Icon

Description automatically generated****Al Imam Mohammad ibn Saud Islamic University

College of Computer and Information Sciences

Computer Science Department

First Semester 1443 H – 2021 G

**Development of Medical Diagnostic Models Using Artificial Intelligence**

**By**

Ali Khalifa Alhawas (439015852)

Abdulilah Saleh Alqasem (439014750)

Mohammad Abdulaziz bin Zouman (439012798)

**Supervisor**

Dr. Sultan Sharidah Aldera

December 4, 2021

**Abstract**

Medical diagnosis in healthcare today carries the risk of misdiagnosis, with illnesses being diagnosed incorrectly or even delayed, causing patients to be in even greater danger. As a result, a fast, accurate, and automated system for medical diagnosis can be extremely beneficial to doctors and to the whole healthcare system. The goal of this project is to use Artificial Intelligence (AI) to build Medical Diagnostic Models (MDMs). A diagnostic model that mimics clinicians' decision-making with high experience will reduce the chances of diagnostic errors and the likelihood of patients' medical conditions deteriorating. By applying the state of the art of AI in actual medical datasets, we hope that our built MDM can increase the accuracy in the medical decision making.

**Table of Contents**

Chapter One Introduction

[1.1 Introduction 10](#_Toc89030255)

[1.2 Aim and Objectives 12](#_Toc89030256)

[1.2.1 Aim 12](#_Toc89030257)

[1.2.2 Objectives 12](#_Toc89030258)

[1.3 Methodology 13](#_Toc89030259)

[1.4 Project timeline 14](#_Toc89030260)

[1.5 Team Qualifications 15](#_Toc89030261)

[1.6 Conclusion 15](#_Toc89030262)

**Chapter Two Literature Review**

[2.1 Introduction 17](#_Toc89030263)

[2.2 Machine Learning 17](#_Toc89030264)

[2.3 Machine Learning in Medicine 19](#_Toc89030265)

[2.4 COVID-19 Diagnostic Models 22](#_Toc89030266)

[2.5 Conclusion 30](#_Toc89030267)

**Chapter Three Methodology**

[3.1 Introduction 32](#_Toc89030268)

[3.2 Data Collection and Preparation 32](#_Toc89030269)

[3.3 Feature Engineering 34](#_Toc89030270)

[3.4 Model Selection 35](#_Toc89030271)

[3.5 Model Evaluation 41](#_Toc89030272)

[3.6 Conclusion 42](#_Toc89030273)

[References 43](#_Toc89030274)

**List of Tables**

[Table 1: Team Qualifications 15](#_Toc89447044)

[Table 2: Related Work 22](#_Toc89447045)

[Table 3: Strawman imputation. 33](#_Toc89447046)

[Table 4: Feature Engineering Techniques 35](#_Toc89447047)

[Table 5: Performance Metrics 42](#_Toc89447048)

**List of Figures**

[Figure 1: Steps for Building Machine Learning Models 14](#_Toc89447158)

[Figure 2: Project Timeline 14](#_Toc89447159)

[Figure 3: Traditional Programming vs Machine Learning 18](#_Toc89447160)

[Figure 4: Machine Learning Types [7] 19](file:///D:\Desktop\Courses\مشروع1%20-%20CS492\FINAL%20REPORT\FINAL\CS-F-2021-Development_of_Medical_Diagnostic_Models_Using_Artificial_Intelligence.docx#_Toc89447161)

[Figure 5: Logistic Regression Algorithm [28] 36](#_Toc89447162)

[Figure 6: Support Vector Machine Algorithm [30] 37](#_Toc89447163)

[Figure 7: K-Nearest Neighbors Algorithm [32] 38](#_Toc89447164)

[Figure 8: Decision Tree Algorithm [35] 39](#_Toc89447165)

[Figure 9: Random Forest Algorithm [37] 40](#_Toc89447166)

[Figure 10: Simplified Structure of XGBoost Algorithm [40] 41](#_Toc89447167)

**List of Abbreviation**

ARDS Acute Respiratory Distress Syndrome

AUC Area Under the Curve

AI Artificial Intelligence

AST Aspartate Aminotransferase

BN Bayes Net

CRP C-Reactive Protein test

CR Classification via Regression

CLD Color Layout Descriptor

CT Computed Tomography scan

CNN Convolutional Neural Networks

DT Decision Tree

ECG Electrocardiogram

GBDT Gradient Boosted Decision Trees

GBT Gradient Boosted Trees

IBK Lazy Classifier

J48 Decision Tree

KNN K-Nearest Neighbors

LASSO Least Absolute Shrinkage and Selection Operator

LDA Linear Discriminant Analysis

LR Logistic Regression

L% Lymphocyte

ML Machine Learning

MDM Medical Diagnostic Model

MNB Multinomial Naïve Bayesian

NB Naïve Bayes

NN Neural Networks

N% Neutrophil

NYPH New York-Presbyterian Hospital

OSR San Raffaele Hospital

PART Rule Learner

PR Precision Recall

RF Random Forest

ROC Receiver Operating Characteristic curve

RFE Recursive Feature Elimination

RT-PCR Reverse Transcription Polymerase Chain Reaction

RIDGE Ridge Regression

SARS Severe Acute Respiratory Syndrome

SIRM Society of Medical and Interventional Radiology

SVM Support Vector Machine

TF-IDF Term Frequency Inverse-Document Frequency

WCM Weill Cornell Medicine

WBC White Blood Cell count

XGBoost eXtreme Gradient Boosting

*Chapter One*

Introduction

Introduction

# **Introduction**

People with medical conditions rely on doctors to give them the correct medical treatment, using medications, operations, or any kind of treatment the doctors prescribe. The process which doctors use to determine the medical condition of patients is called a medical diagnosis.

Doctors diagnose patients based on a medical feature space that includes symptoms, medical history, physical examination, and testing. Additionally, the diagnosis result is heavily dependent on doctor’s judgment, derived from the doctor’s knowledge and experience.

