2021

A Differential diagnostic tool for obstructive lung diseases in adults using classification models.

Ochieng, Daphne Faith
School of Computing and Engineering Sciences
Strathmore University

A Differential Diagnostic Tool for Obstructive Lung Diseases in Adults Using Classification Models

Daphne Faith Ochieng

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Information Technology at Strathmore University

School of Computing and Engineering Sciences
Strathmore University
Nairobi, Kenya

December, 2021

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Daphne Faith Ochieng

DaphfO

13/07/2021

Approval

The thesis of Daphne Faith Ochieng was reviewed and approved by the following:

Professor Ismail Ateya Lukandu, D.Sc.,
Senior Lecturer, School of Computing and Engineering Sciences,
Strathmore University

Dr. Julius Butime,
Dean, School of Computing and Engineering Sciences,
Strathmore University

Dr. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Abstract

The non-specificity of affective symptoms between asthma and chronic obstructive pulmonary disease is a key challenge in medical practice. This is even more complicated when physicians fail to recognize the co-existence of these two diseases, a condition known as asthma-chronic obstructive pulmonary disease overlap. This occurrence mainly affects smokers, middle-aged and older age groups. Failure to diagnose a disease correctly and in a timely manner often leads to administration of wrong drug therapy, delayed treatment or wasted financial resources. To reduce the risk of misclassification, a differential diagnostic tool that can differentiate among patients with asthma, chronic obstructive pulmonary disease and asthma-chronic obstructive pulmonary disease overlap was developed based on measurements of spirometry, blood eosinophil count and smoking history. The study employed a judgmental sampling technique to draw 184 samples from the National Health and Nutrition Examination Survey (NHANES) database based on three classes of obstructive lung diseases. Rapid application development methodology was used as the software methodology upon which the architecture was designed and developed. A comparative analysis was made between a number of classification algorithms including K-nearest neighbour, support vector machines, logistic regression, multilayer perceptron and random forest. The results demonstrate that the differential diagnosic tool can correctly classify patients with obstructive lung diseases with an accuracy of 93.94%, showing an automated approach that would aid physicians in making preliminary diagnoses, leading to optimization of time resources, lower medical device costs and better patient outcomes. Further research regarding the tool's improvement should focus on using a more robust dataset and evaluation in liaison with a physician in a real clinical setting.

Key words: Differential Diagnosis (DDx), Asthma, COPD, Asthma-COPD Overlap, ACO, Obstructive Lung Diseases.

Dedication

This research work is dedicated to my parents and siblings who have been a continued source of motivation and encouragement all through the undertaking of my Master's program.

Acknowledgments

I am highly grateful to the Almighty God for all His love, kindness, guidance, knowledge and wisdom.

I would like to give my special thanks to my supervisor, Professor Ismail Ateya Lukandu, for his guidance and support throughout the project and Dr. Vincent Omwenga for his assistance during the research.

Table of Contents

List of Tables

List of Figures

List of Equations

Abbreviations/Acronyms

ACO	-	Asthma - Chronic Obstructive Pulmonary Disease Overlap
ACOS	-	Asthma - Chronic Obstructive Pulmonary Disease Overlap Syndrome
BD	-	Bronchodilator
COPD	-	Chronic Obstructive Pulmonary Disease
DDx	-	Differential Diagnosis
FN	-	False Negative
FP	-	False Positive
FEV_1	-	Forced Expiratory Volume in the First Second
FVC	-	Forced Vital Capacity
GINA	-	Global Initiative for Asthma
GOLD	-	Global Initiative for Chronic Obstructive Lung Disease
HRCT	-	High-Resolution Computed Tomography
KNN	-	K-Nearest Neighbour
LLN	-	Lower Limit of Normal
MLP	-	Multilayer Perceptron
NHANES	-	National Health and Nutrition Examination Survey
PEF	-	Peak Expiratory Flow
PFT	-	Pulmonary Function Test
SVM	-	Support Vector Machine
SMOTE	-	Synthetic Minority Oversampling Technique
TP	-	True Positive
UML	-	Unified Modelling Language

Definition of Terms

De-identification – The removal of identifying information from a dataset such that individual data cannot be linked to particular individuals (Garfinkel, 2015).

Diagnosis – Identification of the cause of a disease through an assessment of a patient's clinical history, physical examination, and diagnostic imaging/lab data; and the title of the finding (Cook & Décary, 2020).

Differential diagnosis – Identification of the right diagnosis from a number of probable conflicting diagnoses (Cook & Décary, 2020).

Chapter 1 : Introduction

1.1 Background

Respiratory or lung diseases are pathological conditions that prevent the lungs form working properly. They range in severity from mild conditions such as common cold and pharyngitis to life-threatening conditions such as pulmonary embolism, acute asthma, and lung cancer among others. Lung diseases can be categorized in a variety of ways, including by the etiology of the disease, by the tissue involved or by the type and pattern of associated signs and symptoms. Most lung diseases fall under these three types; airway diseases, lung tissue diseases and lung circulation diseases. Airway diseases are those that affect the tubes that carry oxygen and other gases into and out of the lungs, causing a blockage of these tubes/airways. Lung tissue diseases affect the structure of the lung tissue while lung circulation diseases affect the blood vessels in the lungs (Kalhan et al., 2018). Airway diseases can further be classified into obstructive or restrictive. Obstructive lung diseases are conditions that make it difficult for an individual to exhale all the air from their lungs. They include asthma, chronic obstructive pulmonary disease [COPD] which encompasses emphysema and chronic bronchitis, bronchiectasis and cystic fibrosis. On the other hand, individuals with restrictive lung disease have a difficult time filling their lungs fully with air as their lungs are restricted from expanding fully. Some common examples are asbestosis, pulmonary fibrosis and sarcoidosis (Asp, 2020).

Obstructive lung diseases are among the world's leading causes of death and morbidity, with the most common being asthma and COPD (Soriano et al., 2017). Asthma is a heterogeneous disease, characterized by chronic airway inflammation and defined by symptoms such as shortness of breath, wheezing, chest tightness and cough that vary overtime in intensity, together with variable expiratory airway obstruction (Global Initiative for Asthma [GINA], 2019). COPD is a disease characterized by persistent respiratory symptoms and airflow obstruction owing to airway and/or alveolar abnormalities as a result of substantial exposure to toxic particles/gases, influenced by host factors including abnormal lung growth (Global Initiative for Chronic Obstructive Lung Disease [GOLD], 2020).

Asthma and COPD are both extremely prevalent conditions with two different clinical diagnoses but conflicting clinical features. For instance, since they both cause the airways to swell, they can both cause shortness of breath, cough and wheezing. Asthma is most common among children from preschool age but may crop at an adult stage while COPD primarily affects

people over the age of 40 with a smoking history. It is also evident that some cases of COPD are due to long-term asthma, maternal smoking and exposure to noxious gases. Clinical features of both conditions together with persistent airflow limitation can exist within an individual. GINA and GOLD (2017) has identified this as asthma-COPD overlap (ACO), noting that ACO is not a single illness and therefore earlier term used asthma-COPD overlap syndrome (ACOS) is not recommended.

Asthma and COPD have become common in many countries and raised major concerns due to their high mortality and morbidity in severe cases. Asthma prevalence is common in children, while morbidity and mortality rates are much higher in adult population (Abramson, Perret, Dharmage, McDonald & McDonald, 2014). The Global Burden of Disease study in 2016 estimated that 339 people are asthmatic while over 80% of asthma-related deaths occur in developing countries. Another 100 million asthma patients are anticipated by 2025 (Vos et al., 2017). COPD on the other hand had a prevalence of 251 million cases in 2016, and caused 3.17 million deaths in 2015. It remains the fifth and sixth cause of deaths in low and middle-income countries respectively. Measures on prevalence and mortality rates of these two have become difficult due to misclassification and lack of consensus on definitions. However, COPD mortality rates and incidence increase significantly with age (Soriano et al., 2017).

These two diseases have no cure, but are manageable and treatable to ease symptoms and lessen chances of developing complications. However, most times they are both poorly managed especially in older people due to confusion with each other. Distinguishing between the two or even recognizing comorbidity between them has proven difficult for physicians especially in elderly patients and smokers. De Marco et al. (2013) notes that patients who have ACO have a faster progression of the disease and poor quality of health in general than those with either asthma or COPD alone.

Badnjevic, Gurbeta and Custovic (2018) note that the application of computer-based medical diagnostic techniques has been on the rise. This has steadily improved the quality of health care. The key aim of technology involvement is to confirm an initial diagnosis or facilitate a diagnosis in cases where there are uncertainty thereby reducing cases of misdiagnosis. A number of algorithms have been developed in the differential diagnosis of obstructive lung diseases that has resulted to better prognosis, treatment and outcomes for patients. Al Sallakh, Rodgers, Lyons, Sheikh and Davies (2018) for instance, devised a study protocol to develop a

classification model using latent class analysis of health data to identify people with asthma, COPD and ACO.

Understanding characteristics, key differences and similarities between asthma, COPD and ACO is crucial to facilitate a differential diagnosis or recognize comorbidity. This is important as the focus treatment and expected outcome for these diseases are different. Furthermore, administering the correct drug therapy for these conditions can significantly improve patient outcomes and reduce the burden of these diseases.

1.2 Problem Statement

Despite the availability of measures, tests and guidelines to appropriately identify asthma and chronic obstructive pulmonary disease, there still exists a diagnostic confusion between the two, and a challenge in recognizing comorbidity i.e. asthma-COPD overlap, particularly in smokers and older age groups. There is no specific test that is definitive for these diseases due to their complexity. The most effective method for distinguishing the two is spirometry which is prone to errors and not efficient when used alone (Rogliani, Ora, Puxeddu & Cazzola, 2016).

GOLD endorses the spirometry criterion with a fixed cut-off prone to over diagnosing a significant number of older individuals presenting respiratory symptoms as having chronic obstructive pulmonary disease, while also contributing to their relative under diagnosis of asthma (Abramson et al., 2014). Researchers also know little about asthma prevalence in adult population. This represents the shortage of data and greater challenge in discriminating asthma from COPD in older age groups (Global Asthma Report, 2018). The distinction is quite well defined in theory, but is much less clear in practice.

1.3 Objectives

1.3.1 General Objective

The aim of this study is to develop a tool that can be used to discriminate asthma from COPD and ACO using classification models that uses patient history and results from objective tests to construct the model parameters.

1.3.2 Research Objectives

i. To review the characteristics of asthma, chronic obstructive pulmonary disease, and asthma-chronic obstructive pulmonary disease overlap.

ii. To review existing approaches used in hospitals and existing models that have been developed the differential diagnosis of obstructive lung diseases.
iii. To develop a differential diagnostic tool using classification models.
iv. To test the functionality of the application.

1.4 Research Questions

i. What are the characteristics of asthma, chronic obstructive pulmonary disease, and asthma-chronic obstructive pulmonary disease overlap?
ii. What are the existing approaches used in hospitals and existing models that have been developed in the differential diagnosis of obstructive lung diseases?
iii. How will a tool for differential diagnosis using classification models be developed?
iv. How will the functionality of the application be tested?

1.5 Justification

It is clear as the light of day that differentiating asthma from COPD in middle-aged and older people including the elderly is still a challenge in practice. It is even more challenging to identify presence of an overlap of the two conditions within an individual. Yet determining a prognosis and applying an effective treatment to improve patient outcome is highly reliant on physicians making the right diagnosis. Diagnostic confusion between the two or uncertainty in recognizing comorbidity may lead to application of wrong drug therapy or delayed treatment that poses significant risks to patients.

The proposed solution seeks to minimize the risk of misclassification and misdiagnosis of asthma, COPD and ACO patients by assisting physicians confirm an initial differential diagnosis or facilitate one in cases of uncertainty through the use of quality health data and information. Correct differential diagnosis will not only save patients costs spent on unnecessary medications but also save them from the complications that may arise from misdiagnosis. Findings of this study would contribute towards the development of better health management systems through improved information use.

1.6 Scope

The study is limited to only identifying/classifying patients as having asthma, COPD or ACO. It does not aim to identify or classify other lung diseases or other common diseases that

normally overlap with asthma or COPD. Classification algorithms to be used include K-nearest neighbour (KNN), support vector machines (SVM), logistic regression, multilayer perceptron (MLP) and random forest. Datasets to be used for training will be limited to cases initially confirmed as asthma, COPD and ACO.

Chapter 2 : Literature Review

2.1 Introduction

The diagnostic process, often called clinical reasoning, involves identification of the cause of a disease through an assessment of a patient's clinical history, physical examination, and diagnostic imaging/lab data; and the title of the finding. Differential diagnosis on the other hand involves identification of the right diagnosis from a number of probable conflicting diagnoses. The ability to prevent misdiagnosis is important if true management of a disease is to be reached. Clinicians therefore need a high index of suspicion (high order thinking) that goes far beyond memorizing measures and tests, specificity and sensitivity, and a list of set medical diagnoses (Cook & Décary, 2020).

2.2 Theoretical Review

2.2.1 Hypothetico-Deductive Model

There exist a number of models used in clinical reasoning namely dual process diagnostic reasoning model, pattern recognition, hypothetico-deductive model, integrative model for clinical reasoning among others. Differential diagnosis implements certain aspects of the hypothetico-deductive method. It is the most prevalent principle of diagnostic reasoning. This model posits that early in the clinical encounter with a patient, physicians generate certain hypotheses based on indications they find and subsequently gather data to confirm or refute these hypotheses (Donner- Banzhoff, 2018). The four stages of this reasoning process are hypothesis generation, evaluation, refinement and hypothesis verification.

Kovacs and Croskerry (as cited in Medford-Davis, Singh & Mahajan, 2018) note that in the first stage, one or more diagnostic hypotheses are generated early or before the encounter begins. Even though this begins early, hypothesis generation can be activated at all stages of the interaction. At this stage, the heuristics, disease prevalence, and the acuity of the patient's condition are important. In the second stage, a framework is created to guide in gathering more information using either confirmation or elimination strategies. Non-emergency practitioners approach this with an aim to confirm a hypothesis while emergency practitioners aim at eliminating potentially life-threatening diagnostics.

