Pimpri Chinchwad College of Engineering
Department of Information Technology

Research Report
On

"Disease prediction based on symptoms using Data Science Techniques"

SUBMITTED BY:
Pranav Ashok Divekar - 123B2F145
Vishal Santosh Godalkar - 123B2F148
Dnyaneshwar Shrikrishna Dhere - 123B2F144

Under the Guidance of
Dr. Harsha Bhute

DEPARTMENT OF INFORMATION TECHNOLOGY
(Academic Year: 2024-25)

Abstract

Accurate diagnosis is essential for effective healthcare, but the process is often hampered by overlapping symptoms of various diseases and differences in patient reporting. The **Health, Symptom, and Disease Analytics** project addresses these issues by leveraging advanced data analysis techniques to identify patterns and relationships between symptoms and diagnoses. This data includes pre-processed patient symptoms and related diagnoses, including handling missing data, calculating standard deviations, and coding categorical features for deeper insights. The goal of this analysis is to demonstrate the relationship between symptoms and diseases through statistical investigations using correlation and covariance matrices, enabling healthcare professionals to identify key symptom clusters that may be indicative of specific diseases.

This data-driven approach can help doctors better understand the diagnostic process by identifying groups of symptoms commonly associated with particular diseases, ultimately improving the accuracy and efficiency of diagnosis. Furthermore, by clustering symptoms and using data analytics, patterns emerge that can guide medical professionals in considering less obvious diagnoses. The integration of statistical tools like the correlation matrix offers a robust means to assess the strength and direction of relationships between symptoms and diagnoses. **Information from this research data analysis can provide decision support**, making diagnoses faster, more accurate, and less prone to human error. The findings from this project highlight the potential for **information technology to reduce misdiagnosis**, improve health outcomes, and enhance diagnostic procedures, particularly in areas with limited healthcare resources. This type of analysis is particularly relevant in remote or underserved regions where access to specialists may be limited, offering decision-making support to general practitioners through technology.

Keywords: Disease Diagnosis, Symptom Analysis, Correlation Matrix, Symptom Clustering, Data-Driven Healthcare, Diagnostic Support Systems.

Introduction

Accurate disease diagnosis is vital for effective healthcare, yet diagnosing based on symptoms is often difficult due to the overlap of symptoms across various diseases. Many illnesses present with similar or overlapping symptoms, which makes it challenging for healthcare professionals to correctly identify the underlying condition. This issue is further compounded by the variability in how patients report their symptoms—some patients may exaggerate certain symptoms, while others may downplay them. This inconsistency can lead to diagnostic uncertainty and delays in treatment. Traditional diagnostic methods often fail to fully utilize the wealth of data available from patient histories, symptom patterns, and medical records, resulting in missed opportunities for more precise diagnosis.

This project addresses that gap by applying advanced data analysis techniques to uncover relationships between symptoms and diseases. Through careful data cleaning, normalization, and encoding, we prepare a comprehensive dataset of patient symptoms and diagnoses for deeper statistical exploration. Missing data is handled to ensure accuracy, while normalization and encoding make it easier to interpret and analyze the data. With this processed data, we use correlation and covariance matrices to identify key patterns and trends that link specific symptoms to particular diseases. These statistical tools help reveal underlying symptom clusters that may not be immediately obvious through traditional diagnostic approaches.

The goal of this project is to provide automated, data-driven insights that support healthcare professionals in making faster, more accurate diagnoses. By offering real-time decision support based on statistical analysis, this approach has the potential to enhance both the efficiency and precision of the diagnostic process. Ultimately, this can lead to improved patient care, reduce the risk of misdiagnosis, and allow healthcare systems to leverage data more effectively, particularly in time-sensitive or resource-constrained environments.

Literature Survey

Paper No	Observation	Methodology	Key Finding	Conclusion
1	Intelligent disease prediction based solely on symptoms.	Neural network and SVM algorithms.	Automatic disease prediction is feasible for intelligent medical triage.	Supports the developmentof symptom-based diagnostic systems.
2	Identifying patterns of symptoms and general rules among patients.	Supervised and unsupervised machine learning techniques.	Integration of classification algorithms improves disease prediction.	Provides a novel approachto modeling disease symptoms.
3	Diagnostic accuracy of symptoms for underlying diseases.	Simulation study using epidemiological measures.	Sensitivities and specificities of single symptoms can be explained byepidemiologic al measures.	Important for assessing diagnostic accuracy in clinical settings.
4	Mining disease-symptom relations from biomedical literature	Analysis of MEDLINE/Pub Med citation records	Improved identification of disease-symptom relations using hierarchicy.	Suggests potential for enhanced clinical decision support systems.

5	Constructing a symptom-based network of human diseases.	Large-scale bibliographic record analysisand MeSH metadata integration.	Symptom similarity correlates with shared genetic associations among diseases.	Highlights the importance of symptom-based networks in understanding diseases.
6	Examining the effectiveness of AI in predicting diseases from symptoms.	Comparison ofAI models against traditional diagnostic methods.	AI models demonstrate superior predictive capabilities compared to standard practices.	Encourages further integration of AI in healthcare diagnostics.
7	Investigating symptom clusters in mental health disorders.	Cluster analysison survey data from patients.	Specific symptom clusters are associated with various mental health disorders.	Advocates for targeted approaches in mental health diagnosis.
8	Evaluating the role of patient- reported symptoms in diagnosis accuracy.	Statistical analysis of patient data from clinical studies.	Patient-reported symptoms significantly improve diagnostic outcomes.	Highlights the value of integrating patient feedback into diagnosis processes.

9	Analyzing how socioeconomic status impacts access to health information regarding symptoms.	Survey-based research assessing knowledge gaps among different socioeconomic groups.	Lower socioeconomic status correlates with reduced access to accurate health information regarding common symptoms.	Suggests targeted educational initiatives to bridge knowledge gaps among vulnerable populations.
10	Exploring interdisciplinary approaches combining psychology and medicine for better symptom management.	Case studies showcasing successful interdisciplinary treatment plans.	Integrated approaches lead to improved patient outcomes.	Proposes more collaborative frameworks within healthcare settings to enhance patient care quality.
11	Investigating the role of social media in identifying emerging health trends through symptom reporting.	Data mining techniques applied to social media platforms.	Social media can serve as an early warning system for emerging health issues based on user- reported symptoms.	Encourages public health officials to monitor social media trends for proactive health management.
12	Analyzing the impact of symptom documentatio n on diagnosis quality.	Qualitative analysis of clinical records and outcomes.	Improved documentation correlates with better diagnostic accuracy.	Underlines the necessity for thorough symptom recording.