A diagnosis that is missed, delayed or wrong is known as a diagnostic error. Diagnostic errors if not identified in time can lead patients’ medical conditions to progress to a level that is hard or even impossible to treat. Diagnostic errors are a consequence of human errors.

Misdiagnosis of a patient's medical issues is unfortunately common. According to a U.S. hospital report, diagnostic errors account for 60% of all medical errors [1]. As a result, medical diagnosis is a critical and crucial subject. A delayed diagnosis is one that is made later than it should be. Cancer diagnosis is frequently delayed, which can worsen a patient's health and cause treatment to be delayed [2]. Wrong diagnosis occurs when an earlier diagnosis appears to be erroneous, such as when a patient suffering from a heart attack is diagnosed with heartburn [2], [3]. Missed diagnosis happens in individuals with unexplained diseases, and chronic fatigue or chronic pain patients are at high risk of having their diagnosis missed [2].

According to [4], cancer misdiagnoses occur at an alarming rate; for every 6000 patients, doctors fail to diagnose 71 with cancer, causing the disease to advance to the point where it is uncurable and treatment requires more significant medical intervention. Rare

Introduction

disease takes an average of five years to diagnose with today's diagnostic techniques [5]. The incorrect or delayed diagnosis is one of the most serious safety concerns in healthcare today.

We can improve the current medical diagnostic models by using Artificial Intelligence (AI) technologies, to predict if a patient currently has a specific medical condition or may develop one. Diagnostic prediction models are used to detect if a disease or illness is already present in a patient. Prognostic prediction models estimate patient’s risk of developing a certain disease or illness.

The primary purpose of employing AI technologies in medical diagnostics is to eliminate the possibility of human error. As well, improving the speed and accuracy for detecting medical conditions of patients.

We will apply AI methods to develop a diagnostic model that can mimic doctors’ decision-making process with high experience, by using large datasets of patients data combined with the help of machine learning algorithms that finds relevant patterns in the data to determine accurately the presence of a specific medical condition.

In the following, we will introduce our **Aims and objectives** in section (1.2), and the **Methodology** used to achieve them in section (1.3). Then, a **Timeline** that shows when each project milestone is going to be achieved in section (1.4). At the end of this chapter, we will finally present the **Team Qualifications** in section (1.5), and the **Conclusion** in section (1.6).

Introduction

# **Aim and Objectives**

## **Aim**

Our aim in this project is to build a medical diagnostic model, a software capable of analyzing data and identifying patients’ medical conditions, based on the techniques of Artificial Intelligence (AI).

## **Objectives**

Our goal will be met by accomplishing the following **objectives**:

* Establish a medical domain background on the prediction models by reviewing the state-of-the-art research in the medical fields.
* Collecting machine learning workflow requirements including programming tools and datasets.
* Designing an experimental and comparative analysis.
* Building and evaluating machine learning based diagnostic models.

Introduction

# **Methodology**

To begin building our machine learning model, we must consider four important processes: data collection and preparation, feature engineering, algorithm selection, and performance measuring. See Figure 1.

In the data collection and preparation step, we will investigate and obtain a dataset that we will use to build and evaluate our model. We must prepare the dataset before we train our model with it.

In the second step, feature engineering, we need to find the important relevant feature space by which a diagnostic model can reach its highest possible discrimination level in a given condition. There are many techniques used in feature engineering to prepare the dataset for the algorithm to get better performance.

In the third step, algorithm selection, we must select the appropriate learning algorithm. This selection is based on the dataset's characteristics, such as the size of the training data or the number of features, and many other factors that we must consider when selecting potential algorithms for our model before evaluating and comparing their performance.

In the final step, performance measuring, we will evaluate our model’s performance using several performance metrics such as Accuracy and the Area Under the Curve (AUC). These four steps are repeated until we get a satisfying model performance.

We will use python libraries such as scikit-learn, pandas, NumPy, and Matplotlib to help us complete these steps.

Finally, to improve the development of our model, we should look at diagnostic models from previous studies, examine their solutions, and compare the findings.

Introduction

Diagram

Description automatically generated

Figure 1: Steps for Building Machine Learning Models

# **Project timeline**

Timeline

Description automatically generated

Figure 2: Project Timeline

Introduction

# **Team Qualifications**

Table 1 presents the qualifications of each member of the group.

Table 1: Team Qualifications

|  |  |
| --- | --- |
| **Student** | **Qualifications** |
| Ali Alhawas | * Expert in J**ava** programing language. * Expert in **Python** programming language. * Intermediate in **C** programming language. * Has an experience with **MySQL** database management system. |
| Abdulilah Alqasem | * Expert in **Java** programing language. * Expert in **Python** programming language. * Intermediate in **C** programming language. * Did a project on interactive AI, about Constraint Satisfaction problems. |
| Mohammad Zouman | * Expert in **Java** programing language. * Expert in **Python** programming language. * Intermediate in **C** programming language. * Did a project on interactive AI, about Constraint Satisfaction problems. |

# **Conclusion**

The high number of diagnostic errors in today's healthcare cries for a solution or improvement. By acknowledging that human error is the cause of diagnostic errors, we can develop a medical diagnostic model that support doctors in making their diagnostic decisions.

*Chapter Two*

Literature Review

Literature Review

# **Introduction**

In this chapter we will go through an overview of Machine Learning in section (2.2) which is a subfield of Artificial Intelligence, then we will see how machine learning canhelpus in the fields of medicine in section (2.3). After that we will present some of the related work done for building models that specifically diagnose COVID-19 and compare their results in section (2.4). Diagnosing COVID-19 is selected to be our current medical case where we aim to improve. Then a brief conclusion section (2.5) that summarizes this chapter.

# **Machine Learning**

Machine learning (ML) is a branch of computer science derived from both the study of pattern recognition in data and the computational learning theory in artificial intelligence. If Artificial Intelligence (AI) is the science of making machines intelligent, then machine learning is a technology that enables computers to execute certain tasks intelligently by learning from examples. As a result, rather than following pre-programmed rules, these systems can carry out complicated processes by learning from data [6].