In the third stage more information is collected. Hypotheses that were generated earlier are refined, increasing their precision while dropping the irrelevant ones. Prioritization of hypotheses generated is based on prevalence while their confirmation or elimination is based on

the severity with which the patient's symptoms present themselves. This stage may occur simultaneously with hypothesis evaluation. Finally, during hypothesis verification, the hypothesis is considered based on its consistency with the patient's presentation before a working diagnosis is acknowledged (Kovacs and Croskerry (as cited in Medford-Davis et al., 2018)).

2.2.2 Bayes' Theorem

Bayes' theorem is the statistical basis for differential diagnosis and provides one way to overcome data uncertainty. This theory gives an account of the probability of an event based on the prior knowledge of the conditions that might be related to the event. It is stated as follows;

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Equation 2.1: Bayes' Theorem

Where;

i. A and B are events and $P(B) \neq 0$.
ii. $P(A \mid B)$ is probability of A given that B is true.
iii. $P(B \mid A)$ is probability of B given that A is true.
iv. P (A) and P (B) are the independent probabilities of A and B.

González-López et al. (2016) applied this theorem in the development of a Bayesian belief network algorithm that was used in the differential diagnosis of anterior uveitis. The authors included eleven common causes of the disease (ankylosing spondylitis, idiopathic, psoriatic arthritis, sarcoidosis, inflammatory bowel diseases, reactive arthritis, Behçet, tuberculosis, Posner-Schlossman syndrome, Fuchs' heterochromic cyclitis and juvenile idiopathic arthritis), used these in the correct classification of common cases of anterior uveitis and validated in a randomized study of 200 patients with anterior uveitis. This study will apply classification algorithms that apply Bayes' rule in assigning obstructive lung diseases into three classes.

2.2.3 Test Theory

Another theory that has been proposed for use in diagnosis and classification problems is the concept of the testors, i.e. the Test Theory created by Yablonskii and Chegis. It was used as a tool for analysing problems to do with fault control and diagnosis in circuits (Chikalov et al., 2012). The development of this theory has been strongly connected to two kinds of problems; the detection of faults in electrical circuits and the classification and recognition of objects. Applications include document summarization and text categorization (Sanchez-Diaz, Lazo-Cortes & Piza-Davila, 2012). Test theory is used in supervised classification problems to select relevant features. Typical testors are discordant feature subsets that retain the capacity of the original set of features to distinguish objects (Rodríguez-Diez, Martínez-Trinidad, Carrasco-Ochoa & Lazo-Cortés, 2017).

The concept of testors is as follows; Let TM = {Q1, Q2,…, Qn}, TM ⊆ U (U= universe set of objects), be a training matrix with n objects, defined as n features R = {X1, X2, …, Xn } and assigned into c classes {C1, C2, …, CC}. Each feature Xi ∈ R takes values in a set Mi, i = 1,…,n. A dissimilarity criterion of comparison Di: Mi x Mi → {0,1} is linked to each Xi (0 indicates similarity, 1 indicates dissimilarity). Each object Q ∈ TM has associated a c-tuple of membership, which describes the belonging of the object Q to the classes C1,...,Cc. This is denoted as α(Q). Then, α(Q) = (α1(Q),…,αc(Q)), where αi(Q) = 1 means that Q ∈ Ki and αi(Q) = 0 means that Q ∉ Ki (Sanchez-Diaz et al., 2012).

Ortíz-Posadas, Martínez-Trinidad and Ruiz-Shulcloper (as cited in Alba-Cabrera, Godoy-Calderon & Ibarra-Fiallo, 2016) proposed a methodology based on the test theory in developing a criterion for differential medical diagnosis. Using this theory, they obtained a minimum set of equally discriminant features (typical testors) among diseases that were considered. Then, it was applied together with classification algorithms that used typical testors to allow physicians more flexibility in making differential diagnosis of diseases. This study will apply the concept of testors to perform feature selection and determine feature relevance in the accurate classification of patients with obstructive lung diseases into their distinct classes.

2.3 Characteristics of Obstructive Lung Diseases

Asthma and COPD are caused by different factors. Studies have elucidated that early asthma attack leads to COPD. Notably, children with chronic asthma and decreased growth of lung function are prone to developing COPD at adulthood. Asthma manifestation has been found to set during early development stage while COPD at adult stage and most of the patients always have a history of asthma attack. Early lung function is a great indicator of growth of lung function at later stages in life (McGeachie et al., 2016).

Even though in adult population, COPD has always been associated by smoking of cigarettes not all smokers develop the disease and even non-smokers do. Smith et al. (2018) found that genetic differences in lung anatomy might serve as markers to help distinguish individuals with low, but stable, early-life function, and those who are especially at risk of developing COPD due to a decrease in lung function as a consequence of smoking.

Bui et al. (2018*)* deduced that childhood illnesses such as bronchitis, pneumonia, asthma and exposure to parental smoking are also linked to COPD and three quarter of the disease are majorly with children with risk factors that are exacerbated at childhood. In China, it was found that COPD prevalence increased significantly with age with 2.1% in individuals aged between 20-39 years compared to 13.1% in their counterpart that were in the age bracket of 40 years and above. The disease was more prevalent in rural residents (9.6%) than in urban residents (7.4%). This same pattern has been observed among non-smokers (Fang et al., 2018).

On the other hand, asthma can attack at any stage of life but is more prevalent in children as they progress to adulthood compared to attack during later stages of life. It is possible that exhaled volatile organic compounds could qualify as biomarkers to detect early signs of asthma (de Benedictis & Attanasi, 2016). Normally asthma may range from mild to severe cases which may be fatal or lead to COPD. Kashanian, Mohtashami, Bemanian, Moosavi and Lakeh (2017) in their study concluded that history of maternal asthma was the most influential factor on childhood asthma development. This was accompanied by antibiotic exposure in utero, vaginal bleeding during pregnancy and older maternal age.

Environmental factors like pesticides and smoking by caregivers are the major sole cause of asthma in adults while alcohol intake during gestation or lactation was found not to be associated with early manifestation of asthma. A multivariate analysis performed by Hallit, Raherison, Waked and Salameh (2017) revealed that there was an elevated risk of asthma for children residing in

North and South Lebanon and areas where pesticides were commonly used. Smoking pipe during pregnancy was also identified as a major risk factor for developing childhood asthma. The studies sufficiently show that asthma in early adulthood is majorly influenced by early history of exposure to its risk factors during childhood whereas at later stages it is influenced by allergens present in the environment e.g. dust, dust mites, animal dander, foods or food additives, moulds etc.

Asthma-COPD overlap has recently received a great deal of attention. It occurs in adults who exhibit both COPD and asthma symptoms. Distinguishing asthma from COPD and ACO has proven difficult since it requires an expert practice. It should be treated as an independent subtype with distinctive features, identified by inherited environmental factors and not genetical factors (Park et al., 2019).

ACO is not well understood and subsequently its characteristics and prevalence are not well understood. Inoue et al. (2017) conducted a study on patients with fixed airflow obstruction. ACO patients were found to be 9.2% and 4.2% among the 1,008 outpatients medically treated for COPD. Individuals subjected to higher levels of air pollution had almost three times greater probabilities of developing ACO. It was concluded that minimizing exposure to these levels of air pollution may reduce the risk of developing ACO, identifying air pollution as a major risk factor.

Table 2:1: Characteristics of Obstructive Lung Diseases (from Manian, 2019, p.1763)

Asthma	**COPD**	**Asthma-COPD Overlap**
Early onset	Late onset	Asthma patients who also exhibit irreversible obstruction of the airways owing to structural changes
History of atopy	Smoking history	Smoking asthma patients in whom there is prominent neutrophilic inflammation
Intermittent symptoms	Slowly progressive symptoms	COPD patients with eosinophilic inflammation
Significant bronchodilator response	No significant bronchodilator response	
Eosinophilic	Neutrophilic inflammation	

inflammation		
Positive inhaled therapy response	Poor response to inhaled therapy	

2.3.1 Signs and Symptoms

COPD symptoms do not appear immediately unless a significant amount of lung damage has occurred and usually gets worse as the deterioration increases, especially if the patient continues to smoke. Signs and symptoms include: shortness of breath mostly during physical activities, languor, and tightness in the chest, wheeze, chronic and productive cough that is white, greenish or yellow, frequent respiratory infections, unintentional loss of weight later on, swelling in ankles, legs or feet. COPD patients are more likely to experience exacerbations, where their symptoms become severe and persistent (Gentry & Gentry, 2017).

While COPD is determined by significant lung damage, severity and intensity of asthma determines its signs and symptoms. Some people exhibit the symptoms daily while others occasionally during the year. They include tightness in the chest, coughing at night or early morning, wheezing and shortness of breath (Weinberger, Cockrill & Mandel, 2017).

ACO occurs as a result of combination of both diseases in an individual therefore signs and symptoms can be thought of as a synergy of the two. They include difficulty in breathing, frequent coughing, wheezing, chest tightness, production of excess phlegm, tiredness, inability to tolerate exercises and shortness of breath when performing routine tasks (Miravitlles, 2017).

2.4 Approaches Used in Differential Diagnosis

2.4.1 Spirometry

Spirometry measures the volume/amount or speed (flow) of air that one can inhale or exhale. It is also called pulmonary function test (PFT). A patient breathes normally while remaining seated, then takes a deep breath and exhales as fast and hard as possible. From this, the forced expiratory volume in the first second (FEV_1) of exhalation is measured in comparison to the entire volume of air that can be expelled in a forced expiration (forced vital capacity [FVC]). This should be repeated three times for consistency. Peak expiratory flow (PEF) is the maximum flow that can be exhaled while blowing out at a steady rate (Moore, 2012).

Asthma and COPD are characterized by airflow obstruction, which is a reduced FEV_1 and a reduced FEV_1/FVC ratio, such that FEV_1 is a value below the 5^{th} percentile in adults or

less than 80% of that predicted in children and adolescents 5 to 18 years of age, and FEV_1/FVC is less than 0.7 lower limit of normal [LNN] (Langan & Goodbred, 2020). Table 2.2 illustrates an approach to interpretation of PFTs.

Table 2:2: Pulmonary Function Test Interpretation (from Johnson & Theurer, 2014, p.363)

Test Results Based on Age		
FVC	**FEV_1/FVC Ratio**	**Suggested Diagnosis**
5 - 8 years: ≥ 80% Adults: ≥ LLN	5 - 18 years: ≥ 85% Adults: ≥ LLN or ≥ 70%	Normal
5 - 18 years: ≥ 80% Adults: ≥ LLN	5 - 18 years: < 85% Adults: < LLN or < 70%	Obstructive defect
5 - 18 years: < 80% Adults: < LLN	5 - 18 years: ≥ 85% Adults: ≥ LLN or ≥ 70%	Restrictive pattern
5 - 18 years: < 80% Adults: < LLN	5 - 18 years: < 85% Adults: < LLN or < 70%	Mixed pattern

GINA and GOLD guidelines provide further recommendations that differentiate asthma from COPD using spirometry. This includes a post-bronchodilator (BD) test or reversibility test that utilizes spirometry and include asthma-COPD overlap as shown in Table 2.3 below.

Table 2:3: Spirometric Measures in Obstructive Lung Diseases (from GINA & GOLD, 2017, p.14)

Spirometric Factor	**Asthma**	**COPD**	**Asthma-COPD Overlap**
Normal FEV_1/FVC Pre/post BD	Fits the diagnosis	Incompatible with diagnosis	Incompatible unless other evidence of airway limitation is provided
Post-BD FEV_1/FVC < 0.7	Indicates airflow obstruction that may improve	Necessary for diagnosis (GOLD criteria)	Usually present

Post-BD $FEV_1 \geq 80\%$ predicted	Fits the diagnosis	Fits the GOLD classification of mild airflow obstruction if post-BD therapy is < 0.7	Fits the diagnosis of mild ACO
Post-BD $FEV_1 < 80\%$ predicted	Fits the diagnosis	Indicates severity of airflow obstruction	Indicates severity of airflow obstruction
Post-BD increase in $FEV_1 \geq 12\%$ and 200ml from baseline(reversible airflow limitation	Typical at some point in the course of asthma	Usual and more probable when there is a low FEV_1	Usual and more probable when there is a low FEV_1
Post-BD increase in $FEV_1 > 12\%$ and 400ml from baseline(marked reversibility	High risk of asthma	Uncommon in COPD	Fits the diagnosis

2.4.2 High-Resolution Computed Tomography

High-resolution computed tomography (HRCTs) refers to a computed tomography method in where thin-slice photographs of the chest are obtained and post-processed in a reconstruction algorithm with high-spatial-frequency. This method offers exquisite lung detail images suitable for the evaluation of diffuse interstitial lung disease, more particular the ones that involve the interstitium (Bell, 2020).

Park et al. (as cited in Grimm, 2016) conducted a study in an attempt to differentiate asthma from COPD using HRCT scans. The results showed that bronchial walls were 2.3 mm thicker than usual in bronchial asthma and 0.9 mm thicker than usual in COPD. The wall thickness to luminal diameter ratio was not associated with clinical characteristics such as history of smoking, period of symptoms or physiological rates such as FEV_1. HRCT findings of tubular bronchiectasis, emphysema (a type of COPD), and mosaic lung attenuation were associated to a significant history of asthma symptoms, impaired lung function, and reduced bronchial hyper responsiveness. In conclusion, it is possible to discriminate asthma from COPD from the data although this is speculative in individual cases.

2.4.3 Physical Examination

Physical examinations rarely diagnose asthma, COPD and ACO. In cases of COPD there are usually no physical signs of airflow obstruction until significant lung damage or disability has occurred while in asthma physical examination may turn out normal unless the patient is experiencing exacerbations. In asthma, during exacerbations, overt wheezing becomes prominent, but the lungs get overinflated as the condition gets worse.

