CIND 820: Big Data Analytics Project

The Chang School of Continuing Education, Toronto Metropolitan University

Title: Predicting Disease from Symptoms Using Machine Learning

Divya Arora

Student Id: 500334710
Email: divya.arora@torontomu.ca

Date of Submission: June 05, 2023 (Literature Review)

Lead Instructor: Tamer Abdou

INTRODUCTION

The use of machine learning algorithms for disease detection and prediction has made significant advancements in recent years in the healthcare industry. The creation of machine learning algorithms that can accurately predict diseases based on the symptoms displayed by patients is of great interest. The traditional method of disease diagnosis relies on the knowledge of medical experts who, in order to make informed decisions evaluate patient symptoms, medical history, and diagnostic tests (Mahata et al., 2023). Factors like human error, time consumption, cost and limited availability restrict the accuracy of human diagnostics (Mahata et al., 2023). On the other hand, machine learning algorithms can process huge amounts of data and spot complex patterns improving diagnostic accuracy and can assist medical experts in making more informed decisions, leading to early detection and personalized treatment plans (Mahata et al., 2023).

The idea behind machine learning-based disease prediction models is making use of historical patient data. The dataset in this project consists of patient's symptoms and the corresponding disease diagnosis. When this data is fed into a machine learning algorithm, the model can learn to recognize patterns and associations between symptoms and diseases. Based on the symptoms displayed by a new patient, the model is able to estimate the likelihood of various diseases based on the acquired knowledge.

Therefore, as advancements in technology and data availability continue, machine learning models for disease prediction using symptoms hold great potential to transform healthcare by enhancing diagnostic accuracy, improving patient outcomes and improving global healthcare (Mahata et al., 2023). Moreover, it can assist the medical experts to prioritize urgent cases and ultimately reduce mortality rates.

LITERATURE REVIEW

This section describes the published articles on predictive analysis as well as the methods the authors used. These studies may be relevant to the dataset being utilised in this project.

As per the article, "Multiple disease prediction using Machine learning algorithms", in order to assess the efficacy of therapeutic medicines, data mining for healthcare is an interdisciplinary study field that emerged from database statistics. Diabetes-related heart disease is one form of heart disease that affects diabetics (Arumugam et al., 2021). Diabetes is a chronic disease that arises from either inadequate insulin production by the pancreas or inappropriate insulin utilisation by the body. Cardiovascular disease, commonly known as heart disease, is a group of illnesses that harm the heart or blood arteries (Arumugam et al., 2021). There are several data mining classification methods for predicting heart disease, however not enough data exists to predict heart disease in a person with diabetes. Since the decision tree model consistently outperformed the naive Bayes and support vector machine models, the authors optimised it for the best performance in identifying the risk of heart disease in people with diabetes (Arumugam et al., 2021).

The article, "Stratification of Parkinson Disease using python scikit-learn ML library", talks about whether a patient has parkinsons disease or is in good health (Kolte et al., 2019). Parkinson's disease is a condition that affects the central nervous system and affects how well the body moves. The symptoms of this chronic illness get worse with time and are commonly experienced by older

people. This illness may be examined using general machine learning techniques, which offer varying degrees of accuracy (Kolte et al., 2019). The most accurate alternative is picked in order to assess if the patient has the ailment or not since it will yield the greatest results (Kolte et al., 2019). The Parkinson disease dataset with repeated acoustic characteristics had 48 associated features for 240 people. Whether the patient has Parkinson's disease or is in good health is indicated by the target "status" column, which has a value of 0 or 1. Since healthcare applications typically require more precision and cannot be compromised, the ideal application is chosen with the maximum degree of accuracy possible. The naive Bayes classifier, gradient boosting, and support vector machines are the main models applied in this research (Kolte et al., 2019). By evaluating the characteristics of the patients, these methods can be quite effective for doctors in predicting the condition. The dataset was subjected to the chosen techniques and principal component analysis, and the following degrees of accuracy were attained: Decision tree (71.3%), Support Vector Machine (81.2%), and Logistic Regression (78.7%). Gaussian Naive Bayes classifier has the highest accuracy (86.25%) out of all the tested models (Kolte et al., 2019).

The article, “Disease Prediction using Machine Learning Algorithms” is similar to this project as the researcher’s dataset is similar to my dataset but their work is not being completely replicated as the code for this project is written in python and their source code is not available in their article. In the later section of this report, proposed work and methodology, algorithms are listed that are proposed to be used such as Logistic Regression, Random Forest, KNN and K-means clustering. Whereas, machine learning algorithms used by (Grampurohit & Sagarnal, 2020b) are Decision Tree, Random Forest, Naïve Bayes.