Machine Learning is computationally demanding and often requires a substantial amount of training data. It entails repetitive training to increase the learning and decision- making abilities of algorithms. Figure 3 represents how ML models work in comparison to traditional software.

There are several types or approaches for developing Machine Learning models. A suitable method is chosen based on the dataset provided to the computer. Figure 4

Literature Review

summarizes the four major machine learning types, with an example of the type of dataset provided.

Diagram

Description automatically generated

Figure 3: Traditional Programming vs Machine Learning

Literature Review

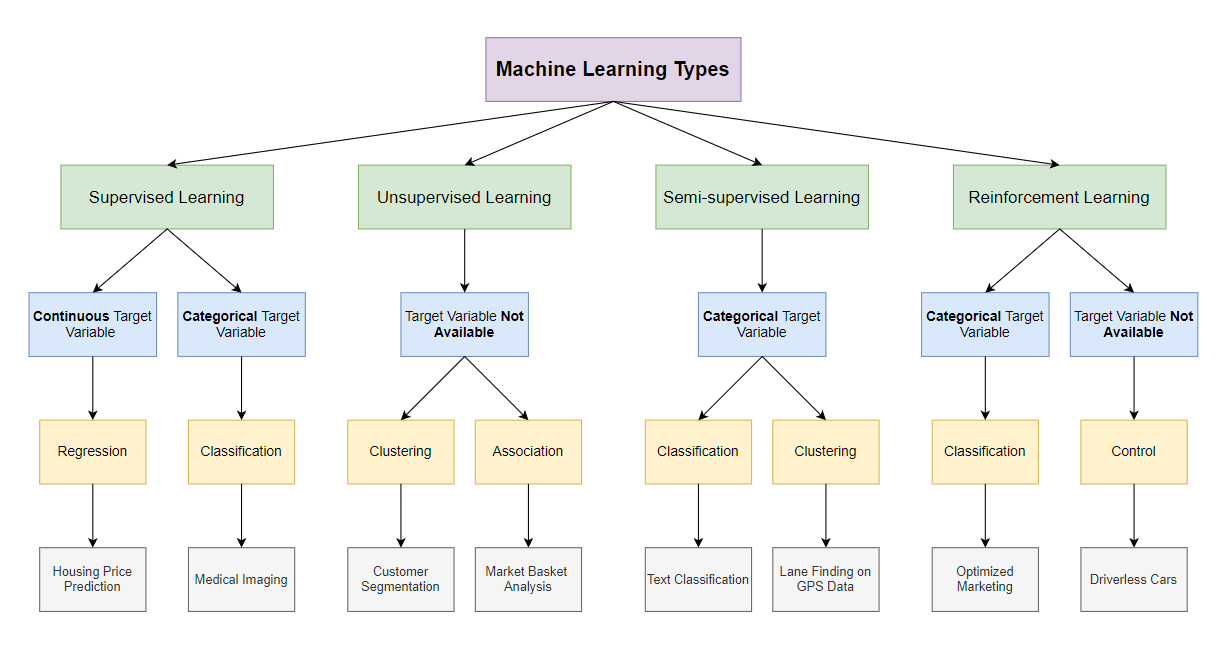


Figure 4: Machine Learning Types [7]

# **Machine Learning in Medicine**

Medical technologies driven by Artificial Intelligence (AI) are rapidly transforming into clinically applicable answers. Smartwatches, smartphones, and other mobile monitoring devices are feeding deep learning algorithms with increasing amounts of data, which may be used in a variety of medical purposes [8]. AI is now used in clinical settings, such as the detection of atrial fibrillation, epilepsy seizures, and hypoglycemia, and also the

Literature Review

diagnosis of illness based on histopathological examination or medical imaging, for instance [8]. Patients have been eager for augmented medicine to be implemented since it provides more personalized care [8]. Nevertheless, physicians have been hesitant because they were not prepared for such a revolution in clinical practice [8]. This phenomenon also necessitates the use of conventional clinical studies to validate these new techniques, along with a discussion of medical curriculum improvements in light of digital medicine, including ethical considerations of continuing linked monitoring [8].

As AI is increasingly becoming a vital aspect of modern healthcare, it should come as no surprise that there are a bunch of commonly used ML applications in the medical field at the current time. Let us take a closer look to some of the applications:

1. **Medical Imaging**

Convolutional Neural Networks (CNN) and some other Deep Learning models were used by gastroenterologists to analyze images from endoscopy and ultrasonography and recognize aberrant structures [9]. Studies show that a machine learning algorithm does the job as well as a trained specialist. While a human physician has an 86.4% sensitivity and 90.5% specificity, deep learning algorithms have 87.0% sensitivity and 92.5% specificity [10]. The Microsoft InnerEye is among the most successful cases of ML in medical imaging.

1. **Diabetes**

Using Machine Learning (ML) we can monitor the glucose level continuously of patients with type 1 or 2 diabetes, to see real-time interstitial glucose measurements and get information on the direction and pace of change in their blood glucose levels, using a continuous glucose monitoring device instead of a traditional blood glucose meter can benefit you in many ways [11].

Literature Review

* have a better control over your blood sugar levels on a daily basis
* have fewer low blood glucose emergencies
* fewer finger sticks are need

1. **Atrial Fibrillation**

One of the first applications of AI in medicine was the detection of atrial fibrillation. Kardia is a smartphone application that allows users to monitor their ECGs and detect atrial fibrillation. As shown in the REHEARSE-AF study, remote ECG monitoring with kardia is better than routine care when it comes to identifying atrial fibrillation in ambulatory patients [12]. However, there were several drawbacks to the software, and their limitations include false positives caused by movement artifacts and barriers to adoption for patients with atrial fibrillation, especially the elderly.

1. **Robotic Surgery**

With robotic surgery, small, precise movements are possible, making it more efficient than standard endoscopic techniques. As a result of that, the surgeon may perform an operation that formerly required open surgery through a little cut [13]. The robotic arm has the benefit of making it easier for the surgeon to use surgical tools via an endoscope once it has been placed in the abdomen [14]. Furthermore, the surgeon gets a better view of the area where the procedure will be conducted [13].