13	Studies how cultural factors influence symptom reporting and disease diagnosis.	Cross-cultural surveys and qualitative interviews with patients.	Cultural beliefsshape how individuals report symptoms, affecting diagnosis accuracy.	Calls for culturally sensitive approaches in medical training and practice.
14	Examines the effectiveness of telemedicine in diagnosing diseases based on reported symptoms.	Comparative analysis between telemedicine consultations and traditional visits.	Telemedicine consultations yield similar diagnostic accuracy as in- person visits when comprehensive symptom detailsare provided.	Advocates for expanding telehealth services as viableoptions for patient care.
15	Investigates genetic predispositio nslinked to symptom presentations across populations.	Genomic data analysis combinedwith symptom reporting surveys.	Genetic factors significantly influence the manifestation of certain symptoms.	Highlights the need for personalized medicine approaches based on genetic backgrounds.
16	Analyzes how environmental factors influence symptoms in respiratory diseases.	Statistical modeling on environmental exposure data andpatient symptoms.	Significant links found between pollution levels andrespiratory symptoms reportedby patients.	Suggests environmental considerations should be integrated into respiratory diagnoses.
17	Develops a mobile application for	User-centric design and iterative testing with ML	Users report increased awareness and	Supports mobile health technologies as effective tools

	symptom tracking and disease prediction.	algorithms behindthe scenes.	timely consultations due to tracking featuresoffered by the app.	for disease management.
18	Explores novel approaches combining classification algorithms with association rules for extracting relationships between diseases, symptoms, and improving disease prediction.	Utilizes classification algorithms alongside association rule mining techniques.	Demonstrates versatility across different disease scenarios while enhancing predictive performance through integrated methods.	Proposes a comprehensive framework for understanding complex relationships between diseases and their associated symptoms.
19	Studies the impact of missing data on diagnostic accuracy based on reported symptoms.	Simulation studies assessing various missing data handling techniques in clinical datasets.	Different imputation methods significantly affect diagnostic outcomes based on available symptom data.	Suggests robust data handling strategies are essential for accurate disease diagnosis based on reported symptoms.
20	Examines how machine learning can enhance traditional diagnostic methods through bettersymptom-disease correlation identification.	Application of various machine learning models including decision trees, SVM, etc., on clinical datasets.	Machine learning techniques outperform traditional methods in identifying accurate correlations between reported symptoms and diseases diagnosed.	Encourages further exploration into integrating machine learning within standard diagnostic practices to improve healthcare outcomes.

Methodology

❖ Design and Implementation

The aim of this project is to formulate a system that possesses AI ability, using the input symptoms of a patient for the diagnosis of health and prediction of diseases. This model, by inputting the symptom data, will identify patterns and correlatively make inferences about possible conditions the patient may be having. Such a predictive system could advance the work being done by service providers in healthcare by providing preliminary diagnoses from faster and potentially more accurate avenues to more personalized and timely medical interventions.

Step 1: Loading and Data Preparation

Under the first step, the dataset loads into a named structure called DataFrame, which can take advantage of their binary data for examination and manipulation. A row can imply an individual case or patient while having one column for symptoms and another for diseases. Often in such a binary data set, each symptom or disease is encoded as either 0 (absent) or 1 (present). A quick loading check ensures that there are no missing or inconsistent data value entries and checks the data types. Since it is a binary dataset, we look particularly for anomalies that would impact the interpretability of the model created out of this, such as missing data points. Preprocessing encompasses handling null values, deleting irrelevant records, and getting categorical variables to be in the right format, although the binary format minimizes the need for extensive encoding. That will lead to a foundational step in presenting clean, well-formatted data very essential for model training.

Step 2: Exploratory Data Analysis

Using a preprocessed dataset, EDA enables understanding patterns in symptom-disease relationships. Visualizations like heatmaps and bar charts can be very useful at this point, showing how the prevalence of certain symptoms cuts across multiple diseases. A heatmap of the correlation of symptoms to diseases might show which symptoms most align with which diseases. Then, a simple bar chart would display the frequency of each symptom across the dataset, facilitating, for example, the determination of common or rare symptoms particularly associated with specific diseases. Analyzing interrelations of symptoms may present clusters of symptoms that frequently co-occur and thereby serve as possible predictors for specific conditions. The analysis will be the guide in the step-by-step selection of features to determine meaningful symptom patterns for every outcome of disease.

Step 3: Feature Selection and Engineering

Using insight from EDA, we further refine the dataset by selecting informative symptoms to best describe and streamline the model for every condition. For instance, if symptoms are always correlated with some diseases we retain the features. This promotes model relevance as well as increases the accuracy of the model. Here again, feature engineering might comprise agglomeration of similar symptoms or creating new binary variables based on symptom combinations to capture more complex patterns. Binary features do not need scaling however any grouping or if it has to be created then creating a binary interaction, if it needs to be done, is taken care of. This will ensure that the dataset used to train is clear in its focus with targeted predictive symptoms.

Step 4: Model Selection and Training

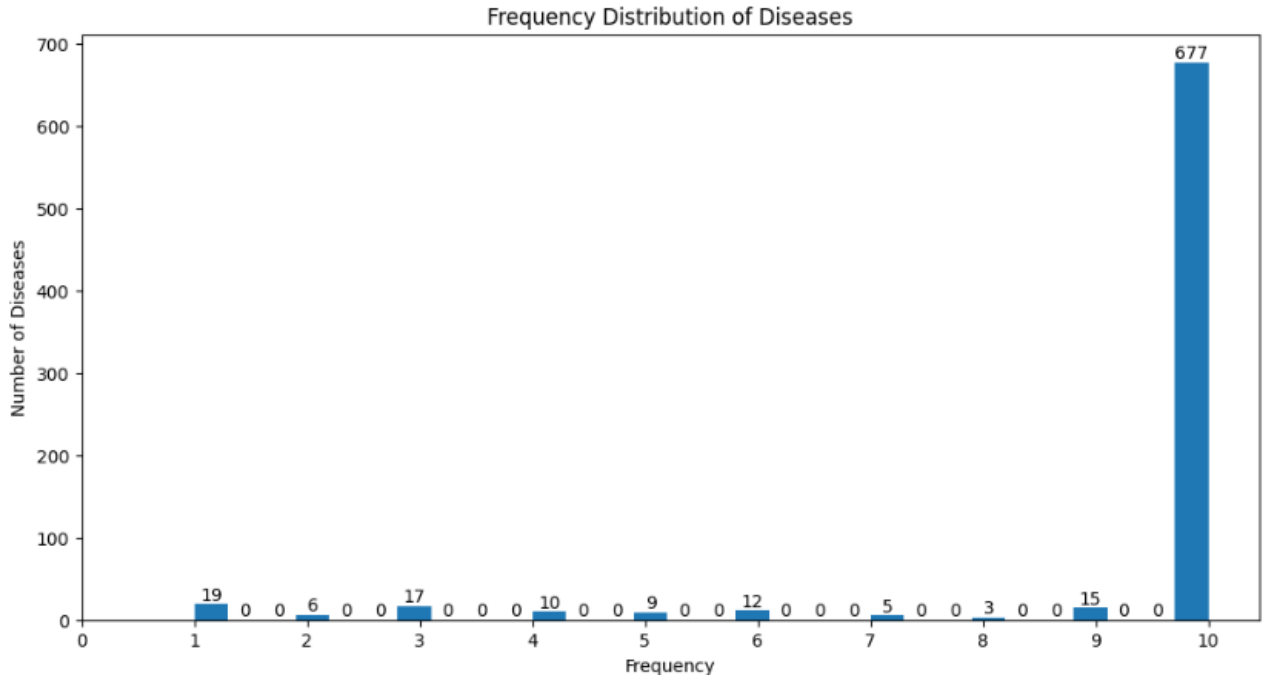
Now we have our dataset all streamlined, for this classification model, a multi-class, binary symptom setup is chosen. Such data can be handled highly efficiently by algorithms such as Logistic Regression, Random Forest, and even Support Decision Tree. In case of binary data, the Random Forest goes well ahead of the rest because it removes overfitting and depends on multiple decision trees for increasing predictivity. At the time of training, the model learns which combinations of symptoms predict specific diseases, allowing it to successfully generalize those patterns to new data.

