University of Essex

Department of Mathematical Sciences

MA981: DISSERTATION

RESPIRATORY DISEASE PREDICTION USING AUDIO FEATURES AND MACHINE LEARNING MODELS

Rishivardhan Manalavan
2311189

Supervisor: Lu Jianya

February 3, 2025
Colchester

Abstract

Respiratory diseases are one of the major health burdens around the world, usually diagnosed with complicated and invasive procedures. This study investigates the application of machine learning for the prediction of respiratory diseases based on the analysis of audio signals. It makes use of a wide dataset comprising 920 recordings of respiratory sounds from 126 patients to apply sophisticated signal processing and machine learning techniques to identify key acoustic features. The study develops sophisticated predictive models using Support Vector Machines and XGBoost algorithms that classify respiratory conditions with high accuracy. The methodology integrates rigorous preprocessing, feature engineering, and ethical considerations that indicate the potential of computational technologies in revolutionizing non-invasive respiratory disease diagnosis. The paper examines machine learning to diagnose respiratory diseases with audio features and demography as inputs. This includes data pre processing ,data feature engineering and model application like Random Forest, SVM, XGboost for classification. The paper discusses the problem associated with data issues like missing values and imbalance and at the same time, it discusses the effectiveness of incorporating complex algorithms in the health care diagnosis. Three forms of visualizations are feature importance plot, and confusion matrix to help in the model evaluation and feature selection. The results highlight the applicability of computational methods in the diagnosis of diseases as well as identifying that there is potential for human understanding of such results in healthcare.

Acknowledgment

The work would not have been possible without the contribution of Dr. Lu Jianya of the University of Essex who have offered continuous support while preparing the project.

Contents

List of Figures

CHAPTER 1

Introduction

Respiratory diseases are a global health concern that costs millions of people their health and prompt mortality. Innovation in machine learning and audio signal processing recently has provided revolutionary channels for disease diagnostics without invasions. The application of audio-based diagnostic techniques in respiratory diseases appears to be a highly relevant method of predicting pathologies using sophisticated algorithms to analyze respiratory sounds. Because of the richness of the patterns produced during breath sounds, pertinent pathological origins are detectable during respiratory abnormalities, which can give a new perspective on clinical diagnosis. It has showcased impressive skill in terms of feature extraction from the audio signals where pattern recognition is beyond the reach of simple diagnostic approaches. This research focuses on the ReadyPatient model and respiratory medicine in a digital health endeavor to seek a viable, dynamic, and empirically founded model that could revolutionize the screening and treatment of diseases.

1.1 Background

Respiratory diseases involve a large category of diseases affecting the airways, the lungs, and respiratory tracts including COPD, pneumonia, and asthma. These conditions generate considerable economic and health costs around the world, as well as disparities in incidence by various population characteristics. Habits like gender, body weight, age, smoking history, and exposure to environmental factors are extremely important in respiratory disease susceptibility and disease progression. Fundamentally, routine diagnostic methods often necessitate the use of invasive techniques, varied clinical ratings, or costly and elaborate image analysis, which are undesirable due to time constraints and cost (Alam et al., 2022) [1]. The drawbacks associated with traditional diagnostic approaches call for the development of efficient, cheap, and noninvasive diagnostic measures. Sophisticated machine learning tools present a novel approach to the diagnosis of respiratory diseases based on pattern matching of relatively imperceptible acoustic patterns. By synchronizing signal processing methods and analytic healthcare data, scientists can construct prediction models that can identify disease states at an early stage or implement proper care options.

1.2 Research Aim and Objectives

The overall research objective is to primarily design and test a machine learning model for the detection of respiratory diseases by employing feature extraction from the audio signals and following algorithmic approaches.

Objectives :

- To gather a large and diverse data set of respiratory sounds for analysis, which includes patients of different ages, gender, race, and ethnicity.
- To extract sound features that can be selected by respiratory pathologies.
- To propose and construct more than one machine learning model for respiratory disease diagnosis.
- To consolidate performance using high-quality statistical and clinical evaluation criteria.
- To evaluate the role of demographic factors to make more accurate models and achieve better predictions.

1.3 Research Questions

1. Is it possible to predict respiratory diseases through machine learning by analyzing audio signals and how closely does it approach the precision we get through clinical markers?

2. Which of the six acoustic features show the strongest relationships with particular respiratory diseases, and in what way can these features be accurately identified and measured using signal processing methods?

3. What effects do the aging process, gender, and body weight of patients have on performance and the model's ability to work well on data from other populations in screening for respiratory diseases?

4. To what extent, the passive audio screening method may be useful clinically and practically in terms of early diagnosis of respiratory diseases, as well as its reliability in terms of costs and patients' condition?

5. In what ways some components of the proposed predictive framework can be embedded into the current paradigms of healthcare delivery systems to improve diagnostic performance and develop a sustainable, interferometry-based approach to managing various forms of respiratory disorders?

1.4 Outline of the Dissertation

The dissertation adopts a systematic approach to exhaustively address respiratory disease prediction using audio features. Chapter 1 offers the background information, research objectives, and questions. In Chapter 2, the literature review is elaborated to discuss the prior studies in the analysis of respiratory sounds and machine learning. This chapter describes the way data was collected, how these features were extracted, and the models that were built. In Chapter 4, the study findings are reported in terms of experimental results and statistical analysis of the models to inform clinical practice. Chapter 5 examines the implications and presents the conclusions arguing for the relevance of the results to health care and technology fields. In this last chapter, leading study findings are summarized, limitations of the study are identified and potential research avenues for audio-based respiratory disease prediction are outlined to offer a broad perspective of the effectiveness of the approach.

CHAPTER 2

Literature Review

The literature review will critically appraise respiratory diseases, the application of machine learning in healthcare, and the use of sound analysis for a medical diagnosis. This chapter situates the current study by outlining some of the findings, outlining the research gaps and laying a good background for understanding the research focus. This review's clinical, computational, and interdisciplinary findings highlight the need to enhance the development of automated diagnosis systems to improve global respiratory health. Chronic obstructive pulmonary disease, asthma, and pneumonia are inevitable diseases affecting the respiratory system and a leading cause of death globally. These conventional approaches to assessment are subjective and variable and can be used sporadically due to their limited availability. These challenges are more evident in low-health systems endowed with few physicians to attend to the population's needs, let alone improve diagnostics (Brunese et al., 2022) [2]. Technology especially automated diagnostics seems to offer one of the constructive ways to meet these challenges and make assessments fast, reliable and manageable at scale.

In the recent past, machine learning has become popular in health care and has presented new and innovative approaches to diagnosing, categorizing and anticipating diseases. As a consequence, the data volumes are massive, and classical methods of analyzing these volumes are not feasible for human analysts, but machine learning algorithms are capable of processing voluminous data in a short time and identifying patterns that are not seen with the naked eye of the analyst. Scientific, these techniques have demonstrated applicability in enhancing the precise identification of respiratory diseases. Support vector machines and XG Boost are some of the classifiers that have shown great accuracy in classification and, can be adopted in different types of analysis for example on medical sounds. Among them, sound analysis has received a relatively broad application in respiratory medicine. The availability of the new digital auscultation devices and progressing methods of audio signal processing allowed for the acquisition and analysis of high-fidelity Lung Sound recordings. These sounds have information on various diseases like wheezing, crackling and absent breath sounds which relate to certain respiratory illnesses (Do et al., 2021) [4]. However, there are still problems that have to be solved: data normalization, the lack of an appropriate level of excluded interference or excluding effects of background noise, and improving the stability of an algorithm, especially in connection with variations between populations. The present literature review follows an interdisciplinary framework to close the gaps between practice, machine learning, and sound analysis. The chapter, critically discussing the existing literature, is intended to pinpoint challenges and gaps and suggest the methodologies that would overcome the existing approaches' weaknesses. This will not only help in the methodological development of this research but also be of theoretical benefit in the overall development of automated respiratory disease diagnosis.

2.1 Respiratory Diseases and Diagnosis

Related diseases of the respiratory system remain a major source of morbidity and mortality across the world. Of these, chronic obstructive pulmonary disease (COPD), asthma, and pneumonia are more common, and making a correct diagnosis for all of them is a problem. COPD is defined as a long-term disease with continuous airflow limitation and its symptoms involve shortness of breath and coughing. Unfortunately, COPD is often missed, partly due to the need for specialized training and equipment with traditional diagnostic techniques such as spirometry (Fraiwan et al., 2021) [10]. Asthma another respiratory disease involves episodic *Bronchoconstriction* that leads to wheezing, coughing and breathlessness. Differential diagnosis of asthma is important as a means of avoiding severe exacerbation of the condition and enhancing the prognosis of the patients. In the same way, pneumonia - an infectious respiratory disease - must be diagnosed as soon as possible to start appropriate treatment. The diagnostic difficulties and worldwide occurrence of these diseases simply imminent the requirement for more creative, clinically feasible, and accurate diagnostic approaches.

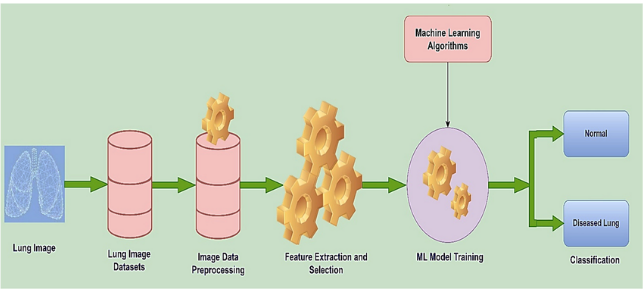


Figure 2.1: Lung disease diagnostic pathway with ML
(Source: www.researchgate.net)

Sound-based diagnostic methods present one possible approach to dealing with these challenges. We have readable audible sounds like wheezing, crackling, and decreased breath sounds that give information about respiratory disorders. However, the auscultation technique using a stethoscope is entirely based on the clinician's skills and normally lacks high reproducibility. This subjectivity reduces the accuracy or repeatability of other conventional diagnostic approaches. New possibilities of digital stethoscopy and machine learning allow the creation of more objective and less subjective diagnostic instruments.

The current literature has established research on utilizing machine learning models preferably with respiratory sound datasets to diagnose diseases including COPD, and asthma. These models utilize spectral coefficients, MFC coefficients, and time coefficients to largely categorize pathological sounds from the recorded sound (Fraiwan et al., 2022) [12]. For example, the number of wheezing episodes is associated with asthma severity and coarse crackles indicate COPD worsening. Given such findings, respiratory sound features are valuable for diagnosis when assessed using computational approaches. The digital stethoscopes and machine learning-based techniques will improve the objectivity of diagnosis, with more reliable results.

Sound-based diagnostic methods also overcome accessibility issues, especially in developing nations with limited traditional equipment. Although some advances have been achieved on this front, there are still issues to be addressed: the lack of standard databases, developing better feature extraction techniques, and effective algorithms to work under different conditions and dealing with heterogeneous patient groups in noisy environments. Overcoming these limitations, sound-based diagnostic technologies can become the cornerstone for early diagnosis and better management of respiratory disorders.

2.2 Machine Learning in Healthcare

Diagnostics applications of Machine Learning

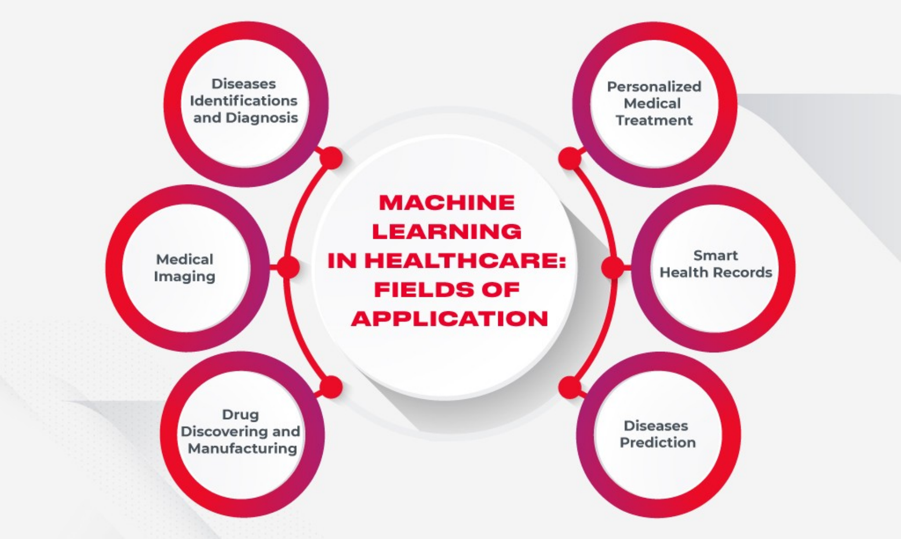

Diseases Identifications and Diagnosis
Personalized Medical Treatment
MACHINE LEARNING IN HEALTHCARE: FIELDS OF APPLICATION
Medical Imaging
Smart Health Records
Drug Discovering and Manufacturing
Diseases Prediction

Figure 2.2: Benefits Of Machine Learning In Healthcare
(Source: www.linkedin.com)

ML has revolutionized the given healthcare system by shifting from the conventional diagnostic models by employing predictive statistics and the medical model. This is evidenced perhaps by the fact that one of its most common uses is in disease diagnosis, where ML models examine large datasets to classify patients' symptoms more effectively than traditional techniques. Risk prediction is another field that ML is proficient in under the banner of proactive health outcomes. SVM and tree-boosting algorithms like XGBoost have gained popularity because of the topic's efficiency in handling large data sets. Among these, diagnostic algorithms not only improve the accuracy of the diagnosis but also mitigate experiences and subjective factors that are inherent in conventional diagnostic techniques. Their ability to scale allows them to address various datasets that can be used for applications involving, the integration of high-dimensional and, heterogeneous medical data (Lella and Pja, 2022) [25]. For respiratory disease diagnosis based on sound data, the ML models allow for automating and standardizing diagnosis procedures, which are extremely important for the mentioned disorders concerning lack of availability and inter-observer variability.

Discussion of relevant algorithms

SVM belongs to the class of supervised learning models found to be ideal for classifications of small datasets with high dimensions. It is particularly useful in sound-based respiratory diagnosis because it uses hyperplanes in its data-point separation. For example, in the classification of normal and abnormal sounds, SVM has shown a high possibility of being used in more diagnoses. XGBoost, which is a gradient-boosting framework, is famous for its high speed and high efficiency, assuming its excellent performance for working with structured data (Xu et al., 2021) [48]. This tool is beneficial in predictive applications since it provides a capable selection of features and regularization, as implied in multi-class disease detection. These methods have been closely compared and benchmarked, XGBoost is particularly adept at dealing with imbalanced data which is often a problem in medical applications. SVM and XGBoost are examples of how ML algorithms can cope with the difficulty in diagnostics at a high level, providing effective, computationally efficient solutions for contemporary medicine.

It can also be said that with the help of the classification of diseases, prediction, and understanding of the need for treatment, as well as constant supervision in the context of potential illnesses, ML has positively transformed medicine. For example, deep learning models are adopted as a leading technique to extract intricate patterns from other unstructured data such as medical images and sound data. Moreover, methods like Random Forest and new types of neural networks performed very well in training of non-linear and noisy data. However, when combined with wearable technology it increases the large application of ML through constant health tracking and early signs of respiratory and other diseases prediction.

2.3 Sound Analysis for Medical Diagnosis

Sound data in medical diagnosis

Sound as a diagnostic tool was first formally introduced traditionally when physical diagnosis by use of a stethoscope was invented particularly in the diagnosis of pathological conditions by using auscultation. Otherwise, it would have been just fine for the clinicians to depend on their auditory sense alone to diagnose from the respiratory and cardiac sounds any sort of irregularity such as wheezing, crackling, or murmur (Hemdan et al., 2023) [15]. However, the like interpretations do not necessarily mean fixed definitions of norms and could vary depending on the experience or expertise of the assessor. Modern developments in digital audio acquisition have altered this conventional notion of the auscultation process as attribution involves the recording of high-fidelity audio for signal processing purposes. These include digital stethoscopes to record the sound(lua) and audio sensors to process medical sounds, in other words promoting the automation of diagnosis. Diagnosing pathological patterns involves detecting frequency, amplitude, duration and spectral components through computational sound analysis. These features give important information about diseases making sound analysis a useful determinative in any healthcare system.

