Al Imam Mohammad ibn Saud Islamic University
College of Computer and Information Sciences
Computer Science Department
First Semester 1443 H – 2022 G

Development of Medical Diagnostic Models Using Artificial Intelligence

By

Ali Khalifa Alhawas (439015852)
Abdulilah Saleh Alqasem (439014750)
Mohammad Abdulaziz Zouman (439012798)

Supervisor

Dr. Sultan Sharidah Aldera

14 May 2022

Abstract

Background: The healthcare sector always carries the risk of medical errors. When illnesses are diagnosed incorrectly or even delayed, many lives are to be in 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). In this project we are focusing on COVID-19 because, of the global influence it had.

Methods: From three different feature spaces (Symptoms, Laboratory, and Imaging), there are six datasets considered in our research project: 1) The Israeli Ministry of Health Dataset. 2) COVID-19 and Influenza Patients Dataset. 3) Albert Einstein Hospital, Brazil Dataset. 4) Italian Society of Medical Radiology Dataset. 5) COVID-19 Chest Xray Dataset. 6) COVID-19 Radiography Database. Using such datasets, we built seven diagnostic Machine Learning (ML) based models to diagnose COVID-19 infected individuals. The used ML algorithms are: 1) Logistic Regression (LR). 2) Support Vector Machine (SVM). 3) K-Nearest Neighbors (KNN). 4) Naïve Bayes (NB). 5) Decision Tree (DT). 6) Random Forest (RF). 7) eXtreme Gradient Boosting (XGB). We evaluated the diagnostic models by using three metrics, which are: precision, recall, and f1-score.

Results: With f1-scores ranging from 46% to 92%, precision ranging from 55% to 92%, and recall ranging from 53% to 92%, our models performed differently across datasets and feature spaces. Our top model built using the COVID-19 and Influenza Patients Dataset, and the RF classifier, with precision, recall and f1-score of 92%.

Conclusion: Our work, beside related works, have showed that 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.

Table of Contents

List of Tables

List of Figures

List of Abbreviation

AP	Anteroposterior
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
IRD	Influenza Research Database
J48	Decision Tree

KUH	Kepler University Hospital
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
PA	Posteroanterior
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

1.1 The Scope

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. 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 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. Also, improving the speed and accuracy for detecting medical conditions of patients.

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).

1.2 Aim and Objectives

1.2.1 Aim

Our aim in this project is to build medical diagnostic models. Software capable of analyzing data and identifying patients' medical conditions, based on the techniques of AI.

1.2.2 Objectives

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

- **Conducting a literature review:** we will establish a medical domain background on the prediction models by reviewing the state-of-the-art research in the medical fields.
- **Building a machine learning based diagnostic modelling pipeline:** we will collect machine learning workflow requirements including programming tools and datasets.
- **Designing and performing an experimental and comparative analysis:** we will build diagnostic models by using multiple datasets representing multiple medical feature spaces before evaluating such models and comparing them to other related work.

1.3 Methodology

To build our ML models, we must consider four important processes: Data Collection and Preparation, Feature Engineering, Modeling, and Evaluation (Figure 1).

In the data collection and preparation step, we will investigate and obtain multiple datasets 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.

In the third step, modeling, we will create 7 models using 7 classifiers. Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and eXtream Gradient Boosting (XGB) are the classifiers we will train, test, assess, and adjust on each dataset.

In the fourth and final step, evaluation, we will measure the performance of our models using several performance metrics such as Recall and Precision. These four steps are repeated until we get a satisfying model performance.

To perform these steps, we will utilize Python libraries such as scikit-learn, pandas, NumPy, and Matplotlib.

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

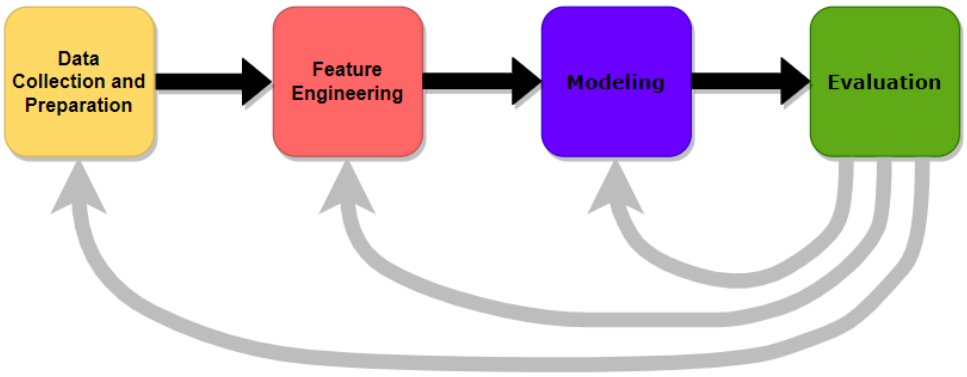

Figure 1: Steps for Building ML Models.

1.4 Project timeline

Figure (2) depicts our project stages in chronological sequence, allowing us as project managers to see the complete project in one spot.

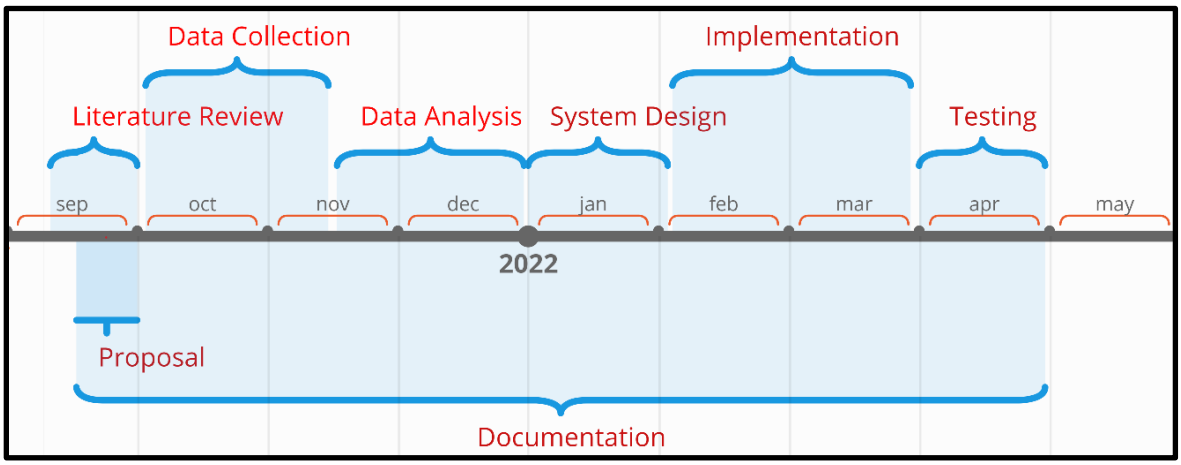

Figure 2: Project Timeline

1.5 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.

1.6 Report structure

We created a medical domain background on the ML prediction models in the Background chapter (Chapter two) by reviewing existing research in the healthcare industry.

We emphasized our chosen datasets and the techniques we used to prepare them, as well as our learning algorithms and evaluation metrics, in the Methodology chapter (Chapter three).

We tested the performance of our diagnostic models and compared it to other relevant published work in the Results chapter (Chapter four).

In the Discussion chapter (Chapter five), we presented our experience looking for and collecting datasets, as well as evaluating their quality. We presented our findings, along with the best model for each feature space and the most predictive features, and we compared our findings to previous research. We talked about the difficulties we had while working on our project.

Chapter Two

Background

2.1 Introduction

The widespread availability of advanced hardware and cloud computing has led to a wider deployment of machine learning in several domains of human life, ranging from social media recommendations to Self-driving cars and yet its demand is just going to increase. ML role in healthcare have a big future because of the volume of data gathered for each patient. So, it is no surprise that there is a wide array of effective ML applications in the healthcare field right now.

In this chapter, in section (2.1), the major topics relevant to our research questions are introduced. Then, we will explain our search strategy in section (2.2), report the related works in section (2.3), examine such works in many levels in section (2.4), and finally conclude in section (2.5).

2.1.1 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 AI is the science of making machines intelligent, then ML is a subfield that enables such machines to learn from the experience without human intervention in terms of being partially programmed machines. As a result, rather than following pre-programmed rules, these systems can carry out complicated processes by learning from data [6].

ML 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 ML models (Figure 4). A suitable method is chosen based on the dataset provided to the computer and the target problem.

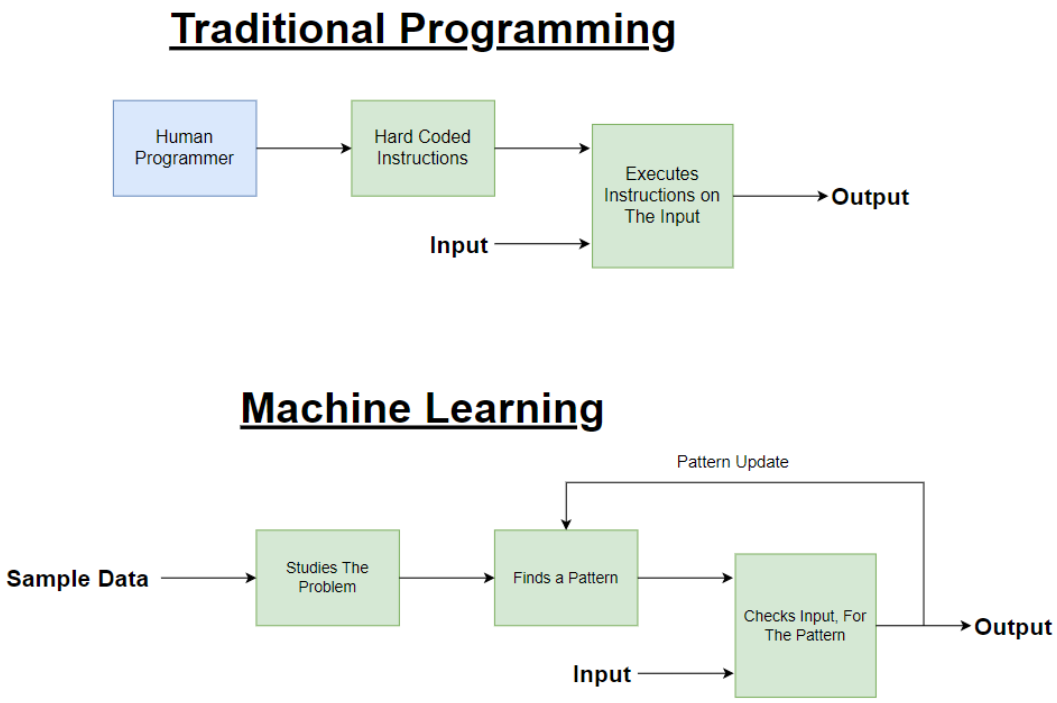

Figure 3: Traditional Programming vs Machine Learning

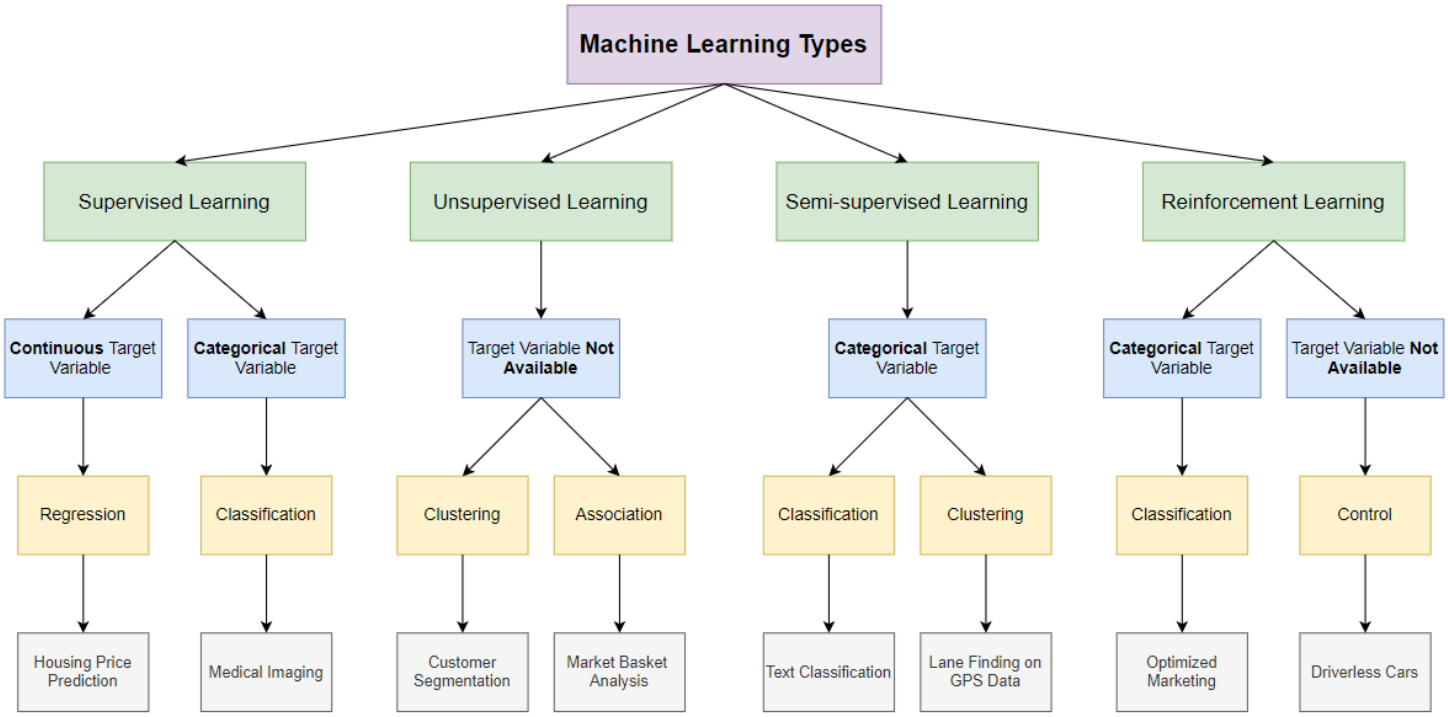

Figure 4: Machine Learning Types [7]

2.1.2 Machine learning in medicine

Medical technologies driven by AI are rapidly transforming into clinically applicable answers. Smartwatches, smartphones, and other mobile monitoring devices are feeding machine 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. AI is also helping in the 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 involves the utilization of conventional clinical studies to validate these new procedures, as well as a debate of medical curriculum enhancements in light of digital medicine. [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]. 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.

2. Diabetes

We can use ML to continuously monitor the glucose levels of individuals with type 1 or type 2 diabetes. Using a continuous glucose monitoring device instead of a typical blood glucose meter that can help you observe real-time interstitial glucose readings and gain information on the direction and rate of change in your blood glucose levels [11].

3. 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, Kardia may produce false positives outcomes caused by movement artifacts and barriers to adoption for patients with atrial fibrillation, especially the elderly.

4. 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].

2.1.3 COVID-19 & Machine Learning

The COVID-19 pandemic has had an influence on almost every area of human civilization in every geographical region. As a result, huge worldwide efforts have been undertaken to diagnose the virus as soon as possible and to treat it as soon as possible in order to prevent the infection from spreading. The current gold standard test in COVID-19 diagnosis is reverse transcription polymerase chain reaction (RT-PCR) [49], however it is time consuming, expensive, requires specialist equipment, and has a false-negative rate of around 20% [50].