Literature Review

# **COVID-19 Diagnostic Models**

The COVID-19 pandemic has had a global impact, and as a result of that, there has been a lot of AI-related studies on COVID-19. As we progress through this section, we'll examine several scientific papers seeing their algorithms, methods, and results, as we advance to start building our model. We used National Library of Medicine (NLM) database and Google Scholar search engine to collect the related papers. Table 2 summarizes these papers.

Table 2: Related Work

| **Reference** | **Dataset** | **Classification** | **Algorithm** | **Results** |
| --- | --- | --- | --- | --- |
| [15] | Data form 21 hospitals, 413 patients with COVID-19 and 1050 with influenza. | Classify patients into COVID-19/ Influenza cases. | XGBoost | AUC :97.7%  Sensitivity: 92.5%  Specificity: 97.9% |
| RIDGE | AUC of 96.6%  Sensitivity: 87%  Specificity: 92% |
| RF | AUC: 95.3%  Sensitivity: 100%  Specificity: 90% |
| LASSO | AUC: 96.3%  Sensitivity: 85%  Specificity: 93% |
| [16] | An open-source GitHub repository contains data for 212 patients. | Classify the patients according to their viruses: COVID-19, ARDS, SARS, and ARDS with COVID-19. | LR | Precision: 94%  Recall: 96%  F1: 95%  Accuracy: 96.2% |
| MNB | Precision: 94%  Recall: 96%  F1: 95%  Accuracy: 96% |
| SVM | Precision: 82%  Recall: 91%  F1: 86%  Accuracy: 90.6% |
| DT | Precision: 92%  Recall: 92%  F1: 92%  Accuracy: 92.5% |
| [17] | This study includes X-Ray images of 85 patients in Wuhan, China. | Classify the patients with COVID-19 from the others. | K-NN | Precision: 96.5%  Recall: 96.5%  F1: 96.4%  ROC: 98.9% |
| [18] | This study evaluated 114 patients at the Taizhou hospital from January 17, 2020, to February 1, 2020. | Classify patients into positive/negative cases. | BN | Precision:67%  Recall: 71.9%  F1:65.3%  ROC:67.5% |
| LR | Precision:80.4%  Recall: 80.7%  F1:80.5%  ROC:78.2% |
| IBK | Precision:73.1%  Recall: 72.8%  F1:72.9%  ROC:64.9% |
| CR | Precision:83.7%  Recall: 84.2%  F1:83.7%  ROC:87.3% |
| PART | Precision:75.3%  Recall: 76.3%  F1:75.7%  ROC:71.9% |
| J48 | Precision:74.2%  Recall: 73.7%  F1:73.9%  ROC:72.2% |
| [19] | From 17 to 30 of March 2020, data were gathered from 235 patients at the Hospital Israelita Albert Einstein in São Paulo, Brazil. | Classify patients into positive/negative cases. | SVM | AUC: 84.7%  Sensitivity: 67.7%  Specificity: 85%  F1: 72.4% |
| RF | AUC: 84.7%  Sensitivity: 67.7%  Specificity: 85%  F1: 72.4% |
| NN | AUC: 84.4%  Sensitivity: 74.2%  Specificity: 80%  F1: 74.2% |
| LR | AUC: 84.3%  Sensitivity: 74.2%  Specificity: 82.5%  F1: 75.4% |
| GBT | AUC: 84.2%  Sensitivity: 80.6%  Specificity: 80%  F1: 78.1% |
| [20] | Positive chest X-ray images for COVID-19 from various papers were obtained from the SIRM COVID-19 Database, the NCV 2019 Dataset. 2905 images of chest X-ray pneumonia were integrated. | Classify patients into positive/negative cases. | KNN | Accuracy: 95.8%  Sensitivity: 92.3%  Specificity: 97.4%  F1: 94% |
| SVM | Accuracy: 98.9%  Sensitivity: 89.4%  Specificity: 99.7%  F1: 96.7% |
| DT | Accuracy: 96.1%  Sensitivity: 93.8%  Specificity: 97.7%  F1: 94.5% |
| CNN | Accuracy: 97.1%  Sensitivity: 94.6%  Specificity: 98.3%  F1: 95.7% |
| [21] | The data were collected from the New York Presbyterian Hospital/Weill Cornell Medicine (NYPH/WCM) from March 11 to April 29, 2020, with 5,893 patients. | Classify patients into positive/negative cases. | GBDT | AUC: 85.4%  Sensitivity: 76.1%  Specificity: 80.8% |
| RF | AUC: 84.3%  Sensitivity: 73.5%  Specificity: 81.8% |
| LR | AUC: 80.9%  Sensitivity: 71.1%  Specificity: 75.6% |
| DT | AUC: 70.4%  Sensitivity: 61.8%  Specificity: 73.2% |
| [22] | The collected data was 75,991 US veterans at the US Department of Veterans Affairs (VA) from 8 March to 22 July 2020. | Classify patients into positive/negative cases. | XGBoost | Accuracy: 86.4%  Sensitivity: 82.4%  Specificity: 86.8% |
| [23] | The data collected from patients that tested for SARS-CoV-2 at the Kepler University Hospital in Austria, from March 1, 2020, until April 30, 2020 | Classify patients into positive/negative cases. | RF | Accuracy: 81%  Sensitivity:60%  Specificity: 82%  AUC: 89.8% |
| [24] | The dataset used for this study consisted of routine blood test results of 1,925 patients at the San Raffaele Hospital from February 19, 2020, to May 31, 2020 | Classify patients into positive/negative cases. | LR | Accuracy: 89%  Sensitivity: 91.5%  Specificity: 87%  AUC: 90.5% |
| NB | Accuracy:86%  Sensitivity: 82.5%  Specificity: 89%  AUC: 88% |
| KNN | Accuracy:86%  Sensitivity: 79%  Specificity: 92.5%  AUC: 86.5% |
| RF | Accuracy:90%  Sensitivity: 89%  Specificity: 92.5%  AUC: 92% |
| SVM | Accuracy: 89%  Sensitivity: 90.5%  Specificity: 88.5%  AUC: 91% |

