SIDTER: Prototype Early diagnosis system for respiratory diseases assisted by AI with human supervision in the process

Juan Sebastian Manquillo[1[0009−0007−8234−6898]], Sebastian Muñoz Carvajal[1[0009−0004−7653−1171]], Juan Diego Giraldo Muñoz[1[0009−0007−5715−4202]], Yeison Daniel Restrepo[1[0009−0009−9804−0353]], Robert Alejandro Beru[1[0009−0009−0419−1143]], Yilmar Mogollon[1[0009−0005−8972−9096]], Oscar Santiago López Erazo[1[0000−0002−8588−8365]], Juliana Maria López Erazo[1,3[0009−0001−2125−6553]], Juliana Delle Ville[2[0009−0007−7888−7544]], Luis Freddy Muñoz[1[0000−0002−8172−0530]], Leandro Antonelli[2,4[0000−0003−1388−0337]], and Cesar Collazos[3[0000−0002−7099−8131]]

[1] Fundación Universitaria de Popayán, Colombia {juan.manquillo, sebastian.munoz, juandiego.giraldo, yeison.restrepo, robert.beru, yilmar.mogollon}@estudiante.fup.edu.co {santiago.lopez,freddy.munoz}@docente.fup.edu.co {shulejuliana}@gmail.com

[2] Universidad Nacional de La Plata, Argentina {jdelleville,gmaltempo,lanto@lifia.info.unlp.edu.ar}

[3] Universidad del Cauca, Colombia {ccollazo, yulianaml}@unicauca.edu.co

[4] CAETI Centro de Altos Estudios en Tecnología Informática, Argentina

Abstract. Human health is a fundamental pillar of individual and collective well-being, as it determines people's ability to reach their potential, contribute to social progress and carry out daily activities. According to the World Health Organization (WHO), health implies not only the absence of disease, but also a complete state of physical, mental and social well-being. Respiratory diseases such as influenza, the common cold, COPD, asthma, pneumonia and allergic rhinitis significantly impact human health by compromising respiratory function, generating acute symptoms and causing chronic complications. These conditions reduce physical capacity, impair quality of life and generate socioeconomic burdens. Early and accurate diagnosis is essential to mitigate their impact, as diseases such as COPD affect more than 200 million people worldwide. However, challenges such as limited access to medical care, unspecified symptoms, continuous exposure to risk factors and delays in referral to specialized centers still persist. In this context, artificial intelligence (AI) presents itself as a key ally for early diagnosis, improving clinical accuracy and optimizing time-consuming tasks. In response to these challenges, SIDTER is presented: a prototype AI-assisted early diagnosis system for respiratory diseases with medical supervision and validation. It aims to support physicians, improve clinical diagnostic capabilities and strengthen patient-physician interaction.

Keywords: Respiratory Diseases · AI · Random Forest · Diagnosis · Human-Patient Interaction.

1 Introduction

Human health is a pillar of individual and group well-being, as it determines people's ability to reach their potential, contribute to social progress, and carry out daily activities. Human health encompasses not only the absence of disease, but also a complete state of mental, social, and physical well-being, as defined by the World Health Organization (WHO) [1]. On the other hand, maintaining good health strengthens resilience in the face of emotional and/or environmental challenges, improves quality of life, and reduces the burden of chronic diseases. For the latter, it is important to consider factors such as a balanced diet, access to medical services, and physical activity [2].

Respiratory diseases such as influenza, the common cold, COPD, asthma, pneumonia, and allergic rhinitis have a significant impact on human health because they compromise the function of the respiratory system, causing acute symptoms as well as chronic complications [Duran]. Respiratory diseases affect human well-being because they limit physical capacity, impair quality of life, and generate socioeconomic burdens. Furthermore, failure to treat them in a timely manner can lead to serious complications and compromise daily autonomy [3].

According to the above, accurate diagnosis of respiratory diseases is important because it is essential to mitigate their impact on global health [4]. Conditions such as asthma, COPD, pneumonia, and influenza

are among the leading causes of morbidity and mortality worldwide. Diseases such as COPD affect 200 million people worldwide, according to the WHO [5]. Timely and accurate diagnosis can help initiate early treatment, improve quality of life, and reduce complications. However, some challenges remain and will be described below: (i) limited access to health services, (ii) nonspecific symptoms of diseases, (iii) continued exposure to risk factors, and (iv) delays in referral to specialized centers. Doctors are frequently confronted with respiratory diseases, which is why it is important to develop AI-supported tools to assist them in their work. These types of diseases are common in Colombia and occur in two peaks per year, which can leave hospitals without resources. In order to make triage more effective, a tool is needed to rule out hospitalization or outpatient care [4], [6], [7].

Artificial intelligence (AI) can be a key ally in the early diagnosis of diseases, as it could help streamline this process by acting as an efficient support for doctors with the aim of enhancing clinical accuracy and optimizing time-consuming tasks. This means that AI not only speeds up the process, but also frees up doctors to focus on communicating with patients, physical evaluation, and making fundamental decisions. Therefore, AI acts as a support by processing complex information while medical professionals maintain their role of interpreting results, contextualizing findings, and defining treatments. In short, this synergy improves diagnostic capacity in high-demand healthcare settings while ensuring that human judgment is not replaced [8].

Therefore, we present SIDTER: Prototype Early Diagnosis System for Respiratory Diseases assisted by AI with human supervision in the process, with the aim of supporting physicians by contributing to improving diagnostic capacity for respiratory diseases and enhancing physician-patient interaction. This study is divided as follows: State of the Art, Results, Discussion, Conclusions, and References.