There has been a great deal of interest in investigating the potential of ML tools to combat the COVID-19 pandemic by contributing to disease diagnosis and prognosis, forecasting, prevention, treatment and management, disease surveillance, and antiviral drug

discovery. Even an experienced physician finds it difficult to extract all the information contained in routine blood tests. ML algorithms, on the other hand, can learn and distinguish between diverse patterns found in patients' diagnostic data. Therefore, we are concentrating on constructing ML models for COVID-19 detection.

2.2 Search Strategy

The literature search and inclusion criteria for studies were manually curated from several sources such as PubMed and Goggle scholar using the keywords in [10], which are: "COVID-19", "coronavirus", "SARS-CoV-2", "diagnosis", "diagnostic", "Algorithms", "Model" and "Machine Learning". We examined 80 scientific papers discussing the COVID-19 diagnostic prediction models and selected a total of 50 publications based on the number of citations, all publications were published between January 4, 2020, and January 7, 2021. All publications that were written in a language other than English and were unrelated to the study subject were excluded from our analysis. We identified 13 studies after conducting a manual review, figure (5) demonstrates the search technique.

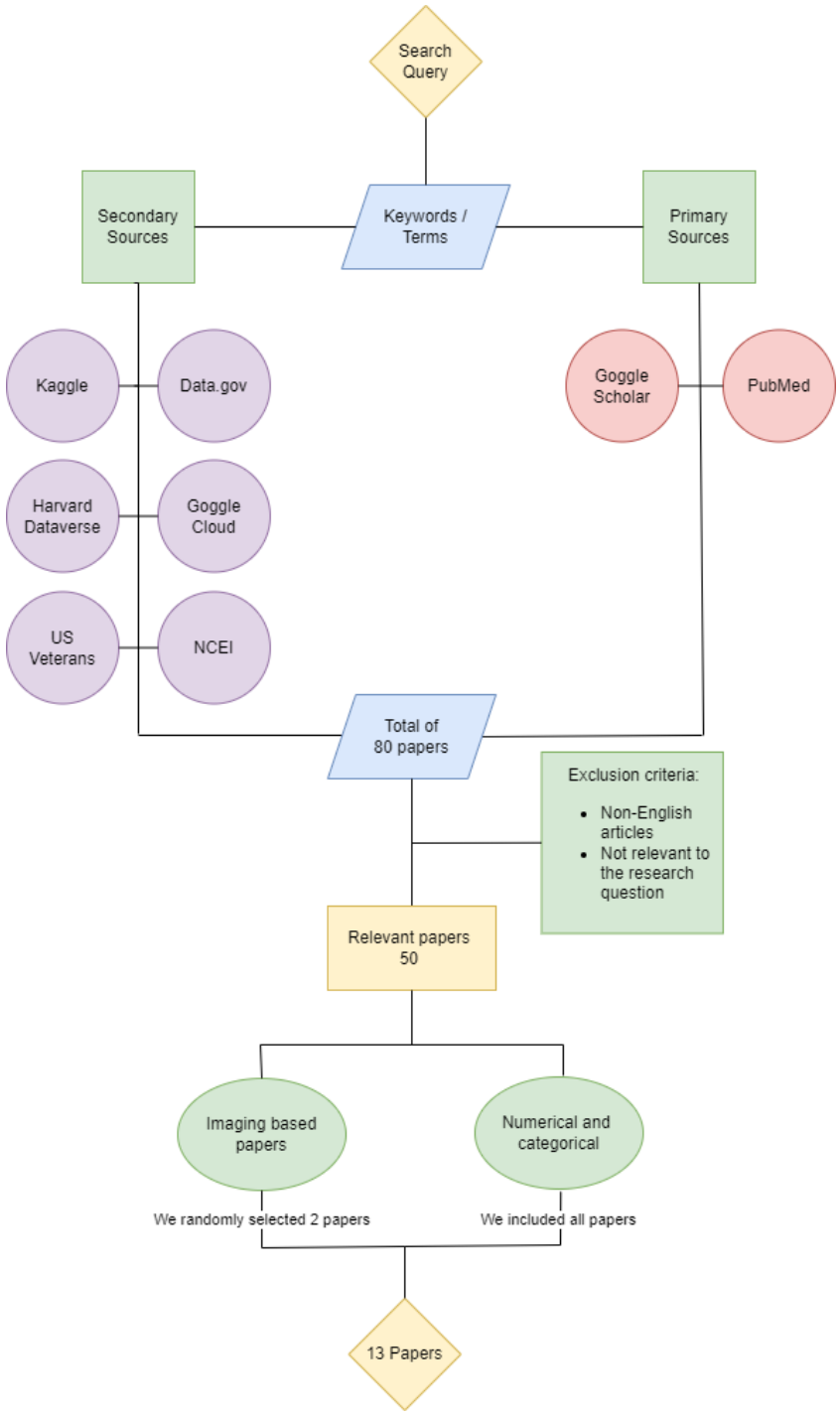

Figure 5: Search Strategy.

2.3 Related Works

In order to gain a complete understanding of the process of developing our model, we had to review the relevant literature. We examined their algorithms, techniques, and findings. In the following, Table 2 summarizes the findings of the scientific papers that were chosen.

Table 2: Performance comparison of existing models.

Reference	Feature Space	Model	Results		
			Precision	Recall	F1-Score
[15]	Laboratory	XGB	-	92.5%	92.8%
		RIDGE	-	87%	-
		RF	-	100%	-
		LASSO	-	85%	-
[16]	Symptoms	MNB	94%	96%	95%
		LR	93%	96%	95%
		SVM	82%	92%	86%
		DT	93%	93%	93%
		Bagging	93%	93%	93%
		Adaboost	85%	92%	88%
		RF	93%	94%	92%

		SGB	93%	94%	92%
[17]	Imaging	KNN	96.5%	96.5%	96.4%
[18]	Laboratory	BN	67%	71.9%	65.3%
		LR	80.4%	80.7%	80.5%
		IBK	73.1%	72.8%	72.9%
		CR	73.1%	72.8%	72.9%
		PART	83.7%	72.8%	72.9%
		J48	74.2%	73.7%	73.9%
[19]	Laboratory	SVM	-	67.7%	72.4%
		RF	-	67.7%	72.4%
		NN	-	74.2%	74.1%
		LR	-	74.2%	75.4%
		GBT	-	80.6%	78.1%
[20]	Imaging	KNN	-	92.3%	94%
		SVM	-	89.4%	96.7%
		DT	-	93.8%	94.5%
		CNN	-	94.6%	95.7%
[21]	Laboratory	GBDT	-	76.1%	-
		RF	-	73.5%	-

		LR	-	71.1%	-
		DT	-	61.8%	-
[22]	Laboratory	XGB	-	82.4%	-
[23]	Laboratory	RF	-	60%	-
[24]	Laboratory	SVM	-	90.5%	-
		RF	-	89%	-
		KNN	-	79%	-
		NB	-	82.5%	-
		LR	-	91.5%	-
[42]	Symptoms	GBC	-	85.7%	-
[44]	Laboratory	ERLX	-	96.8%	-
		ANN With SMOTE	-	43%	-
		BN	-	96.8%	-
		SVM	-	68%	-
[46]	Laboratory	XGB	-	75%	-
		RF	-	69%	-
		NN	-	60%	-
		LR	-	58%	-
		SVM	-	57%	-

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. They divided the data into 80% training and 20% testing, 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. After performing an ANOVA test on a set of 48 features such as age, gender, and CT scan results, they selected 27 clinical variables of high significance. 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%, recall of 92.5%, AUC of 97.7% and f1-score 92.8%, 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 10-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.

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%, f1-score of 96.4%.

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, such as the Lymphocyte (L%), White Blood Cell count (WBC), Neutrophil (N%). The classifiers were tested using 10-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. The results by the classification showed an average for recall of 80.7%, precision of 80.4% and f1-score of 80.5%.

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, such as age, gender, hemoglobin, platelets, and red blood cells. 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. They used the confusion matrix to evaluate their model and obtained a recall of 80.6% and f1-score of 78.1%

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 a recall of 89.4%, and f1-score of 96.7%.

In [21], the authors created a Machine Learning (ML) model that combined 27 standard laboratory tests with patient demographic information. 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% and a recall of 76.1%.

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 and white blood cell count were the most important variables. The model reached an accuracy of 86.4%, specificity of 86.8%, and a recall of 82.4%.

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 with more than 25% missing values were eliminated. 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 5-fold cross-validation loop. The RT-PCR test results were predicted using Random Forest (RF) classifier with an accuracy of 81%, AUC of 89.8% and recall 60%.

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. The model recreated the SARS-CoV2 test results with a recall of 90.5 %.

In [42], the authors used gradient-boosting model developed using decision-tree as a base-learners. The data was gathered from the Israeli Ministry of Health the Dataset contained 278K people who tested for SARS-CoV-2. During the early months of Israel's COVID-19 outbreak. The gradient-boosting predictor dealt with missing values in the dataset. LightGBM is used to train the gradient-boosting predictor. The training-test sets included data from 51,831 participants who had been tested (of whom 4769 were confirmed to have COVID-19). The most important features are contact with confirmed (a case that comes in touch with a positive case), headache, fever, cough respectively. The model recreated the SARS-CoV2 test results with a specificity of 79.1% and a recall of 85.7%.

In [44], The authors employed an ensemble machine learning methodology to create a single model by combining different classification classifiers. The study's data was collected from 5644 individuals who were hospitalized in the Albert Einstein Israelita Hospital in Saulo Paulo, Brazil. More than 100 laboratory tests including blood tests, urine tests, SARS-CoV-2 test, rt-PCR test, and others were gathered throughout Mar 28, 2020, to Apr 3, 2020.

KNNImputer algorithm was used to handle null values. The Albert Einstein dataset contains 108 characteristics; in the proposed model, 18 features were chosen based on their importance in recognizing COVID-19 in clinical studies. The total dataset was split it to 80 percent for the training set and 20% for the test set. SMOTE from the imblearn Python package is used in the suggested model. To oversample the minority class, SMOTE balances the distribution of the dataset by producing minority class entities at random. The model recreated the SARS-CoV2 test results accuracy of 80.3%, recall of 84.4%, specificity of 94.8%.

In [46], the authors used Five Machine Learning (ML) algorithms: Logistic Regression (LR), Neural Network (NN), Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (XGBoost). They used data from 5644 patients seen at the Hospital Israelita Albert Einstein in São Paulo, Brazil in the early months of 2020. They set up a systematic model development, validation, and evaluation pipeline. The multistage pipeline consists of preprocessing, model development, model selection, and model evaluation stages. They discarded any features with more than 99.8% missing data during the preprocessing stage. They derived the models from the preprocessed training fold data, optimized the various types of predictive models, and did a hyperparameter search with m runs for each of them during the model development stage. They tested their predictive performance against the held-out validation fold that was not used for model development during the model selection stage. They ranked all models according to their Evaluated predicted performance and choose the best candidate model. They choose the best model from the model selection step and evaluate it exactly once against the test fold in the model evaluation stage. The XGBoost had the best sensitivity and specificity for predicting SARS-CoV-2 test results, with 75% recall and 49% specificity.

To summarize the related work, we will describe the findings and the best algorithms. It was evident that there were three categories of datasets: imaging, laboratory, and symptoms.

In the imaging datasets, the KNN classifier was the best-performing algorithm to detect the COVID-19 virus. The main drawbacks of imaging are that it has a high cost and that it is time-consuming. The laboratory datasets' results tested well with the XGBoost classifier. The laboratory tests provided more categorical and continuous features for the XGBoost classifier. And finally, for symptom datasets, the best-performing algorithm was the MNB. The MNB algorithms handle both continuous and discrete data. It is also highly scalable in terms of the number of predictors and data points. It is fast and can be used to make real-time predictions.

2.4 Conclusion

It is clear that machine learning applications have a vital role in medicine. The reviewed related work showed the great potential of machine learning-based diagnostic models in enhancing diagnostic performance. Applying more algorithms, performing more methods at all levels from data preparation to modeling evaluation, and comparing different models using different criteria are all necessary actions to improve the current view of machine learning in medicine and, more specifically, to assist the global community in mitigating the impact of COVID-19.

Chapter Three

Methodology

3.1 Data Collection

After our overview of the datasets utilized in the Background chapter, there were three types of datasets in the related work section (2.3). The first one focuses mainly on CT-scan and X-rays Imaging to detect COVID-19. The second one was primarily focusing on the results of a laboratory blood test. The third one focused on routine questions from the patients. In this section we will go through the potential datasets for our project (Table 3).

Table 3: Papers and datasets

Reference	Dataset	Sample Population	Accessibility	Obtained
[15]	COVID19 and influenza patients Dataset	Not specified	Public	Yes, dataset found at a GitHub project.
[16]	open-source GitHub repository	Saudi Arabia, Australia, Egypt and 39 other countries.	Public	Yes, dataset found at a GitHub project.
[17]	Covid Chestxray Dataset	Italy, Korea, Sweden, Australia, China, Taiwan, USA	Public	Yes, dataset found at a GitHub project.

[18]	Taizhou hospital of Zhejiang Province in China	China	Private	No, not included in the paper.
[19]	Hospital Israelita Albert Einstein in Brazil	Brazil	Public	Yes, Found at Kaggle.com
[20]	Italian Society of Medical Radiology and Interventional (SIRM)	Italy, Korea, Sweden, Australia, China, Taiwan, USA	Public	One part of the dataset found at GitHub project, and the other in Kaggle.com
[21]	New York Presbyterian Hospital/Weill Cornell Medicine (NYPH/WCM)	United States	Private	No, not included in the paper.
[22]	The US Department of Veterans Affairs (VA) healthcare system	United States	Private	No, not included in the paper.

[23]	Kepler University Hospital (KUH)	Austria	Private	No, not included in the paper.
[24]	San Raphael Hospital (OSR)	Italy	Private	No, not included in the paper.
[42]	The Israeli Ministry of Health Dataset	Israel	Public	Yes, Found at a GitHub Repository
[43]	COVID-19 Radiography Database	Qatar, Bangladesh, Pakistan, Malaysia	Public	Yes, Found at Kaggle.com
[44]	Hospital Israelita Albert Einstein in Brazil	Brazil	Public	Yes, Found at Kaggle.com
[45]	Italian Society of Medical Radiology and Interventional (SIRM) & Influenza Research Database (IRD)	Italy	Public	Yes, Found at dataverse.harvard.edu

[46]	Hospital Israelita Albert Einstein in Brazil	Brazil	Public	Yes, Found at Kaggle.com

3.1.1 Selected Datasets

Throughout our journey, we encountered several challenges in terms of acquiring a dataset. Most of the datasets mentioned in the scientific papers are private. We did contact each paper regarding acquiring the dataset, but there was no response. On the other hand, the publicly available datasets had their own problems. Several of them had problems with missing values due to the recency of the disease that we are identifying in our project. The majority of the papers were published during the outbreak of COVID-19, and the data that was collected in the papers was imbalanced or highly biased. Nevertheless, after several steps of preprocessing, we decided that the following datasets were the best in terms of quality:

1. The Israeli Ministry of Health Dataset

This dataset was translated from Hebrew to English by a group of students from the University of Tel Aviv. There were many reasons why we chose this specific dataset out of the others. One of the reasons was that the dataset had a large number of patients. It also had one of the most important features in any symptom dataset, which is "contact with confirmed." It also has a small number of missing values compared to others, and the features weren't as biased as the rest of the datasets.

