College of Computer and Information Sciences
Computer Science Department

Surveillance and Analysis Artificial Intelligence Model of Seasonal Flu Diseases: Flu Vigilant

CSC 497/6 – Final Report

Prepared by:

Mashael Aljebreen	438202740
Nouf Alharthi	441201274
Raseel Alhaqbani	441201421
Enas Alzahrani	441201082

Supervised by:

Dr.Mai Alzamel

Research project for the degree of Bachelor in Computer Science
Third Quarter 1444

I.Acknowledgements

We extend our deepest appreciation and heartfelt thanks to Dr. Mai Alzamel, our esteemed supervisor, for her invaluable guidance and unwavering support throughout our research journey. Her mentorship has not only been a source of immense knowledge but has also played a pivotal role in enhancing our research skills.

We would also like to acknowledge our families and friends for their constant encouragement and unwavering support, which has been instrumental in helping us push through the challenges of this project. Without their belief in us, we would not have been able to persevere, and we are deeply grateful for their unwavering faith in us. Finally,

We express our gratitude for the opportunity to work on this project, and we are thrilled to share our findings with the world.

II.English Abstract

In 2019, the COVID-19 pandemic began, resulting in the deaths of millions of people worldwide. The pandemic caused catastrophic economic and healthcare costs, resulting in the greatest crises over two centuries. With numerous patients being screened for COVID-19, computer-assisted detection can effectively improve clinical workflow efficiency and reduce the prevalence of infections among humans. Therefore, there is a need for public health surveillance to detect outbreaks of disease. In this research, we developed a machine learning model to predict the outbreaks of COVID-19. We used a linear regression model to predict the anomaly number of COVID-19 cases in the Riyadh region. The model was trained using Google Trend datasets and evaluated using registered real cases Once the model was trained, it is used to generate predictions for the number of COVID-19 cases in the future. If the model predicts a significant increase in the number of COVID-19 cases, an early warning will be sent to medical centers in the affected region. This will allow medical centers to prepare for the increase in cases, such as by increasing the number of beds in the hospital and ordering more supplies. By implementing this innovative machine learning approach, we believe our research has significant potential to help save lives and reduce the economic impact of the COVID-19 pandemic. Our findings highlight the power of AI in public health surveillance.

III.Arabic Abstract

في عام 2019، بدأ وباء كوفد-19 في الانتشار وأسفر عن وفاة الملايين من الأشخاص في جميع أنحاء العالم. وتسبب الوباء في تكاليف اقتصادية وصحية كارثية تعتبر من أكبر أزمات القرنين الماضيين. يمكن للكشف المساعد باستخدام الحاسوب تحسين كفاءة الكشف عن مثل هذه الزيادة وتقليل انتشار العدوى بين البشر. لذلك، هناك حاجة إلى المراقبة الصحية العامة لكشف تفشي الأمراض. في هذا البحث، قمنا بتطوير نموذج تعلم الآلة للتنبؤ بتفشي فيروس كورونا المستجد. استخدمنا نموذج الانحدار الخطي للتنبؤ بعدد حالات كوفد-19 في منطقة الرياض. حيث تم تدريب النموذج على مجموعة بيانات تاريخية تضمنت عدد مرات البحث عن اعراض مرض كوفد-19 في محرك بحث قوقل. بمجرد تدريب النموذج سيتم استخدامه لتوليد توقعات لعدد حالات كوفد-19 في المستقبل. إذا تنبأ النموذج بزيادة كبيرة في عدد حالات كوفد-19، فسيتم إرسال إنذار مبكر إلى المراكز الطبية في المنطقة المتأثرة. وهذا سيتيح للمراكز الطبية التحضير للزيادة في الحالات، مثل زيادة عدد الأسرة في المستشفى وطلب مزيد من الإمدادات. نحن نعتقد أن هذا المشروع له القدرة على إنقاذ الأرواح وتقليل الآثار الاقتصادية لجائحة كوفد-19 والأمراض الموسمية.

Table of Contents

List of Tables

List of Figures

List of Notations

AI	**Artificial Intelligence**
ML	**Machine Learning**
LR	**Linear Regression**

MLR	**Multivariate Linear Regression**
MAE	**Mean Absolute Error**
MSE	**Mean Squared Error**
TP	**True Positive**
FP	**False Positive**
TN	**True Negative**
FN	**False Negative**
SVM	**Support Vector Machine**
DNN	**Deep Neural Network**
CNN	**Convolutional Neural Networks**
NB	**Naive Bayes approach**
RF	**Random Forest**
KNN	**K-Nearest Neighbors**
DT	**Decision Tree**
IRD	**Influenza Research Database**
EMC	**Epidemiological Monitoring Center**
LightGBM	**Light Gradient Boosting Machine**
XGBoost	**Extreme Gradient Boosti**

Chapter 1: Introduction

In December 2019, Wuhan, China, received the first report of the coronavirus disease COVID-19. Later, in January 2020, the World Health Organization declared it a "public health emergency of international concern". By June 2020, the virus had spread to 213 countries, causing nearly 7 million cases of COVID-19 and over 400,000 fatalities[1].

Most virus-infected individuals develop mild to moderate respiratory illnesses and recover without special care. However, some individuals become extremely ill and require specialized medical attention. Older and those with underlying medical conditions like cancer, diabetes, chronic respiratory disease, or cardiovascular disease are more likely to experience severe illness. Anyone can get sick, become gravely ill, or die from COVID-19 at any age [2].

The number of patients who visit health centers may increase significantly due to the COVID-19 pandemic, making it challenging for the facilities to care for them and take the necessary actions, especially since they are not always prepared with the beds and equipment. Since many patients are increasingly using Google to search for information about their symptoms before visiting the hospital, and this is because Google is a convenient and easy way to access medical information, also they can search for information about their symptoms at any time of day or night, and they can do so from the comfort of their own home, additionally Google provides a wealth of information about a variety of medical topics, including symptoms, diagnosis, treatment, and prevention.

Our objective is to develop a linear regression model that helps predict cases of the spread of Covid-19 to alert health centers and help them to be prepared to receive large numbers of patients. This will be accomplished by utilizing Google Trend, which is an online service that tracks the evolution of search phrase popularity, and this information can be utilized to spot trends that are of relevance to the general public, such the spread of viruses, also analyzes a sample of Google web searches to determine how many searches were conducted over a certain period [3].

1.1 Problem Statement

COVID-19 is currently one of the world's most widespread and persistent viruses. A fever, cough, and loss of taste, and loss of smell are the most common COVID-19 symptoms. There is a higher risk of severe illness in elderly patients and those with underlying medical problems. When Covid-19 infections are on the rise and healthcare facilities are not prepared, problems such as overcrowded medical facilities, the spread of the infection, and a lack of oxygen equipment and certain medications will arise.

Accurately estimating the magnitude and timing of seasonal disease incidence peaks aids in comprehending the underlying causes and, potentially, in the development of interventions [4]. COVID-19 vaccines are being produced to effectively control the pandemic and prevent the loss of thousands of lives [5]. Therefore, doctors need to forecast the Covid-19 virus and research it in advance to prevent disasters and fatalities by creating an artificial intelligence surveillance and analysis model for the Covid-19 virus, AI models can provide real-time insights into the spread of the virus and help healthcare providers prepare for potential surges in demand for medical resources. Our research aims to address these issues.

The graph below demonstrates the relationship between google searches for covid and the number of cases. We looked at google trends which gave us an idea of how many times people are searching for covid symptoms (the blue line). We noticed a while later that the more people are searching, the higher number of confirmed cases (the red line).

Figure 1: Comparison of the number of Google searches for "covid" between March and July 2020-2021 vs cases and mortalities of COVID-19.

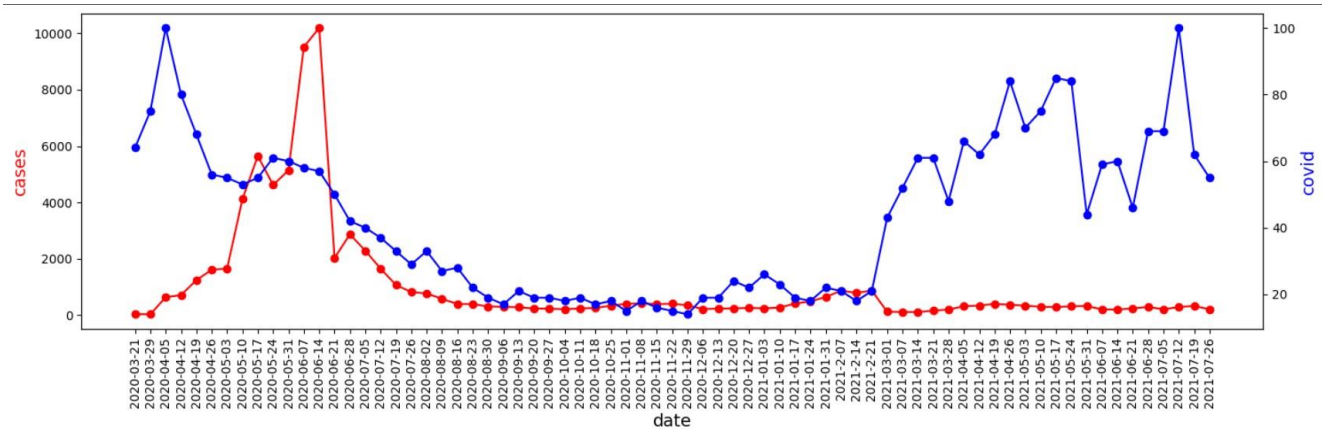


Figure 2: Comparison of the number of Google searches for “fever” between March and July 2020-2021 vs cases and mortalities of COVID-19.

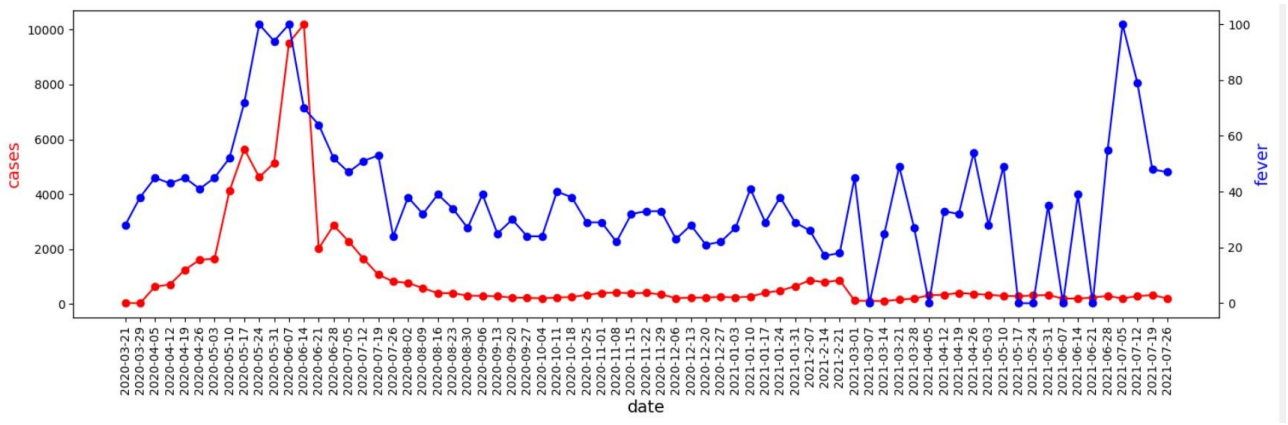
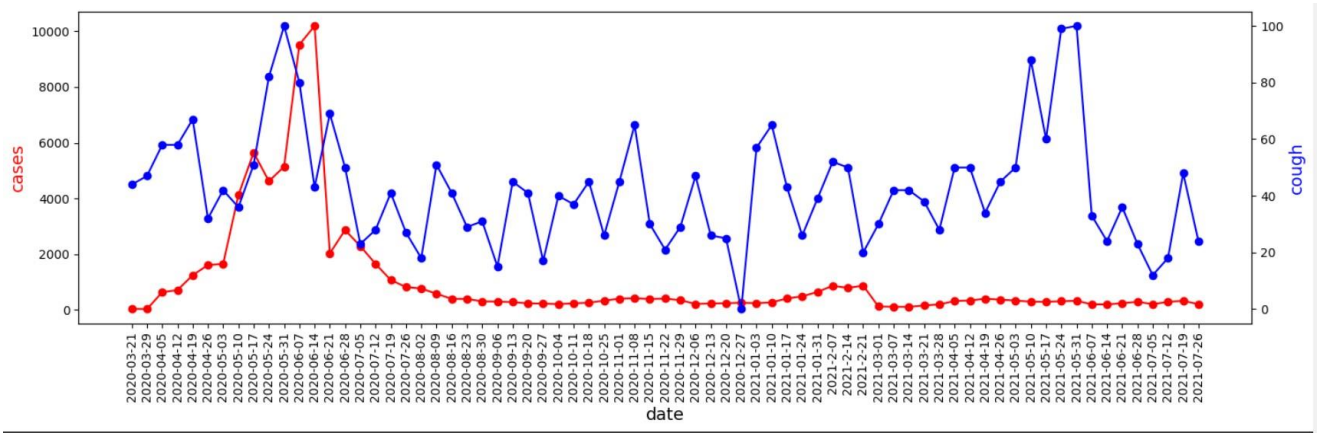


Figure 3: Comparison of the number of Google searches for “cough” between March and July 2020-2021 vs cases and mortalities of COVID-19.



