Gisma University of Applied Sciences

Assessment Submission Form

Student Number (If this is group work, please include the student numbers of all group participants)	GH1025910
Assessment Title	AI-Powered Methodology for early Diagnosis-Predicting Disease Probability using User Input
Module Code	M598
Module Title	Master Dissertation
Module Tutor	Prof.Dr.Mohammad Mahadavi
Date Submitted	20-12-2024

Title: AI-Powered Methodology for early Diagnosis-Predicting Disease Probability using User Input

Table of contents

Table of contents

Abstract

Healthcare has changed because of the quick developments in the artificial intelligence (AI) and machine learning (ML), which have opened doors to new techniques for illness prediction and prevention. The primary objective of this thesis is to develop a machine learning-based disease prediction system that employs user input data for example that includes medical history, and symptoms. The system analyses patterns in these inputs to predict the likelihood of diseases, providing patients and medical professionals with a useful, non-invasive diagnostic tool. In order to train and assess several machine learning algorithms, this study uses a labelled dataset of symptoms and illnesses. The most accurate and interpretable model is then chosen for deployment. By bridging the gap between early symptom assessment and expert medical diagnosis, the technology hopes to lessen the strain on healthcare systems and enable prompt responses.

In order to guarantee that the suggested system is reliable, easy to use, and trustworthy, this thesis also examines ethical issues, data security, and model interpretability. This research adds to the expanding field of AI in personalized healthcare by fusing state-of-the-art machine learning techniques with easily navigable user interfaces.

1 Introduction

The diagnosis, treatment, and management of medical conditions are being revolutionised by recent developments in machine learning (AI) in healthcare. Healthcare systems may now anticipate and proactively handle health problems thanks to predictive modelling, which has emerged as an effective instrument for sickness identification. This thesis explores the creation of a framework for disease prediction based on machine learning (ML) and user-provided inputs, including medical history and symptoms. Conventional diagnostic methods frequently depend on extensive clinical knowledge and medical testing, which can be resource- and time-intensive. ML-powered solutions, on the other hand, provide a quicker and more scalable option by analysing enormous volumes of data to find patterns and connections that human practitioners might not see right away. These technologies are particularly helpful in environments with limited resources and restricted access to cutting-edge medical diagnostics.

This study aims to address the growing need for early and accurate disease prediction by:
1. Designing a user-friendly interface to collect health-related inputs.
2. Developing ML models capable of mapping user symptoms to potential diagnoses.
3. Evaluating the models based on accuracy, interpretability, and real-world applicability.

To train the ML models, the study makes use of a structured dataset that includes symptom-disease mappings. In order to make sure the system is both efficient and moral, it also takes into account issues like feature selection, dataset imbalance, and model transparency.

Our goal is to develop a disease prediction system by the end of this study that will help people take control of their health and assist medical professionals. By laying the groundwork for more easily accessible, effective, and customized medical services, this thesis supports the continuous attempts to use AI for preventative healthcare.

1.1   Conceptual Framework
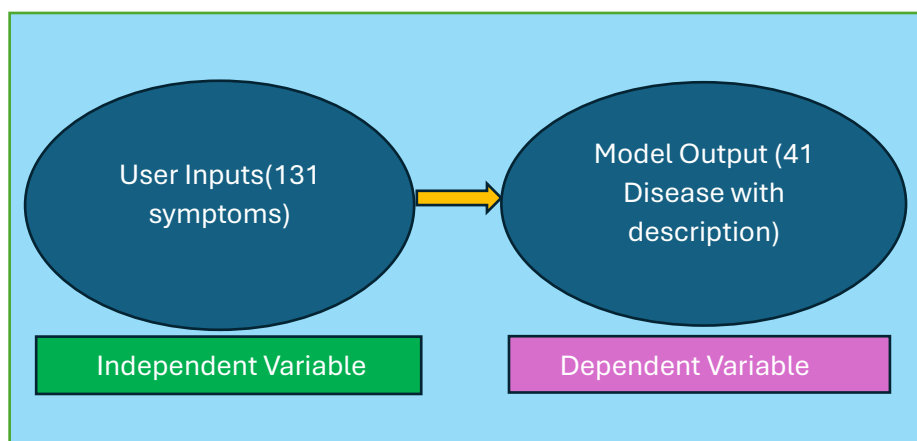
Fig 1: Conceptual framework

- **Independent Variable (IV):** User inputs include health habits, demographic information (e.g., age, gender), and symptoms. The characteristics or predictors that the model uses to infer possible diseases are represented by these inputs.
- **Dependent Variable (DV):** prognosis or likelihood of a disease. Based on user input, this is the result the model forecasts, represented as the likelihood or categorization of one or more diseases.

To estimate the illness prognosis (DV), the relationship is modelled using machine learning techniques, which process and analyse user inputs (IV). To generate precise predictions for fresh data, the model uses patterns found in the training dataset.

1.2 Project Architecture
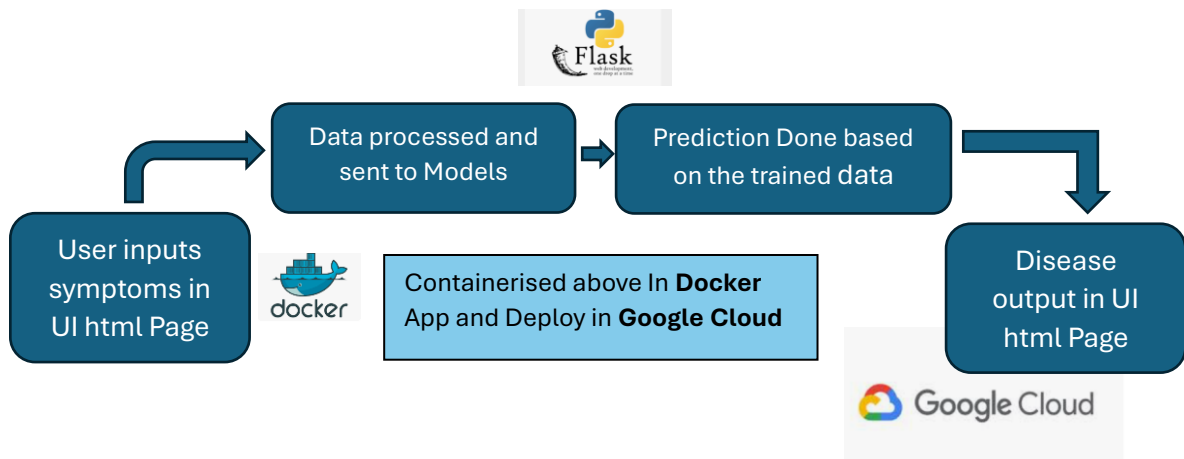
Fig 2. Project Architecture

To produce an effective and user-friendly tool, the illness prediction system's architecture combines several essential elements. To put it briefly, the disease prediction system is made to employ user-input data by combining cloud deployment, machine learning models, and web technologies in a smooth manner. The primary elements of this architecture consist of:

User Interface (UI):
The HTML page used to construct the frontend offers a user-friendly interface for users to enter their symptoms.