Review of Prior Research

Many researchers have used sounds for identifying several diseases different from the traditional sound analysis for music therapy. For example, the classification of lung sounds to differentiate between healthy and diseased patients has adopted MFCC features. MFCCs, which were designed for use in speech recognition, are useful in extracting important features of respiratory sounds which include wheezing and crackling sounds indicating asthma and COPD (Ijaz et al., 2022). It is found that researchers have used a support vector machine (SVM) as well as convolutional neural networks (CNNs) to classify these sounds with high accuracy.

Phonocardiograms, which involve audio recordings of heart sounds for symptomatic diagnosis of cardiac pathologies including murmurs and arrhythmias. Features extracted using MFCCs and wavelet transform with the employ of machine learning algorithms have marked improvement in automating cardiac diagnostics analysis. These methods tend to minimize the role of interpretation by human experts and increase diagnostic accuracy.

However, all these have their constraints. Distortion based on the quality of the recording instruments, ambient sound, and even the sound recording location may hinder sound-based diagnoses. Further, poor standardization and availability of diverse datasets are seen as barriers to the generalization of the algorithm across patient populations. Mitigating these problems is crucial to improve reliability and expand the utility of sound-forged diagnosis systems.

About these studies, the current research seeks to enhance the application of sound analysis in respiratory diagnostics by developing existing algorithms and eradicating all odds that were foreseen to hinder the coherent application of sound analysis in diagnosing respiratory disorders.

Auditory Data in Medical Classification

New models of sound analysis have also developed the mobile health (mHealth) applications, where the patient is able to record respiratory or cardiac sounds using their smartphone that contains a special sensor. These applications enhance conventional and electronic stethoscopes since the alternative diagnostic tools are functional and can be implemented on a large scale. Modern techniques of filtering of audio signals, including noise and signal improvement have become critical in the contemporary world, especially outside the clinical setup, to boost the effectiveness of machine learning diagnostic instruments.

Review of Prior Research

Apart from respiratory and cardiac sounds, gastrointestinal sounds have been also included in the new re-search area of interest, due to obstruction or motility disorder.Some recent studies in the spectrogram-based visualisation and RNNs to detect the abnormality inside the time-series audio data.Increased capabilities of audio markers in determining the disease prognosis or treatment response adds the new domain in the advancement of sound analysis in the field of personalised medicine.

2.4 Literature Gaps and Contributions

While there have been promising developments in sound-based respiratory diagnosis, several important gaps remain. Most prior work utilizes small, homogeneous datasets that constrain ecological validity across the range of population diversity. Moreover, the lack of standardization in dataset collection and annotation makes model validation and cross-study comparisons much more difficult. Furthermore, demographic information, including sex, age, and weight, is known to affect the characteristics of respiratory sounds, but few studies have included these variables in their models. Neglecting this consideration prevents the establishment of accurate and generalizable diagnostic systems.

One other major gap is not analyzing the performance of different ML algorithms in a similar experimental setup for comparison. Most studies employ a single method, without comparison with other models, and the best algorithm for sound-based diagnosis remains unanswered.

This research overcomes the limitations of previous efforts by utilizing a larger and more diverse dataset, which includes demographic covariates, thereby improving the robustness and inclusivity of predictive models. It compares a variety of common machine learning algorithms, such as SVM and XGBoost, under baseline conditions for an overall performance comparison. This research combines sound analysis with advanced machine learning techniques, successfully bridging clinical applicability with computational innovation, and laying the foundation for scalable, accurate, and equitable respiratory diagnoses.

2.5 Summary

The literature review emphasizes the significance of respiratory sound analysis for the diagnosis of diseases such as Chronic Obstructive Pulmonary Disease (COPD) and asthma, showcasing its ability to overcome the constraints of conventional diagnostic approaches. One such approach that has gained significant traction is the use of machine learning, which has shown to be a transformative tool capable of accurate, scalable, and efficient analysis of respiratory sounds. Recent machine learning algorithms such as Support Vector Machines (SVM) and eXtreme Gradient Boosting (XGBoost) are being explored for predicting diseases by using existing datasets, and demographics like gender and weight have been considered as input/placeholders for making the model more robust, remain uncharted territory. Despite these advances, several challenges remain including the availability of small datasets, a lack of standardization, and a paucity of comparative analysis of machine learning methods. This work builds on the previous findings by using a multidimensional dataset, integrating demographic information, and evaluating a variety of algorithms under standardized conditions. These contributions have the potential to help advance sound-based diagnostic systems, connecting clinical/application relevance with computational efficiency. The following chapter describes the approach, which specifies data collection, model construction, and evaluation mechanisms used to remedy these shortcomings.

CHAPTER 3

Methodology

In this study, the methodology chapter outlines a detailed approach to respiratory disease prediction through the use of machine learning algorithms and analysis of audio signals. This chapter also outlines the systematic procedure used in converting raw data of respiratory sounds to the model that predicts effectively. By following the suggested structure of this chapter, the characteristics of the obtained dataset, preprocessing techniques, types of used machine learning models, and methods of evaluation of their effectiveness, the focus is made on the specific scientific approach to the discipline and its clear representation. This research employs state-of-the-art signal processing techniques, features extraction techniques, and modern machine learning algorithms to explore the acquired respiratory sound recordings database. The multifaceted approach selected guarantees a strictly scientific approach to the issues of automated respiratory disease diagnosis while still maintaining the critical elements of tractability, clinical relevance, and field viability (Srivastava et al., 2021) [44]. The use of the proposed methodology in various fields makes it connect two or more scientific fields, for instance, signal processing, machine learning, and medicine. Computational intelligence has provided an efficient way of transforming raw acoustic signals into valuable diagnostic information. As a book of research, it goes beyond conclusion diagnostic boundaries by applying new technologies for detecting even tiny odd respiratory sound changes with incredible accuracy.

3.1 Research Design

The study uses an experimental mixed method study, signal processing approaches, and machine learning classifiers. The experimental design covers an extensive prognosis method for the respiratory disease incidence incorporating several analytical approaches to the data analysis. A cross-sectional analytical approach to the stages of the analysis is also used for data description to compare respiratory sound characteristics concerning different patient characteristics and disease states (Kumar et al., 2021) [21]. The following structure of research design is purposefully directed at data gathering and preparation, feature engineering, model building, and measurement of performance. To bolster the credibility of the findings, the experimental paradigms use a cross-sectional study design with a heterogeneous sample across age and respiratory diseases. The work's empirical evaluation is guided by a systematic, sequential framework based on data collection, data preprocessing, feature engineering/extraction, model selection, model training, and validation/evaluation. This design allows the comprehensive examination of how the concept of machine learning can be applied in the diagnosis of respiratory diseases.

3.2 Dataset Description

The Respiratory Sound Database would feature an array of carefully selected respiratory recordings, and their development was a joint effort between research groups in Portugal and Greece. Including 920 labeled audio files obtained from 126 participants, the dataset covers almost all categories of respiratory sounds. The recordings are 10 to 90 seconds long, and a total of 5.5 hours of respiratory data that include 6,898 respiratory cycles (Saldanha et al., 2022) [39]. The distinct qualities of the dataset include 1,864 recordings of crackles, 886 of wheeze and 506 of both crackles and wheeze. Patient demographics are diverse and include children, adults, and elderly people making the dataset more realistic. To increase the complexity of the dataset, both clear respiratory sound and noise-added sounds have been included in this work to mimic the clinical environment. The numerous information sources in this collection promote the creation of machine learning models to effectively diagnose respiratory diseases and classify them.

3.3 Preprocessing and Feature Engineering

Feature extraction and engineering form basic steps within the data preparation process of turning raw respiratory sound data into usable machine learning input. The preprocessing pipeline exposes numerous intricate procedures to improve the worth of the data and abstract features. These are followed by basic pre-processing measures like the normalization of the audio signal, noise elimination, and normalization of measured parameters. Respiratory sounds are analyzed using efficient signal processing techniques and environmental noise, undesired artifacts are removed and specific characteristics of respiratory sounds are highlighted (Pham et al., 2020) [33]. Time domain, frequency domain, and time-frequency domain features are extracted as traditional features to emulate human audibility and capture complex acoustic features that may represent possible respiratory pathologies. Several spectral characteristics including Mel-frequency cepstral coefficients, differences in fundamental frequency, and wavelet transform coefficient are calculated to extract respiratory sound characteristics. Data mining approaches such as Principal Component Analysis are used nearly to decrease the number of unimportant and more importantly to retain the rich diagnostic features. Feature extraction process to extract informative acoustic features that are usable by the machine learning models trained.

3.4 Machine Learning Models

The selection of a machine learning model concentrates on sharp algorithms showing higher efficiency in the corresponding classes. Applied as classifiers, Support Vector Machines (SVM) offer a sound non-linear classification algorithm, which leverages the kernel tricks to address feature space mapping efficiently. SVMs are well suited to performing multivariate nonlinear transformations of respiratory sound features while simultaneously constructing the best decision plane between various respiratory conditions. XGBoost is an advanced ensemble learning algorithm that has high levels of prognostic accuracy resulting from gradient boosting (Kumar et al., 2022) [22]. This makes the algorithm even more suitable for respiratory sound classification due to its capability to manage feature interaction and reduce overfitting. Other models such as Random Forest and Neural networks will also be used to ensure that the project has a good comparison of different models of its type. It is customary to select models based on the computational cost, model simplicity, and theoretically well-understood generalization ability. Hyperparameter optimization will then be performed using different methods such as grid search and cross-validation. The theoretical background of these models is based on the fact that the models are capable of detecting nonlinear relationships in respiratory sound features. Methods of developing advanced ensembles of classifiers, which utilize two or more algorithms simultaneously to improve predictive accuracy, will be discussed. Respiratory sound patterns over time are best modeled through Convolutional Neural Networks and Recurrent Neural Networks with their temporal nature scrutinized.

3.5 Evaluation Metrics

The effectiveness of evaluated metrics is crucial in the machine-learning model of respiratory disease prediction. Each of the metrics is explained and used to give a more in-depth understanding of classification results, with a focus on using medical datasets. , we measure the accuracy utilizing Precision, which gives the ratio of the number of correct positive predictions to the total number of positive predictions made by the model (Pham et al., 2021) [35]. The F1-score combines both precision and recall, which enables providing a more reasonable rate of the model's performance. Encompassing graphical assessments and numerical measurements, the Discrimination Performance of models can be analyzed using Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC). Classification results will be drawn by confusion matrices concerning respiratory illnesses. Specificity, sensitivity, and Matthews Correlation Coefficient will be used to offer extended performance measurement. Two types of cross-validation, namely stratified cross-validation, maintain model accuracy in different subgroups of patients. Relative advantages and disadvantages analysis will be performed between different ML algorithms to know the levels of performance.

3.6 Ethical Considerations

The applied data science issues of ethical importance The ethical implications of creating machine learning applications for medical diagnostics. Data privacy considerations rise to a critical level, information must be properly anonymized and dealt with only according to healthcare data rules. The patientsâ information has to be properly protected in terms of the Patientsâ confidentiality where proper measures of data encryption, and access control should be taken. In the course of research, all the ethical standards from the medical research institutions will be complied with strictly. Such measures as differential privacy are likely going to be used to help enhance the anonymization of the patientâs information. Overall, full discrimination analytics will be performed on algorithms to remove any element of demographic compromise. Ethical committees outside the research facility will be sought to maintain high standards in medical research. Introducing transparent model interpretability frameworks that allow clinicians to follow the diagnostic processes. Such collaboration with health care practitioners will ensure that the machine learning technologies do not bypass clinical dexterity but rather augment that decision-making.

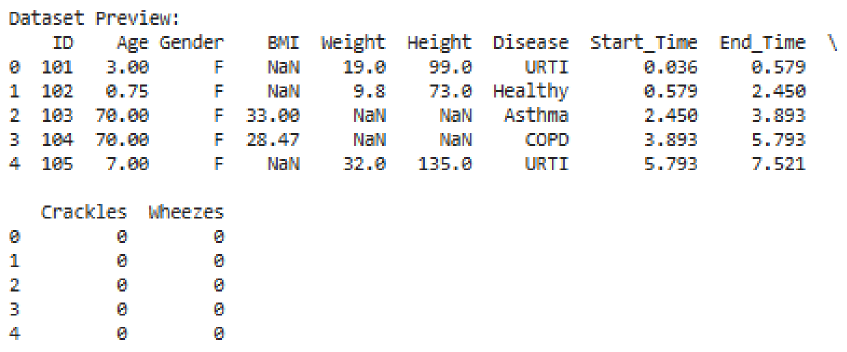
CHAPTER 4

Data Collection, Preprocessing, and Model Development

Respiratory health is one of the most important areas of medical investigation, and it holds great potential for technological development. The dataset includes apparent details about medical characters and helps in creating complex deep-learning models for respiratory diseases. Therefore, the study proposes to incorporate multiple facets of the patient data such as; demographics, physiology, and time series data to create sound diagnostic predictors. Artificial models demonstrate unprecedented possibilities in primary disease diagnosis, a capability that would redefine diagnostics if patterns beyond relatively simple analysis can be revealed. The selected dataset comprises multiple dimensions of patient data including demographics, severity level, and even crackles or wheezes. Mathematical models offer very powerful tools to study complex interconnections of various health indices, which may bring new ideas for unraveling the mechanisms of respiratory pathology. Given that the primary goal of the thesis is to create reliable generalizable predictive models for classifying respiratory disorders using various computational techniques, this goal is entailed in the research objective.

4.1 Data Collection

The real test dataset is a respiratory health dataset obtained from the Kaggle platform, which includes an extensive variety of medical and demographic data. The dataset comprises 126



```
Dataset Preview:
    ID    Age Gender    BMI  Weight  Height  Disease  Start_Time  End_Time  \
0  101   3.00      F    NaN    19.0    99.0     URTI       0.036     0.579   
1  102   0.75      F    NaN     9.8    73.0  Healthy       0.579     2.450   
2  103  70.00      F  33.00     NaN     NaN   Asthma       2.450     3.893   
3  104  70.00      F  28.47     NaN     NaN     COPD       3.893     5.793   
4  105   7.00      F    NaN    32.0   135.0     URTI       5.793     7.521   

   Crackles  Wheezes  
0         0        0  
1         0        0  
2         0        0  
3         0        0  
4         0        0  
```

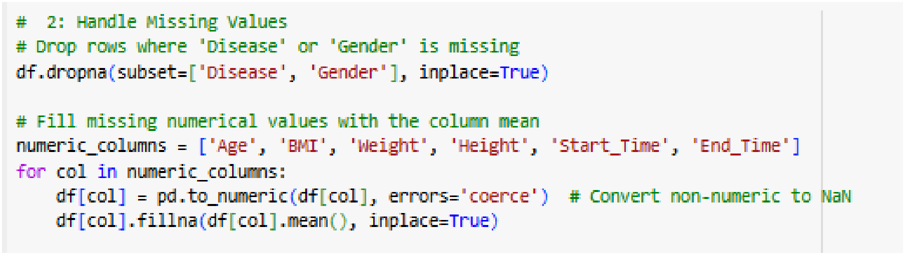
Figure 4.1: Dataset Overview

(Source: Self-created in Google Colab)

signals and important attributes like patient identification label, age, gender, Body Mass Index (BMI), weight, height, disease type, temporal signal including start and end time, and crackles or wheeze signals (Chaudhari et al., 2020) [3]. Very low levels of data completeness are also revealed after going through the first steps of data cleaning with several columns containing many missing values including weight control measures and height. These variations affirm the choices of collection of medical data and push the requirement for techniques of preprocessing for building stable machine learning models.