1.2 Goals and Objectives

Our goal is to build an artificial intelligence model for surveillance and analysis of the COVID-19 pandemic in Riyadh by using a Linear regression ML model to classify whether there is an anomaly increase in the number of people with COVID-19 virus by using a real dataset from the King Abdullah Petroleum Studies and Research Center to compare it with the results of our model which will help to alarm the healthcare centers early.

The study will aim to achieve the following objectives:

- Review previous literature on the problem to clearly define the problem's constraints and scope.
- Obtaining real datasets from King Abdullah Petroleum Studies and Research Center.
- Collecting an open source dataset of COVID-19 virus incidence rate from google.
- Pre-processing the datasets.
- Studying and investigating existing state-of-the-art ML classification models to select the most suitable ones to tackle our problem.
- Implementing the proposed model.
- Conducting a set of experiments on the chosen dataset.
- Performing multiple performance measurements to evaluate the performance of the models used in the experiments.

1.3 Proposed Solution

In this research, we aim to take advantage of Google's trend to conduct an early warning to health centers of the COVID-19 outbreak in Riyadh. This will be accomplished by developing a linear regression model, which has been shown to be effective and is easier to implement than other models based on several reasons where Linear regression can be used to predict an outcome variable based on one or more predictor variables. This can be useful in determining the relationship between variables and making predictions about future outcomes. It helps to determine causal relationships between variables. For example, we can use it to determine whether a change in one variable causes a change in another variable. And can compute results quickly [6]. Also, our datasets have a normal distribution as shown below.

Figure 4: Normal distribution of fever in Eastern region.

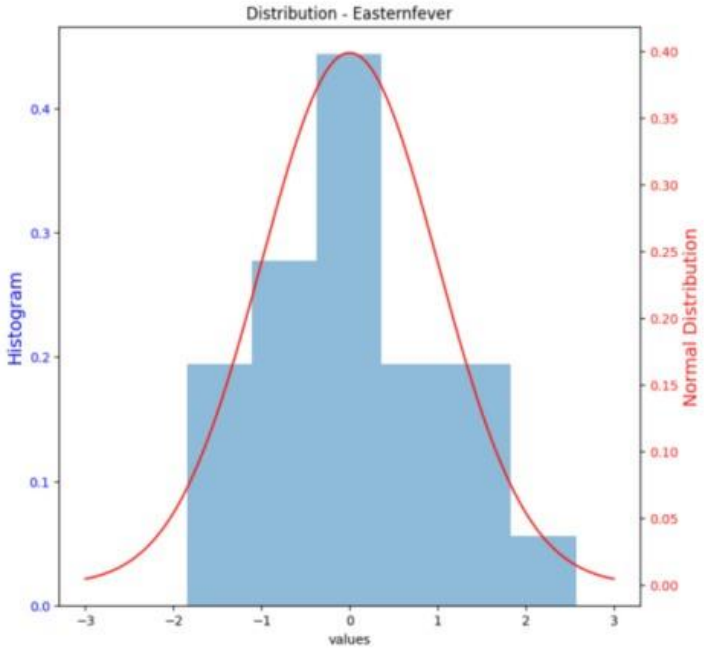
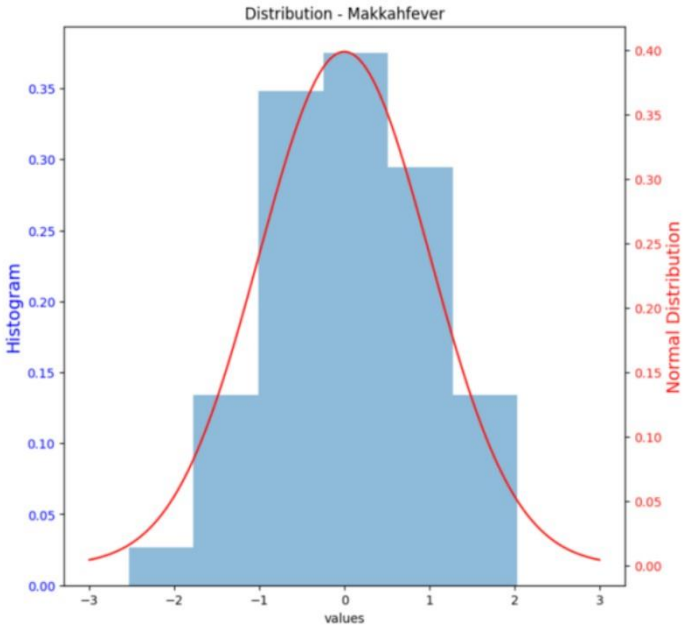


Figure 5: Normal distribution of fever in Makkah.



Normal distribution is a statistical distribution that is symmetrical and bell-shaped, with the majority of the data clustering around the mean. In linear regression, it is assumed that the errors follow a normal distribution. This assumption is important to ensure that the estimates of the regression coefficients are unbiased and the statistical inferences made on them are reliable. Therefore, linear regression assumes that the dependent variable follows a normal distribution [7].

This research aims to develop an easy-to-use model for predicting COVID-19 infection in patients and aiding health centers in accommodating large patient populations.

1.4 Research Scope

The scope of this research centered on COVID-19 due to data availability. And aim to warn the health care centers before the spreading occurs of seasonal diseases in Riyadh region by using a google trend dataset to predict if there are anomaly numbers.
In addition to COVID-19, this study aims to cover other common seasonal ailments such as influenza, asthma, and other related diseases.

1.5 Research Significance

The significance of this research lies in its aim to leverage advanced technology to improve health outcomes and mitigate the impact of disasters [8]. By developing an artificial intelligence model that analyzes Google search data for symptoms of seasonal diseases, this research has the potential to significantly reduce the incidence of health crises and disasters. The model can predict the anomalous increase in cases and warn hospitals, enabling them to proactively prepare for disease outbreaks. This proactive approach can help contain and minimize the spread of disease, thereby leading to improved health outcomes and a reduction in mortality rates. Overall, this research has far-reaching implications for public health and disaster management.

1.6 Ethical and Social Implications

The social and ethical implications of a particular decision or course of action are its effects on society and its members. Some people argue that new technologies always have social and ethical implications and that we should consider those implications before allowing new technology to be used. Others argue that we should not worry about new technologies' social and ethical implications but instead focus on the benefits they bring to society [9]. In this research, there have been no breaches of ethical standards as it pertains to social implications of ethical principles. We are utilizing an open-source dataset that allows for widespread usage and collaboration, with the primary aim of safeguarding patients from any potential risks.

1.7 Report Organization

Eight chapters make up the remainder of this research. The background, including the important and relevant concepts, is demonstrated in Chapter (2). The methods used in earlier studies on the detection of diseases are reviewed in Chapter (3). Our approach to solving the problem is described in Chapter (4). The experimental design can be found in Chapter (5). The implementation and the limitations of our methodology also the problems that accrued during this phase are presented in Chapter (6). Chapter (7) shows the experiment's details and the discussion of the results. The report is concluded in chapter (8).

1.8 Project Details

Table 1: Hardware and Software requirements.

	Requirements	Needed for
Hardware	Laptop or PC	All parts of the project
Software	IDE Google colab	Coding and creating the model
	WhatsApp & Outlook & Teams	Communication
	Google Scholar	Gathering information
	Word document	Documentation

1.9 Project Timeline

Table 2: Rough timeline (1).

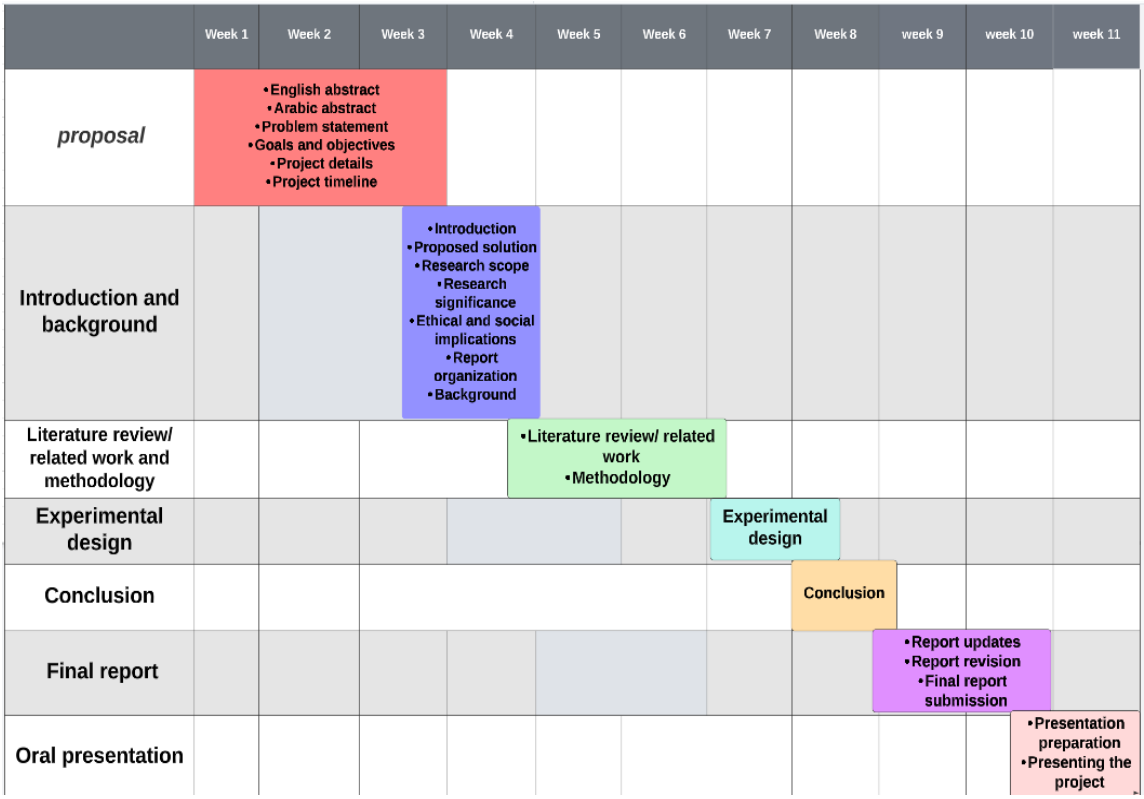
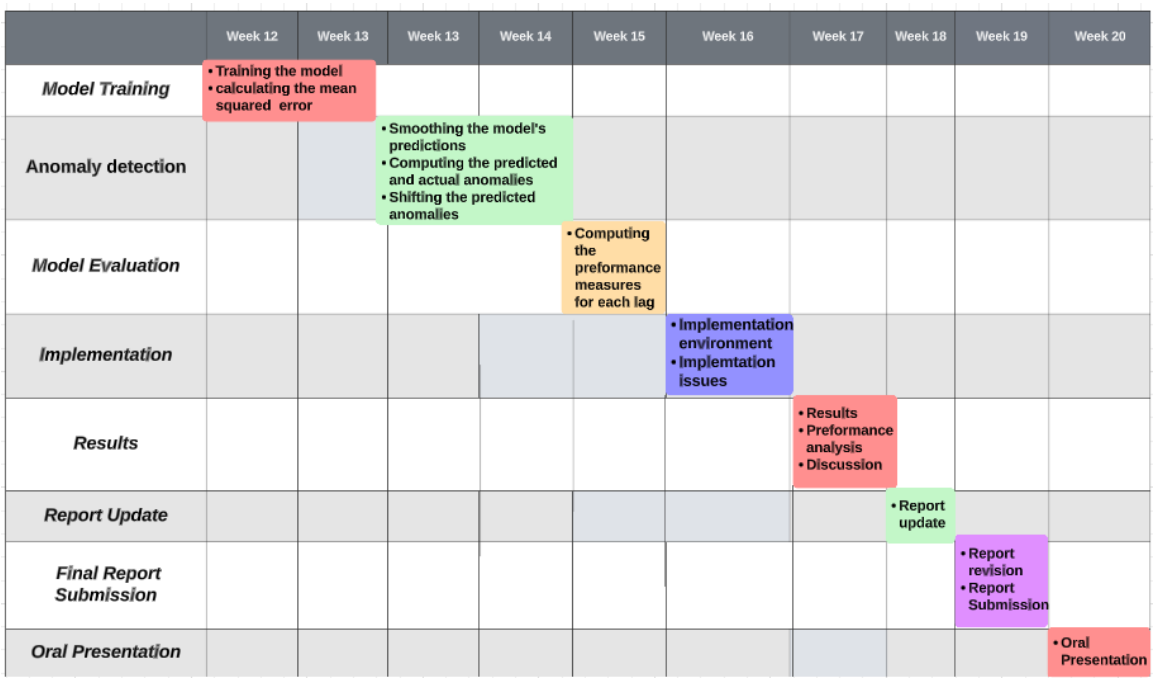


Table 3: Rough timeline (2).