Web Framework:
The lightweight Python web framework Flask is used to manage user requests. In order to produce predictions, it analyses the input data and communicates with the machine learning models.

Machine Learning Models:
Based on the input symptoms, the system predicts diseases using pre-trained machine learning models (such as Random Forest and Deep Learning models). Large datasets of symptoms and the diseases that correspond to them are used to train these algorithms.

Containerization:
To ensure consistency across various environments, the complete application is containerized using Docker. Additionally, this makes the application's scaling and deployment easier.

Cloud Deployment:
The application is hosted on Google Cloud Platform (GCP). The containerized application's availability, scaling, and deployment are controlled via the serverless platform Google Cloud Run.

Prediction Output:
Users are given quick and easy access to diagnostic information by having the projected disease and its description shown back on the user interface.

This design demonstrates how to combine cloud deployment, machine learning models, data processing, and user input to create a reliable and effective disease prediction system. It is a useful tool for early disease detection and individualized healthcare since it uses sophisticated algorithms and cloud services that guarantee scalability, dependability, and ease of access.

2 Literature Review

The application of machine learning (ML) in disease prediction is a rapidly evolving field that combines advancements in artificial intelligence, healthcare data availability, and computational power. This review highlights key studies, methods, and trends in disease prediction using user input data, providing a foundation for this thesis.

2.1 The Role of the Machine Learning in Healthcare

With its major advantages in predictive analytics, machine learning has revolutionized the healthcare industry. By analysing large and intricate datasets, it empowers healthcare providers to make data-driven decisions. While some studies have concentrated on non-imaging data, such as symptoms, demographics, and lab results, to forecast diseases, others, like Esteva et al. (2017), show that machine learning (ML) is effective in predicting dermatological problems using image data (Topol, 2019).

2.2 Symptom-Based Disease Prediction

Since they employ user-provided inputs to directly determine the likelihood of diseases, symptom-based prediction models are among the most useful applications of machine learning. Das et al. (2018) found that decision trees and Naïve Bayes algorithms can be used to map symptoms to diseases with a high degree of accuracy. Similarly, Liu et al. (2020) showed the resilience of symptom-based approaches by using ensemble techniques like Random Forests to forecast diseases like diabetes and cardiovascular disorders.

2.3 Common Algorithms and Techniques

Machine learning models frequently employed in disease prediction includes:

- Logistic Regression: Effective for binary classifications, such as predicting the presence or absence of a disease (Rajkomar et al., 2018).

- Random Forests and Gradient Boosting: These ensemble techniques excel in handling structured data and multi-class problems, as seen in studies predicting infectious diseases (Smith et al., 2020).

- Neural Networks: Deep learning models have been applied to more complex problems, such as early cancer detection, but they require large datasets for optimal performance (Kourou et al., 2015).

2.4 Dataset Challenges

When creating precise prediction models, data quality is essential. Data disruption, class imbalance, and missing values are a few examples of problems that can seriously affect performance. To overcome these obstacles, methods including feature selection, data augmentation, and SMOTE (Synthetic Minority Oversampling Technique) have been effectively used (Chawla et al., 2002).

2.5   Model Interpretability and User Trust

Concerns have been expressed about the use of ML models in healthcare because to their "black-box" nature. Research highlights the value of interpretable models that can clarify the logic behind predictions, like decision trees and explainable AI (XAI) frameworks (Rudin, 2019). Gaining the trust of both healthcare providers and patients depends on transparency.

2.6   Ethical and Privacy Considerations

There are moral and legal issues with using private health information in machine learning algorithms. To protect user privacy, GDPR compliance, data anonymization, and secure storage procedures are essential (McMahan et al., 2017). Furthermore, resolving dataset biases is crucial to provide fair healthcare solutions to a range of demographics.

2.7   Applications of AI in Personalized Healthcare

The expanding trend of AI-driven tailored healthcare is demonstrated by recent studies. For example, ML algorithms are currently integrated into wearable technology and smartphone apps to make real-time predictions about conditions like hypertension and arrhythmia (Ahmed Al Kuwaiti, 2021). These technological advancements have improved accessibility and enabled people to take charge of their health management.

2.8   Machine Learning Models for Disease Prediction

By analysing user-input data, including symptoms, medical history, and vital signs, machine learning algorithms have been widely utilized to forecast a variety of diseases. A comprehensive analysis (Delpino, F. M. et al., 2022). demonstrated how machine learning can accurately forecast the presence, course, and determinants of specific chronic diseases. According to the review, models that performed the best and were most employed were random forest, deep neural networks, k-nearest neighbours, and Naive Bayes.

2.9   Use of Big Data and Multimodal Data

The accuracy of disease prediction models has been greatly improved by the combination of big data and multimodal data. A disease prediction tool that made use of both structured and unstructured hospital data was presented in a paper that was published in IRJMETS (Loni Kalbhor et al., 2022). To provide very accurate predictions about conditions including diabetes, liver disease, and heart disease, the platform used machine learning algorithms like CNN-MDRP (Multimodal Disease Risk Prediction). This method successfully handled missing and partial data, overcoming the shortcomings of earlier systems.

2.10 Early Identification and Risk Stratification

Many diseases have been identified early thanks in large part to machine learning. A thorough review (Md. Manjurul Ahsan et al. 2002) described how machine learning analyses intricate disease mechanisms and underlying symptoms to aid in the early identification of a variety of illnesses. The use of various algorithms, illness categories, data types, and evaluation metrics were among the latest trends and techniques in machine-learning-based disease diagnosis that were emphasized in the review.

2.11 Algorithmic Approaches and Performance

The ability of several machine learning algorithms to forecast diseases has been compared. In a paper published in the International Journal of Engineering, K. Gaurav et al. (2023) suggested a model that included the Random Forest, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) algorithms. Compared to cutting-edge techniques, our model demonstrated superior accuracy and dependability, especially when it came to forecasting conditions like diabetes and heart disease. Weighted KNN and Naive Bayes algorithms have also been investigated; their scalability and feature independence have been cited as benefits and drawbacks.