2. COVID-19 and Influenza Patients Dataset

The dataset contains several features that were important. One of them is the risk factors, which means if the patient has any other medical issue, for example, heart disease, which can be deadly for people that have that risk factor. Some other features that were also important were the sore throat feature, these features contributed more than most in a diagnostic model.

3. Albert Einstein Hospital, Brazil Dataset

The dataset contains anonymized data from patients visited at the Hospital Israelita Albert Einstein in So Paulo, Brazil, who had samples collected to perform the SARS-CoV-2 RT-PCR and other laboratory tests during their visit. As there are a considerable number of samples and features, which we can use and manipulate, this dataset lends itself well to the development of prediction models.

4. Italian Society of Medical Radiology Dataset (SIRM)

This dataset is made up of COVID-19 samples from the Italian society of medical and intervention radiology society (SIRM) database and flu samples from the influenza research database (IRD). The dataset has a balanced distribution of COVID-19 and flu samples, which is quite good because most datasets have an imbalanced sample class distribution.

5. COVID-19 Chest Xray Dataset

This data was gathered from both public and private sources, including hospitals and clinicians. There are several reasons for selecting this dataset, but the most significant three are that it is well-organized, comes from various countries, and there are a few missing images.

6. COVID-19 Radiography Database

In conjunction with medical doctors, a group of researchers from Qatar, Bangladesh, Pakistan, and Malaysia established a database of chest X-ray images for COVID-19 positive cases, as well as normal and viral pneumonia images. The enormous number of images for COVID-19 and normal cases is why this dataset was chosen.

3.2 Data wrangling

The data wrangling process is comprised of several steps that vary depending on the dataset. It is important to first prepare the dataset before data modeling in order for the data provided in the dataset to be used as input parameters for a machine learning algorithm. Below, we will discuss the distinctions in the feature space of each type of dataset in the section **Feature Space**, and then, in the section **Data Wrangling Procedures**, we will explain our steps in preparing the data for modeling.

3.2.1 Feature Space

Because each type of dataset is likely to have its own feature space, the preprocessing steps might vary. The feature space of each dataset type is summarized below.

1. Symptoms Based Datasets

In symptom-based datasets, the feature space is dominated by binary features, such as whether you have a cough or not, if you contacted someone who was confirmed with

COVID-19 or not, and so on. It can have some continuous features, including the patient's temperature, or graphical features, such as the region.

2. Laboratory Based Datasets

Blood samples, urine samples, and bodily tissues make up the majority of the feature space of laboratory-based datasets. It has both continuous and discontinuous features. The number of features you can get from these types of datasets may be enormous, especially when you consider the number of the available lab tests.

3. Imaging Based Datasets

Since a dot represents each pixel in an image, the feature space of an image is a graph of the pixel values of a set of data files. The X-ray images are the feature space for imaging datasets. Covid-19 or Normal are the labels on the X-ray images.

3.2.2 Data Wrangling Steps

Data wrangling is the procedure for preparing raw data for use in a machine learning model. It's the first and most important stage in building a machine learning model. The number of steps will vary depending on the particular dataset. For our six datasets, they appear to have a few steps in common. Each dataset's preprocessing steps are listed in Table 4.

Table 4: Data Wrangling steps for each dataset.

Dataset	Data Wrangling Steps
The Israeli Ministry of Health Dataset	• Handling Missing Values. • Handling Inappropriate Types of Features. • Handling Categorical Features. • Handling Imbalanced Target Variable. • Feature Selection
COVID-19 and Influenza Patients Dataset	• Handling Missing Values. • Handling Bias Features. • Handling Categorical Features. • Imbalanced Target Variable. • Feature Selection
Albert Einstein Hospital, Brazil Dataset	• Handling Missing Values. • Handling Zero Variance Features. • Handling Categorical Features. • Feature Scaling. • Imbalanced Target Variable. • Feature Selection
Italian Society of Medical Radiology Dataset	• Handling Missing Values. • Handling Categorical Features. • Handling Zero Variance Features. • Feature Scaling. • Feature Selection
COVID-19 Chest Xray Dataset	• Handling Missing Images • Imbalanced Target Variable • Image Preprocessing • Feature Engineering • Feature Scaling

COVID-19 Radiography Database	• Imbalanced Target Variable • Image Preprocessing • Feature Engineering • Feature Scaling • Feature Selection

1. The Israeli Ministry of Health Dataset

The dataset consists of 278848 patients in rows and 15 different categorical features. It is one of the largest datasets that we discovered throughout our search.

I. Handling Missing Values

The dataset has few missing values. The only column with a large number of missing values is 'age_60_and_above', which has 45.7%. We chose to remove the column due to the high number of missing data. The rest of the features with missing values have less than 1%.

II. Handling Inappropriate Types of Features

Four

of the ten features in the dataset were classified as objects when they should have been classified as integer, since having a cough should be one (integer) and not having one should be zero (integer), this misclassification resulted in multiple zeros and multiple ones in the exploratory data analysis phase.

III. Handling Categorical Features

As we know to use some algorithms only accept numerical columns therefore, we performed one-hot encoding to convert categorical features to numerical features. Using get dummies from the pandas library, we chose one hot encoding specifically because we don't want the model to assume that there is any ordering in features such as gender or test indication. For example, if we use label_encoder, it will give a specific class the number 0 and another class the number 1, and the third class will get the number 2, as a result, the model will make certain assumptions on this.

IV. Handing Imbalanced Target Variable

After executing all of the preparation processes, the dataset shape has 23394 rows and 8 columns. The target variable ' test result ' has 16248 negative tests and 7146 positive tests, which is clearly imbalanced. Training a model on an imbalanced dataset result in a highly biased model that may classify some classes poorly and some others with a high success rate.

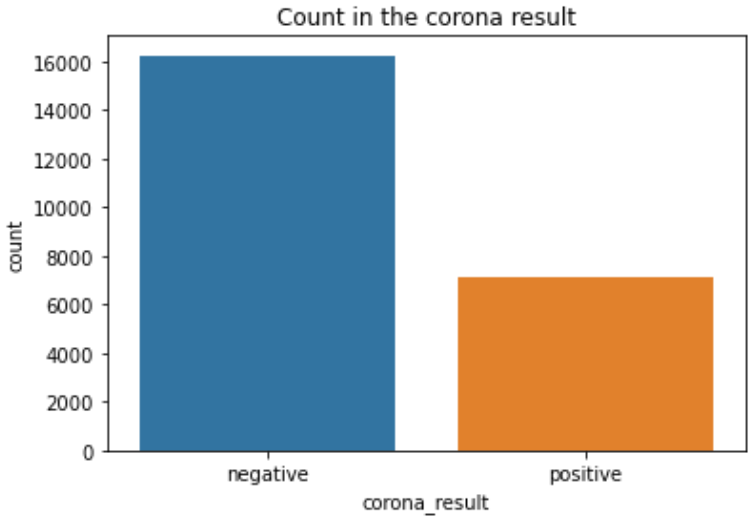

Figure 6: Imbalanced Target variable of the Israeli ministry of health dataset.

We decided to use several methods to solve a problem. The first one is under-sampling the majority class, the second one is over-sampling of the minority class, and finally an ensemble method to handle that issue.

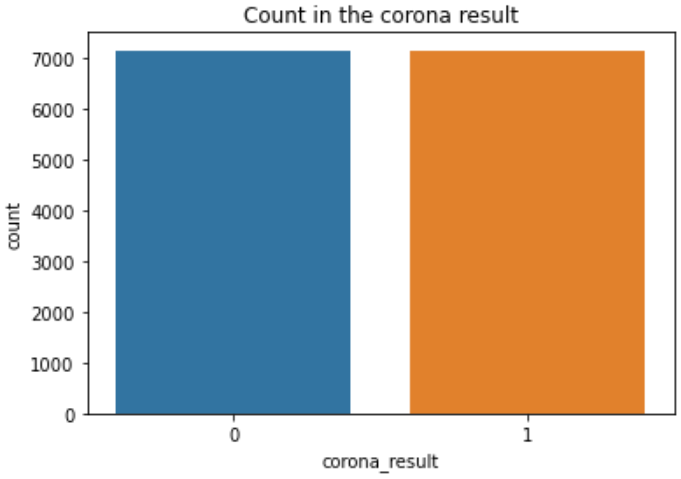

Figure 7: Count in the Target variable after performing under sampling of the Israeli ministry of health dataset.

V. Feature Selection

We dropped all columns that had missing values of more than 25%, which was only one feature, ‘age_60_and_above'. After performing all the prior processing steps, the dataset has 8 features. We performed pearson correlation on the data set to dissect the correlation between the features, we chose the pearson correlation method because is the most widely used correlation statistic to measure the degree of the relationship between two variables [51].

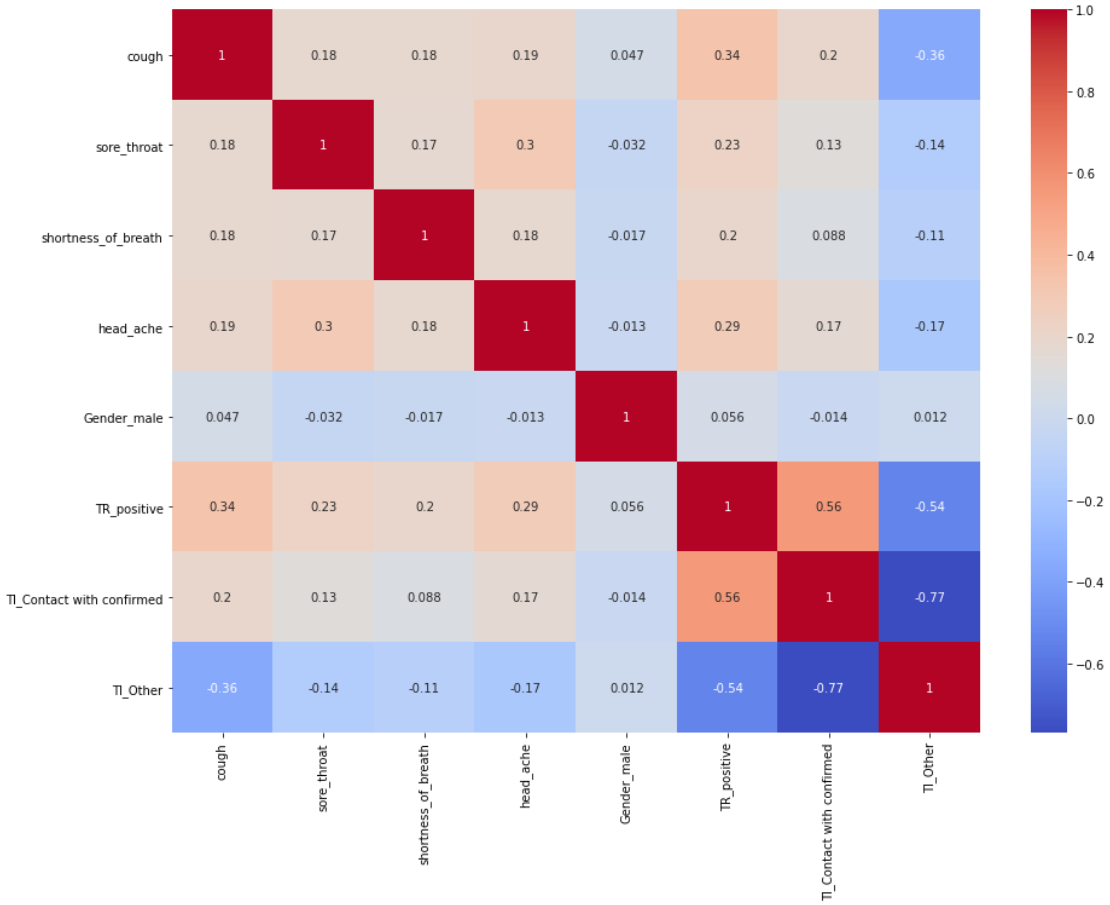

Figure 8: correlation Matrix of The Israeli Ministry of Health Dataset.

There is indeed a strong correlation between the target variable TR-positive and confirmed contact or any feature derived from the test indication (Figure 8). We can see that gender made no difference at all, therefore we decided to drop the gender column.

2. COVID-19 and Influenza Patients Dataset

The dataset is comprised of two separate datasets that have been aggregated together. The first one has 1485 rows, the second one has 413 rows, and both have 51 columns.

I. Handling Missing Values and Handling Bias Features.

Multiple columns were with no significant missing values had less than 6% missing values. We discarded 31 features with missing values more than 85%, We then divided the dataset back into its original form, into two independent datasets. Then, because the first data set contains patients infected with the coronavirus and the second dataset has patients infected with H1N1, we examined each dataset individually. We discovered several features with a large bias, such as the 'region' feature in the coronavirus dataset, which has 23% missing values, where in the H1N1 has 100% missing values in the same feature. If we do any imputation before removing the bias feature, the model will be utterly biased. several other features have the same issue. After all the pre-processing steps, the dataset shape is 7 columns and 1485 rows. Since 7 of the features are categorical, we used simple imputer with the most frequent to fill in the missing values.

II. Handling Categorical Features

As we mentioned before we need to transform category information into numerical features, we used one-hot encoding by getting dummies out of the Pandas library. We choose one hot encoding because we don't want the model to presume any ordering.

III. Imbalanced Target Variable

Following completion of all preparation operations, the dataset shape is 1485 rows and 7 columns, with the target variable 'Diagnosis', and the dataset contains 1072 patients diagnosed with H1N1 and 413 with Coronavirus, which is obviously unbalanced. Training a model on an imbalanced dataset can impact the model accuracy as explained above.

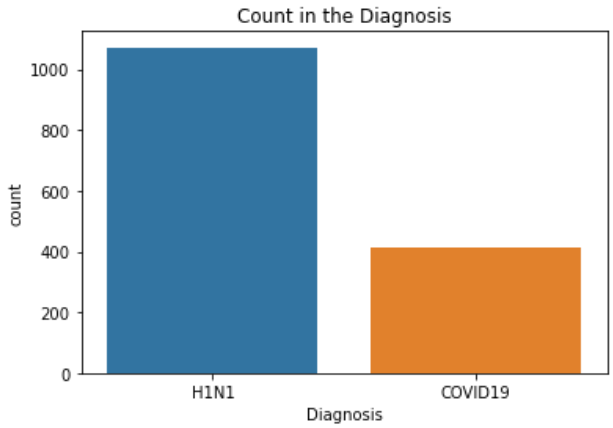

Figure 9: Imbalance in the targets feature for the COVID-19 and influenza patients' dataset.

We chose to employ the under sampling of the majority class method (Figure 10).

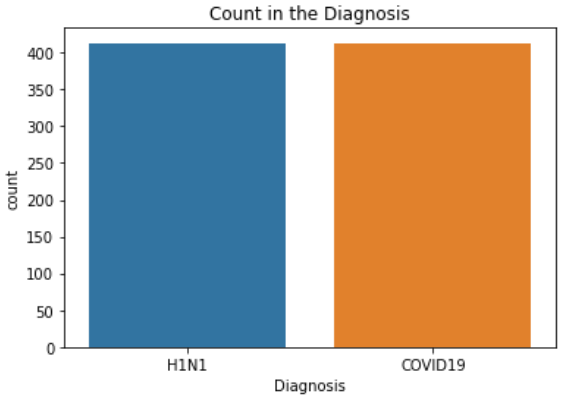

Figure 10: Distribution after performing under sampling for COVID-19 and influenza patients' dataset.