Chapter 2: Background

This chapter discusses the Machine learning (ML) concept and introduces the concepts, terms, algorithms, notations, and tools utilized for this project.

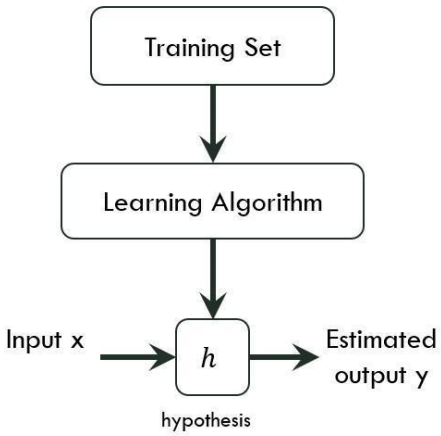
2.1 Machine Learning

Arthur Samuel is credited for coining the term "Machine learning" in 1959. Machine learning is a branch of artificial intelligence (AI) and computer science that uses data and algorithms to simulate human learning and improve accuracy to increase evidence-based decisions in many spheres of society, such as health care [10].

In 1998, Tom Mitchell posed the "Learning" Problem: "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E" [11].

Figure 6 below demonstrates how the machine learning model works:

Figure 6: General representation of machine learning model.



1. Pass the training set into a learning algorithm.
2. The algorithm outputs a function that takes x as input.
3. Outputs the estimated value of y.

ML has different kinds of algorithms to train the model based on learning type, which are:

2.1.1 Supervised Learning

Supervised learning enables machines to classify problems based on related data fed into the machines. Machines are fed with data repetitively until the machines can perform accurate classifications. During supervised learning, a machine is given training data in data mining parlance, based on which the machine does classification. This type of learning algorithm uses outcomes from historical data sets to predict output values for new, incoming data. The algorithm is given labeled training data as inputs and shows the correct answer as outputs.

Supervised learning has two types:

1. Classification: output is discrete.
2. Regression: output is real-valued.

Real-life classifications, such as disease classification, are complex tasks, the machines need appropriate data and several iterations of learning sessions to achieve reasonable abilities [12].

2.1.2 Unsupervised learning

Unlike supervised learning, unsupervised learning cannot be applied directly to regression or classification problems since input data can be found, but output data cannot be generated. Unsupervised learning is aimed at finding the underlying structure of a dataset, grouping it by similarities, and compressing it.

2.1.3 Semi-Supervised Learning

Semi-Supervised learning uses both labeled and unlabeled data for learning [13]. Even though semi-supervised learning is a middle ground between supervised and unsupervised learning, it mainly operates on unlabeled data, typically mixed with a few labels. Since labels are expensive, it may have few labels for corporate purposes.

2.1.4 Reinforcement Learning

Reinforcement learning is a technique that uses feedback to learn how to behave in an environment. It involves performing actions and seeing their results. Positive feedback is given to the agent on each excellent action, whereas negative feedback or penalties are given for each bad action. Reinforcement learning can apply to problems involving sequential decision-making or long-term goals, such as gameplaying and robotics.

2.2 Data Preprocessing

Data preprocessing is the process of analyzing, filtering, transforming, and encoding data in order for it to be suitable and easy to process by a machine learning model [14]. This improves the quality of the model by cleaning up the input data. Data preprocessing is concerned with:

Missing Values: A dataset that is gathered from real life is very limited and might be incomplete, leading to missing data values. Since a machine learning model can't process missing values, one could handle this problem by dropping the samples with missing values, replacing the missing values with zeros, or replacing the missing values with the mean.

Feature Scaling: When dealing with multiple features, these features will likely have completely different ranges, making it difficult for a machine-learning model to compute. Scaling techniques include normalization, where the features are scaled between 0 and 1 or -1 and 1, and standardization, where the new value is calculated by subtracting the mean from the actual value and then dividing it by the standard deviation. The equations for normalization (13) and standardization (14) are as follows:

$$x_{new} = \frac{x - min(x)}{max(x) - min(x)} \qquad (1)$$

$$x_{new} = \frac{x - \mu}{\sigma} \qquad (2)$$

Feature Encoding: Some data cannot be processed by machine learning models, such as names, so we need to encode the data for the machine to understand it. Some classic encoding techniques are: One hot encoding which works by assigning vectors to each category where 1 means that the feature is present and 0 means that it's absent. Target means encoding works by replacing categorical values with the mean1 of the target, and binary encoding works by converting a categorical to its binary representation.

Feature Reduction: Using redundant features will increase the complexity of the algorithm. Therefore, we perform feature reduction to reduce the complexity without losing important information. Feature reduction techniques include principal component analysis, which finds the direction of maximum variance and projects it onto a new subspace with fewer dimensions.

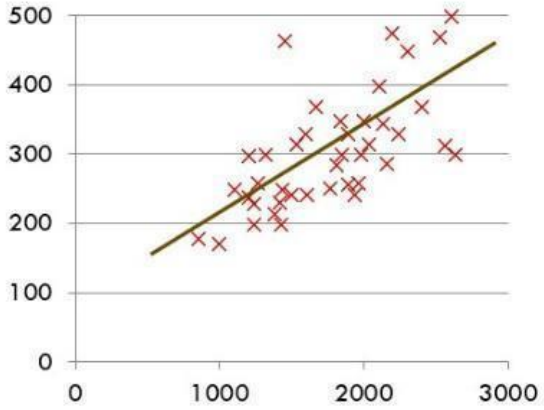
2.3 Linear Regression

Linear regression is one of the most common and comprehensive statistical machine-learning algorithms used to find a linear relationship between one or more predictors [12]. It separates input vectors into classes using linear (hyperplane) decision boundaries to group items by predicting similar feature values into groups [15]. There are two types of linear regression depending on the number of variables which are:

1. Univariate linear regression: linear regression with one variable.

$$h_\Theta(x) = \Theta_0 + \Theta_1 x_1 \qquad (3)$$

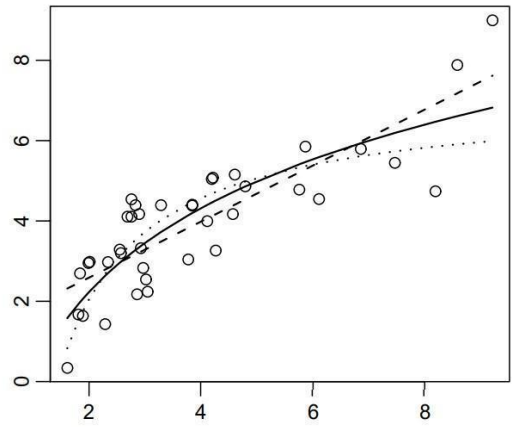
Figure 7: Univariate linear regression.



2. Multivariate linear regression (MLR): linear regression with more than one variable [15].

$$h_\Theta(x) = \Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 + \ldots + \Theta_n x_n \qquad (4)$$

Figure 8: Multivariate linear regression.



2.4 Optimization

Optimization is training a machine learning model in iterations, where each iteration works on improving accuracy and minimizing the error margin to reach a global minimum or maximum [16]. Optimizing machine learning models to achieve the best results is very important. Several optimization algorithms include gradient descent, stochastic gradient, and conjugate gradient.

2.5 Smoothing

Smoothing is a technique used in statistics and data analysis to remove noise from a dataset and create a smoother version of the data. This is often achieved through the use of a mathematical function to create a smooth curve that represents the underlying trend or pattern in the data.

$$F = \alpha A + (1 - \alpha)B \qquad (5)$$

where:

- F is the smoothed value at time t.
- A is the observed value at time t.
- α is the smoothing parameter, which controls the weight given to the current observation versus the smoothed value from the previous time step.
- B is the smoothed value from the previous time step.

In practice, the value of alpha is typically chosen based on the characteristics of the data being analyzed, as well as the desired level of smoothing. A smaller value of alpha will lead to a smoother (more heavily smoothed) estimate, while a larger value of alpha will lead to a more reactive (less smoothed) estimate. The goal of smoothing is to reveal the underlying trend or pattern in the data by removing the extraneous noise or fluctuations. Smoothing can be useful in various applications, including signal processing, image processing, forecasting, and predictive modeling [17].

2.6 Anomaly Detection

Anomaly detection is the process of identifying patterns in a dataset whose behavior differs from what is expected. These unusual behaviors are also known as anomalies or outliers [18].

$$\text{Anomaly detection} = \frac{Daily\ current\ case - previous\ case}{|Average\ of\ cases|} > \sigma \qquad (6)$$

Sigma (σ) is the symbol for standard deviation (7) , which is determined by taking the square root of the variance. It can be used to detect anomalies in normally distributed data. It is useful for detecting outliers, which are most likely to be found at either end of the bell curve. For example, student grades or annual income across a population, which are likely to have a normal distribution and follow a bell curve pattern.

Low standard deviation is associated with data whose values do not vary widely (closer to zero). In addition, a high standard deviation is associated with a data set whose values span a wide range [19].

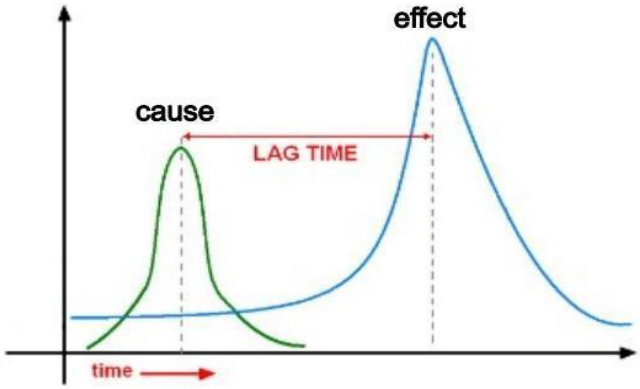
$$\text{Standard déviation} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \underline{x})^2}{n-1}} \qquad (7)$$

In this case, x_i represents the value of the i^{th} point in the dataset, and $\underline{x}$ is the mean value of the dataset. n represents the number of data points in the dataset [20].

2.7 Time Lag

Time lag is a scientific term utilized to describe the temporal delay between a cause, which refers to the act of searching for a symptom, and an effect, which pertains to the number of Covid-19 cases. This phenomenon is relevant in our project as it provides an understanding of the amount of time that elapses between the cause and its subsequent effect on the outcome variable. This analytical tool is useful in providing meaningful insights into the associations and patterns that underlie the variables. Time lag analysis is a widely adopted approach in various fields such as economics, finance, and cases of diseases to discern correlations and forecast future outcomes.

Figure 9: Time lag diagram.



2.8 Performance Measurements

Performance measurements are used to evaluate the quality of machine learning models. They are used to measure how well a model generalizes on new data. Some performance measures include:

Mean Absolute Error (MAE): Measures the average of the absolute differences between actual values and predicted values. The equation is as follows:

$$MAE = \frac{\Sigma \mid x_i - x|}{N} \quad (8)$$

Mean Squared Error (MSE): Measures the average of the squared differences between actual values and predicted values. The equation is as follows:

$$MSE = \frac{\Sigma (x_i - x)^2}{N} \quad (9)$$

R-squared: Measures the proportion of variance of the dependent variable by the independent variable [21]. The equation is as follows:

$$R - squared = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - y)^2} \quad (10)$$

True positive (TP): The number of correct anomalies predicted.

False positive (FP): The number of incorrect anomalies predicted.

True negative (TN): The number of correct non anomalies predicted.

False negative (FN): The number of incorrect non anomalies predicted.

Accuracy: overall how often are the predictions correct. The equation is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

Precision: how often the model was correct when predicting yes. The equation is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

Recall: how many actual positives the model predicted. The equation is as follows:

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

F1-score: weighted average of precision and recall. The equation is as follows:

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \quad (14)$$

Chapter 3: Literature Review/Related Work

There are an increasing number of studies conducted in our research domain that will be discussed in this chapter.

In Harvey et al.'s study[22], machine learning classifiers were used to construct predictive models for asthma based on data collected from the Centers for Disease Control and Prevention website. The variables of sex, difficulty breathing, allergies, and medication were found to have the highest correlation with asthma, with logistic regression showing the strongest connection at 99.99% accuracy. The Naive Bayes model had an accuracy of 82.7%, while KNN had a specificity of 98.280%, recall of 41.234% and accuracy of 90.28%. The random forest classifier achieved 98.9% accuracy. Overall, the logistic regression classifier produced the highest prediction accuracy 99.99%.

Similarly, in Alkouz et al.'s study[23], a deep learning model called Deepluenza was introduced for detecting influenza-related tweets in social media to provide early insight into influenza outbreaks. Deepluenza, based on the BERT base multilingual model, achieved 99% accuracy and an F1-score of 98% for influenza reporting. Their results showed that Deepluenza is highly accurate in identifying influenza reports in social media compared to real-life data collected by health authorities.

