judgment on occasions when the numerical tests can be excessively sensitive, and the numerical test tends to be more objective, reducing the dependence on visual interpretation (CHEN et al., 2019). At first, the visual analysis was done by means of a histogram with a normality line and a QQ-plot (or

quantile-quantile plot), shown in Figures 27 and 28, respectively.

Figure 27 – Histogram with normality line of the mean accuracy groups of each model. In this case, the graphs with lines more similar in shape to a bell tend to indicate a normal distribution. The lines that deviate from this pattern can be considered non-normal.

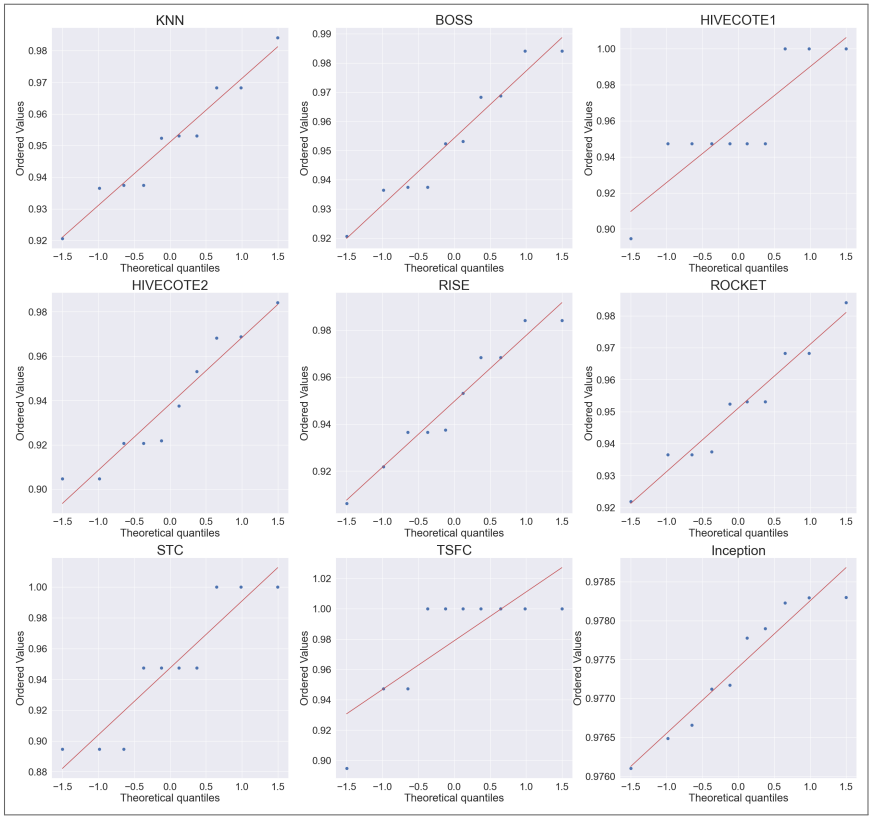


Source: Author

Analyzing the histogram, it is possible to note that most of the sets present a bell-shaped curve (Gaussian curve) in the distribution of the data, with some divergences in relation to the histogram of InceptionTime and Hive-Cote 1 and 2, creating undulations a little different from the shape of a bell on the distribution of the data. To better understand these two points, we can also analyze the QQ-plot, which shows the normality of the data as the points of its distribution get closer to the trend line drawn on the graph. In this sense, it is possible to note that some models present points that are relatively distant and with a different orientation from the inclined line, becoming possible candidates for representing a non-normal distribution of the data.

As a result of the visual analysis, it can be interpreted that the HC1, STC, and TSFC models do not follow a normal distribution of the data, which in itself would already indicate the use of a non-parametric test for the evaluation of the results. However, to obtain a greater sensitivity of the analyses, a numerical statistical normality test was also applied. The most appropriate for the problem in question was the Shapiro-Wilk test. According to Hossin e Sulaiman (2015), this method is more appropriate for small sample sets, smaller than 50, although it can also be used for large sets. Methods like Kolmogorov-Smirnov are more appropriate for large samples or samples equal to 50. Both tests use as a null hypothesis the statement that the data are derived from a set of normal distribution. This hypothesis is accepted when $p > 0.05$,

Figure 28 – QQ-plot graph showing the distribution of the data of the means of each model. In this type of graph, when the points move further away from the straight line, deviating its direction, this suggests that the distribution is moving away from normality. On the other hand, when the points are more aligned with the line, this suggests a normality of the data.



Source: Author

Table 9 – Result of the Shapiro-Wilk normality test.

	KNN	BOSS	HC1*	HC2	RISE	ROCKET	STC*	TSFC*	Inception Time
p	0.77949	0.47213	0.01227	0.27595	0.54311	0.73962	0.03521	0.00021	0.26794
S	0.95944	0.93242	0.79406	0.90926	0.93910	0.95601	0.83184	0.64968	0.90806

*HC1, STC, and TSFC did not show a normal distribution according to the Shapiro-Wilk normality test, with $p < 0.05$

consolidating the data as normally distributed. As the sample set in this study is equal to 10 for each set, the Shapiro-Wilk test was applied to each group of repetitions, resulting in the p-values listed in Table 9.

As can be seen, the HIVE-COTE1, Shaplet Transform Classifier, and TimeSeries Forest Classifier classifiers did not show a normal distribution according to the Shapiro-Wilk test, as observed in Table 9. In any case, with this result, we confirm the need to apply a non-parametric test, considering that only some groups follow a normality distribution. In this sense, according to McKight e Najab. (2010), the most appropriate non-parametric test for

this case is the Kruskal-Wallis test, due to the number of examples in the groups being small and equal. For the execution of the test, the following hypotheses were considered:

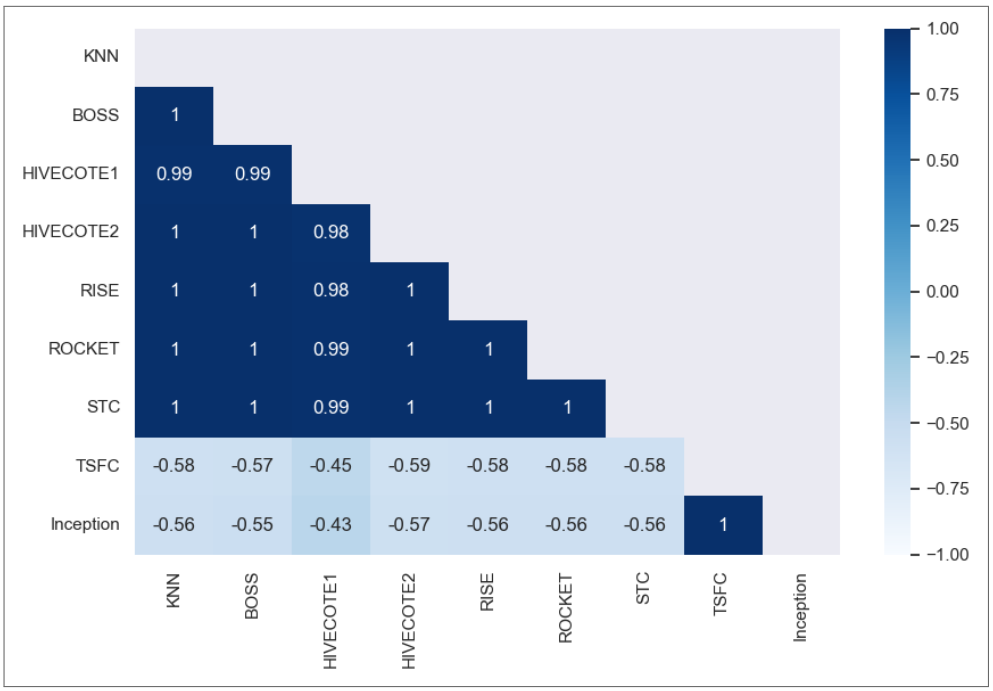
- H0: All models have relatively equal means in terms of classification accuracy;
- H1: At least one of the models differs from the others in terms of mean classification accuracy.

Where H0 is the null hypothesis, which assumes that all models have equal performance, and H1 is the alternative hypothesis, which in this case is the difference in performance of at least one of the models in relation to the others. For this test, a p-value < 0.05 indicates the rejection of the null hypothesis, showing the existence of a significant difference between the evaluated samples. Thus, with the application of Kruskal-Wallis to the set of results, a $p =$ 2.49E-02 was obtained, being less than 0.05, demonstrating with 95% confidence that there is evidence to reject H0 and accept the hypothesis that at least one of the models differs from the others in mean validation accuracy.

Knowing of the existence of this difference between the models, the next step was the application of a post-hoc test to identify which models are statistically different from each other, since the non-parametric test only indicates the existence of this difference, not the relationship between the sets. For this stage, the Nemenyi test was used, which according to Pohlert (2014) is one of the most used post-hoc tests after the application of Kruskal-Wallis. As briefly explained, this method performs a pairwise investigation of each analyzed set, returning the p-values for each relationship between the evaluated groups. The values vary between -1 and 1, with p-values < 0.05 indicating an effective statistical difference between the samples according to the test, and the closer to 1, the more it demonstrates the similarity between them. Figure 29 represents a correlation matrix that crosses the results obtained by the Nemenyi method.

As can be seen, there is a great similarity between most of the models that have a lower precision mean, not presenting a significant statistical discrepancy between them. However, it can be said that there is no significant difference between the Inception Time (FAWAZ et al., 2020) and Time Series Forest Classifier (BABAYEV; WIESE., 2021) models. Each has a great difference in relation to the models with more distinct accuracy, but a great similarity is evident between some with closer accuracy (which was already expected), dividing the models into distinct groups of relevance. Thus, it is possible to identify the difference between the

Figure 29 – Correlation graph of the results of the application of the Nemenyi post-hoc test on the set of results of each model. In this type of graph, the further from 1, it means that the elements are more divergent, that is, they are statistically different.

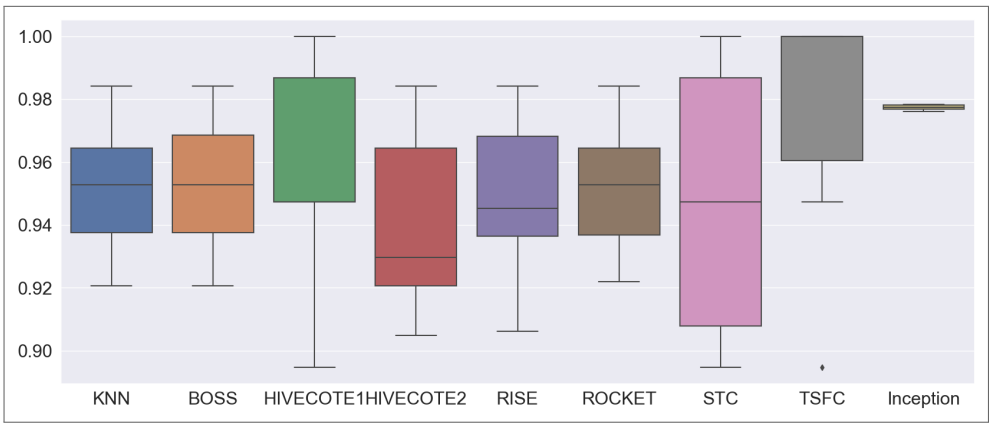


Source: Author

models, with Inception Time and TSFC not presenting a significant difference between them. In this case, even without a significant statistical difference, Inception Time stands out from the others with means of Accuracy, Precision, Recall, and F1 above 95% in all the analysis sets, making it the most promising choice for the final classification of the model of volatiles emitted by *Candida* species. These differences can also be identified from a boxplot (Figure 30), where, through the quartiles of the values, it is possible to get an idea of the differences about each of the groups of results.

In Figure 30, it is possible to note a great variation in the results obtained by each model, with InceptionTime being responsible for the greatest consistency in the values of its repetitions. With this, it is possible to have more security in the results obtained by this model, aiming for a verification with fewer variations. In this case, even without a significant statistical difference, InceptionTime stands out in relation to the others with means of Accuracy, Precision, Recall, and F1 close to 100% in all the analysis sets, making it the most promising choice for the final model for the classification of the volatiles emitted by the species.

Figure 30 – Boxplot of the values of each group of results of the accuracy of the used models. For each presented model, it is possible to see the variation of the results in relation to the median, with the model with the least variation of values being the InceptionTime model.



Source: Author

5.2.2 Scenario 2: Validation with Blood Broth

For the blood broth samples, the artificial intelligence models were evaluated using data derived from the Prototype version of the E-nose, with and without the application of *oversampling*. The *oversampling* strategy was employed to deal with the species imbalance observed in the data, while the original data were used to evaluate the performance without adjustment (MOHAMMED et al., 2020). The metrics used to evaluate the performance included accuracy, precision, F1-score, sensitivity (Recall), specificity, and execution time (Time (S)). The tested models included time series approaches such as KNeighbors Time Series Classifier (KNTC) (LEE et al., 2012), Random Interval Spectral Ensemble (RISE) (FLYNN et al., 2019), ROCKET Classifier (DEMPSTER et al., 2020), Time Series Forest Classifier (TSFC) (DENG et al., 2013), Inception Time (WANG et al., 2020b), and traditional models such as DecisionTree Classifier (DTC), KNeighbors Classifier (KNN), Random Forest Classifier (RF), SVC, and XGBClassifier (XGBC). The selection of time series models was based on the classifiers that showed the best performance in the previous approach, which used culture samples. For the traditional models, the strategy was to apply models from different families to understand if simpler models could have a performance as good as the more complex models in the current literature. To provide a better understanding, the general results of the tested models are presented in Table 10 and Table 12, which detail the metrics for the strategies with and without *oversampling*, for the Training/Validation and Test data, respectively.

Table 10 – The table compares the **training and validation performance** of different classification models, both traditional and time series, with and without *oversampling*, using the metrics of accuracy, precision, F1-score, recall (sensitivity), specificity, and standard deviations. It also presents the execution times (s) of each model in both scenarios. The best performances in each metric are highlighted in bold.

Category	Classifier	With Oversampling							Without Oversampling						
		Acc.	Prec.	F1	Recall	Spec.	Std Dev	Time(s)	Acc.	Prec.	F1	Recall	Spec.	Std Dev	Time(s)
Time Series	KNeighbors Time Series	98.85%	98.76%	98.80%	98.92%	99.77%	0.0225	5.39×10^{20}	94.44%	96.18%	94.28%	93.76%	98.78%	**0.0208**	2.1×10^{21}
	RISE	97.15%	97.03%	97.13%	97.32%	99.43%	0.0173	5.85×10^{20}	95.37%	96.73%	94.78%	93.94%	99.00%	0.0275	3.8×10^{20}
	ROCKET	**98.86%**	**99.10%**	**99.02%**	**99.10%**	**99.77%**	0.0175	6.17×10^{20}	**98.15%**	**97.77%**	**97.74%**	**97.99%**	**99.59%**	0.0232	4.7×10^{21}
	Time Series Forest	97.13%	96.48%	96.63%	97.04%	99.44%	0.0235	7.51×10^{20}	95.42%	96.58%	94.78%	94.92%	99.05%	0.0276	5.6×10^{20}
	Inception Time	96.82%	96.85%	96.72%	96.66%	99.36%	**0.0084**	4.66×10^{20}	94.85%	94.40%	94.45%	96.16%	98.98%	0.0844	2.1×10^{21}
Traditional	Decision Tree	94.29%	94.59%	93.76%	93.56%	98.86%	0.0298	0.794573	88.06%	88.02%	84.17%	85.38%	97.52%	0.0510	0.62
	KNeighbors	89.15%	89.48%	88.96%	88.86%	97.81%	0.0374	**0.369690**	85.32%	81.77%	79.95%	80.22%	97.01%	0.0367	**0.38**
	Random Forest	97.70%	**97.84%**	**97.49%**	97.32%	99.52%	0.0263	3.30×10^{20}	92.59%	92.40%	89.49%	88.67%	98.47%	0.0359	2.7×10^{16}
	SVC	**97.70%**	97.16%	97.41%	**97.83%**	**99.56%**	**0.0141**	1.31×10^{21}	93.59%	96.13%	91.56%	90.28%	98.68%	0.0275	1.1×10^{21}
	XGBClassifier	96.55%	96.26%	96.10%	96.01%	99.31%	0.0190	2.77×10^{21}	88.99%	87.36%	83.40%	85.53%	97.83%	0.0360	2.2×10^{21}

Table 12 – The table compares the **test performance** of different classification models, both traditional and time series, with and without *oversampling*, using the metrics of accuracy, precision, F1-score, recall (sensitivity), specificity, and time (s). It also presents the execution times of each model in both scenarios. The best performances in each metric are highlighted in bold. All models were run using their default settings.

Category	Classifier	With Oversampling						Without Oversampling					
		Accuracy	**Precision**	**F1-Score**	**Recall**	**Specif.**	**Time (s)**	**Accuracy**	**Precision**	**F1-Score**	**Recall**	**Specif.**	**Time (s)**
Time Series	KNeighbors Time Series	**97.27%**	**97.35%**	**97.19%**	**97.37%**	**99.46%**	6.62×10^{20}	94.12%	**94.07%**	**94.67%**	**96.16%**	98.86%	2.72×10^{21}
	RISE	94.55%	94.65%	94.40%	94.54%	98.92%	1.16×10^{21}	85.29%	89.44%	81.97%	81.38%	97.07%	7.46×10^{20}
	ROCKET	93.64%	95.33%	93.62%	93.70%	98.73%	1.05×10^{21}	**95.59%**	91.67%	93.52%	96.88%	**99.17%**	7.90×10^{20}
	Time Series Forest	89.09%	90.47%	89.21%	88.81%	97.81%	1.23×10^{20}	91.18%	90.50%	91.73%	93.68%	98.27%	0.97
	Inception Time	**97.25%**	**96.26%**	**96.69%**	**97.29%**	**99.48%**	1.86×10^{20}	91.18%	93.57%	92.80%	93.15%	98.14%	1.92×10^{21}
Traditional	Decision Tree	96.36%	96.43%	96.34%	96.42%	99.28%	0.0047	85.29%	79.71%	81.38%	88.27%	97.16%	0.0033
	KNeighbors	90.00%	90.26%	89.61%	89.51%	98.01%	0.0117	85.29%	84.40%	83.87%	84.32%	97.07%	0.0100
	Random Forest	98.18%	98.14%	98.14%	98.20%	99.64%	**0.0095**	**95.59%**	90.55%	**92.24%**	**95.23%**	**99.19%**	**0.0087**
	SVC	**98.18%**	**98.20%**	**98.22%**	**98.33%**	**99.64%**	0.0251	92.65%	**94.70%**	88.97%	86.88%	98.40%	0.0224
	XGBClassifier	96.36%	96.24%	96.24%	96.39%	99.28%	0.3867	92.65%	85.39%	87.25%	92.12%	98.66%	0.0812

In traditional models, the application of *oversampling* resulted in a significant increase in all metrics compared to the original data, suggesting that balancing the species contributed to improving the models' discrimination ability. This improvement, following the balancing of classes, suggests that the initial class imbalance may have slightly limited the models' ability to generalize effectively to unseen data, a limitation that was alleviated by the oversampling technique. In contrast, some time series models showed less robustness in both conditions, performing worse than some traditional models, regardless of the application of *oversampling*. A factor that may explain this slight difference is the amount of data. Because they are more complex models, they require a larger number of samples to outperform simpler models. Training time was also a distinguishing factor for traditional models, which took much less time to complete.

Even with some traditional models outperforming the time series ones, overall, the ROCKET model showed the best performance among the evaluated approaches, reaching an accuracy of 98.86%, a sensitivity (recall) of 99.10%, and a specificity of 99.77% in the training/validation set with *oversampling*. In the group without *oversampling*, these indicators were 98.15%, 97.77%, and 99.59%, respectively. Among the traditional models, the SVC had the best performance, with values very close to ROCKET, reaching an accuracy of 97.70%, a sensitivity of 97.83%, and a specificity of 99.56% for the group with *oversampling*. It was also the best traditional model in the original data group, with 93.59%, 96.13%, and 98.68% for the same indicators, respectively. On the other hand, ROCKET was the slowest model to complete training among all evaluated models in the original dataset and one of the worst in the group with *oversampling*.

Observing the data from the test set, we notice some interesting differences. ROCKET is no longer the best-performing model, losing its position to the KNeighbors Time Series Classifier and Inception Time in the group of time series models, and to the SVC and Random Forest Classifier in the group of traditional models. It is with the test data that we evaluate the efficiency of the model, as it is exposed to information it has never seen before, representing the scenario closest to a real context. In this sense, the more complex models, despite showing excellent performance in all metrics, are on par with the SVC and Random Forest Classifier, with less than 1% difference in all metrics compared to the best time series models. In addition, the time to obtain the prediction was often much shorter (0.009520s and 0.025079s, respectively), demonstrating that they are lighter alternatives compared to time series models.

Table 14 – Values for all metrics (accuracy, precision, F1-score, recall (sensitivity), specificity) collected by species for the SVC classifier, executed for data with and without *oversampling*. The table shows that there is a drop in the model's performance in relation to the *glabra_parapsi* and negative species. This suggests that this model needs more instances of these species to have a better performance in these specific contexts.

Species	ID	With Oversampling					Without Oversampling				
		Accuracy	Precision	F1-Score	Recall	Specif.	Accuracy	Precision	F1-Score	Recall	Specif.
glabrata	0	99.09%	94.44%	97.14%	100.00%	98.92%	97.06%	100.00%	88.89%	80.00%	100.00%
parapsilosis	1	99.09%	100.00%	97.30%	94.74%	100.00%	94.12%	91.30%	91.30%	91.30%	95.56%
glabra_parapsi	2	100.00%	100.00%	100.00%	100.00%	100.00%	98.53%	100.00%	66.67%	50.00%	100.00%
cryptococcus	3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
negative	4	99.09%	94.74%	97.30%	100.00%	98.91%	95.59%	76.92%	86.96%	100.00%	94.83%
albicans	5	99.09%	100.00%	97.56%	95.24%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

5.2.2.1 Best Models and Metrics by Species

To better understand how the models behave for each species, the same metrics used for training and testing were collected for the different species, focusing only on the models with the best performance in the testing stage. This step aims to identify which species the models make the most mistakes on and whether any species imbalance may raise concerns about the quality of the model's prediction. In this sense, among the tested models, the SVC stood out with the best overall results, reaching an accuracy of 98.18%, a precision of 98.20%, an F1-score of 98.22%, a sensitivity of 98.33%, and a specificity of 99.64% in the dataset with *oversampling*. In the group without *oversampling*, these indicators were lower, with an accuracy of 92.65%, a precision of 94.70%, and an F1-score of 88.97%.

The detailed metrics by species for this model are presented in Table 14. The results show that, while the species *C. albicans*, *C. neoformans*, *C. glabrata*, and *C. parapsilosis* had high sensitivity and precision, the species *glabra_parapsi* (the union of *Candida parapsilosis* and *glabrata* in the same Petri dish) and negative (uninfected sample) showed slightly lower performance, possibly due to the smaller number of represented samples, only in the group without *oversampling*. For the balanced data, all species were correctly identified by the model, without any species having any metric below 94%.

Among the time series models, Inception Time was chosen to be evaluated individually with the species, as it was one of the models with the best overall performance in the metrics and took the least time during the testing phase (among the time series models). In this case, it is notable that the model performs better, even with the unbalanced data. There is a slight drop in the numbers for the same species that presented difficulties for the SVC; however, this drop was much less significant in a general context. Based on this information, even with the

Table 15 – Values for all metrics (accuracy, precision, F1-score, recall (sensitivity), specificity) collected by species for the Random Forest classifier, executed for data with and without *oversampling*. As in Table 14, there is a drop in performance in the absence of *oversampling* for some species, suggesting the need for additional training instances.

Species	ID	With Oversampling					Without Oversampling				
		Accuracy	Precision	F1-Score	Recall	Specif.	Accuracy	Precision	F1-Score	Recall	Specif.
glabrata	0	100.00%	100.00%	100.00%	100.00%	100.00%	97.94%	96.00%	92.84%	90.00%	99.31%
parapsilosis	1	96.55%	89.23%	90.09%	91.05%	97.69%	93.53%	97.61%	87.89%	82.17%	99.33%
glabra_parapsi	2	99.82%	99.00%	99.49%	100.00%	99.78%	99.56%	100.00%	90.00%	85.00%	100.00%
cryptococcus	3	98.18%	97.50%	93.50%	90.00%	99.57%	95.00%	76.78%	85.18%	100.00%	94.43%
negative	4	98.55%	94.11%	95.58%	97.22%	98.80%	97.21%	87.15%	91.29%	97.00%	97.24%
albicans	5	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

original data containing a slight imbalance in relation to some species, there are alternatives that can efficiently handle this problem without major disadvantages. Table 15 shows the result of this evaluation.

5.2.2.2 Statistical Comparison Between Strategies with and without Oversampling

To evaluate whether the application of *oversampling* significantly impacts the performance of the classifiers, statistical tests were performed comparing the results before and after the technique. First, the Shapiro-Wilk test was performed to check the normality of the distributions of the two groups (with and without *oversampling*). Then, depending on the normality of the data, the paired t-test was applied for normal distributions and the Wilcoxon test for non-normal distributions (IMAM et al., 2014; PROUDFOOT et al., 2018).

The results of the normality tests indicate that, for most classifiers (see Figure 31), at least one of the groups does not follow a normal distribution (Shapiro-Wilk values < 0.05). This justifies the use of the Wilcoxon test for these cases, which is a non-parametric test suitable for comparisons between paired samples without assumptions of normality. Among the evaluated classifiers, those that showed statistically significant differences after the application of *oversampling* include:

- RISE: The paired t-test resulted in a p-value of 0.045, indicating that *oversampling* significantly impacted the performance of this classifier.
- KNTC: The Wilcoxon test showed a p-value of 0.0098, pointing to a significant difference between the two groups.
- Inception Time: The p-value obtained in the Wilcoxon test was 0.0019, reinforcing the

presence of a statistically significant impact.

- XGBC: The paired t-test obtained a p-value of 0.0018, indicating a significant difference between the groups.
- SVC: The p-value of 0.0039 obtained in the Wilcoxon test confirms a statistically significant difference.
- DTC: The paired t-test resulted in a p-value of 0.0074, showing a significant impact.
- RFC: The Wilcoxon test revealed a p-value of 0.0273, indicating a significant difference.

On the other hand, some classifiers, such as TSFC, ROCKET, and KNN, did not show statistically significant differences, with p-values above the significance level of 0.05. This suggests that, for these models, the introduction of *oversampling* did not result in statistically relevant improvements in performance.

In summary, the results indicate that the impact of *oversampling* varies among the different classifiers, being more pronounced in some models than in others. This observation reinforces the importance of individually evaluating the effectiveness of data balancing techniques before their final implementation.

5.2.2.3 Analysis of Time Series and Traditional Data

Time series models showed greater effectiveness in detecting patterns associated with *Candida* infections, especially in the dynamic analysis of the signals collected by the Electronic Nose. Among them, the KNeighbors Time Series Classifier and Inception Time achieved the best performances, reaching 97.27% and 97.25% accuracy, respectively, when *oversampling* was applied. These models maintained a high generalization ability even without *oversampling*, reinforcing their robustness in the face of unbalanced data.

In contrast, traditional models exhibited competitive performance, with a highlight for the Random Forest Classifier and the SVC, which reached 98.18% accuracy with *oversampling*. However, without this balancing technique, their accuracy dropped significantly, evidencing the dependence of these models on specific pre-processing steps to deal with unbalanced data. The application of *oversampling* had a widespread positive impact, especially for traditional models, significantly improving accuracy, precision, and F1-score. This highlights the importance of balancing strategies to optimize performance in scenarios with unequal distribution of species.