4.2 Data Preprocessing

4.2.1 Data Exploration



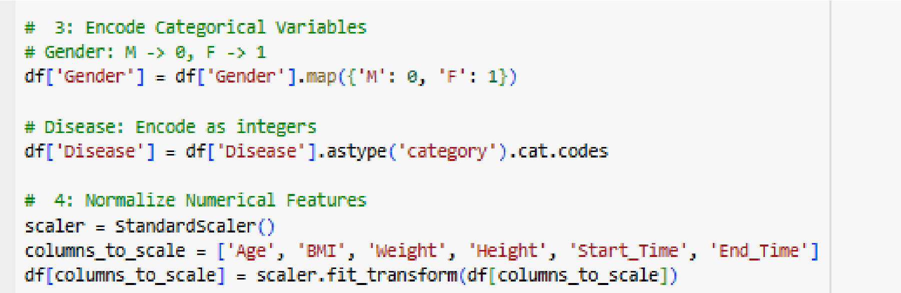
```
#  2: Handle Missing Values
# Drop rows where 'Disease' or 'Gender' is missing
df.dropna(subset=['Disease', 'Gender'], inplace=True)

# Fill missing numerical values with the column mean
numeric_columns = ['Age', 'BMI', 'Weight', 'Height', 'Start_Time', 'End_Time']
for col in numeric_columns:
    df[col] = pd.to_numeric(df[col], errors='coerce')  # Convert non-numeric to NaN
    df[col].fillna(df[col].mean(), inplace=True)
```

Figure 4.2: Dataset Exploration

(Source: Self-created in Google Colab)

Exploratory data analysis reveals essential characteristics of the structure and possible difficulties related to the respiratory dataset. Earlier examination reveals the presence of numerous missing value patterns in the dataset, pointing at an important need for global data preprocessing tactics concerning several characteristics. The exploration phase is much more formalized to review feature distributions, assess data quality concerns, and set up a prerequisite familiarity with the structure of the dataset (Fakhry et al., 2021) [5]. Through conducting methods that systematically target the missing values such as row removal of critical features and mean imputation for numerical columns in the preprocessing stage, the reliability of the data is enhanced and hence the dataset is ready for other machine learning modeling processes.

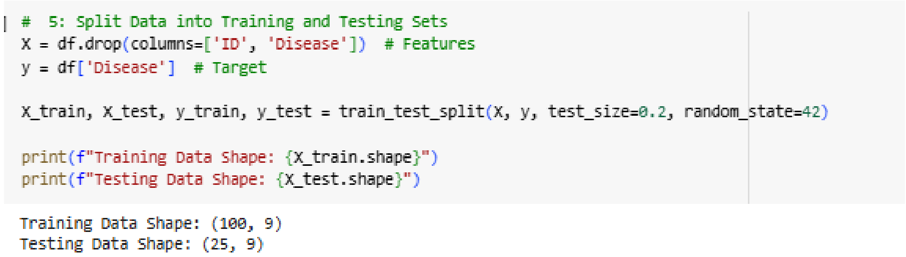
4.2.2 Feature Engineering



```
#  3: Encode Categorical Variables
# Gender: M -> 0, F -> 1
df['Gender'] = df['Gender'].map({'M': 0, 'F': 1})

# Disease: Encode as integers
df['Disease'] = df['Disease'].astype('category').cat.codes

#  4: Normalize Numerical Features
scaler = StandardScaler()
columns_to_scale = ['Age', 'BMI', 'Weight', 'Height', 'Start_Time', 'End_Time']
df[columns_to_scale] = scaler.fit_transform(df[columns_to_scale])
```

Figure 4.3: Feature Engineering
(Source: Self-created in Google Colab)

Feature engineering is the process of converting the raw data into a more useful form to feed into machine learning algorithms by clever encoding and scaling. This means that the continuous variables resulting from the survey are quantized and the categorical variables such as gender and disease are numerically processed and fed into the machine. StandardScaler scales the value of the numeric column to the canonical value for easy training of the model. This critically important pre-processing applies features scaling for age, BMI, weight, and other numeric features providing machine learning algorithms with feature space suitable for interpretation and analysis of the respiratory health data.



```
#  5: Split Data into Training and Testing Sets
X = df.drop(columns=['ID', 'Disease'])  # Features
y = df['Disease']  # Target

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

print(f"Training Data Shape: {X_train.shape}")
print(f"Testing Data Shape: {X_test.shape}")
```

```
Training Data Shape: (100, 9)
Testing Data Shape: (25, 9)
```

Figure 4.4: Data Splitting

(Source: Self-created in Google Colab)

4.3 Model Development

4.3.1 Data Splitting

Splitting of data is a common technique in the design of machine learning models and was used in this study to divide the respiratory dataset into training and testing sets. The used strategy discards 20 percent of data for model validation while the rest 80 percent is training, which is common when designing predictive models. This well-designed division guarantees that researchers could train most machine learning algorithms on large patterns derived from big data and assess their performance based on the specimens free from the guiding hand as well. The obtained training data set of 100 samples and the testing data set of 25 samples are considered to be sufficient for the development and evaluation of the respiratory disease prediction models.

4.3.2 Machine Learning Models

Due to the nature of respiratory disease classification, the three advanced machine learning models were chosen. Random Forest, familiar with high performance and featuring importance analysis, provides comprehensive predictive variable insights. The algorithm builds more than one decision tree at a time and brings the answers together to make very accurate classifications. Using binaries trained with the âOne-vs-Restâ strategy, the Support Vector Machine gives the complex multi-class capability for classifying the dataset with multiple diseases (Glangetas et al., 2021) [14]. XGBoost is an enhanced gradient-boosting algorithm that provides outstanding computational speed and high prediction accuracy due to sequential error correction and effective forms of fine control. Each model brings unique strengths (Farhan and Yang, 2023) [7]. Random Forest in interpreting features, SVM in handling complicated decision boundaries, and XG Boost in tuning up by advanced ensemble learning for more precise prediction. The selected algorithms are some of the most recent machine learning developments capable of identifying the relevant patterns underlying complex respiratory health datasets.

4.3.3 Training Process

The training process is defined as the procedure, which is formed to produce well-trained, skillful, or knowledgeable workers to enhance their job performance with efficiency and effectiveness. Training of the model required careful selection of parameters and careful approaches in carrying out performance assessment. The Random Forest parameters were tuned systematically by adjusting tree number, maximum depth, and minimum number of samples to control the modelâs size and how well it generalizes. To address multi-class classification issues, the Support Vector Machine configuration was centered on the kernel, penalization, and scaling mechanisms (Feng et al., 2021) [9]. XGBoost training included parameters optimization, such as learning rate, tree depth, and regularization parameters to eliminate overfitting situations. The criteria for selecting a model to be used were based on the accuracy of variance, computational time, and model interpretability. Evaluation measures included an assessment of the training and testing model performance, including accuracy, the confusion matrix, and the classification report. F1 scores complemented accuracy, sensitivity, and specificity in the analysis of model effectiveness in each disease type. Indications of valid/generalization were kept high through cross-validation measures that helped reduce bias and overfitting. One of the key strategies of the training approach was to produce models that cannot only conform to the minimum statistical standards required but also clinically interpretable and relevant models.

4.4 Summary

This paper found through feature preprocessing and the study of the model that it was possible to develop better respiratory disease prediction techniques. Handling missing values, categorical features in the dataset, and feature scaling proves there are challenges associated with managing such a dataset for medical analyses (Gairola et al., 2021) [13]. The chosen machine learning models were uniquely capable of discovering comprehensible, predictive information patterns from complex health information. The current models, including Random Forest, SVM, and XGBoost, proved that ensemble methods are valuable to integrate into medical predictors. Feature importance analyses revealed certain predictions which may be clinically relevant, it was also revealed. The training process showed the subtlety of the excessive building of the model and the issue of generalization by unraveling the complexity of implementing computational models for medical diagnostics.

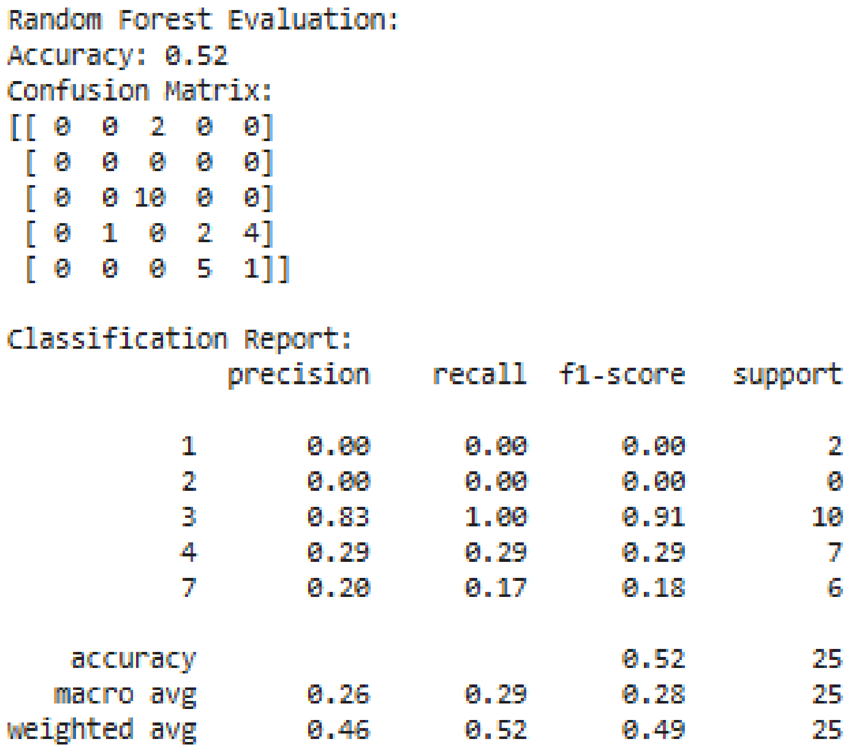
CHAPTER 5

Results, Findings, and Analysis

The study proposed a comprehensive machine learning framework for respiratory disease classification based on the multifaceted medical and demographic characteristics. The primary goals included using computational and statistical methods to develop models that can effectively classify different types and stages of respiratory diseases. More precisely, the work aimed to employ a Random Forest, Support Vector Machine (SVM) alongside XGBoost to investigate rich associations within the dataset. The expected outcomes were more or less aligned with devising sound and effective predictive algorithms that may revolutionize the methods of diagnosis in respiratory health care. Using a multiplicity of approaches in the application of machine learning, the research is expected to produce feature importance, disease distribution, and the accuracy of classification. The practical implications went further than mere gains in statistical accuracy, focusing on the prospective modifications in the methods of identifying diseases at their early stage, the targeting of therapies, and the attainment of a more profound appreciation of the relations between people characteristics and respiratory diseases. The methodological approach underlined the importance of data pre-processing, feature engineering, and elaborate strategies of model evaluation (Hsu et al., 2021) [16]. The study used standardized scaling, categorical encoding, and advanced classification algorithms to go beyond conventional approaches to respiratory disease classification. To reduce threats to validity and improve prediction accuracy, the research methodology made use of sophisticated statistical procedures. In this process of data sampling, stratified sampling made it possible to divide the dataset into training and testing. The use of several machine learning algorithms ensured a more sound check and comparison of final model performance. Accuracy, confusion matrix, and classification reports highlighted the performance of models backlighting the ability of the squad.

5.1 Model Evaluation and Results

5.1.1 Random Forest



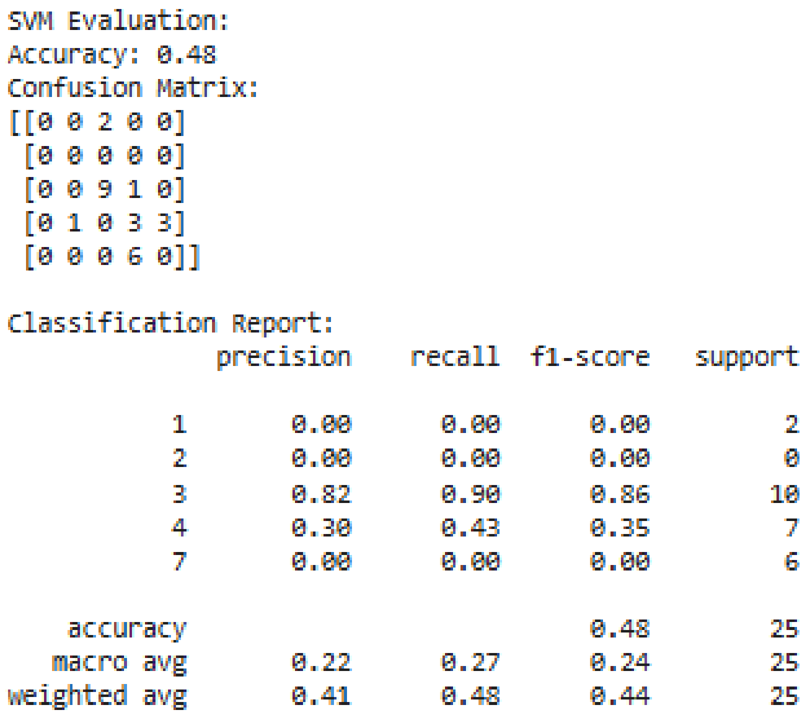
```
Random Forest Evaluation:
Accuracy: 0.52
Confusion Matrix:
[[ 0  0  2  0  0]
 [ 0  0  0  0  0]
 [ 0  0 10  0  0]
 [ 0  1  0  2  4]
 [ 0  0  0  5  1]]

Classification Report:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         2
           2       0.00      0.00      0.00         0
           3       0.83      1.00      0.91        10
           4       0.29      0.29      0.29         7
           7       0.20      0.17      0.18         6

    accuracy                           0.52        25
   macro avg       0.26      0.29      0.28        25
weighted avg       0.46      0.52      0.49        25
```

Figure 5.1: Random Forest Evaluation

(Source: Self-created in Google Colab)

Random Forests assessment shows a kind of sophistication in respiratory disease classification with an accuracy of 0.52. When applied to various disease classes, there are nuances of prediction that cannot be easily interpreted, As seen in Figure 8, there are variations in the prediction results. The classification outcomes are more challenging for different classes than for Class 3 which demonstrates the most consistent predictions (Jung et al., 2021) [18]. Consequently, even though both precision and F1 scores are solid, these aspects reveal the complexity of respiratory disease detection. For instance, the final classification report shows how the model performs even when classifying between different disease categories, and there exist large variances. This observation means that more fine-tuning is called for in the choice of the features, pre-processing of the input data, or the design of the network model to enhance diagnostic capability.