Alex J Ocampo et al.'s study[24] suggests the feasibility of using search queries to predict disease outbreaks by using malaria-related Google searches to predict existing malaria surveillance trends of Thailand. They used four models to fit monthly official case counts and found that the stepwise model, where search query terms were health-related but not specific to malaria, was the best model with an overall accuracy of 93%. The study indicates the potential of search query-based prediction models for disease outbreaks.

Alharbi et al.'s study[25] evaluated the relationship between asthma attacks and five weather variables: high temperature, low temperature, precipitation level, humidity, and thunderstorms. They found that these variables are effective triggers of asthma attacks and built a linear regression model to evaluate their impact, achieving an accuracy of 87.13%. The study highlights the significance of considering weather variables in asthma management. Overall, these studies demonstrate the potential of machine learning, deep learning, and search query-based prediction models for disease and outbreak management.

Numerous studies have been conducted to develop machine learning models for the early detection and accurate diagnosis of deadly viruses, including COVID-19 and Dengue fever. Aftab et al. [26] (2022) proposed using tailored deep learning models for classifying COVID-19 and Influenza, using a dataset of publicly available x-ray images. The evaluation phase showed that the proposed LSTM model outperformed the CNN model

with 98% accuracy, indicating the effectiveness of deep learning techniques for virus detection.

Park et al.[27] (2021) developed a new ML model for disease prediction through the classification of diseases using laboratory test results. They created an optimized ensemble model by combining a DNN model with two ML models, resulting in an f1-score of 81%, 92% prediction accuracy, 78% precision, and 88% recall. They also used a confusion matrix and the SHAP value method to analyze feature importance, resulting in an accurate disease classification pattern.

Alexandro et al. [28] (2021) proposed an accurate classification model based on DL and CNN for the characterization and classification of Influenza type A based on the protein sequence's ability to infect a specific host. Their proposed model achieved an overall accuracy of 99% and an f-score of 98%, demonstrating the effectiveness of their proposed model for virus classification.

Bilal Abdualgalil et al.[29] (2022) used machine learning models to analyze clinical data from patients with Dengue fever. They used five models and evaluated each model using the same performance measures, with the ETC model achieving the highest accuracy, f1-score, precision, recall, and AUC measures. Their study demonstrates how machine learning models can accurately diagnose and detect deadly viruses such as Dengue fever, providing an early indication of the disease and facilitating prompt treatment.

Satyabrata, Aicha et al. [30] (2019), this study discussed the use of Twitter data to predict malaria epidemic outbreaks. The researchers manually labeled 205 malaria-related tweets and combined them with 2295 other malaria-related tweets and 2500 unrelated tweets to create a training set. They built several models, including random forest and SVM, and found that the SVM model performed the best, with an accuracy of 96% and an F-score of 97%.

Finaly, Yousef et al. [31] (2020), this study proposed a new heart disease prediction model based on the Naive Bayes algorithm and several machine learning techniques such as SVM, KNN, decision tree, and RF. The proposed approach employs the Naive Bayes technique to select the best subset of features for the next classification phase and handle the high dimensionality problem by avoiding unnecessary features and selecting only the important ones. In this study, they used and compared several classification algorithms (DT, SVM, RF, and KNN) and found that the combination of the Naive Bayes feature selection approach and the SVM-RBF classifier had the highest accuracy of 98% in predicting heart disease.

Table 4: Summary of the literature review.

Ref	Year	Dataset	Model(s)	Accuracy	Precision	Recall	F1-score
[22]	2019	Centers for Disease Control & Prevention website	KNN	0.90	-	0.41	-
			Random Forest	0.98			
			Naive Bayes	0.82			
			Linear Regression	0.99	-	-	-
[23]	2022	Twitter	BERT base Multilingual Model	0.99	-	-	0.98
[24]	2013	Search query terms from physicians	Physician model Linear regression	0.6	-	-	-
[24]	2013	Health related search Query terms	Stepwise Model Linear regression	0.93	-	-	-
[25]	2019	Backdated 2010 - 201	Linear Regression	0.87	-	-	-
[26]	2022	X-ray images	LSTM	0.98	-	-	-
[27]	2021	Laboratory tests	Ensemble Model	0.92	0.78	0.88	0.81

		optimized					
[28]	2021	IRD	CNN	0.99	-	-	0.98
[29]	2022	EMC	ETC	0.99	0.989	0.991	0.99
[30]	2019	Twitter	SVM	0.96	0.98	0.96	0.97
			XGB	0.87	0.87	0.95	0.91
[31]	2020	Cleveland Heart Disease database	NB-SKDR	0.98	-	-	-

As shown in Table 4, the linear regression showed high results of accuracy and effectiveness, and since it is easy to implement compared to other models and shows good results, after we ensured from the normal distribution of the data we decided to use linear regression to develop the model.

All studies discussed in this chapter demonstrate the potential of machine learning in predicting and preventing diseases. And showed that social media data can be utilized for disease surveillance, and proposed a new approach to feature selection and classification for disease prediction. And highlight the importance of developing accurate and efficient prediction models to improve healthcare outcomes.

Chapter 4: Methodology

This chapter describes the research workflow, which comprises four major phases: research's data collection, pre-processing, model training, and model evaluation. Also, this chapter explains how the proposed method provides a reliable answer by testing the hypothesis.

4.1 Data collection

In this research, we used two datasets, one to predict the results and another to evaluate and compare these results.

The first dataset is an open-source Google Trends dataset that collects the number of search times for COVID-19 virus symptoms in the Google search engine over a specific period in different regions of Saudi Arabia. Google Trends is an online service provided by Google that analyzes search inquiry popularity in various regions and different languages. Google Trends interface provides time-based graphs of the search results. It includes several analytical features, including comparing multiple search queries, tracking various words and phrases typed into Google's search engine, categorizing, and organizing data, and breaking down information by location [32].

- Link to the google trend dataset source:
https://trends.google.com/trends/explore?q=covid&geo=SA

The second dataset represents the real cases from King Abdullah Petroleum Studies and Research Center (KAPSARC). It was used to evaluate and compare the results of the proposed model.

- Link to the King Abdullah Petroleum Studies and Research Center dataset source: https://datasource.kapsarc.org/pages/publications_datasets/

4.2 Preprocessing

During this phase of our study, we aimed to standardize the training and test data by applying the z-score function, which would enable us to normalize the features in our dataset and ensure they were on the same scale. However, upon conducting a thorough analysis, we discovered that the dataset we had collected from Google Trend was already normalized, and the King Abdullah Petroleum Studies and Research Center data had already been standardized. As such, we determined that further normalization or standardization was not necessary for our model, and we proceeded to use the data in its current form.

4.3 Model Training

4.3.1 Linear Regression

The linear regression model was used as a classifier in that phase. In statistics and machine learning, linear regression is the best-known and most well-understood algorithm. It is the most widely used for predictive modeling due to its emphasis on minimizing error or making as accurate predictions as possible at the expense of explainability [33].
In regression modeling, the independent characteristics are based on a target class. According to the King Abdullah Petroleum Studies and Research Center, linear regression performs a task based on a given independent variable (x), which represents COVID-19 symptoms [33] :

1. Covid.
2. Fever.
3. Cough.

To estimate a value for a dependent variable (h) representing predicted COVID-19 infections. A linear relationship between x (input) and h (output) is observed as a result of this method. The formula (4) illustrates in chapter (2) how h is related to x. [33]:

$$h_{\Theta}(x) = \Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 \qquad (15)$$

In this case, h_{Θ} represents the predicted output of COVID symptoms for Riyadh. Furthermore, the input variables x_1 and x_2 are employed as predictors of COVID-19 symptoms for the 12 regions other than Riyadh. Θ_0 represents the intercept and $\Theta_1$represents the slope of the regression line [34].

Throughout the training regimen, we executed the linear regression formula (15) to analyze data from 12 different regions, focusing on the relationship between COVID-19 symptoms and COVID-19 predicted cases. The analysis is conducted by looping through 6 iterations. During each iteration, we compute the r-square using the formula (10) of every region and arrange them in ascending order. Following this, utilize formula (9) to select two regions that display the smallest values of MSE. These two regions are subsequently employed in linear regression and are eliminated from future iterations. The main objective is to examine the MSE associated with COVID-19 symptoms and determine its significance in the dataset.

4.3.2 Smoothing

After obtaining the prediction result of the proposed model, we smoothed the training predicted result. The purpose of smoothing time series data is to reduce noise and better visualize the underlying trend in the data. By smoothing the data, it is easier to identify patterns and observe trends over time.

The smoothing formula (5), used is the rolling mean with a given window size. The window size in the smoothing mathematical function represents the number of adjacent data points that are used to calculate the smoothed value at a particular point. depende on the amount of noise in the data and the desired level of smoothing. Generally, larger window sizes provide more smoothing but can also lead to loss of important detail or resolution in the data. Conversely, smaller windows provide less smoothing and better detail but can also be more affected by noise. It is implemented using the Pandas library's rolling() function, which takes the window size as an argument and applies the rolling mean to the given data [35].

4.3.3 Anomaly Detection

To identify anomalies, we applied formula (6) and labeled each day with an anomaly as 1, and otherwise with a 0. We calculated the predicted anomalies by using the standard deviation of the differences between the model's predictions and multiplying the resulting values by factors ranging from 1 to 2 in increments of 0.05. To ensure the accuracy of our predictions, we first smoothed the model's predictions based on formula (5). Additionally, we calculated the actual anomalies using formula (6) and the standard deviation of the differences between actual cases.

4.3.4 Lag

In this study, we performed a time-shift analysis on the predicted anomalies over a period of 6 weeks, with a one-week lag shifting at each iteration. Following this, we computed the performance measures of each lag to determine the optimal time frame to be used in the testing phase.

4.4 Model Testing

4.4.1 Linear Regression

Based on the ultimate iteration during the training process, the ultimate pair of territories denoted as x1 and x2 were utilized to achieve the most optimal line of best fit during the testing phase, whereby our model was evaluated using the corresponding testing data of these cities.

4.4.2 Smoothing

In some cases, smoothing can be useful for reducing the impact of random variability and outliers in the data, leading to more reliable and accurate predictions. However, in certain research projects, it may not be necessary to use smoothing techniques for testing predictions. In this research, the data is already relatively smooth and free of significant noise, and adding extra smoothing actually introduces bias and compromises the accuracy of the predictions.

4.4.3 Anomaly Detection

A variety of symptoms were tested, including cough, covid, and fever. We shifted the predicted anomalies by one week 6 times and calculated the predicted anomalies using the standard deviation, multiplied by 1.

In the present research, the decision to not smooth the testing predictions was made based on the characteristics of the dataset. After careful analysis, we found that the testing results without smoothing demonstrated higher accuracy levels and aligned better with the research question. Therefore, the decision was made to conduct further analysis without smoothing in order to produce more reliable and useful predictive models.

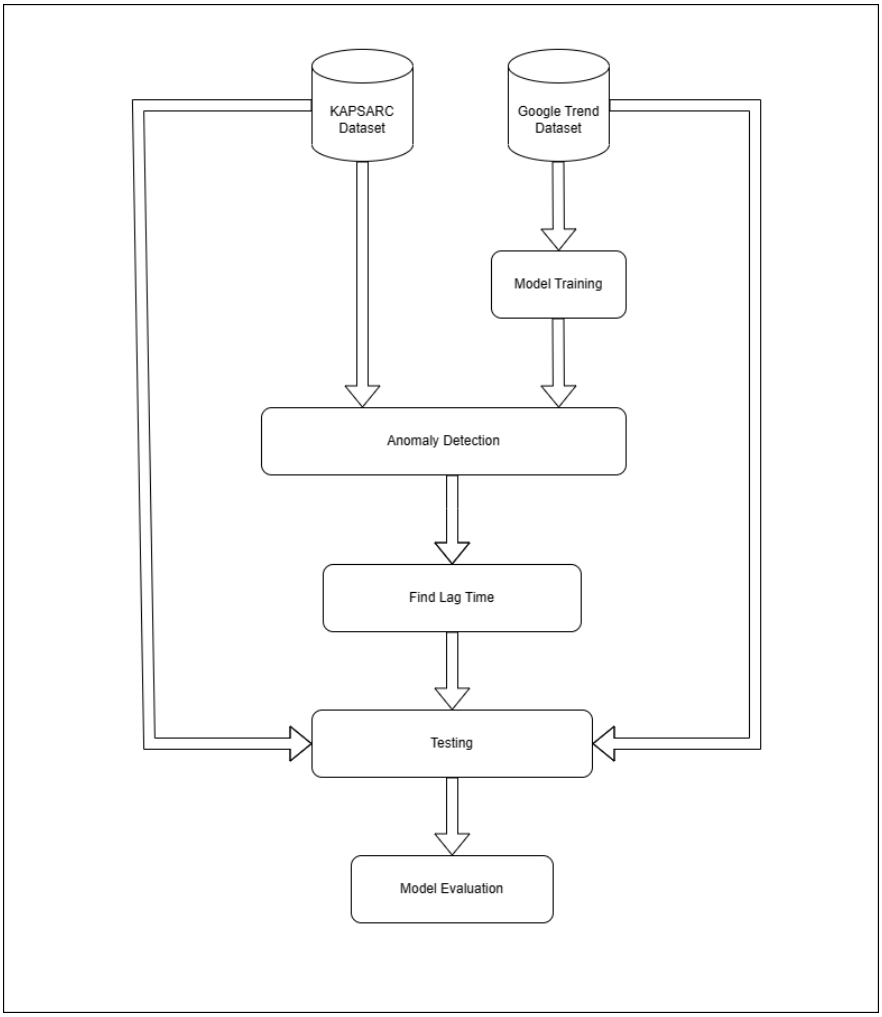
4.4.4 Lag

Based on the identified optimal lag of the training phase, we determined the optimal time frame for the testing phase to be inclusive of all lags up to and including the optimal training lag. This approach was motivated in part by the significant variation observed in the size and characteristics of different waves across the dataset. By incorporating multiple lags in the testing phase, we aimed to increase the robustness and generalizability of our early warning system, and to account for potential variations in the dynamics of the disease over time.