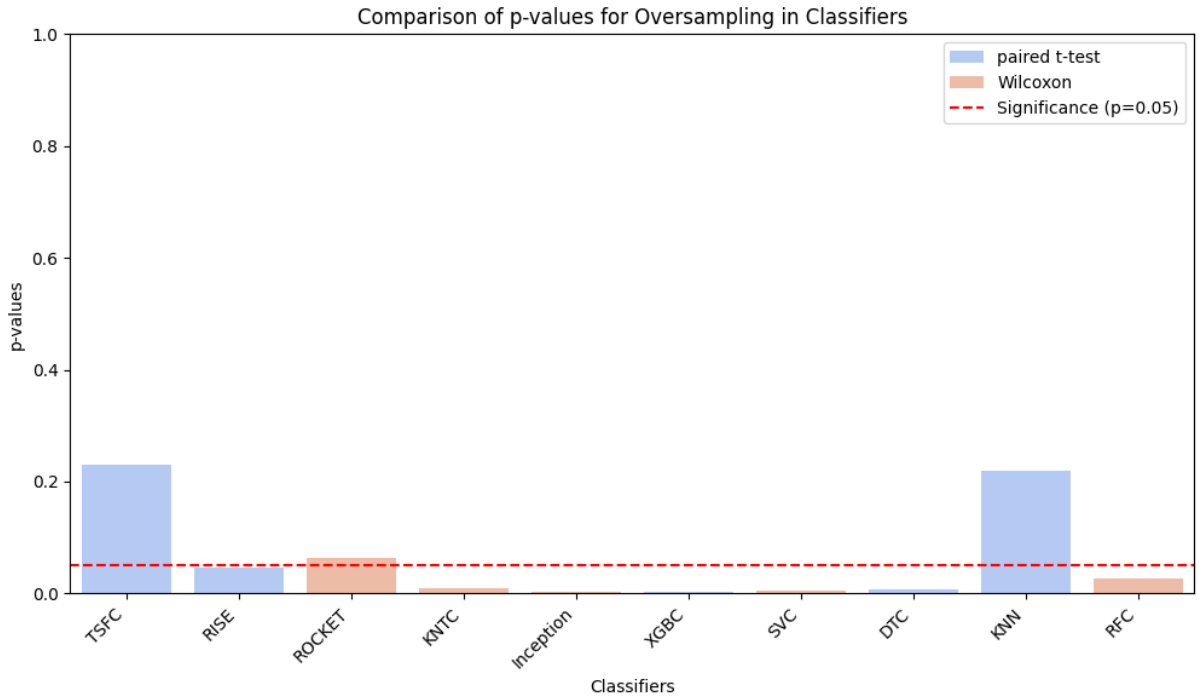


Figure 31 – Comparison of the p-values of the applied statistical tests, indicating which models were significantly impacted by the application of oversampling. The red line represents the confidence interval. Models with bars above this interval were not significantly affected by the use of the oversampling strategy.

Overall, time series models demonstrated greater flexibility and consistency in the analysis of non-stationary data, making them promising alternatives for the rapid and accurate diagnosis of fungal infections caused by *Candida*, without the need for extensive pre-processing adjustments.

These findings further emphasize the variable impact of *oversampling* on different classification models, reinforcing the need for personalized pre-processing approaches. As the results demonstrate, while traditional models benefit significantly from *oversampling*, time series models tend to exhibit an inherent robustness to species imbalance. This observation naturally leads to a more in-depth discussion about the comparative strengths and weaknesses of these two methodological approaches. The next section explores these aspects in more detail, analyzing not only the classification accuracy but also the computational efficiency and clinical applicability, ultimately aiming to determine the most suitable models for real-world diagnostic scenarios.

5.2.2.4 Main discussions

The results obtained with the AI models show significant performance differences between traditional and time series-based models. The application of oversampling proved to be effective in improving the classification metrics, reducing the impact of the imbalance between species in traditional models. In contrast, for time series models, this technique did not show such a pronounced impact, suggesting that these models maintain good performance regardless of the data balancing. This robustness is particularly relevant in clinical contexts, where the distribution of data can be highly variable.

Among the evaluated models, ROCKET showed excellent performance in the training and validation phases, standing out mainly in the version with oversampling, where it obtained the highest accuracy (98.86%) and sensitivity (99.10%). However, when exposed to the test data, its performance was surpassed by the KNeighbors Time Series Classifier and Inception Time among the time series models, and by the SVC and Random Forest Classifier among the traditional models. This suggests that, despite the high performance in the validation phase, some models may have a lower generalization ability, possibly due to the specific characteristics of the training set.

The analysis of the processing time revealed that traditional models are considerably faster than time series models. The SVC and Random Forest Classifier showed significantly shorter prediction times (0.009520s and 0.025079s, respectively), making them viable alternatives for applications that require computational efficiency without major compromises in accuracy. However, despite being slower, time series models like Inception Time offer a greater feature extraction capacity, proving to be more suitable for detecting subtle temporal variations in the E-Nose signals.

In the evaluation of metrics by species, it was observed that the SVC showed consistent performance, reaching an accuracy of 98.18% in the set with oversampling. However, in the set without oversampling, there was a noticeable drop, highlighting the importance of balancing species so that traditional models can detect subtle patterns in the E-Nose signals. In contrast, Inception Time showed greater robustness to the imbalance of the data, maintaining good metrics even without the application of oversampling. This suggests that time series-based models inherently capture temporal dependencies that contribute to their resilience in different data conditions.

From a clinical point of view, the ability to quickly detect fungal pathogens directly from

blood broth represents a transformative advance. Traditional culture-based methods for isolating and identifying pathogens in the bloodstream can take days, delaying treatment decisions. The E-Nose/AI approach based on blood broth reduces this time to minutes, bringing substantial benefits to antifungal management and patient treatment, especially in intensive care environments. By eliminating the need for sample preparation and fungal isolation, this method allows for faster interventions, which can be crucial for improving clinical outcomes and reducing the mortality associated with fungal infections.

In this scenario, diagnostic accuracy and speed are critical factors. While models like SVC and Random Forest Classifier offer a balance between performance and computational efficiency, more specialized approaches, such as KNeighbors Time Series Classifier and Inception Time, deserve further exploration. When evaluating the performance of the models by species, it becomes evident that more robust models, such as Inception Time, stand out in capturing temporal variations of the signals, even when some species have a smaller number of samples.

Deep learning models generally demonstrate a better generalization ability in scenarios with unbalanced datasets when compared to linear models like the SVC. This characteristic makes them especially suitable for complex medical diagnoses. However, there is a wide set of strong alternatives for model selection, which increases the reliability, portability, and scalability of the system. This is particularly important in challenging clinical conditions, where rapid and accurate diagnoses are essential for the proper management of patients and for therapeutic decisions.

When compared to the culture-based methodology, the results are quite similar. However, the blood broth-based approach has a significant advantage by eliminating the need for sample processing, thus avoiding exposure to VOCs from external sources, such as the agar used in fungal cultivation. This can lead to more accurate readings and improve the ability of the models to correctly identify patterns of volatile compounds.

Although the results are promising, there are several considerations for the practical application of this approach. The *drift* of the sensors, a known challenge in Electronic Noses, can affect the responses over time due to prolonged use or environmental factors, possibly requiring periodic recalibrations. In addition, the cross-sensitivity to VOCs from different microorganisms or external sources can influence the accuracy of the classification, reinforcing the importance of robust pre-processing techniques and the inclusion of diversified samples. Despite the high accuracy in the evaluated sets, multicenter validations are necessary to ensure the consistency of the results in different clinical scenarios and collection protocols. Continuous

efforts of refinement and broad validation are essential to increase the reliability and scalability of this method for practical use.

Another important aspect to consider is the comparison of this approach with established diagnostic techniques, such as MALDI-TOF. Although MALDI-TOF is considered the gold standard in microbial identification, due to its accuracy and reliability, it requires sample preparation, specialized equipment, and trained personnel — factors that can limit its accessibility in certain health environments. In contrast, the E-Nose/AI approach offers a fast, non-invasive, and low-cost alternative, reducing the diagnosis time and enabling early interventions. However, more studies are needed to comprehensively establish its competitive advantages and address possible gaps in sensitivity and specificity compared to traditional methods (PATEL, 2019).

In summary, considering the points above, the optimization of model selection involves balancing accuracy, computational efficiency, and clinical applicability. The analysis shows that, while traditional models benefit from the balancing between species, time series-based models, such as Inception Time, show greater robustness in the face of data imbalance, making them particularly suitable for clinical practice, where speed and diagnostic reliability are fundamental. The following section consolidates the main contributions of the study, discussing their implications and pointing out potential directions for future research and applications.

5.3 EVALUATION OF THE XAI ENSEMBLE METHOD AND THE *DIAGNOSE.AI* TOOL

To evaluate the XAI Ensemble module, two types of procedures were conducted: (i) a quantitative evaluation of the explanatory performance, by means of direct comparison tests, ablation study, and sensitivity analysis; and (ii) a quali-quantitative evaluation of the usability and interpretability of the results produced by the XAI Ensemble VOCs method. In this second stage, the Nielsen Heuristics (NIELSEN, 1994) were used as a reference, combining the quantitative analysis of the response percentages with the qualitative analysis of the participants' feedback. The following subsections detail the results obtained in each of these stages.

5.3.1 Quantitative Evaluation of the Method

To evaluate the performance of the method in different data contexts, the tests were conducted on the two datasets. The analysis was structured in three sequential stages, following

the best practices for validating explainability methods (HOOKER et al., 2019): (i) a direct comparison of the agreement between the LIME, SHAP, and Grad-CAM methods; (ii) an ablation study to evaluate the contribution of each method to the ensemble's consensus; and (iii) a sensitivity analysis to test the robustness of the explanation against perturbations in the input data (ALVAREZ-MELIS; JAAKKOLA, 2018).

To ensure a comparative evaluation on both datasets, we adopted a specific sampling strategy. For the *Sensitivity Analysis*, we randomly selected six samples from the test set of each dataset, ensuring the representation of all six classes in both contexts. The results presented in this section correspond to the mean and standard deviation calculated from these subsets. For the other subsections – *Direct Comparison* and *Ablation Study* – we selected the same representative instance, *Candida parapsilosis*, for the contexts of **blood broth** and **culture**. This species was chosen consistently in both datasets due to its high prevalence in the dataset and the high confidence exhibited by the model in its predictions, thus allowing for a direct comparison of the behavior of the explainers under different data conditions.

5.3.1.1 Agreement between Explanation Methods (Direct Comparison)

The initial analysis focused on quantifying the level of agreement between the three XAI methods on the two datasets. This addresses the well-known "disagreement problem" documented in the literature (KRISHNA et al., 2022). Table 16 summarizes the agreement counts, highlighting the most relevant features identified in each experiment.

Table 16 – Feature Agreement Count between XAI Methods for Blood Broth and Culture experiments.

Experiment	**Sensor (Feature)**	**Agreement Count**
Blood Broth	MQ-138	3
	MQ-3	3
	MQ-7	2
Culture	TGS-2611	2
	TGS-823	2
	TGS-822	2

The primary output of XAI methods is the **feature importance score (weight)**, which quantifies the contribution of each input variable (sensor) to the model's prediction. The analysis of these scores across multiple methods allows the establishment of consensus and disagreement regarding feature relevance.

For the **blood broth** data, a **unanimous consensus** (count = 3) is observed for the **MQ-138** and **MQ-3** sensors, indicating that, regardless of the approach of each method, both are consistently identified as highly relevant.

- **Example of Agreement (Consensus):** This unanimous consensus is a clear example of **strong agreement**, as all three XAI methods ranked MQ-138 and MQ-3 as the most important features.

The MQ-7 sensor also showed relevance, being identified by two of the three methods.

- **Example of Disagreement (Partial Consensus):** This result, where one method disagreed with the other two (count = 2), illustrates a point of **disagreement** among the XAI methods regarding the relative importance of MQ-7.

In contrast, the analysis of the **culture** data did not result in a unanimous consensus. However, a partial consensus (count = 2) was found for a different set of sensors: **TGS-2611**, **TGS-822**, and **TGS-823**.

These combined results suggest that, although the specific features of greatest importance may vary depending on the experimental medium, the ensemble's consensus approach is consistently effective in identifying a central set of relevant features in both scenarios. This increases robustness against potentially misleading explanations, as demonstrated by (SLACK et al., 2020).

5.3.1.2 Contribution of the Methods to the Ensemble (Ablation Study)

To understand the influence of each method on the result of the ensemble, an ablation study was performed (HAMEED et al., 2022) on both datasets. The ensemble was executed in its full configuration (LIME + SHAP + Grad-CAM) and in three ablation configurations, each removing one of the methods. Table 17 compares the consensus features identified in each scenario for both experiments.

The results of the ablation of the **blood broth** data reveal that the **MQ-3** sensor is the most stable feature, remaining in the consensus even with the removal of LIME or Grad-CAM, which consistently confirms it as the most robust and reliable feature. The **MQ-138** sensor also consistently appears as a relevant feature in multiple ablation scenarios, reinforcing its importance.

Table 17 – Consensus Features Identified in the Ablation Study for Blood Broth and Culture experiments.

Experiment	Ensemble Configuration	Identified Consensus Features
Blood Broth	**Complete (LIME + SHAP + GRAD)**	**{MQ-3, MQ-138, MQ-7}**
	Ablation (without SHAP)	{MQ-138, MQ-7}
	Ablation (without Grad-CAM)	{MQ-135, MQ-3}
	Ablation (without LIME)	{MQ-138, MQ-3}
Culture	**Complete (LIME + SHAP + GRAD)**	**{TGS-2611, TGS-822, TGS-823}**
	Ablation (without SHAP)	{TGS-822, TGS-823}
	Ablation (without Grad-CAM)	**{TGS-822, TGS-823}**
	Ablation (without LIME)	{TGS-2611}

For the culture data, a different and very revealing pattern emerges. The **TGS-822** and **TGS-823** sensors show remarkable stability, forming the consensus not only when SHAP is removed, but also when Grad-CAM is removed. This strongly suggests that this pair represents the central and most reliable features for this dataset, consistently identified by the LIME and SHAP combination. In addition, the removal of LIME isolates **TGS-2611** as the only consensus feature, indicating a unique contribution of the SHAP and Grad-CAM pair.

This comparative study highlights the value of ablation analysis. While the blood broth experiment points to a single highly dominant feature (MQ-3), the culture experiment reveals a stable "pair" of features (TGS-822, TGS-823). This demonstrates that the specific context of the data critically influences how the XAI methods interact and which features are highlighted, reinforcing the need for an ensemble approach to achieve a comprehensive and reliable explanation.

5.3.1.3 Robustness of the Explanation (Sensitivity Analysis)

The evaluation of an explanation's quality often relies on assessing its fidelity (how well it reflects the model's behavior) and its stability or robustness (how much the explanation changes when the input data is slightly perturbed). In this study, we primarily evaluate the ensemble's explanation by focusing on two key aspects: Robustness through Sensitivity Analysis and Inter-Method Consistency.

The robustness of the ensemble's explanation was evaluated by introducing Gaussian noise

to the original test instances, an essential "sanity check" to validate the fidelity of the explanation (ADEBAYO et al., 2018). This perturbation approach simulates minor, real-world variations in the input data, and a stable explanation should not change drastically. The stability of the explanation was measured quantitatively for the two experimental configurations. The results are summarized in Table 18.

Table 18 – Explanation robustness metrics from the sensitivity analysis for both Blood Broth and Culture experiments. The reported values correspond to the average results obtained among the 6 species in each dataset. For each species, 3 samples were generated with noise perturbation. First, the average of the 3 perturbations was calculated per species; then, the overall average and standard deviation were calculated among all 6 species for each experimental configuration.

Data Source	Metric	Average Value	Standard Deviation
Blood broth	Jaccard Index	0.917	0.105
	Kendall's Tau	0.530	0.314
	Overlap Coefficient	0.981	0.045
Culture	Jaccard Index	0.972	0.062
	Kendall's Tau	0.815	0.270
	Overlap Coefficient	0.981	0.041

The results of both experimental configurations provide strong evidence of the method's robustness. For the **blood broth data**, the high values observed for the Jaccard Index (91.7%) and the Overlap Coefficient (98.1%) indicate excellent stability in the *set* of features identified by the ensemble method, even under noise perturbation. Additionally, the moderate value of Kendall's Tau (53.0%) suggests a reasonable level of consistency in the *ranking* of the importance of the features.

These findings are reinforced by the analysis of the **culture data**, which showed even higher stability metrics. With a Jaccard Index of 97.2% and a remarkably high Kendall's Tau of 81.5%, this second experiment demonstrates that the ensemble not only consistently identifies the same central features but also maintains its relative importance ranking with high fidelity. The consistency of the high Overlap Coefficient (98.1% in both experiments) highlights the reliability of the method. The strong performance on two different data sources significantly increases confidence in the ensemble approach, demonstrating that it is robust both in identifying *which* features are relevant and in maintaining a coherent *order of importance*, aligning well with the central objectives of explainable AI.

To comprehensively evaluate the reliability of the ensemble approach in different data contexts, we performed a second evaluation step: a consistency analysis between methods on

the two datasets. This analysis compares the output of the ensemble with each individual explanation technique (LIME, SHAP, and Grad-CAM). High consistency indicates that the ensemble's explanation is not only stable to noise but also highly aligned with the core findings of its constituent methods. The results, detailed in Table 19, reflect the alignment of the ensemble with its constituent methods in each scenario.

Table 19 – Similarity Between the Ensemble and Individual XAI Methods for the Blood Broth and Culture Datasets (Mean and Standard Deviation of 6 repetitions). Here, μ represents the mean and σ the standard deviation.

Comparison	**Jaccard**		**Overlap**		**Kendall's Tau**	
	μ	σ	μ	σ	μ	σ
Blood Broth						
Ensemble vs Grad-CAM	0.792	0.188	0.944	0.136	-0.556	0.544
Ensemble vs LIME	0.875	0.137	1.000	0.000	1.000	0.000
Ensemble vs SHAP	0.792	0.188	0.944	0.136	0.444	0.655
Culture						
Ensemble vs Grad-CAM	0.751	0.201	0.910	0.150	-0.215	0.600
Ensemble vs LIME	0.850	0.152	0.980	0.025	0.950	0.050
Ensemble vs SHAP	0.763	0.195	0.925	0.141	0.350	0.710

The results demonstrate a consistently high agreement in the selection of features (Jaccard Index and Overlap Coefficient) for both datasets, indicating that the ensemble and the individual methods tend to identify a similar set of important features. However, significant variations arise when analyzing the ranking of these features, measured by Kendall's Tau, revealing a context-dependent behavior of the explainers.

For the **blood broth** dataset, the ensemble exhibits a perfect and stable agreement with LIME (1.000 ± 0.000), a moderate positive agreement with SHAP (0.444 ± 0.655), and a notable moderate negative correlation with Grad-CAM (-0.556 ± 0.544).

In the **culture** dataset, this divergence in ranking is even more pronounced. Although the agreement with LIME remains very high and stable (0.950 ± 0.050), the correlations with SHAP (0.350 ± 0.710) and Grad-CAM (-0.215 ± 0.600) become weaker and exhibit even greater variability, as indicated by the high standard deviations.

This divergence, especially the instability observed in the correlations with SHAP and Grad-CAM in both scenarios and its intensification in the culture dataset, reinforces the justification for the use of an ensemble. It successfully harmonizes complementary — and sometimes

conflicting — perspectives from different explanation paradigms, resulting in a more robust and reliable feature identification in various analytical contexts.

5.3.1.4 Comparative Performance and Computational Cost Analysis of the XAI Ensemble

The validation of an XAI architecture for clinical application depends not only on the robustness of its explanations but also on its computational efficiency and adaptability to different contexts. This section presents a comparative performance analysis of the XAI *ensemble*, evaluating its cost (execution time) and benefit (relevant sensors identified) in the two experimental scenarios: the controlled culture environment (ATCC) and the clinical blood broth scenario.

The objective is to demonstrate the methodology's flexibility and identify the trade-offs between explanatory completeness and practical viability, a critical factor for implementation in the DiagNose.AI system. To facilitate direct comparison, the results from both scenarios are consolidated into a single table.

Table 20 – Comparative Cost-Benefit Analysis of the XAI Ensemble in the Culture and Blood Broth Scenarios.

Experimental Scenario	**Ensemble Configuration**	**Time (s)**	**Consensus Sensors Identified**
Culture (ATCC)	LIME + SHAP + Grad-CAM (Full)	2248.95	TGS-2611, TGS-822, TGS-823
	LIME + Grad-CAM	**21.38**	TGS-822, TGS-823
	LIME + SHAP	2152.37	TGS-813
	SHAP + Grad-CAM	2119.81	TGS-2611
Blood Broth	LIME + SHAP + Grad-CAM (Full)	1273.65	MQ-138, MQ-3, MQ-7
	LIME + Grad-CAM	**16.79**	MQ-138, MQ-7
	LIME + SHAP	1290.86	MQ-3
	SHAP + Grad-CAM	1266.50	MQ-138, MQ-3

The consolidated analysis of Table 20 allows for three main conclusions about the XAI *ensemble* architecture:

1. Computational Cost and the Consistent Pattern of SHAP: Data from both scenarios unequivocally confirm that the SHAP technique is the main performance bottleneck. In both experiments, any combination including it increases the execution time to tens of minutes (between 21 and 37 minutes). In contrast, the LIME + Grad-CAM combination proves

to be extremely fast in both contexts, with times of 21.38s (culture) and 16.79s (blood broth). The faster execution in the blood broth scenario, despite a similar number of cycles, can be attributed to the lower number of sensors in the Prototype E-nose (4 sensors) compared to the Suitcase (7 sensors), which reduces the data dimensionality and, consequently, the processing cost.

2. Framework Adaptability to Different Hardwares: It is crucial to highlight that each experimental scenario used a distinct E-nose device, each equipped with an exclusive set of sensors: the TGS family for the culture experiment and the MQ family for the blood broth experiment. The analysis of the XAI *ensemble* (Table 20) demonstrates that the methodology was able to identify the most pertinent sensors within each of these distinct hardware sets. This finding indicates the Framework's ability to adapt to different hardware configurations and extract the most informative features available in each architecture. The method's robustness is evidenced by its ability to operate effectively and provide coherent explanations, regardless of the underlying sensors.

3. The Trade-off between Completeness and Efficiency: The comparison between the scenarios reinforces the existence of a strategic trade-off.

- The full *ensemble* (with all three methods) consistently provides the most comprehensive and hardware-adapted explanation, making it the ideal choice for in-depth research analyses where time is not a limiting factor.

- The LIME + Grad-CAM configuration establishes itself as the best cost-benefit option for practical application. It offers an ultra-fast (under 30 seconds) and robust explanation in both scenarios, capturing a significant subset (50% for culture and 66% for blood broth) of the most important sensors identified by the complete analysis on their respective hardware.

The main limitation identified in this analysis is the high computational cost associated with SHAP, which points to a clear direction for future work: the investigation of alternative explainability techniques. The objective would be to find or develop a method that, by replacing SHAP in the *ensemble*, can maintain or even enhance the robustness and depth of the explanation but with significantly superior performance. Such an advancement would allow the full *ensemble* configuration to be used more viably in time-critical scenarios, eliminating the current trade-off and consolidating an explanatory solution that is both complete and agile.

In summary, the comparative analysis, with complete data from both scenarios, validates the XAI Ensemble architecture as a robust and, most importantly, flexible and adaptable solution. The Framework's ability to operate on different hardware and provide contextual explanations, combined with the option to configure the *ensemble* to prioritize either analytical completeness or response speed, gives DiagNose.AI the versatility needed to transition between the research environment and real-time clinical application.

5.3.2 Key Discussions on the XAI Ensemble VOCs Method

The results demonstrate that the proposed XAI Ensemble provides explanations that are both robust and context-sensitive, addressing the well-known disagreement problem in single-method approaches (KRISHNA et al., 2022). In the **blood broth** dataset, **MQ-3** and **MQ-138** consistently emerged as the most reliable features, while in the **culture** dataset, the pair **TGS-822** and **TGS-823** showed the greatest stability. This context-dependent behavior reinforces the need for an ensemble strategy.

The ablation study revealed that the removal of a single method significantly alters the consensus, highlighting the risks of relying on a single technique. By integrating LIME, SHAP, and Grad-CAM, the ensemble harmonizes complementary perspectives and mitigates the vulnerabilities of individual explainers (SLACK et al., 2020).

The sensitivity analysis further confirmed the robustness, with high Jaccard and Overlap values ($>$0.91 and $>$0.98) ensuring stability in the set of identified features, even under perturbations. Although the feature rankings (Kendall's Tau) varied, the central set of relevant features remained consistent, a more reliable basis for practical applications (ALVAREZ-MELIS; JAAKKOLA, 2018; ADEBAYO et al., 2018).

Furthermore, the performance analysis highlights a critical trade-off between computational cost and explanatory completeness. The full three-method ensemble, while offering the most comprehensive feature set, is computationally expensive due to the overhead imposed by the SHAP method, with execution times of 1273.65s for blood broth and 2248.95s for culture data. This high cost, primarily attributed to SHAP's extensive calculation time, necessitates a new literature survey to identify alternative explanation methods that offer comparable explanatory completeness with a drastically lower computational cost. In contrast, the partial ensemble of LIME and Grad-CAM provides an ultra-fast alternative, yielding explanations in just 16.79s and 21.38s, respectively. This demonstrates the Framework's flexibility, allowing for a configuration

optimized for speed that still captures a significant subset of the key features, making it viable for time-sensitive clinical applications.

Finally, the ensemble aligned more stably with LIME, while SHAP and Grad-CAM exhibited greater variability, reinforcing the benefit of combining multiple paradigms. Overall, the Framework delivers explanations that are not only interpretable but also verifiably robust and reliable, essential for high-risk domains like healthcare.

5.3.3 Quantitative Evaluation - An Analysis of Nielsen's Heuristics

To evaluate the usability and identify the most effective way to present the information to the user, a medium-fidelity prototype was developed[1]. The evaluation was guided by Nielsen's Heuristics (NIELSEN, 1994) and supplemented by detailed qualitative feedback. 11 responses were collected through an online questionnaire from two distinct groups of participants: technical experts (computing, design, software testing – 45.45%) and experts in the prototype's application domain (Medicine, Biomedicine, and Biology – 54.55%). The questionnaire combined closed questions to measure the heuristics and open questions to capture perceptions and suggestions.

The quantitative analysis of the responses revealed a generally positive perception regarding the prototype's usability. Notably, no heuristic was evaluated exclusively as "Does not meet" by all participants, suggesting that the prototype demonstrates a basic level of compliance with all evaluated heuristics. However, the variation in responses between "Yes" (meets the heuristic) and "Partially" (partially meets the heuristic) indicates specific areas that require attention and refinement:

- **H1 - Visibility of system status:** The vast majority (81.82%) considered it "Yes," demonstrating that the prototype provides adequate feedback about what is happening.
- **H2 - Match between system and the real world:** Showed a division, with 63.64% "Meets" and 36.36% "Partially." This suggests that, although the language is generally understandable, some technical terms need adaptation.
- **H3 - User control and freedom:** The majority (72.73%) evaluated it as "Yes," indicating ease of navigation and undoing actions, but the 27.27% "Partially" point to areas for refinement.

[1] <https://bit.ly/3HCN0ia>

- **H4 - Consistency and standards:** The vast majority (81.82%) indicated that the prototype follows consistent conventions, although 18.18% "Partially" suggest points for improvement.
- **H5 - Recognition rather than recall:** Predominantly (72.73%), users considered that the prototype facilitates recognition, with 27.27% indicating "Partially."
- **H6 - Aesthetic and minimalist design:** Received the highest percentage of "Yes" (90.91%), showing that the design is considered clean and focused.
- **H7 - Help and documentation:** The responses were the most divided, with 63.64% "Yes" and 36.36% "Partially," suggesting that the accessibility or clarity of the help could be improved.