The following are some things to take note of to aid in diagnosis of COPD; noting the patient's posture, respiratory rate, discoordinate thoraco-abdominal motion, active abdominal muscle expiratory efforts, and clubbing of the fingers. An examiner can tap on the patient's chest wall (percussion) to detect whether there is too much air present in the lungs. The cardiac function, peripheral oedema, vein distension, central cyanosis, and chest wall anomalies should also be considered in depth (Spencer & Krieger, 2013).

2.5 Related Works

Different models have been proposed in classifying asthma, COPD and ACO. The following highlights a few of the models that have been developed using machine learning algorithms.

2.5.1 An Expert Diagnostic System to Identify Obstructive Lung Diseases

Badnjevic et al. (2018) developed an expert diagnostic system based on artificial neural networks and fuzzy logic algorithms to classify patients as either healthy or having asthma/COPD. The models achieved 96.66% classification accuracy using data from 3657 patient records and verified on 1650 patients. Measurements were based on lung function test results and information about patient's symptoms. The results demonstrated a correct identification of asthma/COPD patients with a sensitivity of 96.45% and specificity of 98.71%. 98.71% of patients with normal lung function were also correctly classified.

However, even though the results were promising the application of neural networks would have been much more effective when applied to a larger amount of training data set that did not happen in this case. Other methods like gradient boost and random forest may have performed better than neural networks and there might be a problem of over fitting as well. It also predicted a patient as either COPD/asthmatic or normal. It did not show the clear distinction

for those who may be having asthma only or COPD separately which other machine learning techniques like Bayes classifier would achieve better (Gareth, Daniela, Trevor & Robert, 2013).

2.5.2 Differentiating Asthma from Chronic Obstructive Pulmonary Disease

Maravic et al. (2019) adopted a supervised machine learning approach to a dataset consisting of 976,584 visits (307,976 patients) that were diagnosed with asthma or COPD associated with a treatment prescription for a period of three years. Most of the patient and visits in the database corresponded to those who were diagnosed with asthma with 84.1% and 77.8% respectively. 75% of the dataset was used as training data and 25% as verification data to evaluate the accuracy of the model. Parameters used to construct the model were those available in dispensation databases, i.e., age, gender, type/dosing/presentation of the prescribed drug, and date of the prescription. 87.2% and 86.4% of asthma and COPD patients were properly classified respectively. These results seemed promising but as Badnjevic et al. (2018) suggests, neural networks would be the most ideal for a situation like this since the data set was very large. This approach also did not specify the type of machine learning classification algorithm used and the reader is left to ponder what method was applied.

2.5.3 COPD Classification Using Machine Learning Algorithms

This study aimed at automating the process of COPD classification with higher performance and faster approach. Vora and Shah (2019) conducted a comparative study between SVM and K-nearest neighbour to perform classification of COPD severity level. Model parameters used include spirometry test results, data about side effects, hypersensitiveness and auscultation of the patient. The forecasts completed by the SVM calculation delivered a high accuracy of 96.97%, precisely classifying COPD severity level. KNN obtained an accuracy of 92.30%, which was low compared to the accuracy obtained using SVM algorithm. Goel and Mahajan (2017) note that KNN is used mostly as a multi-class classifier whereas standard SVM separate binary data belonging to either of one class. This study only majored on classifying patients' COPD severity level.

2.5.4 Asthma/Chronic Obstructive Pulmonary Disease Classification

Kaplan et al. (2020) investigated eleven supervised machine learning algorithms that would perform better in diagnosis and classification of COPD, asthma and ACO using over 60 clinical features from a US electronic medical record database for a cohort of patients ≥35 years

who had a specialist diagnosis of asthma, COPD or both on ≥2 occasions. It appeared that extreme gradient boosting model using 12 clinical features with Bayesian hyper-parameter optimization performed well compared to the rest, achieving sensitivity 0.98, 0.98 and 0.78 and an F1-score of 0.98, 0.98 and 0.84 in diagnosing COPD, asthma and ACO respectively. The model proved to be performing well by classifying patients into three categories. This approach has however not been evaluated in other countries and settings.

2.5.5 Identifying Patients with Asthma-Chronic Obstructive Pulmonary Disease Overlap

This study proposed the use of latent class analysis to classify a patient as having ACO, asthma only or COPD only in electronic health record data. Latent class analysis divides a sample into classes or cluster related to set of observed variables, assuming that the variables can be explained by additional measurement errors and hidden categorical variables that are able to divide the sample into predefined number of distinct classes (Howard, Rattray, Prosperi & Custovic, 2015). Al Sallakh et al. (2018) hypothesized that their approach would differentiate between the three cases distinctively. This would also offer an opportunity to evaluate how clustering might perform based on these surrogate variables in comparison to disease markers (Ghebre et al., 2015). This study only provided a protocol to be followed in said classification and results were not revealed or made available.

2.5.6 Summary

Table 2:4: Summary of Reviewed Models

Model/System	**Features**	**Specificity (%)**	**Sensitivity (%)**	**Reported Accuracy (%)**	**Reference**
An Expert Diagnostic System to Identify Obstructive Lung Diseases	Patient symptoms, lung function test results	98.71	96.45	96.66	Badnjevic et al. (2018)
Differentiating	Age, gender,	-	-	86	Maravic et

Asthma from Chronic Obstructive Pulmonary Disease	type/dosing/ presentation of prescribed drug, date of prescription				al. (2019)
COPD Classification Using Machine Learning Algorithms	Spirometry test results, data about side effects, hypersensitiveness and auscultation of the patient	-	-	SVM – 96.97 KNN – 92.30	Vora and Shah (2019)
Asthma/Chronic Obstructive Pulmonary Disease Classification	Spirometry results, pack years, body mass index, symptoms, allergic rhinitis, chronic rhinitis	-	COPD– 0.98 Asthma-0.98 ACO – 0.78	-	Kaplan et al. (2020)

2.6 Research Gap

Among all the models reviewed, only two aimed at achieving classification of the diseases into three distinct groups i.e. asthma, COPD and ACO using extreme gradient boosting and latent class analysis while one identified respiratory patients as either having a normal lung function or asthma/COPD without distinctively differentiating the two. One other model only focused on classifying COPD severity level and the other classifying a patient as having either asthma or COPD without including ACO. This study aims to improve and contribute to these solutions by applying KNN, SVM, logistic regression, MLP and random forest algorithms and choosing the best model to classify the diseases into three distinct groups.

2.7 Conceptual Framework

The study applies the attributes of the updates of Bellingegni et al. (2017) comparative analysis model which are also the key success factors for successful implementation of classification algorithm to compare support vector machine, linear discriminant analysis, non-logistic regression, and multilayer perceptron. Unlike their approach this study will focus on KNN, SVM, logistic regression, MLP and random forest to classify patients with obstructive lung diseases into three distinct groups and determine which model will perform well in predicting the patient outcome correctly using performance metrics. The best performing model will then be chosen in order to determine the adoption factor and effectiveness of treatment outcome. Input parameters will include age, gender, FEV_1/FVC ratio, blood eosinophil count and smoking history.

Table 2:5: Operationalization of Variables

	Variable	**The Operationalized Variable**
Dependent Variable	Differential diagnosis	Identifying the proper diagnosis among asthma, COPD and ACO
Independent Variable	Classification models	KNN, SVM, MLP, logistic regression, random forest

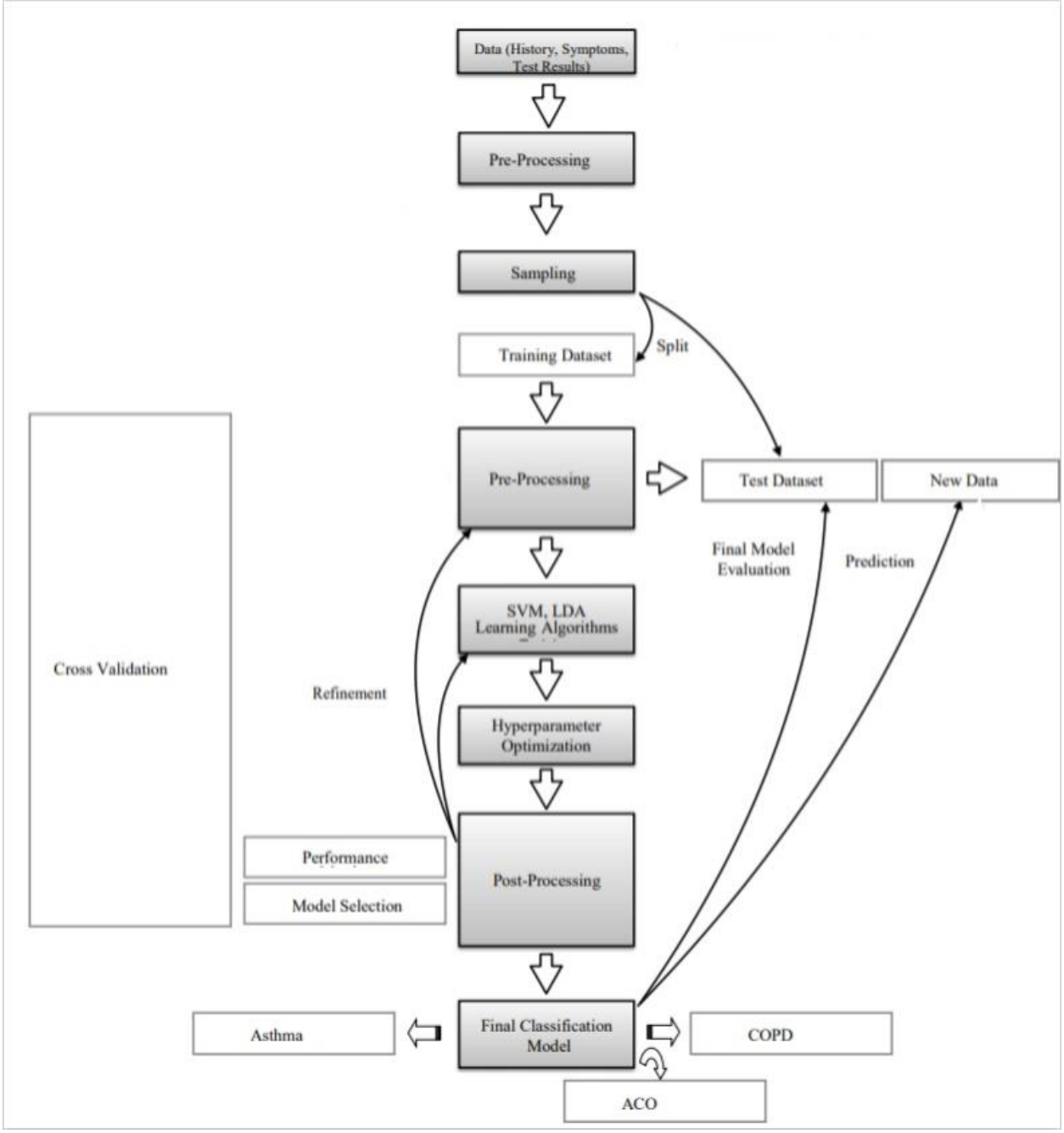


Figure 2:1: Conceptual Framework

Chapter 3 : Research Methodology

3.1 Introduction

Research methodology outlines techniques used in the identification, selection, processing, and analysis of information regarding a topic, enabling the critical evaluation of the overall reliability and validity of a study. This includes the research design, data gathering and analysis (Seltman, 2015). Research methods are aimed at finding solutions to research problems while the methodology aims at utilizing the correct techniques in order to find solutions. In order to satisfy the objectives of this study, the research design will involve consideration of sources of data, methods and tools for data collection, quantitative data analysis, model and application development.

3.2 Research Design

The aim of the research design is to provide a suitable framework for a study. This study employed an experimental method. Ayash (2014) says that the experimental method within information technology is used in a number of different fields such as artificial neural networks, natural languages, analyzing performances and behaviours among others. In an experimental research, values of dependent variables can change due to a change in the independent variables. The study employed this method to measure and submit these changes to statistical analysis to describe the relationships between the variables and analyze performance of models when submitted to the chosen algorithms.

3.3 System Development Methodology

The application to deploy the model was developed using the rapid application development methodology. This is a form of agile methodology that gives priority to rapid prototype releases and iterations, stressing the use of software and user input over strict planning and documenting requirements.

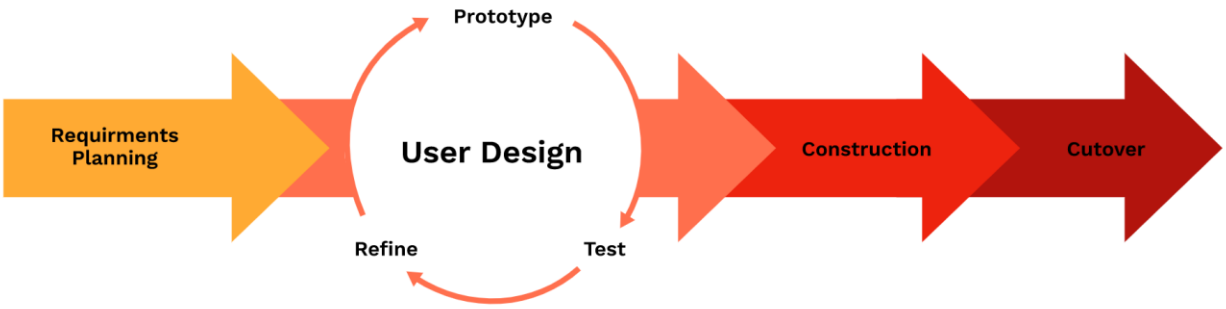


Figure 3:1: Rapid Application Development Methodology (from Roadmunk, 2019)

i. Requirements Planning

This stage entailed collection of relevant secondary data on patients with a physician diagnosis of asthma, COPD and ACO to enable the development of the classification model.

ii. User Design

In this stage, the design of the prototype and architecture of the proposed solution was developed. Unified Modelling Language (UML) diagrams were designed to describe flow of information and interaction of the different system components. Application used in drawing the UML diagrams was draw.io.

iii. Construction

The prototypes were converted to a working model following the information in the user design stage. This was followed by system testing to ascertain functionality. The development language used was Python.

iv. Cutover

The working model was implemented and deployed, with more testing tasks to ensure functionality in the classification and prediction of patients with obstructive lung diseases.