4.5 Model Evaluation

In this phase, we used anomaly detection results from the training and testing to obtain the TP, TN, FP, and FN values for both predicted results and the real dataset to evaluate the proposed model using the performance measurements: accuracy (11), precision (12), recall (13) and f-score (14). This was done to ensure the reliability of the proposed model. [38]. Figure 10 below explain the flow of methodology phases:

Figure 10: Graphical representation of the methodology.



Chapter 5: Experimental Design

5.1 Dataset

As mentioned in the Methodology section, we used Google Trends and the King Abdullah Petroleum Studies and Research Center as the dataset sources for this research from 2020-03-21 to 2021-07-26.

Sample from Google trend for the "covid" worksheet, as shown in Figure 11, was used as a line chart graph to predict the results.

Figure 11: Google trend for "covid" worksheet.

A	B	C
Date	covid	
4/1/20	68	
4/8/20	83	
4/15/20	60	
4/22/20	50	
4/29/20	50	
5/6/20	42	
5/13/20	35	
5/20/20	46	
5/27/20	63	
6/3/20	43	
6/10/20	59	
6/17/20	46	
6/24/20	43	
7/1/20	39	
7/8/20	38	
7/15/20	28	
7/22/20	30	
7/29/20	22	
8/5/20	29	
8/12/20	26	
8/19/20	25	
8/26/20	32	
9/2/20	18	
9/9/20	15	
9/16/20	16	
9/23/20	22	
9/30/20	14	

Sample from King Abdullah Petroleum Studies and Research Center worksheet as given in Figure 12 was used as a line chart graph to evaluate and compare these results.

Figure 12: King Abdullah Petroleum Studies and Research Center worksheet.

D	E	F
Date	Cases (person)	
2020-04-01	7	
2020-04-08	83	
2020-04-15	84	
2020-04-22	157	
2020-04-29	440	
2020-05-06	194	
2020-05-13	478	
2020-05-20	714	
2020-05-27	611	
2020-06-03	675	
2020-06-10	1431	
2020-06-17	1442	
2020-06-24	241	
2020-07-01	397	
2020-07-08	364	
2020-07-15	208	
2020-07-22	143	
2020-07-29	114	
2020-08-05	93	
2020-08-12	86	
2020-08-19	59	
2020-08-26	56	
2020-09-02	37	
2020-09-09	38	
2020-09-16	39	
2020-09-23	36	
2020-09-30	26	

The dataset has been partitioned into two distinct subsets for the purpose of machine learning model development and evaluation. Specifically, the dataset has been divided into a training set and a testing set. The training set constitutes the initial 80% of the dataset, and it has been utilized to train the machine learning model. The remaining 20% of the dataset has been allocated to the testing set, which has been employed to evaluate the performance of the trained model.

5.2 Experiments

In that step, we experimented with the linear regression model mentioned in the methodology, using data from two sources: Google Trends and the King Abdullah Petroleum Studies and Research Center.

5.2.1 Training

The linear regression model was trained according to formula (15) as shown in chapter (4) to obtain the best fit for the data, so that it could make accurate predictions for future data. The training data is Google Trend search from 2020-03-21 to 2021-04-12 for these symptoms:

1. Cough.
2. Covid.
3. Fever.

We chose this period because it captures the time period when internet searches for COVID-19 symptoms spiked, and official case reports began to emerge. Moreover, the data exhibits an upward trend, characterized by a curvilinear trajectory.

As outlined in the Methodology section, we performed a time-shift analysis on the predicted anomalies of our proposed model for 6 weeks, using standard deviation of differences between model predictions and multiplied by factors varying from 1 to 2. We ensured prediction accuracy by smoothing model predictions and calculating actual anomalies, along with standard deviation of differences between actual cases.

5.2.2 Testing

The testing data is Google Trend search from 2021-04-19 to 2021-07-28 for these symptoms:

1. Cough.
2. Covid.
3. Fever.

The symptoms of cough, Covid, and fever were subjected to testing, and standard deviation multiplied by 1 was utilized to predict anomalies. The optimal testing phase was determined to encompass all lags up to and including the optimal training lag, taking into account the variation observed in the magnitude and features of different waves across the dataset.

Chapter 6: Implementation

In the following chapter, we provide a comprehensive overview of the implementation environment utilized in this project. Our discussion includes a detailed description of the tools employed for data preparation, the framework utilized, and the programming platform utilized. Additionally, we provide a comprehensive analysis of the challenges and issues faced during the implementation process.

6.1 Implementation Environment

The proposed system is implemented via Google Collaboratory, a programming platform that enables the development and execution of Python code within a browser-based environment. This platform provides an efficient and user-friendly experience for performing various tasks, including deep learning and data analysis.

The following libraries used in order to implement the models:

- NumPy

Large multidimensional arrays are now supported in the python programming language with NumPy which is a highly optimized and free open-source library. In addition to these arrays, NumPy also provides a collection of high-level mathematical functions. They include basic linear algebra, random simulation, Fourier analyses, trigonometric operations, and statistical operations [36].

- Sklearn

In Python, Scikit-learn (Sklearn) is the most useful and robust machine learning library. Python-based consistency interface provides machine learning and statistical modeling tools including classification, regression, clustering and dimensionality reduction [37].

- Matplotlib

Matplotlib is an amazing Python library for 2D array plotting. Multi-platform data visualization library Matplotlib is built on NumPy arrays and designed to work with broader SciPy stacks. Using visualization allows us to easily digest large amounts of data in digestible visuals, which is one of its greatest advantages. Matplotlib includes several plot types, such as lines, bars, scatters, histograms, etc [38].

- Pandas

Pandas is an open-source library in Python used for data manipulation and analysis. It provides easy-to-use data structures and data analysis tools that allow you to work with data efficiently. The main data structures in Pandas are the Series and DataFrame. A Series is a one-dimensional array-like object that can hold any data type such as integers, strings,

etc. A DataFrame is a two-dimensional labeled data structure consisting of rows and columns. It is similar to a spreadsheet or SQL table and is the most commonly used data structure in Pandas. One of the significant advantages of Pandas is the ability to handle large data sets. Pandas is built on top of NumPy, which is an efficient numerical computing library. It means that Pandas can handle large data sets faster and efficiently compared to traditional programming methods [39].

- klearn.linear_model library

The sklearn.linear_model library in Python provides a range of linear models for regression, classification, and other tasks. It is a part of the popular Scikit-learn library, which is widely used in machine learning and data analysis. The library includes several types of linear models [40].

- sklearn.metrics

The sklearn.metrics module includes the following key functions [41]:

- Confusion_matrix: computes the confusion matrix for a classification problem.
- Accuracy_score: computes the accuracy of a classification problem.
- Precision_score: computes the precision score for a classification problem.
- Recall_score: computes the recall score for a classification problem.
- F1_score: computes the F1-score for a classification problem.
- Classification_report: generates a report that includes precision, recall, f-score, and support for each class in a classification problem.

- Matplotlib.pyplot

Matplotlib.pyplot is a library in Python that is primarily used for data visualization. It provides a convenient interface for creating a variety of charts including line plots, scatter plots, bar plots, histograms, and more. Matplotlib.pyplot allows users to customize the color, marker style, and labels of charts to help visualize data in a meaningful way [42].

6.1 Implementation Issues

We found some problems in Understanding how this machine-learning model can make its predictions. It took us some time to understand and analyze it. Also,It's been a challenge to troubleshoot and fix the execution errors in our code, which has been causing frustration and delaying our project timeline. Consequently, to solve this problem, Google Collaboratory is used in order to provide additional services and computing resources including Scikit-learn for regression.

Chapter 7: Results and Discussion

The objective of this chapter is to present and analyze the findings of our research in detail. The analysis provided in this chapter will assist us in gaining a better understanding of the research problem under study and offering insights into possible solutions or avenues for further investigation.

7.1 Results

7.1.1 Training

The results were acquired by implementing the linear regression formula (15), as depicted in the figures that illustrate six iterations of a linear graph displaying two distinct regions. Upon analyzing the presented graphs for "covid", we can infer that the initial iteration contains the two regions, Al Madinah and Al Jouf regions with Riyadh having an MSE value of 15 and depicting poor results. Conversely, the final iteration showcases the two regions, Makkah and Eastern regions with Riyadh having an MSE value of 6 signifying the optimal results and providing us with the best fit line. From the aforementioned, we can conclude that in every loop where the r-square is increasing, the MSE decreases, resulting in the highest precision in our measurements.

Figure 13: shows a scatter plot comparing the Al Madinah and Al Jouf regions with Riyadh.

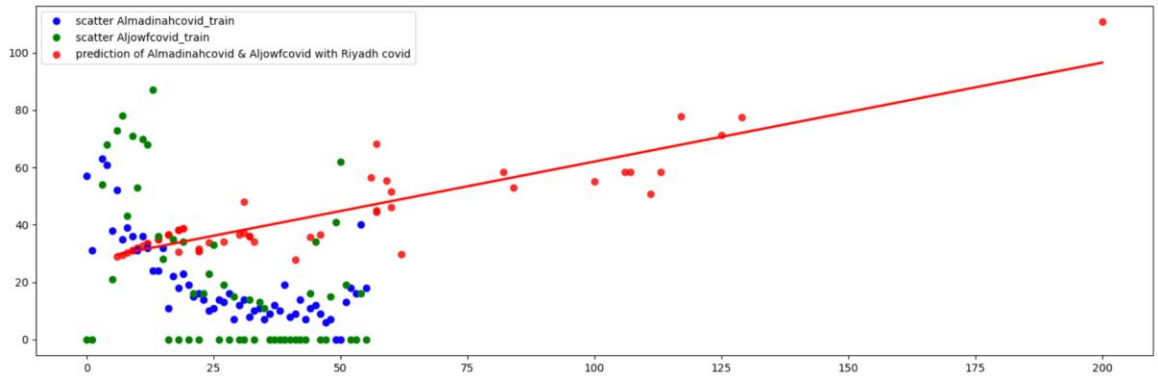


Figure 14: shows a scatter plot comparing the Hail and Jazan regions with Riyadh.

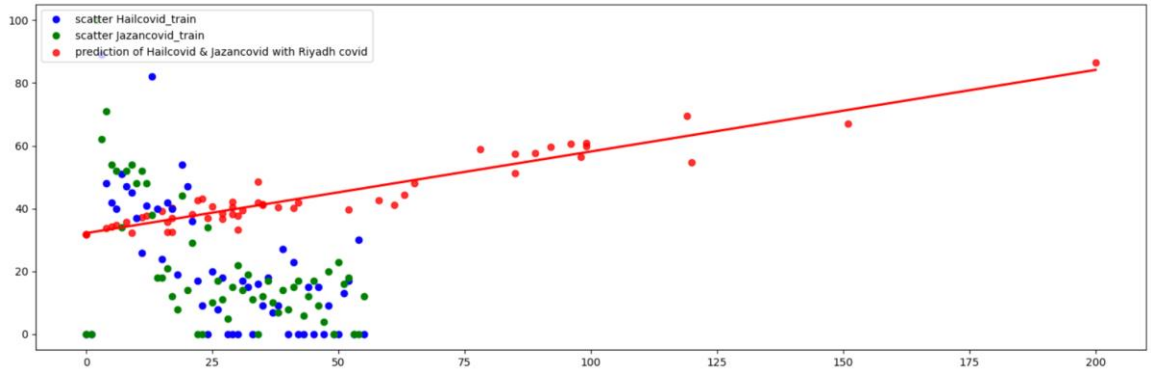


Figure 15: shows a scatter plot comparing the Aseer and Albahah regions with Riyadh.