2 Related Work

The development of artificial intelligence (AI) and machine learning (ML) models has significantly advanced research into the diagnosis and prediction of respiratory diseases. Different approaches have been proposed in the literature, highlighting the use of audio biomarkers, clinical data, and advanced classification techniques. Kapetanidis et al. (2024) conducted a systematic review on the use of audio analysis and AI in the diagnosis of respiratory diseases. They evaluated 75 studies that use convolutional neural networks (CNN) and support vector machines (SVM) for cough detection, respiratory symptoms, and voice analysis. They highlighted advances in COVID-19 detection with 99% accuracy, although they identified challenges in model generalization and integration into clinical systems [9]. Becerra Yoma and Mendoza Inzunza (2024) proposed a system for assessing dyspnea through telephone voice analysis, using CNN and LSTM. They obtained a classification error of 0.94 points with respect to the mMRC scale, with false positive and false negative rates of 11% and 5%, respectively. They pointed out the importance of strengthening the model with additional data and its application in public health [10]. Bhattacharya et al. (2024) introduced the Coswara dataset, collecting 65 hours of recordings of respiratory sounds and symptoms for remote screening of COVID-19. They trained a BLSTM classifier with an AUC of 91.5%, identifying challenges in device variability and recording environments [11]. Kumar et al. (2024) implemented Random Forest and explainable artificial intelligence (XAI) methods such as SHAP and LIME in predicting survival rates in pediatric respiratory diseases. They achieved 96% accuracy, highlighting the importance of interpretability in clinical models. They indicated the need to expand the database and consider deep learning to improve generalization [12]. Ochieng (2024) designed a differential diagnosis tool for chronic obstructive pulmonary diseases (COPD, asthma, and ACO), using spirometry data and smoking history. He applied KNN, SVM, and Random Forest, achieving an accuracy of 93.94%. However, he highlighted the diagnostic confusion between pathologies and the lack of validation in clinical settings [13]. Feng et al. (2024) reviewed the application of AI in asthma and COPD, highlighting the use of neural networks, logistic regression, and latent class analysis. They underscored the potential of these models in phenotype classification and exacerbation prediction, but pointed out the need for external validation and greater inclusion of genetic and clinical data [14]. Jackins et al. (2024) compared Random Forest and Naive Bayes in the prediction of clinical diseases such as diabetes and cancer, demonstrating that Random Forest achieved an accuracy of 92.40%. They highlighted the need to optimize processing time and explore real-time models [15]. Ramalingam and Chinnaiyan (2024) conducted a comparative analysis of ML and DL in COPD classification. They used CNN, SVM, and KNN on computed tomography (CT) images,

achieving accuracies above 90%. They identified limitations in clinical interpretability and the need for more diverse datasets [16]. Anupriya and Thangavelu (2024) developed LUCAGO, a lung cancer detection system that combines the Grey Wolf Optimization (GWO) algorithm with Random Forest. They achieved an accuracy of 92.62%, highlighting the efficiency of the hybrid method. However, they recommended future improvements in adaptive models for different age groups [17]. Yenurkar et al. (2024) used XGBoost and Random Forest to predict the severity of COVID-19, achieving 98% accuracy. They designed a web-based tool for clinical decision-making, although they noted limitations in considering dynamic factors such as viral variants and treatments. These studies demonstrate the potential of machine learning in the diagnosis of respiratory diseases, highlighting the need to improve interpretability, integration into clinical settings, and robustness of models through the use of multimodal data [18].

Unlike the reviewed state-of-the-art works, which mostly focus on a single pathology (e.g., COVID-19, asthma, or COPD) and a specific type of data (audio, images, or spirometry), the SIDTER prototype integrates a comprehensive approach that encompasses six common respiratory diseases (influenza, common cold, COPD, asthma, pneumonia, and allergic rhinitis), allowing for broader and more useful differential diagnoses in clinical practice. Furthermore, it combines a machine learning model (Random Forest) with advanced preprocessing techniques such as SMOTE for data balancing, normalization, and variable encoding, overcoming the problems of imbalance and overfitting noted in previous studies. A notable strength of this work is the integration of a robust development methodology, including the Scrum framework and user-centered design (UCD), along with the use of Mini QAW for the collection and prioritization of key non-functional requirements such as reliability and efficiency. The implementation of the Random Forest model is a sound decision, as this algorithm stands out for its ability to handle large volumes of data quickly and accurately. Furthermore, the attention paid to data manipulation and preparation, such as the application of SMOTE for class balancing, demonstrates a deep understanding of the challenges of machine learning in medical settings. Another distinguishing feature is the prototype's modular architecture, which facilitates maintenance, scalability, and code reuse, as well as the incorporation of a graphical interface developed under user-centered design principles. The interface was evaluated using the SUS scale by a physician and a systems engineer, obtaining scores of 97.5 and 92.5, respectively. Finally, the proposal includes dual technical validation, using performance metrics, and clinical validation, through analysis of real-life scenarios by a healthcare expert, ensuring not only algorithmic accuracy but also the medical consistency of the predictions, an aspect rarely considered in the analyzed research.

3 Methods

The prototype was developed using the Scrum framework (Highsmith and Cockburn)[19]. This methodology allowed us to define the roles assumed by each team member. A set of four sprints was established for the development of the prototypes, with a duration of two weeks each. This allowed us to define the requirements of the current prototype, which are necessary inputs for the construction of the system architecture.

In the article by De Gooijer [20], it is found that good quality requirements help to make the right architectural decisions, but gathering this type of requirement is not always easy. The Quality Attribute Workshop (QAW) allows requirements to be gathered effectively, but its organization can be cumbersome and/or costly. For the design of the prototype architecture, it was decided to use Mini QAW, as this version of the workshop is more agile and designed to be easily understood by novice collaborators, making it an excellent option for teams practicing agile methods. Mini QAW is widely used around the world and is establishing itself as a standard tool for software architects.

As is well known, it is important to obtain a good dataset for the development of AI models because it determines the quality, generalization of results, and accuracy. For this reason, the strategy outlined in the article by Goyal and Mahmoud is used, which mentions the exploration of the use of synthetic data to improve the training of AI models in the field of health. In the study, they use techniques such as generative adversarial networks (GANs) and variational autoencoders (VAEs) that manage to generate datasets that accurately mimic the patterns found in respiratory diseases. They also discuss the Synthetic Minority Oversampling Technique (SMOTE), which is used as a technique to address data imbalance by generating new synthetic samples in the minority class, helping to improve model performance. This approach allows models

to be trained without relying exclusively on real clinical data, improving model generalization and preserving patient privacy. The challenge lies in ensuring that the synthetic data is sufficiently diverse. In addition, it must be representative and avoid bias in order to optimize its application in clinical settings [21].

For the prototype design, the user-centered design (UCD) methodology [22] was used, which is defined as an approach to products and systems in which user expectations, needs, and limitations are prioritized in the design phases. The model follows the iterative process scheme of User-Centered Design, which can be seen below:

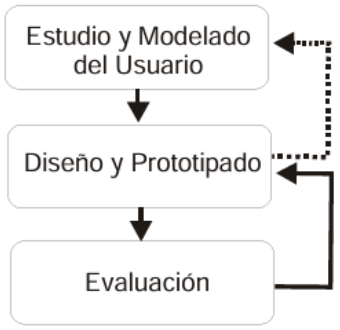

Fig. 1. Iterative process: User study and modeling, design and prototyping, and evaluation.

The model defines the following stages:

- **User Study and Modeling**: User needs and behaviors are analyzed, and profiles and models based on the research are created.
- **Design and Prototyping**: Ideas are generated, initial representations are created, and prototypes are generated to visualize solutions.
- **Evaluation**: Prototypes are tested with real users, and data is collected to iterate and improve the design.

The evaluation will use the System Usability Scale (SUS)[23], which is a standardized method for evaluating the usability of a digital system, product, or service through a 10-question survey with answers on a scale of 1 to 5.