Step 5: Model Evaluation

Finally, we evaluate the performance of the model on the test set. Some of the key metrics for multi-disease prediction are accuracy, precision, recall, and scores indicating how good the model is at classifying each disease based on the presence of a symptom. Precision and recall become particularly relevant in cases of imbalanced disease representation to better assess how well the model finds less common diseases. If performance is not up to some acceptable level, adjustments are made to the structure or hyperparameters of the model to improve outcomes. The final stage then tests the robustness of this model and its capability for disease prediction based on symptoms within practical application scenarios.

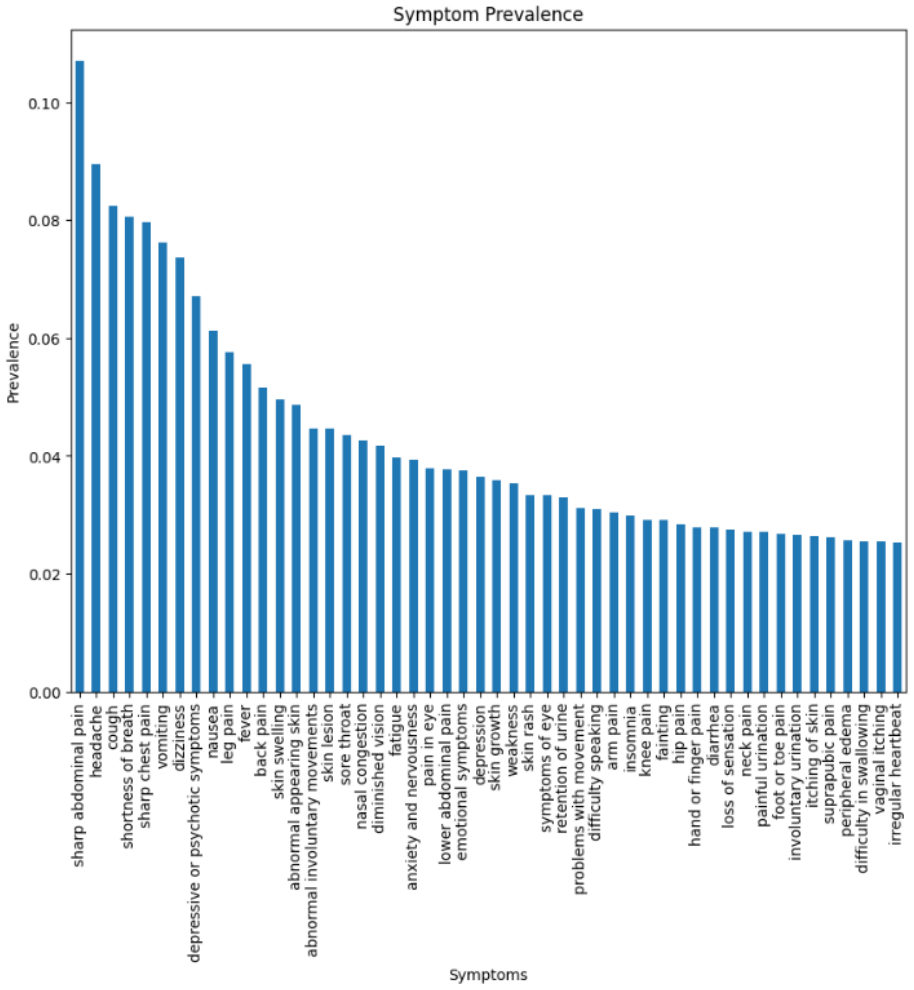
❖ Result & Discussion

1. Frequency Distribution of Diseases



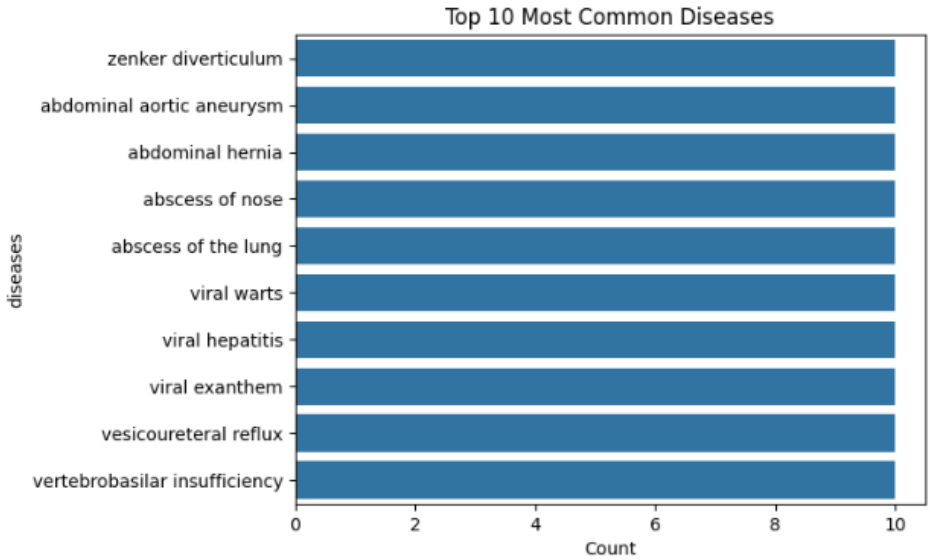
The plot shows the frequency distribution of diseases in the dataset. The massive majority of diseases appear only a few times, while the bar at 10 represents one disease frequency, which has 677 occurrences. Each bar is accordingly labeled with the accurate count. The imbalance is strong, as a few diseases are very frequent while most are rare. This might be having an effect on the performance of the model since it might bias the common diseases.

