# **Introduction**

COVID-19 (Coronavirus) is a novel infectious disease resulted in unprecedented hazards emerged in the Wuhan city (Hubei Province, China) in December 2019. The virus spread rapidly outside China and the World Health Organization (WHO) and was recognized as an epidemic in March 2020 [1]. The epidemic has caused dramatic loss of life around the world. According to the data available in the worldometer, to date, there are more than 49 million of confirmed cases, and more than 1 million death have been reported, affecting 216 countries. The epidemic has not only become a treat to human health but also it affects the entire economy by creating disruptions in social relationships, economic stability, educational activities, political decision making. Fever, dry cough, tiredness, headache, muscle pain, shortness of breath is some of the clinical symptoms of COVID-19 [2]. Individuals who are 70 years of age or over also individuals with underlying health conditions such as, kidney diseases, respiratory disease are at higher risk of being positive [3].

In this challenging situation, identifying the individuals who are positive and who are at high risk are critical to restraint the spread of the novel disease. Lack of vaccines and lack of effective remedial measures have accelerated the use of artificial intelligence and machine learning on a variety of fronts in determining viral dissemination pattern, rapid and accurate diagnosis, development of new curative approaches and identification of people most prone to the disorder [4]. The artificial intelligence [5] and machine learning techniques [6] uses Electronic Health Record (EHR) [7, 8] where EMR is a databases that provides information such as geographic, demographic and clinical measures of patients used in the healthcare sector are being very for the analysts to identify the affected people or those whose are highly vulnerable to be positive. These records provide accurate and up-to-date health-related information of patients, which is created, managed, and consulted by authorised clinicians and staff across the health care organisation.

Machine learning (ML) technique plays a vital role in epidemiological analysis and forecasting during the outbreak situations[9, 10] in determining the epidemiological patterns and plan early steps to prevent the spread of the virus. In our study we are using a synthetic EHR database in determining the COVID-19 positive patients by using a Machine learning algorithm called Gaussian Naïve Bayes Model. The EHR database provided a realistic example of real-world data collected from several major hospitals and regional clinics across Washington. It shows a 10-year medical history of 9,500 patients who had at least one COVID-19 test, 800 of whom tested positive. The database includes information such as observation period, conditions and measurement, personal information, hospital visit etc. The information is sorted based on its relevance and a ML algorithm called Naïve Bayes model is built in order to predict the result of an initial COVID-19 which classifies to which class (positive or negative )a patient belong to by calculating a prediction score from the developed model. The objective of the study is to give insight about the individuals who are testing COVID-19 positive outside the hospital and identifying the factors that makes an individuals to be positive or identify the high risk factors for hospitalization of individuals based on the information sorted from the EHR database. This analysis can help the healthcare people to improve the medical decision making and communication between the health care professional to restraint the spread of the virus.

# **Literature Review**

The Electronic Health Record is a database used by the medical sector and is created and maintained by medical personnel. Medical professionals expect the quality and efficiency of the medical industry to improve through meaningful medical decisions and mutual communication between the medical personnel [11]. In the current situation of the COVID-19 pandemic, Machine learning techniques using EHR data is useful for many studies in analysing and in determining the people who are likely to be positive.

Many studies are carrying on using the EHR data in COVID-19 analysis. A study [12] was conducted on the basis of EHRs of 17 million adult NHS patients in the context of the factors associated with the COVID-19 related hospital death and the result says that a quantified range of factors for death from COVID-19 is available in the EHR database but some are not available. But the drawback in the study is the EHR data available is not up to date. So, in order to get much accurate results up to date database is needed.

Research by [13] was conducted to describe the implementation of critical technology support to optimize clinical management of COVID-19 epidemics based on the information available in the EHR database and as a result an outline was designed and implemented that would help to identify the individuals who are at high risk to the pandemic. Another study by [14] analyse the COVID-19 risk and outcomes in patients who are suffering from substance use disorders based on the data available on the EHR in the US using statistical analysis and the result says that individuals with substance use disorders (SUD) and opioid use disorders (OUD) are having increased risk for COVID-19 and its outcomes says to treat the individuals with SUD in advance to control the pandemic.