It is calculated by adding the contributions of each item, adjusting them according to their type (positive or negative), and multiplying the result by 2.5 to obtain a score between 0 and 100.

A high SUS (85-100) indicates excellent usability, while a low SUS ($<$50) suggests significant difficulties for users. It is a fast, reliable, and widely used tool in user experience (UX) evaluation.

3.1 Functional requirements

Predicting diseases using artificial intelligence: As a user, I want to record my symptoms in the system in order to obtain a prediction of the respiratory disease I may have, with a diagnosis based on artificial intelligence.

- The first step in user modeling is to identify and define the system's audience.
- As a user of the system, I want the system to use the Random Forest model to predict a possible disease, in order to receive quick guidance on my health status.
- As a developer of the system, I want to ensure the validity of the data provided by the user, in order to prevent prediction failures caused by erroneous or out-of-range information.

Displaying results in the graphical interface: As a system user, I want the graphical interface to show me the prediction results in a clear and understandable way.

4As a system user, I want the graphical interface to show me the prediction results in a clear way, so that I can easily understand the possible disease detected. **Data loading and preprocessing:** As a system developer, I want the system to extract information from an Excel file and prepare it for the model to ensure that the data is organized before training.

- As a system developer, I want the system to extract information from an Excel file and prepare it for the model to ensure that the data is properly organized before training.
- As the system developer, I want the categorical variables to be encoded with LabelEncoder so that the classification model can process them properly.
- As the system developer, I want the symptoms entered by the user to be converted into a format suitable for the model so that the prediction is accurate.

Model splitting and training: As a system developer, I want to split the data into training and test sets to evaluate the model's performance before implementation.

- As a system developer, I want the data to be split into training and test sets using train_test_split, to effectively evaluate the model's performance.
- As the system developer, I want to train a classification model using RandomForestClassifier to make accurate disease predictions.
- As the system developer, I want to evaluate the model's accuracy using accuracy_score to measure its performance and improve it if necessary.

3.2 First Step: Selecting non-functional requirements based on user stories

Efficiency. The system must be fast and efficient because users expect immediate responses when they enter their symptoms. Using a Random Forest model helps make predictions agile without compromising accuracy. In addition, the code has been optimized to process data without wasting resources, making the application run smoothly and without annoying delays. **Reliability.** Since this system suggests possible illnesses based on the user's symptoms, it is crucial that its predictions are as reliable as possible. To this end, the model has been trained with a large amount of data and its performance has been validated by dividing the data into training and testing sets. This ensures that the system not only memorizes the data, but actually learns to make good predictions in different cases. **Security.** This system handles users' personal information, such as age, symptoms, and health status. Protecting this data is essential. Currently, basic measures have been taken, such as validating the information entered to avoid errors, but in the future, if the system is implemented on a real platform, it will be key to include encryption, authentication, and security protocols to protect user privacy. **Usability.** Not all users are experienced with technology or artificial intelligence, so the system must be easy to use. It was designed so that anyone can enter their symptoms clearly and simply, with direct questions and easy-to-understand answers.

3.3 Step two: Prioritizing non-functional requirements.

In an artificial intelligence system applied to healthcare, reliability and efficiency are essential, as they directly impact user confidence and experience.

Reliability is key because the system must provide consistent and accurate predictions. In a medical context, an error or inaccurate prediction could cause confusion or even affect decisions about a person's health. To ensure reliability, the model has been trained with a robust dataset, and strategies such as model validation and proper exception handling have been implemented.

On the other hand, efficiency is crucial for the system to function in real time. Users need quick responses when entering their symptoms, without long wait times. To achieve this, a Random Forest algorithm has been used, which is capable of processing large volumes of data quickly without compromising accuracy. In

addition, code optimization and the use of appropriate data structures ensure that calculations are performed quickly and without wasting resources.

Within the MiniQAW framework, these two quality attributes have been selected as drivers, i.e., key factors that influence the system's architectural decisions. Although usability and security are also important, without a reliable and efficient system, the tool would lose its purpose and be of no use to users.

3.4 Step Three: Constraints between non-functional requirements

Reliability - Performance efficiency

- **Reliability**: The model must provide consistent and reproducible predictions with a low margin of error.
- **Performance efficiency**: The aim is to minimize execution time and consumption of computational resources.
- **Constraint**: Increasing reliability through stricter cross-validations or data augmentation with SMOTE can increase training time and memory demand.

Reliability - Usability

- **Reliability**: Ensure that results are accurate and reproducible.
- **Usability**: The interface must be intuitive for users without technical knowledge.
- **Constraint**: Explaining results in detail or adding additional checks to improve reliability can increase the number of interactions required, making the user experience slower or less intuitive.

Reliability - Security

- **Reliability**: Ensure that the model delivers stable and consistent predictions.
- **Security**: Protect the user's personal and medical data.
- **Constraint**: Applying additional security mechanisms, such as user authentication, can generate additional processes that impact the availability of the system in real time.

Performance Efficiency - Security

- **Performance Efficiency**: Minimize processing time to improve the user experience.
- **Security**: Ensure that user data is processed securely.
- **Restriction**:Implementing data encryption can slow down preprocessing and prediction, impacting the speed of the system.

Performance Efficiency - Usability

- **Performance Efficiency**:Maintain low response times when processing user input.
- **Usability**:Allow the user to enter data easily without long wait times.
- **Restriction**:Reducing model complexity to improve speed can decrease prediction accuracy, affecting user confidence in the results.

Security – Usability

- **Security**:Prevent the system from being vulnerable to input data manipulation or attacks.
- **Usability**:Allow the user to enter data easily without long wait times.
- **Restriction**:Implementing stricter validation measures (such as data entry restrictions) can cause the user to have to repeat actions if they enter incorrect information, reducing the user experience.

3.5 Step 4: Non-functional requirements scenarios

Usability: **Scenario 1: Given:** A user accesses the diagnostic system. **When:** They try to enter their symptoms in the prediction form. **Then:** The system should display an intuitive form with clear questions and easy-to-select options. **Scenario 2: Given:** A doctor wants to review a previous diagnosis. **When:** They enter the patient's ID in the search bar. **Then:** The system should display the patient's information and diagnostic history in an organized and quick manner. **Scenario 3: Given:** A user receives the prediction result. **When:** They want to better understand their diagnosis. **Then:** The system should display a clear and accessible description of the predicted disease. **Performance Efficiency: Scenario 4: Given:** A user enters their symptoms into the system. **When:** The AI model must process the data and generate a diagnosis. **Then:** The system must generate a result quickly and efficiently. **Scenario 5: Given:** A doctor reviews the diagnoses generated by the AI. **When:** An error is detected in the data. **Then:** The system must notify the administrator for review. **Reliability: Scenario 6: Given** A user accesses the interface to obtain a diagnosis. **When:** The system experiences a prediction failure. **Then:** The system must send a notification to the technical team for prompt resolution of the problem. **Security: Scenario 7: Given:** a user accesses the interface to obtain a diagnosis. **When:** the system experiences a prediction failure. **Then:** the system must send a notification to the technical team for prompt resolution of the problem.