2.12 Clinical Applications and Future Trends

These models have a wide range of clinical uses, such as automating the process of referring patients to specialists and creating individualized treatment regimens. According to K. Gaurav et al. (2023), the suggested models can help the healthcare sector by accurately forecasting ailments, which will lessen the problems that patients have with availability and affordability.

In summary, machine learning models have shown significant promise in predicting diseases based on user input, leveraging big data, multimodal data, and various algorithms to achieve high accuracy. These models are poised to improve clinical decisions, streamline healthcare facilities, and enhance patient outcomes.

Key Gaps and Future Directions:

Even while current research shows encouraging outcomes, there are still several gaps:
1. Model generalizability is limited by the scarcity of rare disease datasets.
2. Not enough research has been done on the integration of ML models with user-friendly interfaces.
3. It's still difficult to strike a balance between interpretability and model accuracy. By creating a transparent, symptom-based machine learning model for disease prediction and incorporating it with an interactive user interface, this thesis fills these shortcomings and makes it accessible and usable even for non-experts.

This section outlines the systematic approach to developing a machine learning (ML) model for disease prediction based on user input. The methodology is divided into several key phases: data preprocessing, feature selection, model development, evaluation, and deployment.

3 Methodology

3.1 Data Collection and Preprocessing

3.1.1 Data Source

In order to support research and development in predictive healthcare analytics, the Symptom-Disease Prediction Dataset (SDPD) is a thorough collection of structured data that connects symptoms to a variety of diseases. This dataset attempts to offer a solid basis for the creation of symptom-based disease prediction models, drawing inspiration from the approach used by well-known organizations like the Centres for Disease Control and Prevention (CDC). The collection includes a wide variety of symptoms drawn from credible medical literature, expert consensus, and clinical observations.

Link to data: https://data.mendeley.com/datasets/dv5z3v2xyd/1

The dataset symbipredict_2022.csv contains a comprehensive set of symptoms and their corresponding disease diagnoses. Here is an explanation of each column, their scientific importance, and how they contribute to predicting diseases based on user inputs.

Column Names and Descriptions: -

1. Itching

Description: This Indicates whether the patient is suffering from itching or not.

Scientific Importance: The Itching can be a symptoms of various skin conditions, allergies, or systemic illnesses such as liver or kidney disorders. It is essential for differential diagnosis.

2. Skin_rash

Description: This Indicates there is presence of a skin rash.

Scientific Importance: Skin rashes can be sign of infections, autoimmune diseases, or allergic reactions. They are important for diagnosing conditions like psoriasis, eczema or infectious diseases.

3. Nodal_skin_eruptions

Description: This Indicates the presence of the nodal skin eruptions, which are raised and solid elevations on the skin.

Scientific Importance: These eruptions can be associated with the conditions like tuberculosis, lymphoma, or other infections, making them significant for diagnosis.

4. Continuous_sneezing

Description: This indicates whether the patient is experiencing continuous sneezing.

Scientific Importance: Constant sneezing may indicate respiratory infections, allergies, or other disorders that impact the nasal passages.

5. Shivering

Description: This indicates whether the patients are experiencing the shivering.

Scientific Importance: Shivering can be a symptom of hypothermia, fever or other systemic infections.

6. Chills

Description: This lets you know if the patient is feeling cold.

Scientific Importance: Fever is frequently accompanied with chills, which may be a sign of influenza, infections, or other systemic diseases.

7. Joint_pain

Description:This shows that joint discomfort is present.

Scientific Importance: Arthritis, trauma, and systemic conditions like lupus or rheumatoid arthritis, for example, can all cause joint discomfort.

8. Stomach_pain

Description: This shows that there is abdominal ache.

Scientific Importance: Gastrointestinal conditions like gastritis, ulcers, or inflammatory bowel disease may be indicated by stomach pain.

9. Acidity

Description: This Indicates the presence of gastritis or acidity is present.

Scientific Importance: Peptic ulcers, gastroesophageal reflux disease (GERD), and other gastrointestinal conditions can all be accompanied by acidity.

10. Ulcers_on_tongue

Description: Indicates the presence that there are tongue ulcers present.

Scientific Importance: Tongue ulcers can be associated with oral infections, vitamin deficiencies, or systemic diseases like Behçet's disease.

11. Muscle_wasting

Description: shows a reduction of muscular mass or muscle atrophy.

Scientific Importance: Malnutrition, neurological illnesses, and muscular dystrophy can all cause muscle loss.

12. Vomiting

Description: Indicates whether the patient is experiencing vomiting.

Scientific Importance: Food poisoning, gastrointestinal infections, and systemic diseases like diabetes or kidney disease can all cause vomiting.

13. Burning_micturition

Description: shows that there is a burning feeling when urinating.

Scientific Importance: Urinary tract infections (UTIs) and other genitourinary disorders are frequently linked to this symptom.

14. spotting_urination

Description: This Indicates the presence of spotting or blood in the urine.

Scientific Importance: This may be a sign of kidney stones, urinary tract infections, or more severe diseases including kidney or bladder cancer.

15. fatigue

Description: This indicates the presence of excessive tiredness or fatigue.

Scientific Importance: Anaemia, diabetes, thyroid issues, and chronic fatigue syndrome are just a few of the many illnesses that can be linked to exhaustion, which is a general complaint.

16. weight_gain

Description: This indicates whether the patient has experienced an abnormal gain in weigh.

Scientific Importance: Hormonal imbalances, metabolic diseases, and other systemic illnesses may be linked to weight gain.

17. Anxiety

Description: This indicates the presence of the anxiety in patient.

Scientific Importance: Anxiety may be a symptom of some underlying physical illness or a separate mental health issue.

18. Cold_hands_and_feets

Description: This indicates whether the patient's hands and feet abnormally cold.

Scientific Importance: Raynaud's disease, poor circulation, and other vascular disorders can all cause cold extremities.

19. Mood_swings

Description: This indicates the symptoms of mood swings.

Scientific Importance: Hormonal changes, other systemic disorders, and mental health conditions like bipolar disorder can all be linked to mood swings.

20. weight_loss

Description: This indicates whether the patient has experienced abnormal weight loss.

Scientific Importance: Inadvertent weight loss may indicate a serious illness such as metabolic problems, cancer, or tuberculosis.

21. Restlessness

Description: This indicates the presence of the restlessness.

Scientific Importance: Anxiety disorders, sleep difficulties, and other neurological issues can all manifest as restlessness.

22. lethargy

Description: This indicates the presence of lack of energy or lethargy.

Scientific Importance: Numerous illnesses, such as infections, neurological diseases, or metabolic abnormalities, might be linked to lethargy.

23. Patches_in_throat

Description: This indicates the presence of patches in the throat area.