IV. Feature Selection

We discarded any feature with missing values of more than 85%. After all the preprocessing steps, the dataset shape is 7 columns. We performed Pearson correlation on the data set to dissect the correlation between the features, we selected the Pearson correlation for reasons we specified before.

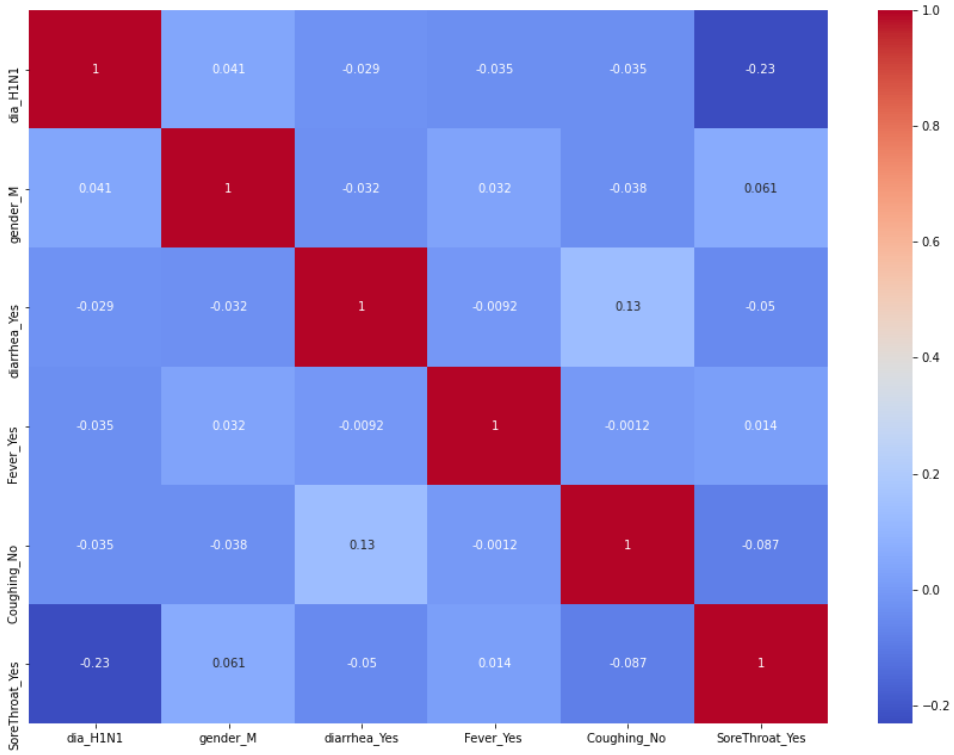

Figure 11: Correlation matrix of the COVID-19 and influenza patients' dataset.

We can observe that the output feature "dia_H1N1" column has a correlation with the 'sore_throat' column, while the 'coughing' column has zero to no correlation with the target variable in particular.

3. Albert Einstein Hospital, Brazil Dataset

The Albert Einstein Hospital dataset consisted of 5644 samples and 111 variables. The variables are divided into 4 dependent variables and 107 independent variables. The dependent variables are: 'SARS-Cov-2 exam result', 'Patient admitted to regular ward', 'Patient admitted to semi-intensive unit' and 'Patient admitted to intensive care unit'. In

our implementation, we only used 1 dependent variable, 'SARS-Cov-2 exam result' to predict COVID-19 infection.

I. Handling Missing Values

Other than 'Patient Age quantile' and 'SARS-Cov-2 exam result', all other features have missing values (Figure 12). The least feature with missing values is 76.01% empty, that is 1354 out of 5644. See figure (12) for a chart of how many values are present in each feature. Our approach into dealing with missing values is: 1) we first shrunk the dataset approximately to the least feature with missing values, 'Parainfluenza 1' which results in a 1352 sample dataset. Figure (13) presents the new feature value count. 2) Then we dropped features with more than 75% missing values, which leaves us a new dataset with 1352 samples and 32 features. 3) Finally, we used KNN imputation with k = 2 to fill in the missing values.

II. Handling Zero Variance Features

There was only 1 feature with 1 unique value, 'Parainfluenza 2' (zero-variance feature), which we dropped.

III. Handling Categorical Features

16 of the 31 features are categorical and have no missing values. We transformed categorical features using dummy encoding, which doubled the number of categorical features from 16 to 32.

IV. Feature Scaling

We standardized all features to have a mean of 0 and a variance of 1. This is important for algorithms such as SVM and KNN, where the distance between data points is critical.

V. Imbalanced Target Variable

The shape of the dataset after the above preprocessing now has 1352 samples and 47 features. The target variable has a class distribution of 1240 negative samples and 112 positive samples, which is very imbalanced, and training a model on a highly imbalanced set will only lead to a model that predicts the minority class poorly or simply a biased model.

So, to solve this issue, we created 11 splits of the dataset. Each set has all the 112 positive samples and a unique set of 112 negative samples. We ran 7 classifiers and used 5-fold cross validation to test and train the models. Each model is trained and tested on each split of the dataset, and every split produces 5 results using 5-fold cross validation. From the 5 results, we take the average, which then gives us the average result for one split. This is done for every split, which means 11 results will be there. We take the average of the 11 splits to produce the finale result for a specific classifier.

VI. Feature Selection

Selected features were those with fewer than 75% missing values. There are two independent features, “Hemoglobin” and “Hematocrit” who had high correlation, and therefore the one with lowest feature importance score, “Hemoglobin,” was dropped.

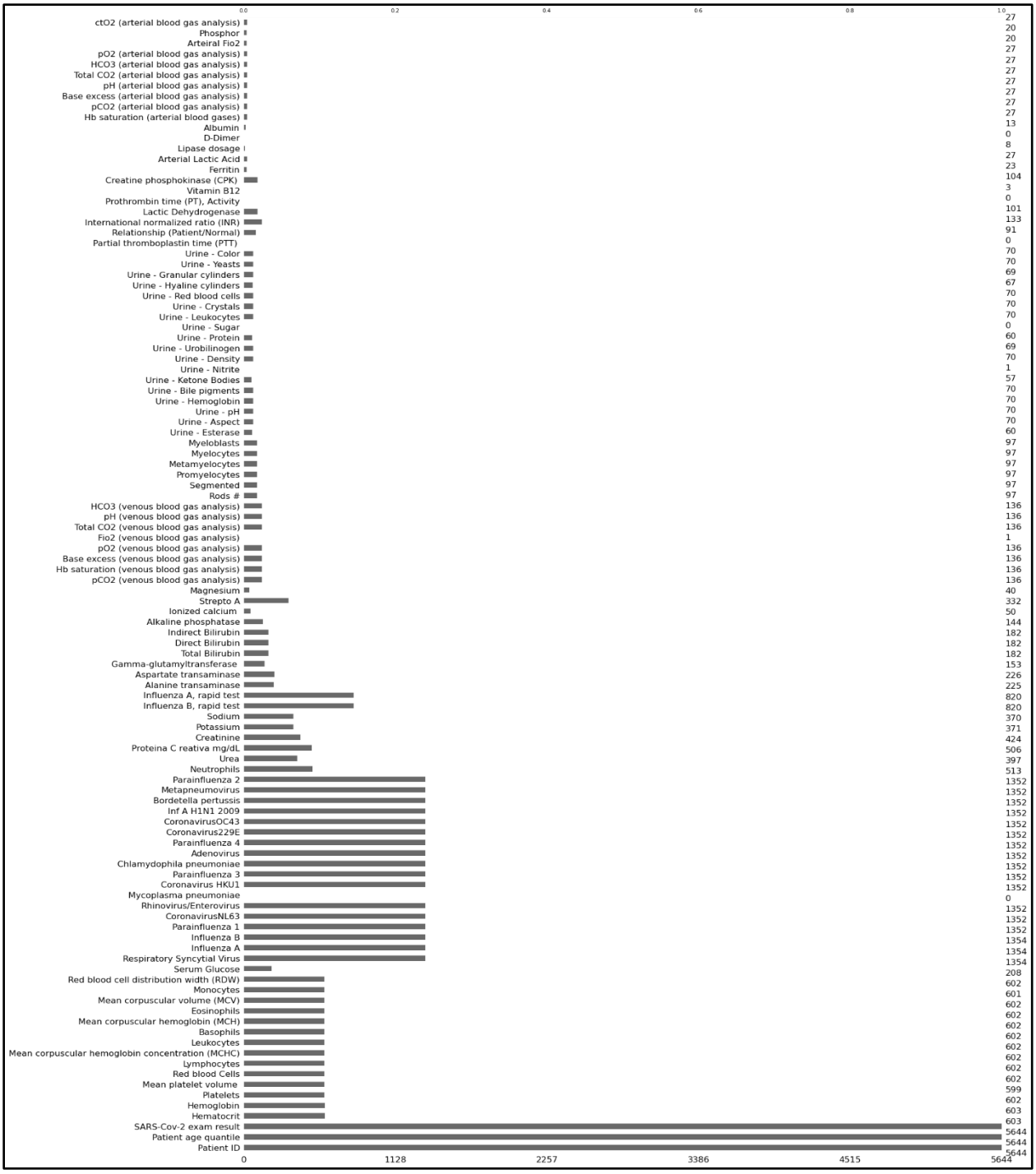

Figure 12: Bar chart shows features value count of Albert Einstein Hospital, Brazil Dataset.

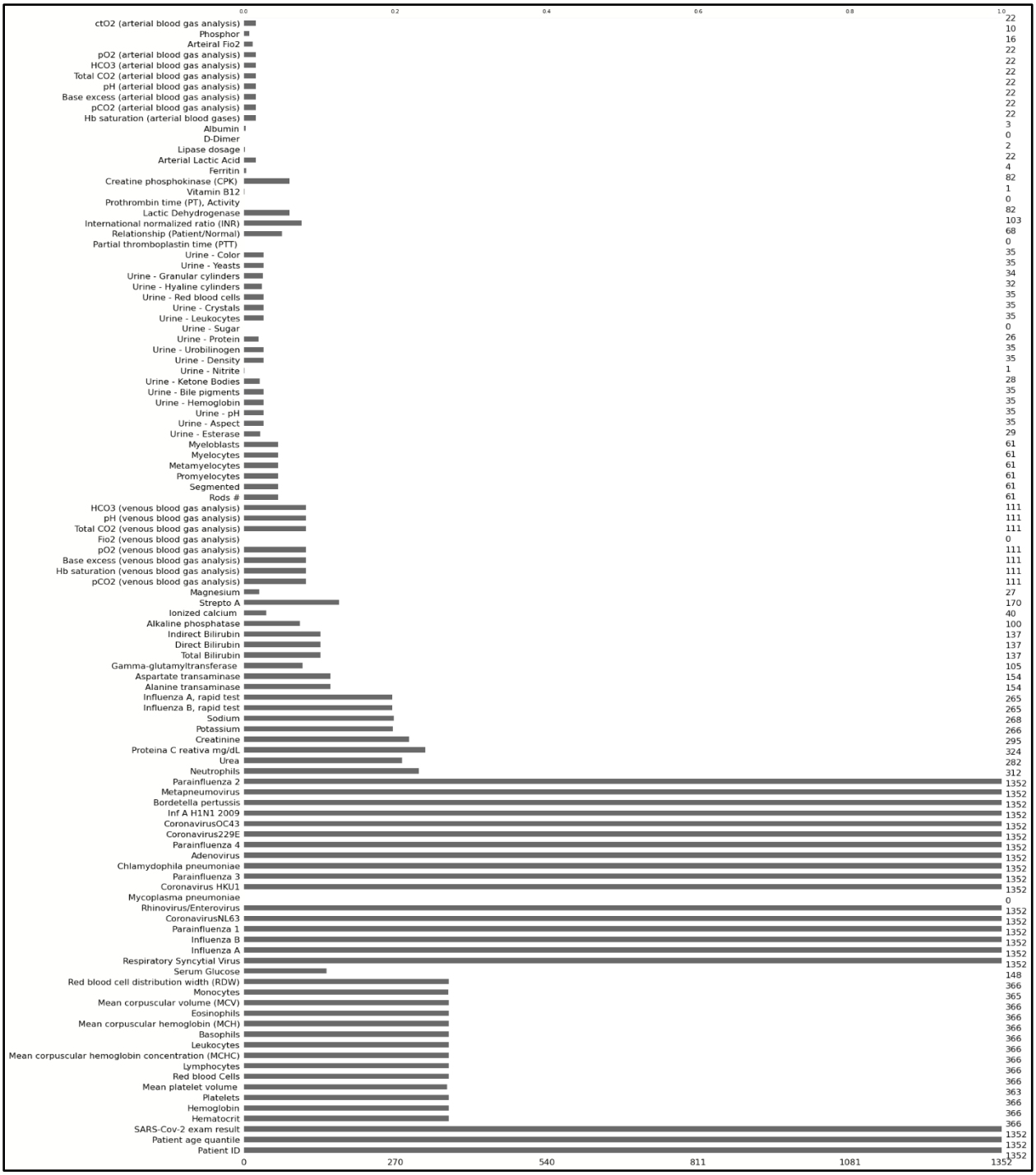

Figure 13: Bar chart of the features value count after shrinking the Albert Einstein Hospital, Brazil Dataset.

4. Italian Society of Medical Radiology Dataset (SIRM)

The SIRM dataset has 130 samples and 18 variables, 1 dependent variable 'Decision label' and 16 independent variables if you exclude sample id 'Number'. Out of the 130 samples, 68 of them were COVID-19 positive and the remaining 62 were Flu positive, so we set the 68 COVID-19 cases as positive or '1' and the 62 Flu cases as negative or '0'.

The people who constructed the dataset assigned missing values with a star symbol '*' instead of leaving them empty. To see how many missing values we are dealing with, we replaced every '*' with a NaN value. Figure (14) presents a bar chart of feature value counts.

I. Handling Missing Values.

Features with more than 50% missing values were dropped. The remaining 9 features are listed in Figure (15) for the updated feature value count. We used Simple Imputer with a mean strategy to fill in the missing values.

II. Handling Zero Variance Features

There was 1 feature, 'High risk zone' which has only 1 unique value and therefore dropped.

III. Handling Categorical Features

6 of the 8 features are categorical, so we manually transformed them into numerical features.

IV. Feature Scaling

We standardized all features to have a mean of 0 and a variance of 1.

V. Feature Selection

Selected features were those with less than 50% missing values.

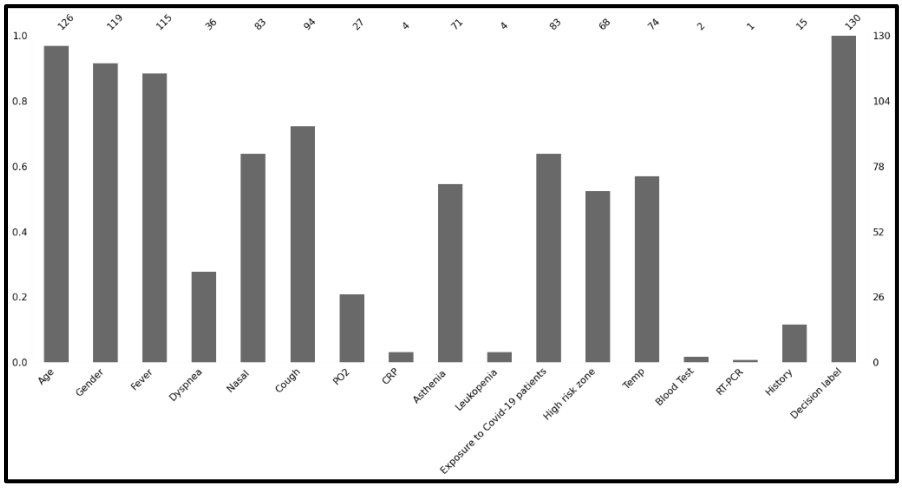

Figure 14: Bar chart of features value count for the SIRM dataset.