4 System architecture

The architecture of the respiratory disease prediction system based on symptoms and patient characteristics is organized into modules, each with a specific function. This modular configuration simplifies maintenance, scalability, and code reuse, enabling efficient integration between data processing, model training, and the user interface.

– **Module for Loading Information (`data_loader.py`)**

- This module is responsible for:
  - ∗ The dataset used contains 500 records and 10 attributes that describe clinical and demographic information about patients with various respiratory diseases. Variables include the type of disease diagnosed, relevant symptoms (fever, cough, fatigue, shortness of breath), and patient data (age, gender, blood pressure, and cholesterol level). It also includes an outcome variable that indicates whether the case is positive or negative for the condition studied. A small fraction of the data was collected from public health management report files (General Data), serving as a real-world basis for modeling. The remaining records were synthetically generated using the SMOTE (Synthetic Minority Over-sampling Technique) technique, with the goal of balancing classes and expanding the dataset without losing coherence in clinical patterns. This approach provides a balanced dataset that combines real and synthetic data to train and evaluate machine learning models, such as Random Forest, for the prediction and classification of respiratory diseases.
  - ∗ Loading the dataset in Excel format (xlsx).
  - ∗ Reading the specific data sheet, ensuring that all necessary columns are present and available.
  - ∗ Checking for null values in the data and, if found, deleting or correcting them as appropriate.
  - ∗ Finally, the system loads a clean `DataFrame` ready for processing.
- **Input:** Excel file path.
- **Output:** `DataFrame` with raw data.

– **Data Preprocessing Module (`preprocessing.py`)**

- In this module, different conversions are performed to convert the data into a format suitable for the model:
  - ∗ Coding of categorical variables: converts symptoms and characteristics (Fever, Cough, Fatigue, Shortness of breath, Gender) into numerical values using `LabelEncoder`.

    - Conversion of ordinal values: for the variables "Blood Pressure" and "Cholesterol", values are mapped to 0 for (Low), 1 for (Normal), and 2 for (High).
    - Class balancing: in case the dataset contains more samples of certain diseases than others, `SMOTE` is used for synthetic data generation in order to prevent the model from favoring the majority classes.
    - Numerical data normalization: by using `StandardScaler`, it is possible to ensure that variables such as age are on a comparable scale.
  - **Input:** `DataFrame` with raw data.
  - **Output:** `X` (features) and `y` (disease labels) ready for training.
- **Model Training Module (`model_training.py`)**
  - This module is in charge of:
    - Splitting the data into training (80%) and testing (20%).
    - Training a machine learning model using `RandomForestClassifier`.
    - Optimizing hyperparameters (such as number of trees and model depth) to improve model performance.
    - Evaluating the model using metrics such as accuracy, recall, and F1-score to ensure that it predicts adequately.
  - **Input:** `X_train`, `y_train`.
  - **Output:** A trained model ready to make predictions.
- **Prediction Module (`predictor.py`)**
  - At this point, the information entered by the user is processed to make a prediction:
    - The application receives the user's input data (symptoms, age, gender, etc.).
    - Converts the input values to the same numerical representations that were used in training.
    - Performs the prediction with the previously trained model.
    - Finally, it converts the prediction to a readable diagnosis, returning the name of the disease instead of a number.
  - **Input:** Data entered by the user.
  - **Output:** Name of the predicted disease.
- **Graphical Interface Module (`gui.py`)**
  - This module is in charge of managing the visual section of the application:
    - It initially creates the main window with `CustomTkinter`.
    - It defines the UI elements: buttons, text entries, drop-down menus, etc.
    - It captures the values entered by the user and directs them to the prediction module.
    - Finally, it displays the result in a clear and easy-to-understand way.
  - **Input:** User interaction with the graphical interface.
  - **Output:** Diagnostics displayed on the screen.
- **Main File (`main.py`)**
  - This file orchestrates the whole flow of the application and serves as the entry point when executing the application:

* It loads the data.
* Preprocesses the dataset.
* Trains the model (or loads a previously trained one).
* Starts the graphical interface to receive data from the user and make predictions.

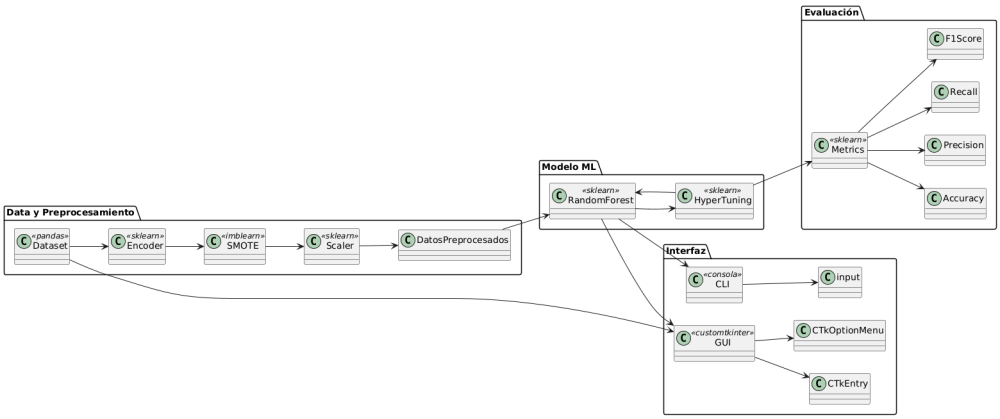

Fig. 2. System Architecture

5 Results

SIDTER: Prototype AI-assisted early diagnosis system for respiratory diseases with human supervision in the process.

Functioning of the application

The respiratory disease prediction system uses machine learning techniques to analyze symptoms and patient characteristics, assisting physicians in the diagnosis.

– **Import of the Libraries**

- Different Python libraries are used for data processing, model training, and the creation of the graphical interface:
  * **pandas** and **numpy**: for data manipulation and numerical calculations.
  * **scikit-learn**: for data preprocessing, training, and model evaluation.
  * **imblearn**: for balancing data with SMOTE.
  * **CustomTkinter**: for the creation of the graphical interface.

– **Data Loading and Preprocessing**

- An Excel file is loaded with respiratory disease data (symptoms and categorical variables).
- The categorical variables are manipulated to convert them into numerical values using `LabelEncoder`.
- Numerical values are assigned to the variables *Blood Pressure* and *Cholesterol Level*.
- The target variable *Disease* is coded.