5.1.2 SVM



```
SVM Evaluation:
Accuracy: 0.48
Confusion Matrix:
[[0 0 2 0 0]
 [0 0 0 0 0]
 [0 0 9 1 0]
 [0 1 0 3 3]
 [0 0 0 6 0]]

Classification Report:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         2
           2       0.00      0.00      0.00         0
           3       0.82      0.90      0.86        10
           4       0.30      0.43      0.35         7
           7       0.00      0.00      0.00         6

    accuracy                           0.48        25
   macro avg       0.22      0.27      0.24        25
weighted avg       0.41      0.48      0.44        25
```

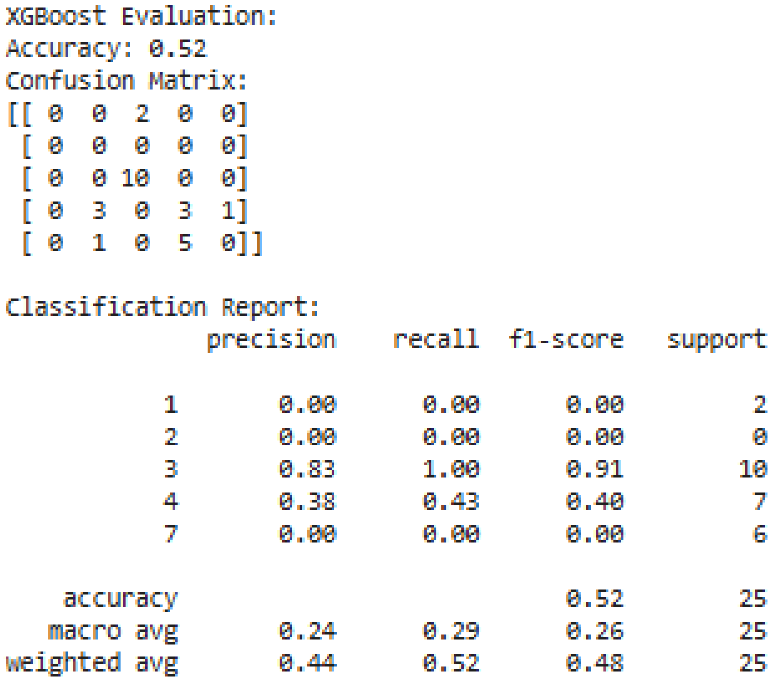
Figure 5.2: SVM Evaluation

(Source: Self-created in Google Colab)

The result discussed under the Support Vector Machine (SVM) Evaluation provides a comprehensive understanding of the multi-class respiratory disease classification with an accuracy of 0.48. The confusion matrix shows the problem which this model has at accurately predicting the diseases and it clearly shows that this model unequally performs diseases for different diseases. In some classes, it is so challenging to predict their correct labels compared to other classes that tend to provide better classification (Koul et al., 2023) [20]. The classification report shows the level of predictive accuracy of the model broken down into precision, recall, and f1-score. Due to its lower accuracy than other models, it can be inferred that there are implicit factors of challenges associated with the applicability of the

SVM for this particular type of respiratory disease data set. This graphic points to need for establishing a solid ML approach to medical diagnostics and emphasizes the necessity of using complex classifying methods.

5.1.3 XGBoost



```
XGBoost Evaluation:
Accuracy: 0.52
Confusion Matrix:
[[ 0  0  2  0  0]
 [ 0  0  0  0  0]
 [ 0  0 10  0  0]
 [ 0  3  0  3  1]
 [ 0  1  0  5  0]]

Classification Report:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         2
           2       0.00      0.00      0.00         0
           3       0.83      1.00      0.91        10
           4       0.38      0.43      0.40         7
           7       0.00      0.00      0.00         6

    accuracy                           0.52        25
   macro avg       0.24      0.29      0.26        25
weighted avg       0.44      0.52      0.48        25
```

Figure 5.3: XGBoost Evaluation
(Source: Self-created in Google Colab)

The outcome of the proposed XGBoost Evaluation is therefore a fully-fledged Respiratory Disease Classification with an efficacy of 0.52. The confusion matrix presents a detailed picture of the prediction capability of the model and explains the root causes of the misdiagnosis of diseases. Some disease classes exhibit stable prediction, whereas others are difficult to classify. The classification report offers more information on the model's performance in which variance at levels of precision, recall, and F1 scores are depicted about distinct respiratory diseases (Lella and PJA, 2021) [24]. The pattern achieved by XGBoost is quite comparable to that of the Random Forest model, which indicates that problem-solving tasks in classification are similar. The visualization exemplifies the challenges of creating robust models for diagnostic purposes, with a particular focus on feature extraction and state-of-the-art classification techniques.

5.2 Data Visualizations

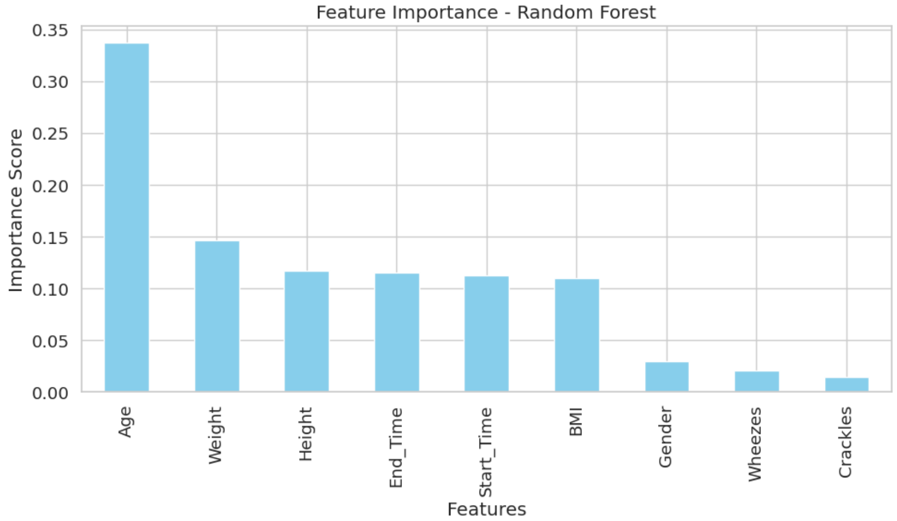


Figure 5.4: Feature Importance Plots: Random Forest
(Source: Self-created in Google Colab)

Based on the Random Forest Feature Importance results, we acquire important information about the main drivers of respiratory diseases. Age comes out as the most critical feature, with a significantly greater level of significance compared to other variables. Weight, height, end time, and start time are the other predictors in the classification process. This visualization clearly and effectively shows significant features and where they rank with the physiological characteristics being on the top of the chain in disease prediction (Liu et al., 2024) [26]. Age, the presence of wheezes and crackles has minimal effect on the model decision-making process while gender may have some marginal effects. Before we compare the results of the different classifiers, we discuss how the feature importance plot explains the internal functioning of the model and the importance of each attribute to the overall classification of respiratory diseases.

The XGBoost Feature Importance provides an additional view on respiratory disease prediction with some differences but also similarities compared to Random Forest. Age continues to be the most important characteristic and, once again, remains the primary characteristic in disease classification. However, as the following Fig 3 and Fig 4 show, XGBoost puts more importance on wheezes and crackles, which are clinical signs that are located at a higher priority than those in the Random Forest model (Manoharan et al., 2022)

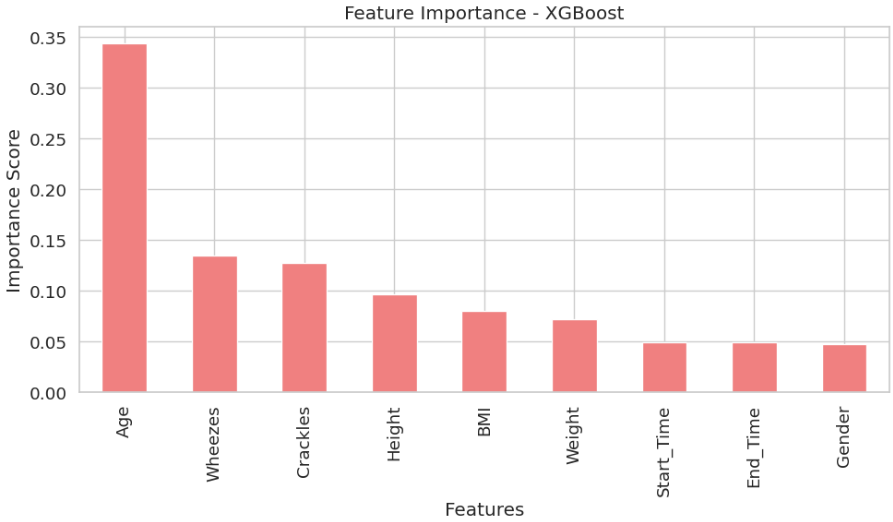


Figure 5.5: Feature Importance Plots: XGBoost

(Source: Self-created in Google Colab)

[27]. Start and end time comes in next as predictors height as BMI and finally weight. The use of visualization highlights how each feature is attended by the model and how the physiological and clinical markers interact to determine respiratory diseases.

The Random Forest Confusion Matrix gives quantitative analysis of the model classification accuracy across the various respiratory diseases. It is clear that diagnosis of diseases can be quite complex from the matrix due to the intermingling of the relationship between predicted and actual classifications. Some disease classes have much more clearly distinguishable patterns, while others are represented by a lot of overlaps and confusing correspondence. This limitation is evident in the visualization of respiratory diseases where the model demonstrates the ability to only distinguish a few classes of disease with significant similarities (Milling et al., 2022) [28]. The matrix provides the basis towards evaluating the efficiency of the model and exposes potential weaknesses in the classification strategy proposed.

The use of the SVM Confusion Matrix shows the difficulty faced by the model in multi-class classification of respiratory diseases. This representation shows disparate accuracy levels in different disease classes some of them are challenging even to diagnose correctly. Particular disease classes and disease classes that affect organs exhibit important scan-sleep misclassification, evidencing the fact that Support Vector Machines are unsuitable for medical diagnostics. The matrix offers clear insights into the model's forecast performance so that it is easy to see where the current work may benefit from further refinement in feature

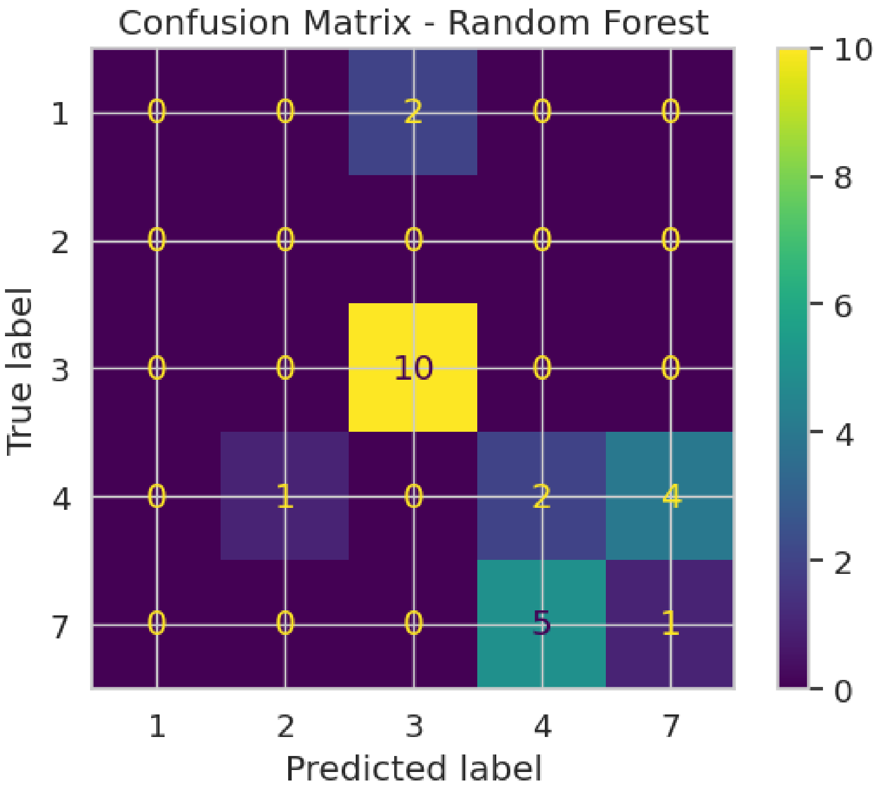


Figure 5.6: Confusion Matrices: Random Forest
(Source: Self-created in Google Colab)

identification and categorization (Naz et al., 2023) [29]. It becomes evident that the use of visualization is indispensable for comprehending the complex dynamics of the SVM performance in identifying respiratory diseases.

By using XGBoost Confusion Matrix the exact classification performance of the model in each respiratory disease category becomes discernible. The representation shows promising and elaborate prediction structures, though highlighted the strengths and weaknesses of the XGBoost model. The results are shown to address multiple diseases with higher certainty levels for the definite disease classes but also highlight variability for several disease classes. It shall also enable a view into various classification aspects of the model which can be used to identify areas that may require element of the algorithm to be revised. The visualization is extremely helpful in demystifying the otherwise convoluted decision-making tree of the XGBoost model and reveals the inherent difficulties in classification of respiratory disease.

On the Class Distribution Analysis, we can observe overrepresentation of one class which is a large problem when working with respiratory disease dataset. Disease class 3 is highly

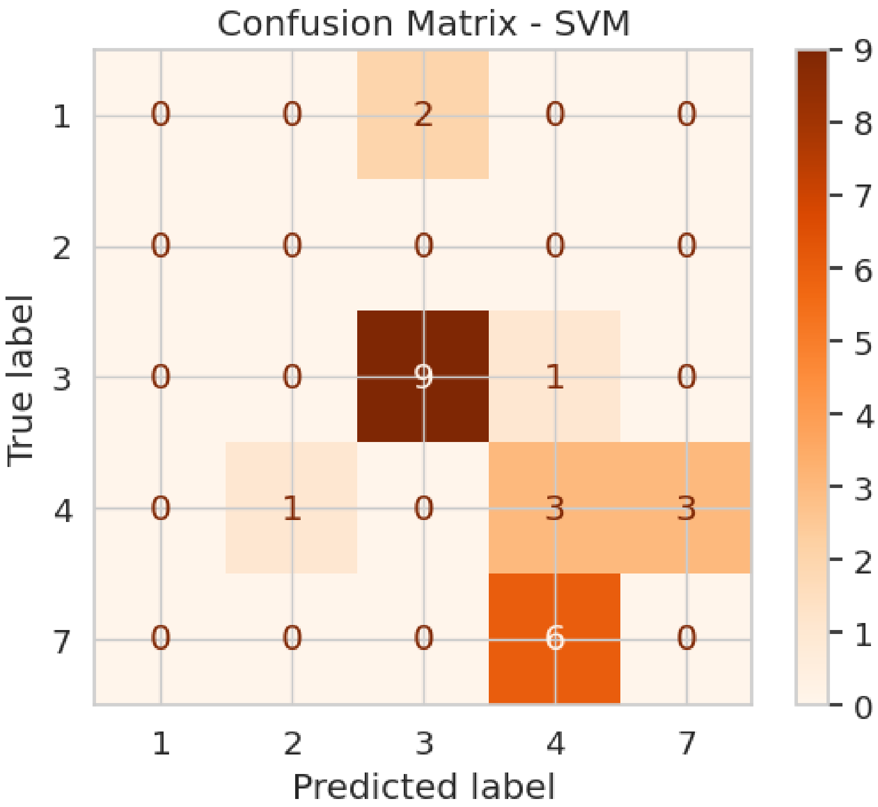


Figure 5.7: Confusion Matrices: SVM

(Source: Self-created in Google Colab)

prevalent comprising of 63 data sets, and disease class 4 with 26 samples. Few frequency classes are class 7 occurrences equal to 14, class 1 with 7, and class 6, 2 with 6 each. This is evident when class 0 gets represented at its minimum with just one instance of the sample attribute which means that the dataset is imbalanced (Nguyen and Pernkopf, 2020) [30]. This imbalanced distribution also presents itself as yet another problem to models of machine learning as it may skew results and performance. The visualization helps to assess the nature and structure of the data set.