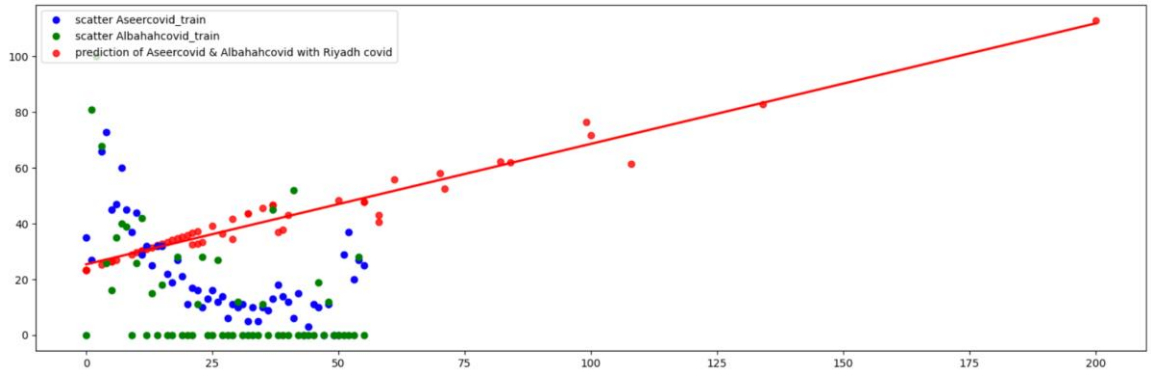


Figure 16: shows a scatter plot comparing the Tabuk and Northern Borders regions with Riyadh.

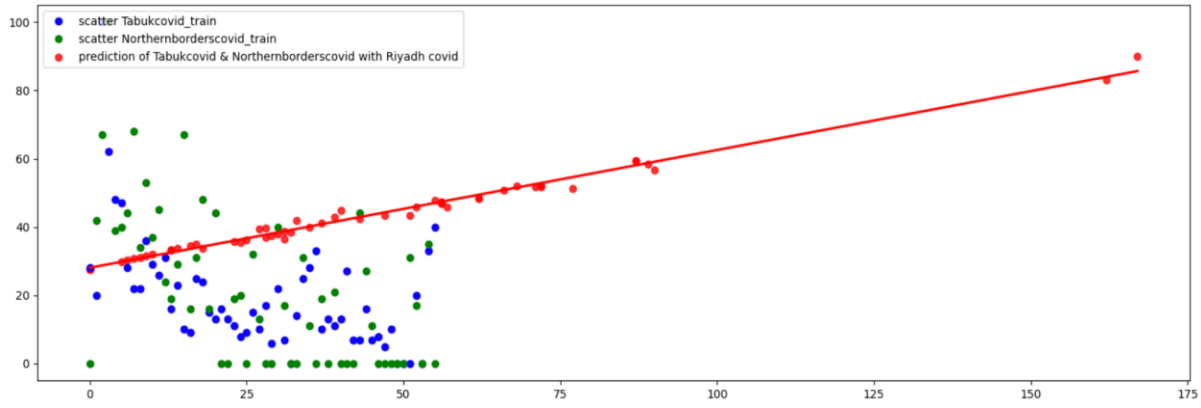


Figure 17: shows a scatter plot comparing the Najran and Al Qassim regions with Riyadh.

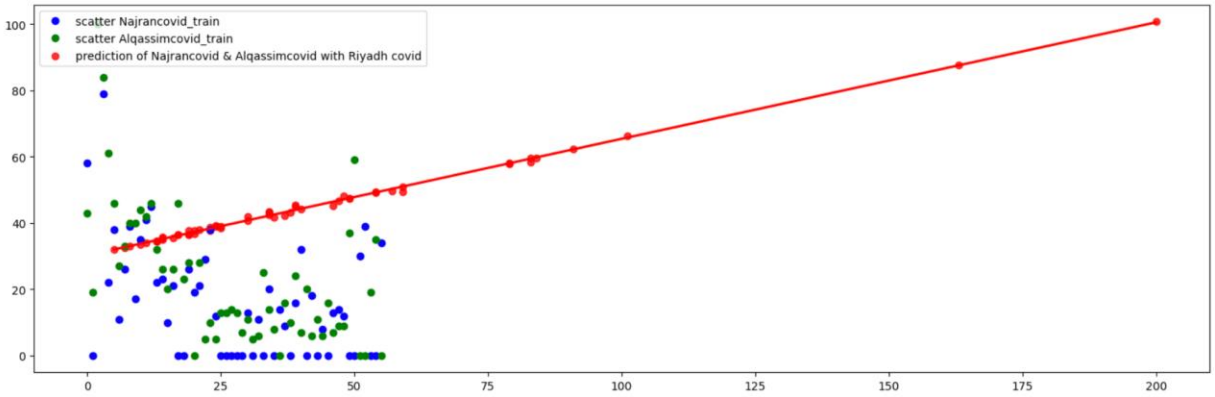
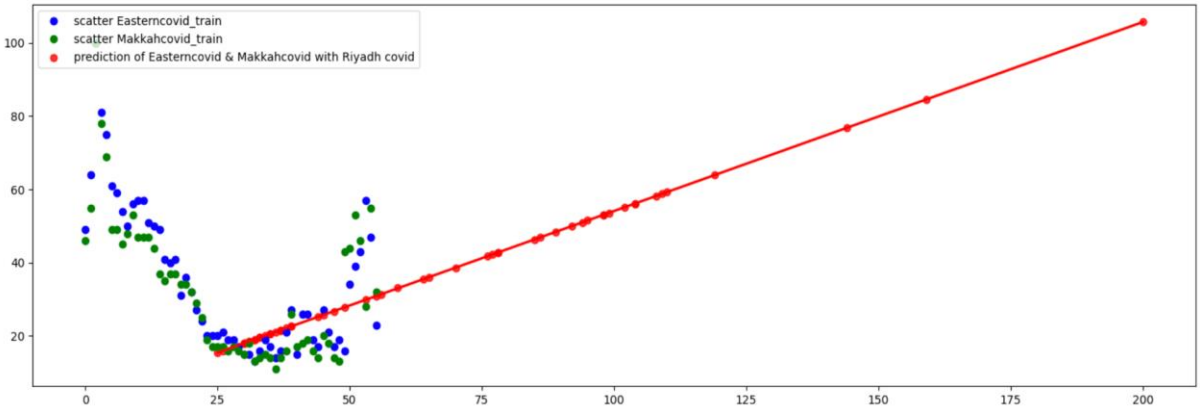


Figure 18: shows a scatter plot comparing Makkah and the Eastern region with Riyadh.



Based on the obtained results, as shown in Figure 18, it was determined that the search term with the lowest MSE is “covid”. The regions considered in the final iteration include Makkah and the Eastern regions. This value will be employed for lag calculation in subsequent predictions.

The graphs in figures 19, 20, and 21, depict the MSE for three distinct symptoms. The figure with the highest accuracy level is figure 21 which is for the “covid” search term.

Figure 19: Plot the MSE for “fever”.

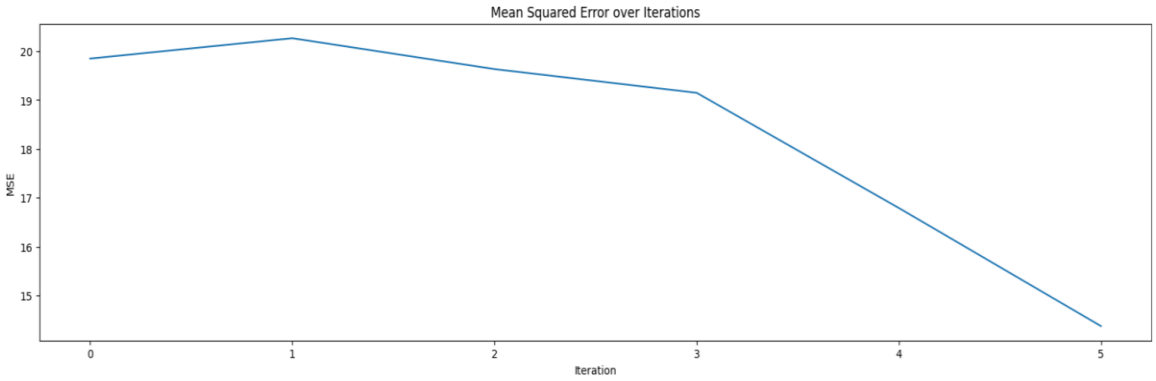


Figure 20: Plot the MSE for “cough”.

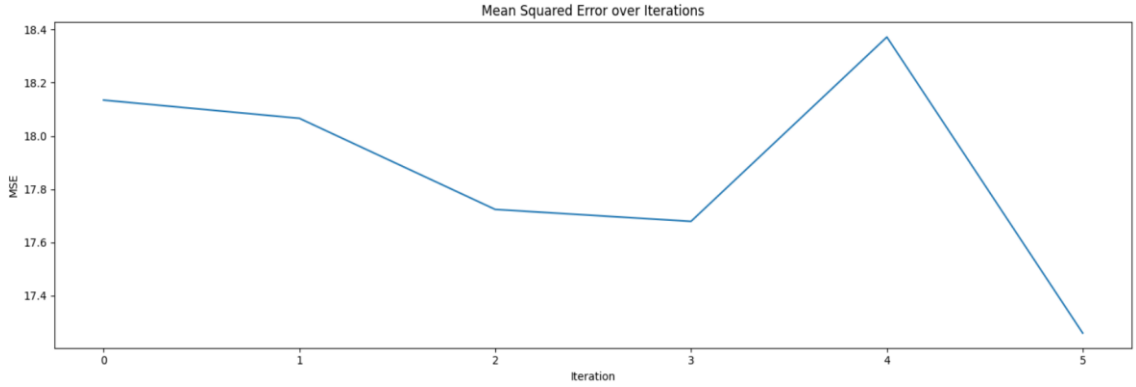


Figure 21: Plot the MSE for “covid”.

Figures 22 and 23 exhibit the predicted outcome before and after smoothing. It is evident that the application of smoothing techniques provides a clearer visualization of the impact prediction. These results highlight the significance of implementing appropriate techniques to enhance the precision and dependability of predictive models.

Figure 22: prediction result of "covid" before smoothing.

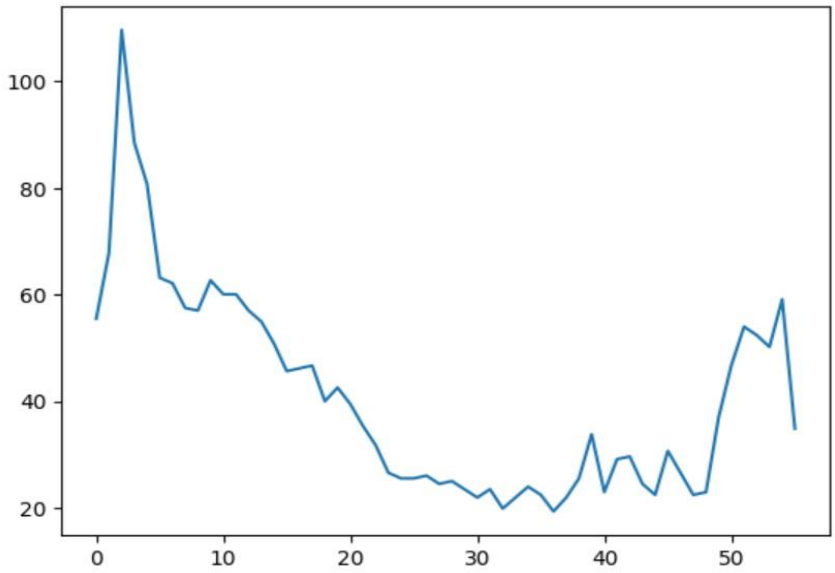
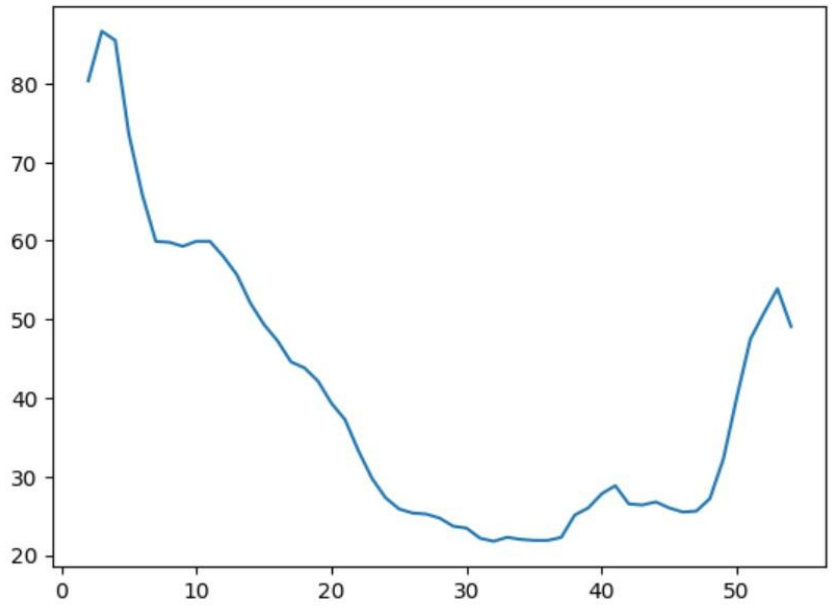


Figure 23: prediction result of "covid" after smoothing.