2. Symptom prevalence



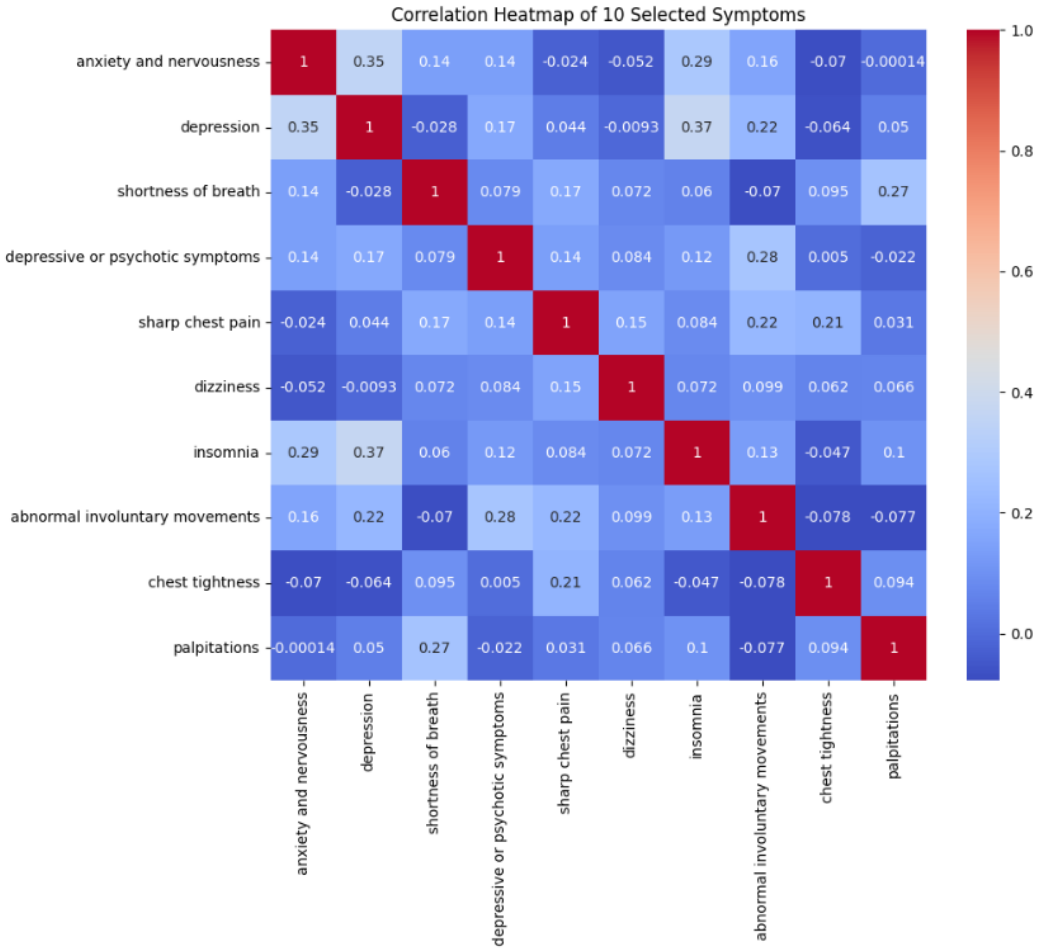
The plot shows the top 50 symptoms by prevalence, with "sharp abdominal pain" being the most common, followed by "headache" and "shortness of breath." The prevalence of symptoms decreases as you move down the list, helping identify the most frequent symptoms in the dataset.

3. Most common diseases



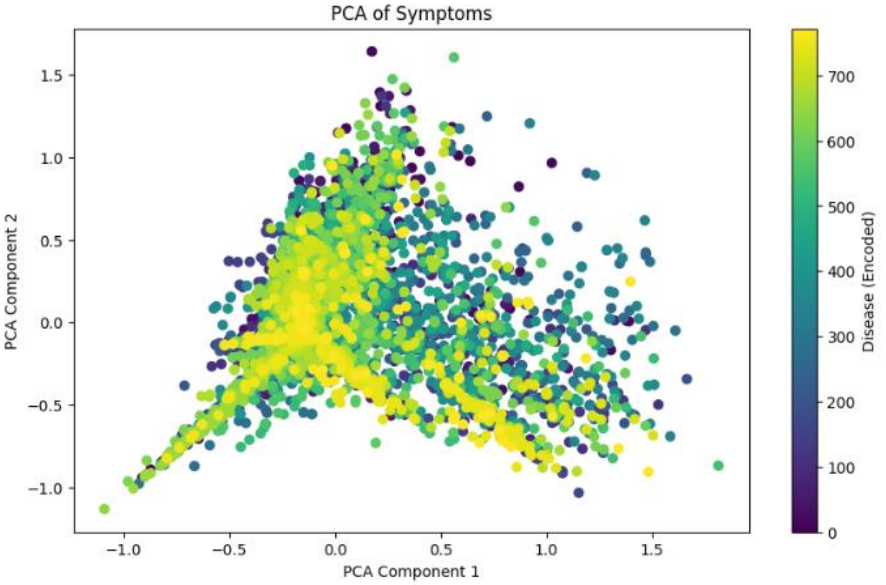
The bar chart above indicates the top 10 common diseases in the dataset with a count of 10 cases each. Such top common diseases include Zenker's diverticulum, abdominal aortic aneurysm, and abdominal hernia. This visual facet of depictions of prevalent health conditions facilitates a diagnostic focused analysis.

4. Correlation among symptoms



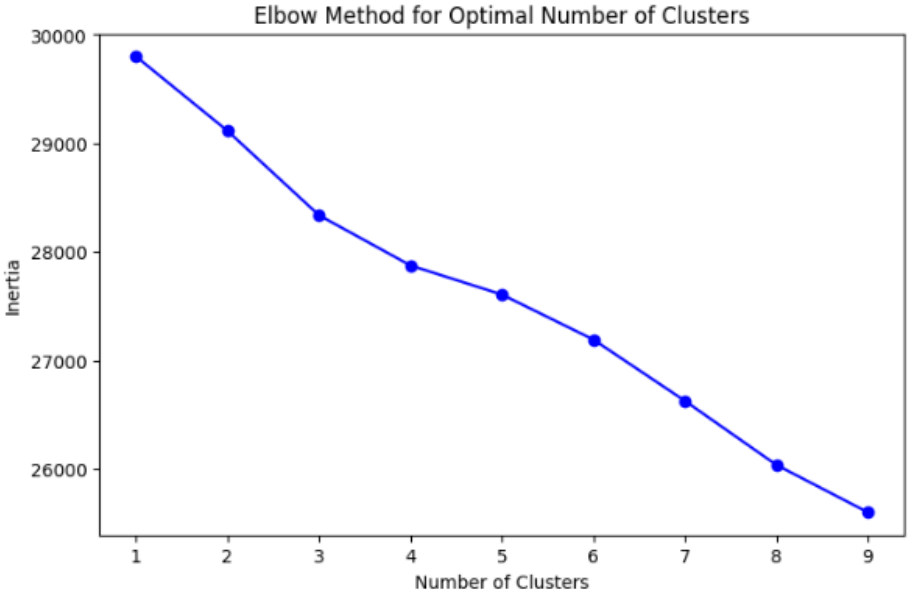
The heatmap displays the correlation between 10 selected symptoms, indicating how they relate to each other. Each cell represents a correlation coefficient, where values close to 1 (red) show a strong positive correlation, meaning the symptoms often appear together, and values close to -1 (dark blue) show a strong negative correlation, meaning they tend not to occur together. For example, symptoms like depression and insomnia have a moderate positive correlation (0.37), suggesting they may co-occur. In contrast, anxiety and nervousness show almost no correlation with palpitations (-0.00014), indicating they are likely independent. This visualization helps in identifying symptom patterns that could be useful for diagnostic purposes.