– **Data Balancing and Normalization**

- SMOTE is used to balance unbalanced classes in the dataset.

- The numerical variable *Age* is normalized with `StandardScaler`.

– **Dataset Partitioning**

- The dataset is separated into training (80%) and test (20%) sets using `train_test_split`.

– **Model Training**

- The **Random Forest Classifier** algorithm is used with previously defined hyperparameters.
- The model is adjusted to work with the training dataset.

– **Disease Prediction**

- The user's symptoms and characteristics are captured through the graphical interface.
- A transformation is performed on the user's inputs to create a `DataFrame` compatible with the model.
- The model prediction is obtained, and the result is displayed on the interface.

– **Graphical Interface (CustomTkinter)**

- An interface with a modern and accessible design is configured.
- Options are added to enter symptoms and characteristics such as age, gender, blood pressure, and cholesterol level.
- A button that executes the disease prediction is implemented.
- Finally, the result of the prediction is displayed on the screen.

5.1 Scenario 1

The patient is a 64-year-old man who presents with fever, cough, fatigue, and shortness of breath. His blood pressure and cholesterol levels are within normal ranges. The combination of acute respiratory symptoms along with fever suggests an infectious process. Considering the patient's age and lack of history of asthma, the most likely diagnosis is pneumonia, a pathology that frequently manifests with this clinical picture in older adults. (See Figure 3)

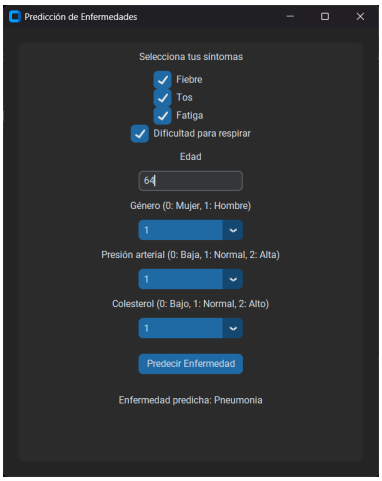

Fig. 3. Scenario 1

5.2 Scenario 2

This is a 31-year-old woman with fever, cough and respiratory distress, but without fatigue. Her vital signs (blood pressure and cholesterol) are normal. Unlike pneumonia, asthma does not usually present with fever; however, in this case, the presence of fever could indicate a concurrent infection. Nevertheless, the absence of fatigue and young age suggest a diagnosis of exacerbated asthma, possibly triggered by an infectious or allergic factor. (See Figure 4)

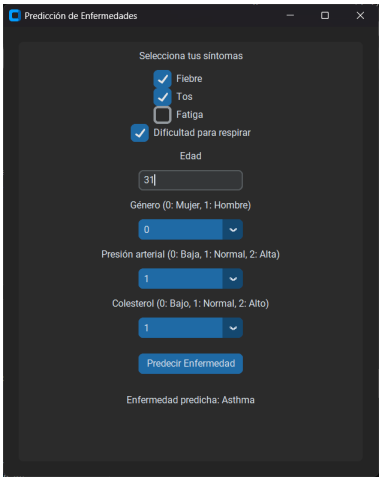

Fig. 4. Scenario 2

5.3 Scenario 3

A 67-year-old woman presents with shortness of breath and fatigue, but no fever or cough. Her blood pressure and cholesterol parameters are normal. Advanced age and the presence of chronic symptoms such as fatigue and dyspnea, without signs of infection, are indicative of chronic obstructive pulmonary disease (COPD). This diagnosis is common in older patients, especially with a history of smoking or exposure to lung irritants. (See Figure 5)

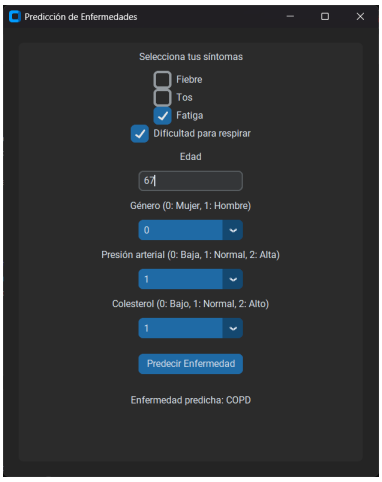

Fig. 5. Scenario 3

5.4 Scenario 4

A 26-year-old man has shortness of breath, but no fever, cough or fatigue. All his clinical parameters (blood pressure, cholesterol) are within normal. The absence of systemic symptoms and the patient's youth point to a non-infectious etiology. The most plausible diagnosis is allergic rhinitis, a condition that can cause nasal obstruction and dyspnea, especially in the context of allergen exposure. (See Figure 6)

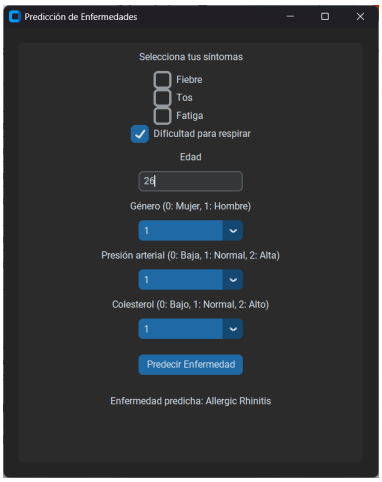

Fig. 6. Scenario 4

5.5 Scenario 5

A 26-year-old man presents with cough and fatigue, but no fever or respiratory distress. His blood pressure and cholesterol are normal. The combination of cough and fatigue, in the absence of high fever, suggests a mild viral infection. The most appropriate diagnosis is influenza, which can manifest with respiratory symptoms and malaise without serious complications in young, healthy patients. (See Figure 7)

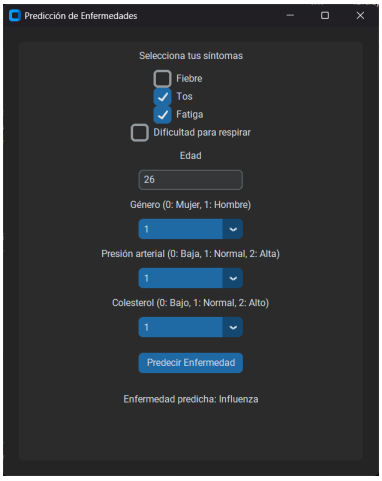

Fig. 7. Scenario 5

5.6 Scenario 6

A 12-year-old boy has a cough but no fever, fatigue or respiratory distress. All his clinical parameters are normal. The isolated cough, along with the absence of systemic symptoms, is characteristic of a common cold, a mild viral infection common in the pediatric population. This condition usually resolves spontaneously without requiring aggressive pharmacological intervention. (See Figure 8)