There are a few studies which uses machine learning techniques such as Logistic Regression [15], Linear Discriminant Analysis (LDA) [16], K-Nearest Neighbours (KNN) [17], Support Vector Machine (SVM) [18] etc used in the classification of the severity of the infection, prognosis of the outcome of the disease and the need for hospitalization. A study by [19] uses machine learning techniques to predict the a poor prognosis in positive COVID-19 patients and possible outcomes and the result of this study shows that the disease outcome can be predicted with a ROC AUC (Area under the receiver operating characteristic curve) of 0.92. However, the model was built on a small sample of data with less features at the beginning stage of the epidemic and these small sample size might affect the model’s performance. Larger datasets can improve the training phase and improve predictive performance.

A study by [20] propose the use of machine learning and deep learning models with the aim of understanding the daily behaviour, predicting the future behaviour of COVID-19 in the nation using real-time information from the Johns Hopkins dashboard and the result of the study shows that the polynomial regression results a minimum root mean square error score over the other proposed approaches in the forecasting of the COVID-19 transmission. However, if proliferation follows the predicted trend of the PR model, it will result in enormous loss of life. [21] proposed machine learning approaches in detecting COVID-19 positives using clinical text data. Feature engineering techniques like Term frequency/inverse document frequency (TF/IDF), Bag of words (BOW) and report length are used to extract features and then the features were supplied to machine learning classifiers such as Logistic regression and Multinomial Naïve Bayes and the results shows 96.2% accuracy. However, the model used small sample size of data, also the study says that with higher number of data the efficiency of the model can be increased further.

Various forecasting models are built in a study by [22] using machine learning algorithms, and their performance is calculated and evaluated. The result of the study says that Random Forest Regressor and Random Forest classifier surpass the other machine learning models like SVM, KNN+NCA, Decision Tree Classifier, Gaussian Naïve Bayesian Classifier, Multilinear Regression, Logistic Regression and XGBoost Classifier. Some study uses X-ray images in determining the COVID-19 positive patients. A study by [23] introduced a deep convolutional neural network(CNN) for the disclosure of COVID-19 cases from chest X-ray images. Based on the chest X-ray images from COVID-19 tested positive and negative patients, a CNN model is developed, which diagnose COVID-19 from chest radiography images.

ML can be used for COVID-19 diagnosis which requires a lot of research effort, but it is not yet widely used. Since less work is done on diagnosis and prognosis using EHR database and machine learning algorithm to classify COVID positive patients and in determining the high-risk factors that makes an individual highly vulnerable to be positive. In our study we are performing a machine learning classification technique called Gaussian Naïve Bayes Model [24] which is less explored in the analysis of COVID-19. So, we built a model using Gaussian Naïve Bayes algorithm to predict the result of an initial COVID-19 test and its corresponding prediction accuracy based on 10 years of clinical records data from 9500 patients who at least had one test for COVID-19, of which 800 patients have tested positive.

# **Approach & Methodology**

# **Data**

A synthetic Electronic health record which is a database used by hospitals to manage health records of patients is developed by the challengers and made available protecting the patient’s privacy which provides a realistic example of actual data collected from multiple large hospitals and regional clinics across Washington. It represents 10 years of clinical records from 9500 patients who at least had one test for COVID-19, of which 800 patients have tested positive.

There are multiple clinical tables and a challenge table among which we have limited our usage to few tables focusing on condition\_occurrence that describes a record of medical conditions of the person, measurement table that contains records, i.e. structured values obtained through examination or testing samples or pathology reports etc. Along with these, person table that stores demographic information of patients which helps to uniquely identify a patient and gold\_standard file which stores the true status of patients to calculate the accuracy of classification model are also considered [25, 26].

# **Data Pre-processing**

The data is filtered based on the most common COVID-19 symptoms and human health conditions of people at high risk to get effected to COVID-19 [22],[27]

|  |  |  |
| --- | --- | --- |
| Most Common | Moderate | Severe |
| * Fever * Dry cough * Fatigue | * Loss of taste or smell, * Nasal congestion, * Conjunctivitis (also known as red eyes) * Sore throat, * Headache, * Muscle or joint pain, * Different types of skin rash, * Nausea or vomiting, * Diarrhea, * Chills or dizziness. | * Shortness of breath, * Persistent pain or pressure in the chest, * High temperature (above 38 °C). |

Table 1: Symptoms of COVID-19

|  |
| --- |
| Health Conditions |
| * Age >60 years * Blood pressure * Oxygen saturation * Heart Diseases * Diabetes |

Table 2: Health Conditions of patients

from condition\_occurrence and measurements tables based on the descriptions provided in the data\_dictionary file by using the concept\_id and created a feature matrix. The age is calculated from the birth\_datetime in person table and further filtered people in the age of high risk, whose body temperature and blood pressure are higher than the usual and oxygen saturation is low. More variables are further added based on the medical conditions of the patient with heart diseases, diabetes etc. All this information is brought together to a feature set with the help of person id, and concept id’s as X, and the gold standard values of those persons which describe demographic info are stored as Y to train a model from the training set, in the same way, data is filtered from evaluation to test the model results.