3.4 Target Population and Sampling

The population for a survey is defined as the entire set of units for which the survey data are to be used to draw conclusions (Gupta, 2014). The target population for this study was adults diagnosed with obstructive lung disease, more specifically asthma, COPD or ACO in the

National Health and Nutrition Examination Survey (NHANES) database (2011-2012). The database was considered ideal for its completeness and availability of the three conditions, each as mutually exclusive disease as other databases only provided data on asthma and COPD alone without including ACO.

Lomax and Hahs-Vaughn (2013) define a sample as a subset of population selected by a defined procedure. A judgmental sample of 184 participants with obstructive lung disease was selected from the database. Reddy and Ramasamy (2016) note that the authority for the judgmental sampling techniques lies in the choice of information rich cases for an in-depth analysis. For instance, in this study, the goal was to develop a differential diagnostic tool for obstructive lung diseases in adults. Therefore, the researcher had to learn a lot more by concentrating in depth on understanding the characteristics of a smaller number of adult patients with a presentation of asthma, COPD and ACO rather than collecting information on a larger number of patients presenting a wider range of respiratory diseases while also including those specified for the study. This was the most effective method in this case because it allowed for the deliberate selection of patients presenting typical characteristics of obstructive lung diseases. Furthermore, it is a low-cost, convenient and time-saving strategy, requiring no special knowledge of statistics (Taherdoost, 2016). However, this technique also has its disadvantages. It is not scientific, prone to researcher bias regardless of the method used to collect the data and it can be challenging to defend the representative nature of the sample. Moreover, there is no way to determine the reliability of the researcher, making it impossible to determine if there is an error in the information presented (Sharma, 2017). Research bias was minimized through the use of accepted GINA and GOLD criteria for identifying patients with a presentation of obstructive lung diseases. Figure 3.2 illustrates the identification of the study population.

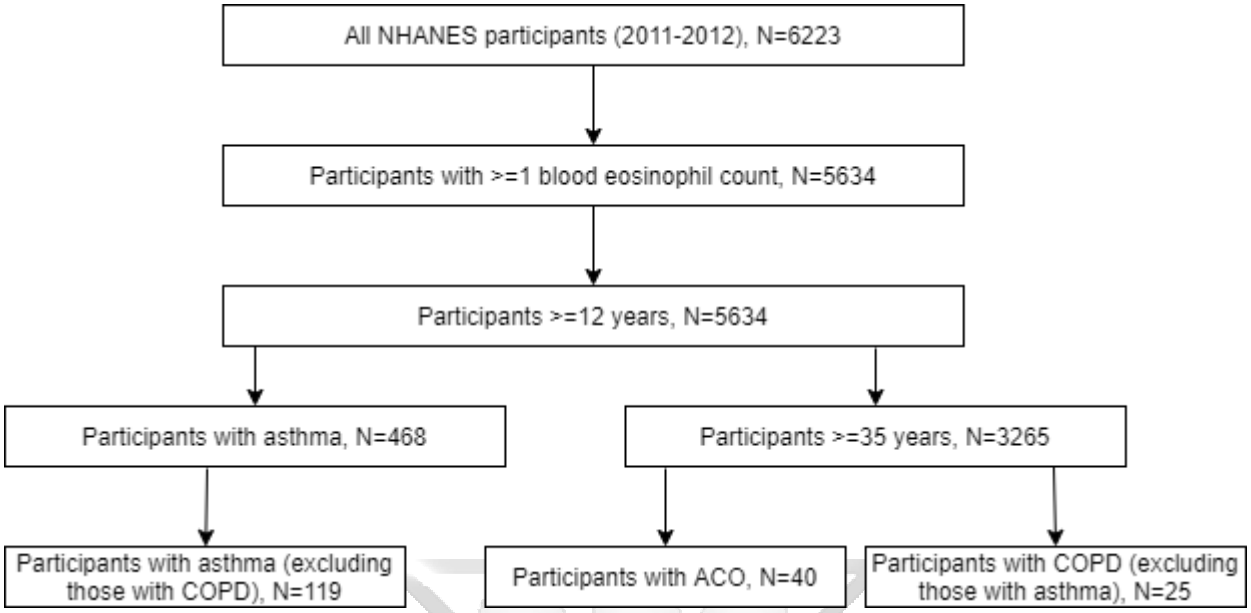


Figure 3:2: Study Population Identification

3.5 Data Collection and Analysis

The dataset required for this study was retrieved through the use of the statistical analysis system universal viewer application provided by NHANES. Using this same application, the data was converted to a comma separated values format then subjected to statistical analysis using python pandas including mean, median and standard deviation for data that was continuous and proportions and frequencies for categorical data. A correlation matrix was used to measure the linear relationships between the variables.

The data retrieved from this database for the study was already de-identified and sub-datasets include demographic, lab, physical examination and questionnaire. Information collected therefore includes age, sex, health disability, spirometry measures, questionnaire responses and peripheral blood eosinophil counts. De-identification of data mitigates privacy risks to individuals thereby supporting the secondary use of data. Patients in the NHANES dataset were classified into three groups according to the following criteria for data labelling;

i. Asthma – patients who reported a diagnosis of asthma or who have had asthma attack in the last 12 months.
ii. COPD – patients with an $FEV_1/FVC < 0.7$ even after administration of bronchodilator therapy, or if they had emphysema or chronic bronchitis.

iii. ACO – patients who meet at least one characteristic from both asthma and COPD group.

3.6 Model Development

The following steps were used in the development of the model;

i. Data pre-processing - This stage involved data engineering and feature engineering to select the features most relevant for the prediction task.
ii. Model training – The model was trained using KNN, SVM, logistic regression, MLP and random forest algorithms.
iii. Model validation and testing – The generated models were validated against the validation dataset and tested with the testing set.

3.7 Research Quality, Validity and Reliability

Reliability and validity are principles used in assessing the quality of a study. They are used to determine how well a method measures something. Reliability is the consistency of a measure while validity indicates the precision of a measure (Middleton, 2019). To achieve this, a wide array of sample sizes was used to increase precision. In addition, the model's performance was evaluated using performance metrics. Since the study was tackling a multi-class classification issue, a confusion matrix was used in defining the notions of true positive (TP), false negatives (FN) and false positives (FP) for each class. These metrics were calculated as shown in Equation 3.1 below;

$$Accuracy = \frac{\sum_{i=1}^{N} TP_i}{Total} \times 100\%$$

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{TP_i}{TP_i + FP_i}\right] \times 100\%$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i} \times 100\%$$

$$F - Measure = \frac{2.Precision.Recall}{Precision + Recall} \times 100\%$$

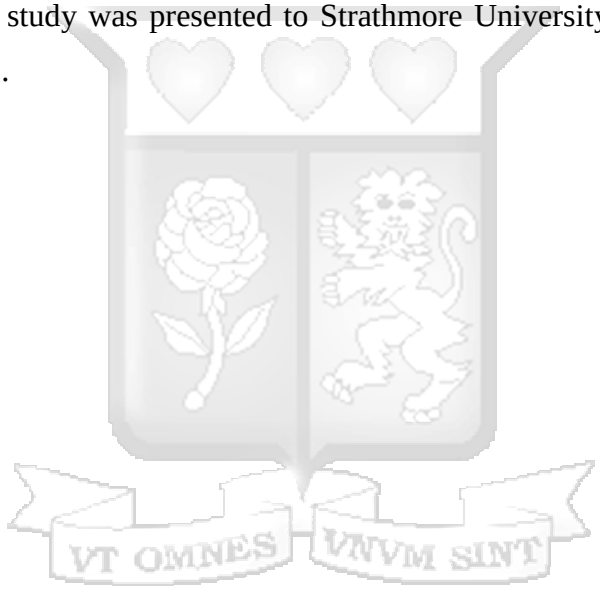
Equation 3.1: Performance Metrics (from Abidine, Fergani & Ordóñez, 2016, p.312)

3.8 Ethical Considerations

This study did not require collection of primary data. The study required secondary data retrieved from the NHANES database from previously conducted surveys that obtained approval from respective authorities. Since these datasets were already de-identified, informed consent was not required. The data was used solely for research purposes. Moreover, all cited authors were acknowledged to avoid plagiarism and give credit for their work. Ethical approval was obtained from Strathmore University Institutional Ethics Review Committee.

3.9 Dissemination and Utilization of Result

The result of this study was presented to Strathmore University, School Of Computing and Engineering Sciences.



Chapter 4 : System Analysis, Design and Architecture

4.1 Introduction

System analysis is the process of understanding and specifying in detail what a system should do. Its purpose is to understand and document the essential characteristics of a system under study. System design on the other hand is specifying in detail how the many components of the system should be physically implemented. The overall goal of system analysis and design is to fully understand the needs or requirements of a system under study and have a flawless construction. Subsequent sections in this chapter details how this was achieved through a comprehensive requirements' analysis and exploration of system design and architecture using UML diagrams such as use case, data flow, sequence, entity-relationship and class diagrams.

4.2 Requirements Analysis

4.2.1 Functional Requirements

Functional requirements define the functions of a system or its components, i.e. what a system must/must not do. A function is a set of inputs, behaviour and outputs.

i. The tool should be able to register a new user i.e. physician.
ii. The tool should accept a patient's input data entered manually by the physician.
iii. The tool should be able to classify a patient's condition based on patient characteristics from input data.
iv. The tool should recommend to the doctor the potential condition the patient is suffering from.

4.2.2 Non-functional Requirements

Non-functional requirements specify a system's operational capabilities and constraints that affect user experience. These include attributes such as system reliability, security, performance, maintainability, usability, scalability among others. The following are the attributes that the tool must have:

i. Security: Users of the tool will need to hold a logon identification and password. Any modifications like insert, update and delete can be synchronized quickly and executed by the administrator.

ii. Performance: The tool should provide a diagnostic report in less than a minute. It should accommodate at least 10 people at once. The user interface should acknowledge within five seconds.

iii. Maintainability: The tool should offer efficiency for data back-up and track and keep a log of every mistake.

iv. Reliability: The system should be available at all times.

4.3 System Architecture

The proposed system architecture as illustrated in Figure 4.1 below outlines the general layout of the tool. The physician should be able to access the tool through an internet enabled device on which it is installed. The user interface will enable the physician to log in, input patient data and submit it. The model pre-processes this data and classifies the patient then gives the physician a diagnostic recommendation.

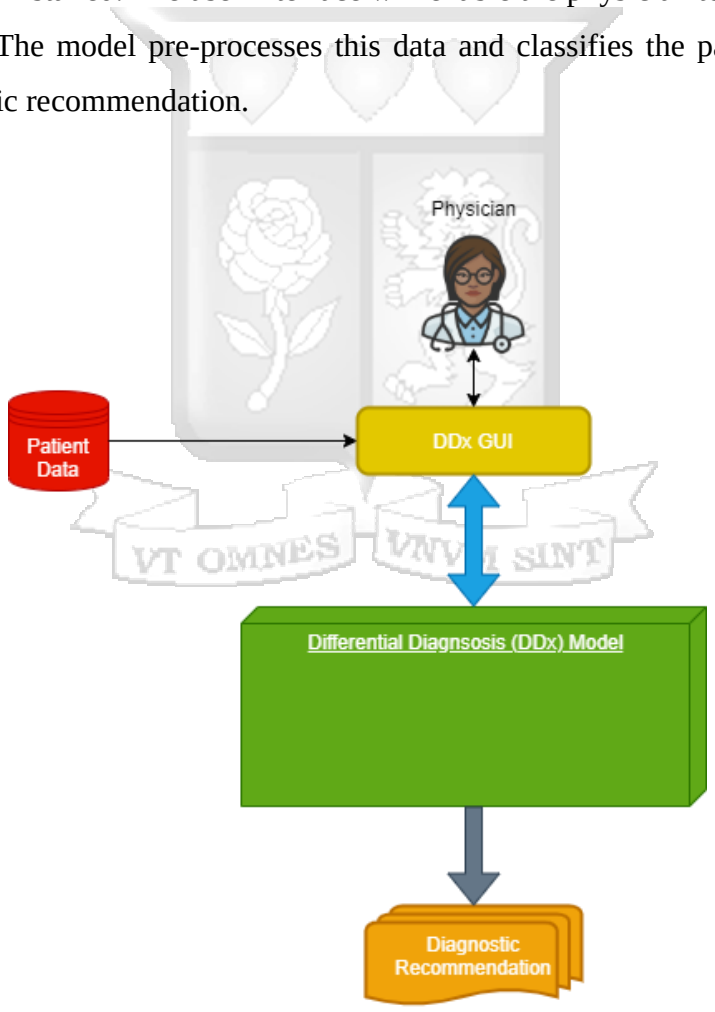


Figure 4:1: System Architecture

4.4 System Design

4.4.1 Use Case Diagram

A use case diagram is used to describe the functionality of a system (use cases) and the interaction between the users (actors) and the system. Use cases are used in the analysis to articulate the high-level requirements of the system. Figure 4.2 illustrates the interactions between various actors and the proposed differential diagnostic tool.