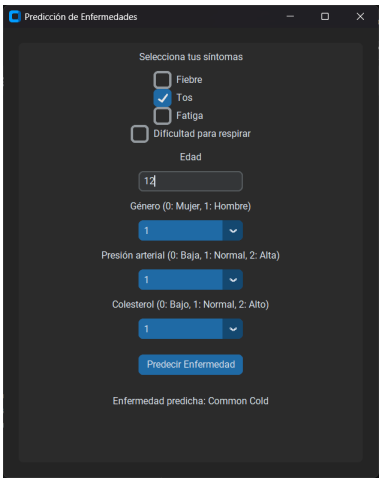

Fig. 8. Scenario 6

5.7 Confusion Matrix

The confusion matrix is a tool that allows the performance of a classification model to be evaluated by showing, for each real class, how many times it was correctly identified (on-diagonal values) and how many times it was confused with other classes (off-diagonal values). In the present study, the model developed with Random Forest Classifier presents acceptable performance, correctly identifying several diseases, although with room for improvement in those with similar symptoms. This is reflected in the obtained metrics: an accuracy of 0.5868 (more than half of the predictions are correct), a precision of 0.6767 (good ability to avoid false positives), a recall of 0.5868 (moderate detection of positive cases), and an F1-score of 0.5629 (acceptable balance between precision and recall). While the model is functional and useful as an initial diagnostic aid, it clearly has room for improvement through hyperparameter optimization, dataset enrichment, and more robust preprocessing techniques, which would increase its discriminatory capacity and reduce errors between diseases with similar clinical characteristics (See Figure 9).

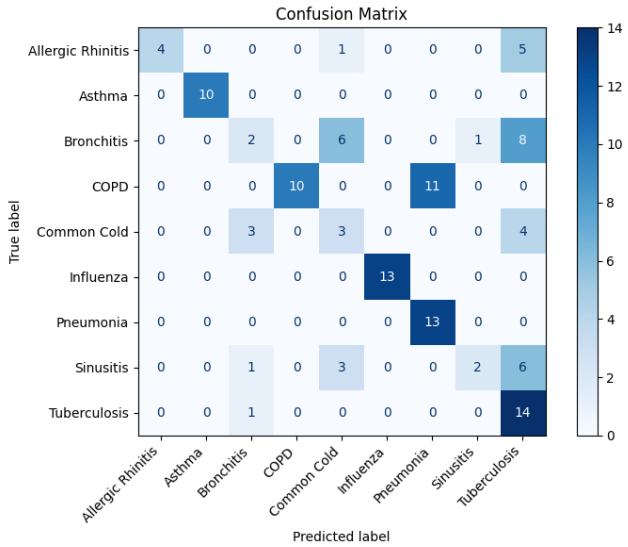

Fig. 9. Confusion Matrix

6 Usability evaluation by experts

6.1 SUS application by a doctor

The SUS questionnaire was also administered by a physician with six years of clinical experience, specializing in the diagnosis and treatment of respiratory diseases. In addition to evaluating the prototype's usability, the physician analyzed the results obtained by the Random Forest Classifier model, with the goal of providing a well-founded medical opinion on the system's accuracy and usefulness in clinical practice. His in-depth knowledge of medicine and direct contact with patients allowed him to assess whether the model's predictions were consistent with patterns observed in reality, as well as to identify potential limitations in its application. This dual participation in both the evaluation of the interface and the clinical validation of results ensures that the assessment obtained reflects not only technical criteria but also its applicability and relevance in the real-life medical context. (See Table 1)

Questions: Rating	**1 (No)**	**2**	**3**	**4**	**5 (Yes)**	**Count**
1. I think I would like to use this system frequently.				X		4
2. I found the system unnecessarily complex.					X	4
3. I thought the system was easy to use.					X	4
4. I think that I would need the support of a technical person to use this system.					X	4
5. I found the various functions in this system were well integrated.					X	4
6. I thought there was too much inconsistency in this system.				X		4
7. I imagine that most people would learn to use this system very quickly.					X	4
8. I found the system very cumbersome to use.				X		4
9. I felt very confident using the system.					X	4
10. I needed to learn a lot of things before I could get going with this system.					X	4
Total						39
SUS Score						97.5

Table 1. (SUS) Evaluation: Doctor

6.2 SUS application by a Systems Engineer

The SUS questionnaire was administered by a systems engineer with over 10 years of experience in software design and development, both in academia and industry. His specialization in systems architecture and artificial intelligence gives him a comprehensive perspective, combining technical knowledge, experience in user-system interaction, and quality criteria in application development. This background allows him to rigorously and objectively evaluate the prototype's usability, interpreting the results not only from the end-user perspective but also considering design, scalability, and optimization aspects, thus ensuring that the analysis obtained is reliable and aligned with industry best practices. (See Table 2)

Questions: Rating	**1 (No)**	**2**	**3**	**4**	**5 (Yes)**	**Count**
1. I think I would like to use this system frequently.				X		4
2. I found the system unnecessarily complex.				X		4
3. I thought the system was easy to use.					X	4
4. I think I would need the support of a technical person to use this system.				X		3
5. I found the various functions in this system were well integrated.				X		3
6. I thought there was too much inconsistency in this system.			X			3
7. I imagine that most people would learn to use this system very quickly.				X		4
8. I found the system very cumbersome to use.				X		4
9. I felt very confident using the system.					X	4
10. I needed to learn a lot of things before I could get going with this system.					X	4
Total						37
SUS Score						92.5

Table 2. (SUS) Evaluation: Systems Engineer

The average score resulting from applying the SUS to two users (a Physician and a Systems Engineer) is 95, which indicates that it is an intuitive, efficient, and easy-to-use application. However, experts suggested improvements such as: (i) For the gender, blood pressure, and cholesterol fields, instead of handling numeric data, the value associated with the number must be entered in a character string; and (ii) it is necessary to validate the age within a range by adjusting the parameters so that only valid responses can be entered according to the range from 3 to 100 years.

Having a systems engineer and a physician involved in the implementation of the SUS questionnaire is essential because it allows the prototype to be evaluated from two complementary and critical perspectives for its success: the technical and the clinical. The engineer contributes his or her expertise in software architecture, usability, and good development practices, ensuring that the system is robust, efficient, and easy to use. The physician, for his or her part, validates that the tool meets the real needs of the healthcare environment, interpreting the results and interaction flows according to their relevance to clinical practice and patient safety. This combination ensures a comprehensive evaluation, balancing technical quality and medical relevance, and increasing the reliability of the usability analysis obtained.