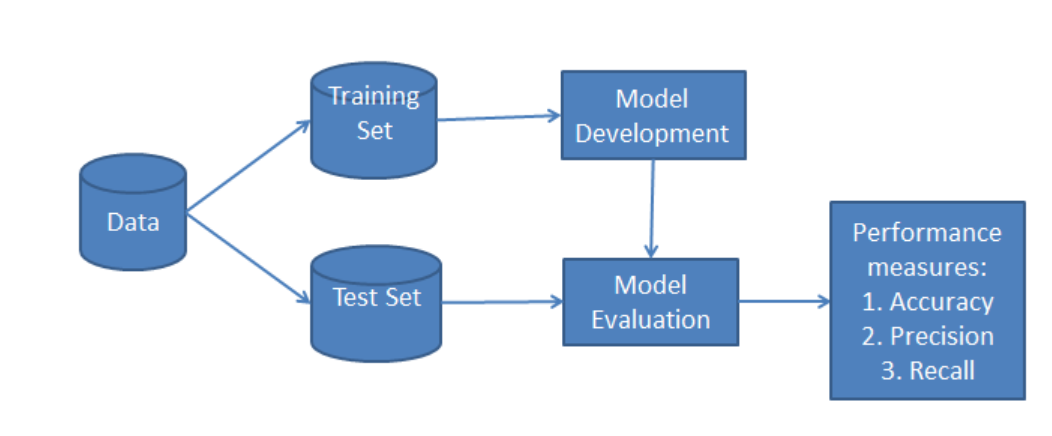


Figure 1: Classification workflow

A training and evaluation sets are provided to train on a specific set and evaluate on the other without the need to perform a split operation which are downloaded from synapse.