5.3.4 Qualitative Results: Perceptions and Suggestions from Users

The qualitative analysis of the feedback revealed valuable insights. Many participants praised the clear and detailed way the system explains the AI results, highlighting the use of graphs and legends as facilitators. Health professionals mentioned its effectiveness in the rapid and safe identification of microorganisms, stressing the importance of comparison with other results. The general ease of use was a recurring positive point. However, several suggestions for improvement were proposed. Several users suggested UI improvements, such as a more refined design, more centralized elements, and less "cartoonish" and more professional icons. The clarity of the information was a concern, with some pointing out difficulty in understanding certain graphical elements and the initial screen. Health professionals emphasized the need for greater detail about the process and technology, similar to established products, including information on sensitivity, analytical reproducibility, and model logic. Understanding the technical scope was considered crucial for appropriate analogies. The usability evaluation indicated strengths (clarity of the AI explanation, minimalist design) and areas for improvement (UI/UX, clarity of information, methodological detail). The final result of the implemented system can be observed in Appendix B of this project.

5.3.5 Validation and Advancement of the Framework: From Culture to the Clinical Scenario

The application of the DiagNose.AI Framework in two distinct contexts — in laboratory cultures (ATCC) and directly in clinical blood broth samples — allowed not only to validate the methodology but also to demonstrate its evolution and practical relevance. The transition from the controlled scenario to the clinical one revealed significant methodological advances.

The main advance refers to the diagnostic response time. While the culture-based approach requires preparation and incubation steps, the application of the Framework to blood broth eliminates these prerequisites, allowing for an almost immediate diagnosis. This optimization represents a drastic reduction in analysis time and laboratory resources, aligning the solution with the urgent needs of the clinical environment.

Additionally, the clinical validation demonstrated the robustness and flexibility of the Framework. The use of a more compact and portable Electronic Nose, with a reduced set of sensors, proved the adaptability of the methodology. Notably, even with a more simplified hardware, the approach not only maintained its effectiveness but also obtained superior accuracy results (98.18% vs. 97.46%), indicating that the computational core of the Framework is capable of extracting significant patterns even under different hardware conditions and greater sample complexity.

Table 22 below synthesizes the main results, connecting each validation scenario to the performance of the best predictive model and the main sensors identified by the XAI Ensemble architecture. This consolidated view demonstrates the Framework's ability not only to classify with high precision but also to provide interpretable insights into the patterns that guided its decisions.

Table 22 – Synthesis of the DiagNose.AI Framework validation, correlating the model's performance (training and testing) with the main sensors identified by the XAI methodology in each experimental scenario.

Database	**Best Model**	**Accuracy (Train)**	**Accuracy (Test)**	**Ref. Species (XAI)**	**Relevant Sensors (via XAI)**
Culture-UFPE	InceptionTime	97.74%	97.46%	*C. parapsilosis*	TGS-2611, TGS-813, TGS-822, TGS-823
BloodBroth-UC	SVC	97.70%	98.18%	*C. parapsilosis*	MQ-138, MQ-3, MQ-7

6 CONCLUSION

This work presented the development, implementation, and validation of the DiagNose.AI Framework, a systematic methodology for the identification of microorganisms from the analysis of Volatile Organic Compounds (VOCs). In response to the critical need for fungal infection diagnostics that are faster, more accurate, and accessible, this work focused on consolidating a complete, robust, and fundamentally explainable workflow.

The methodology was validated in scenarios of increasing complexity, starting in a controlled laboratory environment with fungal cultures and evolving to a clinical context close to the real application scenario, with the use of blood broth. This validation was made possible by the use of two distinct versions of the Electronic Nose — one more robust and the other more portable — proving the Framework's adaptability to different hardware conditions and sample types. A fundamental pillar of this trajectory was the construction of two novel databases, with VOC data from culture and blood broth, which not only served as a foundation for the experiments but also stand as a valuable resource for the scientific community.

The effectiveness of the Framework's analytical components was proven by expressive results. The time-series classification models showed high predictive power, achieving accuracies above 97% in both validation and test scenarios. However, the main computational contribution lies in the pioneering Ensemble XAI architecture. By complementarily integrating the LIME, SHAP, and Grad-CAM techniques within this context, the explainability methodology proved capable of providing consistent and multifaceted interpretations. Its robustness was confirmed in sensitivity and ablation studies, which demonstrated high stability in identifying the most relevant sensors (Jaccard Index > 0.91), even under perturbation of the input data. In addition, the performance analysis revealed the ensemble's practical flexibility, offering an ultra-fast configuration capable of delivering robust explanations in under 30 seconds (in the scenario without SHAP), establishing a crucial balance between analytical depth and the speed required for clinical applications.

Furthermore, the approach fills important gaps in the literature by generating textual and semantically rich explanations, with an interface designed with a focus on the medical domain and based on the needs of end-users, ensuring that the interpretations are not only correct but also useful and actionable for specialists.

Despite the promising results, it is recognized that the consolidation of the Framework

opens new research fronts and challenges to be faced. Generalization to an even larger number of species and the investigation of mixed infections are important future steps. Issues inherent to sensor technology, such as stability and the possibility of drift over time, will demand the continuous refinement of the models and the possible creation of a calibration protocol. The expansion of the database through multicenter studies will be fundamental to further increase the solution's robustness.

In summary, this project fulfills its objective by delivering a complete, end-to-end validated methodological Framework. The successful integration of an acquisition protocol, data engineering, predictive modeling, and, crucially, a robust explainability approach, represents an innovative solution with significant clinical impact. The adoption of technologies like that of the DiagNose.AI Framework has the potential to transform diagnostic practice, reduce associated mortality from infections, and consolidate the application of trustworthy Artificial Intelligence in the clinical environment.

7 MAIN CONTRIBUTIONS AND FUTURE PERSPECTIVES

The conclusion of this study has resulted in a set of methodological, software, and data contributions that establish a new approach for the identification of microorganisms. The consolidated work is presented below, detailing the milestones achieved and the promising lines of research that originate from it.

The main contributions developed and validated throughout this research are:

- **A Methodological Pattern for Sensor-based Systems:** The main contribution to Computer Science is the design of a complete, end-to-end methodological framework. This pipeline, which systematically integrates an acquisition protocol, data engineering, time-series modeling, and a native Explainable AI layer, establishes itself as a **replicable methodological pattern**. Its structure can be adapted for other areas that rely on the interpretation of sensor data, such as environmental sensing, Industry 4.0, and IoT networks.

- **Advances in Explainable AI for Multivariate Time Series:** A pioneering XAI architecture based on an ensemble method was developed. This approach is not limited to health diagnostics; it represents an advance for the analysis of multivariate time series in any domain where model transparency and reliability are crucial. Its applications extend beyond healthcare and can be employed, for example, in anomaly detection in sensor networks or in data analysis for predictive maintenance.

- **Creation and Availability of Novel Databases:** A fundamental pillar was the generation and characterization of two new databases for the analysis of VOCs from *Candida spp.*, a resource that was previously scarce in the literature. The *Culture-UFPE* database was created in a controlled laboratory environment, cultivating *Candida* isolates over different time periods. The *BloodBroth-UC* database, in turn, was developed in a laboratory using infected blood broth samples, an approach that significantly approximates the real-world application scenario of the solution and expands the variability of the samples, making the dataset more heterogeneous, especially when new collections are conducted in Brazil.

- **Scientific Production and Academic Recognition:** The relevance of the research was validated by the scientific community through the publication of an article in the journal

Scientific Reports of the Nature group (Appendix A), the submission of two new articles (one on the experiments with blood broth and another on the XAI Ensemble method) currently under peer review, and the approval in the CAPES/PrInt program, which enabled the international collaboration with the University of Cincinnati, a fundamental step for the consolidation of the Framework.

7.1 FUTURE PERSPECTIVES AND RESEARCH DIRECTIONS

From the completion of this work, several perspectives arise to guide future research, which can be organized into the following directions:

- **Sensor Drift and System Robustness:** One limitation of current E-nose systems is the gradual drift of sensor responses over time, which can degrade predictive performance. Future work should investigate drift compensation techniques, such as adaptive calibration, transfer learning, or domain adaptation strategies, ensuring the long-term robustness of the Framework.
- **Expansion to New Clinical and Nonclinical Domains:** A natural direction is to apply the methodological pipeline to other clinically relevant microorganisms, such as antibiotic-resistant bacteria, for the detection of microbial biofilms on surfaces and for monitoring microorganisms in laboratory settings. Beyond healthcare, the Framework can be adapted to monitor microorganisms in food safety and agriculture, for example.
- **Integration with State-of-the-Art Models:** While this thesis employed Inception-Time and SVC as core models, recent advances in time-series Transformers (e.g., TimesNet, TST, and PatchTST) represent a promising research avenue. Future studies may explore hybrid architectures combining convolutional and transformer-based models to improve both accuracy and interpretability.
- **Optimization of the XAI Ensemble for Real-Time Performance:** The performance analysis in this thesis consistently identified the SHAP method as a significant computational bottleneck, limiting the practicality of the complete three-method ensemble in time-critical scenarios. A promising research direction is the exploration of alternative, more efficient XAI techniques to replace SHAP. The goal would be to find a method that preserves or enhances the explanatory robustness of the ensemble while drastically

reducing execution time. This would resolve the current trade-off between speed and analytical depth, enabling the most comprehensive version of the Framework to be deployed in real-time diagnostic environments.

- **Multimodal Diagnostic Models:** A high-impact line of research involves the fusion of E-nose data with other modalities, such as clinical metadata or imaging. Multimodal integration may provide richer diagnostic insights, pushing the frontier of AI-assisted diagnosis toward more holistic patient profiles.

- **Usability and Clinical Adoption of XAI:** Although the ensemble-based XAI layer increased interpretability and robustness, its real adoption in clinical settings remains an open question. This thesis has already applied Nielsen's heuristics in the evaluation of a high-fidelity prototype, providing initial evidence of usability. The next step, however, is the validation of the *final system* with physicians in real clinical environments. Future work should involve larger and more diverse sets of healthcare professionals, assessing dimensions such as acceptance, trust, decision-making support, and cognitive workload during actual diagnostic routines. These studies are essential to bridge the gap between experimental validation and practical clinical adoption.

- **Large-scale Validation and Technology Readiness:** The next step for practical application involves multicenter validation studies to assess predictive performance and clinical utility in heterogeneous environments. Crucially, this includes expanding the dataset to encompass a wider range of biological variations, specifically focusing on intra-species variability within *Candida albicans*, *Candida parapsilosis*, and other prevalent species to ensure model robustness. Furthermore, the framework must undergo rigorous double-blind testing compared against current gold-standard methods (such as blood cultures and molecular assays). Finally, usability assessments of the *DiagNose.AI* Framework *in situ* will be fundamental for advancing its Technological Readiness Level (TRL).

In summary, while this research establishes a solid methodological, computational, and experimental foundation, its continuity depends on addressing sensor-related challenges, adopting state-of-the-art models, expanding to new domains, and ensuring clinical usability. These directions point toward the consolidation of E-nose systems as reliable, explainable, and widely adopted diagnostic tools.

REFERENCES

ABDI, H.; WILLIAMS., L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics 2.4*, p. 433–459, 2010.

ADADI, A.; BERRADA, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, v. 6, p. 52138–52160, 2018.

ADEBAYO, J.; GILMER, J.; MUELLY, M.; GOODFELLOW, I.; HARDT, M.; KIM, B. Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2018. v. 31.

AGUIRRE, T. da F.; GARCEZ, K. M.; SCHARDONG, P. L.; DEBLE., A. S. D. O. A importÂncia dos fungos para a humanidade. *ANAIS CONGREGA MIC-ISBN 978-65-86471-05-2*, 2016.

AKSEBZECI, B. H.; ASYALı, M. H.; KAHRAMAN, Y.; ER Özgür; KAYA, E.; ÖZBILGE, H.; KARA, S. Classification of root canal microorganisms using electronic-nose and discriminant analysis. *BioMedical Engineering OnLine, V.9*, p. 1–13, 2010.

ALAM, M. Z.; ALAM, Q.; JIMAN-FATANI, A.; KAMAL, M. A.; ABUZENADAH, A. M.; CHAUDHARY, A. G.; AKRAM, M.; MYCOLOGY., A. H. C. identification: a journey from conventional to molecular methods in medical. Candida identification: a journey from conventional to molecular methods in medical mycology. *World Journal of Microbiology and Biotechnology, V.30*, p. 1437–1451, 2014.

ALVAREZ, C. S.; SIERRA-SOSA, D.; GARCIA-ZAPIRAIN, B.; YODER-HIMES, D.; ELMAGHRABY, A. Detection of volatile compounds emitted by bacteria in wounds using gas sensors. *Sensors 19.7*, p. 1523, 2019.

ALVAREZ-MELIS, D.; JAAKKOLA, T. S. On the robustness of interpretability methods. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2018. v. 31.

ARAúJO, L. A. L. Ensino de biologia: uma perspectiva evolutiva–a evolução como eixo integrador na educação básica. *Genética na Escola, V.16.2*, p. 488–489, 2021.

ARRARTE, E.; GARMENDIA, C. R. G.; WISNIEWSKI, M.; VERO., S. Volatile organic compounds produced by antarctic strains of candida sake play a role in the control of postharvest pathogens of apples. *Biological Control 109*, p. 14–20, 2017.

ARRARTE, E. et al. Volatile organic compounds produced by antarctic strains of candida sake play a role in the control of postharvest pathogens of apples. *Biological Control*, Elsevier, 2017.

ARRIETA, A. B. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, Elsevier, 2020.

BABAYEV, R.; WIESE., L. Benchmarking classifiers on medical datasets of uea archive. *Proceedings of AI Health WWW*, 2021.

BAGNALL, A.; KIRáLY, F.; LöNING, M.; MIDDLEHURST, M.; OASTLER., G. A tale of two toolkits, report the first: benchmarking time series classification algorithms for correctness and efficiency. *arXiv preprint arXiv:1909.05738*, 2019.

BARANTSEVICH, N.; BARANTSEVICH, E. Diagnosis and treatment of invasive candidiasis. *Antibiotics*, MDPI, v. 11, n. 6, p. 718, 2022.

BENDA, N. D. et al. Detection and characterization of kodamaea ohmeri associated with small hive beetle aethina tumida infesting honey bee hives. *Journal of Apicultural Research*, Taylor & Francis, 2008.

BEYDA, N. D.; ALAM, M. J.; GAREY, K. W. Comparison of the t2dx instrument with t2candida assay and automated blood culture in the detection of candida species using seeded blood samples. *Diagnostic microbiology and infectious disease 77.4*, p. 324–326, 2013.

BUXTON, R. *Blood Agar Plates and Hemolysis Protocols*. 2016. American Society for Microbiology (ASM) Protocol.

CAILLEUX, A. et al. Gas chromatography-mass spectrometry analysis of volatile organic compounds produced by some micromycetes. *Chromatographia*, Springer, 1992.

CASTRO, M. C.; ALMEIDA, L. M.; FERREIRA, R. W. M.; BENEVIDES, C. A.; ZANCHETTIN, C.; MENEZES, F. D.; INáCIO, C. P.; LIMA-NETO, R. G. de; FILHO, J. G. A.; NEVES, R. P. Breakthrough of clinical candida cultures identification using the analysis of volatile organic compounds and artificial intelligence methods. *IEEE Sensors Journal 22.13*, p. 12493–12503, 2022.

CAYA, M. V. C.; MARAMBA, R. G.; MENDOZA, J. S. D.; SUMAN, P. S. Characterization and classification of coffee bean types using support vector machine. *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM). IEEE*, 2020.

CHAKRABARTI, A.; MOHAMED, N.; CAPPARELLA, M. R.; TOWNSEND, A.; SUNG, A. H.; YURA, R.; MUÑOZ, P. The role of diagnostics-driven antifungal stewardship in the management of invasive fungal infections: a systematic literature review. In: OXFORD UNIVERSITY PRESS. *Open Forum Infectious Diseases*. [S.l.], 2022. v. 9, n. 7, p. ofac234.

CHE, Z.; PURUSHOTHAM, S.; CHO, K.; SONTAG, D.; LIU, Y. Interpretable deep models for icu outcome prediction. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *AMIA Annual Symposium Proceedings*. [S.l.], 2017. v. 2016, p. 1721.

CHEN, M. et al. Artificial intelligence-based medical sensors for healthcare system. *Advanced Sensor Research*, v. 3, n. 3, p. 2300009, 2024.

CHEN, Z.; CHEN, Z.; SONG, Z.; YE, W.; FAN, Z. Smart gas sensor arrays powered by artificial intelligence. *Journal of Semiconductors vol. 40, no. 11, , doi: 10.1088/1674-4926/40/11/111601*, 2019.

CLANCY, C. J.; NGUYEN, M. H. Finding the "missing 50%" of invasive candidiasis: how nonculture diagnostics will improve understanding of disease spectrum and transform patient care. *Clinical infectious diseases 56.9*, p. 1284–1292, 2013.

COLOMBO, A. L.; GUIMARãES, T. Epidemiologia das infecções hematogênicas por *Candida* spp. *Revista da sociedade brasileira de medicina tropical 36, doi: 10.1590/s0037-86822003000500010*, p. 599–607, 2003.

CORONA, G.; CASCIO, A.; CUSUMANO, E.; PANTALEO, D.; CORDARO, V.; BARBERI, I. A retrospective analyses of candida spp. infections in the intensive care unit. *Pediatric Research 70.5*, p. 433–433, 2011.

DADAR, M.; TIWARI, R.; KARTHIK, K.; CHAKRABORTY, S.; SHAHALI, Y.; DHAMA, K. Candida albicans-biology, molecular characterization, pathogenicity, and advances in diagnosis and control–an update. *Microbial pathogenesis 117, doi: 10.1016/j.micpath.2018.02.028*, p. 128–138, 2018.

DEMPSTER, A. et al. Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, Springer, 2020.

DEMPSTER, A.; PETITJEAN, F.; WEBB, G. I. Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery, v. 34.5*, p. 1454–1495, 2020.

DENG, H. et al. A time series forest for classification and feature extraction. *Information Sciences*, Elsevier, 2013.

ESSER-SKALA, W.; FORTELNY, N. Reliable interpretability of biology-inspired deep neural networks. *NPJ Systems Biology and Applications*, v. 9, n. 1, p. 50, 2023.

FAIYAZUDDIN, M.; RAHMAN, S. J. Q.; ANAND, G.; SIDDIQUI, R. K.; MEHTA, R.; KHATIB, M. N.; GAIDHANE, S.; ZAHIRUDDIN, Q. S.; HUSSAIN, A.; SAH, R. The impact of artificial intelligence on healthcare: A comprehensive review of advancements in diagnostics, treatment, and operational efficiency. *Health Science Reports*, v. 8, n. 1, p. e70312, 2025.

FALCONER, K.; HAMMOND, R.; GILLESPIE, S. H. Improving the recovery and detection of bloodstream pathogens from blood culture. *Journal of Medical Microbiology*, v. 69, n. 6, p. 806–811, 2020.

FALLAHI, S.; BABAEI, M.; ROSTAMI, A.; MIRAHMADI, H.; ARAB-MAZAR, Z.; SEPAHVAND, A. Diagnosis of candida albicans: conventional diagnostic methods compared to the loop-mediated isothermal amplification (lamp) assay. *Archives of microbiology 202.2*, p. 275–282, 2020.

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.

FAWAZ, H. I.; LUCAS, B.; FORESTIER, G.; PELLETIER, C.; SCHMIDT, D. F.; WEBER, J.; WEBB, G. I.; IDOUMGHAR, L.; MULLER, P.-A.; PETITJEAN, F. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery 34.6*, p. 1936–1962, 2020.

FILHO, A. d. D. Capítulo 2: coccidioidomicose. *Jornal Brasileiro de Pneumologia 35*, p. 920–930, 2009.

FLYNN, M.; LARGE, J.; BAGNALL, T. The contract random interval spectral ensemble (c-rise): the effect of contracting a classifier on accuracy. *Hybrid Artificial Intelligent Systems: 14th International Conference, HAIS 2019, León, Spain*, p. 381–392, 2019.

FLYNN, M. et al. The contract random interval spectral ensemble (c-rise): The effect of contracting a classifier on accuracy. *Hybrid Artificial Intelligent Systems: 14th International Conference, HAIS 2019*, Springer International Publishing, 2019.

FURIZAL, F.; MA'ARIF, A.; RIFALDI, D. Application of machine learning in healthcare and medicine: A review. *J Robot Control (JRC)*, v. 4, n. 5, p. 621–631, 2023.

GANCARZ, M.; WAWRZEńCZYK, A.; JANCZAREK, I.; ZABORSKI, D.; GRZYWACZ, T.; TATARA, M. R. Electronic nose and its applications: A survey. *International Journal of Automation and Computing*, v. 16, n. 6, p. 695–715, 2019.

GANGULY, R.; SINGH, D. Explainable Artificial Intelligence (XAI) for the Prediction of Diabetes Management: An Ensemble Approach. *International Journal of Advanced Computer Science and Applications*, v. 14, n. 7, 2023.

GILL, A. Y.; SAEED, A.; RASOOL, S.; HUSNAIN, A.; HUSSAIN, H. K. Revolutionizing healthcare: How machine learning is transforming patient diagnoses - a comprehensive review of ai's impact on medical diagnosis. *Journal of World Science*, v. 2, n. 10, p. 1638–1652, 2023.

GUIDOTTI, R.; MONREALE, A.; RUGGIERI, S.; TURINI, F.; GIANNOTTI, F.; PEDRESCHI, D. A survey of methods for explaining black box models. *ACM Computing Surveys*, v. 51, n. 5, p. 1–42, 9 2019.

GéRON, A. *Mãos à obra: aprendizado de máquina com Scikit-Learn." Keras & TensorFlow: Conceitos, ferramentas e técnicas para a construção de sistemas inteligentes.* [S.l.]: O'Reilly Media, Inc., [S. l.: sn], 2021. v. 2.

HAMEED, I.; SOULY, A.; BOGOSAVAC, S.; BUÇINCA, Z.; GILPIN, L. H. BASED-XAI: Breaking ablation studies down for explainable artificial intelligence. *arXiv preprint arXiv:2207.05566*, 2022.

HAYASAKA, T.; LIN, A.; COPA, V. C.; JR, L. P. L.; LOBERTERNOS, R. A.; BALLESTEROS, L. I. M.; KUBOTA, Y.; LIU, Y.; SALVADOR, A. A.; LIN., L. An electronic nose using a single graphene fet and machine learning for water, methanol, and ethanol. *Microsystems & nanoengineering 6.1*, p. 1–13, 2020.

HERTEL, M. et al. Identification of signature volatiles to discriminate candida albicans, glabrata, krusei and tropicalis using gas chromatography and mass spectrometry. *Mycoses*, Wiley, 2016.

HERTEL, M. et al. Volatile organic compounds in the breath of oral candidiasis patients: a pilot study. *Clinical Oral Investigations*, Springer, 2018.

HERTEL, M.; SCHUETTE, E.; KASTNER, I.; HARTWIG, S.; SCHMIDT-WESTHAUSEN, A. M.; PREISSNER, R.; PARIS, S.; PREISSNER., S. Volatile organic compounds in the breath of oral candidiasis patients: a pilot study. *Clinical Oral Investigations 22*, p. 721–731, 2018.

HOLZINGER, A.; BIEMANN, C.; PATTICHIS, C. S.; KELL, D. B. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

HOLZINGER, A. et al. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley, 2019.

HONG, G.; MILLER, H. B.; ALLGOOD, S.; LEE, R.; LECHTZIN, N.; ZHANG., S. X. Use of selective fungal culture media increases rates of detection of fungi in the respiratory tract of cystic fibrosis patients. *Journal of clinical microbiology, V.55.4*, p. 1122–1130, 2017.

HOOKER, S.; ERHAN, D.; KOH, P. W.; SOKLAKOV, K. A benchmark for interpretability methods in deep neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2019. v. 32.

HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process, v. 5, n. 2*, p. 1, 2015.

HOU, J. et al. Self-explainable AI for medical image analysis: A survey and new outlooks. *arXiv preprint arXiv:2410.02331*, 2024.

HUANG, Y.; LI, X.; WANG, J.; ZHANG, Y. Deep Ensemble Learning for Human Activity Recognition Using Wearable Sensors. *IEEE Sensors Journal*, v. 22, n. 15, p. 15234–15245, 2022.

IMAM, A. et al. On consistency and limitation of paired t-test, sign and wilcoxon sign rank test. *IOSR Journal of Mathematics*, IOSR, 2014.

JENKINS, A. T. A. et al. Detection of volatile compounds emitted by bacteria in wounds using gas sensors. *Sensors*, MDPI, 2019.

KOZEL, T. R.; WICKES., B. Fungal diagnostics. *Cold Spring Harbor perspectives in medicine, V.4.4*, p. a019299, 2014.

KRISHNA, S.; HAN, S.; RAWAL, A. S.; SINGH, A. P.; LAKKARAJU, H. The disagreement problem in explainable machine learning: A practitioner's perspective. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. [S.l.: s.n.], 2022. p. 2197–2213.

KUSBANDHINI, O. J.; WIJAYA, D. R.; HIDAYAT., W. Rice shelf-life prediction using support vector regression algorithm based on electronic nose dataset. *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS). IEEE*, 2021.

LEE, Y.-H. et al. Nearest-neighbor-based approach to time-series classification. *Decision Support Systems*, Elsevier, 2012.

LEMFACK, J. et al. mvoc: a database of microbial volatiles. *Charité Universitätsmedizin Berlin, Bioinformatics*, Charité Universitätsmedizin Berlin, 2024.

LI, F. et al. A risk prediction model for invasive fungal disease in critically ill patients in the intensive care unit. *Asian Nursing Research*, Elsevier, v. 12, n. 4, p. 299–303, 2018.

LI, W.; LIU, H.; XIE, D.; HE, Z.; PI., X. Lung cancer screening based on type-different sensor arrays. *Scientific reports 7.1*, p. 1–12, 2017.

LIMA, A. R.; PINTO, J.; AZEVEDO, A. I.; BARROS-SILVA, D.; JERóNIMO, C.; HENRIQUE, R.; BASTOS, M. de L.; PINHO, P. G. de; CARVALHO., M. Identification of a biomarker panel for improvement of prostate cancer diagnosis by volatile metabolic profiling of urine. *British Journal of Cancer 121.10*, p. 857–868, 2019.

LIN, Q.; WU, H.; YUAN, J.; GU, J. The nearest neighbor classifiers for time series with complex shape features. In: *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*. [S.l.: s.n.], 2019. p. 222–227.

LINES, J.; TAYLOR, S.; BAGNALL., A. Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification. *IEEE 16th international conference on data mining (ICDM). IEEE, 10.1109/ICDM.2016.0133.*, p. 1041–1046, 2016.

LIPTON, Z. C. The mythos of model interpretability: In praise of subjective assessments. *Communications of the ACM*, ACM, 2018.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.

MA, S. *Towards Human-centered Design of Explainable Artificial Intelligence (XAI): A Survey of Empirical Studies*. 2024.

MARCOS, J. Y.; PINCUS, D. H. Fungal diagnostics: review of commercially available methods. *Fungal Diagnostics*, p. 25–54, 2013.

MATYSIK, S.; HERBARTH, O.; MUELLER., A. Determination of microbial volatile organic compounds (mvocs) by passive sampling onto charcoal sorbents. *Chemosphere 76.1*, p. 114–119, 2009.

MCINNES, L.; HEALY, J.; MELVILLE, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXi, v:1802.03426*, 2018.

MCKIGHT, P. E.; NAJAB., J. Kruskal-wallis test. the corsini encyclopedia of psychology. *onlinelibrary, https://doi.org/10.1002/9780470479216.corpsy0491*, p. 1–1, 2010.

MERELLES, L. R. d. O.; SILVA, C. de O.; LUZ, M. P. da; MENEZES, J. E. de; DIAS, V. de S. Previsão de geração de resíduos sólidos para o aterro de aparecida de goiânia (go) por séries temporais. *Engenharia Sanitaria e Ambiental 24*, p. 537–546, 2019.

MIDDLEHURST, M.; LARGE, J.; FLYNN, M.; LINES, J.; BOSTROM, A.; BAGNALL., A. Hive-cote 2.0: a new meta ensemble for time series classification. *Machine Learning 110.11-12*, p. 3211–3243, 2021.