7 Evaluation of results by an expert (Physician)

7.1 Scenario 1: Pneumonia

Patient: Male, 64 years old Symptoms: Fever, cough, fatigue, shortness of breath Risk Factors: Advanced age Clinical Analysis: Pneumonia is an infection of the lung parenchyma characterized by inflammation of the alveoli and accumulation of inflammatory exudate, compromising gas exchange. In this 64-year-old patient, the combination of fever, cough, fatigue and dyspnea is highly suggestive of a pulmonary infection, possibly

of bacterial or viral origin. The fact that blood pressure and cholesterol levels are normal does not rule out the presence of the disease, as these factors are not determinant in the etiology of pneumonia. However, advanced age is a relevant risk factor, as the immune system tends to respond less efficiently with aging. The differential diagnosis would include acute bronchitis, congestive heart failure with pulmonary edema and exacerbation of chronic obstructive pulmonary disease (COPD), if there is a history of smoking.

7.2 Scenario 2: Asthma

Patient: Female, 31 years old Symptoms: Fever, cough, shortness of breath Risk Factors: None evident Clinical Analysis: Asthma is a chronic inflammatory disease of the airways characterized by bronchial hyperreactivity, resulting in episodes of bronchospasm and reversible airflow obstruction. In this case, the patient presents with cough and dyspnea, characteristic symptoms of asthma, but the presence of fever is atypical. Asthma does not usually generate fever unless there is an underlying respiratory infection, such as an associated viral infection or pneumonia. This raises the possibility that the fever is a sign of an infectious trigger, rather than asthma per se. Since the patient has no history of hypertension or high cholesterol, there are no cardiovascular factors predisposing to other chronic pulmonary pathology. However, the differential diagnosis would include acute bronchitis, atypical pneumonia and even allergic rhinitis with bronchial involvement.

7.3 Scenario 3: Chronic Obstructive Pulmonary Disease

Patient: Female, 67 years old Symptoms: Fatigue, shortness of breath Risk Factors: Advanced age Clinical Analysis: COPD is a pathology characterized by persistent airflow obstruction due to chronic lung damage, usually as a consequence of prolonged exposure to irritants, mainly smoking. Fatigue and dyspnea in a 67-year-old female patient suggest possible chronic respiratory failure, which is a common manifestation of advanced COPD. The absence of fever and cough suggests no active respiratory infection, which helps to differentiate it from diseases such as pneumonia.

7.4 Scenario 4: Allergic Rhinitis

Patient: Male, 26 years old Symptoms: shortness of breath Risk Factors: None evident Clinical Analysis: Allergic rhinitis is an inflammatory disease of the nasal mucosa caused by a hypersensitivity response to environmental allergens, such as dust, pollen or dust mites. It can manifest with nasal congestion, sneezing, rhinorrhea and, in some cases, shortness of breath. In this case, dyspnea may be related to allergic rhinitis if nasal congestion is severe enough to generate a sensation of upper airway obstruction. In some patients, the inflammation may extend to the pharynx or even trigger a bronchial response, especially in persons with a history of bronchial hyperresponsiveness or concomitant asthma. Since the patient has no cough, fever or other respiratory symptoms, a respiratory infection is unlikely to be the cause. In addition, normal blood pressure and cholesterol levels rule out a possible cardiovascular cause. While allergic rhinitis is a possible explanation for the dyspnea, the differential diagnosis would also include mild asthma, nasal septal deviation or nasal polyps causing chronic obstruction. To confirm the exact cause, it would be necessary to evaluate whether the patient has a history of allergies, recent exposure to allergens or recurrent seasonal symptoms.

7.5 Scenario 5: Influenza

Patient: Male, 26 years old Symptoms: Cough, fatigue Risk Factors: None evident Clinical Analysis: Influenza is an acute viral infection of the respiratory tract caused by influenza A or B virus. It is characterized by systemic symptoms such as high fever, fatigue, myalgia, headache and respiratory symptoms such as cough and nasal congestion. In this case, the patient presents with cough and fatigue, which is compatible with an influenza-like illness, although the absence of fever is uncommon, as fever is one of the most characteristic signs of influenza. However, in some cases, fever may be mild or absent, especially in young adults without comorbidities. Differential diagnosis would include mild respiratory viral infections such as rhinovirus, adenovirus or even a postviral syndrome with residual fatigue. Mild bronchitis could also be considered if the cough is persistent.

7.6 Scenario 6: Common Cold

Patient: Male, 12 years old Symptoms: Cough Risk Factors: None evident Clinical Analysis: The common cold is a self-limiting viral infection caused by various respiratory viruses, primarily rhinovirus. It is characterized by mild symptoms such as nasal congestion, rhinorrhea, sneezing, coughing and occasional febrile fever. In this case, the isolated presence of cough without fever, rhinorrhea or other symptoms suggests a common cold at a late stage of the infectious process or a mild condition. In children, cough may persist even after resolution of the cold due to post-infectious bronchial hyperresponsiveness. The age of the patient is not a significant risk factor, and the course of the disease is usually benign without complications in most cases.

8 Discussion

Respiratory diseases are one of the main reasons for medical consultations, as they affect people of all ages and can present with similar symptoms, making differential diagnosis difficult. Given their high prevalence, it is essential to have tools that help healthcare professionals distinguish between conditions such as pneumonia, asthma, COPD, influenza, and allergic rhinitis, taking into account key factors such as age and symptoms, such as those you are providing us with at this time. In Colombia, specifically in Cauca, these challenges are accentuated by the presence of two annual epidemiological peaks of acute respiratory infections, which overload the healthcare system and limit hospital resources. For this reason, developing a system that allows for early and accurate identification of respiratory diseases would not only facilitate medical work but also contribute to better distribution of resources and more efficient care for patients. In these cases, we can identify most of the correct answers. For example, in the case of the patient with pneumonia: the clinical picture is consistent, as the patient is 64 years old with fever, cough, fatigue, and dyspnea. In the case of the patient with asthma, I find a difference, as asthma does not usually present with fever, so I believe the picture could suggest pneumonia in a 31-year-old patient with fever, cough, and dyspnea, meaning this patient has a respiratory infection. In the case of COPD: fatigue and dyspnea in a 67-year-old patient are indicative of COPD, a common chronic disease in older adults. In allergic rhinitis: isolated dyspnea in a 26-year-old man may be related to allergic rhinitis, although other obstructive causes should be ruled out first. In influenza: the diagnosis would be more accurate if there were fever, since influenza usually presents with fever, cough, and fatigue. In the common cold: the clinical picture of a 12-year-old child with a cough and no other systemic symptoms is consistent with a common cold, since these viruses do not normally cause fever and are mild. In terms of the AI model created, it can be said that it performs acceptably: (I) with an accuracy of 0.5868, indicating that more than half of the predictions were correct, (ii) a precision of 0.6767, in other words, it has a good ability to avoid false positives, (iii) a recall of 0. 5868, reflecting the correct identification of a considerable proportion of positive cases, and (iv) a balance between precision and recall, known as the f1 score of 0.5629, suggesting stable performance with room for improvement, which suggests that new adjustments to the model parameters or other preprocessing techniques could be explored to improve its performance. At the graphical interface level, there are aspects that could be improved with regard to the presentation and expansion of information, so that it is easy to use for the average user.