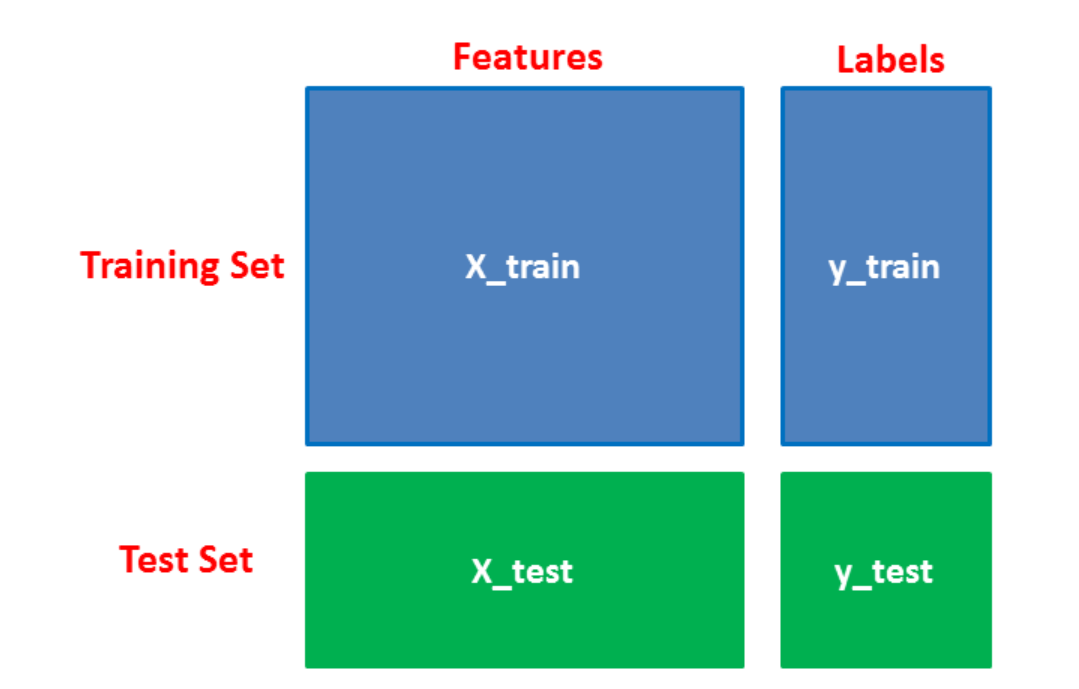


Figure 2: Splitting Data

# **Machine Learning Model**

Machine Learning is a research area to construct computer algorithms that automatically improves with experience is proven to be of great practical value in various domains such as medical. A supervised machine learning algorithm is developed for classification of patients by identifying to which set of categories the patient belongs to. The developed algorithm is applied to the feature set developed from the patients tested for COVID-19 aiming to create models able to predict disease outcome [19].

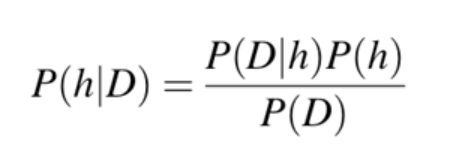
A most straightforward, fast, classical, accurate and reliable classification algorithm suitable for large data which is successfully used in various applications called Naïve Bayes classification model is develop. They have high accuracy on large data sets [28, 29, 30, 31].

The Naive Bayes is automatically achieved by only inducing the numerical parameters of the model. It only requires information about the variables and their corresponding values to estimate probabilities, leading to a computational time complexity that is linear with respect to the amount of training instances. It is also space efficient, requiring only the information provided by two-dimensional tables, in which each entry corresponds to a probability estimated for a given value of a variable. It has provided good results on several domains.

Assumptions/Limitations: The fundamental Naïve Bayes assumption is that each feature makes an independent and equal contribution to the outcome [19].

# **Naïve Bayes Classification model**

Naïve Bayes classification [31] uses Bayes Theorem of probability for class prediction of patients. This theorem works on conditional probability i.e. the probability that it will happen given that a condition has already occurred which gives us the probability of event using its prior knowledge. i.e.



where

* P(h): the probability of hypothesis h being true also known as the prior probability of h (regardless of the data).
* P(D): the probability of the data also known as the prior probability (regardless of the hypothesis).
* P(h|D): the probability of hypothesis h given the data D also known as posterior probability.
* P(D|h): the probability of data d given that the hypothesis h was true also known as posterior probability.

**Working steps of Naïve Bayes classifier:**

**Step 1:** Calculate prior probability for given class labels.

**Step 2:** Calculate conditional probability with each attribute for each class.

**Step 3:** Multiply same class conditional probability.

**Step 4:** Multiply prior probability with step 3 probability.

**Step 5:** Record the probability of a patient to be positive.

**Model Evaluation**

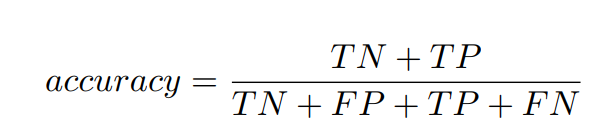
The performance of the learning model is evaluated in terms of accuracy, area under the Receiver Operating Characteristic curve, area under the Precision-Recall curve and finally confusion matrix [19].

**Confusion Matrix:** It is a 2-dimensional matrix with a row and column for each class in which each element shows number of examples for which actual class is the row and predicted class is column.

|  |  |  |
| --- | --- | --- |
| Actual Class | *Predicted Class* | |
|  | ***Positive*** | ***Negative*** |
| **Positive** | True Positive TP | False Negative FN |
| **Negative** | False Positive FP | True Negative TN |

Table 3: Confusion Matrix scheme

**Accuracy:** Ratio of number of correct predictions to total number of samples.



**Area under the curve:** It measures the entire 2-dimensional area underneath the curve which gives an aggregate measure of performance in a single value. AUC values for ROC and PR curves are used in our results.

**Area under Receiver Operating Characteristic AUROC:** It is an evaluation of performance of a classifier which depict the performance without regard to class distribution and error costs.

**Area under Precision-Recall AUPR:** It is an alternative to ROC which does not use number of True Negative results and is a good choice to compare models when facing imbalanced datasets. It shows relationship between precision (predicted positive values) and recall (sensitivity).