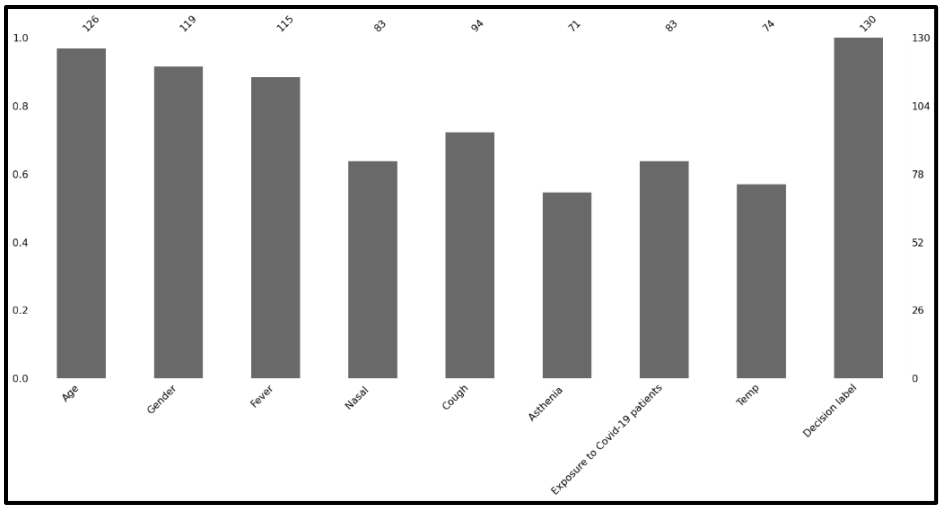

Figure 15: Bar chart of the updated features value count.

5. COVID-19 Chest Xray Dataset

The dataset is formed of 951 COVID-19 patients that have been combined with 5932 normal patients. The dataset has different views of the image, such as the posteroanterior (PA) and anteroposterior (AP) views. We only use the PA view with only 210 COVID-19 images. The reason we only use the PA view since the dataset has many views, with PA being the biggest. The normal images that were combined also form a PA view. The PA view is that the x-ray beam enters through the back of the chest and exits through the front.

I. Handling Missing Images.

In the Excel file, the images are located in the "filename" column. The few missing images in the "filename" are handled by removing the metadata row containing the missing image. COVID-19 has 210 images, but after handling for the missing images, only 188 images remain.

II. Imbalanced Target Variable

The dataset's shape is unbalanced. We simply combined the 188 COVID-19 images at random with the same number of normal images. The model was trained and tested using 10-fold cross validation as the default for imaging feature space.

III. Image Preprocessing

The dataset was preprocessed in stages, starting with grayscale conversion by combining the image's three channels into a single channel with the rgb2gray function. The image is then binarized by converting it to black and white in order to extract the necessary information from it. Following that, we use morphological operations to fill in any holes in any region of the image using the area closing function. After that, the area opening function is used to remove any noise from the background image. Following that, we use the label function to identify all regions of the image in label each region using connected components. In Figure (16), we see how images are preprocessed. The regionprops function can then be used to extract properties from the image's regions. The regionprops table function is then used to specify the properties extracted from the regions. The features are extracted using ten quantifiable properties: perimeter, equivalent diameter, mean intensity, solidity, eccentricity, area, convex area, bbox area, major axis length, minor axis length.

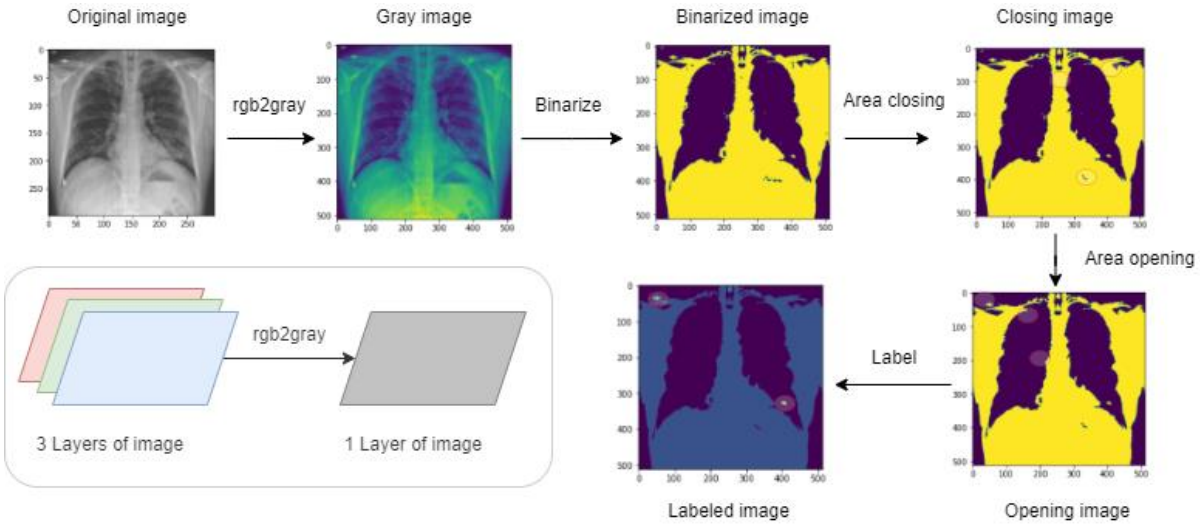

Figure 16: Shows the image preprocessing steps.

IV. Feature Engineering

We used the 10 extracted features to further expand the feature space. We took the extracted features from the image preprocessing phase and calculated the ratio of the new features. We ended up with 6 additional features, for a total of 16 features.

V. Feature Scaling

To avoid one variable from being overly influential, all the feature values were standardized to have a mean of 0 and a variance of 1.

6. COVID-19 Radiography Database

There are 3616 COVID-19 images and 10192 normal images in this dataset. All of the images are from the PA view.

I. Imbalanced Target Variable

We simply combined the 3616 Covid-19 images at random with the same number of normal images because the dataset's form was imbalanced. The model was trained and tested using 10-fold cross validation as the default for imaging feature space.

II. Image Preprocessing

The same image preprocessing was performed as for the previous dataset. Figure (16) shows the image preprocessing steps.

III. Feature Engineering

We performed the same feature engineering on the previous dataset.

IV. Feature Scaling

All features were normalized to have a mean of 0 and a variance of 1.

V. Feature Selection

In this step, we use the confusion matrix to detect the coloration of the features and remove any features with a value of 0.9 or higher. Following this step, the following features are removed: convex area, bbox area, major axis length, minor axis length, perimeter, equivalent diameter, area ratio convex, perimeter ratio minor.

3.3 Data Modeling

In this section, we will demonstrate the methods that we used to develop and evaluate our diagnostic models.

3.3.1 Learning

We built our models using a variety of ML algorithms, including Support Vector Machine (SVM), Decision Tree (DR), and Random Forest (RF). When we developed our ML pipelines, we used the algorithms' default settings (parameters). In developing our models, we considered seven ML algorithms as they are the most used algorithms in the reviewed literature. Each algorithm is described briefly below.

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 (17) illustrates the LR algorithm.

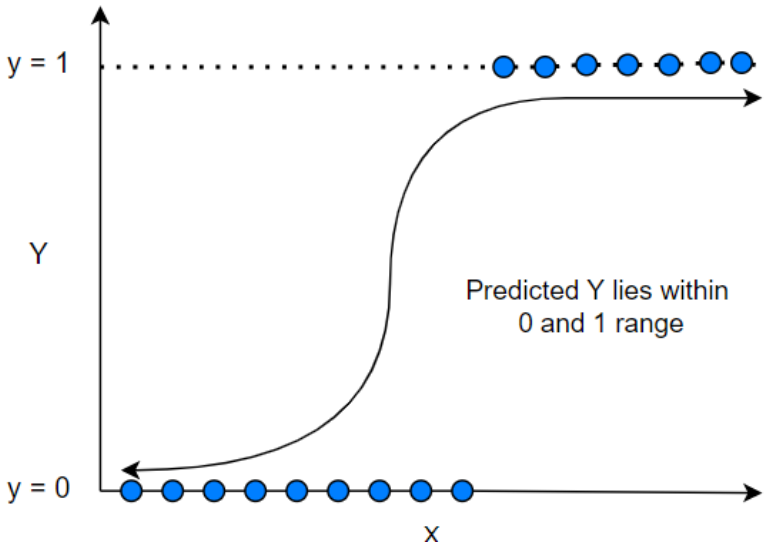

Figure 17: Logistic Regression Algorithm [28].

2. 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 the maximum margin between the two data points from the two classes [29]. Figure (18) illustrates the SVM algorithm.

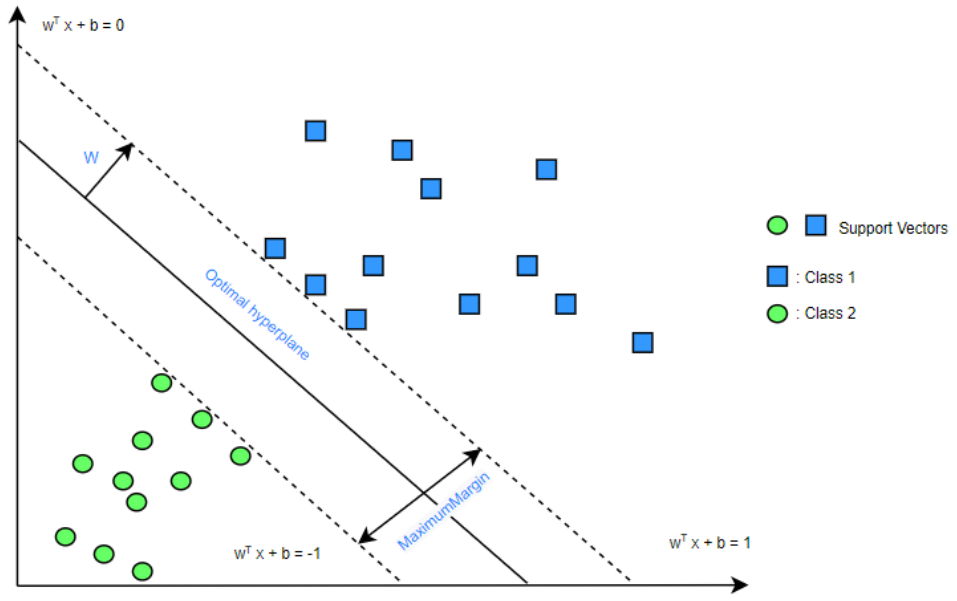

Figure 18: Support Vector Machine Algorithm [30].

3. 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 (19) illustrates the KNN algorithm.

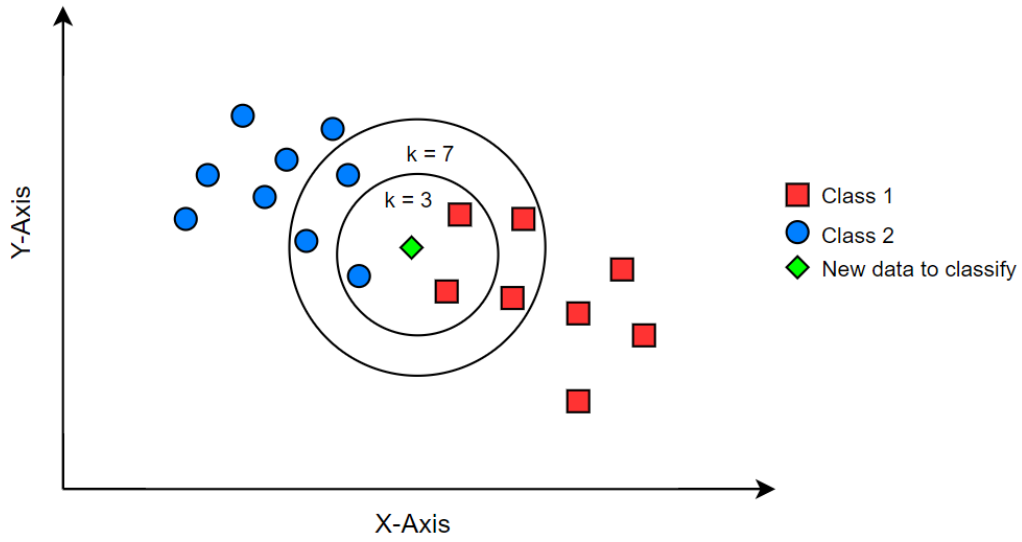

Figure 19: K-Nearest Neighbors Algorithm [32].

4. 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 the NB classifier is to calculate conditional probability [33]:

$$P\text{–}C_k x_1, x_2, \dots, x_n)$$

For each of 'k' possible outcomes or classes C_k. Let x = $(x_1, x_2, \dots, x_n)$. Using Bayesian theorem, we can get [33]:

$$P(C_k|x) = \frac{P(C_k)\,P(x|C_k)}{p(x)} \infty\, p(C_k)P(x|C_k) = p(C_k|x_1, x_2, \dots, x_n)$$

5. 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 (20) illustrates the DT algorithm.

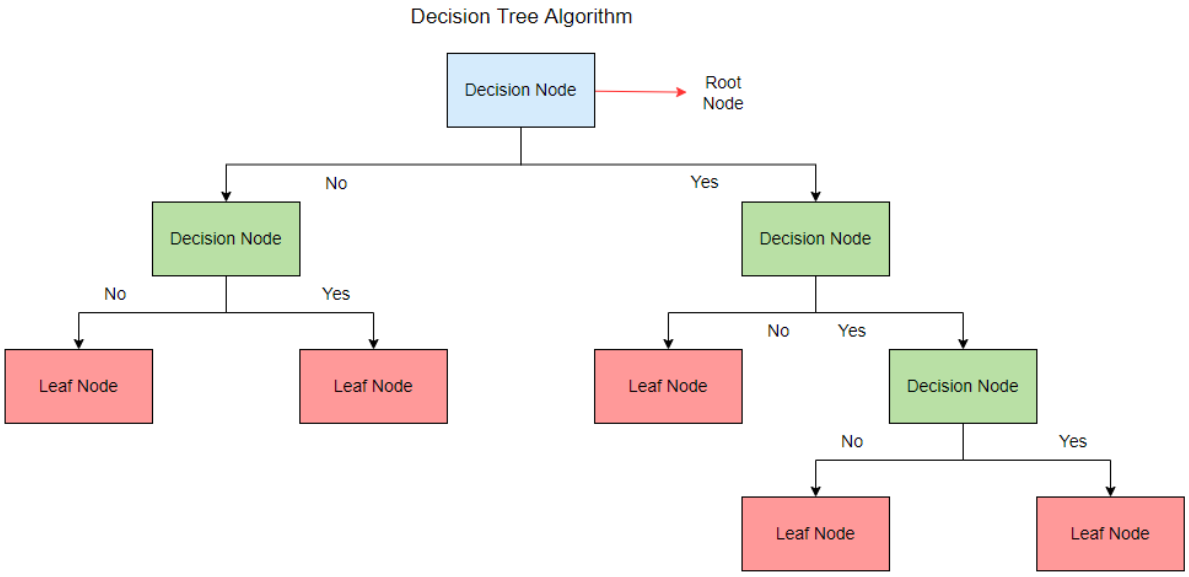

Figure 20: Decision Tree Algorithm [35].