Medical database analysis that is accurate helps with early disease identification, patient care, and social support. Machine learning techniques have been successfully used in many fields, including the diagnosis and prognosis of diseases. By assisting clinicians in early disease prediction and diagnosis, a classifier system developed using machine learning algorithms seeks to considerably aid in the resolution of health-related issues (Grampurohit & Sagarnal, 2020b). A sample set of 4920 patient records with diagnoses for 41 diseases were selected for analysis. The disease prediction system created utilising machine learning techniques including Decision Tree classifier, Random forest classifier, and Naive Bayes classifier is demonstrated in this research project (Grampurohit & Sagarnal, 2020b).

The medical records of 4920 patients who were at risk for 41 diseases due to a cluster of symptoms were used to train the algorithm. 95 out of 132 symptoms have been considered in order to avoid overfitting. Using the K fold cross validation methodology (K=5), the authors evaluated the efficacy of each strategy against the dataset. They came to the conclusion that all three approaches perform incredibly well on the dataset based on their findings. However, Nave Bayes may be slightly superior to the other two algorithms in terms of performance (Grampurohit & Sagarnal, 2020b). It is evident from the historical development of machine learning and its uses in the medical field that methodologies and techniques have developed that make it possible to analyse complex data in an easy and basic manner. The effectiveness of three algorithms on a medical record is carefully compared in this study. Results from each algorithm were up to 95% accurate (Grampurohit & Sagarnal, 2020b).

The article, “Symptoms Based Disease Prediction Using Decision Tree and Electronic Health Record Analysis” shows that in order to distinguish between the patient's illness and more common symptoms, health professionals conduct far too many disease surveys and collect information on patients, the severity of their illness, and their symptoms (Radhika et al., 2020). As a result, useful information hidden in the data set is used to train the model that predicts the disease based on the symptoms. The authors constructed the decision tree classifier model, which is trained using the dataset in a shorter amount of time, after normalising the data using case gradient descent standardisation techniques. After normalisation, their trained model is utilised to predict the disease, along with the degree of confidence, underlying causes, and suggested preventive measures (Radhika et al., 2020).

On the other hand, their technology takes a EHR (Electronic Health Records) as input and converts it into a text file (Radhika et al., 2020). The text file is summarised using the NLTK (Natural Language Toolkit) to help the patient understand the health report.They used health record analysis in this article, which provides customised input and user interaction with the system, to increase confidence (Radhika et al., 2020). To make this system a fully functional health monitoring system, extra modules are offered in addition to the two primary modules (Radhika et al., 2020). These modules include functions for locating doctors, planning doctor visits, and saving and retrieving medical records. The entire system consists of two modules, one for disease prediction and the other for health monitoring. The second module is utilised as one of the training data sets to get the best outcome and boost confidence. In order to collect the user input and provide status updates, this procedure is linked to the user interface. Because the files users supply as input may contain native languages, multilingual summarization makes it straightforward to quickly summarise health records from many parts of the world (Radhika et al., 2020). Currently, only English is used to write the document. This article defines disease prediction and examines some of the associated tasks, such as appointment scheduling and locating the closest medical institution, in addition to defining disease prediction utilising highly personalised training data sets (Radhika et al., 2020).

Disorders affecting the heart, kidneys, breast and brain are the focus of machine learning models (KNN, NB, DT, CNN, SVM, and LR) to assess for disease detection in the article “Disease Prediction Using Machine Learning”. The goal of their study is to investigate the idea that supervised machine learning algorithms can improve healthcare by quickly and correctly identifying diseases (Ferjani, 2020). Accuracy was the most important performance metric, and the three most frequently used prediction algorithms in the literature were SVM, RF, and LR. The CNN model performed the best at predicting common diseases (Ferjani, 2020). Additionally, the SVM model constantly showed greater accuracy for kidney diseases and PD because of its dependability in handling high-dimensional, semi-structured, and unstructured data (Ferjani, 2020). RF showed advantage in the likelihood of correctly identifying the prediction of breast cancer because to its potential to scale efficiently for large datasets and susceptibility to prevent overfitting. To finish, the LR algorithm was the most accurate at predicting heart illnesses (Ferjani, 2020).