# **Findings and Results**

The results that the model delivered on synthetic data were efficient, but the results on the UW dataset are not good enough to predict the patients COVID-19 status. The model delivered an accuracy of 89% when the study was conducted on synthetic data. After which, the model is pushed to docker for evaluation on UW dataset, which is the real dataset. The results of the model were provided by the synapse team after running our model on UW data which delivered an accuracy of 54.9% which is obtained based on the output file generated by the model with a person\_id and score which is the probability of the person being identified COVID-19 positive.

## classification accuracy is usually preferred to measure the performance of our model, but it's not enough to really judge our model. Here we evaluate the model using different types of evaluation metrics. Here are the results that can help understand the performance of the model:

## **On Synthetic data:**

* Classification Accuracy:



Figure 3. Accuracy score on Synthetic Data

The accuracy of the model is calculated by comparing the model’s prediction with the actual values for the test set. It says about the altitude of the measured value relative to the true value. This means that the closer the measurement is to the true value, the more accurate it will be. The builded model is giving an accuracy of 89.8%, which is a good accuracy score for the model prediction.

* Area Under Curve (AUROC curve)

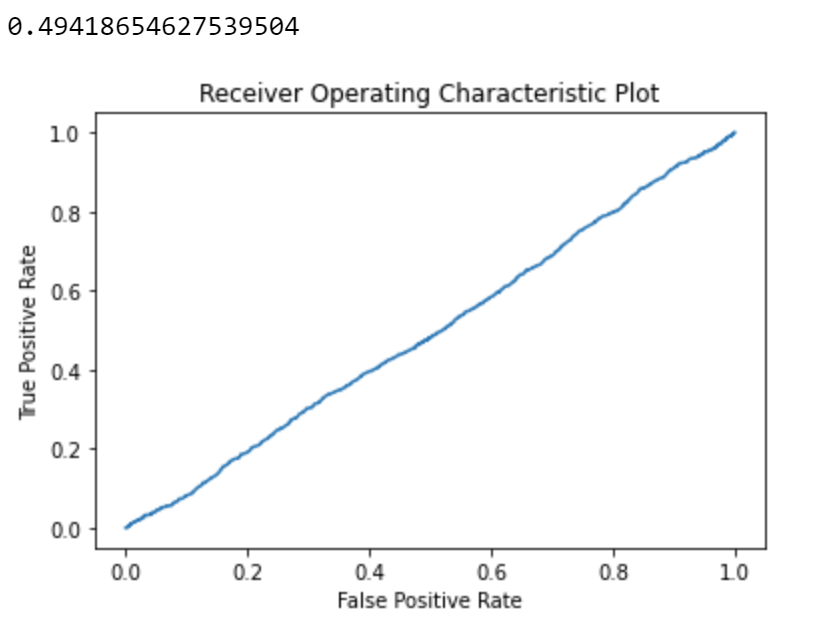


Figure 4: AUROC curve

AUC is the area under the curve of the plot False Positive Rate Vs True Positive Rate at various points of [0, 1]. The FPR and TPR both are computed at varying threshold values such as (0.00, 0.02, 0.04, …., 1.00) and a graph is drawn. The higher the value, the better the performance of our model. The developed model gives a ROC value of 49.4%, which means that the performance of the is model is unsatisfying in prediction.