6. 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 of those subsets. The RF takes the predictions from each tree and predicts the final output based on the majority votes of predictions [36]. Figure (21) illustrates the RF algorithm.

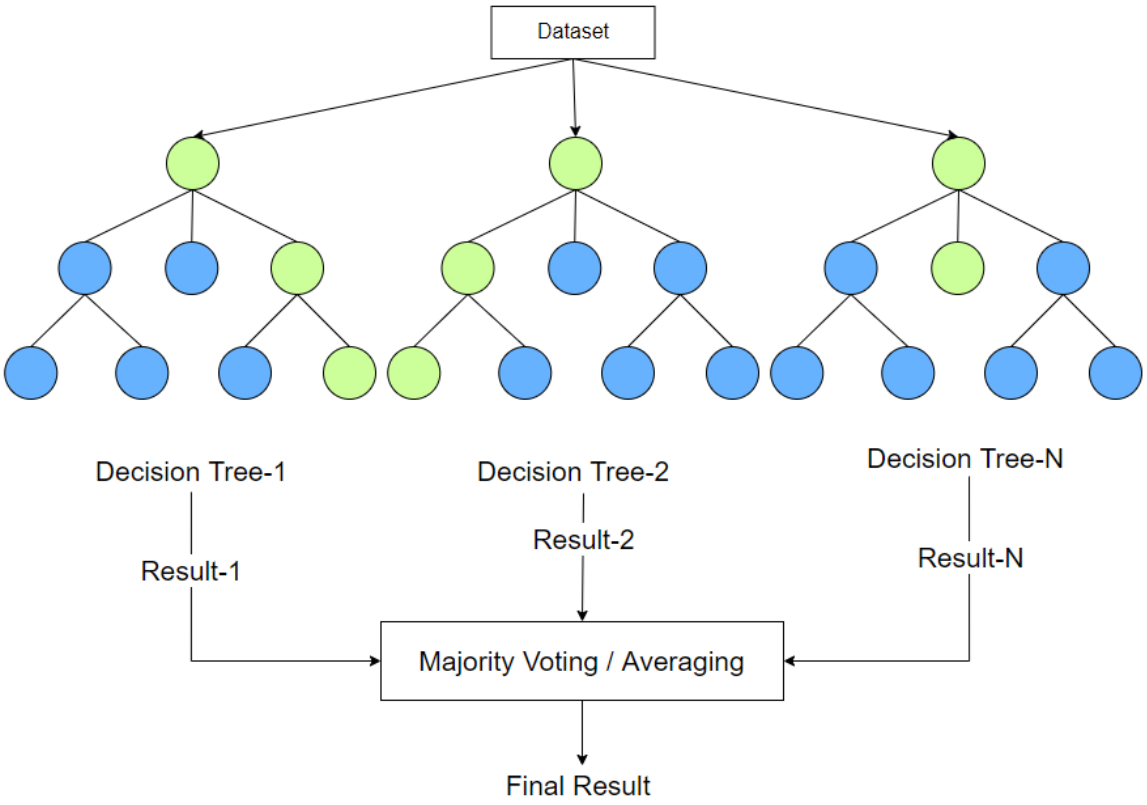

Figure 21: Random Forest Algorithm [37].

7. eXtreme Gradient Boosting (XGB)

XGB is a supervised ML algorithm for classification or regression. The XGB is a decision tree based on ensemble ML algorithms that uses a gradient boosting framework [38]. The XGB 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 (22) illustrates the XGB algorithm.

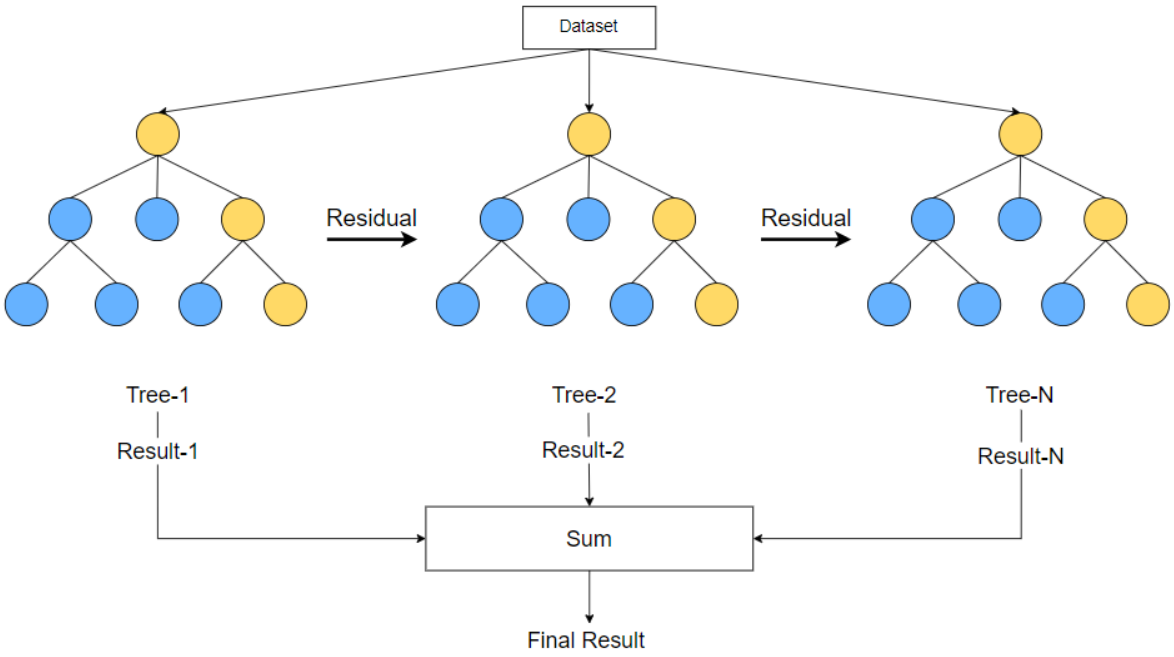

Figure 22: Simplified Structure of XGB Algorithm [40].

3.3.2 Evaluation

Our diagnostic models were evaluated using 3 metrics: precision, recall, and f1-score. These 3 are the most common metrics we found in the literature for identifying how precise the model is in predicting the target variable. Cross validation was used for our models to ensure that they extracted the proper patterns from the data and it is not getting up too much noise. Each metric is defined as follows:

Layman definition: Of all the positive predictions I made, how many of them are truly positive? [41].

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

Layman definition: Of all the actual positive examples out there, how many of them did I correctly predict to be positive? [41].

$$Recall = \frac{TP}{TP + FN}$$

Layman definition: Harmonic mean of precision and recall for a more balanced summarization of model performance [41].

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Chapter Four

Results

4.1 Results

4.1.1 The Israeli Ministry of Health Dataset

Our model's findings were similar to the model described in the paper [42], and the highest performing algorithms was random forest with 86% for precision, 85% for recall, and 85% for f1-score. Table (5) and figure (23) showcase our model performance.

Table 5: The Israeli Ministry of Health Dataset models scores

MODEL	PRECISION	RECALL	F1-SCORE
LR	83%	83%	83%
SVM	85%	84%	84%
KNN	84%	83%	83%
NB	84%	83%	83%
DT	86%	85%	85%
RF	86%	85%	85%
XGB	84%	84%	83%

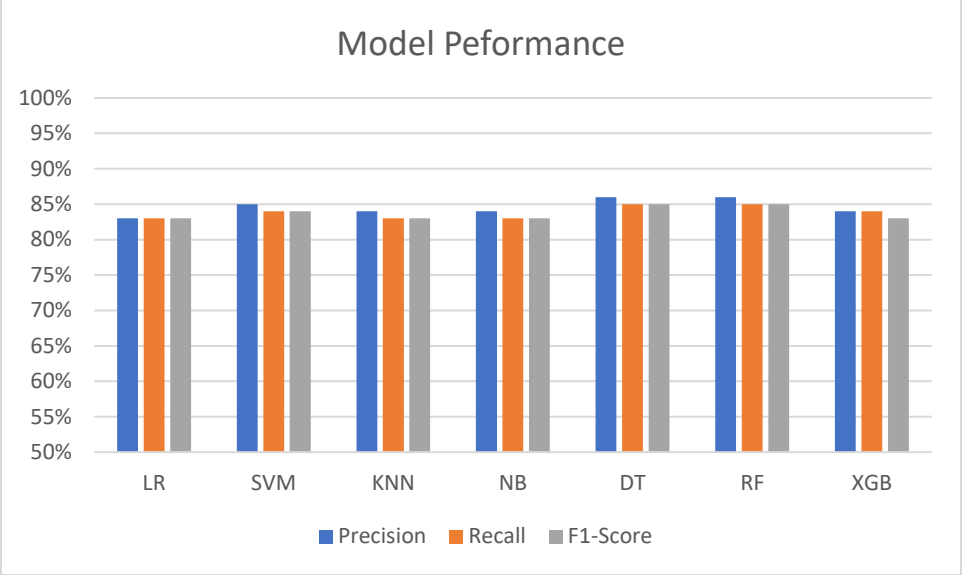

Figure 23: The Israeli ministry of health dataset models performance.

4.1.2 COVID-19 and Influenza Patients Dataset

Our findings were identical to the model provided in the paper [16], and the best-performing models were LR and RF, which scored 92% for precision, 92% for recall, and 92% for f1-score. Table (6) and figure (24) showcase our model performance.

Table 6: COVID-19 and Influenza Patients Dataset models scores

MODEL	PRECISION	RECALL	F1-SCORE
LR	92%	92%	92%
SVM	90%	90%	90%
KNN	72%	72%	72%

NB	72%	72%	72%
DT	91%	91%	91%
RF	92%	92%	92%
XGB	69%	62%	58%

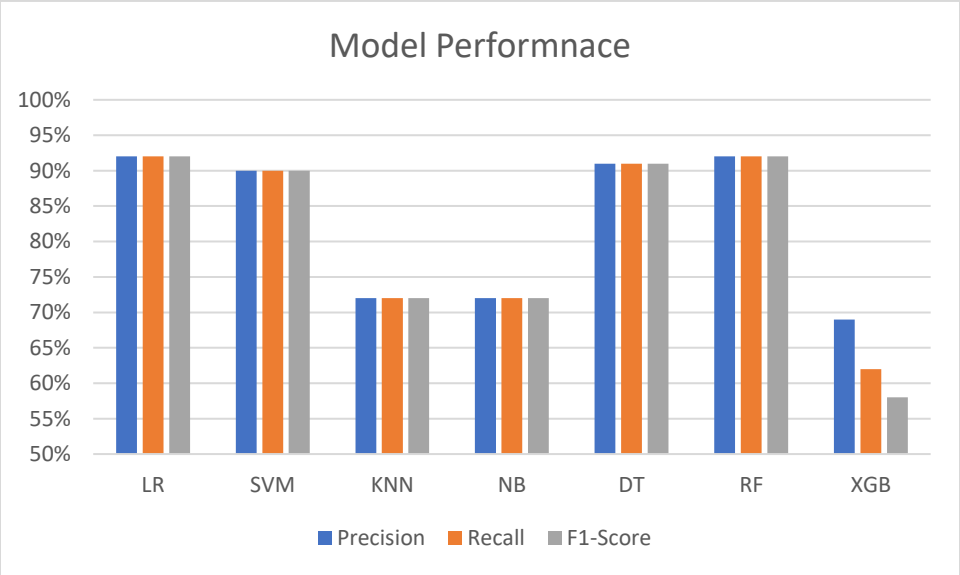

Figure 24: COVID-19 and Influenza Patients Dataset models performance.

4.1.3 Albert Einstein Hospital, Brazil Dataset

SVM has the best performance, with a precision of 81%, recall of 77%, and f1-score of 77%. In [19], their top performing model is RF, with a precision of 80%, recall of 75%, and f1-score of 77%, which is similar to ours. Table (7) and figure (25) highlights the performance results of our models.

Table 7:Albert Einstein Hospital, Brazil Dataset models scores.

MODEL	PRECISION	RECALL	F1-SCORE
LR	79%	76%	75%
SVM	81%	77%	77%
KNN	76%	74%	73%
NB	76%	60%	53%
DT	73%	71%	71%
RF	78%	75%	75%
XGB	77%	75%	75%

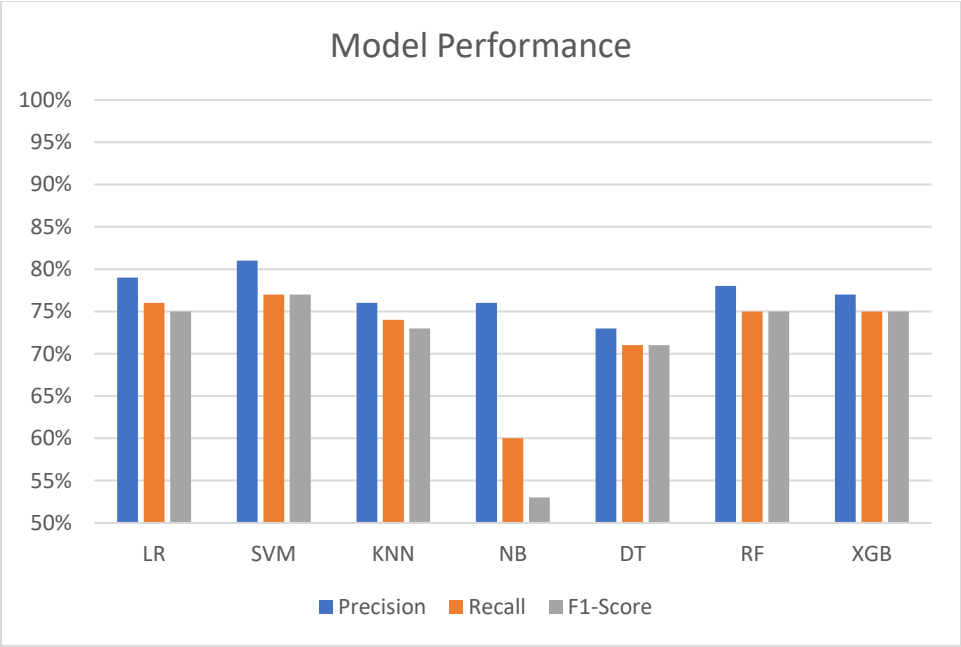

Figure 25: Albert Einstein Hospital, Brazil Dataset models performance.

4.1.4 Italian Society of Medical Radiology Dataset

The RF model scored the highest performance, with a precision of 83%, recall of 82%, and f1-score of 81%. Table (8) and figure (26) highlights the performance results of our models.

Table 8: Italian Society of Medical Radiology Dataset models scores.