Scientific Importance: Patches on the throat may be a sign of more serious illnesses like mouth cancer or infections like tonsillitis.

24. Irregular_sugar_level

Description: This Indicates irregular blood sugar levels than the normal person.

Scientific Importance: Diabetes is characterized by abnormal blood sugar levels, which are frequently linked to other metabolic diseases.

25. cough

Description: This indicates the presence of a coughing in the patient.

Scientific Importance: Allergies, respiratory infections, and long-term illnesses like asthma or chronic obstructive pulmonary disease (COPD) can all cause coughing.

26. High_fever

Description: This indicates the experiencing of the High fever.

Scientific Importance: A high temperature is frequently a sign of an infection, but it can also be linked to systemic illnesses like sepsis.

27. Sunken_eyes

Description: This indicates the presence of sunken/deepened eyes.

Scientific Importance: Dehydration, malnourishment, or other systemic disorders may be the cause of sunken eyes.

28. Breathlessness

Description: This indicates the presence of shortness of breath or breathlessness.

Scientific Importance: Asthma, COPD, and heart problems are among the respiratory disorders that can cause dyspnoea.

29. Sweating

Description: This can be observed by excessive sweating.

Scientific Importance: Sweating excessively may indicate a systemic illness, infection, or hormonal imbalance.

30. Dehydration

Description: This can be observed by the dehydration.

Scientific Importance: Numerous illnesses, such as diabetes, gastrointestinal tract infections, or insufficient fluid consumption, can cause dehydration.

31. Indigestion

Description: This indicates the presence of the indigestion.

Scientific Importance: Peptic ulcers, GERD, and other digestive diseases can all be accompanied by indigestion.

32. Headache

Description: This indicates the presence of the headaches.

Scientific Importance: Numerous illnesses, such as tension headaches, migraines, and more dangerous ailments like meningitis, can be linked to headaches.

33. Yellowish_skin

Description: This indicates yellowish skin of the skin sometimes the eyes and nails (jaundice).

Scientific Importance: Hepatitis or cirrhosis are two examples of liver or biliary tract conditions that can cause jaundice.

34. Dark_urine

Description: This indicates dark-coloured urine.

Scientific Importance: Dehydration, liver or renal disease, and other metabolic abnormalities can all be indicated by dark urine.

35. Nausea

Description: This indicates the presence of the nausea.

Scientific Importance: Motion sickness, gastrointestinal diseases, and systemic illnesses like chemotherapy or pregnancy can all cause nausea.

36. Loss_of_appetite

Description: This indicates a loss of appetite.

Scientific Importance: Several illnesses, including as infections, depression, or long-term ailments like cancer, can cause appetite loss.

37. Pain_behind_the_eyes

Description: This indicates pain behind the eyes.

Scientific Importance: This symptom may be linked to sinusitis, migraines, or other neurological disorders.

38. Back_pain

Description: This indicates the presence of back pain.

Scientific Importance: Back pain can be associated with musculoskeletal issues, herniated discs, or systemic conditions like kidney stones.

39. Constipation

Description: This feature indicates constipation.

Scientific Importance: Dietary practices, gastrointestinal disorders, and systemic diseases like hypothyroidism can all cause constipation.

40. Abdominal_pain

Description: This feature indicates abdominal pain.

Scientific Importance: Numerous illnesses, such as appendicitis, gastrointestinal infections, and other abdominal diseases, might be linked to stomach pain.

41. Diarrhoea

Description: This feature indicates diarrhoea.

Scientific Importance: Food poisoning, gastrointestinal infections, and other systemic illnesses including irritable bowel syndrome (IBS) can all cause diarrhoea.

42. Mild_fever

Description: This feature indicates the presence of mild fever.

Scientific Importance: A mild fever may indicate a little infection or the beginning of a more serious illness.

43. Yellow_urine

Description: This indicates yellow-coloured urine.

Scientific Importance: In addition to being a sign of dehydration or other metabolic disorders, yellow urine can sometimes be normal.

44. Yellowing_of_eyes

Description: This indicates yellowing of the eyes (jaundice).

Scientific Importance: This is an early indication of liver or biliary tract disorders, much like yellowish skin.

45. Acute_liver_failure

Description: This feature indicates the acute liver failure.

Scientific Importance: Acute liver failure is a serious illness that can be brought on by several conditions, such as drug overdose, viral hepatitis, or other liver disorders.

46. Fluid_overload

Description: This feature indicates the fluid overload.

Scientific Importance: Heart failure, renal dysfunction, and other disorders that cause fluid retention can all be accompanied by fluid overload.

47. Swelling_of_stomach

Description: This feature indicates swelling of the stomach.

Scientific Importance: This may be linked to liver illness, other abdominal diseases, or gastrointestinal problems such ascites.

48. Swelled_lymph_nodes

Description: This feature indicates swollen lymph nodes.

Scientific Importance: Infections, autoimmune disorders, and malignancies such as lymphoma can all cause enlarged lymph nodes.

49. Malaise

Description: This indicates a general feeling of malaise or illness.

Scientific Importance: A variety of ailments, such as infections, autoimmune diseases, or chronic illnesses, might be linked to malaise, a vague feeling.

50.Blurred_and_distorted_vision

Description: This indicates the presence of the distorted vision or blurred.

Scientific Importance: Several disorders affecting the eyes or nervous system can cause blurred or distorted vision.

51. Phlegm

Description: This indicates the presence of phlegm or mucus.

Scientific Importance: Phlegm can be a symptom of respiratory infections, allergies, or chronic conditions like bronchitis or COPD.

52. Throat_irritation

Description: This indicates the irritation or discomfort in the throat.

Scientific Importance: Allergies, tonsillitis, and other throat-related disorders can all be linked to throat irritation.

53. Redness_of_eyes

Description: This indicates the irritation redness of the eyes.

Scientific Importance: Allergies, eye infections, and other ocular disorders can all cause red eyes.

54. Sinus_pressure

Description: This feature indicates the pressure or pain in the sinuses.

Scientific Importance: Allergic reactions, sinusitis, and other respiratory disorders are frequently linked to sinus pressure.

55. Runny_nose

Description: This feature indicates a runny nose.

Scientific Importance: A runny nose may indicate a respiratory infection, cold, or allergy.

56. Congestion

Description: This feature indicates the nasal congestion.

Scientific Importance: Allergies, colds, and other breathing-related conditions can all be associated with congestion.

57. Chest_pain

Description: This feature indicates discomfort or chest pain.

Scientific Importance: Respiratory infections, heart problems, and other dangerous illnesses like pneumonia can all cause chest pain.

58. Weakness_in_limbs

Description: This indicates weakness in the limbs.