MIOTTO, N. M. L.; YURGEL, L. S.; CHERUBINI, K.; CAZANOVA., R. F. Métodos laboratoriais de identificação do fungo candida sp. *Revista da Faculdade de Odontologia-UPF 9.1*, 2004.

MOHAMMED, R. et al. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. *2020 11th International Conference on Information and Communication Systems (ICICS)*, IEEE, 2020.

MOLINARO, E. M.; CAPUTO, L. F. G.; AMENDOEIRA., M. R. R. *Conceitos e métodos para a formação de profissionais em laboratórios de saúde*. [S.l.]: Arca fiocruz, 2012. v. 5.

MORATH, S. U.; HUNG, R.; BENNETT., J. W. Fungal volatile organic compounds: a review with emphasis on their biotechnological potential. *Fungal biology reviews 26.2-3*, p. 73–83, 2012.

MORTAZ, E. Imbalance accuracy metric for model selection in multi-class imbalance classification problems. *Knowledge-Based Systems, v. 210*, p. 106490, 2020.

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*, p. 59–70, 2021.

NASCIMENTO, J. W. A. do. Identificação de bactérias comuns em feridas infectadas (staphylococcus aureus, pseudomonas aeruginosa, enterococcus faecalis e escherichia coli) através de um nariz eletrônico e modelos de inteligência artificial. *MS thesis. Universidade Federal de Pernambuco*, 2022.

NAVARRO-ARIAS, M. J.; HERNáNDEZ-CHáVEZ, M. J.; GARCIA-CARNERO, L. C.; AMEZCUA-HERNáNDEZ, D. G.; LOZOYA-PéREZ, N. E.; ESTRADA-MATA, E.; MARTíNEZ-DUNCKER, I.; FRANCO, B.; MORA-MONTES, H. M. Differential recognition of candida tropicalis, candida guilliermondii, candida krusei, and candida auris by human innate immune cells. *Infection and drug resistance vol. 12, doi: 10.2147/IDR.S197531*, p. 783–794, 2019.

NIELSEN, J. Heuristic evaluation. *Usability Inspection Methods*, John Wiley & Sons, 1994.

PAPPAS, P. G.; KAUFFMAN, C. A.; ANDES, D. R.; CLANCY, C. J.; MARR, K. A.; OSTROSKY-ZEICHNER, L.; REBOLI, A. C. Clinical practice guideline for the management of candidiasis: 2016 update by the infectious diseases society of america. *Clinical Infectious Diseases 62.4*, p. e1–e50, 2016.

PAPPAS, P. G.; LIONAKIS, M. S.; ARENDRUP, M. C.; OSTROSKY-ZEICHNER, L.; KULLBERG, B. J. Invasive candidiasis. *Nature Reviews Disease Primers*, Nature Publishing Group, v. 4, n. 1, p. 1–20, 2018.

PAPPAS, P. G.; LIONAKIS, M. S.; ARENDRUP, M. C.; OSTROSKY-ZEICHNER, L.; KULLBERG., B. J. Invasive candidiasis. *Nature Reviews Disease Primers 4.1*, p. 1–20, 2018.

PATEL, R. A moldy application of maldi : Maldi-tof mass spectrometry for fungal identification. *Journal of Fungi*, MDPI, 2019.

PENG, P.; ZHAO, X.; PAN, X.; YE., W. Gas classification using deep convolutional neural networks. *Sensors 18.1*, p. 157, 2018.

PERL, T.; JüNGER, M.; VAUTZ, W.; NOLTE, J.; KUHNS, M.; ZEPELIN, M. B.; QUINTEL., M. Detection of characteristic metabolites of aspergillus fumigatus and candida species using ion mobility spectrometry–metabolic profiling by volatile organic compounds. *Mycoses 54.6*, p. e828–e837, 2011.

PINI, P.; COLOMBARI, B.; MARCHI, E.; CASTAGNOLI, A.; VENTURELLI, C.; SARTI, M.; BLASI, E. Performance of candida albicans germ tube antibodies (cagta) and its association with (1->3)-β-d-glucan (bdg) for diagnosis of invasive candidiasis (ic). *Diagnostic Microbiology and Infectious Disease 93.1, doi: 10.1016/j.diagmicrobio.2018.07.007*, p. 39–43, 2019.

POHLERT, T. The pairwise multiple comparison of mean ranks package (pmcmr). *R package, v. 27.2019*, p. 9, 2014.

PROUDFOOT, J. et al. Tests for paired count outcomes. *General Psychiatry*, BMJ, 2018.

REZK, N. G.; EL-GHAFAR, H. M. A.; HASSAN, B. M. Xai-augmented voting ensemble models for heart disease prediction: A shap and lime-based approach. *Bioengineering*, MDPI, v. 11, n. 10, p. 1016, 2024.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, p. 1135–1144, 2016.

ROJAT, T. et al. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950*, arXiv, 2021.

ROTH, M. G.; WESTRICK, N. M.; BALDWIN, T. T. Fungal biotechnology: From yesterday to tomorrow. *Frontiers in Fungal Biology*, Frontiers, v. 4, p. 1135263, 2023.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, v. 1, n. 5, p. 206–215, 5 2019.

SAIDI, T.; MOUFID, M.; BELEñO-SAENZ, K. de J.; WELEAREGAY, T. G.; BARI, N. E.; JAIMES-MOGOLLON, A. L.; IONESCU, R. Non-invasive prediction of lung cancer histological types through exhaled breath analysis by uv-irradiated electronic nose and gc/qtof/ms. *Sensors and Actuators B: Chemical vol. 311, doi: 10.1016/j.snb.2020.127932*, 2020.

SAMARANAYAKE, Y. H.; WU, P. C.; SAMARANAYAKE, L. P.; SO, M.; YUEN., K. Y. Adhesion and colonisation of candida krusei on host surfaces. *Journal of medical microbiology 41.4, doi: 10.1099/00222615-41-4-250*, p. 250–258, 1994.

SANTOS, P. S. M. et al. Metabolic profiling of candida auris, a newly-emerging multi-drug resistant candida species, by gc-ms. *Molecules*, MDPI, 2019.

SARMA, S.; UPADHYAY, S. Current perspective on emergence, diagnosis and drug resistance in candida auris. *Infection and drug resistance (2017) vol. 10, doi: 10.2147/IDR.S116229*, p. 155–165, 2017.

SCHLEGEL, U. et al. Towards a rigorous evaluation of xai methods on time series. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, 2019.

SCHLEGEL, U. et al. Towards a rigorous evaluation of xai methods on time series. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, 2019.

SCHäFER, P. The boss is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery, v. 29, n. 6*, p. 1505–1530, 2015.

SELVARAJU, R. R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, IEEE, p. 618–626, 2017.

SEMREEN, M. H.; SOLIMAN, S. S.; SAEED, B. Q.; ALQARIHI, A.; UPPULURI, P.; IBRAHIM., A. S. Metabolic profiling of candida auris, a newly-emerging multi-drug resistant candida species, by gc-ms. *Molecules 24.3*, p. 399, 2019.

SENIN, P.; MALINCHIK, S. SAX-VSM: Interpretable time series classification using SAX and vector space model. *2013 IEEE 13th International Conference on Data Mining*, IEEE, p. 1175–1180, 2013.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning important features through propagating activation differences. *Proceedings of the International Conference on Machine Learning*, PMLR, 2017.

SHTAYAT, M. M. et al. An explainable ensemble deep learning approach for intrusion detection in industrial internet of things. *IEEE Access*, v. 11, p. 115047–115061, 2023.

SILVA, G. A. et al. Candida species (volatile) metabotyping through advanced comprehensive two-dimensional gas chromatography. *Microorganisms*, MDPI, 2020.

SLACK, D.; HILGARD, S.; JIA, E.; SINGH, S.; LAKKARAJU, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. [S.l.: s.n.], 2020. p. 180–186.

SM, C.; Z, Z.; XR, Q.; WT, H.; Y, Z.; W, F.; YL, C.; Y, D. W.; YY, J.; H, A. M. S. The critical role of dectin-1 in host controlling systemic candida krusei infection. *American journal of translational research 11.2*, p. 721–732, 2019.

TAVENARD, R.; FAOUZI, J.; VANDEWIELE, G.; DIVO, F.; ANDROZ, G.; HOLTZ, C.; PAYNE, M. Tslearn, a machine learning toolkit for time series data. *The Journal of Machine Learning Research 21.1*, p. 4686–4691, 2020.

TERRERO-SALCEDO, D.; MARGARET, V. P.-F. Updates in laboratory diagnostics for invasive fungal infections. *Journal of Clinical Microbiology, V.58.6*, p. e01487–19, 2020.

THEISSLER, A. et al. Explainable ai for time series classification: A review, taxonomy and research directions. *IEEE Access*, IEEE, 2022.

TJOA, E.; GUAN, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, IEEE, 2020.

TROVãO, J.; PEREIRA., L. *IntroduÇÃo ao estudo dos microfungos: Guia simples para a iniciação à identificação*. [S.l.]: Departamento de Ciências da Vida - Universidade de Coimbra, 2019. v. 1.

TURING, A. M. Computing machinery and intelligence. mind lix (236). *460. bona fide field of study. he has cochaired the aaai fall 2005 symposium on machine." IEEE Intelligent Systems 2*, 1950.

TURPPA, E.; POLAKA, I.; VASILJEVS, E.; KORTELAINEN, J. M.; SHANI, G.; LEJA, M.; HAICK, H. Repeatability study on a classifier for gastric cancer detection from breath sensor data. *IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE*, 2019.

TóTH, R.; NOSEK, J.; MORA-MONTES, H. M.; GABALDON, T.; BLISS, J. M.; NOSANCHUK, J. D.; TURNER, S. A.; BUTLER, G.; VáGVöLGYI, C.; GáCSER., A. Candida parapsilosis: from genes to the bedside. *Clinical microbiology reviews 32.2, doi: 10.1128/CMR.00111-18.*, p. 1–38, 2019.

VASCONCELOS, P. J. d. M. dentificação de fungos anemófilos, em ambientes abertos, através de um nariz eletrônico e modelos de inteligência artificial. *MS thesis. Universidade Federal de Pernambuco*, 2022.

VERGIDIS, P.; CLANCY, C. J.; SHIELDS, R. K.; PARK, S. Y.; WILDFEUER, B. N.; SIMMONS, R. L.; NGUYEN, M. H. Intra-abdominal candidiasis: the importance of early source control and antifungal treatment. *PLoS One*, Public Library of Science San Francisco, CA USA, v. 11, n. 4, p. e0153247, 2016.

WANG, J. et al. Multilevel wavelet decomposition network for interpretable time series analysis. *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, ACM, 2018.

WANG, Y.; DIAO, J.; WANG, Z.; ZHAN, X.; ZHANG, B.; LI, N.; LI., G. An optimized deep convolutional neural network for dendrobium classification based on electronic nose. *Sensors and Actuators A: Physical 3097, 10.1016/j.sna.2020.111874.*, p. 111874, 2020.

WANG, Y. et al. An optimized deep convolutional neural network for dendrobium classification based on electronic nose. *Sensors and Actuators A: Physical*, Elsevier, 2020.

WEI, W. W. 'time series analysis', in todd d. little. *(ed.), The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2, Statistical Analysis, Oxford Library of Psychology,*, 2013.

WHALEY, S. G.; BERKOW, E. L.; RYBAK, J. M.; NISHIMOTO, A. T.; BARKER, K. S.; ROGERS., P. D. Azole antifungal resistance in candida albicans and emerging non-albicans candida species. *Frontiers in microbiology 7,doi: 10.3389/fmicb.2016.02173*, p. 1–12, 2017.

WONGCHOOSUK, C.; LUTZ, M.; KERDCHAROEN., T. Detection and classification of human body odor using an electronic nose. *Sensors 9.9*, p. 7234–7249, 2009.

WONGCHOOSUK, C.; LUTZ, M.; KERDCHAROEN., T. Detection and classification of human body odor using an electronic nose. *Sensors 9.9*, p. 7234–7249, 2009.

WU, S.; WANG, Y.; ZHANG, Y.; WANG, Z.; LI, Y.; ZHANG, H. Application of artificial intelligence in clinical diagnosis and treatment: an overview of systematic reviews. *Intelligent Medicine*, Elsevier, v. 2, n. 2, p. 88–96, 2022.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, Springer International Publishing, p. 818–833, 2014.

ZHAN, X.; WANG, Z.; YANG, M.; LUO, Z.; WANG, Y.; LI., G. An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction. *Measurement vol. 158, doi: 10.1016/j.measurement.2020.107588*, p. 107588, 2020.

ZHANG, J. D. et al. Interpretable machine learning on metabolomics data reveals biomarkers for Parkinson's disease. *ACS Central Science*, v. 9, n. 5, p. 1035–1045, 2023.

ZOU, L.; SHU, C.; SHI, E.; HE, J.; WANG, J.; WANG, Y. Ensemble image explainable AI (XAI) algorithm for severe community-acquired pneumonia and COVID-19 respiratory infections. *IEEE Transactions on Artificial Intelligence*, IEEE, v. 4, n. 2, p. 242–254, 2022.

Appendices

A - SOFTWARE ARTIFACTS AND COMPUTATIONAL ENVIRONMENT

This appendix details the computational resources, software artifacts, and libraries used for the implementation and validation of the DiagNose.AI Framework. The objective is to ensure the transparency and reproducibility of the experiments conducted in this thesis.

A.1 SOURCE CODE REPOSITORY

All the source code developed for data processing, training of the classification models, and implementation of the *XAI Ensemble* explainability architecture has been documented and is publicly available. Access can be obtained through the following GitHub repositories:

- ***Candida* Identification from culture samples:**

  <https://github.com/michaellopes16/CandidaIdentification.git>

- **Ensemble Xai API:**

  <https://github.com/michaellopes16/EnsembleXaiAPI.git>

- **Explainability Ensemble API:**

  <https://github.com/michaellopes16/ExplainabilityEnsembleAPI.git>

- ***Candida* Identification from blood samples:**

  <https://github.com/michaellopes16/BloodCandidaIdentification.git>

The repository includes the notebooks (*Jupyter Notebooks*) with the analysis scripts and the data files necessary to replicate the results presented in Chapter 5.

A.2 COMPUTATIONAL ENVIRONMENTS

The experiments were executed in two distinct computational environments to ensure flexibility and the validation of the processes in different hardware configurations.

A.2.1 Local Configuration

Most of the initial development and testing was carried out on a personal workstation with the following specifications:

- **Processor:** Core i7
- **RAM:** 24 GB
- **GPU:** NVIDIA GTX 1060

A.2.2 Cloud Environment (Google Colaboratory)

For experiments that demanded greater computational power and to ensure reproducibility in a standardized environment, the Google Colaboratory (Colab) platform was used. The configuration (free version) made available by the platform during the experimental period was:

- **RAM:** Approximately 12.7 GB
- **Disk Storage:** Approximately 78.2 GB
- **GPU:** Google Compute Engine graphic accelerator (e.g., NVIDIA Tesla K80, T4, etc.)

A.3 MAIN TECHNOLOGIES AND LIBRARIES

The implementation of the methodology was carried out using the Python 3 programming language. The main software libraries that supported the development include, but are not limited to:

- **Data Analysis and Manipulation:** Pandas, NumPy
- **Machine Learning and Time Series:** Scikit-learn, Sktime, TensorFlow, Keras
- **Explainable Artificial Intelligence (XAI):** SHAP, LIME, GRAD-CAM
- **Data Visualization:** Matplotlib, Seaborn

The use of these open-source tools was fundamental for the agile development and robust validation of the proposed framework's components.

B - THE DIAGNOSE.AI SYSTEM PROTOTYPE

This appendix details the interface and workflow of the *DiagNose.AI* system prototype, developed as a proof of concept to implement and validate the explainability methodology proposed in this thesis.

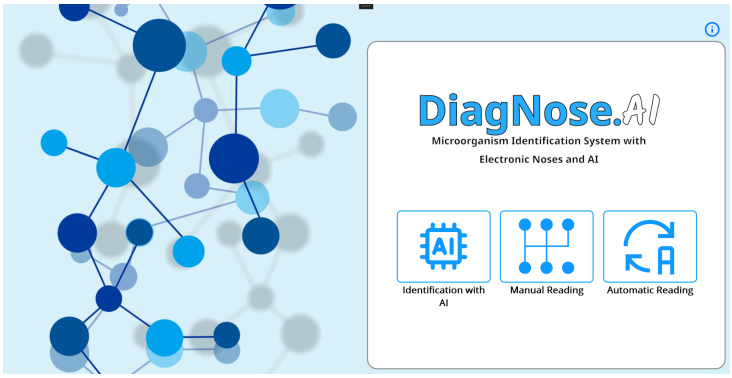
Based on the results and feedback obtained from the usability evaluation, the first version of the *DiagNose.AI* system was developed. In addition to offering the functionality of database creation, the system now also allows for the analysis and prediction of VOCs from samples, featuring the integration of the *XAI Ensemble* structure and an enhanced explainability module. The interface has been refined, adopting a more minimalist style with better-positioned elements. The final prediction screen now presents textual and visual explanations about the prediction process more clearly.

Another important update was the implementation of informational buttons that trigger help pop-ups distributed throughout the application, aiming to guide the user in each step of the process, providing detailed information about the functionality and purpose of each module. Representative images of the final system are presented in Figure 32.

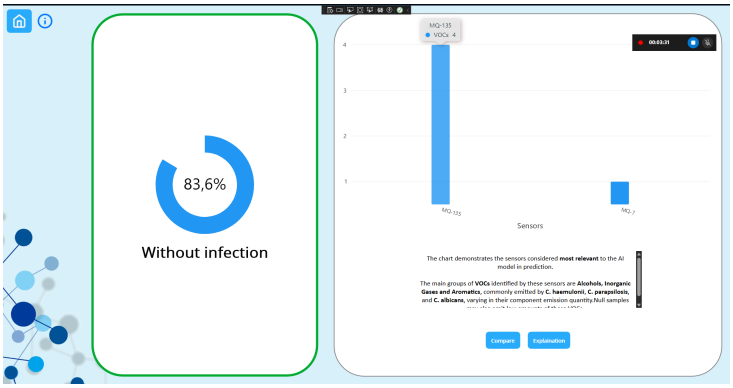
On the application's home screen, the user is presented with three main options, in addition to access to further information. Upon selecting the AI-based sample identification option, the user is directed to a settings screen and then to the sample reading screen (Figure 32(a)). In this step, a complete reading cycle is performed, which includes purging and VOC collection. Detailed descriptions of each step are accessible through informational elements in the interface. After the cycle is completed, the VOC data undergoes preprocessing and is sent for analysis by the *XAI Ensemble* library and the AI model.

In a few seconds, the system returns several results (Figure 32(b)): the sample prediction, the main features (sensors) that influenced the model's decision, an interpretable textual explanation (see an example in the quote below), the mapping of the identified VOCs (Figure 32(c)), as well as data from three similar samples (Figure 32(d)), extracted from the database, which were used as a reference. Thus, the user obtains a comprehensive and transparent view of the model's decision-making process.

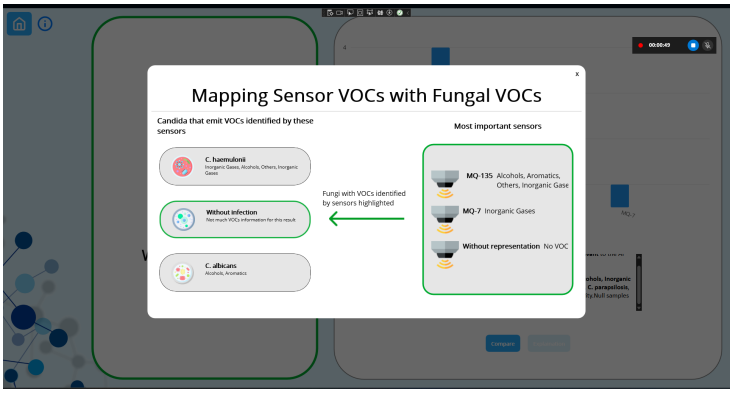
> "The chart demonstrates the sensors considered **most relevant** to the AI model in prediction.
>
> The main groups of **VOCs** identified by these sensors are **Alcohols, Inorganic**



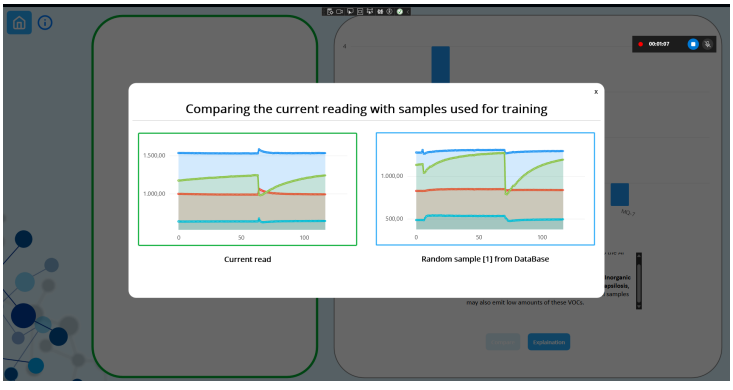
(a) Main Screen



(b) Result and Explainability Screen



(c) VOCs Mapping Screen



(d) Comparison of Current Sample and Database Sample

Figure 32 – Exploration of the main screens of the DiagNose.AI system, illustrating the interaction flow from the initial screen to the presentation of explainable results.

> **Gases and Aromatics**, commonly emitted **by C. haemulonii, C. parapsilosis, and C. albicans**, varying in their component emission quantity. Null samples may also emit low amounts of these VOCs."

Additionally, the system effectively addresses the main bottlenecks identified in the literature regarding currently available explainability methods, such as the absence of textual explanations, the limitation of intuitive graphical visualizations, and the difficulty of interpretation by non-technical users (TJOA; GUAN, 2020; HOLZINGER et al., 2017; CHE et al., 2017). By adopting an approach centered on the interpretation of VOC classification, based on the opinion and needs of end-users, the system promotes a target-audience-oriented explainability.

Annexes

ANNEX A – PUBLICATION IN THE JOURNAL SCIENTIFIC REPORTS

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Breaking barriers in *Candida* spp. detection with Electronic Noses and artificial intelligence

Michael L. Bastos[1]✉, Clayton A. Benevides[2,6], Cleber Zanchettin[1,6], Frederico D. Menezes[3,6], Cícero P. Inácio[4,6], Reginaldo G. de Lima Neto[4,6], José Gilson A. T. Filho[5,6], Rejane P. Neves[4,6] & Leandro M. Almeida[1]✉

The timely and accurate diagnosis of candidemia, a severe bloodstream infection caused by *Candida* spp., remains challenging in clinical practice. Blood culture, the current gold standard technique, suffers from lengthy turnaround times and limited sensitivity. To address these limitations, we propose a novel approach utilizing an Electronic Nose (E-nose) combined with Time Series-based classification techniques to analyze and identify *Candida* spp. rapidly, using culture species of C. *albicans*, C.*kodamaea ohmeri*, C. *glabrara*, C. *haemulonii*, C. *parapsilosis* and C. *krusei* as control samples. This innovative method not only enhances diagnostic accuracy and reduces decision time for healthcare professionals in selecting appropriate treatments but also offers the potential for expanded usage and cost reduction due to the E-nose's low production costs. Our proof-of-concept experimental results, carried out with culture samples, demonstrate promising outcomes, with the Inception Time classifier achieving an impressive average accuracy of 97.46% during the test phase. This paper presents a groundbreaking advancement in the field, empowering medical practitioners with an efficient and reliable tool for early and precise identification of candidemia, ultimately leading to improved patient outcomes.

Infections caused by fungi are a significant issue in the scenario of Intensive Care Units (ICUs), increasing morbidity and the number of deaths in patients who are in a critical state of health[1,2]. The main reason for the occurrence of this type of infection, also described as invasive fungal infections (IFI), is candidiasis, with *Candida albicans* as the primary causative agent, followed by *Candida parapsilosis*, *Candida glabrata*, *Candida krusei* and *Candida tropicalis*[3]. According to reports by[4], approximately 15 species of *Candida* can cause human diseases, and the most common, presented in more than 90% of cases. Furthermore, there have been notable changes in this field, with the emergence of species considered rare or uncommon, such as occurrences with *C. pelliculosa*, *C. haemulonii*, *C. guilliermondii*, *C. lusitaniae*, *C. famata* and *C. auris*[4,5].

Data reported by[5] show that, despite considerable advances in antifungal therapy in recent years, mortality related to Invasive fungal infections (IFIs) in ICUs has been 40 to 60%. One of the factors contributing to this mortality rate is the challenge in recognizing and diagnosing IFIs in the early stages of treatment[5,6]. According to[6], only half of the tested patients were reported to be infected by *Candida* spp. Considering that the result may take 2 to 7 days to be confirmed (in the case of culture-based methods), and given the severity of this condition, a delay of more than 12 hours can increase the risk of mortality.

At present, blood culture is the standard method in the laboratory diagnosis of candidemia, enabling the isolation of the causative agent for identification[7]. Alternative techniques that do not rely on cultures are also used, including polymerase chain reaction (PCR), detection of mannan and beta-D-1,3-glucan antigens (BDG), and enzyme-linked immunosorbent assay (ELISA). It is important to note that some of these approaches involve careful sample preparation, have long response times, entail significant costs, and require professionals with specific expertise[8,9].

[1]Centro de Informática, Universidade Federal de Pernambuco, Recife, PE, Brazil. [2]Comissão Nacional de Energia Nuclear, Centro Regional de Ciências Nucleares do Nordeste, Recife, PE, Brazil. [3]Departamento de Mecânica, Instituto Federal de Pernambuco, Recife, PE, Brazil. [4]Centro de Ciências Médicas, Universidade Federal de Pernambuco, Recife, PE, Brazil. [5]Centro de Ciências Sociais e Aplicadas, Universidade Federal de Pernambuco, Recife, PE, Brazil. [6]These authors contributed equally: Clayton A. Benevides, Cleber Zanchettin, Frederico D. Menezes, Cícero P. Inácio, Reginaldo G. de Lima Neto, José Gilson A. T. Filho and Rejane P. Neves. ✉email: mlb@cin.ufpe.br; lma3@cin.ufpe.br

In addition, we can also mention T2Candida, which combines targeted PCR with T2 magnetic resonance and Matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry (MS). T2Candida allows for early detection of candidemia in patients undergoing antifungal therapy; however, it is not suitable for low-prevalence environments, is costly, and covers only five of the main species[10]. As for MALDI-TOF MS, it is highly successful in identifying clinical samples, but it can be a time-consuming process and heavily relies on the expertise of clinical mycologists handling the samples[8,12]. Furthermore, according to[13], combining the MALDI-TOF MS technique with other methods is often advisable to achieve more accurate and satisfactory results. However, this approach also involves the use of equipment that can be costly and may not always be readily available in various microbiology laboratories, particularly in developing countries[14]. However, alternative methods are based on detecting Volatile Organic Compounds (VOCs) to identify these fungal agents. These methods include Gas Chromatography-Mass Spectrometry (GC-MS), Solid Phase Microextraction (SPME), Simultaneous distillation extraction (SDE), and Selected Ion Flow Tube Mass Spectrometry (SIFT-MS)[15].

Another method that has received some attention and shows potential for development is the Electronic Nose, often called the "E-Nose." This technology combines a variety of gas sensors and uses artificial intelligence to identify patterns of Volatile Organic Compounds (VOCs) and categorize the unique "smell fingerprints" associated with these compounds. This tool is generally built with metal oxide conducting chemical sensors (MOS), which are responsible for identifying the volatile organic compounds released by the odor-emitting components. Its functioning is based on the olfactory function of mammals and has been studied since the 1980s[16]. Like a real nose, the E-nose aims to identify patterns from the VOCs identified by the sensors, whose reading values are analyzed and classified by an artificial intelligence (AI) model. This device typically comprises three main parts: sensors, a signal processing unit, and a pattern recognition system[17].