In [15] For their model, they used the eXtreme Gradient Boosting (XGBoost) algorithm. The data was gathered from 21 hospitals, 413 patients with COVID-19 and 1050 with influenza. To begin, they divided the data into 80% training and 20% testing,

Literature Review

with 5-fold cross-validation. The data were preprocessed by mixing COVID-19 and influenza cases, then deleting any clinical features that were missing from both datasets. They chose 27 clinical variables of high significance from a set of 48 variables. Variables include age, gender, the results of CT scans and chest X-rays, and so on. The receiver operating characteristic curve (ROC) and precision-recall (PR) curves were plotted to assess the results. The area under the curve (AUC) was used to calculate both curves. They performed classification using several Machine Learning (ML) algorithms, including Least Absolute Shrinkage and Selection Operator (LASSO), RIDGE Regression, and Random Forest (RF), and compared the results to those of XGBoost. The results of the XGBoost algorithm were a specificity of 97.9%, sensitivity of 92.5% and AUC of 97.7%, however, none of these algorithms achieved the XGBoost level of accuracy. In addition, they discovered that the most relevant features were temperature, fever, age, coughing, CT scan result and lymphocyte levels.

In [16], authors used algorithms such as Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB), Logistic Regression (LR), Decision Tree (DT) for their Classification task. The data was gathered from an open-source GitHub repository containing information from 212 patients with coronavirus and other virus symptoms. The data is preprocessed by eliminating any unnecessary attributes to make the machine learning algorithms achieve better accuracy. They employed the Term Frequency-Inverse Document Frequency (TF-IDF) technique to extract important features and identified 40 features that can be used to perform classification. The categorization was carried out to classify the patients based on their viruses: COVID-19, ARDS, SARS, and ARDS with COVID-19. The data is divided into 70% for training and 30% for testing. To avoid sampling bias, each algorithm was subjected to a ten-fold cross-validation technique. The results showed that the LR and MNB algorithms outperformed all other algorithms in this research, with precision of 94%, recall of 96%, F1 score of 95%, and accuracy of 96.2%. They suggested that more data might improve the efficiency of their models.

Literature Review

In [17], the authors used the K-Nearest Neighbors algorithm (KNN) for their modeling. The study includes X-Ray images of 85 patients from Wuhan, China. The Color Layout Descriptor (CLD) is used to extract numerical values from medical images. The dataset was split into 90% training and 10% testing. The results by the classification showed an average for precision and recall of 96.5%.

The authors in [18] applied six algorithms: Logistic Regression (LR), Classification via Regression (CR), decision-tree (J48), lazy-classifier (IBk), Bayes classifier (BN) and rule-learner (PART). From January 17, 2020, to February 1, 2020, data was collected from 114 patients at the Taizhou hospital. There were 170 variables in the data, and missing values were eliminated. There are 14 variables (features) selected, including the Lymphocyte (L%), White Blood Cell count (WBC), Neutrophil (N%), and so on. The classifiers were tested using ten-fold cross-validation, with 90% of the data used for training and 10% for testing. The CR classifier outperformed the other five classifiers in predicting COVID-19 cases.

To develop their models [19], the authors used five Machine Learning (ML) algorithms: Gradient Boosting Trees (GBT), Support Vector Machines (SVM), Neural Networks (NN), Logistic Regression (LR), Random Forests (RF). The data was obtained from 235 patients at the Hospital Israelita Albert Einstein in São Paulo, Brazil, between March 17 and March 30, 2020. The algorithms were trained using a total of 15 variables, including age, gender, hemoglobin, platelets, red blood cells, and so on. The data were randomly divided into 70% training and 30% testing. According to the findings, the three most essential variables for the algorithms' predicted effectiveness are the number of lymphocytes, leukocytes, and eosinophils, in that order.

Literature Review

In [20], the authors’ model is based on the Convolution Neural Network (CNN) architecture that extracts discriminative features on chest X-ray images. The dataset was split into two parts: 70% for training and 30% for testing. The data was gathered from the Italian Society of Medical and Interventional Radiology (SIRM). The Decision Tree (DT) classifier is primarily used to handle classification problems. They used the confusion matrix to evaluate their model and obtained an accuracy of 97.1%, sensitivity of 94.6%, and specificity of 98.3%.

In [21], the authors created a Machine Learning (ML) model that combined 27 standard laboratory tests with patient demographic information (age, gender, and race). Data were obtained from 5,893 patients at New York-Presbyterian Hospital/Weill Cornell Medicine (NYPH/WCM) between March 11 and April 29, 2020. The authors employed a 5-fold cross-validation. They tested four algorithms: the GBDT, RF, LR, and DT; the GBDT performed best, with an AUC of 85.4%.

Authors in [22] constructed a machine learning based model to examine the relationship between SARS-CoV-2 test results and the findings of 20 standard laboratory tests. The data was gathered from 75,991 patients who had at least one SARS-CoV-2 RT-PCR test between March 8 and July 22, 2020. They selected XGBoost algorithm due to its accuracy in the test set and its great tolerance for missing data without the requirement for imputation. They divided the dataset into three-quarters for training and cross-validation, while one-quarter was saved for testing. Eosinophil count, Serum ferritin, patient temperature, CRP, white blood cell count, were the most important variables. The model recreated the SARS-CoV2 test results with an accuracy of 86.4%, specificity of 86.8%, and sensitivity of 82.4%.

Literature Review

In [23], using a Machine Learning (ML) algorithm, the authors aimed to create a prediction model that could predict SARS-CoV-2 cases. The data was gathered during a test at the Kepler University in Austria. The data cleaning process included detecting type errors and outliers, as well as imputation of missing information. Variables which have more than 25% missing values are eliminated, and the remaining missing values were imputed using Strawman imputation, that takes median values (continuous variables), or the most often occurring value (categorical values) to fill in for missing data. Grid-search was used to execute the hyper-parameter search in the inner five-fold cross-validation loop. The RT-PCR test results were predicted using Random Forest (RF) classifier with an accuracy of 81% and AUC of 89.8%.