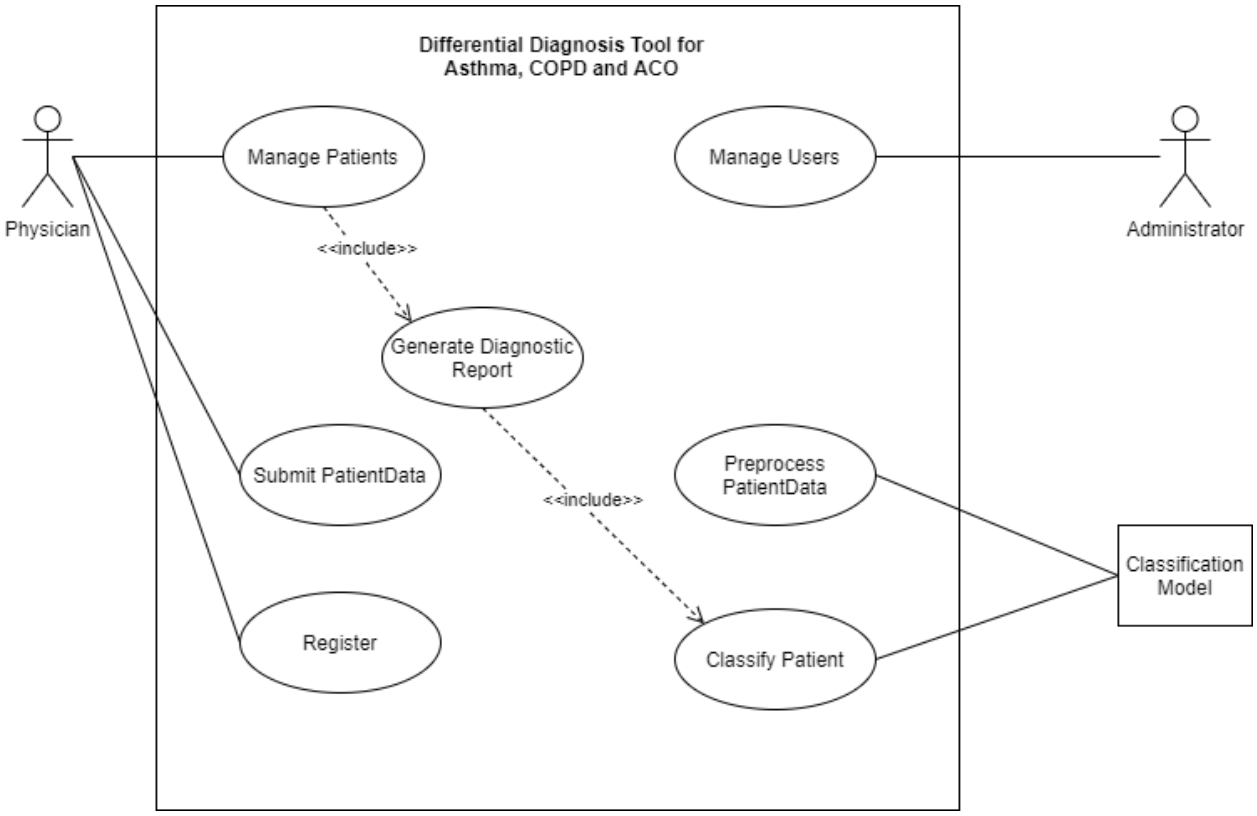


Figure 4:2: Use Case Diagram

The definition of actors in relation to the roles of the tool available to them is shown in Table 4.1 below.

Table 4:1: Use Case Description

Actor	**Use Case**	**Description**
User - Physician	Register	To use the application, a potential user must first register by providing the necessary information.
	Submit PatientData	A registered user is able to input data and submit it to the tool.
	Manage Patients	A registered user can provide further observations and treatment once results are displayed and edit this information.
Administrator	Manage Users	The administrator can view and update users.

4.4.2 Sequence Diagram

Figure 4.3 shows the various interactions between the internal components. It illustrates in a sequential manner how the system achieves its specific objective. The physician logins to the tool and submits a patient's data. Once he\she submits this data, it is run through the model. A recommended diagnosis is then displayed to the physician who can edit or update it to include comments and treatment.

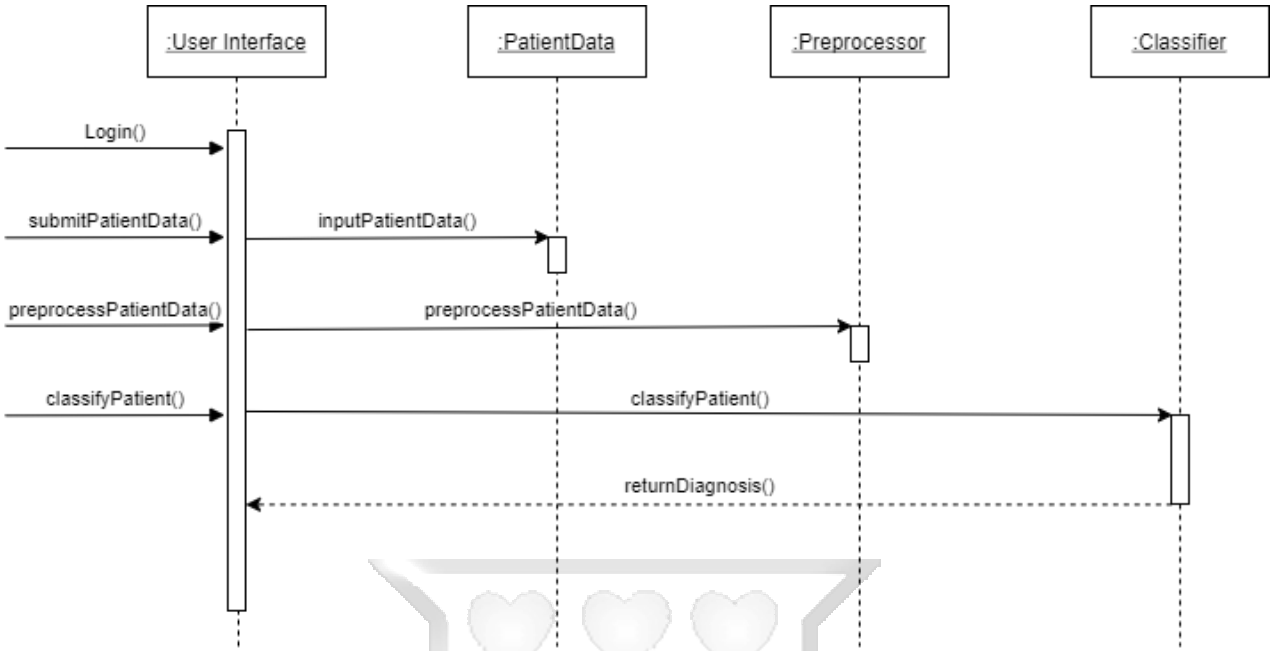


Figure 4:3: Sequence Diagram

4.4.3 Entity Relationship Diagram

Figure 4.4 below shows the entity relationship diagram for the system, illustrating the relationship between the entities in the database. It also highlights the constraints for the tool to ensure reliability and accuracy.

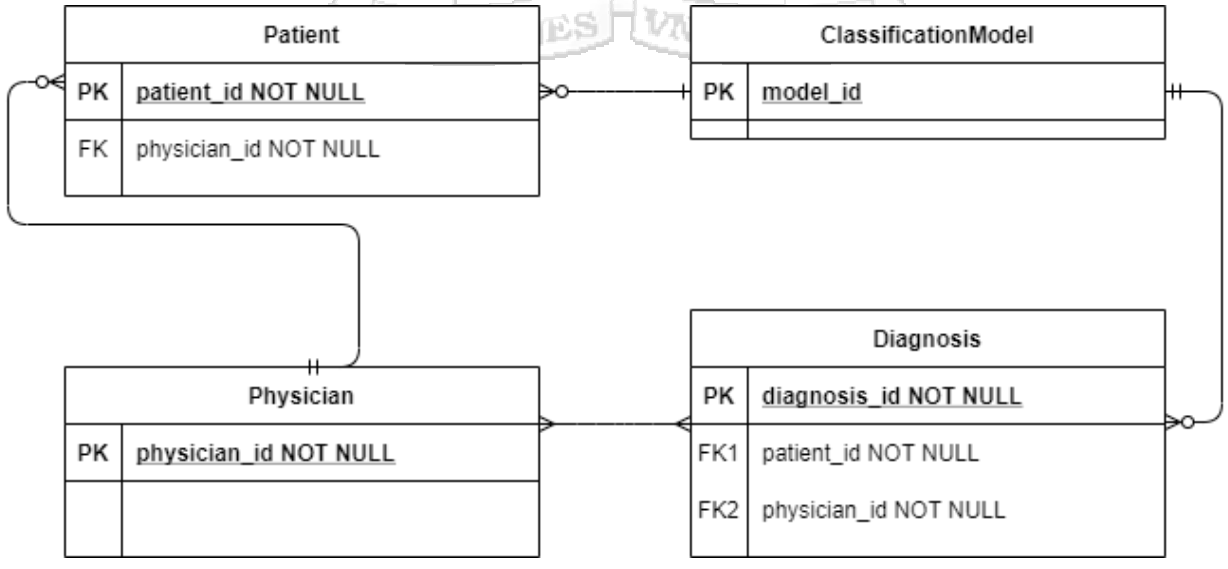


Figure 4:4: Entity Relationship Diagram

4.4.4 Class Diagram

The class diagram represents a static view of the application. A class diagram is not only used for visualization, description and documentation of different aspects of the system but also construction of executable code for the tool or application. It defines the variables and methods contained within it, allowing users to create real-world objects useful for the framework. The classes contain information for the tool's front end and back end. This will allow for expansion or modification of the design class diagrams by future developers when improving the system.

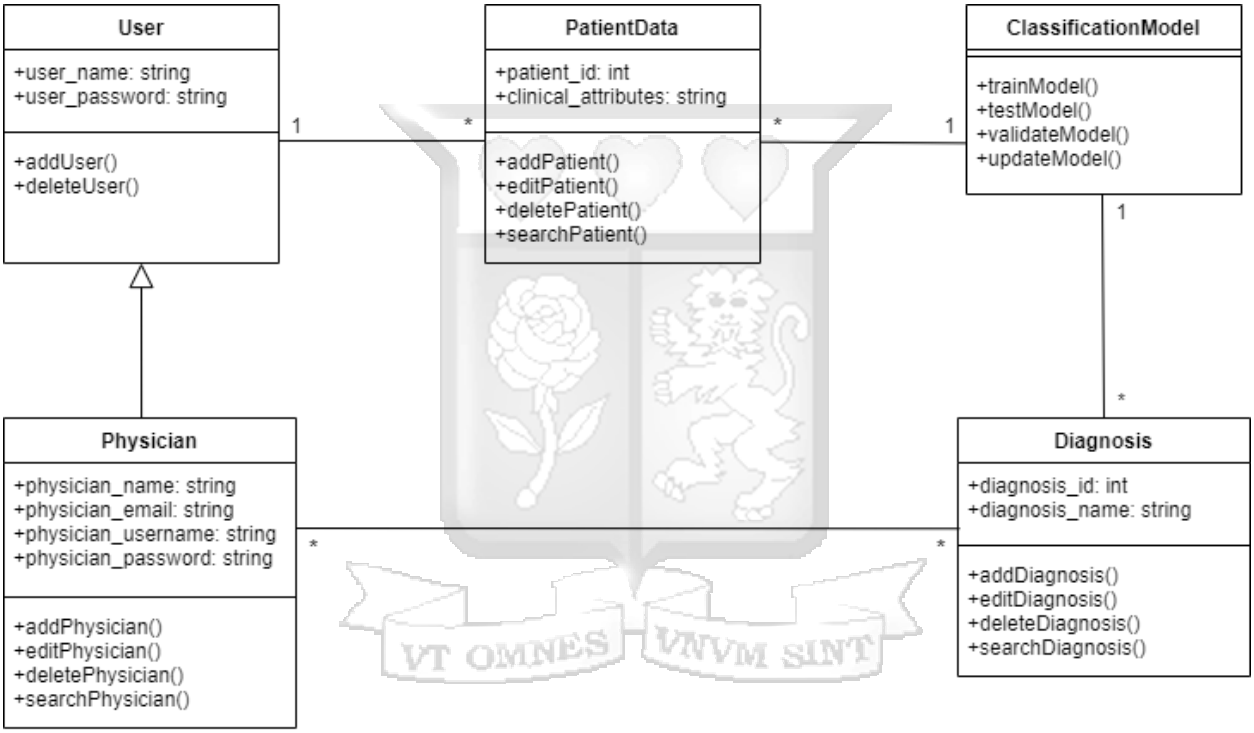


Figure 4:5: Class Diagram

Chapter 5 : System Implementation and Testing

5.1 Introduction

System implementation and testing details the sets of procedures followed to complete the design of the system approved in system analysis and design. The study aimed at developing a tool for the differential diagnosis of asthma, COPD and ACO. Subsequent sections outlined in detail the overall layout of the tool, hardware and software requirements employed to achieve this goal. In addition, it outlines how the model was developed, tested and evaluated to satisfy the tool's functionality requirements. Python was employed as the core development language for the tool. Functional type of testing was carried out to ascertain that the tool achieved its intended goal.

5.2 Hardware and Software Environment

The development of the model was undertaken on a personal computer with Python as the core development language. Python provides a vast number of libraries that were useful in the development of the model including Scikit-Learn and Pandas. Pandas is useful for data manipulation and analysis while Scikit-Learn contains a lot of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction. Jupyter notebook was used as the environment for model development whereas PyCharm was used for the application interface. The hardware and software specifications have been summarized in Table 5.1 below.