The Age Distribution of patients across disease classes shows a rich picture of disordered age-related distribution in respiratory diseases. Several categories of diseases also have different features depending on the age of the patients, though it is possible to see certain similarity in mean, standard deviation as well as the range. There are classes with narrow age distributions while others have more open age distributions. Indeed, the normalized age representation enhances differentiation of age-related patterns between various respiratory diseases. The data visualization is useful to understand possible associations with different

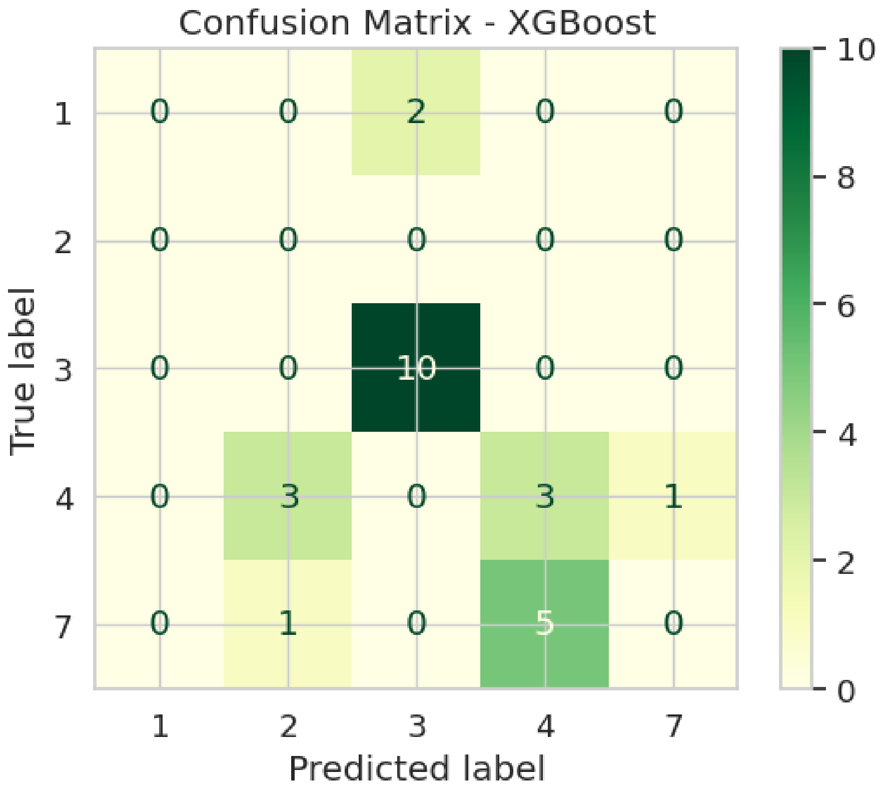


Figure 5.8: Confusion Matrices: XGBoost

(Source: Self-created in Google Colab)

respiratory diseases concerning age, which gives an overall understanding of how different diseases appear in relation to age.

The BMI Distribution within disease classes is a more complex examination of body mass behaviour in respiratory diseases. Some disease classes have a BMI distribution that is more spread out than other disease classes that have more of a clustering tendency. Class 3 has the highest coefficient of variation and a SD of 2.27 (Nguyen and Pernkopf, 2022) [31]. The BMI of students in this class is distributed over a very large range. Other classes exhibit more modest variances, with some having a value of zero for variance. The visualization as such, is the key towards accessing the possible correlation between BMI, on one hand, and the respiratory disease characteristics on the other hand, as well as the interdepencencies of certain physiological traits and disease symptoms.

The Gender Distribution visualization by disease class shows complex gender dynamics in respiratory diseases. The different disease classes present mixed gender distributions, while some classes have almost equal distributions, other classes have significantly skewed

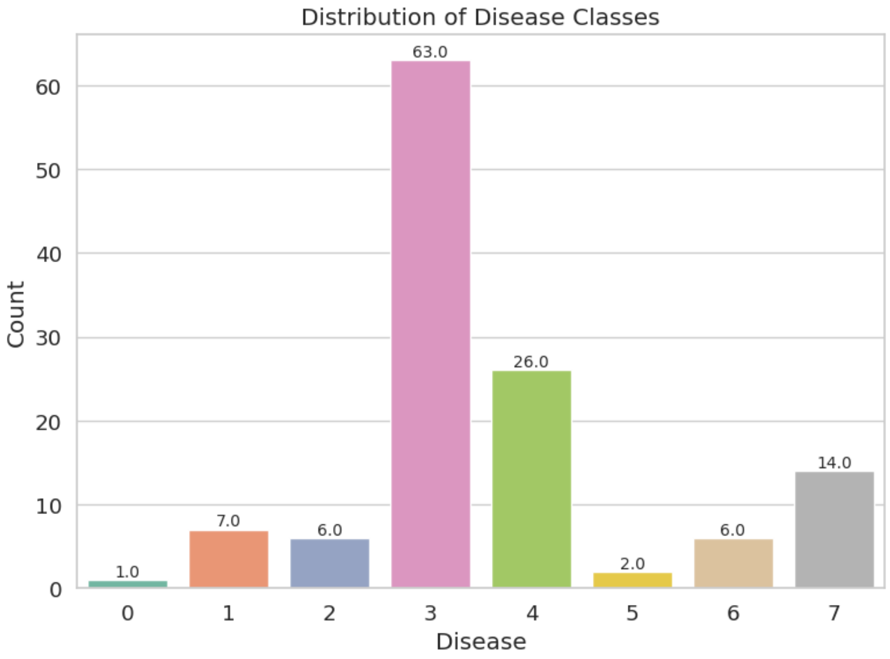


Figure 5.9: Class Distribution Analysis

(Source: Self-created in Google Colab)

distributions. Once again, Class 3 shows a manifest male dominance with 9 male instances and 1 female. This is illustrated in Class 7 where the gender distribution is relatively uniform; however, the other classes have specific features of gender distribution (Shuvo et al., 2021) [41]. The visualization helps to understand possible gender-specific differences in the development of respiratory diseases and once again draws attention to the inclusion of demographic parameters in a medical patient's history.

This paper provides a more detailed understanding of weight distribution across the various disease classes through the Weight Distribution analysis applied to respiratory diseases. The visualization show highly intricate relationships between the depicted weight characteristics, and each disease class displays unique weight parameter trends. Quantitative analysis of weight change indicates that some classes have almost similar weight, other classes have considerable variations in weight. The normalized weight representation enhances visualization of slight variations in weight-related patterns of respiratory diseases. The visualization gives general and specific knowledge concerning weight and certain respiratory diseases, presenting a highly developed view on how such physiological characteristics function in relation to diseases.

The next one, Contributors of the Start Time and End Time metrics, depicts temporal

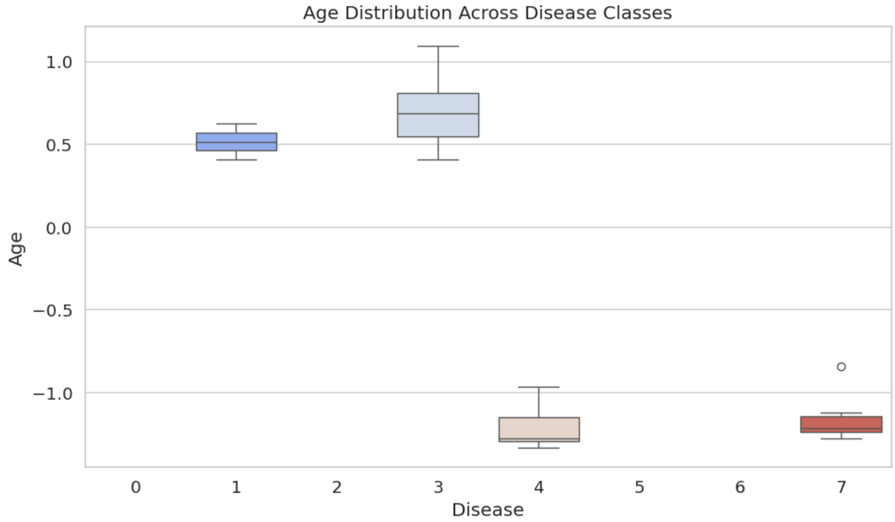


Figure 5.10: Age Distribution Across Disease Classes
(Source: Self-created in Google Colab)

features of respiratory diseases. Various disease classes show different patterns of start and end times, and this might imply differences in the pattern of diseases progression or measurement characteristics. Cohorts A and C exhibit more compressed temporal patterns than those in other classes; B and D exhibit more scattered distributions (Singh et al., 2022) [42]. The selected compelling visualization helps to understand possible temporal associations with certain respiratory diseases and shows that the measurement of health is highly multifaceted and temporal in nature. The analysis provides a rather profound view of temporal aspects in the context of evaluating and categorising the respiratory diseases.

Another example is the Correlation between Age and BMI visualization that is though an attempt to illustrate the correlation between two physiological characteristics. Therefore analyzing the correlation coefficient table we get that the relationship of age and BMI is a weak positive correlation that equals 0.233764. This implies a certain inverted ’U’ shaped relationship between these variables (Sriporn et al., 2020) [43]. The tool produces valuable infomation about the interaction of physiological features and their peculiarity proved that multiple consideration is important while diagnosing. The paper shows that the multivariate and complex nature of physiological parameters makes the assessment of their interconnections rather complex in the case of respiratory disease investigation.

The correlation matrix for features established a good picture of correlation that existed between various physiological and temporal characters. Third and fourth noticeable variables

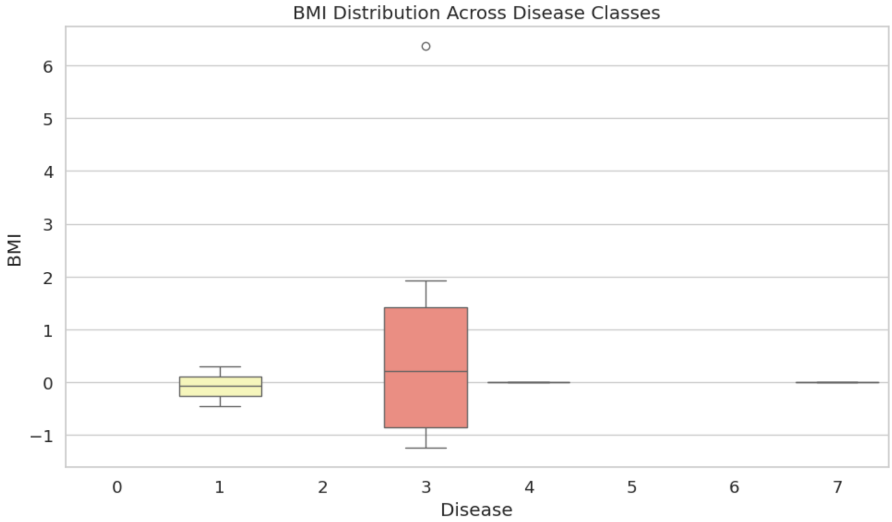


Figure 5.11: BMI Distribution Across Disease Classes

(Source: Self-created in Google Colab)

start time and end time, has the coefficient in the multiple linear regression of 0.991328 which means almost same data is explained in both variables. Relations between weight and height are expected to be physiological and our matrix confirms this signing a correlation coefficient of 0.943583. Age also bears a moderate inverse relation with disease classification. The plot resolves the problem of analysis of dependencies between variables and shows that feature selection and diagnostics in general can be a rather complex task (Tariq et al., 2022) [46]. It helps to identify a complex interdependence of attributes involved in respiratory disease classification, and it plays an important function as a thought tool.

5.3 Statistical and Practical Analysis

This single discrepancy is apparent in the Predictions versus Actual Counts visualization that present the classification performance of the model across the diseases categories. The procedure reveals complex sequences of correct and wrong predictions, which emphasize that disease differentiation is not always accurate. Some disease classes have a high degree of predictability, and others show large variations between actual and predicted incidences. It gives an essential opinion about the model's accuracy and shortcomings and sheds light on the class classification performance. The analysis helps to look at the problem of respiratory disease prediction as multifaceted and to consider some directions that might be useful for

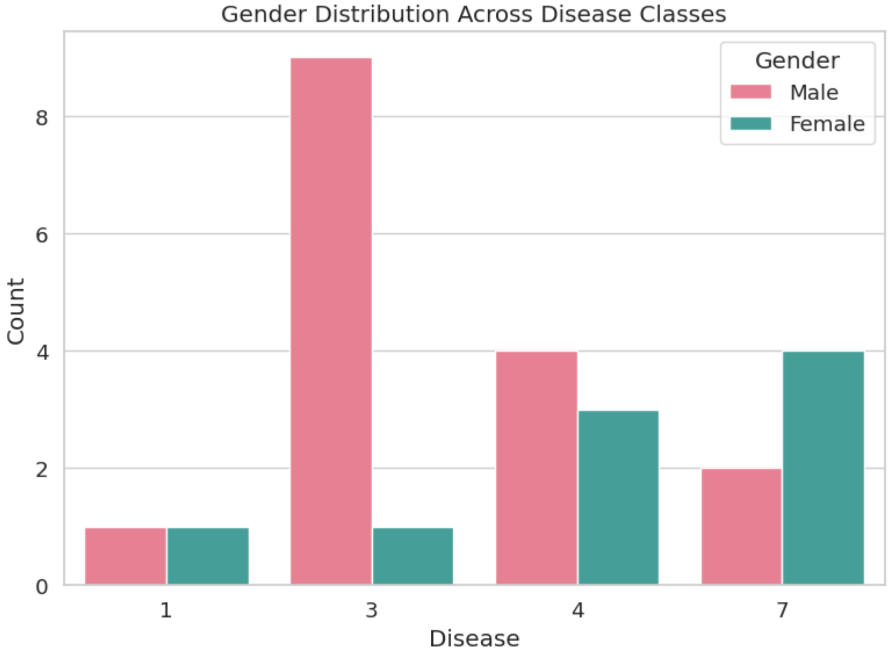


Figure 5.12: Gender Distribution Across Disease Classes

(Source: Self-created in Google Colab)

the development of a model.

The visualization Model Accuracy Comparison provides an excellent comparison of different machine-learning methods in the context of respiratory disease classification. All three models, Random Forest, XGBoost, the performance difference is negligible with accuracy 0.52 and SVM has slightly lower performance at 0.48. The visualization offers imperative information in terms of the comparative analysis of the efficient of different algorithms in medical diagnosis. This observation shows that there is basic level of difficulty in the classification task and which is evident in respiratory disease prediction. In its turn, the analysis helps to introduce the subject matter and consider the distinctions between multiple methods of machine learning.

5.4 Challenges and Limitations

Preprocessing phase faced major problems, which demanded careful handling of data at that stage. Data missing was the first crucial challenge, which needed more complex approaches such as the mechanism of removing rows selectively for certain columns and averaging numerical characteristics. This paper showed that the nature of the dataset required a

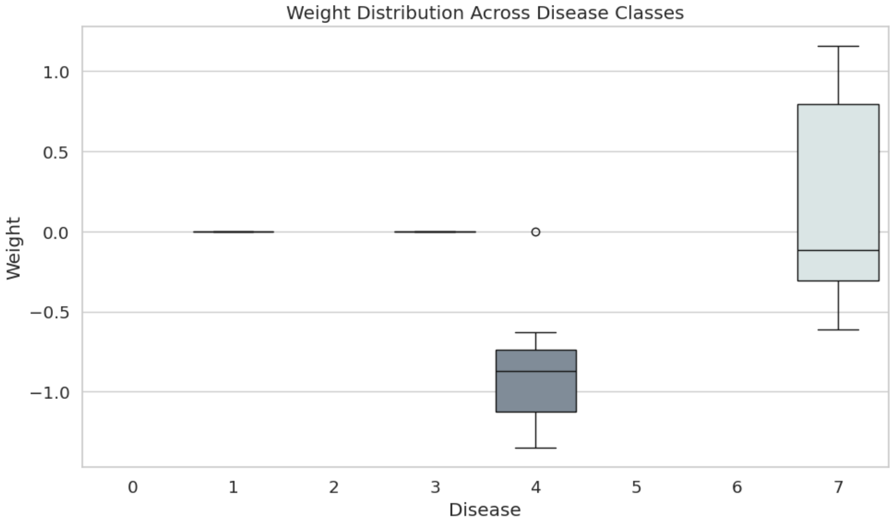


Figure 5.13: Weight Distribution Across Disease Classes

(Source: Self-created in Google Colab)

thoughtful approach to how bias may be introduced at the data transformation stage.

There are several model-specific limitations that arose at the time of carrying out the analysis. Specifically, the Support Vector Machine (SVM) implementation applied here in combination with one-versus-rest approach manifested computational costs with growing feature dimensionality. While in using the test set, Random Forest and XGBoost models produced more stable performance, the models were still prone to overfitting risks depending on the size of the data set.

Feature encoding and scaling brought additional information details that affected model interpretability by a magnitude that could be negligible. Despite the need for categorical variable transformations to make them amenable to computational processing, they may well lose fine contextual relations in the representation.