Following the training of our model, we conducted a time-shift analysis on the predicted anomalies for a total of 6 lags, shifting one week at a time. This process yielded two anomaly signals for cases and one for predictions. After computing the performance measures for each iteration, we determined that applying a standard deviation multiplier between 1 and 2 produced consistent results. Specifically, lag 4, as depicted in figures 24, 25, and 26, produced the best outcomes for all symptoms, with “cough”, “covid”, and “fever” demonstrating confusion matrix values as presented in tables 5,6,7 . Notably, these symptoms exhibited an average precision of 100%, recall of 50%, and f1-score of 67%.

Table 5: Results for lag 0-6 for “cough”.

lag	TP	TN	FP	FN	Precision	Recall	F1-Score
0	0	36	1	2	0	0	0
1	0	37	1	2	0	0	0
2	0	38	1	2	0	0	0
3	0	39	1	2	0	0	0
4	1	41	0	1	1.00	0.50	0.67
5	0	41	1	2	0	0	0
6	0	42	1	2	0	0	0

Table 6: Results for lag 0-6 for “covid”.

lag	TP	TN	FP	FN	Precision	Recall	F1-Score
0	0	36	1	2	0	0	0
1	0	37	1	2	0	0	0
2	0	38	1	2	0	0	0
3	0	39	1	2	0	0	0
4	1	41	0	1	1.00	0.50	0.67
5	0	41	1	2	0	0	0
6	0	42	1	2	0	0	0

Table 7: Results for lag 0-6 for “fever”.

lag	TP	TN	FP	FN	Precision	Recall	F1-Score
0	0	36	1	2	0	0	0
1	0	37	1	2	0	0	0
2	0	38	1	2	0	0	0
3	0	39	1	2	0	0	0
4	1	41	0	1	1.00	0.50	0.67
5	0	41	1	2	0	0	0
6	0	42	1	2	0	0	0

Figure 24: Plot of the predicted anomalies and actual anomalies for “cough” in lag 4.

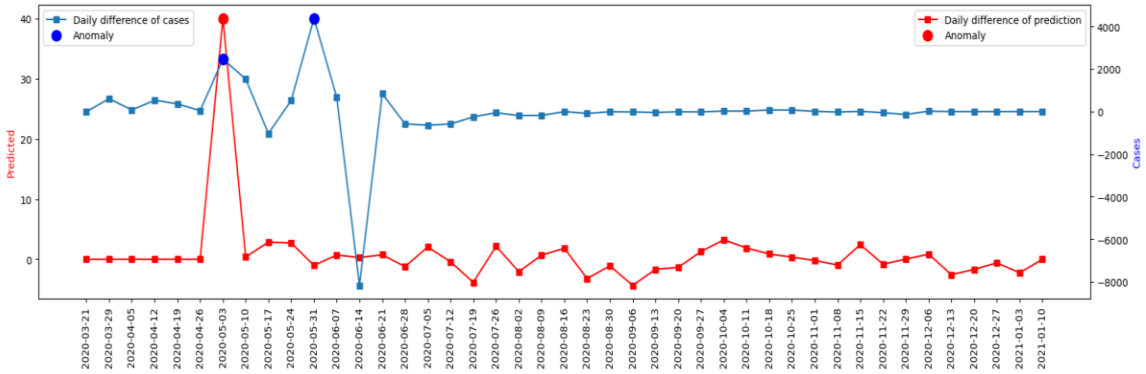


Figure 25: Plot of the predicted anomalies and actual anomalies for “covid” in lag 4.

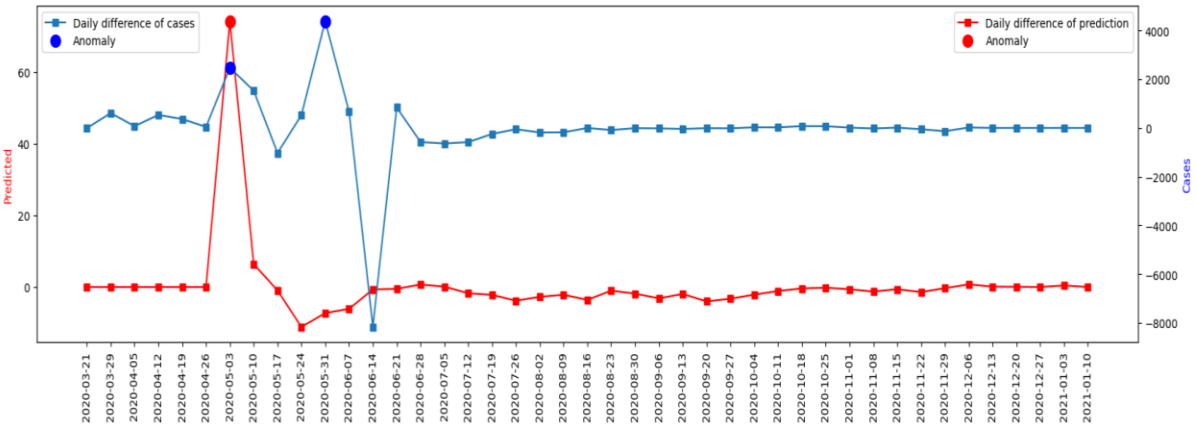
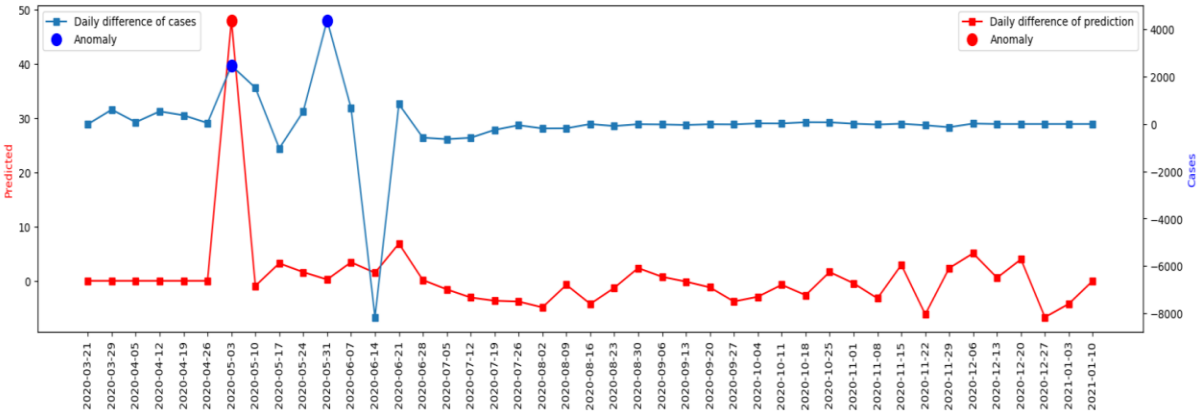


Figure 26: Plot of the predicted anomalies and actual anomalies for “fever” in lag 4.



7.1.4 Testing

Upon conducting tests on our model, a time-shift analysis was performed on the predicted anomalies from 0 up to 4 weeks, with each iteration being shifted by one week. Since the results obtained when applying a standard deviation multiplier ranging from 1 to 2 were consistent in the training phase, we used a standard deviation multiplier of 1. After careful evaluation, we determined that the most effective lag value for our analysis was "covid" in lag 4, as shown in Table 5, with a precision of 100%, recall of 100%, and f1-score of 100%.

Table 8: Results for lag 0-4 for "cough".

lag	TP	TN	FP	FN	Precision	Recall	F1-Score
0	0	12	2	1	0	0	0
1	1	14	1	0	0.5	1	0.67
2	0	14	2	1	0	0	0
3	0	15	2	1	0	0	0
4	0	16	2	1	0	0	0

Table 9: Results for lag 0-4 for "covid".

lag	TP	TN	FP	FN	Precision	Recall	F1-Score
0	0	13	1	1	0	0	0
1	0	14	1	1	0	0	0
2	0	15	1	1	0	0	0
3	1	16	1	1	0	0	0
4	1	18	0	0	1	1	1

Table 10: Results for lag 0-4 for “fever”.

lag	TP	TN	FP	FN	Precision	Recall	F1-Score
0	1	12	2	0	0.33	1	0.5
1	0	12	3	1	0	0	0
2	0	13	3	1	0	0	0
3	0	14	3	1	0	0	0
4	0	15	3	1	0	0	0

Figure 27: Plot of the predicted anomalies and actual anomalies for “cough” in lag 4.

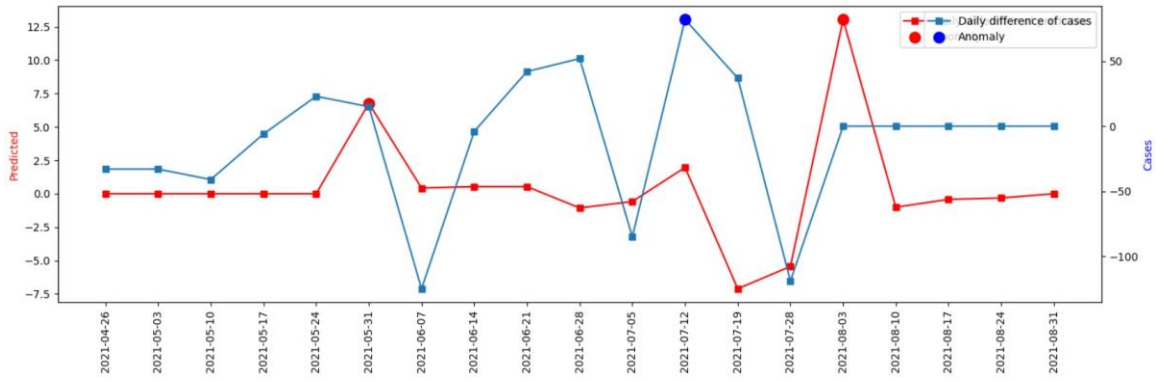


Figure 28: Plot of the predicted anomalies and actual anomalies for “covid” in lag 4.

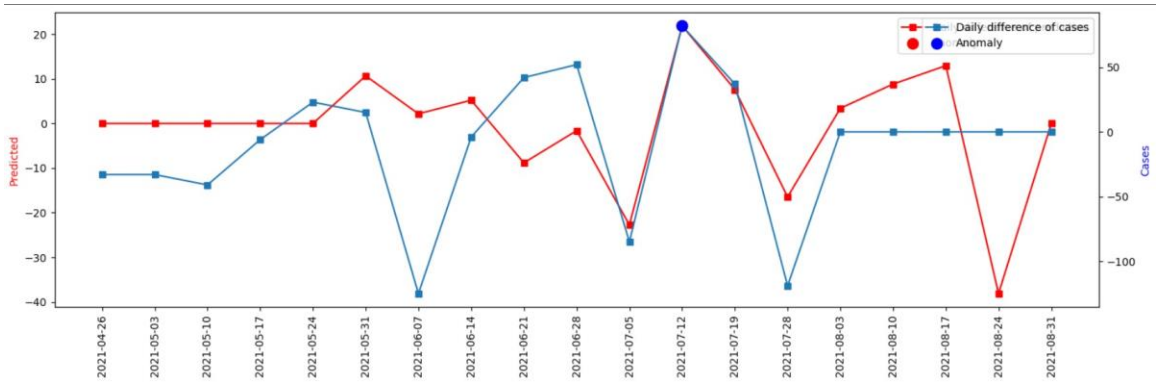
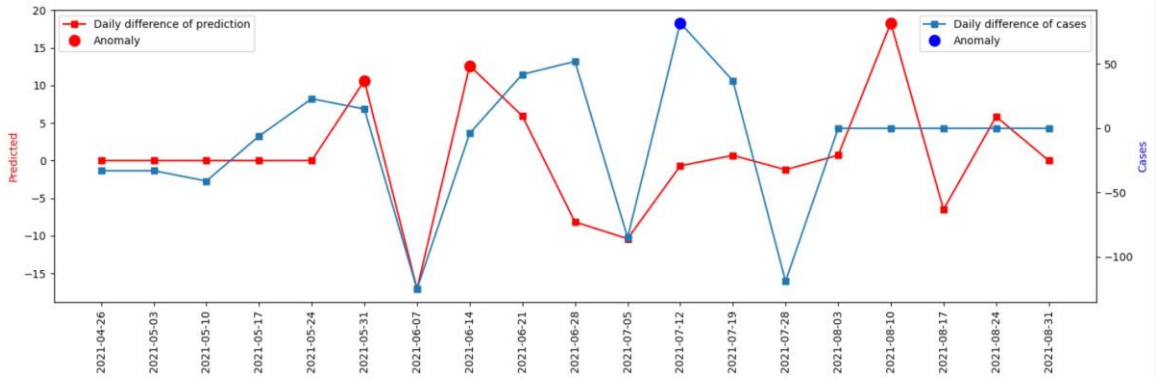


Figure 29: Plot of the predicted anomalies and actual anomalies for “fever” in lag 4.