9 Conclusions

There are various respiratory diseases such as the common cold, influenza, asthma, COPD, pneumonia, and allergic rhinitis that impact human health by affecting respiratory function, causing acute symptoms, and leading to chronic complications. These conditions reduce physical capacity, impair quality of life, and generate socioeconomic burdens. Therefore, a medical specialist is needed to make a quick diagnosis to detect and treat the disease. AI can be a good support for the early diagnosis of diseases, helping the physician to streamline this process. This enhances clinical accuracy and optimizes time-consuming tasks.

That is why this paper presented a prototype tool called SIDTER, a prototype early diagnosis system for respiratory diseases assisted by AI with human supervision in the process. This tool supports doctors by helping to improve their diagnostic capacity for respiratory diseases and contributing to doctor-patient interaction.

It can be seen that of the six scenarios proposed and after expert analysis, the prototype identifies most of the diseases. In short, four of the six scenarios are correctly identified (pneumonia, COPD, allergic rhinitis, common cold), while the other two (asthma, influenza) show improvements suggested by the health expert.

As future work, new adjustments to the model parameters or other preprocessing techniques could be explored to improve its performance. There are also aspects of the user interface that could be improved in terms of the presentation and expansion of information, emphasizing that for the fields of gender, blood pressure, and cholesterol, instead of handling numerical data to enter the data, the value associated with the number must be entered in a character string, and it is necessary to validate the age in a range by adjusting the parameters so that only valid answers can be entered according to the range from 3 to 100 years old. It is also proposed to expand the number of clinical cases evaluated by incorporating a greater diversity of patients in terms of age, range of symptoms, and epidemiological context. This will improve the model's ability to generalize to different clinical profiles and reinforce the validity of the results obtained in real-life settings.

References

1. Organización Mundial de la Salud (OMS), "Documentos básicos," 1948, accessed: Mar. 28, 2025. [Online]. Available: https://apps.who.int/gb/bd/pdf/bd48/basic-documents-48th-edition-sp.pdf
2. ——, "Closing the gap in a generation health equity through action on the social determinants of health commission on social determinants of health," 2008, accessed: Mar. 28, 2025. [Online]. Available: https://www.who.int/publications/i/item/WHO-IER-CSDH-08.1
3. ——, *A handbook on how to implement mBreatheFreely mHealth for asthma and COPD.* World Health Organization, 2018.
4. J. L. Galindo, O. M. G. Morales, D. R. Sánchez, C. Celis-Preciado, and A. C. Arboleda, "Barreras de acceso en la atención de las enfermedades pulmonares intersticiales en colombia," *Saúde e Soc.*, vol. 28, no. 4, pp. 102–112, Dec 2019.
5. Organización Mundial de la Salud (OMS), "Report of the seventeenth annual meeting of the global alliance against chronic respiratory diseases virtual meeting," 2024, accessed: Mar. 28, 2025. [Online]. Available: https://iris.who.int/bitstream/handle/10665/380049/9789240104532-eng.pdf?sequence=1
6. A. Garud, D. Biswas, S. Moitra, and S. Moitra, "Health promotion in the management of respiratory diseases: an indian perspective," *The Lancet Respiratory Medicine*, Dec 2024.
7. O. C. V. P. D. D. Palomino, "Chronic respiratory disease: Considerations within the public health system," *Rev. Ciencias la Salud*, 2007.
8. E. Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again.* Basic Books, 2019.
9. P. Kapetanidis *et al.*, "Respiratory diseases diagnosis using audio analysis and artificial intelligence: A systematic review," *Sensors*, vol. 24, no. 4, p. 1173, Feb 2024.
10. N. B. Yoma and L. M. Inzunza, "Inteligencia artificial aplicada a la medicina respiratoria," *Rev. Chil. enfermedades Respir.*, vol. 37, no. 4, pp. 271–274, Dec 2021.
11. D. Bhattacharya *et al.*, "Coswara: A respiratory sounds and symptoms dataset for remote screening of sars-cov-2 infection," *Sci. Data*, vol. 10, no. 1, p. 397, Jun 2023.
12. R. Kumar *et al.*, "Using explainable machine learning methods to predict the survivability rate of pediatric respiratory diseases," *IEEE Access*, vol. 12, pp. 189 515–189 534, 2024.
13. D. F. Ochieng, "A differential diagnostic tool for obstructive lung diseases in adults using classification models," Jul 2021.
14. Y. Feng, Y. Wang, C. Zeng, and H. Mao, "Artificial intelligence and machine learning in chronic airway diseases: Focus on asthma and chronic obstructive pulmonary disease," *Int. J. Med. Sci.*, vol. 18, no. 13, pp. 2871–2889, 2021.
15. V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "Ai-based smart prediction of clinical disease using random forest classifier and naive bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, May 2021.
16. R. Ramalingam and V. Chinnaiyan, "A comparative analysis of chronic obstructive pulmonary disease using machine learning, and deep learning," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 1, p. 389, Feb 2023.
17. A. Anupriya and A. Thangavelu, "An adaptive approach towards prediction and diagnosis of lung cancer using heuristic grey wolf optimization approach and random forest classifier-lucago," *J. Theor. Appl. Inf. Technol.*, vol. 15, no. 3, 2023, [Online]. Available: www.jatit.org.
18. G. K. Yenurkar *et al.*, "Multifactor data analysis to forecast an individual's severity over novel covid-19 pandemic using extreme gradient boosting and random forest classifier algorithms," *Eng. Reports*, vol. 5, no. 12, Dec 2023.
19. A. Cockburn and J. Highsmith, "Agile software development, the people factor," *Computer (Long. Beach. Calif).*, vol. 34, no. 11, pp. 131–133, 2001.

20. T. D. Gooijer, "Discover quality requirements with the mini-qaw," in *2017 IEEE International Conference on Software Architecture Workshops (ICSAW)*, Apr 2017, pp. 196–198.
21. M. Goyal and Q. H. Mahmoud, "A systematic review of synthetic data generation techniques using generative ai," *Electronics (Switzerland)*, vol. 13, no. 17, Sep 2024.
22. Y. H. Montero, F. Jesús, and M. Fernández, "Propuesta de adaptaciÓn de la metodologÍa de diseÑo centrado en el usuario para el desarrollo de sitios web accesibles," *Rev. Española Doc. Científica*, vol. 27, 2004.
23. J. Brooke, "Sus: A quick and dirty usability scale," 1995, [Online]. Available: https://www.researchgate.net/publication/228593520.