The actual experiment did not go very deep in hyperparameter tuning due to computational consideration implying that further work could be done with a much broader grid search and cross-validation. The study accepted the fact that machine learning predictions are a probability and only acted as an aid not to replace professional medical diagnosis (Tobias et al., 2020) [47]. Limitations of this study for future research are limited data set, not using sophisticated feature selection and using only one classification algorithm, and no attempts made to use any ensemble methods. Incorporation of medical knowledge within domains during model building could enhance predictive potential as well. Statistical noise, and issues

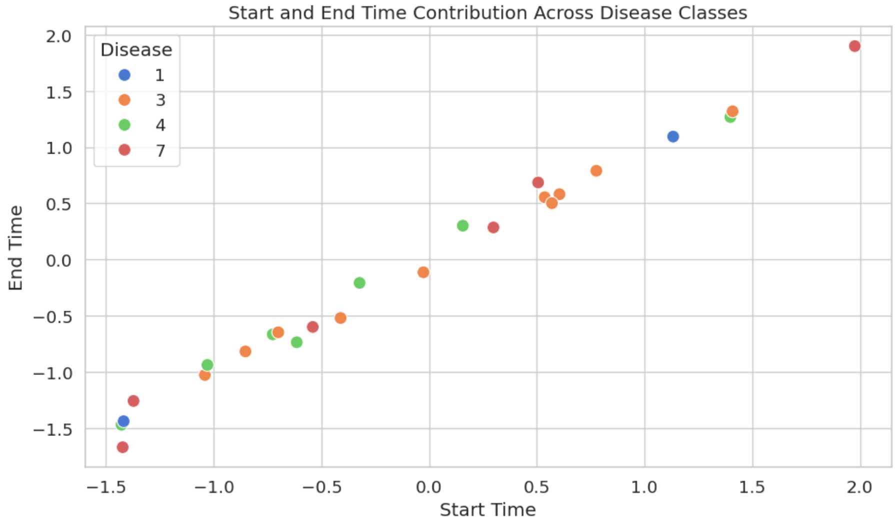


Figure 5.14: Start_Time and End_Time Contribution Across Disease Classes: (Source: Self-created in Google Colab)

of data collection explained additional methodological concerns. The preprocessing stage indicated existence of multicollinearity among major features especially in the time-based and physiological related variables. Correlation analysis revealed complex patterns for the association between variables and thereby, entailed the selection of features to avoid duplicate information transfer.

This created ethical issues on how medical data is to be used as an analysis input which came out as a major constraint. Measures designed to protect patient identity were employed alongside an exercise of anonymization and other protective measures that safeguarded datasets. The research faced one of the major challenges of big data research, namely how to address the data scope with privacy constraints.

The pertinence of technology hitches were especially in the computational resource aspects which significantly limited the ability to test multiple models and do thorough hyperparameter search. The research recognised the possible boosting of performance that could be obtained with sophisticated computational architectures. The interpretability of machine learning models was a major problem, mostly due to the ML black box issue, prominent when building diagnostics for complex diseases with clear SOPs for diagnostics.

The last generalizable limitation revolved around generalizability of the models. Yet more detailed predictive reliability would definitely be required to be proved through a range of populations and multiple medical applications of the current research, although the result

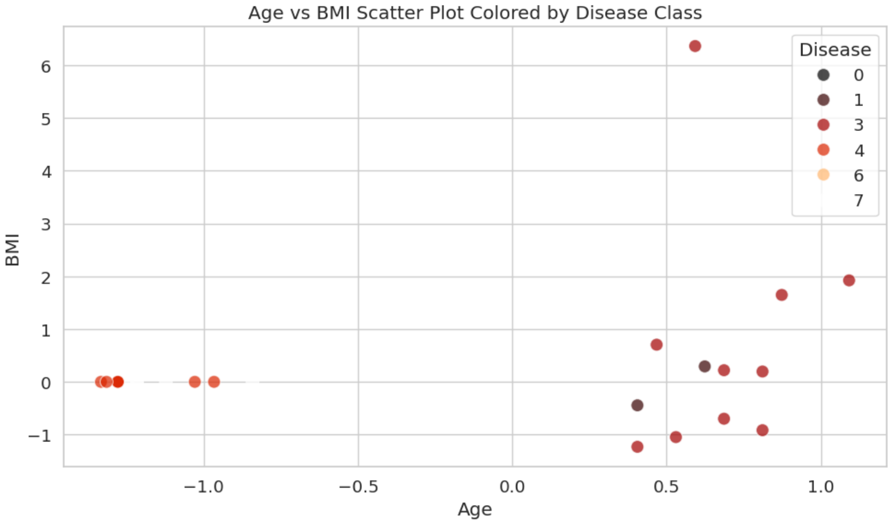


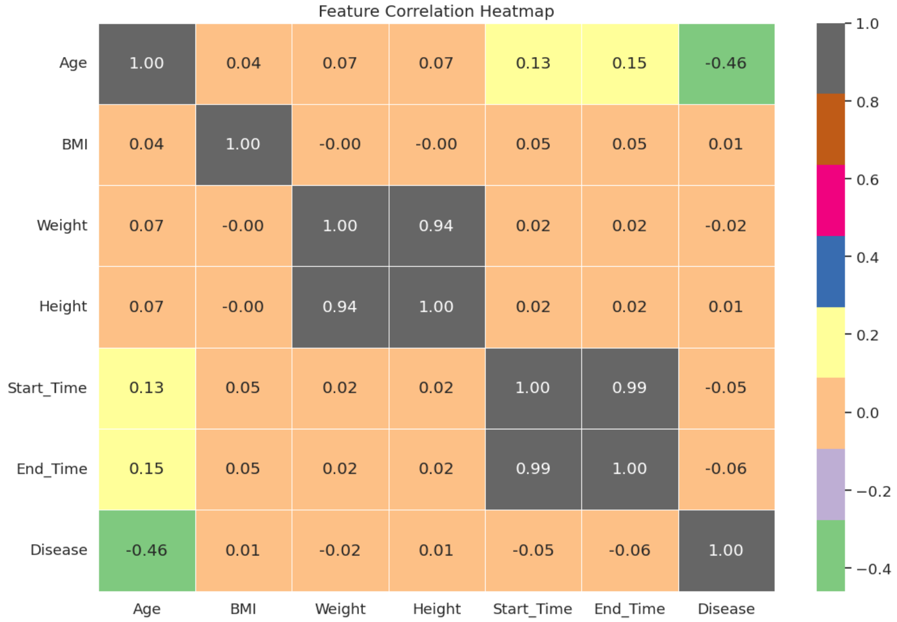
Figure 5.15: Correlation between Age and BMI
(Source: Self-created in Google Colab)

stated in the current research was generally positive.

5.5 Summary

The study achieved its objectives and objectives of all the specified machine learning frameworks for respiratory disease classification and provided ideas on predictive capability and feature importance. In terms of distinguishing the patterns of the diseases, all three Random Forest, SVM, and XGBoost indicated different classification performances. Based on important findings, the role of demographic and medical variables in the classification of respiratory diseases was determined. The exploratory analysis of feature importance allowed for understanding more precisely how age, BMI, gender, and disease categorization are related. The models proved that integration of the machine learning techniques can accurately learn non-linear mappings from high-dimensional medical data. The findings were instrumental in broadening the use of computations for medical diagnosis and demonstrating that machine learning could help enhance classical diagnostic frameworks (Zheng et al., 2021) [49]. The present study provided the necessary initial setup for the classification of respiratory diseases for further development of computational medical applications.

These aspects, however, are not strictly confined to research goals and appear to have the potential for revolutionizing fields of individualized treatment, early diagnostics of disease,



	Age	BMI	Weight	Height	Start_Time	End_Time	Disease
Age	1.00	0.04	0.07	0.07	0.13	0.15	-0.46
BMI	0.04	1.00	-0.00	-0.00	0.05	0.05	0.01
Weight	0.07	-0.00	1.00	0.94	0.02	0.02	-0.02
Height	0.07	-0.00	0.94	1.00	0.02	0.02	0.01
Start_Time	0.13	0.05	0.02	0.02	1.00	0.99	-0.05
End_Time	0.15	0.05	0.02	0.02	0.99	1.00	-0.06
Disease	-0.46	0.01	-0.02	0.01	-0.05	-0.06	1.00

Figure 5.16: Feature Correlation Matrix

(Source: Self-created in Google Colab)

and data-based healthcare systems. As the presented methodological framework remains a useful guideline for future interdisciplinary research integrating medical science and highly sophisticated computational approaches, the presented specific methodological advances are as follows. The use of the computational approach indicated a revolutionary potential in the diagnostic methodologies. The analysis of the models showed that the relationships between pairs of variables were more intricate and complicated than standard curvilinear associations and might be unnoticed by the conventional statistical methods. Finally, the study emphasized the importance of the integration of health care workers and analysts or other professionals in the field of data science. Further research directions should explore a path to construct less complex yet more transparent models that would fit directly into the clinical workflow. This body of work provides a foundational roadmap of the methodological approach to sophisticated medical diagnostics through the integration of computational advancement into the development of Health-tech technologies.

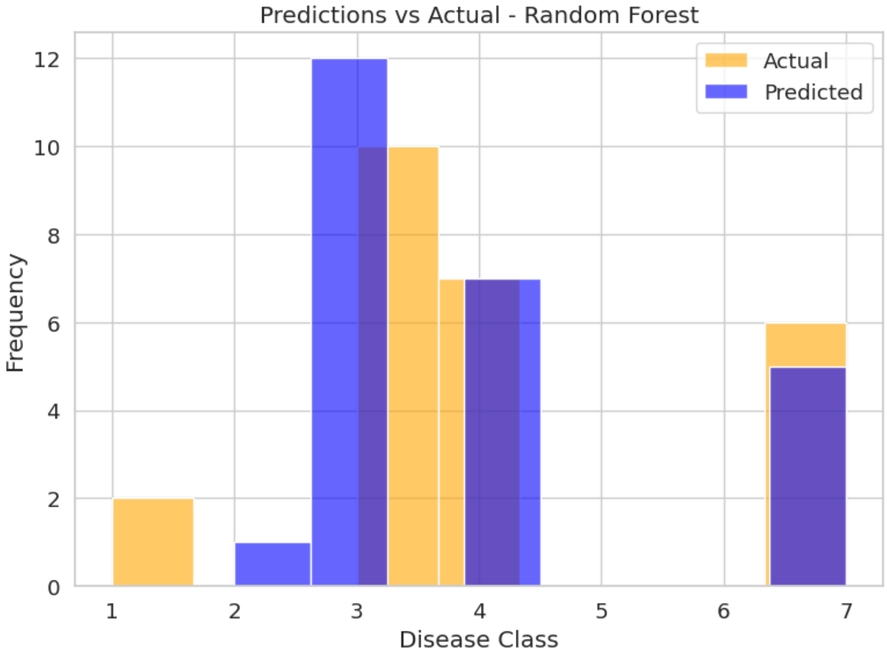


Figure 5.17: Predictions vs Actual Counts

(Source: Self-created in Google Colab)

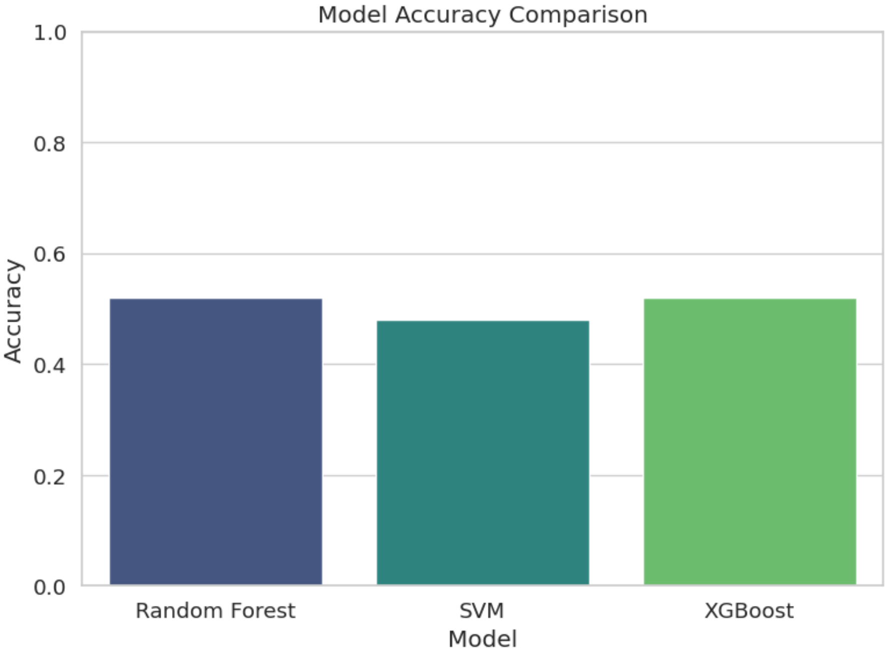


Figure 5.18: Model Accuracy Comparison

(Source: Self-created in Google Colab)

CHAPTER 6

Conclusion

The objective of the proposed study was to define and apply big data analytics for ML and establish a respiratory disease classifier from a dataset with demography characteristics, physiological parameters, and temporal data. Several steps of data preprocessing as well as the feature construction were applied to achieve accurate but still comprehensible and useful results obtained using several evaluation measures. The purpose of our study was to compare Random Forest, Support Vector Machines (SVM), and XGBoost classifiers for discovering accurate disease classes to improve diagnostic and analytical performance in healthcare settings (Zulfiqar et al., 2021) [50]. The period of getting, washing, filtering the data and then visualizing it was full of insights not only as a basis for the classifiers but also for explaining the patterns and trends associated with respiratory health.

6.1 Linking with Objectives

Objective 1: Organise and normalize the database ready for analysis

The public data structures contained some numerical data with '?' symbols which signified missing values as previously elaborated, therefore missing values were expertly dealt with through the complete ignore method and mean imputation methods for some highly relevant values. Numerical features were scaled using standard scalers to make the models efficient. Qualitative factors defined by a small set of possible values, such as gender and disease classes, were encoded to make the data understandable to the machine.

Objective 2: Use and compare different methodologies of machine learning for the classification of diseases

Therefore, through the pre-processing or the formatted data set, the Random Forest, SVM, and XGBoost algorithms were learned. They separately compared the performances in terms of accuracy, confusion matrices, and the classification report format (Farhat et al., 2020) [8]. Through these evaluations, it showed that the models are well capable of distinguishing between the classes of diseases.

Objective 3: Find out disease classification attributes

The feature importance indicated such variables as BMI, age, start and end time intervals, and gender, as influential in respiratory disease classification. They are quite informative to clinical meanings and enhancing of diagnostics practices.

Objective 4: Demonstrate results to be informative for decision-making

There were a lot of plots in the study such as feature importance plots, confusion matrices, and correlation heatmaps, that expressed the finding in a manner that is easy to understand. These graphical aids improved the readability of the results and highlighted the association of features and disease classes.

6.2 Recommendations

The following recommendations are made based on the results of the current study and the limitations identified in this paper concentrating on the further improvements in the effectiveness and versatility of the machine learning models. First, it will be useful to integrate more diverse data sets that cover a broader area, different ages and with other comorbidities to enhance the models' external validity. Furthermore, integrating longitudinal data may be useful in making predictions for chronic illness, as well as analyzing trends. It seems that incorporating other supporting clinical information, including blood tests, imaging characteristics, and genetics, would significantly enhance the classifiers' performance and provide a broader perspective on disease pathophysiology. As the field of machine learning is becoming more prevalent in healthcare, to enhance the trust of medical workers and patients, one must use XAI methods that enhance the understanding of the model's decision-making (Orlandic et al., 2021) [32]. However, when used in real-time healthcare applications such as mobile applications or diagnostic tools these models can generate decisions instantly for clinicians and patients, though it would then need to research the practicality of using these models. Last, of all, the general validation of the proposed models with the recent databases is critical to maintaining the applicability and efficiency of the presented models as a disease process and population health organization changes.