7.2 Performance Analysis

Time Efficiency is influenced by a variety of factors, such as hardware and software. Table 10 shows the running time for the proposed linear regression model.

Table 11: Model Running Time.

Symptoms	**Running Time**
Covid	23 s
Fever	23 s
Cough	19 s

We used Google Colab that has a positive influence on running time. It provides powerful computing resources, which can significantly reduce the time it takes to train machine learning models [43].

In terms of Space Efficiency, Google Colab provides a limited amount of disk space on the virtual machine. The allocated disk space depends on the type of virtual machine selected by the user. The platform provides 100GB of disk space by default for the runtime virtual machine, which is shared among all the notebooks running on the virtual machine [44].

7.3 Discussion

The present study involved an analysis that led us to identify the search term with the least MSE was “covid” which was 6, as evident in figure 21. Notably, the study evaluated various regions, and the final iteration focused on Makkah and the Eastern region. The findings indicate that the Makkah and the Eastern region had the best performance compared to other regions. During the training of our model, the best values were obtained in lag 4 for all three symptoms. When testing our model, we concluded that the most optimal lag value for our analysis was covid in lag 4. This finding demonstrates that our model possesses the capability to provide timely alerts to hospitals and healthcare providers, up to four weeks in advance of a potential COVID-19 outbreak. The study also provided insights into the relationship between COVID search symptoms and the actual number of COVID-19 cases in different regions. The findings suggest that there is a strong correlation between the two variables, which can be used to predict potential COVID-19 outbreaks.

Chapter 8: Conclusion

In this concluding chapter, we summarize the findings of our research on linear regression modeling as a data analysis tool to predict COVID-19 outbreaks, and provide insights into the implications of our research. The use of linear regression has several advantages including the ability to determine causal relationships between variables, make predictions about future outcomes, and compute results quickly.

In the study that has been presented we found that if the volume of queries submitted to Google with keywords like “covid” increases in a region, then after a few weeks, the number of COVID-19 patients in the emergency rooms of hospitals in the corresponding area will rise accordingly. With this discovery, we conclude that we can predict seasonal disease outbreaks and deploy countermeasures one month in advance by forecasting the spread of Covid-19 infections using a linear regression model trained on the Google Trend dataset.

Furthermore, the study's findings have significant implications for public health policy and practice. By leveraging the power of search engine data and linear regression modeling, healthcare providers and policymakers can develop more effective strategies for monitoring and responding to COVID-19 outbreaks. Specifically, the approach can be used to identify potential hotspots and allocate resources accordingly, thereby improving the overall efficiency and effectiveness of public health interventions. Additionally, the use of linear regression modeling can help to reduce the time and cost associated with traditional epidemiological studies, which can take months or even years to complete. Overall, the study's findings demonstrate the potential of data-driven approaches to public health, and highlight the importance of leveraging emerging technologies and methodologies to address complex health challenges.

Future research can build on the findings of this study to improve the predictive power of linear regression models for COVID-19 outbreaks. Since the datasets obtained from the hospital and google trend resulted in few anomalies, it would be beneficial to carry out this research in regions with higher covid infections and greater anomalies. Future studies may also look at the use of real-time data sources, such as social media platforms and mobile applications, to monitor the development of seasonal disease outbreaks and predict their future trajectories.

References

[1] R. M. Ansari and P. Baker, "Identifying the predictors of covid-19 infection outcomes and development of prediction models," Journal of infection and public health,Jun-2021.[Online].Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7970794/#sec0090title.[Accessd:01-Feb-2023].

[2] "Coronavirus," World Health Organization. [Online]. Available: https://www.who.int/health-topics/coronavirus#/tab=tab_1. [Accessed: 04-Feb-2023].

[3] "Google trends: Understanding the data.," Google News Initiative. [Online]. Available:https://newsinitiative.withgoogle.com/resources/lessons/google-trends-understanding-the-data/. [Accessed: 04-Feb-2023].

[4] Barnett , A.G. and Dobson, A.J. (2012) *Analysing seasonal health data*. Erscheinungsort nicht ermittelbar: Springer.

[5] Knobler SL, Mack A, Mahmoud A, Lemon SM. , *The threat of pandemic influenza: are we ready?. Workshop summary*, 2005Washington, DCNational Academies Press.

[6] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," *Big Data Res.*, vol. 2, no. 2, pp. 59–64, 2015.

[7] Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

[8] Tamhane, A. C., & Dunlop, D. D. (2000). Statistics and Data Analysis: From Elementary to Intermediate. Prentice Hall.

[9] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," *Big Data Res.*, vol. 2, no. 2, pp. 59–64, 2015.

[10]*Researchgate.net*.[Online].Available:https://www.researchgate.net/profile/Rachel-Birnbaum-2/publication/269520846_Ethical_and_Legal_Implications_on_the_Use_of_Technology_in_Counselling/links/5547bc940cf2b0cf7ace931b/Ethical-and-Legal-Implications-on-the-Use-of-Technology-in-Counselling.pdf.

[11] "Machine learning: Trends, perspectives, and prospects" *Google.com*. [Online]. Available:https://drive.google.com/drive/folders/1QjgRbmhUU0FCyFc4JIdFPmkOPMuwxE98. [Accessed: 29-Dec-2022].

[12] "Journal of physics: Conference series," *Iop.org*. [Online]. Available: https://iopscience.iop.org/journal/1742-6596. [Accessed: 29-Dec-2022].

[13] G. Bonaccorso, *Machine Learning Algorithms*. Birmingham, England: Packt Publishing, 2017.

[14] *Investopedia.com*. [Online]. Available: https://www.investopedia.com/terms/s/smoothing.asp. [Accessed: 07-Jun-2023].

[15] Aydogan Ebru, Ali Akcayol M. A comprehensive survey for sentiment analysis tasks using machine learning techniques. INnovations in Intelligent SysTems and Applications (INISTA), 2016 International Symposium on. IEEE, 2016.

[16] "Research," *Electrical Engineering and Computer Science*, 17-Aug-2021. [Online].Available:https://lassonde.yorku.ca/eecs/research/.[Accessed:30-Dec-2022].

[17]Third Edition, "Applied Linear Regression," *Edu.ps*. [Online]. Available: http://site.iugaza.edu.ps/biqelan/files/2010/09/S.-Weisberg.-Applied-Linear-Regression-Wiley2005ISBN-0471663794329s.pdf. [Accessed: 30-Jan-2023].

[18] R. M. Ansari and P. Baker, "Identifying the predictors of covid-19 infection outcomes and development of prediction models," Journal of infection and public health, Jun-2021.

[Online].Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7970794/.[Accessed: 23-Jan-2023].
[19] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques,"ProcediaComputerScience,01-Sep-2015.[Online].Available:https://www.sciencedirect.com/science/article/pii/S1877050915023479. [Accessed: 19-Jan-2023].
[20] Wavefront (no date), *Detecting Anomalies with Functions and Statistical Functions.* Availableat:https://docs.wavefront.com/query_language_statistical_functions_anomalies.html(Accessed: 24 January, 2023).
[21]Secherla, S. (2021) *Understanding optimization algorithms in machine learning*, *Medium*.TowardsDataScience.Availableat: https://towardsdatascience.com/understanding-optimization-algorithms-in-machine-learning-edfdb4df766b (Accessed: December 25, 2022).
[22] Ghosh, S. (2022) *A comprehensive guide to data preprocessing*, *neptune.ai*. Available at: https://neptune.ai/blog/data-preprocessing-guide (Accessed: December 25, 2022).
[23]*Researchgate.net*.[Online].Available:https://www.researchgate.net/profile/Sathish-Kumar-26/publication/339404272_Machine_Learning_for_Predicting_Development_of_Asthma_in_Children/links/5e6a7365a6fdccf321d908e0/Machine-Learning-for-Predicting-Development-of-Asthma-in-Children.pdf. [Accessed: 04-Jan-2023].
[24] B. Alkouz, Z. Al Aghbari, M. Al Garadi, and A. Sarker, "Deepluenza: Deep learning for influenza detection from Twitter" *Expert Systems with Applications,* 2022.
[25] Ocampo, A.J., Chunara, R. and Brownstein, J.S. (2013) "Using search queries for malaria surveillance, Thailand," *Malaria Journal*, 12(1). Available at: https://doi.org/10.1186/1475-2875-12-390.
[26] E. Alharbi and M. Abdullah, "AsthmaAttack Prediction based on Weather Factors," *Semanticscholar.org*.[Online].Available:https://pdfs.semanticscholar.org/150c/16c97b9ff6fea503571159eeb3153a57b42a.pdf.
[27] M. Aftab, R. Amin, D. Koundal, H. Aldabbas, B. Alouffi, and Z. Iqbal, "Classification of COVID-19 and Influenza Patients Using Deep Learning" *National Library of medicine*, 2022.
[28] D. J. Park, M. W. Park, H. Lee, Y.-J. Kim, Y. Kim, and Y. H. Park, "Development of machine learning model for diagnostic disease prediction based on laboratory tests,"Nature News, 07-Apr-2021. [Online]. Available: https://www.nature.com/articles/s41598-021-87171-5.[Accessed: 01-Jan-2023].
[29] C. Chrysostomou, F. Alexandrou, M. A. Nicolaou, and H. Seker, "Classification of Influenza Hemagglutinin protein sequences using convolutional neural networks," *arXiv [q-bio.QM]*, 2021.
[30] Abdualgalil, B., Abraham, S. and Ismael, W.M. (no date) *Early diagnosis for dengue disease prediction using efficient machine learning techniques based on clinical data*, *Journal of Robotics and Control (JRC)*. Available at: https://journal.umy.ac.id/index.php/jrc/article/view/14387 (Accessed: January 2, 2023).
[31] Abdualgalil, B., Abraham, S. and Ismael, W.M. (no date) *Early diagnosis for dengue disease prediction using efficient machine learning techniques based on clinical data*, *Journal of Robotics and Control (JRC)*. Available at: https://journal.umy.ac.id/index.php/jrc/article/view/14387 (Accessed: January 2, 2023).

[32] M. M. Yousef, "Heart disease prediction model using naïve Bayes algorithm and machine ..., "2020. [Online]. Available at: https://www.researchgate.net/publication/349353189_Heart_Disease_Prediction_Model_Using_Naive_Bayes_Algorithm_and_Machine_Learning_Techniques.[Accessed: 02-Jan-2023].
[33] Alam, R. (2020) *Normalization vs standardization explained*, *Medium*. Towards DataScience.Availableat:https://towardsdatascience.com/normalization-vs-standardization-explained-209e84d0f81e (Accessed: January 19, 2023).
[34] V.Bhadana, A. S. Jala, and P. Pathak, "A comparative study of machine learning models for COVID-19 prediction ..., IEEE xplore , 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9312112. [Accessed: 19-Jan-2023].
[35] J. Bimasgara, T. Abdullah's, F. Ramdhani, and H. Sumpena, "RELIABILITY OF GOOGLE TREND AS DIGITAL ETHNOGRAPHY TOOLS TO CAPTURE WORLD TREND" *SEMANTIC SCHOLAR*, 2021.
[36] J. Santos, "what is Google Colab?," SmartAI Blog, 14-Mar-2021. [Online]. Available: https://smartai-blog.com/what-is-google-colab/. [Accessed: 02-Feb-2023].
[37] "numpy: Fundamental package for array computing in Python." Accessed: May. 21, 2023. [MacOS, Microsoft :: Windows, POSIX, Unix]. Available: https://www.numpy.org.
[38] "Scikit learn tutorial," Online Courses and eBooks Library, https://www.tutorialspoint.com/scikit_learn/index.htm#:~:text=Scikit%2Dlearn%20(Sklearn)%20is,a%20consistence%20interface%20in%20Python.(accessed Jun. 7, 2023).
[39] "Python: Introduction to matplotlib," GeeksforGeeks, https://www.geeksforgeeks.org/python-introduction-matplotlib/ (accessed Jun. 7, 2023).
[40] A. Bronshtein, "A quick introduction to the 'Pandas' python library," Towards Data Science, 18-Apr-2017. [Online]. Available: https://towardsdatascience.com/a-quick-introduction-to-the-pandas-python-library-f1b678f34673. [Accessed: 04-Jun-2023].
[41] "Medium," *Medium*. [Online]. Available: https://towardsdatascience.com/linear-regression-models-with-sklearn-a80b375b12b2. [Accessed: 04-Jun-2023].
[42] "Medium," *Medium*. [Online]. Available: https://towardsdatascience.com/metrics-in-python-sklearn-and-how-to-implement-them-f4c970c22bf3. [Accessed: 04-Jun-2023].
[43] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.
[44] "Medium," *Medium*. [Online]. Available: https://towardsdatascience.com/how-to-use-google-colab-34ce8b5ce27ca. [Accessed: 06-Jun-2023].