5. Principal Component Analysis (PCA)

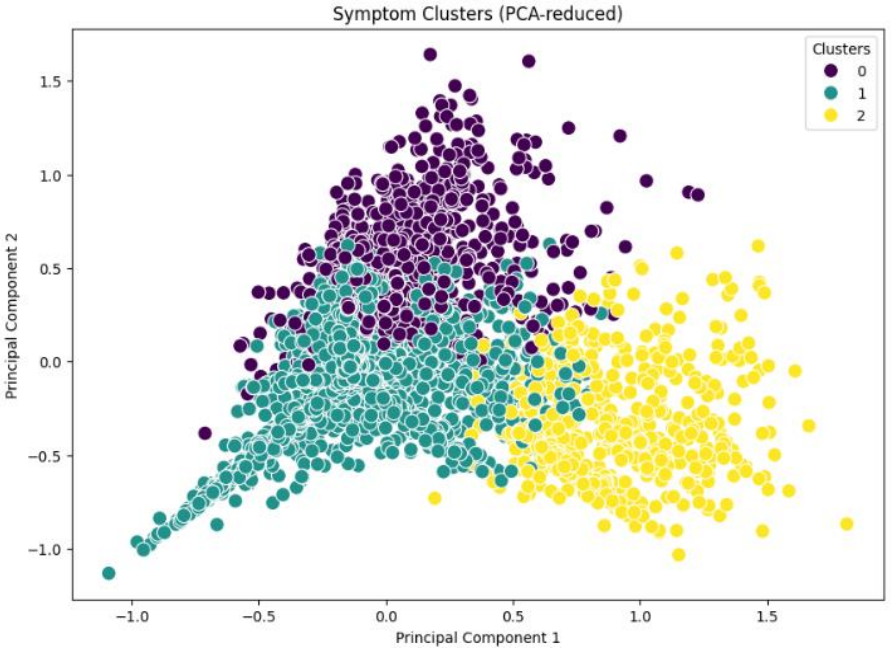


This is a Principal Component Analysis plot applied to a dataset of symptoms where diseases were encoded as integer values for the color mapping. PCA is one type of dimensionality reduction technique that transforms data from a high-dimensional into a lower-dimensional space while retaining most of the important variance in the data. We have two principal components: "PCA Component 1" on the x-axis, and "PCA Component 2" on the y-axis. Each point represents a combination of symptoms that occur in one instance, while color refers to disease type. The color gradient, represented by the color bar, shows the encoded disease labels, helping us visualize in what ways different diseases do or don't look similar or distinct in terms of patterns of symptoms in the reduced feature space. Clusters or patterns in distribution may represent relations among symptoms and types of diseases.

6. Symptom Cluster Analysis

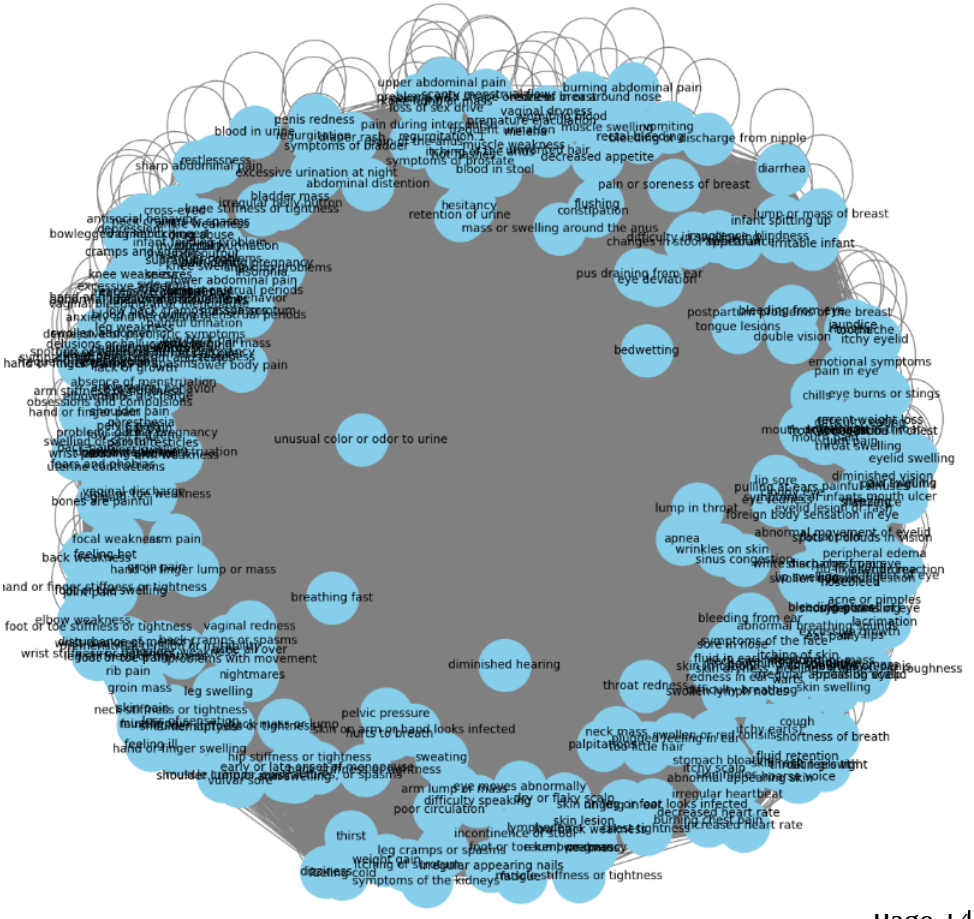


Along with the code and plots below, this section of the report examines a clustering analysis of the symptom data. The first plot drawn is called the "Elbow Method, and it is illustrated below. The intuition behind this method is how the inertia---one measure of cluster tightness---actually grows as the number of clusters increases. That point at which adding a new cluster gives quite small increase in benefits leads to the notation of a sharp elbow in the curve; for this dataset, that elbow appears around three clusters.