Table 5:1: Hardware and Software Specifications

Hardware	**Details**	
Personal Computer	RAM	4.00 GB
	CPU	Core i3
	HDD	500GB
Software	**Libraries**	**Version**
Python 3.9	Scikit-Learn	0.24.1
	Pandas	1.2.4
	Numpy	1.20.2

5.3 Model Development

5.3.1 Data Pre-processing

The data was loaded into a Pandas data frame. Ages in the dataset were divided into five bins; infant, toddler, kid, teens and adult to cross-check if there were different categories. This confirmed that the data was only for adults aged 20 and above.

```
# first lets look if these ages in the dataset are either a Infant(0-2), Toddler (2-4), Kid(4-13), Teens(13-20), Adult(20-11-)

bins= [0,2,4,13,20,110]
labels = ['Infant','Toddler','Kid','Teen','Adult']
data['AgeGroup'] = pd.cut(data['AGE'], bins=bins, labels=labels, right=False)
```

Figure 5:1: Age Cross-Check

A correlation matrix was then created to measure the linear relationships between the variables. Outliers were eliminated for FEV_1/FVC and eosinophil variables as shown in Figure 5.2 below and an encoded variable assigned to the labels.

```
data = data[~(data['FEV1/FVC'] <= 0.55)]
data = data[~(data['FEV1/FVC (%)'] <= 50.0)]
data = data[~(data['EOSINOPHILS (%)'] >= 8.0)]
print(np.shape(data))
```

Figure 5:2: Correlation Matrix

5.3.2 Model Training and Testing

The data was split as follows; training – 70% and testing - 30%, then initially trained with a K-nearest neighbour classifier. Due to the dataset's limitation in size and imbalanced nature, Synthetic Minority Oversampling Technique (SMOTE) was employed. This technique generates synthetic data for the minority class. It works by randomly picking a point from the minority class and computing the k-nearest neighbours for this point. The synthetic points are added between the chosen point and its neighbours.

```
from imblearn.over_sampling import SVMSMOTE

print('Original dataset shape %s' % Counter(y))

svm_sm = SVMSMOTE(random_state=42)

X_res_svm_sm, y_res_svm_sm = svm_sm.fit_resample(X, y)
print('Resampled dataset shape %s' % Counter(y_res_svm_sm))
```

Figure 5:3: SMOTE Sampling

The model was also trained with logistic regression, support vector classifier, random forest classifier and multilayer perceptron as shown in Figure 5.4 below.

```
LR = LogisticRegression(random_state=0, solver='lbfgs', multi_class='multinomial').fit(X_train, y_train)
y_pred_lr = LR.predict(X_test)
print(f"LogisticRegression: ",round(LR.score(X_test,y_test), 4))

SVM =svm.SVC(decision_function_shape='ovo').fit(X_train, y_train)
y_pred_svm = SVM.predict(X_test)
print(f"SVC: ", round(SVM.score(X_test, y_test), 4))

RF = RandomForestClassifier(n_estimators=1000, max_depth=10, random_state=0).fit(X_train, y_train)
y_pred_rf = RF.predict(X_test)
print(f"RandomForestClassifier: ", round(RF.score(X_test, y_test), 4))

NN = MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(150, 10), random_state=1).fit(X_train, y_train)
y_pred_nn = NN.predict(X_test)
print(f"MLPClassifier: ",round(NN.score(X_test, y_test), 4))
```

Figure 5:4: All Classifiers Modelling

The hyper-parameters of the multilayer perceptron classifier were tuned and the model retrained in a grid search method. 5-folds were used for cross-validation. Grid search is a tuning technique that attempts to compute the best hyper-parameter values.

```
mlp_gs = MLPClassifier(max_iter=100)
parameter_space = {
    'hidden_layer_sizes': [(10,30,10),(20,)],
    'activation': ['tanh', 'relu'],
    'solver': ['sgd', 'adam', 'lbfgs'],
    'alpha': [0.0001, 0.05],
    'learning_rate': ['constant','adaptive'],
}

from sklearn.model_selection import GridSearchCV

clf = GridSearchCV(mlp_gs, parameter_space, n_jobs=-1, cv=5)
clf.fit(X_res_svm_sm, y_res_svm_sm) # X is train samples and y is the corresponding labels
```

Figure 5:5: Multilayer Perceptron Modelling

The hyper-parameters of the logistic regression and random forest classifier were also optimized and the models retrained using a grid search. For cross-validation, each model was trained 5 times for each of 20 candidates, totalling 100 fits. A random forest regressor was also introduced. Since a random forest regressor works with data having a numeric or continuous output and they cannot be defined by classes, it was converted to a classification problem by converting the quantity to be predicted into buckets as follows; ACO: output >= 0 and output < 1, Asthma: output >= 1 and output < 2, COPD: output >= 2. The random forest regressor model was trained 3 times for each of 100 candidates, totalling 300 fits.

```
def rand_forest_grid_search(X_res_ada, y_res_ada, n_estimators = [int(x) for x in np.linspace(start = 20, stop = 200, num = 10)],
                                        max_features = ['auto', 'sqrt'],
                                        max_depth = [int(x) for x in np.linspace(10, 80, num = 11)],
                                        min_samples_split = [2, 5, 10],
                                        min_samples_leaf = [1, 2, 4], bootstrap = [True, False]):
    param_grid = {'n_estimators': n_estimators,
                  'max_features': max_features,
                  'max_depth': max_depth,
                  'min_samples_split': min_samples_split,
                  'min_samples_leaf': min_samples_leaf,
                  'bootstrap': bootstrap}

    global best_model
    rf = RandomForestRegressor()

    grid_search = GridSearchCV(estimator = rf, param_grid = param_grid,
                               cv = 3, n_jobs = -1, verbose = 1)
    best_model = grid_search.fit(X_res_svm_sm, y_res_svm_sm)
    print(best_model.best_estimator_)
    print("The mean accuracy of the model is:",best_model.score(X_res_svm_sm, y_res_svm_sm))

def execute_pipeline(X_res_svm_sm, y_res_sm, search_space=[
                {"classifier": [LogisticRegression()],
                 "classifier__penalty": ['l2','l1'],
                 "classifier__C": np.logspace(0, 4, 10)
                 },
                {"classifier": [LogisticRegression()],
                 "classifier__penalty": ['l2'],
                 "classifier__C": np.logspace(0, 4, 10),
                 "classifier__solver":['newton-cg','saga','sag','liblinear'] ##This solvers don't allow L1 penalty
                 },
                {"classifier": [RandomForestClassifier()],
                 "classifier__n_estimators": [10, 100, 1000],
                 "classifier__max_depth":[5,8,15,25,30,None],
                 "classifier__min_samples_leaf":[1,2,5,10,15,100],
                 "classifier__max_leaf_nodes": [2, 5,10]}], cv=5, verbose=0, n_jobs=-1):
    global best_model

    pipe = Pipeline([("classifier", RandomForestClassifier())])

    gridsearch = GridSearchCV(pipe, search_space, cv=cv, verbose=verbose,n_jobs=n_jobs)

    # Fit grid search
    best_model = gridsearch.fit(X_res_svm_sm, y_res_svm_sm)
    print(best_model.best_estimator_)
    print("The mean accuracy of the model is:",best_model.score(X_res_svm_sm, y_res_svm_sm))
```

Figure 5:6: Logistic Regression, Random Forest Classifier & Regressor Modelling

5.3.3 Model Evaluation and Selection

Model evaluation is an integral phase of the model development process. It is the process of assessing the performance of a given model to a given metric with the aim of estimating the generalization accuracy of a model on future unseen or out-of-sample data. There are a number of metrics commonly used for classification problems including accuracy, confusion matrix, precision, recall among others. Accuracy indicates how many correct classifications were made from the total number of classifications. The model configurations developed were evaluated against performance metrics and the best model chosen.

5.4 System Testing

Table 5.2 below shows the functional tests conducted for the registration, log in, classification and management components of the tool.

Table 5:2: Functional Tests

No.	Component	Pre-requisite	Test Case	Expected Results	Observed Results
1	Registration	User has never registered before	User provides required information	a) Account information is saved b) User is taken back to login page	OK
2	Sign/Log in	User account must exist	User enters username and password	User is redirected to the home/landing page	OK
3	Classification	User captures patient data	User captures patient data and submits	Tool/application gives patient information summary including a diagnostic recommendation	OK
4	Manage Patients	Patient is classified	User saves patient data	a) The application saves patient information and physician's summary. b) The physician can edit this information.	OK

5.5 System Modules

The system aimed at providing real-time classification of a patient's condition based on three conditions. Modules that were created include registration, login, classification and management modules.

5.5.1 Registration Module

The registration module allows a new user to create their account as per the set guidelines.

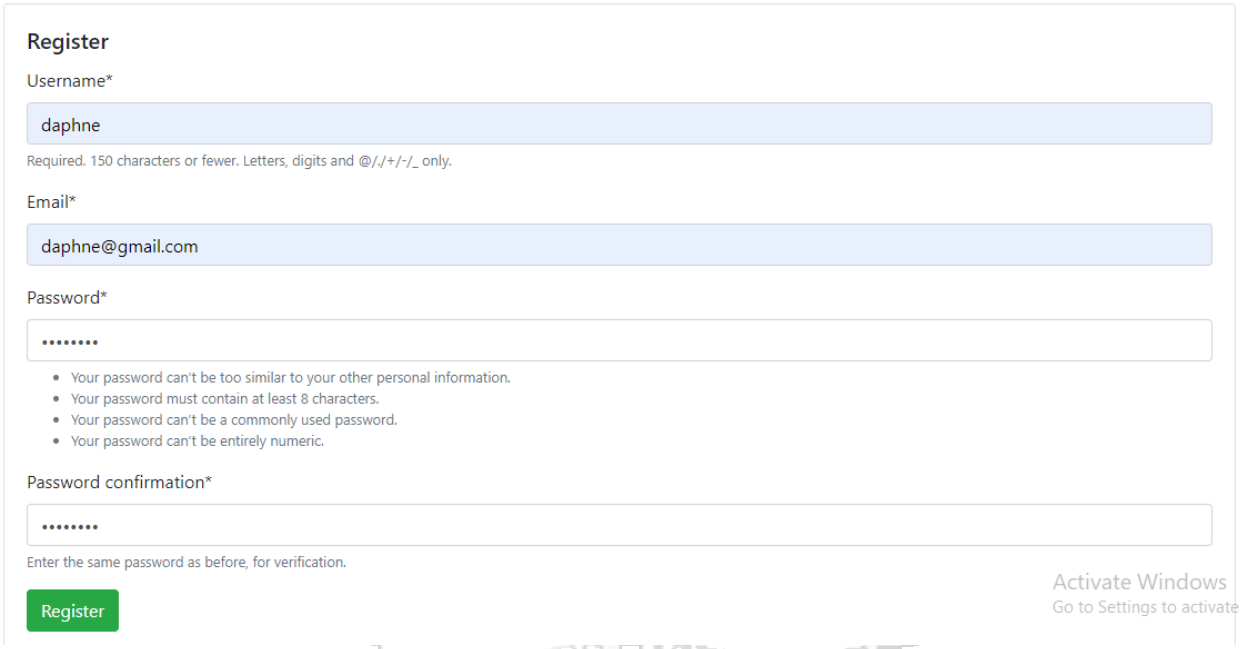


Figure 5:7: Registration Module

5.5.2 Sign/Log In Module

Upon registration, a user is redirected to the log in page to access the tool.

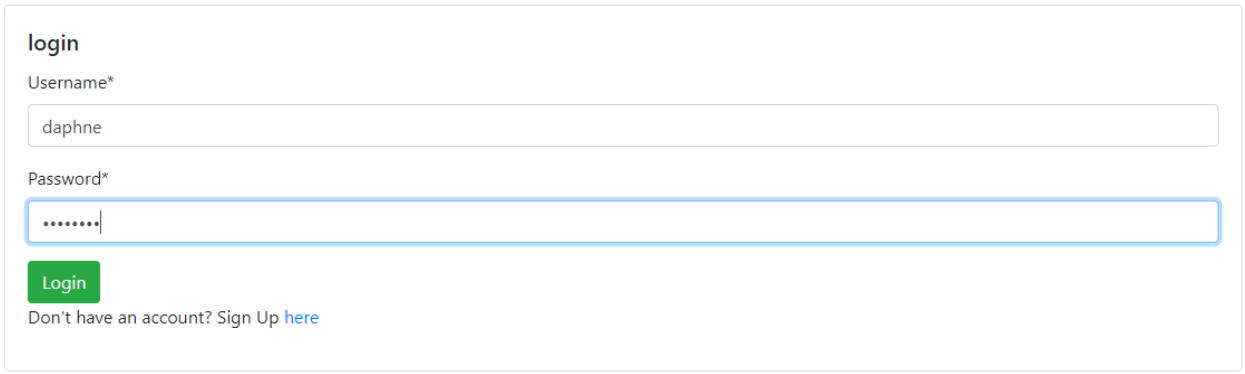


Figure 5:8: Sign/Log In Module

5.5.3 Classification Module

This module allows the user to input patient data and submit to the classification model and then displays the feedback.