The Electronic Nose is already being applied in various domains, from food safety to agricultural applications and disease diagnosis, as[18] mentioned. For a more comprehensive view of these applications, one can delve into studies conducted by[19,20], and[21], which focus on the identification of microorganisms, including fungi and bacteria. Furthermore, research carried out by[22,23], and[24] further extends the exploration of Electronic Nose applications in the food industry. It's also worth highlighting the study by[25], in which a portable Electronic Nose device is employed to diagnose gynecological conditions in a clinical setting rapidly.

In the context of medical diagnosis, Electronic Noses have experienced remarkable advances in recent years, particularly in hardware development and algorithm evolution[18,26]. Medical diagnosis stands out among the fields most benefited by the progress of this technology, as previously mentioned[18]. However, some limitations still require refinement, such as the stability, standardization, and reliability of certain sensors[27,28]. In this regard, efforts are being devoted to enhancing the sensitivity, selectivity, and stability of these devices, with significant progress when these mechanisms are integrated with artificial intelligence and Machine Learning techniques[18,29].

Given the above, it is understood that there is a significant issue regarding the rapid identification of fungi in hospitalized patients and those with a clinical condition that requires extra care[30,5]. Considering that this identification process can be improved, this project proposes using an Electronic Nose to recognize patterns related to fungi of the *Candida* spp. species[31] utilizing control samples collected by ATCC company. This method can be combined with a set of machine learning techniques, enabling quicker and more efficient identification[32], streamlining the decision-making process of health professionals, and, consequently, improving the survival chances of these patients. It is essential to mention that in this initial proof-of-concept study, we are using only culture samples, aiming at the creation and validation of a rapid and efficient protocol that can be replicated in the future for samples of other materials, such as whole blood. To better understand this, the following sections will address the Materials and methods used for the construction of the study, the Results and discussions on its development, and, finally, the Conclusions of the findings of this investigation.

Results

Through the implemented models, a series of experiments were conducted and cataloged using the metrics Accuracy, F1-score, Recall (Sensitivity), Specificity, Precision, and Standard deviation, aiming to identify patterns in the VOCs released by the analyzed *Candida* species. The variety of models covered the different characteristics that the data may have, highlighting the models that best fit the data standard and discarding those with less potential. Initially, all models were applied with the parameters defined by the documentation or in their respective repositories. The possibility of including a parameter validation step for the models was considered. However, given the satisfactory performance of most models and considering the computational cost and time that this step would require, it was deprioritized for the time being.

Regarding the methods used, the primary rationale for using time series models is the temporal nature of the signal reading, with data from each round of the aspiration process being added to the database. The majority of the models used were sourced from the Sktime library. However, due to its uniqueness, Inception Time was the only one implemented independently of the library, as there is currently no tool that simplifies access to its functions and properties. The model code provided by the authors on GitHub had to be modified to accommodate the metrics and dataset of this study.

As a result of the training stage, most of the models achieved 100% accuracy. This is justified due to the reduction of instances that the pre-processing step brought, using the cycles as training elements. Thus, models learn data patterns better as they have less to memorize. In this regard, to prevent overfitting, in addition to adding more data cycles for model training, grid search steps or optimization algorithms can be employed to find better parameters[33]. Another commonly used strategy is the application of more robust models, as was the case with InceptionTime[34], achieving greater consistency at all stages of the process.

In addition to the average value referring to the metrics in the training process, the values referring to the averages of the validation and testing stages of the models were also recorded. There was a moderate decrease

Classifiers	Accuracy	F1-score	Recall (Sensitivity)	Precision	Specificity	Test time (s)
Inception Time	**0,97468**	**0,97605**	**0,97817**	**0,97540**	**0,99513**	1,21489
Random Interval Spectral Ensemble (RISE)	0,65000	0,55758	0,57007	0,56251	0,94223	4,58585
Time Series Forest Classifier	0,67500	0,61261	0,58960	0,61261	0,91312	1,18719
ROCKET Classifier	0,78750	0,78105	0,85764	0,79171	0,93804	4,91326
Shapelet Transform Classifier	0,63750	0,58207	0,60258	0,59832	0,93028	0,61583
K-Neighbors Time Series Classifier	0,75000	0,73245	0,72192	0,81357	0,95635	52,86100
HIVE COTE 1	0,52500	0,40245	0,42669	0,39475	0,94009	11,93961
HIVE COTE 2	0,66250	0,58503	0,60360	0,62833	0,94342	2,57620
BOSS Ensemble	0,63750	0,50525	0,53555	0,49929	0,90165	**0,48432**

Table 1. Result of the model testing stage—test values for the metrics Accuracy, F1-score, Recall (sensitivity), Precision, Specificity, and Test time measured for each model after the training and validation phase. Significant values are in bold.

in the performance of the models between the training phase and the validation and test phases, amidst 7 and 4%. This is because, during training, the models identify *Candida* patterns precisely due to the distinctive nature between each species and the new data division. In the other phases, as they are new data, and the model has never seen them, it is normal and expected that it ends up making more errors, which in no way interferes with its final evaluation. Table 1 demonstrates the data referring to the testing steps of each of the models.

As observed in the result set, the most notable model was Inception Time[34], executed with the standard set of parameters, followed by ROCKET Classifier[35], Time Series Forest Classifier[36] and Random Interval Spectral Ensemble (RISE)[37], respectively. All metrics calculated in Inception Time were near 100%, demonstrating high consistency between the results.

In addition to collecting the metrics, statistical tests were conducted to verify the difference between the results of the different models. Specifically, a normality test was performed with the accuracy results obtained in the 10 repetitions for each model of the validation stage. This was followed by a significance test and a post-hoc test to compare the selected algorithms pairwise.

It can be interpreted that only the Inception Time model does not follow a normal data distribution. It would already suggest using a non-parametric test to evaluate the results. However, to obtain increased sensitivity of the analyses, a numerical test of statistical normality was also applied, where the most suitable test for the problem in question was the Shapiro-Wilk test. According to[38], this method is more suitable for small sample sets smaller than 50, although it can also be used for larger sets. In contrast, methods such as Kolmogorov-Smirnov are ideal for samples larger than or equal to 50. Both tests use as a null hypothesis the statement that the data are all derived from a normal distribution set, accepting this hypothesis when $p>0.05$, confirming the data as normally distributed.

As a result of applying the normality test, the HIVE COTE1, Shaplet Transform Classifier, and TimeSeries Forest Classifier classifiers did not present a normal distribution according to the Shapiro-Wilk test, with p-values equal to 0.01227, 0.03521, and 0.00021, respectively. All these values are less than 0.05.

Indeed, with this result, we confirm the need to apply a non-parametric test, given that only some groups follow a normal distribution. As per[39], the most appropriate non-parametric test for this case is the Kruskal-Wallis test, considering the number of examples in the groups is small and equal. For the execution of the test, the following hypotheses were considered:

- H0: All models have relatively equal means in terms of classification accuracy;
- H1: At least one of the models differs from the others in terms of mean classification accuracy.

Where H0 is the null hypothesis, which assumes that all models have equal performance, H1 is the alternative hypothesis, which is the difference in performance of at least one of the models about the others. For this test, a p-value less than 0.05 indicates the rejection of the null hypothesis, suggesting the existence of a significant difference between the evaluated samples. Thus, applying Kruskal-Wallis to the set of results acquired, a p-value of 2.49E-02 was obtained, which is less than 0.05. This demonstrates that with 95% confidence, there is evidence to reject H0 and accept the hypothesis that at least one of the models differs from the others in mean validation accuracy.

Given this model difference, the next step was applying a post-hoc test to identify which models are statistically different. The non-parametric test only indicates the existence of this difference, not the relationship between the sets. For this step, the Nemenyi test was used, which, according to[40], is one of the most commonly used post-hoc tests after applying Kruskal-Wallis. As briefly explained, this method performs a pairwise investigation of each analyzed set, returning the p-values for each relationship between the evaluated groups. The values vary between -1 and 1, with $p<0.05$ indicating a significant statistical difference between the samples according to the test and values closer to 1 demonstrating similarity. Figure 1 depicts a correlation matrix that crosses the results obtained by the Nemenyi method.

As observed, there is a high similarity between most models with a lower accuracy average, not showing a significant statistical discrepancy between them. However, it can be stated that there is no significant difference

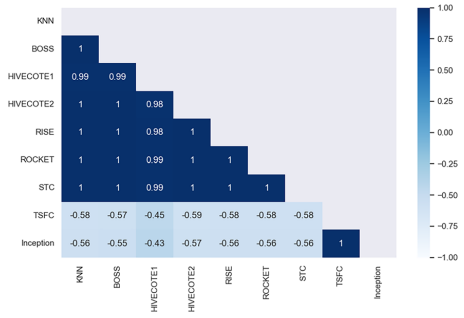


Figure 1. Correlation graph of the results of applying the Nemenyi post-hoc test on the set of results for each model. In this type of graph, when it is farther from 1, the elements are more divergent; that is, they are statistically different.

between the Inception Time[34] and Time Series Forest Classifier[36] models, both of which have p-values much less than 0.05. Each one has a notable difference between the models with more distinct accuracy. However, there is a high similarity between some with closer accuracy (which was expected), dividing the models into different groups of relevance. This way, it is possible to identify the difference between the models, with Inception Time and TSFC not showing a significant difference. Although there is no significant statistical difference, Inception Time stands out with average values for Accuracy, Precision, Recall (sensitivity), Specificity, and F1 above 95% in all analysis sets. Additionally, it boasts an execution time of just over 1 second, making it the most promising choice for the final classification model of volatile compounds emitted by *Candida* species. The entire process of identifying microorganisms, encompassing sample reading and model classification, is completed in approximately 15 minutes.

Discussion

With the results of this study, it is possible to see the effectiveness of using Electronic Noses in the face of such complex problems, including identifying fungi through VOCs emitted by species in culture. In contrast to other solutions, using this technology, in addition to making the process of helping identify fungi cheaper, can speed it up, achieving a satisfactory result within a few hours. Traditional methods use expensive, large machines (challenging to transport), which require a longer time to indicate an accurate result. With the E-nose built with low-cost parts in a compact suitcase, it will be possible to transport it more easily and quickly. The identification speed is up to the AI models being trained because the more accurately they use data with less culture time, the faster their classification returns.

From the first stages of the study, in the visual analysis of the data, it is possible to identify a distinct separation between some species (highlighted in the PCA of the Fig. 4a). This helps to understand which *Candida* species can be better identified by the models and demonstrate a linear separation between some. For example, it's possible to observe in the left part of the projection a cluster of five species (C. *albicans*, C. *glabrata*, C. *haemulonii*, C. *kodamaea ohmeri*, and C. *krusei*), which could be separated by some lines, as well as in the lower right corner, where C. *parapsilosis* and C. *krusei* are located, and in the upper right corner, where C. *albicans* and C. *glabrata* can be found. It's worth noting that some other species within these groups might account for some of the errors recorded by the models during the learning process.

Another critical point is the choice of Time Series for training and data classification. This decision was taken given the temporal characteristic of the data, both for the time of culture of the fungi and for the reading of the volatile emitted by them and captured by the Electronic Nose, based on the process in evidence in Fig. 3c.

All this flow culminated in obtaining outstanding results for the validation and classification phase of the samples, where most of the models achieved an assertiveness above 90%, with emphasis on the Inception Time, with an average of 97.70%, 95.87%, and 97.46% of accuracy in the training, validation, and testing phases, respectively, with very similar values for the other metrics. In the training step, most models reached 100% in all metrics. However, this can be seen as a bias in the data, harming the test step. All this difference was confirmed by the analysis of statistical significance, where through the Shapiro-Wilk normality tests, the Kruskal-Wallis non-parametric test, and the Nemynyi post-hoc test, the difference between the algorithms used was identified.

Although there are still no comparative studies between the E-nose and artificial intelligence in relation to more traditional yeast identification techniques, we can observe a great similarity between the efficiency of the method presented in this work and methods such as MALDI-TOF MS, CHROMagar and Corn meal tween-80 agar, as demonstrated in study[41]. The authors' approach indicates that, even though these techniques are not considered gold standard for yeast identification, they can lead to very promising results for some species, with a performance very similar to that of our study (indicated in Table 1), when compared to the percentage of correct answers. This highlights the importance of using new methods that can fill the gaps left by more traditional methods.

Thus, as seen in the study, it is possible to perceive how powerful Electronic Nose, combined with new Time Series techniques, can yield satisfactory and promising results. Because it is a portable tool with a moderate construction cost compared to current mechanisms - it can reach a wide range of environments in places with fewer resources and difficult to access. These facilitators should broaden the identification process's scope of use, benefiting many people. For the next steps, samples of new species of *Candida*, even rarer, such as *C. tropicalis*, *C. auris*, *C. famata*, *C. pelliculosa*, *C. guilliermondii*, *C. lusitaniae*, and other fungal segments should be added to the dataset, seeking to create more generic and accurate models in the identification of this fungus. Additionally, these new samples will allow for a broader development of specificity tests among fungi, aiming to ensure the absence of false positives in our results. Furthermore, new analyses will be conducted with shorter culture times to determine if further reducing the identification time is possible.

Another critical step for the future will be to expose the Electronic Nose to patient blood and *in situ* samples to identify its efficiency in a scenario closer to its final operation. In this sense, the equipment requires an environment free from high levels of odors to prevent the risk of incorrect readings due to external interference. However, it can be used in a clinic if there is assurance of an environment free from other contaminating odors (e.g., alcohol, perfumes, air fresheners, etc.). This condition may be possible by using a room containing an extractor fan. From there, health professionals will also perform a qualitative assessment to obtain feedback related to the results indicated by the tool.

Methods

This work is an evolution of the project developed by[19], which introduced research on using Electronic Nose and AI to identify *Candida* spp. In the current project, more robust, automated equipment is used that makes it possible to analyze a greater volume of samples. In addition, it also allows the construction of a database of volatile signature patterns and employs advanced AI methods based on Time Series classification. The entire study was developed based on an iterative process of activities, where their execution led to the construction of the final solution. All the code can be found on GitHub (link: https://github.com/michaellopes16/CandidaTimeSeries-Classification.git). It was developed using the Python language on the Jupyter Notebook (in a Core i7 PC, with 16GB of RAM and the GTX 1060 video card) and Google Colab platforms (in your free version). In the initial phases of the research, the primary purpose was to conduct exploratory studies using literary reviews about the main issues related to the work to understand better the state of the art and the best practices for developing the project. In this sense, the course of this section is divided into four stages: Structure and operation of E-nose, Process of sample identification, Analysis and processing of data, and Process of classification of samples.

All these steps seek to select the most promising algorithm for classifying volatiles. In this sense, after the model has been defined, in-place tests must be carried out to ensure its effectiveness in an operational environment. From this, it will be necessary to perform a descriptive study on the use of the solution, aiming to thoroughly analyze its use and better understand its absolute power of contribution, also inserted in this context, a quali-quantitative approach regarding the evaluations.

In this scenario, the project is being developed in partnership with the Mycology department at [Anonimous]. In addition, international alliances are already being prospected, so the collection of samples with different variations can also compose the database under development. The qualitative study should be accomplished through interviews with health professionals to understand the proposal's feasibility better and identify improvement points. Fig. 2 illustrates part of the process related to the sample identification flow, starting from the Acquisition of control samples to the Species identification report.

Structure and functioning of E-Nose

In parallel with constructing the theoretical basis and structuring the problem, the first steps for making the solution were carried out. The database was built from control samples created by ATCC (https://www.atcc.org/about-us), an American company offering quality products and services to the scientific and academic community involving biological materials. These samples were utilized by the Laboratory of Medical Mycology/[Anonimous] for the mycological diagnosis. Then, they were labeled and cultivated in Petri dishes for analysis by the Electronic Nose, developed in partnership with the [Anonimous]. The *E-Nose* identifies the "smell fingerprints" released by the fungi through the Volatile Organic Compounds. In this process, the *E-Nose* uses ten different categories of sensors, seven of them from the manufacturer Figaro Engineering Inc. (TGS826 (Ohm), TGS2611 (Ohm), TGS2603 (Ohm), TGS813 (Ohm), TGS822 (Ohm), TGS2602 (Ohm), TGS823 (Ohm)). The other three are the temperature sensors (Co), pressure (kPa), and humidity (%), used to analyze possible interference of these parameters in the behavior of the samples. A summary of the main functions of the sensors used in the device is in Table 2

To provide greater flexibility in transporting the device, it was built and adapted inside a compact case, with the appropriate seal and structure to withstand all the elements necessary for the Electronic Nose to work. In this case, in addition to the sensors attached to an air chamber on the inside and the on/off button, there is a pump responsible for the suction/injection of gases or air into the chamber, a control valve, and an air filter with activated carbon and, finally, a simple chamber for inserting the Petri dish and collecting the volatile emitted by the microorganisms' reactions. All connections between components and chamber surfaces are made with polytetrafluoroethylene (PTFE) due to its non-stick properties and low coefficient of friction, facilitating cleaning and avoiding the permanence of volatiles between the suction and purge cycles. Fig. 3a presents the Electronic Nose Device used in the experiments.

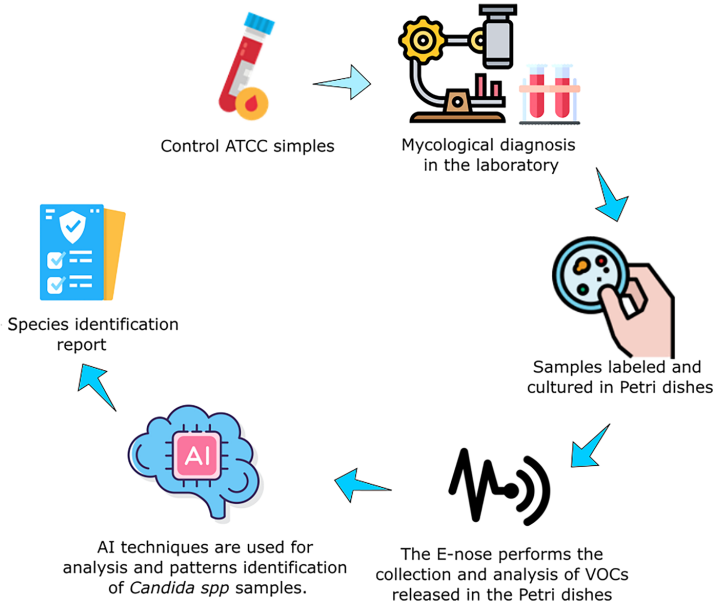


Figure 2. Flow for sample identification and classification. First, control samples derived from the ATCC company are used to analyze and define the mycological diagnosis by the Laboratory of Medical Mycology. With this, the already cultivated species are identified and separated in *Petri* dishes. These cultures are then placed in the E-nose to identify the VOCs. With the collected data, pre-processing routines are executed to use the data already treated by the AI models. At the end, a species identification report is generated.

Sensor	Main Function
TGS826	Ammonia detection
TGS2611-E00	Methane detection
TGS2603	Detection of odors and air contaminants (High sensitivity to series of amines and gases with sulfurous odor and high sensitivity to food odors)
TGS813	Detection of combustible gases (High sensitivity to methane, propane, and butane)
TGS822	Detection of Solvent Vapors (High sensitivity to alcohol and organic solvent)
TGS2602	Detection of air contaminants (High sensitivity to gaseous air contaminants)
TGS823	Detection of Vapors from Organic Solvents (High sensitivity to vapors from organic solvents such as ethanol)

Table 2. Sensors used in the Electronic Nose to identify volatiles emitted by gases generated by the *Candida* species and their functions.

Sample identification process

As briefly mentioned, the first stage of the sample identification process is accomplished by the Medical Mycology Laboratory/[Anonimous], which manipulates samples. After that, the material is labeled with their respective species, cultivated in Petri dishes containing the culture medium Sabouraud Dextrose Agar (see Fig. 3b with an example of samples of *Candida albicans* (URM8368)) and taken for reading by the Electronic Nose, resulting in the generation of the database. The VOCs of species are aspirated with different culture times to increase the heterogeneity of the data and allow better generalization by models in the future. This aspiration at other times also aims to identify whether it is possible to obtain accurate results faster, which is of great importance to help health professionals make decisions.

For each sample collected, the *E-Nose* performs a collection protocol based on three categories of actions, aspiration, stabilization, and purge (cleaning step) (as seen in Fig. 3c), where the completion of all three characterizes the completion of a cycle. For each sample, a volume of three readings per second is collected for 20 seconds in the aspiration phase, for 60s in the stabilization stage, and another 60s in the cleaning phase, totaling an average of 420 readings per cycle in each sensor (for each sample, it is a predefined number of cycles is performed). Considering that numerous samples of the same species are needed to obtain diversity in the data (so that the AI

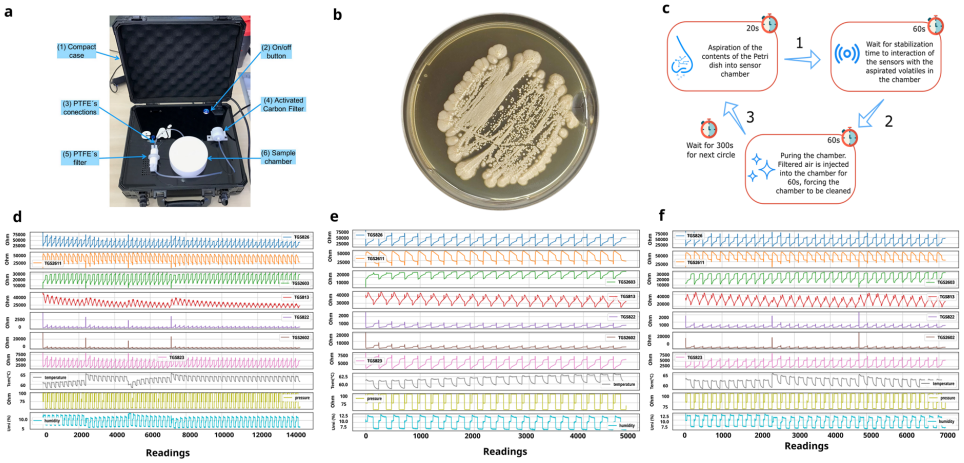


Figure 3. (**a**): Electronic Nose device used in experiments: (1) The Electronic Nose is packaged in a compact box; (2) The on-off switch activates it; (3) All connections are made of PTFE; (4) It has activated carbon filter and (5) PTFE filter; (6) Sample chamber also made of PTFE. (**b**): Example of samples of *Candida albicans* (URM8368) used to create the database. All were cultivated in Petri dishes using Sabouraud Dextrose agar culture medium. (**c**):*E-Nose* collection cycle. (1) Camera suction step (2) Sensor stabilization step (3) Camera cleaning (purge). (**d**): Data from the readings of each sensor over time for the samples of *C. albicans*. (**e**): Data readings from *C. krusei* after one day of culture. (**f**): Data readings from *C. krusei* after two days of culture.

models can satisfactorily learn the patterns of each species), a relevant amount of data was collected in this first step, with 20,189 instances of *C. albicans* (3 isolates - ATCC 14053, URM8368, URM8369), 19,068 of *C. glabrata* (1 isolate - URM6393), 6,989 of *C. haemulonii* (1 isolate - URM6555), 7,067of *C. kodamaea ohmeri* (1 isolate - URM6935), 17,255 of *C. krusei* (3 isolates - ATCC 6258, URM8371, URM6391) and 20,234 of *C. parapsilosis* (3 isolates - ATCC 22019, URM7049, URM7048), totaling 90,802 samples collected in approximately 514 cycles with cultures on different days. There are cycles with different sizes due to inconsistent reading in the E-nose. To solve this, it was necessary to match the cycle sizes, explained in more detail in the Sample classification process section ([Anonymous]-URM is a culture collection affiliated with the Word Federation for Culture Collections).

After the construction of the first version of the database, the need to carry out an analysis of the data was identified, seeking to observe the existence of behaviors or indications of patterns for the different sensors related to each of the species. In addition, this initial check was essential to identify strategies for cleaning and restructuring the base to make its use viable by the learning models.

Data analysis and processing

After generating the data, a descriptive analysis was performed to understand better its behavior and which AI models may be more suitable for identifying the patterns generated by the samples. For this, it was first necessary to analyze and preview the data to get an idea of how they would be about each sensor for each collection of *Candida* spp. After that point, a new database was constructed with the data set of all species collected, with only the sensors considered significant, and with the addition of new columns for labeling the samples about their species and culture time. Another critical point in this information visualization step was using UMAP (Uniform Manifold Approximation and Projection) and PCA (Principal Component Analysis). These two-dimensionality reducers helped to understand the grouping of data better. In this sense, as initial steps for the pre-processing and visualization of information, four relevant points were verified about the data:

- If all sensor data for the same species have similar behavior;
- If there are differences in information between the same species at different collection times;
- If there is a predominance of sensors by species;
- Whether there is a clear division between the data and how it is grouped.

Some graphs with data from all sensors related to the collections of each *Candida* species were generated to analyze the first point. In these, the wave patterns of each collection were observed, following the chronological order of reading, visualized in Fig. 3d for *C. albicans* data.

As can be seen, each of the sensors has a specific wave pattern, varying in well-defined intervals. Some reading peaks in some regions can signal detection errors by the sensors, indicating the presence of possible outliers. Pressure and humidity sensors have an almost constant reading cycle, not interfering at any time with the reading pattern of other gas sensors. The temperature sensor, despite fluctuating a little at some points, also does not interfere with the reading of the other devices, which may be an indication that the alteration of these

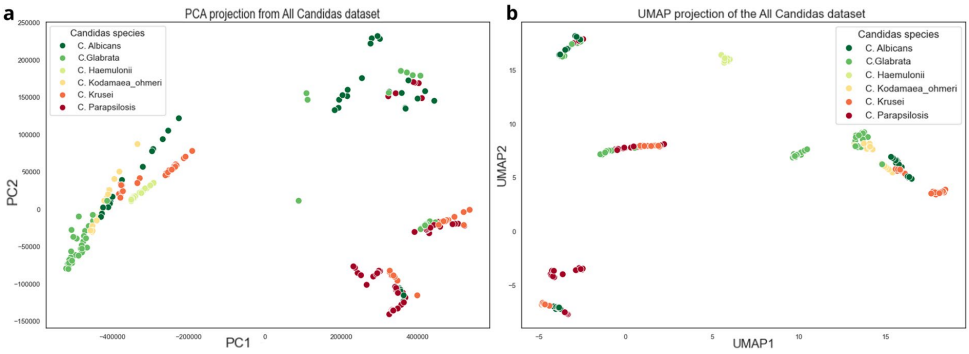


Figure 4. Two dimensions from Principal Component Analysis (**a**) and Uniform Manifold Approximation and Projection for Dimension Reduction (**b**).

parameters does not cause, in this case, any interference in the captures of the others sensors, can be removed from the analysis.

Another important point for this initial analysis is identifying differences between data from the same species but for different collection times. This point helps visually determine if there are significant differences between the readings performed with cultures from different days because the earlier the reading patterns are identified, the better the decision-making process. Fig. 3e and f shows data from one-day and two-day readings for the species *C. krusei*.

As can be seen in Figs. 3e and f, there is a slight distinction between the amplitudes of the waves concerning some sensors from one day to the next. This demonstrates that these devices have a difference in resistance of the volatiles between day 1 and day 2. One hypothesis is that the concentration of gases released by this species changes over time, decreasing in some cases and increasing in others, contributing to the differences in patterns between distinct days. Through this analysis, it is possible to focus on the early cycles of culture analysis, streamlining the decision-making process.

The third point is the possibility of a predominance of a particular sensor per species. This can indicate which sensor can differentiate itself more about each *Candida* species, contributing to the distinction of patterns and selection of the features used in the database consumed by the classification models.