OVERVIEW OF THE DATASET

The dataset used in this project has the following attributes:

- The size of the dataset is 133 columns and 4920 rows. The dataset has been bifurcated into training and testing data. The model will be trained through training dataset and the result will be tested on testing dataset. In this dataset, the 133 columns are split as follows:
    - 132 columns have continuous discrete binary values (Symptoms wherein 0 stands for no symptom and 1 stands for presence of symptom in the patient.)

    - 1 column is qualitative discrete categorical (Prognosis of disease consisting of 41 diseases such as Dengue, Diabetes, Heart attack etc.)

DATA APPROACH

Python programming is being used in this project. Python Programming Code for the Exploratory Data Analysis can be found at: https://github.com/DA-CIND/CIND820

1. DATA COLLECTION AND DATA PREPARATION

- The libraries imported are as follows: NumPy, Pandas, Matplotlib, Seaborn
- Importing the dataset
    - The dataset was gathered from kaggle.com and other websites listed below:
        - https://impact.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html
          (The New York Presbyterian Hospital provides a database of health symptoms that individuals experienced along with the corresponding prognosis admitted during 2004. (*Disease*, n.d.))
        - https://github.com/anujdutt9/Disease-Prediction-from-Symptoms
        - https://www.kaggle.com/datasets/neelima98/disease-prediction-using-machine-learning

- Uploading training data into the dataframe named as df with the help of pandas library.
- Identifying the dimensions of the data.
- Used df. describe() function to check the minimum and maximum values as well as mean and standard deviation.
- Used Numerical Analysis to shows data distribution at percentiles.
    - Results: No presence of outliers. Minimum value observed as 0 and maximum value is 1.
- Finding out the value counts of columns (Symptoms and prognosis)
    - Results: No. of unique symptoms are: 132
      No. of unique prognosis are: 41

- Note: Visualization of correlation was carried out through heatmap but it was not clear as there were 133 columns. So, correlation is not relevant here as meaningful conclusions could not be drawn from it.

2. DATA CLEANING

- Using isna() function, to check NAN values
    - Results: Unnamed133 has 4920 NAN values
- Dropping the NAN values through dropna() function

3. EXPLORATORY DATA ANALYSIS

- Creating a function named value_counts and using a loop to show the count of values (0,1) in all columns. A loop is used because the number of columns were 132 so to check the count of value the process needs to be repeated n number of times
    - Results:
        - Prognosis value_counts

Table 1: Prognosis and its count

Prognosis	Count	Prognosis	Count
Fungal infection	120	Hepatitis C	120
Hepatitis E	120	Alcoholic hepatitis	120
Tuberculosis	120	Common Cold	120
Pneumonia	120	Dimorphic hemmorhoids(piles)	120
Heart attack	120	Varicose veins	120
Hypothyroidism	120	Hyperthyroidism	120
Hypoglycemia	120	Osteoarthristis	120
Arthritis	120	(vertigo) Paroymsal Positional Vertigo	120
Acne	120	Urinary tract infection	120
Psoriasis	120	Hepatitis D	120
Hepatitis B	120	Allergy	120
hepatitis A	120	GERD	120
Chronic cholestasis	120	Drug Reaction	120
Peptic ulcer diseae	120	AIDS	120
Diabetes	120	Gastroenteritis	120
Bronchial Asthma	120	Hypertension	120
Migraine	120	Cervical spondylosis	120
Paralysis (brain hemorrhage)	120	Jaundice	120
Malaria	120	Chicken pox	120
Dengue	120	Typhoid	120
Impetigo	120		

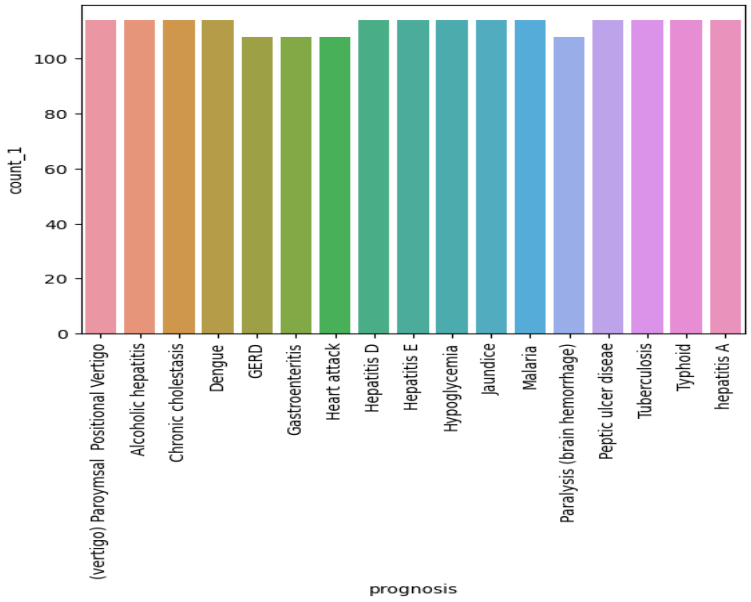
- Symptoms having frequency above 500 are:

Table 2: Symptoms and its count

Symptoms	Count	Symptoms	Count
Itching	678	yellowish_skin	912
skin_rash	786	dark urine	570
Chills	798	Nausea	1146
joint_pain	684	loss_of_appetite	1152
Vomiting	1914	abdominal_pain	1032
Fatigue	1932	Diarrhea	564
Cough	564	yellowing_of_eyes	816
high_fever	1362	Malaise	702
Sweating	678	chest_pain	696
Headache	1134		

- Univariate Analysis
  - Creating a function called 'analysis' for visualization of symptoms occurring in diseases through barplot
  - Using For Loop and If Else function in the dataframe through the following process: When the column is not equal to prognosis then the count of symptoms will be taken using groupby (filter) on prognosis
  - Next step is to create barplot using seaborn library and matplotlib to show the visualization of barplot
    - Results: Below is bar graph of symptom vs prognosis. Symptom is shown on the y axis whereas prognosis is shown on x axis.

Bar Graph 1: Vomiting (Symptom)



Analysis shows that there are some symptoms which are common to a lot of disease like:

- Vomiting is a common symptom in: (vertigo) Paroymsal Positional Vertigo, Alcoholic Hepatitis, Chronic Cholestasis, Dengue, GERD, (Hepatitis A,D,E), Malaria, Tuberculosis etc.
- Itching is a common symptom in Chicken Pox, Chronic Cholestasis, Drug Reaction, Fungal Infection, Hepatitis B, Jaundice.
- Skin Rash is common in Acne, Chicken Pox, Dengue, Drug Reaction, Fungal Infection, Impetigo, Psoriasis.
- Fatigue is a common symptom in: Chicken Pox, Common Cold, Diabetes, Dengue, Typhoid etc.

This shows that these symptoms are common in many diseases irrespective of the disease being communicable or non communicable. So, a pattern cannot be observed from this analysis.

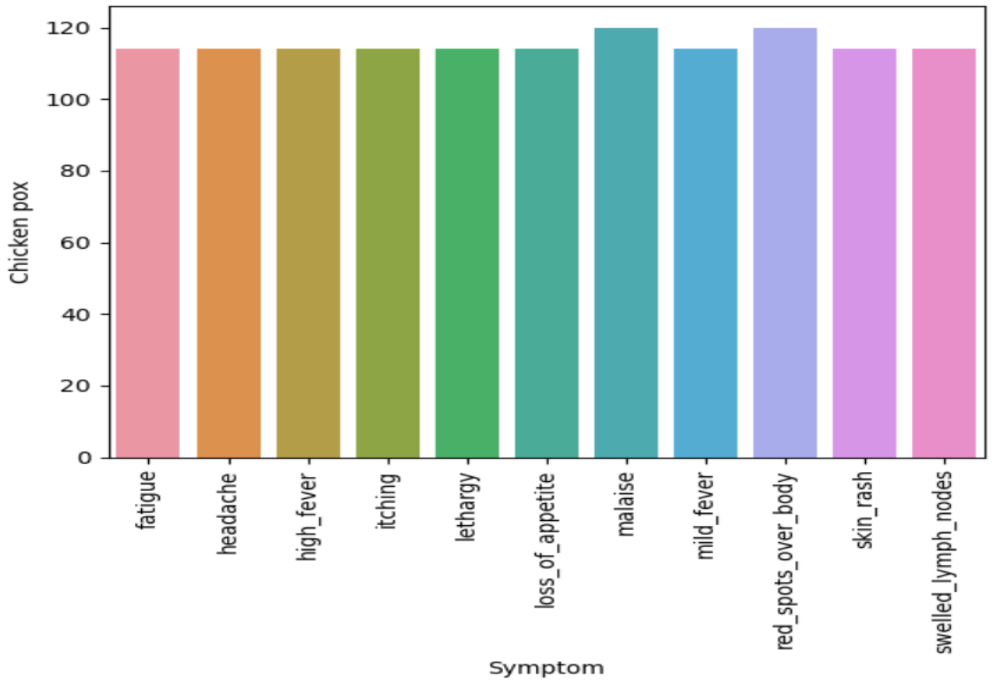
Now moving on to the analysis of the symptoms which are unique in some diseases which are as follows:

- Nodal_skin_eruptions is a symptom of fungal infection
- Continuous_Sneezing is a symptom of allergy and common cold
- Muscle_wasting is a symptom of AIDS
- Weight_gain, cold_hand_and_feet, enlarged_thyroid, swollen_extremeties are symptoms of Hypothyroidism
- Irregular_sugar_level is a symptom of diabetes
- Pain_behind_eyes is a symptom of dengue
- Constipation, pain_during_bowel_movements, pain_in_anal region, Bloody_stool are symptoms of Dimorphic hemorrhoids (piles)
- Fluid_overload and swelling_of_stomach are symptoms of Alcoholic Hepatitis

This shows that some symptoms are unique to a particular kind of disease as described above. Therefore, it can be concluded that from a unique symptom, early detection of disease can be predicted.

- Bivariate Analysis
    - Creating a pivot table in the data frame named as df_tidy_2 to show the count of symptoms in the diseases.
    - Creating function called bi_analysis for visualization of bivariate analysis between symptoms and prognosis.
    - Seaborn and Matplotlib libraries are used for visualization.
        - Results:
          Considering the Bar graph between Prognosis and symptoms. Prognosis is represented on y axis and Symptoms are represented on x axis.

Bar Graph 2: Chicken Pox (Prognosis)



Analysis shows the following:

- Paroymsal Positional Vertigo symptoms: headache, loss of balance, nausea, spinning mo vements, unsteadiness, vomiting.
- AIDS symptoms: extra_marital_contacts, high_fever, muscle_wasting, patches in throat
- Acne symptoms: blackheads, pus_filled_pimples, scurring, skin_rash
- Alcoholic hepatitis symptoms: abdominal_pain, distension_of_abdomin, fluid_overload1, history_of_alcoholic_consumption, swelling_of_stomach, vomiting_yellowish_skin
- Allergy symptoms: chills, continuous_sneezing, shivering, watering_from eyes
- Arthritis symptoms: movement_stiffness, muscle_weakness, painful_walking, stiff neck, swelling_joints
- Bronchial Asthma symptoms: breathlessness, cough, family_history, fatigue, high_fever, mucoid_sputum
- Cervical spondylosis symptoms: back_pain, dizziness, loss_of_balance, neck_pain, weak ness_in_limbs
- Chronic cholestasis symptoms: abdominal_pain, itching, loss_of_appetite, nausea, yellow ing_of_eyes, yellowish_skin
- Common Cold symptoms: chest_pain, chills, congestion, continuous_sneezing, cough, fat igue, headache, high_fever, losss_of smell, malaise, muscle_pain, phlegm, redness_of_ey es, runny_nose, sinus_pressure, swelled_lymph_nodes, throat_irritation

4. PROPOSED WORK

The aim of this project is to find appropriate machine learning models for disease prediction utilising a dataset of symptoms. The proposed models would use the symptom data to determine if a given disease is likely to exist or not. The objective of this research is to enhance early disease identification and detection, allowing for prompt intervention and treatment.

- **Feature Selection:** Next step will be application of feature engineering techniques. Feature encoding will be done for the 'Prognosis' column (categorical variables) and numerical features will be normalized. Feature engineering techniques will be applied to extract relevant information from the symptom data, potentially including feature selection or dimensionality reduction.
- **Model Selection:** Evaluation and comparison of various machine learning models appropriate for classification tasks. Models such as Logistic Regression, Random forests, KNN will be considered. The selection will be based on the dataset characteristics, model complexity and performance metrics.
- **Model Training and Validation:** Train and validate the selected models. Using appropriate techniques, such as cross-validation, to ensure robust evaluation of model performance. To improve performance, adjust the model hyper parameters using methods like grid search or random search.
- **Evaluation Metrics:** Using suitable evaluation metrics in the context of disease prediction. Common metrics include accuracy, precision, recall, F1 score etc.
- **Model Comparison and Selection:** Compare the performance of different models using the chosen evaluation metrics. Analyze their strengths, weaknesses, and limitation. Select the model(s) that achieve the highest predictive accuracy and meet the project's requirements.