Scientific Importance: Muscular dystrophy, neurological disorders, and other systemic ailments can all be linked to limb weakness.

59. Fast_heart_rate

Description: This indicates a fast heart rate (tachycardia).

Scientific Importance: Anxiety, heart problems, and other systemic disorders can all be indicated by a rapid heartbeat.

60. Pain_during_bowel_movements

Description: This indicates a pain during the bowel movements.

Scientific Importance: Haemorrhoids, other rectal disorders, or gastrointestinal problems including constipation may be linked to this symptom.

61. Pain_in_anal_region

Description: This indicates pain in the anal region.

Scientific Importance: Anal fissures, haemorrhoids, and other rectal disorders may all trigger anal pain.

62. Bloody_stool

Description: This feature indicates the presence of blood in the stool.

Scientific Importance: Severe illnesses including colon cancer, ulcers, or gastrointestinal bleeding can all be indicated by bloody stool.

63. Irritation_in_anus

Description: This feature indicates irritation or discomfort in the anus.

Scientific Importance: Haemorrhoids, anal fissures, and other rectal disorders can be linked to anal discomfort.

64. Neck_pain

Description: This feature indicates the neck discomfort or pain.

Scientific Importance: Injuries, musculoskeletal problems, and other disorders affecting the neck can all cause neck pain.

65. Dizziness

Description: This feature indicates the dizziness or light-headedness.

Scientific Importance: Low blood pressure, vestibular problems, and other systemic diseases can all be linked to dizziness.

66. Cramps

Description: This feature indicates the muscle cramps.

Scientific Importance: Dehydration, electrolyte imbalances, and other muscular disorders can all cause muscle cramps.

67. Bruising

Description: This feature indicates easy bruising.

Scientific Importance: Easy bruising may be a sign of vitamin deficiencies, bleeding disorders, or other illnesses that interfere with blood coagulation.

68. Obesity

Description: Indicates obesity.

Scientific Importance: A number of chronic illnesses, such as diabetes, heart disease, and several forms of cancer, are made more likely by obesity.

69. Swollen_legs

Description: This feature indicates the swelling in the legs.

Scientific Importance: Leg swelling may indicate heart failure, fluid retention, or other vascular disorders.

70. Swollen_blood_vessels

Description: This feature indicates the swelling of blood vessels.

Scientific Importance: Allergies, vascular disorders, and other inflammatory processes can all be linked to enlarged blood vessels.

71. Puffy_face_and_eyes

Description: This feature indicates the puffiness of the face and eyes.

Scientific Importance: Allergies, fluid retention, and other disorders affecting the face and eyes can all cause puffiness.

72. Enlarged_thyroid

Description: This feature indicates an enlarged thyroid gland.

Scientific Importance: Thyroid conditions including goitre or thyroiditis can manifest as an enlarged thyroid.

73. Brittle_nails

Description: This feature indicates brittle nails.

Scientific Importance: Thyroid issues, dietary deficits, and other systemic diseases can all manifest as brittle nails.

74. Swollen_extremeties

Description: This indicates swelling in the extremities.

Scientific Importance: Heart failure, fluid retention, and other vascular disorders can all be linked to swelling in the extremities.

75. Excessive_hunger

Description: This indicates the excessive hunger.

Scientific Importance: Diabetes and other metabolic diseases that impact appetite regulation may cause excessive hunger.

76. Extra_marital_contacts

Description: This indicates the if the patient has had extra-marital contacts.

Scientific Importance: Sexually transmitted infections (STIs) and other sexually transmitted diseases may be diagnosed using this information.

77. Drying_and_tingling_lips

Description: This indicates the drying and tingling of the lips.

Scientific Importance: Allergies, nutritional deficiencies, and other disorders affecting the skin and mucous membranes may be linked to this symptom.

78. Slurred_speech

Description: This indicates the slurred speech.

Scientific Importance: Speech slurring may indicate neurological problems such as stroke, drunkenness, or other brain illnesses.

79. Knee_pain

Description: This indicates the knee pain or discomfort.

Scientific Importance: Injuries, musculoskeletal problems, and other diseases affecting the knee joint can all be linked to knee pain.

80. Hip_joint_pain

Description: This feature indicates the hip joint discomfort or pain.

Scientific Importance: Arthritis, musculoskeletal disorders, and other hip-related illnesses can all cause hip joint pain.

81. Muscle_weakness

Description: This feature indicates muscle weakness.

Scientific Importance: Muscular dystrophy, neurological disorders, and other systemic ailments can all be linked to muscle weakness.

82. Stiff_neck

Description: This feature indicates a stiff neck.

Scientific Importance: Meningitis, musculoskeletal disorders, and other disorders affecting the neck can all cause a stiff neck.

83. Swelling_joints

Description: This feature indicates the swelling in the joints.

Scientific Importance: Injuries, arthritis, and other joint-related disorders can all be linked to joint swelling.

84. Movement_stiffness

Description: This feature indicates stiffness during movement.

Scientific Importance: Stiffness in movement can be a sign of arthritis, musculoskeletal problems, or other disorders that impair joint mobility.

85. Spinning_movements

Description: Indicates dizziness or spinning sensations (vertigo).

Scientific Importance: Vestibular illnesses, inner ear issues, and other neurological conditions can all be linked to vertigo.

86. loss_of_balance

Description: This feature indicates loss of balance.

Scientific Importance: A vestibular disorder, a neurological condition, or other problems affecting coordination might cause loss of balance.

87. Unsteadiness

Description: This indicates unsteadiness or lack of stability.

Scientific Importance: Vestibular abnormalities, neurological conditions, and other problems affecting balance and coordination can all be linked to unsteadiness.

88. Weakness_of_one_body_side

Description: This feature indicates weakness on one side of the body.

Scientific Importance: This symptom may be linked to many brain illnesses or neurological conditions such as multiple sclerosis or stroke.

89. loss_of_smell

Description: This indicates loss of smell (anosmia).

Scientific Importance: Smell loss may be a sign of neurological disorders, upper respiratory infections, or other olfactory system problems.

90. Bladder_discomfort

Description: This feature indicates discomfort or pain in the bladder.

Scientific Importance: Urinary tract infections (UTIs), bladder stones, and other genitourinary disorders can all be linked to bladder discomfort.

91. Foul_smell_of_urine

Description: Indicates a foul or unusual smell of urine.

Scientific Importance: Urine with an unpleasant Odor may indicate kidney stones, UTIs, or other genitourinary disorders.

92. Continuous_feel_of_urine

Description: This feature indicates a continuous feeling of needing to urinate.