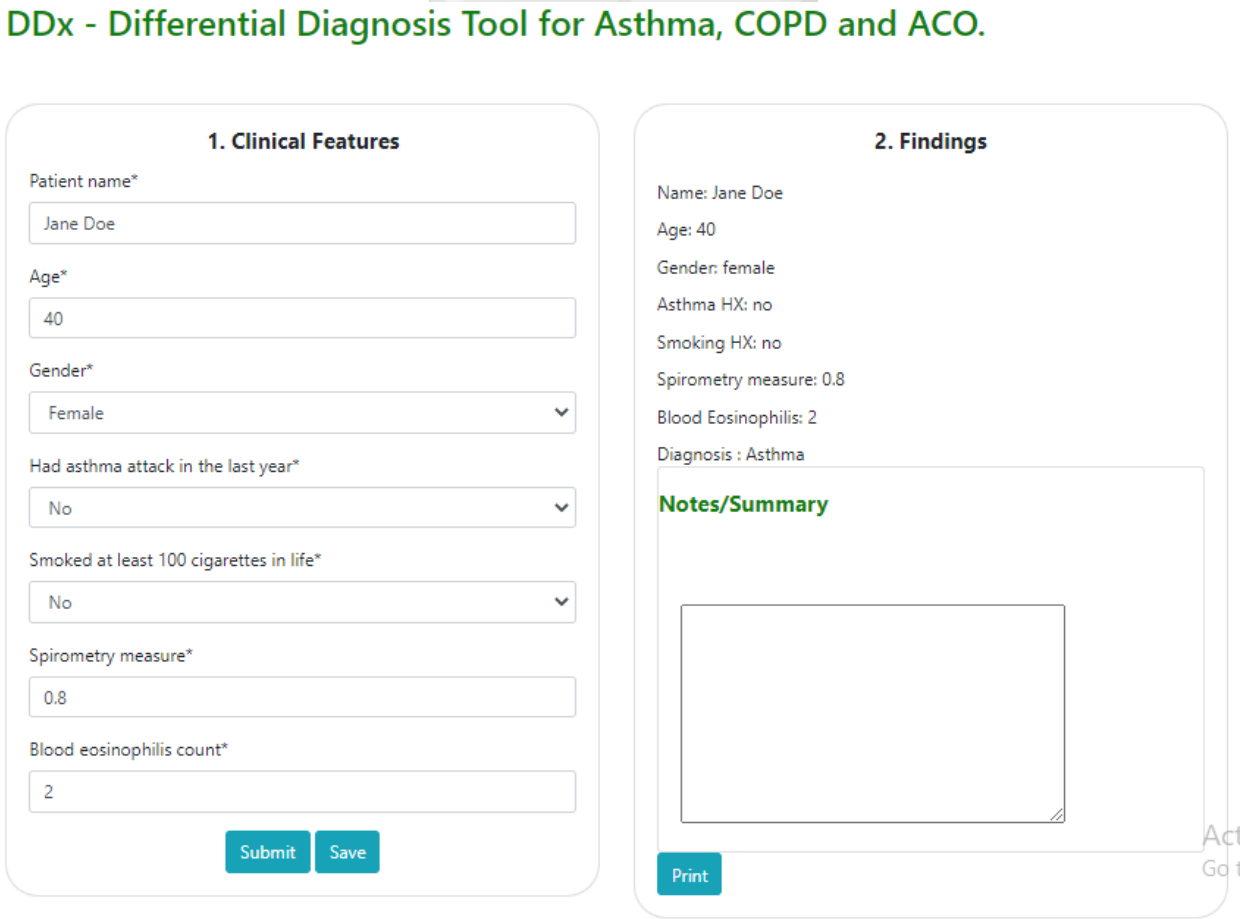


Figure 5:9: Classification Module

5.5.4 Management Module

This module allows the user to view previously diagnosed patients whose records are stored on the system. The user can choose to edit information from this module.

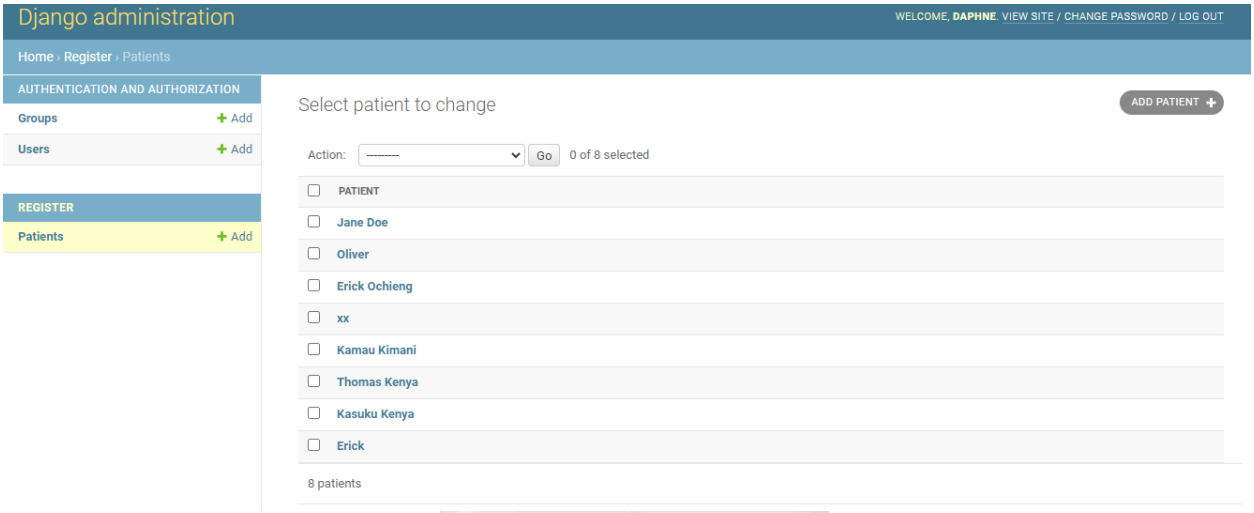


Figure 5:10: Management Module

Chapter 6 : Discussions

6.1 Introduction

The purpose of this was to develop an understanding of obstructive lung diseases leading to the development of a differential diagnostic tool for asthma, COPD and ACO using classification models. This chapter includes a discussion of major findings as related to the literature on obstructive lung diseases and what implications may be valuable for use in the differential diagnosis of these conditions. The chapter concludes with a discussion of the limitations of the study.

6.2 Results

There have been several studies focusing on differential diagnosis of obstructive lung diseases using various types of classifiers including K-nearest neighbour, support vector machines, artificial neural networks, fuzzy logic, latent class analysis, extreme gradient boosting among others. These studies have applied different structures for various respiratory diseases diagnosis, including asthma, COPD and ACO, using datasets with various features. Consistent with a number of studies including Kaplan et al (2020) previous' research in this area, a differential diagnostic tool for asthma, COPD and ACO based on classification models was developed.

Datasets obtained from the NHANES database (2011-2012), were pre-processed, trained and tested using a number of machine learning models namely K-nearest neighbours, logistic regression, random forest, support vector machines and multilayer perceptron. The training set was such that we needed to use data for three different classes to be able to validate the model. The testing set was used to ensure that the model's output values were nearly identical to the values initially in the dataset. Performance metrics that were used in the evaluation of these models are accuracy, precision, recall and F1-score. These metrics did not only help in selecting the best model but also in validating the model. Theoretical underpinnings of learning differ between machine learning methods. As a result, the variety of experiments performed aided in the selection of the best model.

K-nearest neighbour model achieved an accuracy of 68.08% as shown in Figure 6.1 below in a fast and non-explicit initial training phase. K-nearest neighbour is usually considered a lazy learning, non-parametric algorithm.

Figure 6:1: K-Nearest Neighbour Performance

The K-nearest neighbour algorithm presents a number of advantages and disadvantages for the tool. As a lazy learner it has no training period. It neither learns anything in the training period nor derives any discriminative function from the training data. It stores the training dataset and learns from it only at the time of making real time predictions. This makes the algorithm much faster than other algorithms that require training e.g. logistic regression, SVM etc. Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm and therefore easy to implement (Kumar, 2019). The algorithm generalized fairly well since the data set was also limited in size and dimension. However, there is a chance that it could have also made wrong predictions since the dataset was not standardized and scaled before training. The present results achieve lower accuracy in comparison to Vora and Shah's (2019) work that dealt with comparative study between SVM and K-nearest neighbour to perform classification of COPD severity level.

The model was also trained against logistic regression, support vector classifier, random forest classifier and multilayer perceptron classifier. The performances of these models have been summarized in Table 6.1 below.

Table 6:1: All Classifiers Performance

Classifier	**Accuracy (%)**
Logistic Regression	65.75
Support Vector Machine	46.58
Random Forest Classifier	76.71
Multilayer Perceptron	43.84

The support vector classifier achieved a very low accuracy in classifying asthma, COPD and ACO as compared to an accuracy of 96.97% achieved by Vora and Shah (2019) in classification of COPD severity level. Support vector classifiers are more effective in small or medium-sized datasets with a high dimensional space. This dataset was small in size. It however had a low dimensional space that could possibly have contributed to the low performance of the model. At this point, the support vector classifier was dropped and the multilayer perceptron classifier was subjected to further training after the hyper-parameters were optimized.

Precision, recall and F1-score metrics were used to measure the performance of the multilayer perceptron classifier after further optimization. The performance of this model is summarised in Table 6.2. Whereas Kaplan et al. (2020) achieved F1-scores of 0.98, 0.98 and 0.84 using a Bayesian algorithm, the present study achieves F1-scores of 0.54, 0.63 and 0.65 in diagnosing COPD, asthma and ACO respectively. Multilayer perceptron classifiers are capable of generalization. They identify an unknown pattern by comparing it to other patterns with similar characteristics. This means that noisy or incomplete inputs will be classified based on their resemblance to pure and complete inputs. The prediction accuracy based on this classifier was however unreliable owing to the size of the dataset.

Table 6:2: Multilayer Perceptron Performance

	Precision	**Recall**	**F1-Score**
ACO	0.65	0.50	0.57
Asthma	0.63	0.53	0.58
COPD	0.54	0.93	0.68

Further experiments were conducted with logistic regression and random forest classifier and regressor. These models were also subjected to hyper-parameter tuning using a grid search method and cross-validated to ensure optimum performance. Logistic regression was considered as it does not require the input features to be scaled, is highly interpretable and is easy to regularize. It is also easy and very efficient to train. Logistic regression is a good base line to measure the performance of a more complex algorithm like random forest because of its simplicity and the fact that it can be implemented very quickly. However, it is also an algorithm known for its vulnerability to overfitting (Bertsimas & King, 2017). It reported an accuracy of 68.87% as shown in Table 6.3 below.

Table 6:3: Logistic Regression, Random Forest Classifier and Regressor Performance

Classifier	**Accuracy (%)**
Logistic Regression	68.87
Random Forest Classifier	82.15
Random Forest Regressor	93.94

Random forest classifier reported an accuracy of 82.15%. The classifier reduces overfitting in decision trees and helps in improving the accuracy. The random forest regressor proved to be the best model, reporting an accuracy of 93.94%. Despite this advantage, the random forest model also has some drawbacks. Firstly, it requires a significant amount of computational power and resources as it builds numerous trees to combine their outputs. Secondly, it takes a long time to train because it combines many decision trees to determine the class. Lastly, due to the ensemble of decision trees, it is difficult to interpret and fails to

determine the importance of each variable (Dogru & Subasi, 2018). These results are consistent with the claim that the application of computer-based medical diagnostic techniques has improved the quality of healthcare (Badnjevic et al., 2018).

Chapter 7: Conclusion and Recommendations

7.1 Conclusions

For decades, we have faced challenges in diagnosing and managing various chronic respiratory diseases due to a population lifestyle that includes smoking and air pollution, among other factors. Standardized procedures for the diagnosis and management of these diseases are available, but these procedures have yet to produce results that significantly reduce mortality and morbidity rates. Consequently, new and innovative approaches to dealing with this type of disease are required.

This study intended to develop a differential diagnostic tool for asthma, COPD and ACO using classification models as such a tool would be useful in reducing cases of misdiagnosis. To enable the successful implementation of the tool, it was necessary to understand the characteristics of these conditions. Additionally, a critical understanding and review of the steps and tools applied and challenges encountered in diagnosing the conditions was necessary. Considering the success of classification algorithms had in other similar studies, it was anticipated to yield similar results when applied to this specific study.

During development of the tool, the models were trained with 184 samples acquired from 2011 to 2012 in the NHANES database. Classification techniques used were K-nearest neighbour, support vector machines, logistic regression, multilayer perceptron and random forest. Performance metrics that were used in the evaluation of these models are accuracy, precision, recall and F1-score. K-Nearest Neighbour model achieved an accuracy of 68.08% in a fast and non-explicit initial training phase. Despite the dataset's limitation in size and dimension, a non-contributing factor to the low performance of the K-Nearest Neighbour algorithm, the dataset was not scaled or standardized at this point.

Logistic regression, support vector machines, random forest and multilayer perceptron reported an accuracy of 65.75%, 46.58%, 76.71% and 43.84% respectively. At this point, the support vector classifier was dropped and the multilayer perceptron classifier was subjected to further training after hyper-parameter tuning. The classifier reported precision values of 0.65, 0.63 and 0.54 for ACO, asthma and COPD respectively. Further experiments were conducted with logistic regression, random forest classifier and random forest regression. These models were also subjected to hyper-parameter tuning using a grid search method and cross-validated to ensure optimum performance. Random forest regressor proved to be the best model, reporting an

accuracy of 93.94%. It is considered a powerful model due to its ability to limit overfitting without substantially increasing error due to bias. With the random forest model, the tool was developed and tested to incorporate real-time diagnosis. These results show an automated approach in the differential diagnosis of obstructive lung diseases that would aid physicians in making preliminary diagnoses, leading to optimization of time resources, lower medical device costs and better patient outcomes.

7.2 Limitations of the Study

There are at least three potential limitations concerning the results of this study. A first limitation concerns the population setting of the data collected. The NHANES database consists of only data collected in the United States. It would be beneficial to test the models' output on data obtained from different populations. A second potential limitation is that the current model is limited in dataset size and therefore there could be a probability that the model is over fitted. Lastly, evaluation was not performed in liaison with a physician in a real clinical setting. Despite these limitations, the present study has enhanced our understanding of the relationship between asthma, COPD and ACO. The research can be seen as one among the many steps towards improving the differential diagnosis of obstructive lung diseases.

7.3 Recommendations

Based on the research findings, the researcher recommends the following actions to be taken:

i. Train the model with a more robust dataset including more features such as signs and symptoms and results from other objective tests other than spirometry and blood eosinophil count that could be important biomarkers in the diagnosis of the conditions.
ii. Study locations for data collection be expanded to other geographical areas.
iii. Healthcare institutions in partnership with policy–makers should make de-identified patient data available on online databases to facilitate the development of more robust healthcare systems.
iv. Efficient training of health officials on spirometry examination to improve results.
v. To facilitate the ease of adoption of the tool, the project should be carried out in liaison with physicians in a real clinical setting.

7.4 Suggestions for Future Research

In terms of future research, it would be useful to extend the current findings by examining the following:

i. In cases where it is adopted by a hospital, it can be integrated or plugged into an existing hospital information management system as a subsystem.
ii. It can be extended to also include younger patients and other similar lung diseases.
iii. It can be designed from a patient's perspective where features required to do a differential diagnosis are not limited to spirometry examinations and other lab tests.