MODEL	PRECISION	RECALL	F1-SCORE
LR	75%	75%	74%
SVM	80%	80%	77%
KNN	74%	73%	71%
NB	56%	58%	46%
DT	73%	75%	74%
RF	83%	82%	81%
XGB	74%	74%	74%

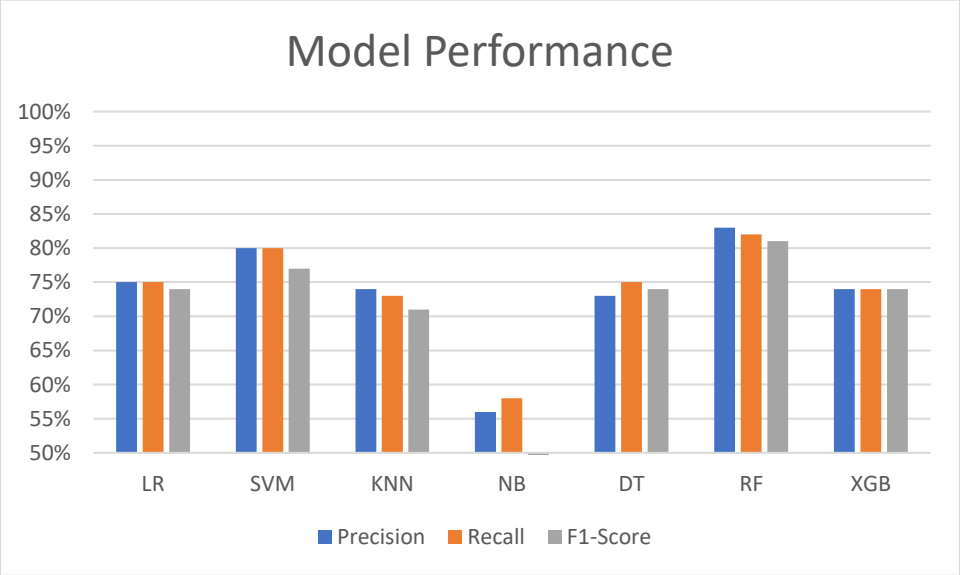

Figure 26: Italian Society of Medical Radiology Dataset models performance.

4.1.5 COVID-19 Chest Xray Dataset

Our data demonstrated that precision, recall, and f1-score were all low. In comparison, in [17], their KNN achieved great precision, recall, and an f1-score, which had 96.5%, 96.5%, and 96.4%, respectively. Our best classifier was RF, which had 72% precision, 71% recall, and a 71% f1-score. Tables (9) and figure (27) illustrate the results.

Table 9: COVID-19 Chest Xray Dataset model score (all features)

MODEL	PRECISION	RECALL	F1-SCORE
LR	65%	61%	60%
SVM	70%	66%	65%
KNN	64%	64%	63%
NB	59%	59%	58%
DT	66%	66%	66%
RF	72%	71%	71%
XGB	71%	70%	70%

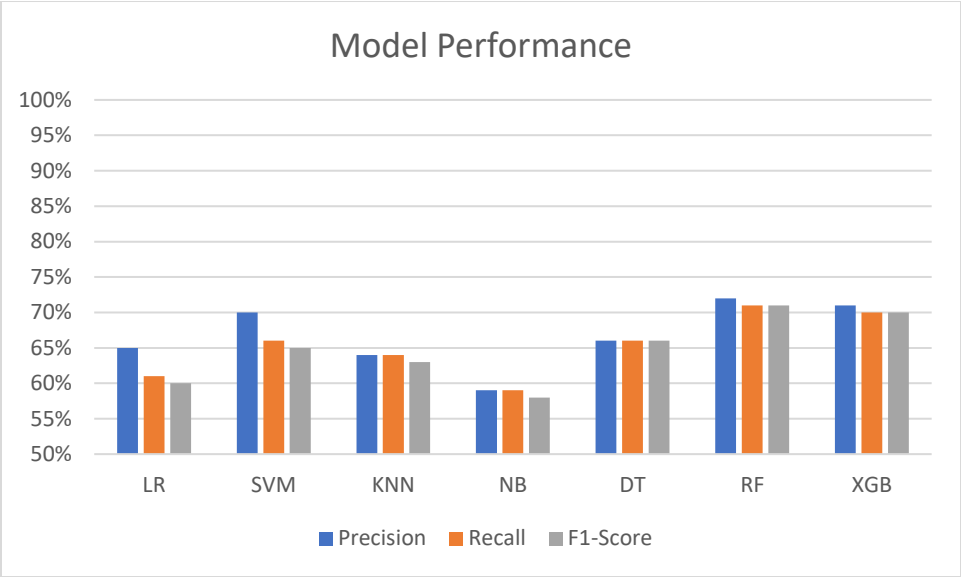

Figure 27: COVID-19 Chest Xray Dataset models performance.

4.1.6 COVID-19 Radiography Database

According to our results, precision, recall, and f1-score were all poor. The top classifier was RF, which achieved 70 % precision, 69% recall, and an f1-score of 69%. Table 10 demonstrate the findings and figure (28) will illustrate the table.

Table 10: COVID-19 Radiography Database models scores (all features)

MODEL	PRECISION	RECALL	F1-SCORE
LR	57%	57%	57%
SVM	65%	65%	65%
KNN	67%	67%	67%

NB	55%	53%	49%
DT	65%	65%	65%
RF	70%	69%	69%
XGB	66%	65%	65%

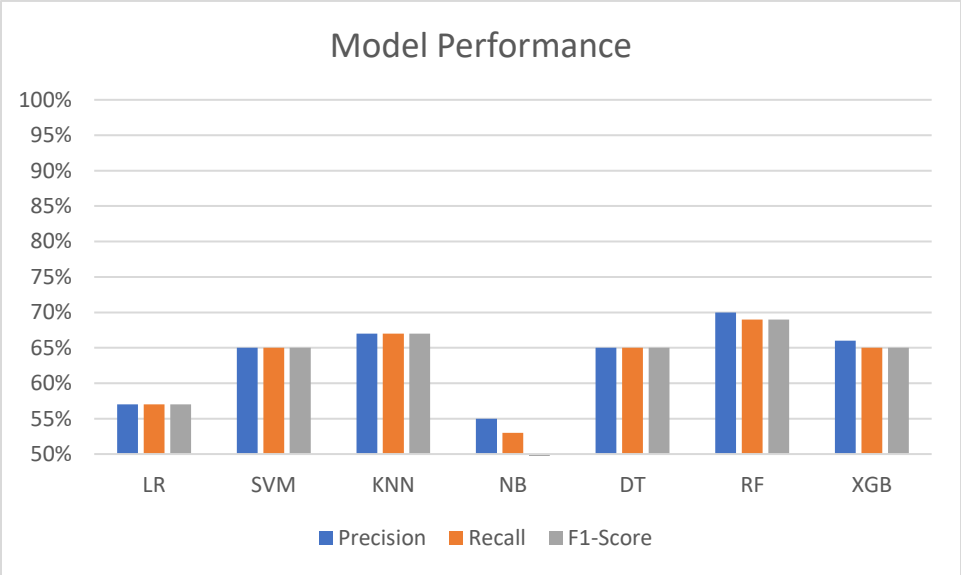

Figure 28: COVID-19 Radiography Database models performance (all features).

4.2 External validation

The practice of evaluating the original prediction model on a new group of patients to check whether it performs properly is known as external validation. The easiest approach to comparing the performance of two models is to validate them using entirely independent,

external data. An external validation was used for the purpose of testing the imaging feature space since the features are the same for both of our datasets in the imaging feature space. We trained the model using the bigger dataset, the COVID-19 Radiography Database. The model was then evaluated using the smaller dataset, the COVID-19 Chest X-ray Dataset. Our findings were evaluated using the best classifier RF, which had a precision of 60.4%, 63.8%, and 65.1% f1-score.

4.3 Overall Performance Summary

as a part of our research, we want to investigate if a specific algorithm can perform better than others in all different feature spaces, we can observe in figure (29) that the random forest classifier performed well compared to other algorithms with an average of 80% precision, 79% recall and 78.8% f1-score.

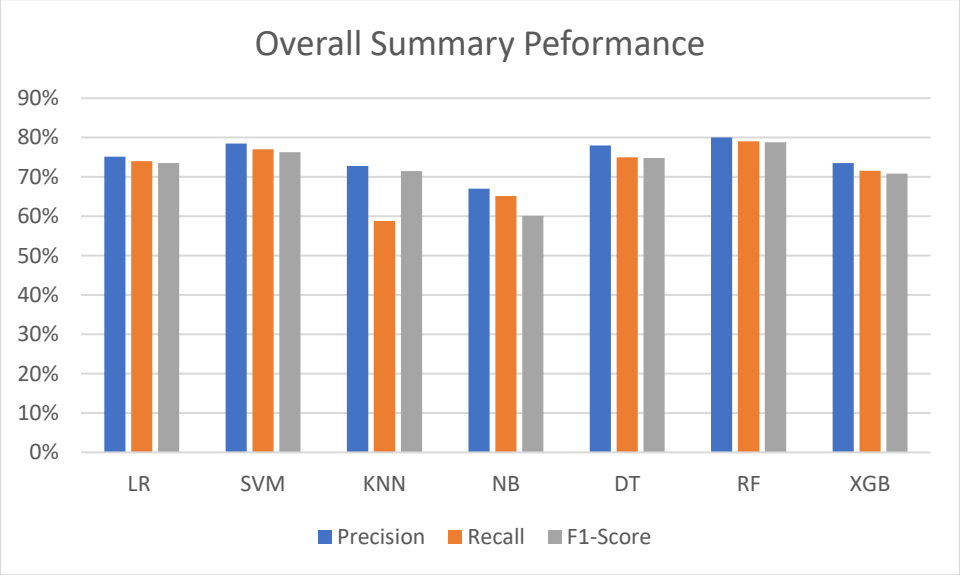

Figure 29: models overall performance summary for all feature spaces.

Chapter Five

Discussion

5.1 Data availability & quality

Different forms of data have different levels of accessibility. The accessibility and quality of data are two of the most essential components of data management. One of the challenges we face is the unavailability of public data sources. The majority of sources are private, and when we contact them, they either reject us or tell us to locate a public source.

Our goal was to obtain a large and high-quality laboratory dataset, but due to time and resource limitations, we had to settle for what was available. While looking for a laboratory dataset, we started to investigate the other two categories: symptoms and imaging, where we found a large and high-quality dataset. The symptom dataset was accessible, and the greatest part was that we found a massive dataset. On the other hand, the imaging dataset was accessible everywhere, but due to its high cost, we did not prioritize it.

5.2 Diagnostic modeling

Our models' performance varied across datasets and feature spaces, with f1-scores ranging from 46% to 92% and an overall average of 72.3%. Our models performed better in the symptoms feature space than in the laboratory and imaging-based feature spaces, with the laboratory coming in second and the imaging feature space coming in third.

With precision, recall, and f1-score of 92%, our LR and RF models developed using the COVID-19 and Influenza Patients Dataset performed the best.

5.3 Related Work Comparison

It is critical to compare our work to previously published studies to gain more perspective. Figure (30) illustrates the comparison of our models with the published related work for each feature space.

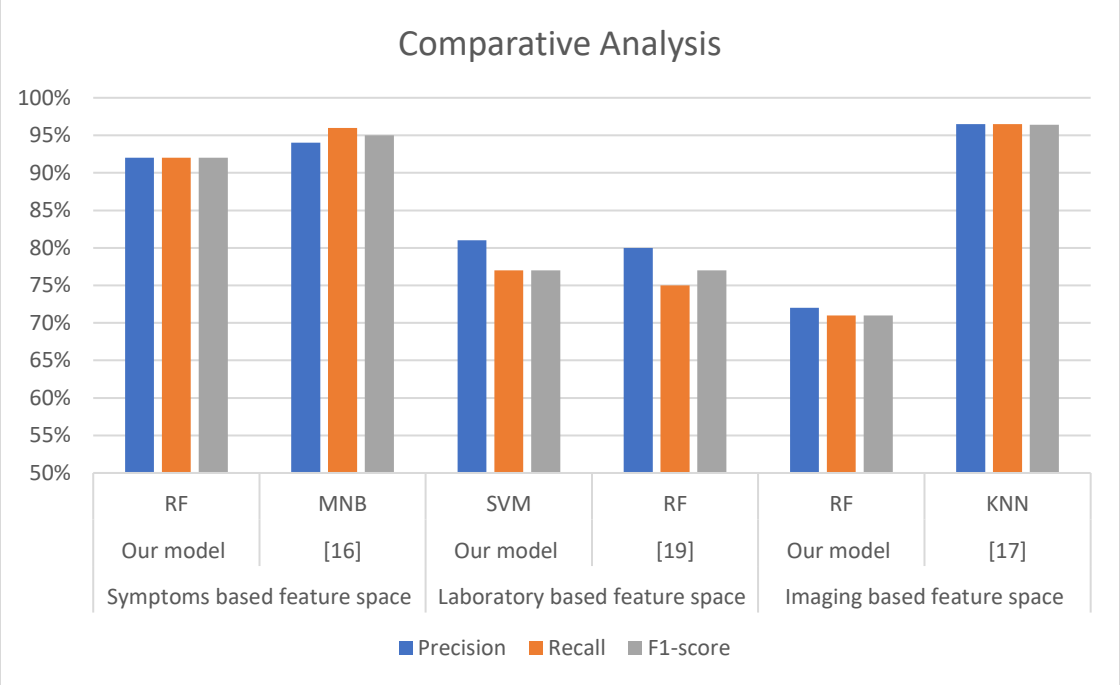

Figure 30: Comparison between our best performing models and the papers models.

For the symptoms feature space, we relied on data that was available to the public, which has some restrictions, biases, and missing information for some of the variables. We demonstrated that training and testing a model after filtering out symptoms of a strong bias. And we achieved extremely high accuracy.

It is important to understand that a one-to-one comparison is not possible due to differences in datasets, methods, and various simulation environments. Our findings were similar to the papers [42,16]. We can see that the datasets were aggregated between two different datasets and the authors didn't count for that. They imputed missing values for the entire dataset, which means assuming it was collected by the same third party, but after splitting the dataset into its original form, we can see that some features have high missing values in patients who were diagnosed with H1N1 and a low number of missing values in the patients who diagnosed with COVID-19. We can see that in their feature selection they didn't take that into account and that may cause the model to be biased. The most important features in the symptoms were the temperature and contact, which confirmed the same features were mentioned in both papers.

In the laboratory-based datasets, our best performing models were RF, SVM, and XGB, with f1-scores ranging from 75% to 81%. In [15], the best performing model was XGB with f1-score of 92.8%, which could confirm that XGB models when used on laboratory-based feature space it performs best. The most important features in our model were: leukocytes, platelets, age, eosinophils, and monocytes, which agrees with the other relevant papers. The 5-fold cross-validation technique was used for all algorithms to investigate the generalization of our model from training data to unseen data and to make sure our model was not overfitted.

The imaging dataset findings were pretty low compared to the related work findings. Our best algorithm was RF, with an average of 70.5% precision, 70% recall, and 70.5% f1-score. We think that the method we used to preprocess the images is the main cause of the poor results. Using other commonly used methods in the current literature, such as CNN, appeared

to be more helpful as they can be customized while with our used method the whole image details were considered in the preprocessing. Our initial intention was to consider all the details of the image because we are not medical experts. On the other hand, the best algorithm was from [17], which is a KNN classifier with 96.5% precision, 96.5% recall, and 96.4% f1-score. They used the CLD features to obtain numerical values from the medical images. We believe their findings were exaggerated because the dataset was so small.

5.4 Limitations

Every study has limitations. One of the limitations of this work is that our review of the literature was not systematic, which means that we did not evaluate all published articles that addressed COVID-19 diagnostic models. We conducted a nonsystematic review, which is defined as a critical assessment and evaluation of some but not all search studies addressing a specific issue [48].