Scientific Importance: This symptom may be linked to genitourinary disorders, bladder problems, or UTIs.

93. Passage_of_gases

Description: This feature indicates the passage of gas (flatulence).

Scientific Importance: Food intolerance, irritable bowel syndrome (IBS), and other digestive disorders can all be accompanied by gasping.

94. Internal_itching

Description: This feature indicates internal itching or discomfort.

Scientific Importance: Several ailments, such as allergies, infections, or systemic disorders, might be linked to internal itching.

95. Toxic_look_(typhos)

Description: This feature indicates a typhoid-like symptoms or toxic appearance.

Scientific Importance: Serious diseases like typhoid fever or other systemic illnesses may be linked to this symptom.

96. Depression

Description: Indicates depression or depressive symptoms.

Scientific Importance: A mental health illness, depression can also be a sign of other underlying physiological disorders.

97. Irritability

Description: This feature indicates the irritability or mood changes.

Scientific Importance: Hormonal changes, mental health issues, and other systemic disorders can all be linked to irritability.

98. Muscle_pain

Description: This feature indicates the muscle pain or myalgia.

Scientific Importance: Infections, musculoskeletal disorders, and other systemic ailments can all manifest as muscle discomfort.

99. Altered_sensorium

Description: This feature indicates an altered state of consciousness or sensorium.

Scientific Importance: A changed sensorium may be a sign of infections, neurological disorders, or other systemic diseases that impact the brain.

100. Red_spots_over_body

Description: This feature indicates the red spots or rashes over the body.

Scientific Importance: Skin infections, allergic responses, and other systemic diseases can be linked to red patches.

101. Belly_pain

Description: This feature indicates the abdominal pain or discomfort.

Scientific Importance: Injuries, disorders of the abdomen, or gastrointestinal problems can all cause abdominal pain.

102. Abnormal_menstruation

Description: This indicates the abnormal menstrual cycles or symptoms.

Scientific Importance: Unusual menstruation may indicate uterine fibroids, hormone abnormalities, or other gynaecological disorders.

103. Dischromic_patches

Description: This feature indicates the presence of discoloured patches on the skin.

Scientific Importance: Vitiligo, melasma, and other pigmentation disorders are among the skin problems that can be linked to dyschromic patches.

104. Watering_from_eyes

Description: This feature indicates excessive tearing or watering from the eyes.

Scientific Importance: Allergies, blocked tears ducts, other eye infections can all cause watery eyes.

105. Increased_appetite

Description: This feature indicates an increase in appetite.

Scientific Importance: Diabetes, hyperthyroidism, and several drugs can all be linked with greater appetite.

106. polyuria

Description: This feature indicates excessive urination.

Scientific Importance: Polyuria can be a symptom of diabetes, kidney disorders, or other conditions affecting fluid balance.

107. Family_history

Description: This feature indicates the presence of a relevant family history.

Scientific Importance: When determining a person's genetic susceptibility to a particular disease, family history is essential.

108. Mucoid_sputum

Description: This feature indicates the presence of mucoid (thick and sticky) sputum.

Scientific Importance: Chronic illnesses like cystic fibrosis or respiratory infections may be linked to mucoid sputum.

109. Rusty_sputum

Description: This feature indicates the presence of rusty-coloured sputum.

Scientific Importance: Sputum that is rusty may indicate pneumonia or other severe respiratory diseases.

110. Lack_of_concentration

Description: This feature indicates difficulty concentrating.

Scientific Importance: Several of the neurological or psychological disorders can manifest as a lack of focus.

111. Visual_disturbances

Description: This feature indicates the presence of visual disturbances.

Scientific Importance: Migraines, neurological problems, and ocular conditions can all be linked to visual abnormalities.

112. Receiving_blood_transfusion

Description: This feature indicates if the patient has received a blood transfusion.

Scientific Importance: When evaluating the risk of specific diseases or transfusion responses, blood transfusion history is crucial.

113. Receiving_unsterile_injections

Description: This feature indicates if the patient has received unsterile injections.

Scientific Importance: When determining one's risk of contracting diseases like HIV or hepatitis, this knowledge is essential.

114. Coma

Description: This feature indicates if the patient is in a coma.

Scientific Importance: Coma is a serious illness that can be brought on by several things, such as metabolic problems or brain trauma.

115. Stomach_bleeding

Description: This feature indicates the presence of stomach bleeding.

Scientific Importance: Bleeding in the gastrointestinal tract may indicate gastritis, ulcers, or more serious disorders including stomach cancer.

116. Distention_of_abdomen

Description: This feature indicates abdominal distention or bloating.

Scientific Importance: Numerous gastrointestinal disorders or other systemic illnesses may be linked to abdominal distention.

117. History_of_alcohol_consumption

Description: This feature indicates a history of alcohol consumption.

Scientific Importance: When determining the risk of liver disease and other alcohol-related health problems, past alcohol use is crucial.

118. Fluid_overload

Description: This feature indicates fluid overload in the body.

Scientific Importance: Fluid overload may be a symptom of a condition that impairs fluid balance, such as heart failure or kidney illness.

119. Blood_in_sputum

Description: This feature indicates the presence of blood in sputum (haemoptysis).

Scientific Importance: Sputum with blood may indicate lung cancer or other severe respiratory diseases like pneumonia.

120. Prominent_veins_on_calf

Description: This feature indicates prominent veins on the calf.

Scientific Importance: Veins that are noticeable may indicate venous insufficiency or other vascular disorders.

121. Palpitations

Description: This feature indicates the sensation of a racing or pounding heart.

Scientific Importance: Several of the cardiac arrhythmias and other heart disorders can be linked to palpitations.

122. Painful_walking

Description: This feature indicates pain while walking.

Scientific Importance: Peripheral neuropathy, vascular illnesses, but or musculoskeletal disorders can all cause painful walking.

123. Pus_filled_pimples

Description: This feature indicates the presence of pus-filled pimples.

Scientific Importance: Acne or infections of the skin may be indicated by pus-filled pimples.

124. Blackheads

Description: This feature indicates the presence of blackheads.

Scientific Importance: Acne is frequently linked to the common skin problem known as blackheads.

125. scurrying

Description: This feature indicates the presence of scurrying (likely meant to be "scarring" or scurfy skin).

Scientific Importance: Scurfy skin is linked to a variety of dermatological disorders.

126. Skin_peeling

Description: This feature indicates peeling of the skin.

Scientific Importance: Skin peeling may indicate allergic comments, sunburn, or other skin disorders.

127. Silver_like_dusting

Description: This feature indicates a silver-like dusting on the skin.

Scientific Importance: Psoriasis along with other dermatological conditions may be linked to this symptom.