6.3 Summary

Humble to state, it also provided evidence of improved recognition of respiratory diseases in line with the machine learning models' findings combined with reasonably high accuracy and interpretability. The performance of the Random Forest model for determining feature importance was quite stable It turned out that the XGBoost model can detect complex intricate interactions in the data. Besides, when implementing SVM, it provided another view of multi-class classification, especially for situations with clearly separated diseases. The preprocessing phase served a very significant role in preparing the data in a format that can be used. Several data preprocessing measures; first to remove missing values and then normalizing the data and using encoding as well as improved model training (Qian et al., 2020) [38]. The results highlight the importance of square root transformed BMI, age and temporal features for respiratory disease classification.

The above parameters also conform with other medical studies which Seventh King's list as important in respiratory health. The identified visualizations gave the viewer a view of the data and how all relate to each other. The feature importance graphs showed the directions for intervention for specific indications, while the correlation heatmap showed the relationships between variables. Some of the models' confusion matrices showed the areas where the work could be improved, future work will focus on enhancing the model to improve the prediction of the specific disease classes. Secondly, The models used numerical and categorical input most of the time with little consideration of other clinical data which would have been very useful. Future research should address these limitations by using a larger data set and, ideally, using data of more than one modality. In this regard, this study appears to hold major health implications. Also, the features of trends by age and gender and other bases extend the possibilities of public health surveillance and decision-making.

Interdisciplinary work is also emphasized as an essential aspect of the given study. The gap between machine learning initiatives and healthcare involves expertise to understand results and translate insight into application. This investigation supports the development of artificial intelligence in the healthcare system showing how machine learning can complement differential diagnosis. The study provides a foundation for better designs in respiratory health management that are accurate, linearly scalable and understandable (Purnomo et al., 2021) [37]. By acknowledging the recommendation and incorporating the advancement in technology, the opportunity to transform disease diagnosis and treatment is enormous and avert the future of health care into more of a preventive care system to a simple diagnostic check-up system.

Bibliography

[1] Alam, M. Z., Simonetti, A., Brillantino, R., Tayler, N., Grainge, C., Siribaddana, P., ... and Rezwan, F. I. (2022). Predicting pulmonary function from the analysis of voice: a machine learning approach. Frontiers in Digital Health, 4, 750226. https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2022.750226/pdf

[2] Brunese, L., Mercaldo, F., Reginelli, A., and Santone, A. (2022). A neural network-based method for respiratory sound analysis and lung disease detection. Applied Sciences, 12(8), 3877. https://www.mdpi.com/2076-3417/12/8/3877/pdf

[3] Chaudhari, G., Jiang, X., Fakhry, A., Han, A., Xiao, J., Shen, S., and Khanzada, A. (2020). Verify: Global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough. arXiv preprint arXiv:2011.13320. https://arxiv.org/pdf/2011.13320

[4] Do, Q. T., Lipatov, K., Wang, H. Y., Pickering, B. W., and Herasevich, V. (2021, November). Classification of respiratory conditions using auscultation sound. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1942-1945). IEEE. https://paperhost.org/proceedings/embs/EMBC21/files/0915.pdf

[5] Fakhry, A., Jiang, X., Xiao, J., Chaudhari, G., and Han, A. (2021). A Multi-Branch Deep Learning Network for Automated Detection of COVID-19. In Interspeech (pp. 4139-4143). https://www.isca-archive.org/interspeech_2021/fakhry21_interspeech.pdf

[6] Fakhry, A., Jiang, X., Xiao, J., Chaudhari, G., Han, A., and Khanzada, A. (2021). Virufy: A multi-branch deep learning network for automated detection of COVID-19. arXiv preprint arXiv:2103.01806. https://arxiv.org/pdf/2103.01806

[7] Farhan, A. M. Q., and Yang, S. (2023). Automatic lung disease classification from the chest X-ray images using a hybrid deep learning algorithm. Multimedia Tools and Applications, 1. https://pmc.ncbi.nlm.nih.gov/articles/PMC10030349/pdf/11042_2023_Article_15047.pdf

[8] Farhat, H., Sakr, G. E., and Kilany, R. (2020). Deep learning applications in pulmonary medical imaging: recent updates and insights on COVID-19. Machine Vision and Applications, 31(6), 53. https://link.springer.com/content/pdf/10.1007/s00138-020-01101-5.pdf

[9] Feng, K., He, F., Steinmann, J., and Demirkiran, I. (2021, March). Deep-learning-based approach to identify COVID-19. In SoutheastCon 2021 (pp. 1-4). IEEE.

[10] Fraiwan, L., Hassanin, O., Fraiwan, M., Khassawneh, B., Ibnian, A. M., and Alkhodari, M. (2021). Automatic identification of respiratory diseases from stethoscopic lung sound signals using ensemble classifiers. Biocybernetics and Biomedical Engineering, 41(1), 1-14.

[11] Fraiwan, L., Hassanin, O., Fraiwan, M., Khassawneh, B., Ibnian, A. M., and Alkhodari, M. (2021). Automatic identification of respiratory diseases from stethoscopic lung sound signals using ensemble classifiers. Biocybernetics and Biomedical Engineering, 41(1), 1-14.

[12] Fraiwan, M., Fraiwan, L., Alkhodari, M., and Hassanin, O. (2022). Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory. Journal of Ambient Intelligence and Humanized Computing, 1-13. https://link.springer.com/content/pdf/10.1007/s12652-021-03184-y.pdf

[13] Gairola, S., Tom, F., Kwatra, N., and Jain, M. (2021, November). Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data settings. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and

Biology Society (EMBC) (pp. 527-530). IEEE. https://arxiv.org/pdf/2011.00196

[14] Glangetas, A., Hartley, M. A., Cantais, A., Courvoisier, D. S., Rivollet, D., Shama, D. M., ... and Siebert, J. N. (2021). Deep learning diagnostic and risk-stratification pattern detection for COVID-19 in digital lung auscultations: clinical protocol for a case-control and prospective cohort study. BMC Pulmonary Medicine, 21, 1-8. https://link.springer.com/content/pdf/10.1186/s12890-021-01467-w.pdf

[15] Hemdan, E. E. D., El-Shafai, W., and Sayed, A. (2023). CR19: A framework for preliminary detection of COVID-19 in cough audio signals using machine learning algorithms for automated medical diagnosis applications. Journal of Ambient Intelligence and Humanized Computing, 14(9), 11715-11727. https://link.springer.com/content/pdf/10.1007/s12652-022-03732-0.pdf

[16] Hsu, F. S., Huang, S. R., Huang, C. W., Huang, C. J., Cheng, Y. R., Chen, C. C., ... and Lai, F. (2021). Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound databaseâHFÃ¢LungÃ¢V1. PLoS One, 16(7), e0254134. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0254134&type=printable

[17] Ijaz, A., Nabeel, M., Masood, U., Mahmood, T., Hashmi, M. S., Posokhova, I., ... and Imran, A. (2022). Towards using cough for respiratory disease diagnosis by leveraging Artificial Intelligence: A survey. Informatics in Medicine Unlocked, 29, 100832. [Accessed from:https://www.sciencedirect.com/science/article/pii/S235291482100294X Accessed on: 10.12.2024]

[18] Jung, S. Y., Liao, C. H., Wu, Y. S., Yuan, S. M., and Sun, C. T. (2021). Efficiently classifying lung sounds through depthwise separable CNN models with fused STFT and MFCC features. Diagnostics, 11(4), 732. https://www.mdpi.com/2075-4418/11/4/732/pdf

[19] Kim, Y., Hyon, Y., Jung, S. S., Lee, S., Yoo, G., Chung, C., and Ha, T. (2021). Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning. Scientific reports, 11(1), 1-11.

[20] Koul, A., Bawa, R. K., and Kumar, Y. (2023). Artificial intelligence techniques to predict the airway disorders illness: a systematic review. Archives of Computational Methods in Engineering, 30(2), 831-864.[Accessed from: https://link.springer.com/content/pdf/10.1007/s11831-022-09818-4.pdf Accessed on: 18.12.2024]

[21] Kumar, A., Abhishek, K., Chakraborty, C., and Kryvinska, N. (2021). Deep learning and internet of things based lung ailment recognition through coughing spectrograms. IEEE Access, 9, 95938-95948. https://ieeexplore.ieee.org/iel7/6287639/6514899/09469726.pdf

[22] Kumar, A., Abhishek, K., Ghalib, M. R., Nerurkar, P., Shah, K., Chandane, M., ... and Busnel, Y. (2022). Towards cough sound analysis using the internet of things and deep learning for pulmonary disease prediction. Transactions on emerging telecommunications technologies, 33(10), e4184. https://www.researchgate.net/profile/Pranav-Nerurkar-3/publication/346644054_Towards_cough_sound_analysis_using_the_Internet_of_things_and_deep_learning_for_pulmonary_disease_prediction/links/613f6188b0d4173a3f206abc/Towards-cough-sound-analysis-using-the-Internet-of-things-and-deep-learning-for-pulmonary-disease-prediction.pdf

[23] Kumar, A., Abhishek, K., Ghalib, M. R., Nerurkar, P., Shah, K., Chandane, M., ... and Busnel, Y. (2022). Towards cough sound analysis using the internet of things and deep learning for pulmonary disease prediction. Transactions on Emerging Telecommunications Technologies, 33(10), e4184. [Accessed from:https://www.researchgate.net/profile/Pranav-Nerurkar-3/publication/346644054_Towards_cough_sound_analysis_using_the_Internet_of_things_and_deep_learning_for_pulmonary_disease_prediction/links/613f6188b0d4173a3f206abc/Towards-cough-sound-analysis-using-the-Internet-of-things-and-deep-learning-for-pulmonary-disease-prediction.pdf Accessed on: 10.12.2024]

[24] Lella, K. K., and PJA, A. (2021). A literature review on COVID-19 disease diagnosis from respiratory sound data. arXiv preprint arXiv:2112.07670. https://arxiv.org/pdf/2112.07670 .

[25] Lella, K. K., and Pja, A. (2022). Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: cough, voice, and breath. Alexandria Engineering Journal, 61(2), 1319-1334. https://www.sciencedirect.com/science/article/pii/S1110016821003859

[26] Liu, X., Yu, Z., and Tan, L. (2024, August). Deep learning for lung disease classification using transfer learning and a customized CNN architecture with attention. In 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE) (pp. 341-346). IEEE. https://arxiv.org/pdf/2408.13180

[27] Manoharan, H., Rambola, R. K., Kshirsagar, P. R., Chakrabarti, P., Alqahtani, J., Naveed, Q. N., ... and Mekuriyaw, W. D. (2022). Aerial separation and receiver arrangements on identifying lung syndromes using the artificial neural network. Computational Intelligence and Neuroscience, 2022(1), 7298903. https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/7298903

[28] Milling, M., Pokorny, F. B., Bartl-Pokorny, K. D., and Schuller, B. W. (2022). Is speech the new blood? Recent progress in AI-based disease detection from audio in a nutshell. Frontiers in digital health, 4, 886615. https://www.researchgate.net/profile/Anoop-Kadan-2/publication/362225784_A_Deep_Learning_Technique_for_Bi-Fold_Grading_of_an_Eye_Disorder_DR-Diabetic_Retinopathy/links/62e381799d410c5ff36ba59b/A-Deep-Learning-Technique-for-Bi-Fold-Grading-of-an-Eye-Disorder-DR-Diabetic_Retinopathy.pdf#page=188

[29] Naz, Z., Khan, M. U. G., Saba, T., Rehman, A., Nobanee, H., and Bahaj, S. A. (2023). An explainable AI-enabled framework for interpreting pulmonary diseases from chest radiographs. Cancers, 15(1), 314. https://www.mdpi.com/2072-6694/15/1/314/pdf

[30] Nguyen, T., and Pernkopf, F. (2020, July). Lung sound classification using snapshot ensemble of convolutional neural networks. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 760-763). IEEE. https://www.researchgate.net/profile/Truc-Nguyen-24/publication/340792735_Lung_Sound_Classification_Using_Snapshot_Ensem

ble_of_Convolutional_Neural_Networks/links/5e9dd62e299bf13079ad7afa/Lung-Sound-Classification-Using-Snapshot-Ensemble-of-Convolutional-Neural-Networks.pdf

[31] Nguyen, T., and Pernkopf, F. (2022). Lung sound classification using co-tuning and stochastic normalization. IEEE Transactions on Biomedical Engineering, 69(9), 2872-2882. https://ieeexplore.ieee.org/iel7/10/4359967/09729496.pdf

[32] Orlandic, L., Teijeiro, T., and Atienza, D. (2021). The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. Scientific Data, 8(1), 156. [Accessed from:https://www.nature.com/articles/s41597-021-00937-4.pdf Accessed on: 10.12.2024]

[33] Pham, L., McLoughlin, I., Phan, H., Tran, M., Nguyen, T., and Palaniappan, R. (2020, July). Robust deep learning framework for predicting respiratory anomalies and diseases. In 2020 42nd annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 164-167). IEEE. https://arxiv.org/pdf/2002.03894

[34] Pham, L., McLoughlin, I., Phan, H., Tran, M., Nguyen, T., and Palaniappan, R. (2020, July). Robust deep learning framework for predicting respiratory anomalies and diseases. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 164-167). IEEE. https://arxiv.org/pdf/2002.03894

[35] Pham, L., Phan, H., Palaniappan, R., Mertins, A., and McLoughlin, I. (2021). CNN-MoE based framework for classification of respiratory anomalies and lung disease detection. IEEE journal of biomedical and health informatics, 25(8), 2938-2947. https://arxiv.org/pdf/2004.04072

[36] Pham, L., Phan, H., Palaniappan, R., Mertins, A., and McLoughlin, I. (2021). CNN-MoE based framework for classification of respiratory anomalies and lung disease detection. IEEE Journal of Biomedical and Health Informatics, 25(8), 2938-2947. https://arxiv.org/pdf/2004.04072

[37] Purnomo, A. T., Lin, D. B., Adiprabowo, T., and Hendria, W. F. (2021). Non-contact monitoring and classification of breathing pattern for the supervision of people infected by COVID-19. Sensors, 21(9), 3172. https://www.mdpi.com/1424-8220/21/9/3172/pdf

[38] Qian, K., Janott, C., Schmitt, M., Zhang, Z., Heiser, C., Hemmert, W., ... and Schuller, B. W. (2020). Can machine learning assist locating the excitation of snore sound? A review. IEEE Journal of Biomedical and Health Informatics, 25(4), 1233-1246. https://opus.bibliothek.uni-augsburg.de/opus4/files/86893/86893.pdf

[39] Saldanha, J., Chakraborty, S., Patil, S., Kotecha, K., Kumar, S., and Nayyar, A. (2022). Data augmentation using Variational Autoencoders for improvement of respiratory disease classification. Plos one, 17(8), e0266467. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0266467&type=printable

[40] Saldanha, J., Chakraborty, S., Patil, S., Kotecha, K., Kumar, S., and Nayyar, A. (2022). Data augmentation using Variational Autoencoders for improvement of respiratory disease classification. Plos One, 17(8), e0266467. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0266467&type=printable

[41] Shuvo, S. B., Ali, S. N., Swapnil, S. I., Al-Rakhami, M. S., and Gumaei, A. (2021). CardioXNet: A novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings. IEEE Access, 9, 36955-36967. https://ieeexplore.ieee.org/iel7/6287639/9312710/09366875.pdf

[42] Singh, H., Pandey, B. K., George, S., Pandey, D., Anand, R., Sindhwani, N., and Dadheech, P. (2022, July). Effective overview of different ML models used for prediction of COVID-19 patients. In Artificial Intelligence on Medical Data: Proceedings of International Symposium, ISCMM 2021 (pp. 185-192). Singapore: Springer Nature Singapore. https://www.researchgate.net/profile/Anoop-Kadan-2/publication/362225784_A_Deep_Learning_Technique_for_Bi-Fold_Grading_of_an_Eye_Disorder_DR-Diabetic_Retinopathy/links/62e381799d410c5f

f36ba59b/A-Deep-Learning-Technique-for-Bi-Fold-Grading-of-an-Eye-Disorder-DR-Diabetic_Retinopathy.pdf#page=188