Python's default values for algorithms parameters are represented differently. Default values indicate that if there are no arguments given during the algorithms call, the function argument will assume the values.

There are two benefits to the default values. The first is utilizing the function in a simpler matter since we don't have to worry about the different values of the arguments for the parameters. The second is that we can only assign values to the parameters we choose to change, and the other parameters will have the default argument values.

The datasets have several limitations. For example, some of the datasets didn't include an important feature in that feature space, and they contain a substantial number of missing values, which might have a negative impact on model performance.

Chapter Six

Conclusion

6.1 Conclusion

In conclusion, we demonstrated the use of ML models to predict COVID-19 presence using 3 types of datasets: imaging, laboratory, and symptoms. The ML models were constructed using publicly available data from the published literature. Our models may play an important role in assisting in the identification of SARS-CoV-2 infected patients in areas where RT-PCR testing is not possible due to financial or supply constraints. Our results have illustrated the potential role for these models as a tool to preliminarily identify the high-risk of SARS-CoV-2 infected patients before their RT_PCR results are available. The ML models described in this paper performed comparably to the RT-PCR Test [49], the current gold standard for COVID-19 diagnosis.

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.

[41] K. Leung, "Towards Data Science," [Online]. Available: https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f.

[42] S. D.-R. &. N. S. Yazeed Zoabi, "nature," 4 1 2021. [Online]. Available: https://www.nature.com/articles/s41746-020-00372-6.

[43] T. R. A. K. R. M. M. A. K. Z. B. M. K. R. I. M. S. K. P. A. I. N. A.-E. P. M. B. I. R. Muhammad E. H. Chowdhury, "Kaggle," [Online]. Available: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database.

[44] I. A. A. I. A. M. Maryam AlJame, "Science Direct," 20 10 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914820305992.

[45] A. Hamed, "Harvard Dataverse," 5 5 2020. [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LQDFSE.

[46] A. D. S. B. D. a. S. B. Patrick Schwab, "Research Gate," 5 2020. [Online]. Available: https://www.researchgate.net/profile/Patrick-Schwab/publication/341478619_predCOVID-19_A_Systematic_Study_of_Clinical_Predictive_Models_for_Coronavirus_Disease_2019/links/5ec3c814299bf1c09ac94f0f/predCOVID-19-A-Systematic-Study-of-Clinical-Predictive-Mode.

[47] D. Pierron, "Smell and taste changes are early indicators of the COVID-19 pandemic and political decision effectiveness," *Nature Communications,* 2020.

[48] "HealthyPeople," 6 2 2022. [Online]. Available: https://www.healthypeople.gov/2020/Implement/EBR-glossary#:~:text=Nonsystematic%20Review%3A%20A%20non%2Dsystematic,that%20address%20a%20particular%20issue..

[49] M. Gandhi, "thermofisher," 23 10 2020. [Online]. Available: https://www.thermofisher.com/blog/clinical-conversations/why-qpcr-is-the-gold-standard-for-covid-19-testing/#:~:text=A%20COVID%2D19%20antigen%20test,the%20virus%20in%20their%20body..

[50] M. D. W. P. J. D. M. N. W. M. H. H. M. H. X. M. a. C. X. M. Dasheng Li, "Korean Journal of Radiology," 5 3 2020. [Online]. Available: https://www.kjronline.org/DOIx.php?id=10.3348/kjr.2020.0146.

[51] "Correlation (Pearson, Kendall, Spearman)," *Complete Dissertation by statistics solutions.*

Appendix

A. Modeling and Evaluation

1. The Israeli Ministry of Health Dataset:

Table (11) and figure (31) indicate our model's performance without adjusting the imbalance in the target feature.

Table 1111: The Israeli Ministry of Health Dataset models scores [imbalanced].

MODEL	PRECISION	RECALL	F1-SCORE
LR	80%	81%	78%
SVM	86%	79%	83%
KNN	88%	83%	85%
NB	88%	83%	85%
DT	91%	83%	86%
RF	72%	81%	75%
XGB	73%	78%	72%

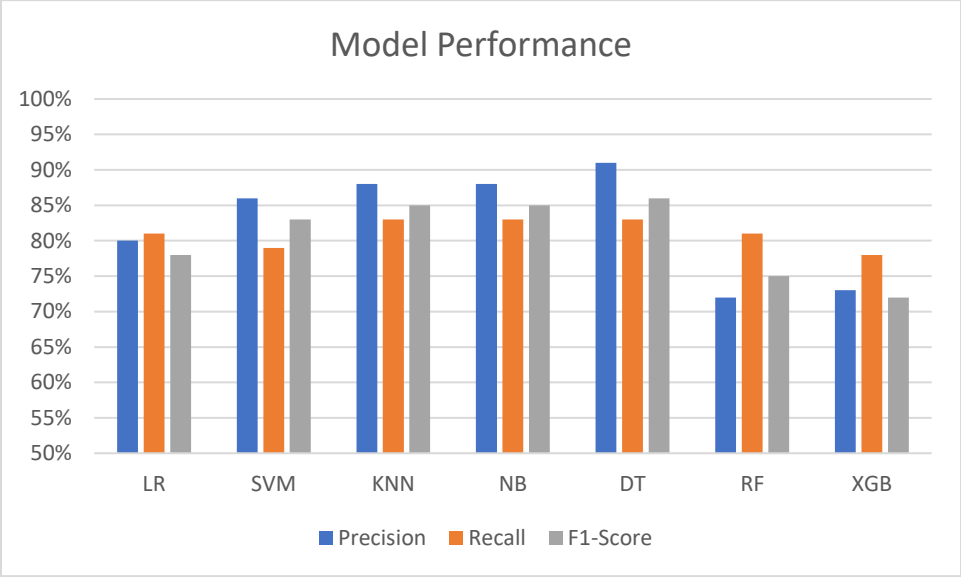

Figure 31: The Israeli ministry of health dataset models performance [Imbalanced].

2. COVID-19 and Influenza Patients Dataset:

Table (12) and figure (32) show our results without taking the imbalance in the target variable into account.

Table 1212: COVID-19 and Influenza Patients Dataset models scores [imbalanced].

MODEL	PRECISION	RECALL	F1-SCORE
LR	81%	83%	77%
SVM	84%	85%	85%
KNN	72%	69%	71%
NB	72%	69%	71%
DT	91%	87%	89%
RF	96%	92%	94%

XGB	75%	80%	78%

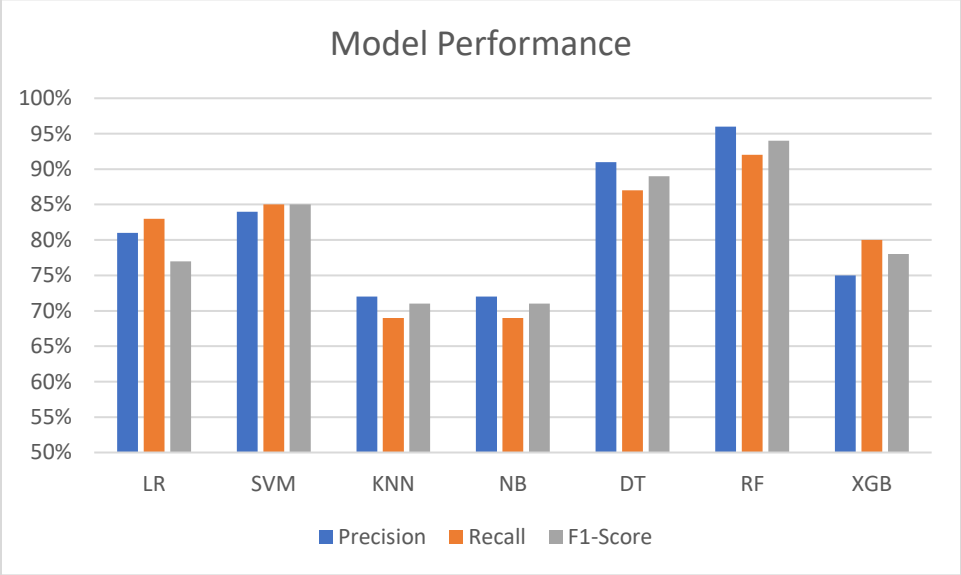

Figure 32: COVID-19 and Influenza Patients Dataset models performance[imbalanced].

3. Albert Einstein Hospital, Brazil Dataset:

Table (13) and figure (33) show model results without accounting for the target variable's unbalanced class distribution.

Table 13: Albert Einstein Hospital, Brazil Dataset models scores [Imbalanced].

MODEL	PRECISION	RECALL	F1-SCORE
LR	57%	51%	50%
SVM	47%	50%	48%
KNN	68%	62%	64%

NB	54%	60%	26%
DT	74%	62%	65%
RF	94%	60%	64%
XGB	87%	57%	61%

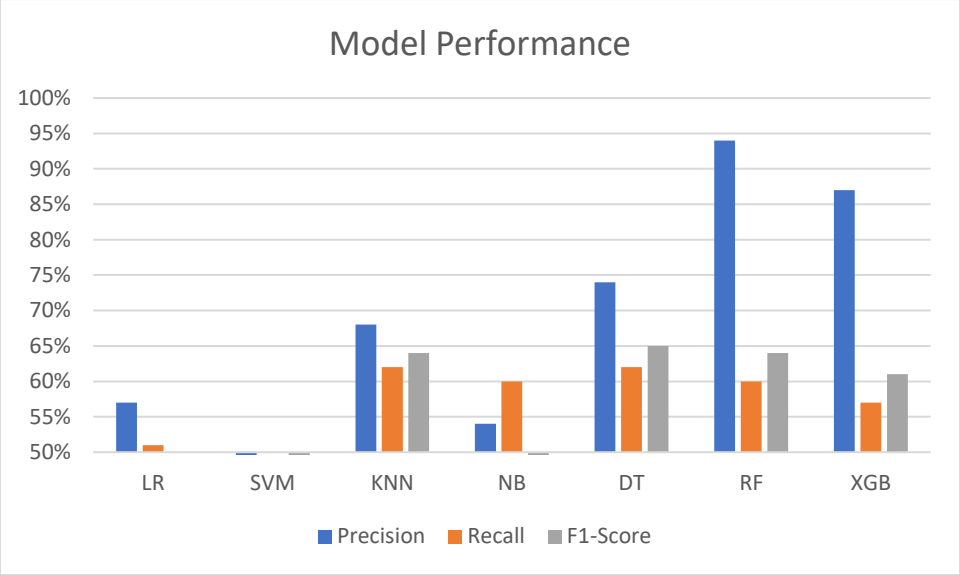

Figure 33: Albert Einstein Hospital, Brazil Dataset models performance [Imbalanced].

4. COVID-19 Chest Xray Dataset:

Table (14) and figure (34) show model results with the features with high correlation dropped.

Table 14 COVID-19 Chest Xray Dataset model score (the features with high correlation dropped)

MODEL	PRECISION	RECALL	F1-SCORE
LR	64%	62%	61%

SVM	69%	68%	68%
KNN	64%	64%	64%
NB	60%	58%	56%
DT	63%	64%	63%
RF	72%	71%	71%
XGB	73%	71%	71%

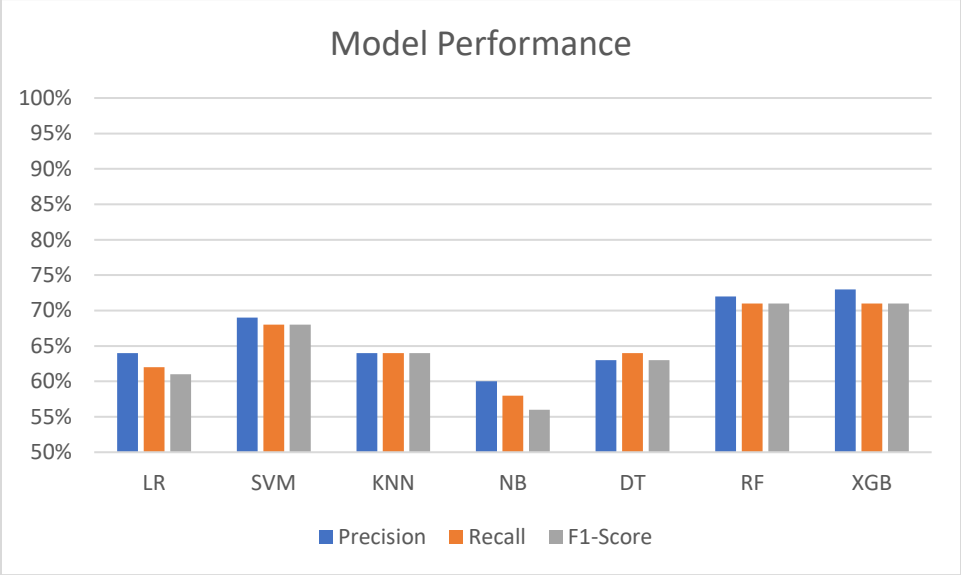

Figure 34: COVID-19 Chest Xray Dataset models performance (the features with high correlation dropped).

5. COVID-19 Radiography Database:

Table (15) and figure (35) show model results with high correlation features dropped.

Table 15: COVID-19 Radiography Database models scores (the features with high correlation dropped)

MODEL	PRECISION	RECALL	F1-SCORE
LR	54%	54%	52%
SVM	65%	65%	65%
KNN	66%	66%	66%
NB	55%	52%	43%
DT	65%	65%	65%
RF	69%	69%	69%
XGB	65%	65%	65%

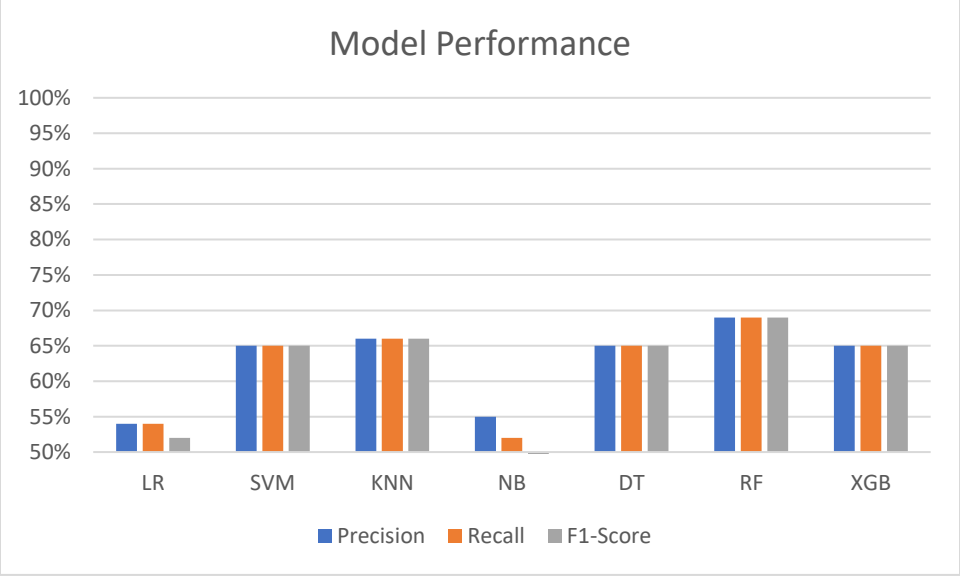

Figure 35: COVID-19 Radiography Database models performance (the features with high correlation dropped).