128. Small_dents_in_nails

Description: This feature indicates small dents or pits in the nails.

Scientific Importance: Changes in the nails may be a sign of a number of systemic diseases or malnutrition.

129. Inflammatory_nails

Description: This feature indicates inflammation around the nails.

Scientific Importance: Infections or other nail-related disorders may be indicated by inflamed nails.

130. Blister

Description: This feature indicates the presence of blisters.

Scientific Importance: Burns, infections, and other skin disorders can all be linked to blisters.

131. Red_sore_around_nose

Description: This feature indicates red sores around the nose.

Scientific Importance: This may be a sign of allergies, infections, or other nasal-related disorders.

132. Yellow_crust_ooze

Description: This feature indicates yellow crusty or oozing lesions.

Scientific Importance: Bacterial skin infections may be indicated by a yellow crust or spilling.

133. Prognosis

Description: This feature indicates the predicted outcome or diagnosis.

Scientific Importance: Based on the symptoms that are displayed, the model seeks to forecast this target variable.

A more thorough and precise diagnosis is made possible by the extensive collection of symptoms and variables these extra columns offer, which can be utilized to forecast several medical disorders.

Diseases and Brief Explanations

1. **Acne**

Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.

2. **AIDS (Acquired Immunodeficiency Syndrome)**

The late stage of HIV (Human Immunodeficiency Virus) infection, known as AIDS, is characterized by significant immune system damage that impairs the body's ability to fight off infections and illness.

3. **Alcoholic Hepatitis**

Alcoholic hepatitis is a liver inflammation brought on by excessive alcohol usage. Jaundice, exhaustion, and stomach ache are among the symptoms. If left untreated, it can result in liver failure and cirrhosis.

4. **Allergy**

An allergy occurs when the immune system overreacts to an innocuous item, like dust mites, pollen, or specific foods. Itching, sneezing, runny nose, and in extreme situations, anaphylaxis, are some of the symptoms.

5. **Arthritis**

Inflammation and discomfort in the joints are symptoms of arthritis. Psoriatic arthritis, rheumatoid arthritis, and osteoarthritis are among the various varieties. Joint discomfort, stiffness, and enema are among the symptoms.

6. **Bronchial Asthma**

Inflammation, airway blockage, and bronchial tube spasm are the hallmarks of bronchial asthma, a chronic respiratory illness. Chest tightness, coughing, wheezing, and shortness of breath are some of the symptoms.

7. Cervical Spondylosis

A disorder called cervical spondylosis affects the neck (cervical spine) as a result of aging. In certain situations, it may result in numbness or weakness in the arms or legs, as well as stiffness and neck pain.

8. Chickenpox

The varicella-zoster virus is the cause of the highly contagious disease known as chickenpox. In addition to producing itchy, blister-like rashes, it can result in consequences like encephalitis or pneumonia.

9. Chronic Cholestasis

A disorder known as chronic cholestasis causes liver injury by reducing or blocking bile flow. Fatigue, itching, and jaundice are some of the symptoms.

10. Common Cold

The upper respiratory tract is impacted by the viral infection known as the common cold. Sneezing, coughing, sore throat, and runny nose are some of the symptoms.

11. Dengue

Dengue is a virus that is spread by mosquitoes and can produce serious flu-like symptoms. High fever, excruciating headache, rash, joint and muscular discomfort, pain behind the eyes, and moderate bleeding are some of the symptoms.

12. Diabetes

Diabetes is a long-term illness that alters how your body uses food as fuel. High blood sugar levels are its defining feature, and medication, diet, and exercise can all help control it.

13. Dimorphic Haemorrhoids (Piles)

Swollen veins in the anus or lower rectum are known as hemorrhoids. They can cause pain, itching, and bleeding during bowel movements and can be external or internal.

14. Drug Reaction

An undesirable or dangerous reaction to a medication is called a drug reaction, sometimes referred to as an adverse drug reaction. From minor skin rashes to serious illnesses, symptoms can vary widely.

15. Fungal Infection

When a fungus overgrows in the body, it can cause a fungal infection. Ringworm, candidiasis, and athlete's foot are common varieties. Skin rashes to more serious systemic infections are possible symptoms.

16. Gastroenteritis

Gastroenteritis is an inflammation of the intestines and stomach that is frequently brought on by bacterial or viral infections. Fever, vomiting, diarrhea, and abdominal pain are some of the symptoms.

17. **GERD (Gastroesophageal Reflux Disease)**

Heartburn, chest pain, and trouble swallowing are symptoms of GERD, a chronic illness in which stomach acid refluxes into the oesophagus.

18. **Heart Attack**

A heart attack happens when the heart's blood supply is cut off, harming the heart muscle. Chest pain, dyspnoea, and pain in the arms, back, neck, jaw, or stomach are among the symptoms.

19. **Hepatitis A**

The hepatitis A virus is the cause of hepatitis A, a liver infection. Jaundice, exhaustion, appetite loss, and stomach pain are some of the symptoms it may induce.

20. **Hepatitis B**

The hepatitis B virus is the cause of hepatitis B, a liver illness. With symptoms resembling those of hepatitis A, it may result in acute or chronic liver disease.

21. **Hepatitis C**

The hepatitis C virus is the cause of hepatitis C, a liver illness. In addition to symptoms including exhaustion, jaundice, and abdominal pain, it frequently results in chronic liver disease.

22. **Hepatitis D**

 Only those who have already caught hepatitis B can develop hepatitis D, a liver infection. It may worsen symptoms and damage to the liver.

23. **Hepatitis E**

The hepatitis E virus is the cause of hepatitis E, a liver infection. Although it is typically acute, pregnant women and those with underlying liver disease may experience severe symptoms.

24. **Hypertension**

High blood pressure, often known as hypertension, is a condition in which the artery pressure is consistently high. If left untreated, it can result in kidney failure, heart disease, and stroke.

25. **Hyperthyroidism**

The disease known as hyperthyroidism occurs when the thyroid gland overproduces thyroid hormone. Weight loss, an accelerated heartbeat, and increased perspiration are some of the symptoms.

26. Hypoglycaemia

A condition known as hypoglycaemia occurs when blood sugar levels are abnormally low. Shaking, light-headedness, perspiration, hunger, and bewilderment are some of the symptoms.

27. Hypothyroidism

When the thyroid gland does not create enough thyroid hormone, it is known as hypothyroidism. Weight gain, exhaustion, dry skin, and hair loss are some of the symptoms.

28. Impetigo

Impetigo is a highly contagious skin condition that primarily affects the mouth and nose, causing red sores on the face. Children are more likely to have it.