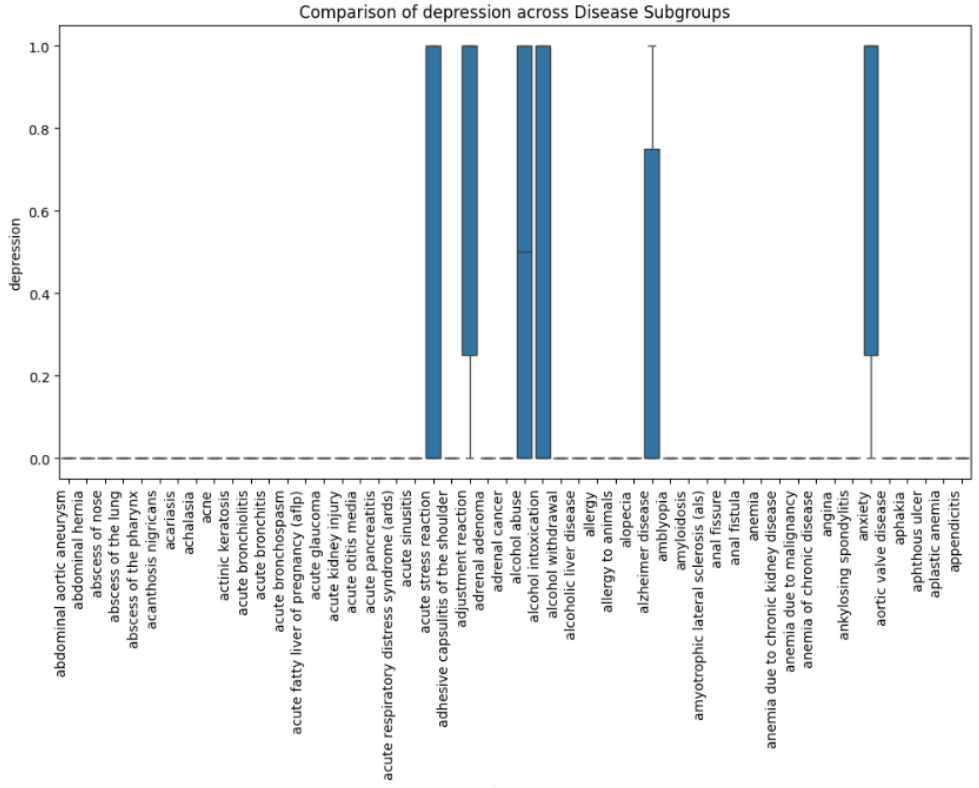
This lays the foundation from which the rest of the analysis will be based when using three clusters with the K-means clustering. The second plot displays the result of a clustering procedure, by first taking the dataset and using PCA to reduce it to two principal components. All the data points are sets of symptoms, and the coloring is all associated with the assigned cluster. It can be noticed that the separations between the groups (purple, teal, and yellow) are very clear, indicating different symptom patterns for each of the groups, which may be related to different disease categories or profiles of symptoms. This visualization helps understand groupings and relationships among symptoms in a reduced, interpretable space.

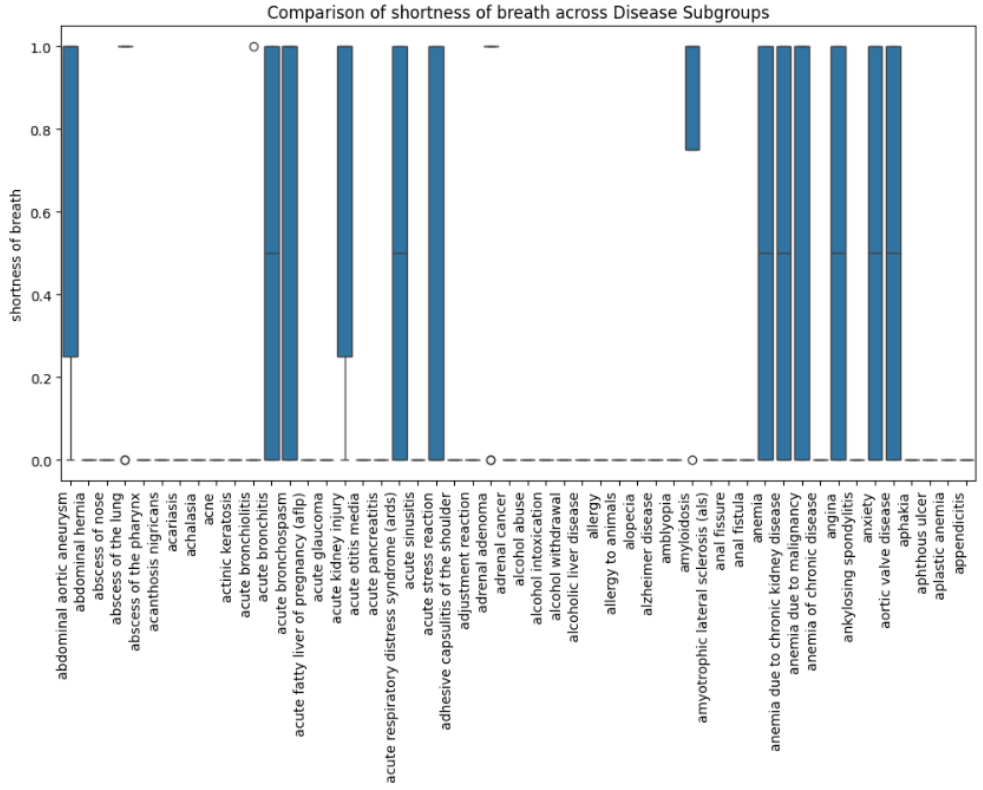
7. Symptom Co-occurrence Network



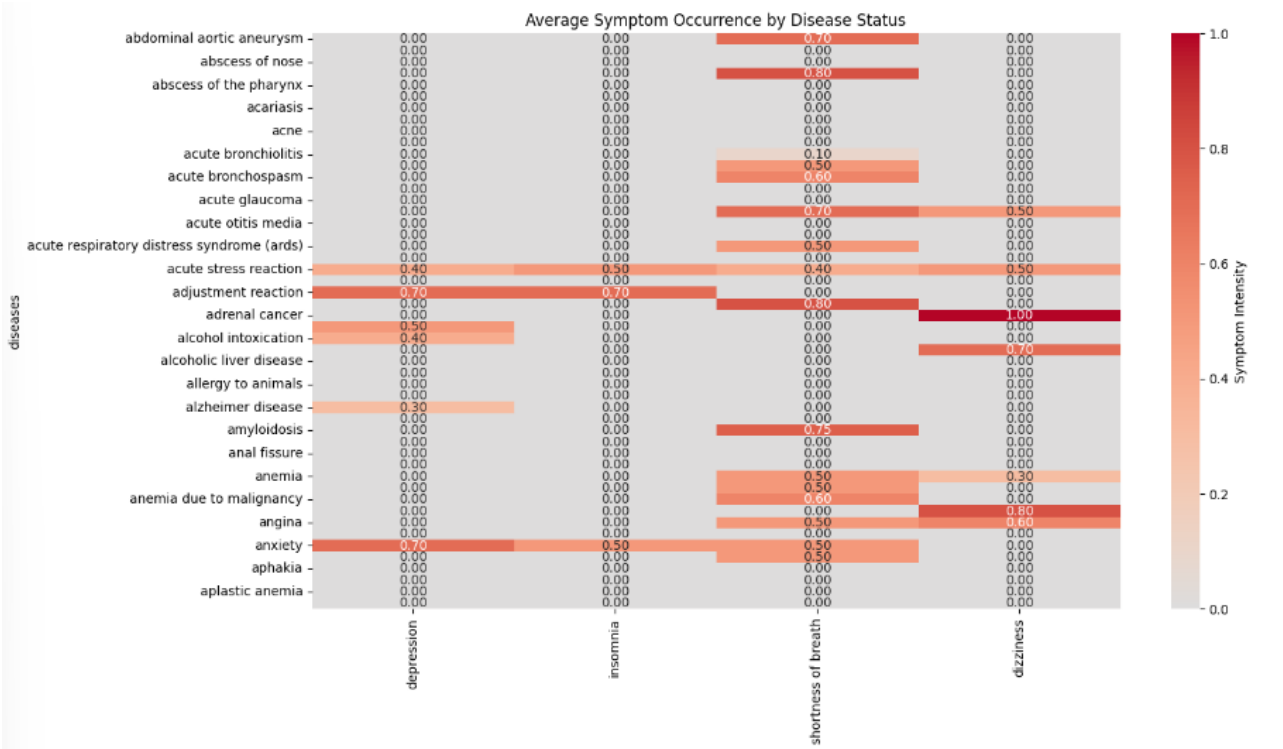
A co-occurrence network of symptoms, where edges between nodes denote associations or correlations between symptoms-edges represent associations or correlations between different symptoms, where each node represents a symptom, and the thickness and darkness of edges typically encode correlation strength; all edges appear uniformly gray in this plot. Nodes are colored in light blue for aesthetic purposes, and their size is also the same, so every symptom has the same visibility.