In [24], the authors developed five machine learning models. Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Support Vector Machine (SVM). The dataset used for this research comprised of routine blood-test results from 1,925 patients admitted to the ED at San Raffaele Hospital between 19 February and 31 May 2020. The training set was used to optimize hyperparameters using a 5-fold stratified cross-validation (via grid search). Hyper-parameter optimization was used to determine the best features to use. They tested five algorithms: LR, NB, KNN, RF, and SVM; concluding that SVM and RF were the best-performing classifiers.

Literature Review

# **Conclusion**

In this chapter, we presented an overview of Machine Learning (ML), outlining the fundamental principles. And then, we showcased the revolution of machine learning in the medical field. We displayed the applications of machine learning and the essential role it now plays in the healthcare system. Lastly, we narrowed the literature review to the application of machine learning in COVID-19 data space. It was clear that there were three types of datasets, the first focuses on CT scans, the second focuses on routine blood test results, and the third type focused on routine questions.

In the CT Scan Imaging datasets, the XGboost classifier was the best-performing algorithm to detect covid-19 virus because, of its accuracy as well as its high tolerance for missing data without the need for imputation. The main drawback of the CT Scan Imaging that it has a high cost, and it is time-consuming. On the other hand, the datasets with blood test results tested well with the Random Forest (RF) classifier. The Routine blood tests provided more categorical and continuous features for the RF classifier. And finally, the dataset with routine questions, the best-performing algorithm was the LR, which offers great accuracy for simple datasets. When the dataset is linearly separable, it usually performs well.

The next chapter considers the potential datasets as well as the model structure, the algorithms and evaluation metrics for our prediction model.

*Chapter Three*

Methodology

Methodology

# **Introduction**

The prior chapters provided background information for our project as well as a list of critical concepts in Artificial Intelligence (AI) including the vital role that machine learning is currently playing in the healthcare systems.

In this chapter, we will present the potential datasets in **Data Collection and Preparation** section (3.2) as well as the methods we might use to deal with missing data, and more. In the **Feature Engineering** section (3.3), we will show why we might need feature engineering techniques with our feature space. In the **Model Selection** section (3.4), we will list all the possible algorithms we are considering in developing our model. For measuring the performance of such models, we listed the performance metrics in the **Model Evaluation** section (3.5). Finally, the **Conclusion** section (3.6) summarizes this chapter.

# **Data Collection and Preparation**

After our overview of the datasets utilized in the Literature Review chapter, there were three types of datasets in the literature review section. The first one focuses mainly on CT-scan Imaging to detect covid-19, and the second one was primarily focusing on the results of a Laboratory blood test, the third one focused on Routine questions from the patients. So, in this section we will go through the potential datasets for our project:

Methodology

**1. CT scan Imaging**

The accuracy of your model will be based on the training images. So, the higher quality of images we can get**,** the higher quality model we can make. We should clean up the data as much as possible. If we are notsure what category an image belongs to, we will not use it. If there are duplicate images, we should remove them. An essential thing to remember is that the images should have the same number of pixels. It is advisable to segregate training images into individual folders based on the category they belong in when working with a large dataset [25]. The downside of using this dataset that it is costly, time-consuming, and some hospitals may not have the machines available [26].

**2. Laboratory blood test**

The dataset contains anonymized data from patients. Features such as serum ferritin, white blood cell count, and eosinophil count are included. The preparation of the dataset includes detection of typos and dropping any outliers; we may use dataset integration, if possible, to resolve some of the missing data problems in potential datasets. The dataset contains a large number of missing values which increase the importance of the imputation method. Table 3 shows an example of Strawman imputation for missing data.

Table 3: Strawman imputation fill continuous variables with the median value and categorical values with the most frequently. occurring value.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Weight | Height>152 | Obese | Age | Weight | Height>152 | Obese |
| 20 | 70 | Yes | No | 20 | 70 | Yes | No |
| 24 | 95 | Yes | Yes | 24 | 95 | Yes | Yes |
| 26 | 68 | No | NO | 26 | 68 | No | NO |
| 25 |  |  | NO | 25 | 70 | Yes | NO |

Methodology

**3. Routine questions**

The dataset contains a low number of features which makes it more difficult to accurately predict the covid-19 virus. The routine questions to the patients are limited, which makes it a small dataset that leads to overfitting. A Machine Learning (ML) model aims to generalize patterns in training data so that it can correctly predict new data that it has never seen before. When a model adjusts too much to the training data, it sees patterns that don't exist, and as a result, it performs badly in predicting new data.

# **Feature Engineering**

The features in the dataset substantially influence the performance of a machine learning model. With feature engineering we can improve the results of our model. There are many situations when feature engineering is needed and for every situation there is a technique you can use to improve your feature space hence a better model result. Table 4 presents the potential feature engineering techniques collected from the related work in section (2.4), that we might use in developing our model.

Methodology

Table 4: Feature Engineering Techniques

|  |  |  |
| --- | --- | --- |
| Technique | Situation | Methods |
| Imputation | Missing values. | Strawman Imputation,  Nearest neighbor. |
| Feature Selection | Irrelevant or redundant features. | CfsSubsetEval algorithm, Recursive Feature Elimination (RFE) algorithm. |
| Feature Extraction | Dimensionality reduction. (Large feature space) | There were no feature extraction methods used in the literature. In section (2.4), but we might consider Linear Discriminant Analysis (LDA)algorithm. |
| Feature Scaling | Standardize features range. | There were no feature scaling methods used in the literature, in section (2.4), but we might consider using one if needed. |

# **Model Selection**

There are several Machine Learning (ML) algorithms to consider when building a model, such as Support Vector Machine (SVM), Decision tree (DR), and Random Forest (RF). As we start developing our ML modeling pipeline, we will initially use the algorithms default settings (parameters), and after evaluating its performance we might need to adjust these settings to get a better performance. We considered seven ML algorithms in building our model. Below is a brief description for each algorithm.