* Precision Recall Curve

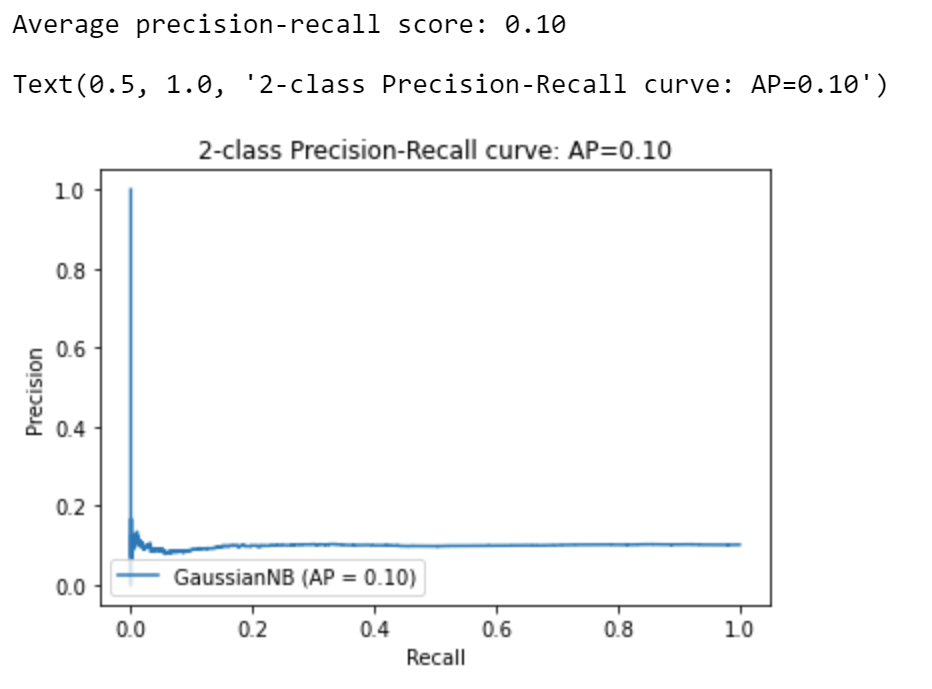


Figure 5: AUPR curve

## **On UW data:**

Graphical user interface, application

Description automatically generated

Figure 7: Results on UW dataset

Docker results provided two scores to evaluate model performance on UW dataset which are:

* AUROC is a performance metric for judging if a patient will be COVID positive or negative, it evaluates the ability of the model to differentiate between positive cases and negative cases. For example, a score like eighty nine percent means that the model can predict up to eighty nine percent of the time if a patient is COVID positive than to a randomly selected patient who is not COVID positive.
* The area under the precision-recall curve is AUPR which is the probability that tells if a patient who is COVID positive is selected from the ranked list, then a record above it on the list will also be positive.

# Discussion

# The study conducted is focused on designing a model on synthetic data and then pushed.

With the goal of designing a model that can predict if a person is COVID-19 positive or not, we filtered a small subset of features from a larger dataset by working on required data files that are the most significant attributes for designing the optimal model based on the study of COVID-19 symptoms and high level risk conditions in patients. By looking at the research papers and the discussion forum on COVID-19 DREAM CHALLENGE and studying the aim of authors of the challenge also helped us to deduce what features can be important for the kind of application they are trying to design.

For being able to run the model on UW dataset the script is designed in a specific format which creates an output file output.csv upon a successful run of the model on docker. This challenge is developed to help the author build a model in their research, which is an application that delivers if a person will be COVID-19 positive or negative upon inputting the feature details of a person considered in the model.

# Limitations

There were certain hinderances to our development phase because of which the model could not be tested rigorously for improvement i.e.

* The model could only be tested once a week, keeping in mind the fair chance to allow the other team to upload their model as well.
* The script involves a particular format and reading from only certain files that the example model contains, this is because in discussion forum we were able to figure out that the files do not have the same name in UW dataset which makes it riskier for the model to pass on the UW dataset.
* The website also has a downtime every week which caused delay in uploading the model to docker.

# **Conclusion**

COVID-19 pandemic has affected everyday life worldwide. The humankind is facing unprecedented economic and social difficulties. In order to identify the individuals who are high risk in advance Electronic Health Record database can be used. Based on the data available from the EHR database, a Machine learning technique is performed to determine which class a patient belongs to. In the study we conducted, we developed a Naïve Bayes model on a selected subset of data from the various clinical measures data available from the synthetic EHR database and the builded model results in an accuracy of 0.8949 on synthetic data and he model results were provided by the Synapses team after running our model on UW data, giving an accuracy of 54.9%, which is not good enough in predicting COVID-19 patient status. In future, to improve the accuracy of the model, more features can be included into the existing model also into some other different machine learning classification model such as Random Forest, Decision Tree, regression models, Neural Networks and more to determine the best performing model in the prediction of COVID-19 patient status.

# **References**

[1] W. H. Organization, "WHO Director-General’s opening remarks at the media briefing on COVID-19. 2020," *Website:/. Accessed,* vol. 19, 2020.

[2] L. Fu *et al.*, "Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: a systematic review and meta-analysis," *Journal of Infection,* 2020.

[3] A. Clark *et al.*, "Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study," *The Lancet Global Health,* vol. 8, no. 8, pp. e1003-e1017, 2020.

[4] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, "Artificial intelligence and machine learning to fight COVID-19," ed: American Physiological Society Bethesda, MD, 2020.

[5] N. J. Nilsson, *Principles of artificial intelligence*. Morgan Kaufmann, 2014.