Some experiments show that the behavior of sensors is based on the resistance caused by the gases emitted by each species at the time of reading by the Electronic Nose. Seeking to identify a predominance of a sensor over the species, it was noted that the TGS2602 and TGS822 sensors have a greater amount of readings spread over different resistance (Ohm) levels for *C. parapsilosis*, with the values of the other *Candida* in regions very similar but quite different from *C. parapsilosis*. The opposite occurred with the TGS2611 and TGS823 sensors, where the other samples had more distributed resistances and *C. parapsilosis* more focused on a region. This all shows that some sensors have predominance about some species; however, to identify different levels of resistance about the other, all reading values end up being relevant, as together they become important characteristics for identifying patterns by models.

After analyzing the data for each species and sensor separately, the need to understand how the entire dataset was grouped was identified. For this, two techniques for dimensionality reduction were applied: PCA (Principal Component Analysis) and UMAP (Uniform Manifold Approximation and Projection). In the case of PCA, according to[42], its main objective is to extract relevant information from a set of tabulated data and convert it into a new set of orthogonal variables called Principal Components. In this sense, it is possible to display similarity patterns in the instances and variables as components in a graphical map. On the other hand, the UMAP, according to[43], is an innovative technique of dimensionality reduction that is based on a theoretical structure of Riemannian geometry and algebraic topology, which makes the results derived from its reduction scalable and easily used on accurate data. Unlike PCA, it performs dimension reduction non-linearly, trying to keep similar cases close together and different cases separate. This study applied a two-dimensional decrease for both techniques, which can be analyzed in Fig. 4.

Analyzing the two projections, we can see small groups built by each species. In the case of PCA, the standard difference from *C. parapsilosis*, *C. albicans* and *C. glabara* for the other *Candida* is evident, as their points are well dispersed from the additional data group, with some samples separated from the leading group. This demonstrates that this species has very particular characteristics and can probably be distinguished by IA models. Although the other species are concentrated in a single region, they are well separated, with not much visible shuffling between them. One visual problem is the existence of the same group in different parts of the PCA image. Some models can find issues to distinguish this behavior. In the graph generated by UMAP, it is already possible to see a separation of the data, with groups of species being made in different regions of the graph. This is explained by how UMAP deals with reduction through algebraic topology and similarity measures. It is essential to highlight that, despite not being grouped in the same region of the graph, species with similar characteristics end up staying close to each other and, because they have very different reading averages within the same species

- due to the differences in sensor readings - the same species may contain data that are not very close, considering that this method does not seek its resizing based on the main components, but on similarity measures.

Finally, knowing how the data are arranged and grouped, the base was prepared for use by Time Series models, modified to 2 dimensions, one of the few ones withstand by most models in this segment. From there, experiments with the models were started, which will be detailed in the following sections.

Sample classification process

With clean and structured data, the models were selected based on the results of the data visualization phase and the study on *Inception Time*[34], which compares it with other state-of-the-art models, including its predecessors, the Hierarchical Vote Collective of Transformation-based Ensembles 1[44] and 2[45] (HIVE-COTE 1 and HIVE-COTE 2). The information visualization showed that the data do not overlap and have a single division between them, so there are not many restrictions on which categories of models to use. Thus, in addition to the techniques already mentioned, the K-Neighbors Time Series Classifier (KNN) was also introduced in the experiments, which implements the K-nearest neighbors for time series[46], the Time Series Forest Classifier (TSFC), implementation of a Time Series Forest using intervals[36], the Shapelet Transform Classifier (STC), which uses transformed discriminatory subseries as a classifier[47], the Random Interval Spectral Ensemble (RISE), built based on trees and different sets of partial and automatic correlation of features[37], the ROCKET Classifier (ROCKET)[35] and BOSS Ensemble (BOSS)[48], all Time Series models that will be used as a classifier, due to the temporal characteristic of the data, translated through the parameter *culture_day* from the base.

As previously mentioned, a total of 90,802 readings of the six species of *Candida* were collected in about 514 cycles; however, to obtain a "smell impression" from data, it was necessary to concatenate all readings of all sensors of a cycle in one row of the dataset, resulting in a new set of 397 instances with 821 columns (now, each sample is related to a cycle). Therefore, the base was divided into training, validation, and test sets, with 60% for the first (238 cycles) and 20% for the other (79 and 80 cycles).

Stratified cross-validation is used to maintain a homogenized proportion of data sampling to ensure that the training set can represent the entire population, avoiding sample bias[33]. For each subset used in training, results were obtained for five metrics: accuracy, recall (sensitivity), F1-Score, precision, and specificity[49]. Accuracy measures the proportion of correct model predictions over the evaluated examples. Recall (sensitivity) is applied to measure the portion of patterns correctly identified by the classification model. Specificity is used to test the ability to determine healthy cases accurately. On the other hand, precision is applied to measure the quantity of correctly predicted positive patterns based on the total amount of predicted patterns in a positive class. Finally, the F1-Score or F1-measure portrays the harmonic mean between precision and recall values[38]. All these metrics are calculated based on the values of true positive (TP), false positive (FP), false negative (FN), and true negative (TN), obtained after the crossing of predicted values with the actual values of each class.

Therefore, at the end of the experimentation process, a statistical analysis using the Shapiro-Wilk normality test, the Kruskal-Wallis non-parametric test, and the Nemenyi post-hoc test was applied to understand the statistical significance between the means of the results and highlight the difference between the models tested, which are detailed in the Results and discussions section.

Data availability

Accession codes and database: The code and datasets generated and analyzed during the current study are available in the 'CandidaIdentification' repository: https://github.com/michaellopes16/CandidaIdentification.git. The research described in the article does not use human tissue, only ATCC standard samples.

Received: 22 June 2023; Accepted: 19 December 2023
Published online: 10 January 2024

References

1. Garbee, D. D., Pierce, S. S. & Manning, J. Opportunistic fungal infections in critical care units. *Crit. Care Nurs. Clin. N. Am.* **29**(1), 67–79. https://doi.org/10.1016/j.cnc.2016.09.011 (2017).
2. Li, F. *et al.* A risk prediction model for invasive fungal disease in critically ill patients in the intensive care unit. *Asian Nurs. Res. (Korean Soc. Nurs. Sci.)* **12**(4), 299–303. https://doi.org/10.1016/j.anr.2018.11.004 (2018).
3. Matthaiou, D. K., Christodoulopoulou, T. & Dimopoulos, G. How to treat fungal infections in icu patients. *BMC Infect. Dis.* **15**(1), 1–8. https://doi.org/10.1186/s12879-015-0934-8 (2015).
4. Pietro, P. Performance of *Candida albicans* germ tube antibodies (cagta) and its association with (1 → 3)-β-d-glucan (bdg) for diagnosis daof invasive candidiasis (ic). *Diagn. Microbiol. Infect. Dis* **93**(1), 39–43 (2019).
5. Clancy, P. G. P. C. K. D. A. C. & Marr, K. Clinical practice guideline for the management of candidiasis: 2016 update by the infectious diseases society of america. *Clin. Infect. Dis.* **62**, 1–50 (2016).
6. Silva, E. Prevalência e desfechos clínicos de infecções em utis brasileiras: subanálise do estudo epic ii. revista brasileira de terapia intensiva. *Revista Brasileira de Terapia Intensiva* **24**(2) (2012).
7. Beyda, N. D., Garey, K. W. & Alam, M. J. Comparison of the t2dx instrument with t2candida assay and automated blood culture in the detection of *Candida* species using seeded blood samples. *Diagn. Microbiol. Infect. Dis.* **77**(4), 324–326 (2013).
8. Fernández-Manteca, M. G. Automatic classification of *Candida* species using raman spectroscopy and machine learning. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **290** (2023).
9. Koulenti, D. Severe *Candida* infections in critically ill patients with covid-19. *J. Intensive Med.* (2023).
10. Lass-Florl, C. Interpretation, pitfalls of biomarkers in diagnosis of invasive fungal diseases. *Indian J. Med. Microbiol.* (2022).
11. Terrero-Salcedo, D. & Powers-Fletcher, M. V. Updates in laboratory diagnostics for invasive fungal infections. *J. Clin. Microbiol.* **58**(6) (2020).
12. Patel, R. A moldy application of maldi: Maldi-tof mass spectrometry for fungal identification. *J. Fungi* **5**(1) (2019).
13. Han, S. S., Jeong, Y. S., Sang-Soo & Choi, S.-K. Current scenario and challenges in the direct identification of microorganisms using maldi tof ms. *Microorganisms* **9**(9) (2021).

14. Mahmoudi, S. Methods for identification of *Candida auris*, the yeast of global public health concern: A review. *J. Mycol. Med.* **29**(2), 174–179 (2019).
15. Morath, S. U., Hung, R., & Bennett, J. W. Fungal volatile organic compounds: A review with emphasis on their biotechnological potential. *Fungal Biol. Rev.***26**(2–3) (2012).
16. Shaposhnik, A. V. & Moskalev, P. V. Wine quality assessment using electronic nose. In *IEEE 2021 Asian Conference on Innovation in Technology (ASIANCON)* 1–5 (2021).
17. Shaposhnik, A. & Moskalev, P. V. Processing electronic nose data using artificial neural networks. In *IEEE 2020 4th Scientific School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR)* 208–209 (2020).
18. Karakaya, O. U., Diclehan, & Turkan, M. Electronic nose and its applications: A survey. *Int. J. Autom. Comput.* **17**(2), 179–209 (2020).
19. Castro, M. C. A. Breakthrough of clinical *Candida* cultures identification using the analysis of volatile organic compounds and artificial intelligence methods. *IEEE Sens. J.* **22**(13), 12493–12503 (2022).
20. Vasconcelos, P. J. d. M. Identificação de fungos anemófilos, em ambientes abertos, através de um nariz eletrônico e modelos de inteligência artificial. MS thesis. Universidade Federal de Pernambuco (2022).
21. do Nascimento, J. W. A. Identificação de bactérias comuns em feridas infectadas (staphylococcus aureus, pseudomonas aeruginosa, enterococcus faecalis e escherichia coli) através de um nariz eletrônico e modelos de inteligência artificial. MS thesis. Universidade Federal de Pernambuco (2022).
22. Shaposhnik, A. V. & Moskalev, P. V. Wine quality assessment using electronic nose. In *IEEE 2021 Asian Conf. on Innov. Technol. (ASIANCON)* 1–5 (2021).
23. Chen, L. Quality assessment of royal jelly based on physicochemical properties and flavor profiles using hs-spme-gc/ms combined with electronic nose and electronic tongue analyses. *Food Chem.***403** (2023).
24. Jiarpinijnun, K. O., Asada & Siripatrawan, U. Visualization of volatomic profiles for early detection of fungal infection on storage jasmine brown rice using electronic nose coupled with chemometrics. *Measurement* **157** (2020).
25. Kuchmenko, T. A. Portable electronic nose system for fast gynecological-conditions diagnosis in consulting room: A case study. *Sens. Actuators B Chem.* **358** (2022).
26. Ye, Z., Li, Q. & Liu, Y. Recent progress in smart electronic nose technologies enabled with machine learning methods. *Sensors* **21**(22) (2021).
27. Scheepers, M. H. M. C. Diagnostic performance of electronic noses in cancer diagnoses using exhaled breath: A systematic review and meta-analysis. *Jama Netw. Open* **5**(6) (2022).
28. Zhang, L., Tian, F. & Zhang, D. *Electronic Nose: Algorithmic Challenges* (Springer, Singapore, 2018).
29. Farraia, M. V. The electronic nose technology in clinical diagnosis: A systematic review. *Porto Biomed. J.***4**(4) (2019).
30. Inácio, C. P. Invasive *Candida* tropicalis infection caused by catheter biofilm in a patient with tongue cancer. *Mycopathologia* **184**, 345–346 (2019).
31. Jiarpinijnun, A., K. O. & Siripatrawan, U. Visualization of volatomic profiles for early detection of fungal infection on storage jasmine brown rice using electronic nose coupled with chemometrics. *Meas. J. Int. Meas. Confed.* **157**, 107561. https://doi.org/10.1016/j.measurement.2020.107561 (2020).
32. Wang, Y. *et al.* An optimized deep convolutional neural network for dendrobium classification based on electronic nose. *Sens. Actuators A Phys.* **157**, 111874. https://doi.org/10.1016/j.sna.2020.111874 (2020).
33. Aurélien, G. Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems. Alta Books **2** (2021).
34. Hassan Ismail, F. Inceptiontime: Finding alexnet for time series classification. *Data Min. Knowl. Discov.* **34**(6), 1936–1962. https://doi.org/10.1016/j.sna.2020.111874 (2020).
35. Angus, D., Webb, G. I. & François, P. Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Min. Knowl. Discov.* **34**(5), 1454–1495 (2020).
36. Rufat, B. & Wiese, L. Benchmarking classifiers on medical datasets of uea archive. *Proc. AI Health WWW* **34**(6) (2021).
37. Michael, F., Bagnall, T. & James, L. The contract random interval spectral ensemble (c-rise): The effect of contracting a classifier on accuracy. *International Conference on Hybrid Artificial Intelligence Systems* 381–392 (Springer, 2019).
38. Mohammad, H. & Sulaiman, M. N. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manag. Process* **5**(2), 1 (2015).
39. McKight, P. E., & Najab, J. Kruskal–wallis test. *The Corsini Encyclopedia of Psychology* 1–1 (2010).
40. Pohlert, T. The pairwise multiple comparison of mean ranks package (pmcmr). *R Package* **27**(2019), 9 (2014).
41. Evren, E. Medically important *Candida* spp. identification: An era beyond traditional methods. *Turk. J. Med. Sci.* **52**(3), 834–840 (2022).
42. Abdi, H., & L. J. W. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2**(4), 433–459 (2010).
43. McInnes, H., M. J. & John, H. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXi **v:1802.03426** (2018).
44. Lines, J., Taylor, S. & Bagnall, A. Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification. In *IEEE 16th International Conference on Data Mining (ICDM)* Vol. 2(4), 1041–1046. https://doi.org/10.1109/ICDM.2016.0133 (2016).
45. Middlehurst, M. *et al.* Hive-cote 2.0: A new meta ensemble for time series classification. *Mach. Learn.* **110**, 3211–3243 (2021).
46. Tavenard, R. *et al.* Tslearn, a machine learning toolkit for time series data. *J. Mach. Learn. Res.* **21**(118), 1–6 (2020).
47. Bagnall, A. *et al.* A tale of two toolkits, report the first: Benchmarking time series classification algorithms for correctness and efficiency. arXiv preprint arXiv:1909.05738 (2019).
48. Patrick, S. The boss is concerned with time series classification in the presence of noise. Data mining and knowledge discovery. In *IEEE 16th International Conference on Data Mining (ICDM)*. Vol. 29(6) 1505–1530 (2015).
49. Mortaz, E. Imbalance accuracy metric for model selection in multi-class imbalance classification problems. *Knowl. Based Syst* **210**, 106490 (2020).

Acknowledgements

This work was supported in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação do Amparo a Ciência e Tecnologia do Estado de Pernambuco (FACEPE). Thank you, Maria Andressa, for helping me write and revise the paper.

Author contributions

M.L. conducted major AI experiments and wrote the paper, C.A. collaborated in the construction of the E-nose, C.Z. collaborated in the writing of the article and data analysis, F.D. collaborated in the construction of E-nose, C.P carried out the analyzes of the *Candidas* species in the laboratory, R.G. collaborated with the analysis of candidates in the laboratory, J.G. collaborated in the construction of E-nose, R.P led the analysis of candidates

in the laboratory, L.M. conducted the AI experiments and collaborated on the writing of the article. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to M.L.B. or L.M.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024

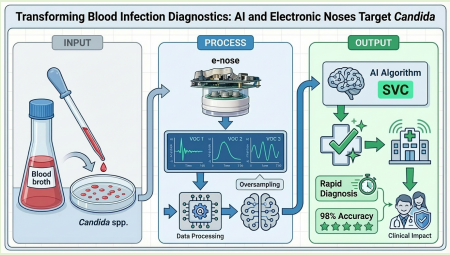
ANNEX B – ARTICLE ACCEPTED FOR PUBLICATION IN IEEE SENSORS.

IEEE Sensors Council IEEE SENSORS JOURNAL, VOL. XX, NO. XX, XXXX 2024 1

Transforming Blood Infection Diagnostics: AI and Electronic Noses Target *Candida*

Michael L. Bastos, Christina Cox, Clayton A. Benevides, Cícero P. Inácio, Rejane P. Neves, Frederico D. de Menezes, Margaret V. Powers-Fletcher and Leandro M. Almeida

***Abstract*— Candidemia is a severe fungal bloodstream infection with high mortality rates, especially in immunocompromised patients. Traditional diagnostic methods, such as blood cultures, suffer from low sensitivity and long turnaround times, delaying targeted antifungal therapy. In this study, we explore the potential of electronic noses (e-noses) combined with artificial intelligence (AI) models to provide a rapid and accurate diagnosis of *Candida* infections directly from blood culture broth samples. Samples were analyzed using an e-nose, generating datasets that underwent preprocessing steps including outlier removal, cycle equalization, feature transformation, and optional oversampling to address species imbalances. To ensure clinical relevance, we employed a dual-scenario validation strategy, assessing both intra-sample sensor stability and inter-sample biological generalization. We tested both traditional machine learning models and time series classifiers, selecting models based on prior research and performance in similar tasks. The best-performing models were evaluated based on accuracy, precision, recall, F1-score, specificity, and computational efficiency. Results demonstrated that while the framework achieved $> 98\%$ consistency in sensor stability tests, the Support Vector Classifier (SVC) emerged as the most robust model for generalization, achieving statistical parity with complex Time Series models like InceptionTime, but with significantly higher computational efficiency. This study highlights the feasibility of AI-enhanced e-noses for rapid *Candida* detection, offering a promising alternative to conventional diagnostics in clinical settings.**



***Index Terms*— Medical computing, Microbiology, Artificial intelligence, Sensors, Volatile organic compounds**

I. Introduction

CANDIDEMIA, an invasive fungal infection caused by yeast species from the *Candida* genus, represents a growing public health issue due to its high mortality rate, which can range from 47% to 60%, depending on the length of hospitalization, especially in immunocompromised patients and those admitted to intensive care units, as highlighted by Schroeder et al. [1] and other studies [2] [3] . In addition, candidemia also incurs high economic costs, ranging from $10,500 to $157,500 per patient. Among the most problematic *Candida* species are *Candida albicans*, *C. glabrata*, *C. tropicalis*, *C. parapsilosis*, and *C. krusei*, responsible for over 90% of invasive fungal disease cases and the third leading cause of bloodstream infections in intensive care units in the United States [4] . Despite advancements in diagnosis and treatment, early detection of candidemia remains a critical challenge in clinical practice. Traditional methods, such as blood cultures, have significant limitations, including low sensitivity and prolonged turnaround times, which often delay the implementation of targeted antifungal therapies. These delays can be particularly detrimental, as candidemia is associated with high mortality rates [5] [6] .

In recent years, the development of innovative technologies, such as electronic noses and artificial intelligence (AI), has emerged as a promising approach to overcome these diagnostic limitations. Electronic noses, inspired by the functionality of the human sense of smell, are devices capable of identifying volatile organic compounds produced by microorganisms,

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. A special thanks goes to everyone at the University of Cincinnati College of Medicine for welcoming me so warmly and for their great contributions to this work.

Michael L. Bastos is with the Universidade Federal de Pernambuco, Centro de Informática, Recife, PE, Brazil and University of Cincinnati, Cincinnati, OH, USA (e-mail: mlb@cin.ufpe.br).

Christina Cox is with University of Cincinnati,College of Medicine, Cincinnati, OH, USA. (e-mail: cox3ci@mail.uc.edu).

Clayton A. Benevides is with the Comissão Nacional de Energia Nuclear, Centro Regional de Ciências Nucleares do Nordeste, Recife, PE, Brazil (e-mail: clayton.benevides@cnen.gov.br).

Cícero P. Inácio is with the Centro de Ciências Médicas, Universidade Federal de Pernambuco, Recife, PE, Brazil (e-mail: ciceropinheiro2000@hotmail.com).

Rejane P. Neves is with the Centro de Ciências Médicas, Universidade Federal de Pernambuco, Recife, PE, Brazil (e-mail: rejadel@yahoo.com.br).

Frederico D. de Menezes is with the Instituto Federal de Pernambuco, Recife, PE, Brazil (e-mail: fredericomenezes@recife.ifpe.edu.br).

Margaret V. Powers-Fletcher is with University of Cincinnati, College of Medicine, Cincinnati, OH, USA. (e-mail: powersmg@ucmail.uc.edu).

Leandro M. Almeida is with the Universidade Federal de Pernambuco, Centro de Informática, Recife, PE, Brazil (e-mail: lma3@cin.ufpe.br).

enabling the rapid and precise differentiation of pathogens. When combined with AI algorithms, these technologies can enhance the sensitivity and specificity of diagnostics, as well as enable rapid diagnosis, leading to a faster and more effective therapeutic response [7] [8] .

This article is an evolution of the work previously developed by Bastos et al. [7] , who used AI techniques and electronic noses to identify *Candida* in culture samples. Although this is a very promising approach and has demonstrated excellent performance, it can still be improved and serve as a basis for new studies. The use of subculture still requires that the samples undergo a period of cultivation and growth, a step that is eliminated when blood culture broth is used directly to read and identify VOCs.

With this in mind, in this new study, the potential of combining electronic noses and artificial intelligence for the identification of *Candida* in blood culture broth samples from infected patients is explored. Preliminary results demonstrate excellent performance of the different models tested, with special attention to the SVC in the family of traditional models, and the KNeighbors Time Series Classifier and Inception Time, from the time series models. Tests with and without oversampling were conducted. Following this Introduction, the next sections will present the Literature Review, Material and Methods, Results, Discussion and, Conclusions and Future Work of the study.

II. Literature Review

The rapid and accurate detection of fungal and bacterial infections is crucial in clinical practice to improve patient outcomes. Recently, innovative approaches combining electronic noses (E-noses) and artificial intelligence (AI) have shown significant potential in different contexts [7]–[13].

For example, both Bastos et al. [7] and Castro et al. [8] have demonstrated that *Candida* species isolated from clinical cultures can be identified through the analysis of VOCs combined with AI methods. In Bastos et al., the Inception Time classifier achieved an average accuracy of 97.46% in the testing phase of clinical isolates. Although this is promising, both studies are limited in the same way, in that the identification approach is dependent upon prior microorganism subculture and isolation to enable VOC collection. This extends the overall diagnostic time and thus limits the potential clinical impact. Eliminating these steps would enable even faster identification, making the method even more efficient in clinical practice.

E-nose analysis has not been studied solely for fungal pathogens, but for bacteria as well. Mohamed et al. [14] compared the effectiveness of an E-nose with the VITEK 2 system in the rapid identification of bloodstream infections caused by two bacterial species, *E. coli* and *K. pneumoniae*. The results indicated that the E-nose not only accelerated the diagnostic process but also showed comparable accuracy to the traditional method. This approach may be particularly advantageous in resource-limited settings where access to automated systems like VITEK 2 is restricted. The main limitation identified was the variability in the E-nose's response depending on the concentration of volatile metabolites and the bacterial growth phase, which can impact diagnostic consistency.

Beyond infection and culture-based identification algorithms, the use of nanomaterial-based sensors to detect patterns of VOCs has been applied for the detection of cancer as well. For example, Peled et al. [15] explored the detection potential for VOCs emitted by cancer cells with specific genetic mutations. Although this study focused on lung cells, it highlights the applicability of E-noses in identifying volatile signatures associated with genetic alterations, suggesting potential for non-invasive and personalized diagnoses across various pathologies. However, the method presents challenges related to specificity, as factors such as diet, metabolism, and other medical conditions can interfere with the composition of detected VOCs, reducing the reliability of the diagnosis in a clinical setting. Similarly, Zhou et al. [16] investigated the detection of gastric cancer through a breath analyzer based on sensors for identifying VOCs exhaled by patients. The study demonstrated that different VOC patterns can be used as non-invasive biomarkers for disease diagnosis, offering a fast and accessible alternative compared to conventional methods such as endoscopy. However, the approach presents challenges related to interindividual variability in volatile profiles, which can be influenced by factors such as diet, metabolism, and preexisting medical conditions, impacting the specificity of the diagnosis.

In the context of invasive pulmonary aspergillosis (IPA) in patients undergoing prolonged chemotherapy-induced neutropenia, de Heer et al. [17] evaluated the feasibility of using electronic nose technology for early, non-invasive detection of this fungal infection. The study demonstrated that the e-nose was able to distinguish between patients with proven/probable IPA and those without, based on the analysis of exhaled volatile organic compounds, offering a promising alternative to traditional diagnostic tools which often require invasive procedures or are limited by low sensitivity. The authors highlighted that this method could potentially be incorporated into clinical workflows for high-risk hematology patients, enabling faster diagnosis and treatment initiation.

However, a key limitation noted was the small sample size and the proof-of-principle nature of the study, which restricts generalizability. In addition, further validation in larger, multicenter cohorts is necessary to assess reproducibility and clinical utility. The performance of the e-nose may also be influenced by external factors such as environmental VOCs or individual variability in breath profiles, posing challenges for standardization in real-world hospital settings [17] .

In summary, the integration of electronic noses and artificial intelligence represents a promising advancement in the rapid and accurate diagnosis of fungal and bacterial infections. These approaches have the potential to transform clinical practice, offering more accessible and efficient diagnostic methods, especially in resource-limited contexts. However, challenges such as sensor calibration , variability of volatile compounds, data quality, and integration with hospital systems must be overcome to ensure the effective implementation of these technologies in routine clinical practice.

In order to translate the advances discussed in the literature into a practical approach, this study followed a rigorous methodological protocol for data collection and analysis. The

implementation of artificial intelligence techniques combined with E-nose sensors required the standardization of experimental processes to ensure the reproducibility and reliability of the obtained results. Thus, the next section details the strategies adopted for sample acquisition, data processing, and predictive model development, enabling a systematic evaluation of the applicability of these technologies in the rapid and accurate identification of fungal infections.

III. Material and Methods

The development of this project was based on two main methodological flows: the methodology for data collection and generation, and the methodology for the development of preprocessing code and model implementation. All the sample collection and data generation steps were carried out under IRB protocol 2020-0313, adhering to all necessary security standards and protocols. All code was developed using the Python programming language, supported by several libraries such as Pandas, Numpy, Sklearn, and Sktime. The latter is the library responsible for providing most of the time series models, except for Inception Time, which was implemented based on the original code from the authors of the project. All experiments were executed on Google Colab using the standard settings of the basic plan (limited GPU access, up to 12GB of RAM, and 100GB of temporary storage). The complete code for the project can be accessed on GitHub , including the code responsible for executing the statistical tests. To better illustrate the two exposed methodologies, Fig. 1 shows the data collection and generation phase, and Fig. 3 the solution development phase.

The first methodological step (Fig. 1) involved the preparation, storage, and analysis of the samples in a controlled environment, ensuring the standardization of procedures and the reliability of the results obtained. Aliquots of remanent broth were collected from clinical blood cultures with known subculture results tested at the [University of Cincinnati Medical Center clinical laboratory]. The aliquots were stored in sterile tubes under refrigeration at a controlled temperature of 4°C, with storage limited to a maximum period of 48 hours to preserve the integrity of the VOCs emitted. For this initial proof-of-concept study, a total of 14 aliquot samples were collected, and the corresponding subculture results were used to label the samples analyzed by the E-nose, defining the target variable of our dataset. In future experiments, a larger number of samples should be included to enhance the robustness and reliability of the validation process. This information was essential for training the AI models, allowing them to identify which sets of VOCs correspond to each type of microorganism. Broth samples included either a single fungal organism isolated in subculture (*C. glabrata*, *C. albicans*, *C. parapsilosis*, or *Cryptococcus neoformans*), a mix of *C. glabrata* and *C. parapsilosis*; or no fungal organisms isolated on subculture (ie. negative).