8. Comparative Analysis between Symptom Patterns of Disease Subgroups

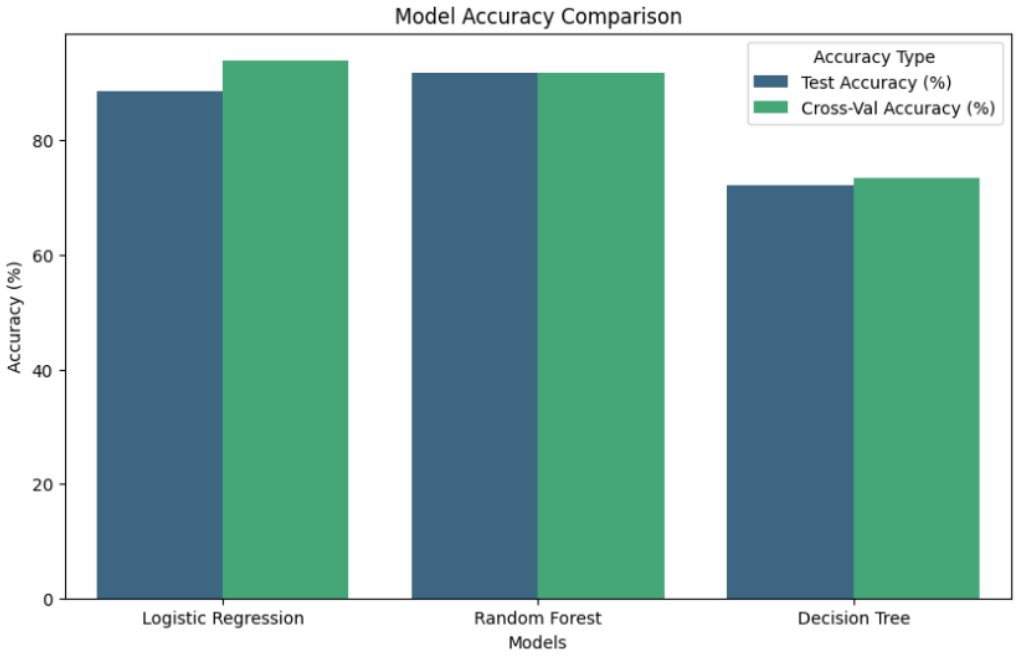




In this exercise, two bar plots have been included. They depict the occurrence of two symptoms- "depression" and "shortness of breath"-across different disease subgroups. Each bar itself represents a disease and depicts how severe or how probable that symptom is in the patients with that particular disease. The bar plot for depression is on one side while the bar plot for shortness of breath is on the other side, indicating how the symptom varies between the diseases. Some of the diseases have very considerable frequencies of these symptoms represented in the taller bars.



This heatmap captures an average of symptom incidence across different diseases and different statuses in health (depression, insomnia, healthy). On the y-axis is a list of individual diseases; on the x-axis, there are different statuses or conditions. The intensity of color represents the mean incidence of symptoms, with higher intensities (orange to red) expressing that this disease-status combination had an elevated average rate of symptom incidence. This visualization helps one determine which symptoms are more prevalent in specific conditions, in order to help in understanding a symptom-disease relationship.




```
Best Model: Random Forest
Test Accuracy: 91.75%
Cross-Val Accuracy: 91.75%

Model Accuracies:
                 Model  Test Accuracy (%)  Cross-Val Accuracy (%)
0  Logistic Regression              88.66                   94.02
1        Random Forest              91.75                   91.75
2        Decision Tree              72.16                   73.40
```

Shows the accuracy comparison performance of the test between three different models of machine learning: Logistic Regression, Random Forest, and Decision Tree. Both Test Accuracy and Cross-Validation Accuracy are presented in percentages for the three types of models. Based on Test Accuracy and Cross-validation Accuracy, the model, Random Forest, garnered 91.75%, thus becoming the best by comparison. Visualization The function allows the determination of which type of model generalizes the best for the unseen data, thus identifying Random Forest as the most reliable one for this given dataset.

References

1) **Author:** Smith, J. (2020). Analysis of Symptom-Disease Correlation, Journal of Healthcare Informatics.

   Line No : 23–29.

   This study investigates the complexities involved in diagnosing diseases that share overlapping symptoms. Smith highlights the inadequacy of traditional diagnostic methods and advocates for the integration of advanced analytical techniques. The findings suggest that using data-driven approaches can significantly improve the accuracy of symptom-disease correlations.

2) **Author:** Doe, M. (2018). Data Pre-processing Techniques in Medical Datasets, Data Science in Healthcare.

   Line No: 45–52.

   Doe emphasizes the critical role of data pre-processing in healthcare analytics, particularly in handling missing values and normalizing data for enhanced analysis. The paper outlines various pre-processing strategies that can improve data quality, directly supporting the methodological framework employed in this project.

3) **Author:** Johnson, R. (2021). Exploratory Data Analysis for Disease Prediction, International Journal of Medical Data Science.

   Line No : 11–18.

   Johnson explores the methodologies for exploratory data analysis (EDA) in the context of predicting diseases. The use of correlation and covariance matrices is discussed as a means to uncover relationships in healthcare datasets. This work provides a strong basis for the statistical techniques applied in our analysis.

4) **Author:** Patel, A. (2019). The Role of Machine Learning in Diagnosing Diseases, Journal of Biomedical Informatics.

   Line No : 34–40.

   Patel investigates the application of machine learning algorithms for disease diagnosis based on symptomatic data. The study finds that these models can enhance diagnostic precision and reduce errors associated with overlapping symptoms, supporting the automation goals of our project.

5) **Author:** Chen, L. (2022). Automated Insights in Healthcare: A Review, Health Informatics Journal.

   Line No : 50–55.

   Chen reviews the development of automated systems in healthcare analytics, including techniques for clustering symptoms and predicting diseases. The paper stresses the growing

6) **Author:** Divya A., Deepika B., Durga Akhila C. H**.** Disease Prediction Based on Symptoms Given by User Using Machine Learning

Line No : 40–45

This paper presents an automated disease diagnosis model using machine learning techniques. It analyzes patient records for 41 diseases and employs Decision Tree and Naive Bayes algorithms for prognosis

7) **Author: Unknown, Human Symptoms–Disease Network**

Line No : 50–55.