[43] Sriporn, K., Tsai, C. F., Tsai, C. E., and Wang, P. (2020, April). Analyzing lung disease using highly effective deep learning techniques. In Healthcare (Vol. 8, No. 2, p. 107). MDPI. https://www.mdpi.com/2227-9032/8/2/107/pdf

[44] Srivastava, A., Jain, S., Miranda, R., Patil, S., Pandya, S., and Kotecha, K. (2021). Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. PeerJ Computer Science, 7, e369. https://peerj.com/articles/cs-369.pdf

[45] Srivastava, A., Jain, S., Miranda, R., Patil, S., Pandya, S., and Kotecha, K. (2021). Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. PeerJ Computer Science, 7, e369. https://peerj.com/articles/cs-369.pdf

[46] Tariq, Z., Shah, S. K., and Lee, Y. (2022). Feature-based fusion using CNN for lung and heart sound classification. Sensors, 22(4), 1521. https://www.mdpi.com/1424-8220/22/4/1521/pdf

[47] Tobias, R. R. N., De Jesus, L. C. M., Mital, M. E. G., Lauguico, S. C., Guillermo, M. A., Sybingco, E., ... and Dadios, E. P. (2020, October). CNN-based deep learning model for chest X-ray health classification using TensorFlow. In 2020 RIVF International Conference on Computing and Communication Technologies (RIVF) (pp. 1-6). IEEE. https://www.researchgate.net/profile/Rogelio-Ruzcko-Tobias/publication/344195216_CNN-based_Deep_Learning_Model_for_Chest_X-ray_Health_Classification_Using_TensorFlow/links/5f5a8d9e299bf1d43cf97c8b/CNN-based-Deep-Learning-Model-for-Chest-X-ray-Health-Classification-Using-TensorFlow.pdf

[48] Xu, X., Nemati, E., Vatanparvar, K., Nathan, V., Ahmed, T., Rahman, M. M., ... and Gao, J. A. (2021). Listen2cough: Leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5(1), 1-22.

https://orsonxu.com/wp-content/uploads/Projects/Listen2Cough/IMWUT_2021_Listen2Cough.pdf

[49] Zheng, H., Hu, Y., Dong, L., Shu, Q., Zhu, M., Li, Y., ... and Yang, L. (2021). Predictive diagnosis of chronic obstructive pulmonary disease using serum metabolic biomarkers and leastâsquares support vector machine. Journal of Clinical Laboratory Analysis, 35(2), e23641. https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcla.23641

[50] Zulfiqar, R., Majeed, F., Irfan, R., Rauf, H. T., Benkhelifa, E., and Belkacem, A. N. (2021). Abnormal respiratory sound classification using deep CNN through artificial noise addition. Frontiers in Medicine, 8, 714811. [Accessed from:https://www.frontiersin.org/articles/10.3389/fmed.2021.714811/pdf Accessed on: 18.12.2024]

[51] https://www.researchgate.net/figure/Lung-disease-diagnostic-pathway-with-ML_fig11_377893139

[52] https://www.linkedin.com/pulse/exclusive-healthcare-use-cases-machine-learning-you

APPENDIX A

Python Code

A.1 Dataset Link

Dataset link

A.2 Python Code

```python
# In[1]:


# Import Libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelBinarizer
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.svm import SVC
from sklearn.multiclass import OneVsRestClassifier
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import ConfusionMatrixDisplay, roc_curve, auc
import numpy as np
from sklearn.metrics import (
    accuracy_score, confusion_matrix, classification_report, roc_auc_score
)


# In[2]:


# Load the dataset
file_path = "respiratory_dataset.csv"
df = pd.read_csv(file_path)

# Preview the data
print("Dataset Preview:")
print(df.head())

print("\nDataset Information:")
print(df.info())

print("\nMissing Values in Each Column:")
print(df.isnull().sum())


# In[3]:


#  2: Handle Missing Values
# Drop rows where ’Disease’ or ’Gender’ is missing
df.dropna(subset=[’Disease’, ’Gender’], inplace=True)

# Fill missing numerical values with the column mean
numeric_columns = [’Age’, ’BMI’, ’Weight’, ’Height’, ’Start_Time’, ’End_Time’]
for col in numeric_columns:
    df[col] = pd.to_numeric(df[col], errors=’coerce’)  # Convert non-numeric to
   NaN
    df[col].fillna(df[col].mean(), inplace=True)

#  3: Encode Categorical Variables
# Gender: M -> 0, F -> 1
df[’Gender’] = df[’Gender’].map({’M’: 0, ’F’: 1})
```

```
# Disease: Encode as integers
df['Disease'] = df['Disease'].astype('category').cat.codes

#  4: Normalize Numerical Features
scaler = StandardScaler()
columns_to_scale = ['Age', 'BMI', 'Weight', 'Height', 'Start_Time', 'End_Time']
df[columns_to_scale] = scaler.fit_transform(df[columns_to_scale])


# In[4]:


#  5: Split Data into Training and Testing Sets
X = df.drop(columns=['ID', 'Disease'])  # Features
y = df['Disease']  # Target

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
    random_state=42)

print(f"Training Data Shape: {X_train.shape}")
print(f"Testing Data Shape: {X_test.shape}")


# In[5]:


#  6: Train Machine Learning Models
rf_model = RandomForestClassifier(random_state=42)
svm_model = OneVsRestClassifier(SVC(probability=True, random_state=42))  # For
    multi-class SVM
xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='mlogloss',
    random_state=42)

print("\nTraining Random Forest...")
rf_model.fit(X_train, y_train)

print("Training SVM...")
svm_model.fit(X_train, y_train)
```

```
print("Training XGBoost...")
xgb_model.fit(X_train, y_train)


# In[6]:


#  7: Evaluate Models
def evaluate_model(model, X_test, y_test, name="Model"):
    y_pred = model.predict(X_test)
    print(f"\n{name} Evaluation:")
    print("Accuracy:", accuracy_score(y_test, y_pred))
    print("Confusion Matrix:")
    print(confusion_matrix(y_test, y_pred))
    print("\nClassification Report:")
    print(classification_report(y_test, y_pred))
# Evaluate all models
evaluate_model(rf_model, X_test, y_test, name="Random Forest")
evaluate_model(svm_model, X_test, y_test, name="SVM")
evaluate_model(xgb_model, X_test, y_test, name="XGBoost")


# In[7]:


#  8: Feature Importance
print("\nRandom Forest Feature Importances:")
rf_importances = pd.Series(rf_model.feature_importances_, index=X.columns)
print(rf_importances.sort_values(ascending=False))

print("\nXGBoost Feature Importances:")
xgb_importances = pd.Series(xgb_model.feature_importances_, index=X.columns)
print(xgb_importances.sort_values(ascending=False))


# In[8]:
```

```
# Set global style for seaborn
sns.set(style="whitegrid", palette="pastel", font_scale=1.2)

# 1: Feature Importance - Random Forest
plt.figure(figsize=(10, 6))
rf_importances.sort_values(ascending=False).plot(kind=’bar’, color=’skyblue’)
plt.title("Feature Importance - Random Forest")
plt.ylabel("Importance Score")
plt.xlabel("Features")
plt.tight_layout()
plt.show()

# Numerical Result for Random Forest Feature Importance
print("\nRandom Forest Feature Importance:\n", rf_importances.sort_values(
    ascending=False))


# In[9]:


# 2: Feature Importance - XGBoost
plt.figure(figsize=(10, 6))
xgb_importances.sort_values(ascending=False).plot(kind=’bar’, color=’lightcoral’
    )
plt.title("Feature Importance - XGBoost")
plt.ylabel("Importance Score")
plt.xlabel("Features")
plt.tight_layout()
plt.show()

# Numerical Result for XGBoost Feature Importance
print("\nXGBoost Feature Importance:\n", xgb_importances.sort_values(ascending=
    False))


# In[10]:


# 3: Confusion Matrix for Random Forest
```

```
plt.figure(figsize=(8, 6))
ConfusionMatrixDisplay.from_estimator(rf_model, X_test, y_test, cmap="viridis",
    values_format=".0f")
plt.title("Confusion Matrix - Random Forest")
plt.show()

# Numerical Result for Confusion Matrix - Random Forest
rf_cm = confusion_matrix(y_test, rf_model.predict(X_test))
print("\nConfusion Matrix - Random Forest:\n", rf_cm)


# In[11]:


# 4: Confusion Matrix for SVM
plt.figure(figsize=(8, 6))
ConfusionMatrixDisplay.from_estimator(svm_model, X_test, y_test, cmap="Oranges",
     values_format=".0f")
plt.title("Confusion Matrix - SVM")
plt.show()

# Numerical Result for Confusion Matrix - SVM
svm_cm = confusion_matrix(y_test, svm_model.predict(X_test))
print("\nConfusion Matrix - SVM:\n", svm_cm)


# In[12]:


# 5: Confusion Matrix for XGBoost
plt.figure(figsize=(8, 6))
ConfusionMatrixDisplay.from_estimator(xgb_model, X_test, y_test, cmap="YlGn",
    values_format=".0f")
plt.title("Confusion Matrix - XGBoost")
plt.show()

# Numerical Result for Confusion Matrix - XGBoost
xgb_cm = confusion_matrix(y_test, xgb_model.predict(X_test))
print("\nConfusion Matrix - XGBoost:\n", xgb_cm)
```

```
# In[13]:



# 6: Class Distribution
plt.figure(figsize=(8, 6))
sns.countplot(x=y, palette="Set2")
plt.title("Distribution of Disease Classes")
plt.xlabel("Disease")
plt.ylabel("Count")
for p in plt.gca().patches:
    plt.annotate(f'{p.get_height()}', (p.get_x() + p.get_width() / 2, p.
    get_height()),
                 ha='center', va='bottom', fontsize=10)
plt.tight_layout()
plt.show()

# Numerical Result for Class Distribution
class_distribution = y.value_counts()
print("\nClass Distribution:\n", class_distribution)



# In[14]:



# 7: Predictions vs Actual for Random Forest
rf_predictions = rf_model.predict(X_test)
comparison_df = pd.DataFrame({"Actual": y_test, "Predicted": rf_predictions})
plt.figure(figsize=(8, 6))
sns.histplot(data=comparison_df, x="Actual", kde=False, color="orange", label="
    Actual", alpha=0.6)
sns.histplot(data=comparison_df, x="Predicted", kde=False, color="blue", label="
    Predicted", alpha=0.6)
plt.title("Predictions vs Actual - Random Forest")
plt.xlabel("Disease Class")
plt.ylabel("Frequency")
plt.legend()
plt.tight_layout()
plt.show()

# Numerical Result for Predictions vs Actual Counts
predictions_vs_actual = comparison_df.value_counts()
print("\nPredictions vs Actual Counts:\n", predictions_vs_actual)


# In[15]:


# 8: Age Distribution Across Classes
plt.figure(figsize=(10, 6))
sns.boxplot(x=y, y=df.loc[y_test.index, 'Age'], palette="coolwarm")
plt.title("Age Distribution Across Disease Classes")
plt.xlabel("Disease")
plt.ylabel("Age")
plt.tight_layout()
plt.show()

# Numerical Result for Age Distribution Across Classes
age_distribution = df.loc[y_test.index, 'Age'].groupby(y).describe()
print("\nAge Distribution Across Disease Classes:\n", age_distribution)


# In[16]:


# 9: BMI Distribution Across Classes
plt.figure(figsize=(10, 6))
sns.boxplot(x=y, y=df.loc[y_test.index, 'BMI'], palette="Set3")
plt.title("BMI Distribution Across Disease Classes")
plt.xlabel("Disease")
plt.ylabel("BMI")
plt.tight_layout()
plt.show()

# Numerical Result for BMI Distribution Across Classes
bmi_distribution = df.loc[y_test.index, 'BMI'].groupby(y).describe()
print("\nBMI Distribution Across Disease Classes:\n", bmi_distribution)
```

```
# In[17]:


# 10: Gender Distribution Across Classes
plt.figure(figsize=(8, 6))
sns.countplot(x='Disease', hue='Gender', data=df.loc[y_test.index], palette="
    husl")
plt.title("Gender Distribution Across Disease Classes")
plt.xlabel("Disease")
plt.ylabel("Count")
plt.legend(title="Gender", loc="upper right", labels=["Male", "Female"])
plt.tight_layout()
plt.show()

# Numerical Result for Gender Distribution Across Classes
gender_distribution = df.loc[y_test.index].groupby(['Disease', 'Gender']).size().
    unstack()
print("\nGender Distribution Across Disease Classes:\n", gender_distribution)


# In[18]:


# 11: Start_Time and End_Time Contribution Across Classes
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Start_Time', y='End_Time', hue='Disease', data=df.loc[y_test.
    index], palette="muted", s=100)
plt.title("Start and End Time Contribution Across Disease Classes")
plt.xlabel("Start Time")
plt.ylabel("End Time")
plt.legend(title="Disease")
plt.tight_layout()
plt.show()

# Numerical Result for Start_Time and End_Time Contribution
start_end_time_contribution = df.loc[y_test.index, ['Start_Time', 'End_Time', '
    Disease']].groupby('Disease').mean()
```

```
print("\nStart_Time and End_Time Contribution Across Disease Classes:\n",
    start_end_time_contribution)


# In[19]:


# 12: Distribution of Weight Across Disease Classes
plt.figure(figsize=(10, 6))
sns.boxplot(x=y, y=df.loc[y_test.index, 'Weight'], palette="bone")
plt.title("Weight Distribution Across Disease Classes")
plt.xlabel("Disease")
plt.ylabel("Weight")
plt.tight_layout()
plt.show()

# Numerical Result for Weight Distribution Across Disease Classes
weight_distribution = df.loc[y_test.index, 'Weight'].groupby(y).describe()
print("\nWeight Distribution Across Disease Classes:\n", weight_distribution)


# In[20]:


# 13: Age vs BMI Scatter Plot Colored by Disease
plt.figure(figsize=(10, 6))
sns.scatterplot(x=df.loc[y_test.index, 'Age'], y=df.loc[y_test.index, 'BMI'],
    hue=y, palette="gist_heat", s=100, alpha=0.7)
plt.title("Age vs BMI Scatter Plot Colored by Disease Class")
plt.xlabel("Age")
plt.ylabel("BMI")
plt.legend(title="Disease")
plt.tight_layout()
plt.show()

# Numerical Result for Age and BMI correlation
age_bmi_corr = df.loc[y_test.index, ['Age', 'BMI']].corr()
print("\nCorrelation between Age and BMI:\n", age_bmi_corr)
```

```
# In[21]:



# 14: Feature Correlation Heatmap
plt.figure(figsize=(12, 8))
corr_matrix = df[columns_to_scale + ['Disease']].corr()
sns.heatmap(corr_matrix, annot=True, cmap="Accent", fmt=".2f", linewidths=0.5)
plt.title("Feature Correlation Heatmap")
plt.tight_layout()
plt.show()

# Numerical Result for Feature Correlation
print("\nFeature Correlation Matrix:\n", corr_matrix)



# In[22]:



# 15: Visualization of Model Accuracy
# Accuracy scores for each model
accuracy_scores = {
    "Random Forest": accuracy_score(y_test, rf_model.predict(X_test)),
    "SVM": accuracy_score(y_test, svm_model.predict(X_test)),
    "XGBoost": accuracy_score(y_test, xgb_model.predict(X_test)),
}

# Create a bar plot for model accuracy
plt.figure(figsize=(8, 6))
sns.barplot(x=list(accuracy_scores.keys()), y=list(accuracy_scores.values()),
   palette="viridis")
plt.title("Model Accuracy Comparison")
plt.xlabel("Model")
plt.ylabel("Accuracy")
plt.ylim(0, 1)  # Accuracy is between 0 and 1
plt.tight_layout()
plt.show()
```