[6] D. Michie, D. J. Spiegelhalter, and C. Taylor, "Machine learning," *Neural and Statistical Classification,* vol. 13, no. 1994, pp. 1-298, 1994.

[7] D. Kalra, "Electronic health record standards," 2006.

[8] N. Menachemi and T. H. Collum, "Benefits and drawbacks of electronic health record systems," *Risk management and healthcare policy,* vol. 4, p. 47, 2011.

[9] G. Kalipe, V. Gautham, and R. K. Behera, "Predicting malarial outbreak using Machine Learning and Deep Learning approach: A review and analysis," in *2018 International Conference on Information Technology (ICIT)*, 2018: IEEE, pp. 33-38.

[10] R. Singh, R. Singh, and A. Bhatia, "Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics," *Int. J. Adv. Sci. Res,* vol. 3, no. 2, pp. 19-24, 2018.

[11] T. R. Schopf, B. Nedrebø, K. O. Hufthammer, I. K. Daphu, and H. Lærum, "How well is the electronic health record supporting the clinical tasks of hospital physicians? A survey of physicians at three Norwegian hospitals," *BMC Health Services Research,* vol. 19, no. 1, pp. 1-9, 2019.

[12] E. Williamson *et al.*, "OpenSAFELY: factors associated with COVID-19-related hospital death in the linked electronic health records of 17 million adult NHS patients," *MedRxiv,* 2020.

[13] J. J. Reeves *et al.*, "Rapid response to COVID-19: health informatics support for outbreak management in an academic health system," *Journal of the American Medical Informatics Association,* vol. 27, no. 6, pp. 853-859, 2020.

[14] Q. Q. Wang, D. C. Kaelber, R. Xu, and N. D. Volkow, "COVID-19 risk and outcomes in patients with substance use disorders: analyses from electronic health records in the United States," *Molecular psychiatry,* pp. 1-10, 2020.

[15] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[16] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," in *Institute for Signal and information Processing*, 1998, vol. 18, no. 1998, pp. 1-8.

[17] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers," *arXiv preprint arXiv:2004.04523,* 2020.

[18] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[19] F. S. H. Souza, N. S. Hojo-Souza, E. B. Santos, C. M. Silva, and D. L. Guidoni, "Predicting the disease outcome in COVID-19 positive patients through Machine Learning: a retrospective cohort study with Brazilian data," *medRxiv,* 2020.

[20] N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms," *medRxiv,* 2020.

[21] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. M. U. Din, "Machine learning based approaches for detecting COVID-19 using clinical text data," *International Journal of Information Technology,* vol. 12, no. 3, pp. 731-739, 2020.

[22] K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar, and Y. N. Pawan, "Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms," *International Journal,* vol. 8, no. 5, 2020.

[23] L. Wang and A. Wong, "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images," *arXiv preprint arXiv:2003.09871,* 2020.

[24] M. Ontivero-Ortega, A. Lage-Castellanos, G. Valente, R. Goebel, and M. Valdes-Sosa, "Fast Gaussian Naïve Bayes for searchlight classification analysis," *Neuroimage,* vol. 163, pp. 471-479, 2017.

[25] COVID-19 DREAM CHALLENGE. Available: <https://www.synapse.org/#!Synapse:syn21849255/wiki/602406>

[26] COVID-19 DREAM CHALLENGE. Available: <https://www.synapse.org/#!Synapse:syn21978034>

[27] Who.int. 2020. *Coronavirus Disease (COVID-19)*. [online] Available at: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses#:~:text=symptoms>. [Accessed 9 November 2020].

[28] Gandhi, R., 2018. Naive Bayes Classifier. [online] Medium. Available at: <https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>. [Accessed 9 November 2020].

[29] Chauhan, N., 2020. Naïve Bayes Algorithm: Everything You Need To Know - Kdnuggets. [online] KDnuggets. Available at: <https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html>. [Accessed 9 November 2020].

[30] Dwivedi, R., 2020. What Is Naive Bayes Algorithm In Machine Learning? | Analytics Steps. [online] Analyticssteps.com. Available at: <https://www.analyticssteps.com/blogs/what-naive-bayes-algorithm-machine-learning>. [Accessed 9 November 2020].

[31] Navlani, A., 2018. Naive Bayes Classification Using Scikit-Learn. [online] DataCamp Community. Available at: <https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn>. [Accessed 9 November 2020].