Methodology

1. **Logistic Regression (LR)**

LR is a supervised ML algorithm used for classification problems. The LR algorithm limits the prediction probability of the action between 0 and 1 by using labeled data with sigmoid functions, also known as logistic functions. It predicts the type of numerical variable based on the relationship with the label [27]. Figure 5 illustrates LR algorithm.

Diagram

Description automatically generated with medium confidence

Figure 5: Logistic Regression Algorithm [28]

1. **Support Vector Machine (SVM)**

SVM is a supervised ML algorithm for classifying subjects into different classes. Many possible hyperplanes could be selected to separate classes of data points. The objective is to choose a hyperplane with maximum margin between the two data points [29]. Figure 6 illustrates SVM algorithm.

Methodology

Chart, scatter chart

Description automatically generated

Figure 6: Support Vector Machine Algorithm [30]

1. **K-Nearest Neighbors (KNN)**

KNN is a supervised ML algorithm for classification or regression. The KNN algorithm predicts the correct class for the test data [31]. The test data are indicated by how close they are to all the train data points by calculating their distances to each other, then it selects the 'K' number of points that is closest to the test data [31]. The KNN algorithm calculates the probability of the test data belonging to the 'K' training data classes and the class holding the highest probability will be selected. Figure 7 illustrates KNN algorithm.

Methodology

Diagram

Description automatically generated

Figure 7: K-Nearest Neighbors Algorithm [32]

1. **Naïve Bayes (NB)**

NB is a supervised ML algorithm for classification based on the Bayes theorem. It is based on the assumption that one feature in a class is independent of the other feature present in the same class [33]. The NB algorithm is a probabilistic classifier, which means it predicts based on the probability of an object. The NB classifier is simple to use, computationally fast and performs well on large datasets. The goal of NB classifier is to calculate conditional probability [33]:

For each of ‘k’ possible outcomes or classes . Let x = . Using Bayesian theorem, we can get [33]:

Methodology

1. **Decision Tree (DT)**

DT is a supervised ML algorithm for classification. The DT algorithm is designed to predict the value of a target variable by using the tree representation based on the tree representation of the problem [34]. The leaf node corresponds to a class label, and the attributes are represented on the internal node. Figure 8 illustrates DT algorithm.

Diagram

Description automatically generated

Figure 8: Decision Tree Algorithm [35]

1. **Random Forest (RF)**

RF is a supervised ML algorithm for classification and regression. The random forest classifier contains multiple decision trees applied to various subsets of a given dataset. It takes the average to improve the predictive accuracy on those subsets. The RF takes the prediction from each tree and predicts the final output based on the majority votes of predictions [36]. Figure 9 illustrates RF algorithm.

Methodology

Diagram

Description automatically generated

Figure 9: Random Forest Algorithm [37]

1. **eXtreme Gradient Boosting (XGBoost)**

XGBoost is a supervised ML algorithm for classification or regression. The XGBoost is a decision tree based on ensemble ML algorithms that uses a gradient boosting framework [38]. The XGBoost algorithm combines several optimization techniques to get perfect results [39]. By using regularization and cross-validation, overfitting can be avoided, and missing data can be handled perfectly [39]. Figure 10 illustrates XGBoost algorithm.

Methodology

Diagram

Description automatically generated

Figure 10: Simplified Structure of XGBoost Algorithm [40]

# **Model Evaluation**

Measuring a model’s performance is essential because we need to know how good the model is at predicting outcomes. From seeing the results of the model’s performance, we can decide if the model is at a satisfying performance or needs improvements to increase its predicting quality. The improvements can be made in the previous steps of building ML models, demonstrated in figure 1. Consequently, we need some metrics to measure our model’s performance. So, we collected the metrics we might consider in evaluating our model’s performance from the related work in section (2.4). Table 5 summarizes these evaluation metrics.

Methodology

Table 5: Performance Metrics

|  |  |
| --- | --- |
| **Performance Metric** | **Description** |
| **Accuracy** | The total proportion of predictions that have been correctly predicted by the model. |
| **Sensitivity or Recall** | The proportion of actual **positive** cases which are correctly predicted by the model. |
| **Specificity** | The proportion of actual **negative** cases which are correctly predicted by the model. |
| **Precision or Positive Predicted Value (PPV)** | The proportion of **positive** cases that were correctly predicted by the model. |
| **Negative Predicted Value (NPV)** | The proportion of **negative** cases that were correctly predicted by the model. |
| **F1 Score** | The F1 score represents the balance of precision and recall. |
| **Area Under the Curve (AUC)** | Reflects how well the model at distinguishing between classes. |

# **Conclusion**

The methodology chapter covers the building blocks for the development of our model. First, we explained the potential datasets we might use to develop our model and how we will prepare them. Then we discussed the feature engineering techniques we might employ in our model's data and, we also presented the algorithms we might use in our model. Lastly, we explained the metrics we will use to evaluate our model's performance.

The development of our model will be completed in Graduation Project 2 (GP2). The model's implementation methods, as well as the programming language libraries used, will be presented. Model testing and evaluating results will also be covered in detail, as will the model's conclusion and future work.