Following storage and prior to the readings, the samples were kept at room temperature for one hour, allowing thermal stabilization and minimizing the influence of temperature variations on the volatilization of the compounds. To ensure

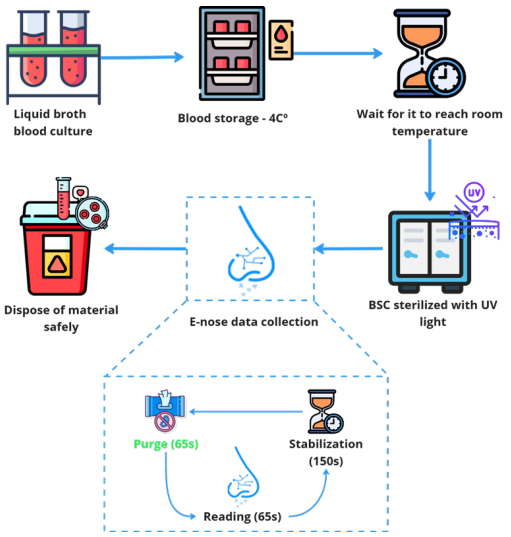


Fig. 1. Experimental Setup for in situ blood sample analysis. Step 1: Collect liquid broth blood culture. Step 2: Store the sample at 4℃. Step 3: Wait for the sample to reach room temperature (25℃). Step 4: Sterilize the collection environment. Step 5: Perform the collection cycle (purging, reading, and stabilization). Step 6: Dispose of the material used.

that the samples were in ideal conditions at the time of measurement, the temperature was checked (around 25 °C) before the start of each reading. The non-*Candida* samples , negative samples, and those with a mixture of more than one species were used to test whether the models could identify potential mixed infections, non-infected samples, and other types of infections other than *Candida*.

Additionally, to ensure optimal sensor performance, the E-nose system was powered on for approximately one hour before initiating the measurements. This preheating period allowed the sensors to reach a stable operating temperature, minimizing baseline drift and reducing fluctuations that could affect VOC detection. During the experiments, the minimum operating temperature recorded by the sensors was approximately 45 °C. Environmental variables were also monitored throughout the procedure: relative humidity ranged from 60% to 75%, and pressure varied between 780 mbar and 820 mbar (slightly lower than standard atmospheric pressure due to the mild vacuum generated by the VOC suction pump). However, after analyzing these variables, they were not included in the final dataset, as they did not influence the measurements. Instead, they were considered solely for evaluating potential environmental interferences, which were not observed during any of the readings. No sensor calibration was performed, and only the raw sensor values were converted into the measurement units used in the analysis.

It is also important to clarify that, in the context of this study, the electronic nose operates as a pattern-recognition device based on relative variations in sensor conductivity rather than on the absolute quantification of specific volatile

compounds. For this reason, no gas chromatography–mass spectrometry (GC–MS or GC×GC–MS) analysis was incorporated into the experimental protocol. The objective of this work was not to identify or enumerate individual VOCs, but rather to capture the global olfactory impression emitted by each microbiologically confirmed sample and use it as input for diagnostic classification. In this framework, traditional analytical chemistry procedures such as calibration curves, linearity assessment, or limits of detection (LOD/LOQ) for specific gases are not applicable. Future studies integrating chromatographic techniques are planned to enable detailed chemical characterization of the biomarkers underlying the observed sensor responses.

In this sense, the experiments were conducted within a biological safety cabinet (Biosafety Cabinet - BSC), ensuring a controlled environment free from external contamination. The BSC was sterilized using UV light for approximately 30 minutes before the start of each experimental cycle, eliminating potential microbiological contaminants. Since the sensors of the E-nose are highly sensitive to external volatile compounds, the use of 70% alcohol for disinfecting the area was avoided, as its vapors could interfere with the detection of VOCs from the samples and compromise the accuracy of the measurements. To prevent cross-contamination, each sample was individually placed in a disposable Petri dish and positioned in the E-nose measurement chamber.

The process steps were Purge, Reading, and Stabilization. In the Purge stage, the E-nose was placed over a Petri dish containing activated charcoal and activated for 65 seconds. This procedure allowed the device to be cleaned, removing any remaining VOCs and preventing cross-contamination between readings. Next came the Reading stage, in which the Petri dish with activated charcoal was replaced with another containing the blood sample.

The E-nose then performed the reading for another 65 seconds. After the reading, the device entered the Stabilization stage, remaining inactive for 150 seconds. This period was necessary for the VOCs to stabilize in the sample. The process was repeated between 10 and 40 times, with variations introduced to assess possible interferences in the VOC quantity after the sample was exposed to air.

All materials used were properly discarded according to current biosafety standards, ensuring the safety of the environment and the researchers involved. Petri dishes, gloves, and other contaminated waste were disposed of in red biosafety bins designated for infectious waste, ensuring compliance with biological material handling protocols. This rigorous methodological approach ensured the collection of reliable and reproducible data, minimizing potential environmental interferences and ensuring the validity of the analyses. The E-nose used was specially developed for the experiments, containing 4 gas sensors (MQ-7, MQ-138, MQ-3, MQ-135) and 3 environmental sensors (temperature, pressure, and humidity).

To assess potential sensor drift effects, these environmental variables were continuously monitored throughout the data acquisition sessions. As shown in Figure 2, the environmental conditions remained highly stable, with no significant long-term drift trends observed. The minor periodic fluctuations

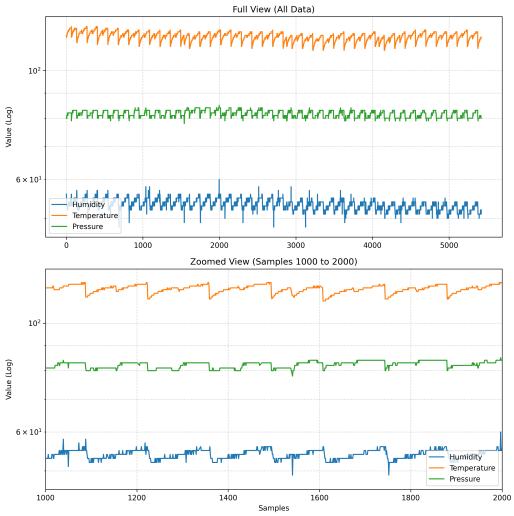


Fig. 2. Time-series monitoring of environmental variables (Temperature, Humidity, and Pressure) acquired simultaneously with gas detection experiments. Data is presented on a logarithmic scale to allow simultaneous visualization of different magnitudes. The stability and parallelism of the curves over time demonstrate the absence of significant sensor drift. Note that the minor periodic oscillations observed reflect the intrinsic dynamics of the electronic nose operating cycles (alternation between purge and sampling stages) rather than external environmental instabilities.

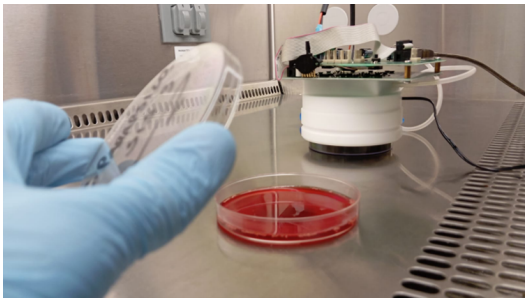


Fig. 3. Experimental setup: Petri dish with blood sample and electronic nose used to read the samples, positioned for the purging stage, on the Petri dish with activated carbon.

visible in the time series correspond to the intrinsic dynamics of the purge and sampling cycles rather than external environmental instability. Given the low standard deviation observed across all environmental parameters (Temperature $\sigma \approx 2.45$, Humidity $\sigma \approx 1.30$, Pressure $\sigma \approx 1.27$), no specific algorithmic compensation for drift was applied, as the experimental conditions were maintained within a controlled range to minimize MOX sensor cross-sensitivity. Fig.2 demonstrates part of the collection process, and the E-nose used in the experiments.

The second stage of the methodology (Fig. 3) involved the development of data preprocessing and the implementation

of predictive models. After the collection of patient samples and their analysis by the electronic nose, each measurement generated a CSV file containing records corresponding to the different phases of the experiment: purging, reading, and stabilization. As an initial step in processing, these files were unified into a single dataset, allowing for a structured approach to handling the information.

The data preprocessing involved a series of steps to ensure its quality and representativeness. Initially, outliers and missing values were removed to prevent negative impacts on the reliability of the analyzes. Next, measurement cycles of different lengths were standardized to maintain the consistency of the time series. After this standardization, each measurement cycle—including the purging, reading, and stabilization phases—was consolidated into a single instance, ensuring that each cycle corresponded directly to an entry in the training dataset (see dashed square in Fig. 3).

In addition to these steps, an oversampling procedure was incorporated to assess the impact of species balancing on model performance. From this point onward, the methodology was divided into two experimental pipelines: one using the original dataset and another employing a balanced version produced through oversampling. The oversampling was performed exclusively on the training partitions using the `RandomOverSampler` algorithm (`random_state = 42`), ensuring that no information from the validation or test sets was introduced, thereby avoiding data leakage. This division allowed a direct comparative analysis of the effect of class balancing on the predictive models, while guaranteeing that all subsequent processing and evaluation steps were conducted identically across both dataset versions.

In the modeling phase, traditional models from different algorithm families, as well as time series models, were selected. The choice of time series models was based on the results obtained in the study by Bastos et al. [7] , prioritizing methods that demonstrated high performance in classifying similar patterns. The traditional models were chosen based on their different training strategies to ensure that the data were evaluated from various perspectives and foundations. This approach allowed for an assessment of whether less complex models could provide competitive results, enabling simpler, more robust, and efficient solutions.

To ensure a rigorous evaluation of the framework's robustness and clinical applicability, the validation strategy was structured into two distinct experimental scenarios. The first scenario, termed Intra-Sample Consistency, employed a Stratified Random Split strategy (Repeated K-Fold with 10 repetitions). In this setup, cycles from the same biological sample could be distributed across both training and testing partitions. This approach was primarily designed to validate the stability of the sensor hardware and the reproducibility of the signal acquisition protocol, ensuring that the E-nose reads the same sample consistently over multiple cycles without signal degradation.

The second scenario, Inter-Sample Generalization, utilized a Grouped Cross-Validation approach to address biological variability. In this configuration, data splitting was strictly anchored to the biological source (sample ID), ensuring that all measurement cycles from a specific patient sample were isolated in the test set while the model was trained on the remaining biological samples. This method acts as a strict 'stress test' for the framework, simulating a real-world point-of-care deployment where the system must classify a completely unknown sample. This prevents data leakage regarding the patient identity and assesses the model's ability to learn generalizable species-specific features rather than memorizing individual sample characteristics.

The performance evaluation of the models was conducted through metrics such as accuracy, precision, F1-score, recall (sensitivity), specificity, and processing time (both training and testing). Additionally, statistical tests were applied to investigate the existence of significant differences between the evaluated groups, analyzing the effectiveness of oversampling and its contribution to improving the models. This statistical analysis helped support the conclusions regarding the impact of balancing on the predictive capability of the algorithms and identify the most appropriate approach for the proposed application.

First, it was necessary to determine the most appropriate normality test, considering the number of repetitions adopted in the cross-validation strategy. According to [18] , when the number of samples is less than 50, the Shapiro-Wilk test is more suitable for assessing data normality. Since cross-validation employed 10 repetitions (the accuracy metric was used as a base value for the tests), this test was selected for normality verification.

Based on the normality test results, it was necessary to choose between two statistical tests: the Wilcoxon test and the paired t-test. Following the recommendations of Imam et al. [19] and Proudfoot et al. [20] , the Wilcoxon test was applied to non-normal distributions, while the paired t-test was used for normally distributed samples. This procedure allowed us to assess the impact of oversampling on the dataset used for training the models.

It is important to clarify that this study focused on the pattern-recognition of the global VOC fingerprint for diagnostic classification, rather than on the quantification of individual compounds. Consequently, analytical performance metrics such as limits of detection (LOD), linearity, and calibration curves for specific gases were not established, as the E-nose operates based on relative resistance changes rather than absolute concentration measurements. Although the AI models demonstrated strong ability to interpret the raw multidimensional sensor responses for classification purposes, future work integrating gas chromatography, mass spectrometry (GC–MS) is recommended to identify and quantify specific VOC biomarkers, enabling a more detailed analytical characterization of the sensor system.

A. Comparison with Previous Method

Compared to the methodology used in the previous study [7] , several significant advancements deserve emphasis. The most important aspect of this comparison lies in the clinical response time induced by the two methods. In the culture-based approach, sample preparation was required, followed

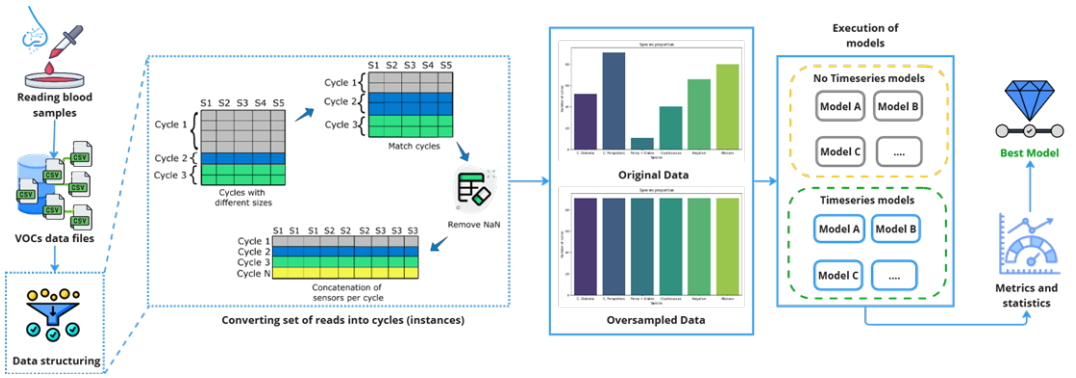


Fig. 4. Workflow of *Candida* identification process using E-nose and AI models. Blood culture aliquots are collected and analyzed by E-nose, which captures the emitted volatile organic compounds (VOCs). The VOC data is stored in files for preprocessing. In this step, a cycle conversion activity is performed, where each set of instances corresponding to a cycle is restructured and merged into a single data row. After that, the original dataset undergoes an oversampling process to address species imbalance, ensuring better model generalization. The original and oversampled datasets are used to train AI models, categorized into traditional classification models and time series classification models. The models are trained using Repeated K-Fold cross-validation with 10 replicates to ensure robustness. The final step involves selecting the best performing models based on evaluation metrics such as accuracy, F1 score, and specificity. The selected model is then deployed for Rapid identification and analysis.

by waiting for the culture incubation period, and only after this time could the results be read. In contrast, by directly using the patient's blood, these steps can be skipped, allowing for almost immediate results right after blood collection. This advancement not only reduces the analysis time but also minimizes the required work and resources, as fungal culture before VOC collection is no longer necessary.

Additionally, a more compact Electronic Nose device was used, featuring fewer sensors and being much more portable. Despite the lower number of sensors and a more manual procedure, the method proved to be just as effective as the one that employed a more robust E-nose with greater sensor redundancy.

Factors such as sensitivity, specificity, accuracy, and precision in both methodologies highlight their potential clinical impact and the benefits they can bring to patients. The blood sample approach, in particular, outperformed the culture-based method (98.18% and 97.46% respectively, for the accuracy metric). These advancements clearly demonstrate the promising nature of this new approach, paving the way for significant growth and clinical impact.

Building upon these methodological advancements, the next step involved evaluating the performance of artificial intelligence models in processing data collected from the electronic nose. The transition from a culture-based approach to direct blood sample analysis not only enhanced clinical efficiency but also introduced new challenges in data handling and model training. To ensure robust and reliable predictions, different classification algorithms were tested, considering both time series and traditional machine learning models. The following section presents the results of these evaluations, highlighting the impact of oversampling strategies, model selection, and key performance metrics in optimizing the diagnostic accuracy of this innovative approach.

IV. Results

The artificial intelligence models were evaluated using data derived from the electronic nose (e-nose). To provide a comprehensive validation, the results are presented in two scenarios: (A) Intra-Sample Consistency, which evaluates sensor stability using random split, and (B) Inter-Sample Generalization, which evaluates biological robustness using Grouped Cross-Validation.

A. Scenario A: Intra-Sample Consistency (Sensor Stability)

The artificial intelligence models were evaluated using data derived from the electronic nose (e-nose), with and without the application of oversampling. The oversampling strategy was employed to address the species imbalance observed in the data, while the original data were used to evaluate performance without adjustment [21]. The metrics used to assess performance included accuracy, precision, F1-score, sensitivity (Recall), specificity, and execution time (Time (S)). The models tested included time series approaches such as KNeighbors Time Series Classifier (KNTC) [22] , Random Interval Spectral Ensemble (RISE) [23] , ROCKET Classifier [24], Time Series Forest Classifier (TSFC) [25], Inception Time [26], and traditional models such as DecisionTree Classifier (DTC), KNeighbors Classifier (KNN), Random Forest Classifier (RF), SVC, and XGBClassifier (XGBC). The selection of time series models was based on the classifiers that performed best in the previous study [7], which used culture samples. For the traditional models, the strategy was to apply models from different families to understand whether simpler models could perform as well as the more complex models in current literature. To provide a better understanding, the overall results for the tested models are presented in Table I and Table II, which detail the metrics for strategies with

and without oversampling, for the Training/Validation and Test data, respectively.

In traditional models, the application of oversampling resulted in a significant increase in all metrics compared to the original data, suggesting that species balancing contributed to improving the models' discrimination ability. In contrast, some time series models demonstrated lower robustness in both conditions, performing worse than some of the traditional models, regardless of the application of oversampling. A factor that may explain this slight difference is the amount of data. As more complex models, they require a larger number of samples to perform better than simpler models. Training time was also a distinguishing factor for the traditional models, which took much less time to complete training.

Even with some traditional models outperforming the time series models, overall, the ROCKET model showed the best performance among the approaches evaluated, achieving an accuracy of 98.86%, a sensitivity (recall) of 99.10%, and a specificity of 99.77% on the training/validation set with oversampling. In the no oversampling group, these indicators were 98.15%, 97.77%, and 99.59%, respectively. Among the traditional models, the SVC performed best, with values very close to ROCKET, achieving an accuracy of 97.70%, a sensitivity of 97.83%, and a specificity of 99.56% for the oversampling group. It was also the best traditional model in the original data group, with 93.59%, 96.13%, and 98.68% for the same indicators, respectively. On the other hand, ROCKET was the slowest model to complete training among all models evaluated in the original data set and one of the worst in the oversampling group.

When looking at the test set data, we notice some interesting differences. ROCKET is no longer the best-performing model, losing its position to the KNeighbors Time Series Classifier and Inception Time in the time series models group, and to the SVC and Random Forest Classifier in the traditional models group. It is with the test data that we assess how efficient the model is, as it is exposed to information it has never seen before, representing the closest scenario to a real-world context. In this sense, the more complex models, despite showing excellent performance in all metrics, are on par with the SVC and Random Forest Classifier, with less than a 1% difference in all metrics compared to the top time series models. Additionally, the time taken to obtain the prediction was often much lower (0.009520s and 0.025079s, respectively), demonstrating that they are lighter alternatives compared to the time series models.

1) Best Performing Models and Metrics by Species: To better understand how the models perform for each species , the same metrics used for training and testing were collected for the different species, focusing only on the models with the best performance in the test stage. This step aims to identify which species the models misclassify the most and whether any species imbalance may raise concerns about the model's prediction quality. In this sense, among the models tested, the SVC stood out with the best overall results, achieving an accuracy of 98.18%, precision of 98.20%, F1-score of 98.22%, sensitivity of 98.33%, and specificity of 99.64% on the oversampling data set. In the no oversampling group, these indicators were lower, with an accuracy of 92.65%, precision of 94.70%, and F1-score of 88.97%.

The detailed metrics by species for this model are presented in Table III. The results show that while the species *C. albicans*, *C. neoformans* , *C. glabrata*, and *C. parapsilosis* had high sensitivity and precision, the *mixed culture* (the union of *Candida parapsilosis* and *glabrata* in the same petri dish) and negative (uninfected sample) showed slightly lower performance, possibly due to the smaller number of samples represented, only in the no oversampling group. For the balanced data, all species were correctly identified by the model, with no species having any metric below 94%.

Among the time series models, Inception Time was chosen to be evaluated individually alongside the species, as it is one of the best-performing models overall in the metrics and took the least time during the test phase (among the time series models). In this case, it is noticeable that the model performs better, even with imbalanced data. There is a slight drop in the numbers for the same species that presented difficulties for the SVC; however, this drop was much less significant in a general context. Based on this information, even with the original data containing a slight imbalance regarding some species, there are alternatives that can efficiently address this problem without major drawbacks. Table IV shows the result of this evaluation.

B. Scenario B: Inter-Sample Generalization (Robustness Analysis)

While Scenario A demonstrated high sensor stability and reproducibility, Scenario B was designed as a stress test to evaluate the framework's ability to generalize to new biological donors. This was achieved using Grouped Cross-Validation, ensuring no data leakage between subjects.

Table V compares the performance of Traditional and Time Series models under this rigorous condition. Unlike the previous scenario, traditional models demonstrated superior robustness. The **SVC** achieved the best overall stability with an accuracy of 76.35%, significantly outperforming complex Time Series models in the low-data regime. However, it is crucial to note that applying Oversampling to the **InceptionTime** model recovered a significant portion of its performance (increasing Accuracy from 52% to 67%), suggesting that Time Series models will benefit most from future dataset expansion.

Table VI details the capabilities of the system for clinical screening. Despite the drop in global metrics compared to Scenario A, the system maintained high efficacy in identifying the most critical pathogen, *C. albicans*. The SVC model achieved nearly perfect metrics for this species (Recall $> 97\%$, Precision 100%). The lower global scores were primarily driven by the difficulty in distinguishing between non-albicans species (e.g., *C. parapsilosis*) and negative controls in a few folds, highlighting the biological variability that will be addressed with larger cohorts.

C. Statistical Comparison Between Oversampling and No-Oversampling Strategies (Scenario A: Intra-Sample Consistency)

To evaluate whether the application of oversampling significantly impacts the performance of the classifiers, statistical

TABLE I

THE TABLE COMPARES THE PERFORMANCE OF TRAIN AND VALIDATION OF DIFFERENT CLASSIFICATION MODELS, BOTH TRADITIONAL AND TIME SERIES, WITH AND WITHOUT OVERSAMPLING, USING THE METRICS OF ACCURACY, PRECISION, F1-SCORE, RECALL (SENSITIVITY), SPECIFICITY, AND STANDARD DEVIATIONS. IT ALSO PRESENTS THE EXECUTION TIMES (S) OF EACH MODEL IN BOTH SCENARIOS. THE BEST PERFORMANCES IN EACH METRIC ARE HIGHLIGHTED IN BOLD. ALL STANDARD DEVIATIONS WERE BELOW 0.084, BOTH WITH AND WITHOUT OVERSAMPLING.

Category	Classifier	With Oversampling							No Oversampling						
		Accuracy	Precision	F1-Score	Recall	Specificity	Deviations	Time (s)	Accuracy	Precision	F1-Score	Recall	Specificity	Deviations	Time (s)
Time Series	KNeighbors Time Series	98.85%	98.76%	98.80%	98.92%	99.77%	0.0225	5.39×10^{20}	94.44%	96.18%	94.28%	93.76%	98.78%	**0.0208**	2.1×10^{21}
	RISE	97.15%	97.03%	97.13%	97.32%	99.43%	0.0173	5.85×10^{20}	95.37%	96.73%	94.78%	93.94%	99.00%	0.0275	3.8×10^{20}
	ROCKET	**98.86%**	**99.10%**	**99.02%**	**99.10%**	**99.77%**	0.0175	6.17×10^{20}	**98.15%**	**97.77%**	**97.74%**	**97.99%**	**99.59%**	0.0232	4.7×10^{21}
	Time Series Forest	97.13%	96.48%	96.63%	97.04%	99.44%	0.0235	7.51×10^{20}	95.42%	96.58%	94.78%	94.92%	99.05%	0.0276	5.6×10^{20}
	Inception Time	96.82%	96.85%	96.72%	96.66%	99.36%	**0.0084**	4.66×10^{20}	94.85%	94.40%	94.45%	96.16%	98.98%	0.0844	2.1×10^{21}
Traditional	Decision Tree	94.29%	94.59%	93.76%	93.56%	98.86%	0.0298	0.794573	88.06%	88.02%	84.17%	85.38%	97.52%	0.0510	0.62
	KNeighbors	89.15%	89.48%	88.96%	88.86%	97.81%	0.0374	**0.369690**	85.32%	81.77%	79.95%	80.22%	97.01%	0.0367	**0.38**
	Random Forest	97.70%	**97.84%**	**97.49%**	97.32%	99.52%	0.0263	3.30×10^{20}	92.59%	92.40%	89.49%	88.67%	98.47%	0.0359	2.7×10^{16}
	SVC	**97.70%**	97.16%	97.41%	**97.83%**	**99.56%**	**0.0141**	1.31×10^{21}	93.59%	96.13%	91.56%	90.28%	98.68%	0.0275	1.1×10^{21}
	XGBClassifier	96.55%	96.26%	96.10%	96.01%	99.31%	0.0190	2.77×10^{21}	88.99%	87.36%	83.40%	85.53%	97.83%	0.0360	2.2×10^{21}

TABLE II

THE TABLE COMPARES THE PERFORMANCE OF TEST OF DIFFERENT CLASSIFICATION MODELS, BOTH TRADITIONAL AND TIME SERIES, WITH AND WITHOUT OVERSAMPLING, USING THE METRICS OF ACCURACY, PRECISION, F1-SCORE, RECALL(SENSITIVITY), SPECIFICITY AND TIME(S). IT ALSO PRESENTS THE EXECUTION TIMES OF EACH MODEL IN BOTH SCENARIOS. THE BEST PERFORMANCES IN EACH METRIC ARE HIGHLIGHTED IN BOLD.

Category	Classifier	WIth Oversampling						No Oversampling					
		Accuracy	Precision	F1-Score	Recall	Specificity	Time (s)	Accuracy	Precision	F1-Score	Recall	Specificity	Time (s)
Time Series	KNeighbors Time Series	**97.27%**	**97.35%**	**97.19%**	**97.37%**	**99.46%**	6.62×10^{20}	94.12%	**94.07%**	**94.67%**	**96.16%**	98.86%	2.72×10^{21}
	RISE	94.55%	94.65%	94.40%	94.54%	98.92%	1.16×10^{21}	85.29%	89.44%	81.97%	81.38%	97.07%	7.46×10^{20}
	ROCKET	93.64%	95.33%	93.62%	93.70%	98.73%	1.05×10^{21}	**95.59%**	91.67%	93.52%	96.88%	**99.17%**	7.90×10^{20}
	Time Series Forest	89.09%	90.47%	89.21%	88.81%	97.81%	1.23×10^{20}	91.18%	90.50%	91.73%	93.68%	98.27%	0.97
	Inception Time	**97.25%**	**96.26%**	**96.69%**	**97.29%**	**99.48%**	1.86×10^{20}	91.18%	93.57%	92.80%	93.15%	98.14%	1.92×10^{21}
Traditional	Decision Tree	96.36%	96.43%	96.34%	96.42%	99.28%	0.0047	85.29%	79.71%	81.38%	88.27%	97.16%	0.0033
	KNeighbors	90.00%	90.26%	89.61%	89.51%	98.01%	0.0117	85.29%	84.40%	83.87%	84.32%	97.07%	0.0100
	Random Forest	98.18%	98.14%	98.14%	98.20%	99.64%	**0.0095**	**95.59%**	90.55%	**92.24%**	**95.23%**	**99.19%**	**0.0087**
	SVC	**98.18%**	**98.20%**	**98.22%**	**98.33%**	**99.64%**	0.0251	92.65%	**94.70%**	88.97%	86.88%	98.40%	0.0224
	XGBClassifier	96.36%	96.24%	96.24%	96.39%	99.28%	0.3867	92.65%	85.39%	87.25%	92.12%	98.66%	0.0812

TABLE III

VALUES FOR ALL METRICS (ACCURACY, PRECISION, F1-SCORE, RECALL(SENSITIVITY), SPECIFICITY) COLLECTED BY SPECIES FOR THE SVC CLASSIFIER, EXECUTED FOR DATA WITH AND WITHOUT OVERSAMPLING. THE TABLE SHOWS THAT THERE IS A DROP IN MODEL PERFORMANCE IN RELATION TO THE *MIXED CULTURE* AND NEGATIVE SPECIES. THIS SUGGESTS THAT THIS MODEL NEEDS MORE INSTANCES OF THESE SPECIES TO PERFORM BETTER IN THESE SPECIFIC CONTEXT.