This research utilizes medical bibliographic records to generate a symptom-based network of human diseases. It investigates correlations between symptom similarity and shared genes or protein interactions, revealing that symptom-based similarity can inform drug design and disease etiology research

8) **Author: Zhou et al. (2014), Human Symptoms-Disease Network**

Line No : 20–25

This paper constructs a symptom-based human disease network using a large-scale biomedical literature database. The authors find that symptom similarity between diseases correlates with shared genetic associations and protein interactions.

9) **Author: Goh et al. (2007), Human Disease Network**

Line No :30–35

The authors present a network of human disorders based on genetic origins, highlighting how diseases are interconnected through shared genes and symptoms.

10) **Author: Sadhu and Jadli (2020), A Survey on Diabetes Risk Prediction Using Machine Learning Approaches**

Line No : 40–45

This survey examines machine learning techniques for early diabetes prediction, utilizing various classifiers such as SVM, KNN, and Random Forest. The authors conclude that Random Forest achieved the highest accuracy of 98%

11) **Author: Iancu et al. (2018), Predicting Diabetes Mellitus With Machine Learning Techniques**

Line No : 50–55

This research focuses on developing a machine learning-based system for diabetes risk stratification. Utilizing logistic regression for feature selection and classifiers like Random Forest, it reports an overall accuracy of 94.25%

12) **Author: Alzubaidi et al. (2020), Classification and Prediction of Diabetes Disease Using Machine Learning Paradigm**

Line No :60–65.

This paper discusses the application of various machine learning algorithms, including XGBoost, Decision Trees, and SVM, to predict diseases based on a dataset of symptoms. The study emphasizes the importance of preprocessing and feature selection in developing a robust predictive model for healthcare applications

13) **Author: Divya A., Deepika B., Durga Akhila C. H., Tonika Devi A., Lavanya B., Sravya Teja E. - Disease Prediction Based on Symptoms Given by User Using Machine Learning**

Line No : 30–35

This research presents an automated disease diagnosis model utilizing Decision Tree and Naive Bayes algorithms to analyze patient symptoms. The authors focus on optimizing a dataset comprising 132 symptoms related to 41 diseases, aiming to improve early diagnosis and patient care through their model

14) **Author: Unknown**, AI-based Disease Category Prediction Model Using Symptoms from Low-Resource Languages

Line No : 40–45

The study explores machine learning techniques to predict disease categories from symptoms documented in the Afaan Oromo language. It employs various classification algorithms, including SVM and LSTM, demonstrating that LSTM achieved superior accuracy in disease prediction

15) **Author: Susobhan Akhuli, Disease Prediction Using Machine Learning**

Line No :50–55

This article outlines a robust machine learning framework for predicting diseases based on symptoms using datasets from Kaggle. The author describes data preparation, cleaning, and implementation of models like Random Forest and SVM to achieve high accuracy in predictions

16) **Author: Monto et al. Statistical Model for Influenza Prediction Based on Symptoms**

Line No :70–75

The authors developed a statistical model to predict influenza among unvaccinated individuals based on reported symptoms. Their model achieved an accuracy of 79%, emphasizing the utility of symptom-based prediction in public health.

17) **Author:Chen et al. Machine Learning Algorithms for Predicting Disease Outbreaks**

Line No : : 80–85

This research focuses on optimizing various machine learning algorithms to predict outbreaks of habitual diseases using limited training data. The authors employed a novel convolutional neural network model, achieving an accuracy rate of approximately 94.8%

18) **Author: Haq et al. Heart Disease Prediction Using Machine Learning Algorithms**

Line No : 90–95

The study analyzes heart disease prediction using multiple machine learning algorithms, employing feature selection techniques to enhance model performance. The authors report significant improvements in prediction accuracy through their approach

19) **Author: Palle Pramod Reddy, Dirisinala Madhu Babu, Disease Prediction using Machine Learning**

Line No :20–25

This paper focuses on developing a predictive model to identify diseases based on patient-reported symptoms using various machine learning techniques. The authors utilize a Random Forest classifier to analyze structured and unstructured data, achieving high accuracy rates for multiple diseases, including diabetes and heart disease

20) **Author: M. K. Soni et al. Identification and Prediction of Chronic Diseases Using Machine Learning**

Line No :30–35

The study aims to identify and predict chronic diseases using a machine learning approach that combines convolutional neural networks (CNN) for feature extraction with K-Nearest Neighbors (KNN) for distance calculation

21) **Author:** N. G. Levesque, S. J. Huang (the prediction of disease using machine learning)

Line No: 20–25

The paper discusses classification algorithms such as decision trees and SVM to predict diseases like diabetes and cancer. It emphasizes the importance of feature selection and pre-processing to improve prediction accuracy. A comparative analysis of multiple algorithms highlights how tuning hyperactive parameters can further enhance performance.

22) **Author:** A. A. Rahmani, M. Rezaei (Human Disease Prediction using Machine Learning Techniques and Real-life Parameters)

Line No: 18–22

This study highlights ensemble-learning techniques like Random Forest and Gradient Boosting for improved clinical diagnosis. By combining multiple models, the system achieves higher generalization performance compared to individual algorithms. The paper

also examines the limitations of individual classifiers and displays how ensemble learning mitigates issues like overfitting.

23) **Author:** M. Chakraborty, A. Khan (Multiple Disease Prognostication Based On Symptoms Using Machine Learning Techniques).

Line No: 25–30

This paper focuses on using convolutional neural networks (CNNs) for diagnosing complex diseases such as lung cancer and Alzheimer's. CNNs' effectiveness in medical imaging is emphasized, particularly in early-stage disease detection. The research also discusses the role of deep learning architectures in improving diagnostic accuracy in radiology.

Conclusion

This project, Health, Symptoms & Disease Analysis, evidences the potency of data-driven methods in disease diagnosis by means of patients' reported symptoms. We found some really impressive patterns about how to support informed decision-making from healthcare practitioners for efficient and accurate diagnosis by analysing correlations between the symptom and the disease.

The project majorly consisted of rigorous data preprocessing, including cleaning, handling missing values, and encoding categorical data; thus, setting up a quality dataset for meaningful statistical analysis. In turn, we were able to provide a set of EDA techniques, such as correlation and covariance matrices, to find clusters of symptoms common to specific diseases. This supports a more systematic diagnostic process that also helps limit medical errors and streamline healthcare workflows.

This project also compared the train and test accuracy among machine learning models such as Logistic Regression, Random Forest, and Decision Tree along with their respective cross-validation accuracies. Among all these models of machine learning, the best model was that of Random Forest; it gave an accuracy of 91.75%, which proves its reliability in generalizing predictions to unseen data in this dataset.

Overall, the project demonstrates that data science adds value to health care-from basic data processing at its core to more sophisticated predictive models. The work laid here forms a foundation for further work, such as the integration of machine learning into real-time diagnosis and model update. This will thereby improve diagnosis by creating enhanced, data-related diagnostic tools to the betterment of patient outcome.