# **References**

|  |  |
| --- | --- |
| [1] | N. P. Taylor, "MedTech Dive: Medical Technology," 24 January 2019. [Online]. Available: https://www.medtechdive.com/news/duke-report-identifies-barriers-to-adoption-of-ai-healthcare-systems/546739/. |
| [2] | "SOCIETY to IMPROVE DIAGNOSIS in MEDICINE," 2021. [Online]. Available: https://www.improvediagnosis.org/what-is-diagnostic-error/. |
| [3] | "Salvi, Schostok & Pritchard P.C.," [Online]. Available: https://www.salvilaw.com/medical-malpractice/types/misdiagnosis/heart-attack/. |
| [4] | "Jonathan C.Reiter," [Online]. Available: https://www.jcreiterlaw.com/posts/cancer-study-doctors-miss-cancer-diagnosis-in-71-out-of-6000-patients/. |
| [5] | K. G. C. R. a. M. T. V. G. Rich Colbaugh, "PMC US National library of Medicine," 5 December 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371307/. |
| [6] | "The Royal Society," [Online]. Available: https://royalsociety.org/topics-policy/projects/machine-learning/. |
| [7] | "proft.me," 14 11 2017. [Online]. Available: https://en.proft.me/2015/12/24/types-machine-learning-algorithms/. |
| [8] | O. L. M. Giovanni Briganti, "Artificial Intelligence in Medicine: Today and Tomorrow," *US National Library of Medicine,* 2020 . |
| [9] | B. C. Yang YJ, "Application of artificial intelligence in gastroenterology. World J Gastroenterol," *The National Library of Medicine,* 2017. |
| [10] | p. A. K. D. prof Lucas M Bachmann, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysi," *THE LANCET Digital health,* 2019. |
| [11] | ,. J. E. Y. R. M. B. Othmar Moser, "Interstitial Glucose and Physical Exercise in Type 1 Diabetes: Integrative Physiology, Technology, and the Gap In-Between," *The National Library of Medicine,* 2018. |
| [12] | W. K. C. A. G. M. B. J. P. C. Halcox JPJ, "Assessment of remote heart rhythm sampling using the AliveCor heart monitor to screen for atrial fibrillation: the REHEARSE-AF study. Circulation," *The National Library of Medicine,* 2017. |
| [13] | "JOHNS HOPKINS MEDICINE," [Online]. Available: https://www.hopkinsmedicine.org/minimally\_invasive\_robotic\_surgery/types.html. |
| [14] | "MEDLINE PLUS," [Online]. Available: https://discord.com/channels/619224954730315776/882577225415786506/916283470647545857. |
| [15] | J. M. N. S. G. C. J. C. J. C. T. L. A. C. O. H. J. X. L. M. W. T. Z. A. L. A. G. T. K. H. S. Z. K. ,. M. A. Y. ,. Wei Tse Li, "Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis," *BMC,* 2020. |
| [16] | S. T. R. Q. R. K. N. R. M. M. U. D. Akib Mohi Ud Din Khanday, "Machine learning based approaches for detecting COVID-19 using clinical text data," *SpringeLink,* 2020. |
| [17] | F. M. F. M. A. S. Luca Brunese, "Machine learning for coronavirus covid-19 detection from chest x-rays," 2020. |
| [18] | S. H. M. A.-E. M. N. A.-K. M. P. Ibrahim Arpaci, "Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms," *SpringerLink,* 2021. |
| [19] | J. L. M. T. H. R. D. A. D. P. C. F. André Filipe de Moraes Batista, " COVID-19 diagnosis prediction in emergency care patients: a machine learning approach," *The Preprint Server for Health Sniences,* 2020. |
| [20] | Z. K. MajidNour, "A Novel Medical Diagnosis model for COVID-19 infection detection based on Deep Features and Bayesian Optimization, December," *ScienceDirect,* 2020. |
| [21] | Y. H. L. V. V. P. A. D. S. A. C. S. E. R.-B. P. V. M. M. C. M. L. R. K. Z. Z. F. W. He S Yang, "Routine Laboratory Blood Tests Predict SARS-CoV-2 Infection Using Machine Learning," *National Library of Medicine,* 2020. |
| [22] | S. P. R. R. C. L. H. P. J. M. F. S. P. E. M. H. Vafa Bayat, "A Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Prediction Model From Standard Laboratory Tests," *OxfordAcdemic,* 2020. |
| [23] | M. M. D. M. C. B. M. K. S. M. J. M. M. Thomas Tschoellitsch, "Machine Learning Prediction of SARS-CoV-2 Polymerase Chain Reaction Results with Routine Blood Tests," *Oxford Academic ,* 2020. |
| [24] | A. C. D. F. C. D. R. D. C. E. S. A. C. E. D. V. G. B. M. L. A. C. Federico Cabitza, "Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests," *National Library of Medicine ,* 2020. |
| [25] | " IBM," [Online]. Available: https://www.ibm.com/docs/vi/scdli/1.2.0?topic=dataset-images-object-classification. |
| [26] | H. L. Fred, "Drawbacks and Limitations of Computed Tomography," *National Library of Medicine,* 2004. |
| [27] | J. Brownlee, "Machine learining Mastery," 1 April 2016. [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/. |
| [28] | A. Pant, "Introduction to Logistic Regression," 22 jan 2019. [Online]. Available: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148. |
| [29] | S. Polamuri, "Dataaspirant," 15 December 2020. [Online]. Available: https://dataaspirant.com/3-support-vector-machine-algorithm/. |
| [30] | R. Gandhi, "towards data science," 7 jun 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47. |
| [31] | A. Christopher, "Start it up," [Online]. Available: https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4. |
| [32] | R. Shah, "Artificial Intelligence," [Online]. Available: https://ai.plainenglish.io/introduction-to-k-nearest-neighbors-knn-algorithm-e8617a448fa8. |
| [33] | "towards data science," 9 sep 2019. [Online]. Available: https://towardsdatascience.com/introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf. |
| [34] | "Scikit Learn," [Online]. Available: https://scikit-learn.org/stable/modules/tree.html. |
| [35] | F. N. Arain, "Devops," [Online]. Available: https://www.devops.ae/decision-tree-classification-algorithm/. |
| [36] | "javaTpoint," [Online]. Available: https://www.javatpoint.com/machine-learning-random-forest-algorithm. |
| [37] | H. Ampadu, "AI Pool," 1 May 2021. [Online]. Available: https://ai-pool.com/a/s/random-forests-understanding. |
| [38] | V. Morde, "towards Data Science," 8 Apr 2019. [Online]. Available: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.. |
| [39] | "EDUCBA," [Online]. Available: https://www.educba.com/random-forest-vs-xgboost/. |
| [40] | G. C. B. C. Weilun Wang, "ResearchGate," December 2020. [Online]. Available: https://www.researchgate.net/figure/Simplified-structure-of-XGBoost\_fig2\_348025909. |