Specie	ID	With Oversampling					No Oversampling				
		Accuracy	Precision	F1-Score	Recall	Specificity	Accuracy	Precision	F1-Score	Recall	Specificity
glabrata	0	99,09%	94,44%	97,14%	100,00%	98,92%	97,06%	100,00%	88,89%	80,00%	100,00%
parapsilosis	1	99,09%	100,00%	97,30%	94,74%	100,00%	94,12%	91,30%	91,30%	91,30%	95,56%
mixed culture	2	100,00%	100,00%	100,00%	100,00%	100,00%	98,53%	100,00%	66,67%	50,00%	100,00%
cryptococcus	3	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
negative	4	99,09%	94,74%	97,30%	100,00%	98,91%	95,59%	76,92%	86,96%	100,00%	94,83%
albicans	5	99,09%	100,00%	97,56%	95,24%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

TABLE IV

VALUES FOR ALL METRICS (ACCURACY, PRECISION, F1-SCORE, RECALL(SENSITIVITY), SPECIFICITY) COLLECTED BY SPECIES FOR THE RANDOM FOREST CLASSIFIER, EXECUTED FOR DATA WITH AND WITHOUT OVERSAMPLING. AS IN TABLE III, THERE IS A DECLINE IN PERFORMANCE IN THE ABSENCE OF OVERSAMPLING FOR SOME SPECIES, SUGGESTING THE NEED FOR ADDITIONAL TRAINING INSTANCES.

Specie	ID	With Oversampling					No Oversampling				
		Accuracy	Precision	F1-Score	Recall	Specificity	Accuracy	Precision	F1-Score	Recall	Specificity
glabrata	0	100,00%	100,00%	100,00%	100,00%	100,00%	97,94%	96,00%	92,84%	90,00%	99,31%
parapsilosis	1	96,55%	89,23%	90,09%	91,05%	97,69%	93,53%	97,61%	87,89%	82,17%	99,33%
mixed culture	2	99,82%	99,00%	99,49%	100,00%	99,78%	99,56%	100,00%	90,00%	85,00%	100,00%
cryptococcus	3	98,18%	97,50%	93,50%	90,00%	99,57%	95,00%	76,78%	85,18%	100,00%	94,43%
negative	4	98,55%	94,11%	95,58%	97,22%	98,80%	97,21%	87,15%	91,29%	97,00%	97,24%
albicans	5	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

tests were conducted comparing the results before and after the technique. First, the Shapiro-Wilk test was performed to check the normality of the distributions of the two groups (with and without oversampling). Then, depending on the normality of the data, the paired t-test was applied for normal distributions, and the Wilcoxon test for non-normal distributions [19] [20].

The results of the normality tests indicate that, for most classifiers (see Figure 1), at least one of the groups does not

TABLE V
SCENARIO B (GROUPED CV): PERFORMANCE COMPARISON OF MODELS (MEAN ± STD). TRADITIONAL MODELS LIKE SVC SHOWED HIGHER STABILITY IN GENERALIZATION, WHILE OVERSAMPLING WAS CRITICAL FOR RECOVERING PERFORMANCE IN DEEP LEARNING MODELS (INCEPTIONTIME).

Model	Oversampling	Accuracy	Precision	Recall	F1-Macro	Time (s)
Traditional Models						
Decision Tree	No	0.6135 ± 0.41	0.4792 ± 0.38	0.4193 ± 0.39	0.4420 ± 0.38	0.0270 ± 0.0055
KNeighbors	No	0.6667 ± 0.42	0.4167 ± 0.32	0.3760 ± 0.31	0.3938 ± 0.31	0.0077 ± 0.0007
Random Forest	Yes	0.6448 ± 0.42	0.5417 ± 0.42	0.4938 ± 0.44	0.5126 ± 0.43	0.218 ± 0.037
SVC	**Yes/No**	**0.7635 ± 0.38**	**0.6875 ± 0.37**	**0.6318 ± 0.41**	**0.6484 ± 0.40**	**0.020 ± 0.001**
Time Series Models						
InceptionTime	No	0.5213 ± 0.51	0.5417 ± 0.50	0.5071 ± 0.52	0.5122 ± 0.52	205.07 ± 12.60
InceptionTime	**Yes**	**0.6738 ± 0.46**	**0.6875 ± 0.45**	**0.6494 ± 0.48**	**0.6601 ± 0.47**	244.54 ± 22.38
RISE	Yes	0.6612 ± 0.32	0.5417 ± 0.33	0.4420 ± 0.36	0.4795 ± 0.35	95.59 ± 159.96
TimeSeriesForest	Yes	0.6448 ± 0.42	0.4375 ± 0.32	0.3849 ± 0.31	0.4053 ± 0.31	18.90 ± 13.29

TABLE VI
SCENARIO B (GROUPED CV): DETAILED PERFORMANCE BY CLASS FOR THE BEST MODELS. NOTE THE HIGH ROBUSTNESS IN DETECTING *C. ALBICANS* EVEN UNDER STRESS TESTS.

Model	Class	Accuracy	Precision	Recall	F1-Score	Specificity
SVC	*C. Albicans*	**0.9916**	**1.0000**	**0.9750**	**0.9873**	**1.0000**
	C. Parapsilosis	0.7764	0.7159	0.6923	0.7039	0.8288
	Negative	0.7679	0.5775	0.6212	0.5985	0.8246
InceptionTime (Over)	*C. Albicans*	0.7257	0.6154	0.5000	0.5517	0.8408
	C. Parapsilosis	0.7257	0.5833	**1.0000**	0.7368	0.5548
	Negative	0.7890	1.0000	0.2424	0.3902	1.0000

follow a normal distribution (Shapiro-Wilk values ¡ 0.05). This justifies the use of the Wilcoxon test for these cases, which is a non-parametric test suitable for comparisons between paired samples without normality assumptions. Among the evaluated classifiers, those that showed statistically significant differences after the application of oversampling include:

- RISE: The paired t-test resulted in a p-value of 0.045, indicating that oversampling significantly impacted the performance of this classifier.
- KNTC: The Wilcoxon test presented a p-value of 0.0098, pointing to a significant difference between the two groups.
- Inception Time: The p-value obtained in the Wilcoxon test was 0.0019, reinforcing the presence of a statistically significant impact.
- XGBC: The paired t-test obtained a p-value of 0.0018, indicating a significant difference between the groups.
- SVC: The p-value of 0.0039 obtained in the Wilcoxon test confirms a statistically significant difference.
- DTC: The paired t-test resulted in a p-value of 0.0074, showing a significant impact.
- RFC: The Wilcoxon test revealed a p-value of 0.0273, indicating a significant difference.

On the other hand, some classifiers, such as TSFC, ROCKET and KNN, did not show statistically significant differences, with p-values above the 0.05 significance level. This suggests that for these models, the introduction of oversampling did not result in statistically relevant improvements in performance.

In summary, the results indicate that the impact of oversampling varies across different classifiers, being more pronounced in some models than in others. This observation reinforces the importance of individually evaluating the effectiveness of data balancing techniques before their final implementation.

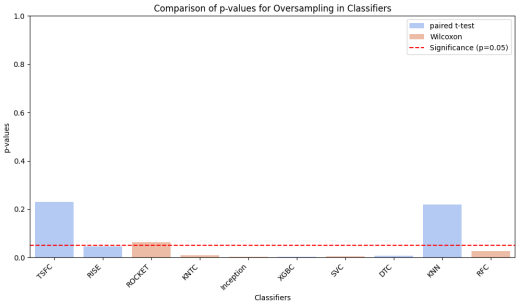


Fig. 5. Comparison of p-values from the applied statistical tests, indicating which models were significantly impacted by the application of oversampling. The red line represents the confidence interval. Models with bars above this interval were not significantly affected by the use of the oversampling strategy.

D. Statistical Robustness in Inter-Sample Generalization (Scenario B)

While the previous analysis (Scenario A) demonstrated the positive impact of oversampling in a controlled random-split environment, the statistical evaluation of the Inter-Sample Generalization scenario (Scenario B) reveals a different dy-

namic regarding biological variability. Using the same statistical rigor (Shapiro-Wilk for normality followed by paired tests), we compared the performance of models with and without oversampling under the strict Grouped Cross-Validation protocol.

In this high-stress testing environment ($N = 8$ independent biological samples), the statistical tests (Wilcoxon Signed-Rank) indicated no significant difference ($p > 0.05$) between the Oversampling and No-Oversampling strategies for the majority of models, including the top performers SVC ($p = 1.00$) and InceptionTime ($p = 0.18$).

This lack of statistical significance, despite the observed increase in mean accuracy for Deep Learning models (e.g., InceptionTime improved from 52% to 67%), can be attributed to the high variance introduced by the biological heterogeneity of the subjects. The standard deviations in Scenario B ($\approx \pm 40\%$) are considerably larger than in Scenario A ($\approx \pm 2\%$), masking the benefits of oversampling from a strictly statistical p-value perspective.

Furthermore, a direct statistical comparison between the best traditional model (SVC) and the best deep learning model (InceptionTime) with oversampling yielded a p-value of 0.715 (Wilcoxon), indicating a "statistical tie". This finding is critical for the proposed framework: it suggests that in resource-constrained clinical settings, the simpler SVC model provides diagnostic power statistically equivalent to complex Deep Learning architectures, validating its use as a highly efficient screening tool.

E. Time Series and Traditional Data Analysis

Time series models showed greater effectiveness in detecting patterns associated with *Candida* infections, especially in the dynamic analysis of signals collected by the electronic nose. Among them, the KNeighbors Time Series Classifier and Inception Time achieved the best performances, reaching 97.27% and 97.25% accuracy, respectively, when oversampling was applied (Scenario A). These models maintained a high generalization capability even without oversampling, reinforcing their robustness in the face of imbalanced data within controlled partitions.

In contrast, traditional models exhibited competitive performance, with an emphasis on the Random Forest Classifier and SVC, which achieved 98.18% accuracy with oversampling. However, without this balancing technique, their precision dropped significantly, highlighting the dependence of these models on specific preprocessing steps to handle imbalanced data. The application of oversampling had a widespread positive impact, especially for traditional models, significantly improving accuracy, precision, and F1-score. This underscores the importance of balancing strategies to optimize performance in scenarios with uneven species distribution.

When evaluating the models under the stricter Inter-Sample Generalization protocol (Scenario B), a crucial shift in performance dynamics was observed. While time series models like InceptionTime suffered a noticeable drop in accuracy without data balancing (falling to $\approx 52\%$), the traditional SVC model demonstrated superior stability, maintaining an accuracy of $\approx 76\%$ even in the absence of oversampling. Furthermore, statistical tests (Wilcoxon Signed-Rank) in this scenario revealed no significant difference ($p > 0.05$) between the best traditional model (SVC) and the best deep learning architecture (InceptionTime) when oversampling was applied. This indicates that, despite the theoretical advantages of deep learning for temporal signals, simpler traditional models may offer equivalent diagnostic power with lower computational cost when generalizing to new biological donors.

Overall, while time series models demonstrate greater flexibility in identifying complex temporal features, the traditional approach proves to be a highly efficient alternative for clinical screening. These findings further emphasize the varying impact of oversampling on different classification models, reinforcing the need for tailored preprocessing approaches. As the results demonstrate, while traditional models significantly benefit from oversampling, time series models tend to exhibit inherent robustness to specie imbalance only when signal consistency is high. This observation naturally leads to a deeper discussion on the comparative strengths and weaknesses of these two methodological approaches. The following section explores these aspects in greater detail, analyzing not only classification accuracy but also computational efficiency and clinical applicability, ultimately aiming to determine the most suitable models for real-world diagnostic scenarios.

V. Discussion

The results obtained with the AI models highlight significant differences in performance between traditional models and time series-based models. The application of oversampling proved effective in improving classification metrics, reducing the impact of specie imbalance on traditional models. In contrast, for time series models, this technique did not have such a pronounced impact, suggesting that these models perform well regardless of data balancing. This robustness is particularly relevant in clinical settings where data distribution can be highly variable.

Among the evaluated models, ROCKET demonstrated excellent performance in the training and validation phase, standing out mainly in the oversampling version, where it achieved the highest accuracy (98.86%) and sensitivity (99.10%). However, when exposed to the test data, its performance was surpassed by the KNeighbors Time Series Classifier and Inception Time among the time series models, and by the SVC and Random Forest Classifier among the traditional models. This suggests that, despite high performance in the validation phase, some models may have a lower generalization ability, possibly due to the specific characteristics of the training set.

The processing time analysis revealed that traditional models are considerably faster than time series models. The SVC and Random Forest Classifier showed significantly lower prediction times (0.009520s and 0.025079s, respectively), making them viable alternatives for applications requiring computational efficiency without significant compromise in accuracy. However, despite their slower processing times, time series models such as Inception Time may offer superior feature extraction capabilities, making them more suited for detecting subtle temporal variations in E-Nose signals.

In the evaluation of metrics by specie, it was observed that the SVC showed consistent performance, with an accuracy of 98.18% in the oversampling set. However, in the set without oversampling, there was a noticeable drop, emphasizing the importance of specie balancing for traditional models in detecting subtle patterns in E-Nose signals. In contrast, Inception Time demonstrated greater robustness to data imbalance, maintaining good metrics even without the application of oversampling. This suggests that time series models inherently capture temporal dependencies that contribute to their resilience in varying data conditions.

From a clinical perspective, the ability to rapidly detect fungal pathogens directly from blood culture broth represents a transformative advancement. Traditional subculture-based methods for bloodstream pathogen isolation and identification can take some days, delaying treatment decisions. The E-Nose/AI approach based on blood culture broth reduces this turnaround time to minutes, offering substantial benefits for antifungal stewardship and patient management, particularly in intensive care environment. By bypassing the need for sample preparation and fungal isolation, this method allows for faster interventions, which may be crucial for improving patient outcomes and reducing mortality associated with fungal infections.

In this scenario, diagnostic accuracy and speed are critical factors. While models like SVC and Random Forest Classifier provide a balance between performance and computational efficiency, more specialized approaches, such as KNeighbors Time Series Classifier and Inception Time, warrant further exploration. When evaluating model performance by specie, it becomes evident that more robust models, such as Inception Time, excel in capturing temporal variations in signals, even when some species have a reduced number of samples.

Deep learning models generally demonstrate superior generalization in handling imbalanced datasets compared to linear models like SVC. This ability makes them particularly suitable for complex medical diagnostics. Nevertheless, a wide array of strong alternatives exists for model selection, enhancing the system's reliability, portability, and scalability. This is especially crucial in challenging clinical conditions, where rapid and precise diagnostics are essential for effective patient management and treatment decisions.

To rigorously validate the framework's applicability in a real-world context with unseen patients, an additional analysis focusing on Inter-Sample Generalization was conducted using Grouped Cross-Validation. Unlike the intra-sample consistency scenario, which confirmed sensor stability with accuracies exceeding 98%, this "stress test" revealed the impact of biological variability inherent in the pilot dataset ($N = 8$). The observed reduction in overall accuracy to the 60-76% range was anticipated and highlights the challenge of generalizing complex metabolic profiles from a limited number of biological donors, rather than indicating sensor limitations.

In this rigorous testing environment, the robustness of the traditional SVC model was particularly notable. While Deep Learning models like InceptionTime experienced a significant performance drop without data balancing (falling to $\approx$ 52%), the SVC maintained a stable accuracy of $\approx$ 76% even without oversampling. Furthermore, statistical analysis (Wilcoxon Signed-Rank test, $p > 0.05$) revealed no significant performance difference between the SVC and the best-performing deep learning architecture when oversampling was applied. This statistical tie suggests that, for the current dataset size, simpler and computationally efficient models like SVC provide diagnostic power equivalent to complex architectures, reinforcing their suitability for resource-constrained point-of-care devices.

When compared to the culture-based methodology explored by Bastos [7], the results are highly similar. However, the blood broth approach offers a significant advantage by eliminating the need for sample processing, thereby avoiding exposure to VOCs from external sources, such as the agar used for fungal cultivation. This can lead to more precise readings and improve the models' ability to accurately identify volatile compound patterns.

While the results are promising, there are several considerations for the real-world application of this approach. Sensor drift, a well-known challenge with electronic noses, can affect sensor responses over time due to prolonged use or environmental factors, possibly requiring periodic recalibration [27]. Additionally, cross-sensitivity to VOCs from various microorganisms or external sources could influence classification accuracy, underlining the importance of robust preprocessing techniques and the inclusion of diverse sample types. Despite high accuracy in the evaluated datasets, multicenter validation is necessary to ensure the consistency of findings across different clinical settings and sample collection protocols. Continued refinement and broader validation efforts are essential to enhance the reliability and scalability of this method for practical use.

Another important aspect to consider is how this approach compares with established diagnostic techniques such as MALDI-TOF. While MALDI-TOF is a gold standard in microbial identification due to its precise and reliable results, it requires sample preparation, specialized equipment, and trained personnel—factors that can limit its accessibility in certain healthcare environments. In contrast, the E-Nose/AI approach offers a rapid, non-invasive, and cost-effective alternative, potentially reducing diagnosis time and enabling earlier intervention. However, further studies are necessary to comprehensively establish its competitive advantages and to address any gaps in sensitivity and specificity compared to traditional methods [28].

In summary, considering the above, optimizing model selection involves balancing accuracy, computational efficiency, and clinical applicability. The analysis highlights that while traditional models benefit from specie balancing, time series-based models, such as Inception Time, exhibit greater robustness to data imbalances, making them particularly suitable for clinical practice where diagnostic speed and reliability are key. The following section will consolidate the study's key contributions, discussing their implications and outlining potential directions for future research and applications.

VI. Conclusions and Future Work

This study represents a significant advancement in the rapid detection of fungal infections directly from blood culture broth samples, eliminating the need for subculture and organism isolation. Compared to traditional subculture-based methods, the proposed approach substantially reduces the turnaround time to results, enabling faster therapeutic interventions and potentially saving lives, particularly among immunocompromised patients. The rigorous dual-scenario validation confirmed the reliability of the proposed solution: the high intra-sample consistency ($> 98\%$) validated the stability of the Framework, while the inter-sample stress test defined the current baseline for biological generalization.

The results underscore the clinical potential of this methodology, showing that direct blood culture broth analysis can overcome the limitations of conventional methods in terms of speed and efficiency. Among the classifiers, the traditional SVC model demonstrated superior stability in the generalization scenario, achieving statistical parity with complex Deep Learning architectures like InceptionTime. This finding is critical for point-of-care implementation, suggesting that computationally efficient models are sufficient for accurate screening without the need for high-end processing hardware. Furthermore, the oversampling strategy positively influenced model generalization, particularly for Deep Learning models, enhancing accuracy in imbalanced data conditions.

Despite these promising findings, some limitations should be acknowledged. The contrast between the near-perfect Framework stability and the reduced accuracy in the Grouped Cross-Validation scenario ($\approx 76\%$) isolates biological variability as the primary challenge. This confirms that the limitation lies in the dataset size ($N = 14$ samples) rather than the sensing capabilities. Addressing these limitations in future research through larger, multicenter datasets and standardized collection protocols will be essential for capturing the full spectrum of intravarietal metabolic profiles. The long-term stability of sensor performance and the potential need for recalibration over time also warrant further investigation.

For future directions, it is crucial to conduct multicenter clinical trials to validate the applicability of this approach across different hospital environments and patient populations. Prioritizing the collection of samples from a larger number of unique donors, rather than just increasing the number of cycles per sample, will be key to bridging the gap between Framework stability and biological generalization. Expanding VOC profiling to include new *Candida* species, such as *C. krusei*, *C. tropicalis*, and *C. auris*, along with increasing the dataset for previously studied species, will refine the method's precision and coverage. Additionally, incorporating samples from *non-Candida* microorganisms, such as bacteria and other fungal species, will enhance model generalization and reduce false positives.

Another important strategy is collecting samples containing multiple *Candida* species, allowing for the evaluation of rare mixed culture and improving the model's ability to identify mixed infections. Furthermore, integrating explainable artificial intelligence (XAI) techniques will provide greater transparency in the decision-making process, increasing healthcare professionals' trust in the technology and facilitating its adoption in clinical practice.

The findings of this study reinforce the potential of combining Electronic Noses and AI as an innovative solution for rapid diagnostics. The clinical implementation of this approach could significantly transform how fungal infections are diagnosed and treated, reducing associated mortality rates and optimizing hospital resources.

Acknowledgment

This work was supported in part by the National Council for Scientific and Technological Development (CNPq), and the Foundation for the Support of Science and Technology of the State of Pernambuco (FACEPE). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. A special thanks to everyone at the University of Cincinnati College of Medicine for having welcomed me so warmly and for their great contributions to this work. In particular, Dr. Margaret Powers-Fletcher and Christina Cox for everything they did during my exchange.

References

[1] Schroeder, M., Weber, T., Denker, T. *et al.*, "Epidemiology, clinical characteristics, and outcome of candidemia in critically ill patients in Germany: a single-center retrospective 10-year analysis". Ann. Intensive Care 10, 142, 2020. doi: 10.1186/s13613-020-00755-8

[2] Mohr, A., Simon, M., Joha, T. *et al.*, "Epidemiology of candidemia and impact of infectious disease consultation on survival and care". Infection 48, pp. 275–284, 2020, doi: 10.1007/s15010-020-01393-9

[3] Raja NS. "Epidemiology, risk factors, treatment and outcome of Candida bloodstream infections because of Candida albicans and Candida non-albicans in two district general hospitals in the United Kingdom", Int J Clin Pract, 75:e13655, 2020, doi: 10.1111/ijcp.13655

[4] Ismail, Wan Nor Ain Wan, *et al.*, "The economic burden of candidemia and invasive candidiasis: a systematic review". Value in health regional issues 21, pp. 53-58, 2020, doi: 10.1016/j.vhri.2019.07.002

[5] Peri, A.M., O'Callaghan, K., Rafiei, N. *et al.*, "Integrating omics techniques and culture-independent systems may improve the detection of persistent candidemia: data from an observational study", Ann Clin Microbiol Antimicrob 23, pp. 75 , 2024, doi: 10.1186/s12941-024-00736-w

[6] Kotey FC, Dayie NT, Tetteh-Uarcoo PB, Donkor ES, "Candida Bloodstream Infections: Changes in Epidemiology and Increase in Drug Resistance. Infectious Diseases: Research and Treatment". 2021;14. doi: 10.1177/11786337211026927

[7] Bastos, M.L., Benevides, C.A., Zanchettin, C. *et al.*, "Breaking barriers in Candida spp. detection with Electronic Noses and artificial intelligence". Sci Rep 14, pp. 956, 2024, doi: 10.1038/s41598-023-50332-9

[8] M. C. A. Castro *et al.*, "Breakthrough of Clinical Candida Cultures Identification Using the Analysis of Volatile Organic Compounds and Artificial Intelligence Methods", in IEEE Sensors Journal, vol. 22, no. 13, pp. 12493-12503, 1 July1, 2022, doi: 10.1109/JSEN.2022.3178346

[9] X. Yang, M. Li, X. Ji, J. Chang, Z. Deng, and G. Meng, "Recognition Algorithms in E-Nose: A Review," in *IEEE Sensors Journal*, vol. 23, no. 18, pp. 20460–20472, Sept. 2023, doi: 10.1109/JSEN.2023.3302868.

[10] S. A. Wulandari, R. Pramitasari, S. Madnasri, and Susilo, "Electronic Noses for Diabetes Mellitus Detection: A Review," in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia, 2020, pp. 364–369, doi: 10.1109/iSemantic50169.2020.9234304.

[11] T. Wang *et al.*, "Classification and Concentration Prediction of VOCs With High Accuracy Based on an Electronic Nose Using an ELM-ELM Integrated Algorithm," in *IEEE Sensors Journal*, vol. 22, no. 14, pp. 14458–14469, July 2022, doi: 10.1109/JSEN.2022.3176647.

[12] V. H. Tran *et al.*, "Breath Analysis of Lung Cancer Patients Using an Electronic Nose Detection System," in *IEEE Sensors Journal*, vol. 10, no. 9, pp. 1514–1518, Sept. 2010, doi: 10.1109/JSEN.2009.2038356.

[13] Q. Li, Y. Gu, and N.-f. Wang, "Application of Random Forest Classifier by Means of a QCM-Based E-Nose in the Identification of Chinese Liquor Flavors," in *IEEE Sensors Journal*, vol. 17, no. 6, pp. 1788–1794, Mar. 2017, doi: 10.1109/JSEN.2017.2657653.

[14] Mohamed, Ehab I., *et al.*, "Electronic nose versus VITEK 2 system for the rapid diagnosis of bloodstream infections", Brazilian Journal of Microbiology 54.4,pp. 2857-2865, 2023

[15] Peled, Nir, *et al.*, "Volatile fingerprints of cancer specific genetic mutations." Nanomedicine: Nanotechnology, Biology and Medicine 9.6, pp. 758-766, 2013.

[16] Leja M, *et al.*, "Sensing gastric cancer via point-of-care sensor breath analyzer." Cancer 127.8, pp. 1286-1292, 2021.

[17] de Heer, Koen, *et al.*, "Electronic nose technology for detection of invasive pulmonary aspergillosis in prolonged chemotherapy-induced neutropenia: a proof-of-principle study.", Journal of clinical microbiology 51.5, pp. 1490-1495, 2013.

[18] Mishra, Prabhaker, *et al.*, "Descriptive statistics and normality tests for statistical data.", Annals of cardiac anaesthesia 22.1, pp.67-72, 2019.

[19] Imam, A. Usman, M., Chiawa, M., "On Consistency and Limitation of paired t-test, Sign and Wilcoxon Sign Rank Test". IOSR Journal of Mathematics, VOl. 10, pp. 01-06, doi: 10. 01-06. 10.9790/5728-10140106, 2014.

[20] Proudfoot JA, Lin T, Wang B, Tu XM., "Tests for paired count outcomes", Gen Psychiatr. Sep 8;31(1):e100004, doi: 10.1136/gpsych-2018-100004, PMID: 30582120; PMCID: PMC6211281, 2018.

[21] R. Mohammed, J. Rawashdeh and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results". 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, pp. 243-248, doi: 10.1109/ICICS49469.2020.239556, 2020.

[22] Lee, Yen-Hsien, *et al.*. "Nearest-neighbor-based approach to time-series classification". Decision Support Systems 53.1, pp. 207-217. doi: 10.1016/j.dss.2011.12.014, 2012.

[23] Flynn, Michael, James Large, and Tony Bagnall. "The contract random interval spectral ensemble (c-RISE): The effect of contracting a classifier on accuracy", Hybrid Artificial Intelligent Systems: 14th International Conference, HAIS 2019, León, Spain, September 4–6, Proceedings 14. Springer International Publishing, 2019.

[24] Dempster, A., Petitjean, F. and Webb, G.I. "ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels". Data Min Knowl Disc 34, pp. 1454–1495, doi: 10.1007/s10618-020-00701-z, 2020.

[25] Deng, Houtao, *et al.*, "A time series forest for classification and feature extraction", Information Sciences 239, pp. 142-153, doi: 10.48550/arXiv.1302.2277, 2013.

[26] Wang, You, *et al.*, "An optimized deep convolutional neural network for dendrobium classification based on electronic nose". Sensors and Actuators A: Physical 307, 111874. doi: 10.1016/j.sna.2020.111874, 2020.

[27] D. Raymond, S. Chaudhri, and K. Shaikh, "Data Processing and Pattern Recognition for Electronic Nose," in *2025 International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India, 2025, pp. 463–468, doi: 10.1109/ICEARS64219.2025.10940421.

[28] Patel, Robin. "A moldy application of MALDI : MALDI-ToF mass spectrometry for fungal identification.", Journal of Fungi 5.1, 2019.