Below is graphical representation of the process to be followed next in this project:

Flowchart 1: Illustration of the process and next steps

5. METHODOLOGY

Machine Learning algorithms proposed for the purpose of this project are as follows:

5.1 LOGISTIC REGRESSION

A popular technique for supervised classification, logistic regression is particularly made for binary classification issues when the target variable denotes the presence or absence of a disease (Uddin et al., 2019). It may be viewed as an extension of an ordinary regression and can only model a dichotomous variable, which often indicates whether an event will occur or not. It aids in determining the likelihood that a brand-new instance belongs to a particular class (Uddin et al., 2019). Given that the result represents a probability, it falls between 0 and 1. As a result, a threshold must be set to separate between two classes in order to utilise it as a binary classifier. For instance, an input instance is classified as "class A" if its probability value is greater than 0.50; otherwise, it is classified as "class B." (Uddin et al., 2019).

5.2 RANDOM FOREST

An ensemble classifier made up of several decision trees is known as a random forest (Uddin et al., 2019b). The training data are usually overfitted when decision trees are grown very deeply, which causes a considerable variation in classification outcomes for a little change in the input data (Uddin et al., 2019b). They are extremely sensitive to their training data, which makes them prone to mistakes on the test dataset. The training dataset's multiple decision trees are learnt using a variety of examples. The input vector of a new sample must pass down with each decision tree of the forest in order to be classified. Each decision tree then delivers a classification result while taking into account a different component of the input vector (Uddin et al., 2019b). The forest then decides on the classification for a numeric classification outcome by taking the average of all the trees in the forest or for a discrete classification outcome by taking the most "votes". The random forest algorithm can decrease the variance caused by the consideration of a single DT for the same dataset since it takes results from several distinct decision tree into account (Uddin et al., 2019b).

5.3 K NEAREST NEIGHBOR (KNN)

The supervised machine learning method k-nearest neighbor (KNN) is primarily used for classification tasks (Uddin et al., 2022b). Using the characteristics and labels of the training data, the KNN forecasts the categorization of unlabelled data. The k nearest training data points (neighbors), which are the ones closest to the testing question, are often used by the KNN method to classify datasets using a training model comparable to the testing query (Uddin et al., 2022b). It then applies a majority vote mechanism to determine which classification should be finalised. Because of its easy adaptability and simple to comprehend design, the KNN algorithm is well known for its application in regression and classification problems with data of various sizes, label numbers, noise levels, ranges, and contexts (Uddin et al., 2022b).

This algorithm’s working is simple and offers the flexibility to be changed in a variety of ways to minimise its weaknesses and challenges, enhance its accuracy, and make it applicable to a wider range of datasets (Uddin et al., 2022b). The typical KNN method has a number of shortcomings that limit its capacity to classify data, such as being unbiased to all of its classification-dependent

neighbours, missing features for computing distances between data points, and accounting for useless dataset elements (Uddin et al., 2022b). However, KNN may be changed in several ways, resulting in a number of KNN forms or variants. The KNN versions differ from one another in a number of computational areas, such as optimising the k parameter, increasing distance calculations, assigning weight to different data points, and reducing training datasets to solve issues (Uddin et al., 2022b).

REFERENCES

Mahata, S., Kapadiya, Y. B., Kushwaha, V., Joshi, V., & Farooqui, Y. (2023). Disease Prediction and Treatment Recommendation Using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, *11*(3), 1232–1237. https://doi.org/10.22214/ijraset.2023.49641

Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2021). Multiple disease prediction using Machine learning algorithms. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.07.361

Kolte, A., Mahitha, B., & Raju, N. V. G. (2019). *Stratification of Parkinson Disease using python scikit-learn ML library*. https://doi.org/10.1109/icese46178.2019.9194627

Grampurohit, S., & Sagarnal, C. (2020b). Disease Prediction using Machine Learning Algorithms. In *2020 International Conference for Emerging Technology (INCET)*. https://doi.org/10.1109/incet49848.2020.9154130

Radhika, S., Shree, S. R., Divyadharsini, V. R., & Ranjitha, A. (2020). *Symptoms Based Disease Prediction Using Decision Tree and Electronic Health Record Analysis*. European Journal of Molecular & Clinical Medicine. https://ejmcm.com/article_1944_cb7aaa34894c921618817c5c40cdaf5d.pdf

Ferjani, M. (2020, December 16). *Disease prediction using machine learning* . Research Gate. https://www.researchgate.net/profile/Marouane-Ferjani/publication/347381005_Disease_Prediction_Using_Machine_Learning/links/5fda5556299bf1408816daa4/Disease-Prediction-Using-Machine-Learning.pdf

Disease. (n.d.). https://impact.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html

Uddin, S., Khan, A. O., Hossain, E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-1004-8

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022b). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-10358-x