References

Abidine, M. B., Fergani, B., & Ordóñez, F. J. (2016). Effect of over-sampling versus under-sampling for SVM and LDA classifiers for activity recognition. *International Journal of Design & Nature and Ecodynamics*, *11*(3), 306-316.

Abramson, M. J., Perret, J. L., Dharmage, S. C., McDonald, V. M., & McDonald, C. F. (2014). Distinguishing adult-onset asthma from COPD: a review and a new approach. *International journal of chronic obstructive pulmonary disease*, *9*, 945.

Al Sallakh, M. A., Rodgers, S. E., Lyons, R. A., Sheikh, A., & Davies, G. A. (2018). Identifying patients with asthma-chronic obstructive pulmonary disease overlap syndrome using latent class analysis of electronic health record data: a study protocol. *NPJ primary care respiratory medicine*, *28*(1), 1-4.

Alba-Cabrera, E., Godoy-Calderon, S., & Ibarra-Fiallo, J. (2016). Generating synthetic test matrices as a benchmark for the computational behavior of typical testor-finding algorithms. *Pattern Recognition Letters*, *80*, 46-51.

Asp, K. (2020, July 22). Obstructive Lung Disease vs Restrictive Lung Disease: Causes, Diagnosis, and Treatment Options [web log].

Ayash, M. M. (2014). Research methodologies in computer science and information systems. *Computer Science*, *2014*, 1-4.

Badnjevic, A., Gurbeta, L., & Custovic, E. (2018). An expert diagnostic system to automatically identify asthma and chronic obstructive pulmonary disease in clinical settings. *Scientific reports*, *8*(1), 1-9.

Bell, D. (2020). HRCT chest | *Radiology Reference Article* | Radiopaedia.org. Retrieved 19 July 2020, from https://radiopaedia.org/articles/hrct-chest-1

Bellingegni, A. D., Gruppioni, E., Colazzo, G., Davalli, A., Sacchetti, R., Guglielmelli, E., & Zollo, L. (2017). NLR, MLP, SVM, and LDA: a comparative analysis on EMG data from people with trans-radial amputation. *Journal of neuroengineering and rehabilitation*, *14*(1), 82.

Bertsimas, D., & King, A. (2017). Logistic regression: From art to science. *Statistical Science*, 367-384.

Bui, D. S., Lodge, C. J., Burgess, J. A., Lowe, A. J., Perret, J., Bui, M. Q., ... & Hamilton, G. S. (2018). Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life. *The lancet Respiratory medicine*, *6*(7), 535-544.

Chikalov, I., Lozin, V., Lozina, I., Moshkov, M., Nguyen, H. S., Skowron, A., & Zielosko, B. (2012). *Three approaches to data analysis: Test theory, rough sets and logical analysis of data* (Vol. 41). Springer Science & Business Media.

Cook, C. E., & Décary, S. (2020). Higher order thinking about differential diagnosis. *Brazilian journal of physical therapy*, *24*(1), 1-7.

de Benedictis, F. M., & Attanasi, M. (2016). Asthma in childhood. *European Respiratory Review*, *25*(139), 41-47.

De Marco, R., Pesce, G., Marcon, A., Accordini, S., Antonicelli, L., Bugiani, M., ... & Pirina, P. (2013). The coexistence of asthma and chronic obstructive pulmonary disease (COPD):

prevalence and risk factors in young, middle-aged and elderly people from the general population. *PloS one*, *8*(5), e62985.

Dogru, N., & Subasi, A. (2018, February). Traffic accident detection using random forest classifier. In *2018 15th learning and technology conference (L&T)* (pp. 40-45). IEEE.

Donner-Banzhoff, N. (2018). Solving the diagnostic challenge: a patient-centered approach. *The Annals of Family Medicine*, *16*(4), 353-358.

Fang, L., Gao, P., Bao, H., Tang, X., Wang, B., Feng, Y., ... & Wang, N. (2018). Chronic obstructive pulmonary disease in China: a nationwide prevalence study. *The Lancet Respiratory Medicine*, *6*(6), 421-430.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

Garfinkel, S. L. (2015). De-identification of personal information. *National institute of standards and technology*.

Gentry, S., & Gentry, B. (2017). Chronic obstructive pulmonary disease: diagnosis and management. *American Family Physician*, *95*(7), 433-441.

Ghebre, M. A., Bafadhel, M., Desai, D., Cohen, S. E., Newbold, P., Rapley, L., ... & Burton, P. R. (2015). Biological clustering supports both "Dutch" and "British" hypotheses of asthma and chronic obstructive pulmonary disease. *Journal of Allergy and Clinical Immunology*, *135*(1), 63-72.

Global Initiative for Asthma and Global Initiative for Chronic Obstructive Lung Disease. (2017). *Diagnosis and Initial Treatment of Asthma, COPD And Asthma-COPD Overlap.*

Global Initiative for Asthma. (2019). *Global Strategy for Asthma Management and Prevention (2019 update).*

Global Initiative for Chronic Obstructive Lung Disease. (2020). *Global Strategy for The Diagnosis, Management and Prevention of Chronic Obstructive Pulmonary Disease*.

Goel, A., & Mahajan, S. (2017). Comparison: KNN & SVM Algorithm. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), 5*.

González-López, J. J., García-Aparicio, Á. M., Sánchez-Ponce, D., Muñoz-Sanz, N., Fernandez-Ledo, N., Beneyto, P., & Westcott, M. C. (2016). Development and validation of a Bayesian network for the differential diagnosis of anterior uveitis. *Eye*, *30*(6), 865-872.

Grimm, L. (2016). Asthma Imaging. *The Medscape Journal of Medicine*. Retrieved from https://emedicine.medscape.com/article/353436-overview

Gupta, D. (2014). Anthropometry and the design and production of apparel: an overview. In *Anthropometry, Apparel Sizing and Design* (pp. 34-66). Woodhead Publishing.

Hallit, S., Raherison, C., Waked, M., & Salameh, P. (2017). Association between caregiver exposure to Toxics during pregnancy and childhood-onset asthma: a case-control study. *Iranian Journal of Allergy, Asthma and Immunology*, *16*(6), 488-500.

Howard, R., Rattray, M., Prosperi, M., & Custovic, A. (2015). Distinguishing asthma phenotypes using machine learning approaches. *Current allergy and asthma reports*, *15*(7), 38.

Inoue, H., Nagase, T., Morita, S., Yoshida, A., Jinnai, T., & Ichinose, M. (2017). Prevalence and characteristics of asthma–COPD overlap syndrome identified by a stepwise approach. *International journal of chronic obstructive pulmonary disease*, *12*, 1803.

Johnson, J. D., & Theurer, W. M. (2014). A stepwise approach to the interpretation of pulmonary function tests. *American family physician*, *89*(5), 359-366.

Kalhan, R., Dransfield, M. T., Colangelo, L. A., Cuttica, M. J., Jacobs Jr, D. R., Thyagarajan, B., ... & Washko, G. R. (2018). Respiratory symptoms in young adults and future lung disease. The CARDIA Lung Study. *American journal of respiratory and critical care medicine*, *197*(12), 1616-1624.

Kaplan, A., Cao, H., Fitzgerald, J. M., Yang, E., Iannotti, N., Kocks, J. W. H., ... & Tsiligianni, I. (2020). Asthma/COPD Differentiation Classification (AC/DC): Machine Learning to Aid Physicians in Diagnosing Asthma, COPD and Asthma-COPD Overlap (ACO). In *D22. COMORBIDITIES IN PEOPLE WITH COPD* (pp. A6285-A6285). American Thoracic Society.

Kashanian, M., Mohtashami, S. S., Bemanian, M. H., Moosavi, S. A. J., & Moradi Lakeh, M. (2017). Evaluation of the associations between childhood asthma and prenatal and perinatal factors. *International Journal of Gynecology & Obstetrics*, *137*(3), 290-294.

Kumar, N. (n.d.). Advantages and Disadvantages of KNN Algorithm in Machine Learning [Web log post]. Retrieved February 23, 2019.

Langan, R. C., & Goodbred, A. J. (2020). Office Spirometry: Indications and Interpretation. *American Family Physician*, *101*(6), 362-368.

Lomax, R. G., & Hahs-Vaughn, D. L. (2013). *An introduction to statistical concepts*. Routledge.

Manian, P. (2019). Chronic obstructive pulmonary disease classification, phenotypes and risk assessment. *Journal of Thoracic Disease*, *11*(Suppl 14), S1761.

Maravic, M., Sigogne, R., Bourdin, A., Roche, N., Mounir, S., Milic, D., ... & Bacry, E. (2019). Differentiating asthma from chronic obstructive pulmonary disease (COPD) in medico-economic databases: myth or reality?.

McGeachie, M. J., Yates, K. P., Zhou, X., Guo, F., Sternberg, A. L., Van Natta, M. L., ... & Cho, M. H. (2016). Patterns of growth and decline in lung function in persistent childhood asthma. *New England Journal of Medicine*, *374*(19), 1842-1852.

Medford-Davis, L. N., Singh, H., & Mahajan, P. (2018). Diagnostic decision-making in the emergency department. *Pediatric Clinics*, *65*(6), 1097-1105.

Middleton, F. (2019). Reliability vs validity: what's the difference?. *Available at: https//www. scribbr. com*.

Miravitlles, M. (2017). Diagnosis of asthma–COPD overlap: the five commandments.

Moore, V. C. (2012). Spirometry: step by step. *Breathe*, *8*(3), 232-240.

Network, G. A. The global asthma report 2018. 2018.

Park, S. Y., Jung, H., Kim, J. H., Seo, B., Kwon, O. Y., Choi, S., ... & Won, S. (2019). Longitudinal analysis to better characterize Asthma-COPD overlap syndrome: Findings from an adult asthma cohort in Korea (COREA). *Clinical & Experimental Allergy*, *49*(5), 603-614.

Reddy, L. S., & Ramasamy, D. (2016). Justifying the judgmental sampling matrix organization in outsourcing industry. *Vidushi, July-December*.

Roadmunk. (2019). [Rapid Application Development Methodology]. Retrieved September 06, 2020, from https://roadmunk.com/guides/types-of-software-development-methodologies/

Rodríguez-Diez, V., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & Lazo-Cortés, M. S. (2017, June). Fast-BR vs. fast-CT_EXT: An empirical performance study. In *Mexican Conference on Pattern Recognition* (pp. 127-136). Springer, Cham.

Rogliani, P., Ora, J., Puxeddu, E., & Cazzola, M. (2016). Airflow obstruction: is it asthma or is it COPD?. *International journal of chronic obstructive pulmonary disease*, *11*, 3007.

Sanchez-Diaz, G., Lazo-Cortes, M., & Piza-Davila, I. (2012). A fast implementation for the typical testor property identification based on an accumulative binary tuple. *International Journal of Computational Intelligence Systems*, *5*(6), 1025-1039.

Seltman, H. J. (2015). Experimental design and analysis.

Sharma, G. (2017). Pros and cons of different sampling techniques. *International journal of applied research*, *3*(7), 749-752.

Smith, B. M., Traboulsi, H., Austin, J. H., Manichaikul, A., Hoffman, E. A., Bleecker, E. R., ... & Guo, J. (2018). Human airway branch variation and chronic obstructive pulmonary disease. *Proceedings of the National Academy of Sciences*, *115*(5), E974-E981.

Soriano, J. B., Abajobir, A. A., Abate, K. H., Abera, S. F., Agrawal, A., Ahmed, M. B., ... & Alam, N. (2017). Global, regional, and national deaths, prevalence, disability-adjusted

life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Respiratory Medicine*, *5*(9), 691-706.

Spencer, P., & Krieger, B. (2013). The differentiation of chronic obstructive pulmonary disease from asthma: a review of current diagnostic and treatment recommendations. *The open nursing journal*, *7*, 29.

Taherdoost, H. (2016). Sampling methods in research methodology; how to choose a sampling technique for research. *How to Choose a Sampling Technique for Research (April 10, 2016)*.

Vora, S., & Shah, C. (2019). COPD Classification using Machine Learning Algorithms.

Vos, T., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., ... & Aboyans, V. (2017). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, *390*(10100), 1211-1259.

Weinberger, S. E., Cockrill, B. A., & Mandel, J. (2017). *Principles of Pulmonary Medicine E-Book*. Elsevier Health Sciences.

Appendices

Appendix A: Ethical Certificate

Strathmore UNIVERSITY

16th December 2020

Ms Ochieng', Daphne
daphne.ochieng@strathmore.edu

Dear Ms Ochieng',

RE: A Differential Diagnosis Tool for Obstructive Lung Diseases in Adults Using Classification Models

This is to inform you that SU-IERC has reviewed and **approved** your above research proposal. Your application reference number is **SU-IERC0928/20.** The approval period is **16th December 2020 to 15th December 2021.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) https://oris.nacosti.go.ke and also obtain other clearances needed.

Yours sincerely,

Dr Virginia Gichuru,
Secretary; SU-IERC

Cc: Prof Fred Were,
Chairperson; SU-IERC

STRATHMORE UNIVERSITY INSTITUTIONAL ETHICS REVIEW COMMITTEE SU-IERC
16 DEC 2020
TEL: +254 (0)703 034 000
P. O. Box 59857 - 00200
NAIROBI - KENYA

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email info@strathmore.edu www.strathmore.edu

Appendix B: Ouriginal Report

Ouriginal

Document Information

Analyzed document	A Differential Diagnosis Tool for Obstructive Lung Diseases in Adults Using Classification Models (Signed) - 122694.docx (D103568465)
Submitted	5/2/2021 8:20:00 PM
Submitted by	
Submitter email	daphne.ochieng@strathmore.edu
Similarity	5%
Analysis address	library.strath@analysis.urkund.com

Sources included in the report

W	URL: https://www.researchgate.net/publication/312109952_Identifying_possible_asthma-COP ... Fetched: 10/21/2019 1:02:33 PM	3
W	URL: https://www.nature.com/articles/s41598-018-30116-2 Fetched: 9/26/2019 2:02:13 PM	5