29. Jaundice

Jaundice is a disorder where elevated bilirubin levels cause the skin and eyes to turn yellow. Haemolytic anaemia, bile duct obstruction, or liver disorders can all be the cause.

30. Malaria

An infected female Anopheles mosquito transmits malaria, a dangerous and occasionally fatal illness. Fever, chills, flu-like symptoms, and in extreme situations, organ failure, are among the symptoms.

31. Migraine

A migraine is a neurological disorder marked by severe, incapacitating headaches, frequently accompanied by light and sound sensitivity, nausea, and vomiting.

32. Osteoarthritis

The most prevalent kind of arthritis, osteoarthritis, is brought on by joint wear and strain. Joint discomfort, stiffness, and enema are among the symptoms.

33. Paralysis (Brain Haemorrhage)

A cerebral haemorrhage, or bleeding in the brain, can cause paralysis. In certain bodily parts, this may result in a loss of sensation and muscular function.

34. Peptic Ulcer Disease

Sores in the stomach or duodenum (the first segment of the small intestine) lining are a symptom of peptic ulcer disease. Abdominal pain, nausea, and vomiting are among the symptoms.

35. Pneumonia

An infection in one or both lungs is known as pneumonia. Coughing, fever, chills, and trouble breathing are some of the symptoms.

36. **Psoriasis**

Psoriasis is an autoimmune disease that causes skin cells to proliferate more quickly, leading to a rapid accumulation of skin cells on the skin's surface. Scales and red, itchy, and occasionally painful patches are formed by the excess skin cells.

37. **Tuberculosis**

The bacterial infection known as tuberculosis (TB) mainly affects the lungs, however it can also spread to other areas of the body. Weight loss, chest pain, and blood in the cough are some of the symptoms.

38. **Typhoid**

Salmonella Typhi is the bacterium that causes typhoid. Fever, headaches, sore throats, and stomach-aches are some of the symptoms.

39. **Urinary Tract Infection (UTI)**

Any infection that affects the kidneys, ureters, bladder, or urethra is referred to as a UTI. Abdominal pain, burning when peeing, and frequent urination are some of the symptoms.

40. **Varicose Veins**

Although they can appear anywhere, varicose veins—large, twisted veins—are most frequently found in the legs. They may result in skin discoloration, discomfort, and edema.

41. **Vertigo**

Vertigo is the sense that you are spinning or that everything around you is spinning. It may be brought on by disorders of the brain and nerve system or issues with the inner ear.

3.1.2   Data Cleaning

To guarantee a dataset's correctness and integrity, data cleansing is a crucial step. In data cleansing, handling missing values is one of the main responsibilities. Several imputation approaches can be used to do this, such as using more complex techniques like predictive imputation or replacing the missing values with the dataset's mean or median. Since missing values can have a substantial impact on machine learning model performance and provide biased results, it is imperative that they be addressed. Identifying and fixing any inconsistent data entry is also crucial. Errors in data entry, system malfunctions, or the combining of datasets from several sources are some of the causes of inconsistent data. These discrepancies have the potential to interfere with the analysis and produce inaccurate results.

Thus, it is essential to eliminate or fix these mistakes in order to preserve the general integrity of the dataset. One may greatly improve data quality and provide more accurate and trustworthy analyses and predictions by carefully managing missing values and maintaining consistency throughout the dataset. In the end, this painstaking procedure guarantees that the data utilized is reliable and strong, providing a strong basis for any further analysis or model construction.

3.1.3 Data Encoding

When working with categorical variables like symptoms, data encoding is an essential step in getting the data ready for machine learning models. To convert these categorical variables into numerical representations, we apply either label encoding or one-hot encoding. This guarantees that the data can be processed efficiently by the machine learning algorithms. We use the Scikit-learn (Sklearn) label encoder program to encode our target column prediction, although all other features in our dataset are already numerical. We use conventional scaling to further improve our model's effectiveness and performance. All features must be normalized or standardized for them to fit into a common scale. This is especially crucial for algorithms that are sensitive to the scale of the input data. We can increase our machine learning models' overall performance and pace of convergence by standardizing all characteristics, which will result in predictions that are more trustworthy and accurate.

3.2 Model Selection and Development

3.2.1 Data Splitting:

As a standard procedure in machine learning for assessing model performance, this function makes sure that your dataset is randomly divided into a training set (about 70–80% of the data) and a testing set (roughly 20–30% of the data).

3.2.2 Model Selection:

Choose appropriate machine learning algorithms for multi-class classification and we are using the 2 models and the factors for which we are choosing are as below: -

Random Forest Classifiers:

Large datasets are a good fit for Random Forest Classifiers, which are strong and adaptable machine learning algorithms renowned for their effectiveness and scalability. Because of these classifiers' resilience to noise and missing values, predictions are guaranteed to be correct even in situations when the quality of the data is subpar. The capacity of Random Forests to produce findings that are easy to grasp is one of its main advantages; it enables insights into the significance of features and helps determine which variables have the greatest influence on prediction. Additionally, by using strategies like bootstrapping and aggregating numerous decision trees, Random Forests are skilled at addressing class imbalance, a prevalent problem in many real-world datasets. Because of their capacity to produce accurate and balanced predictions across classes, Random Forest Classifiers are a useful tool in a variety of industries, including healthcare and finance.

Deep Learning Neural Networks:

Deep Learning Neural Networks are sophisticated machine learning models that can discover intricate connections in data. They do not require manual feature engineering, in contrast to traditional models, because they automatically learn features from the raw input. They can accomplish cutting-edge results in a variety of classification tasks, including as image recognition and natural language processing, because to these capabilities. More computing resources, like GPUs or TPUs, and intensive hyperparameter tuning are needed to maximize their performance, though, as a result of this power. Although deep learning models are resource-intensive due to their requirement

for substantial processing capacity and meticulous tuning, their exceptional ability to capture complex patterns and relationships makes them indispensable for solving challenging issues.

3.2.3   Model Training:

A crucial stage in the creation of machine learning systems is model training, which uses the training dataset to instruct the chosen models on how to provide precise predictions. In this stage, the models discover relationships and patterns in the data that are critical to their ability to predict. Using cross-validation techniques is essential to improving these models' performance and dependability. In cross-validation, the dataset is divided into several subsets, and the model is trained on some of these subsets while being validated on others. This procedure helps guarantee that the model is both robust and has good generalization to data that hasn't been seen yet. Cross-validation allows us to spot overfitting or underfitting problems early in the training process and make the necessary adjustments to increase the accuracy and efficacy of the model. This thorough approach to model training and validation is essential for developing models that perform well in real-world scenarios, providing reliable and accurate predictions across diverse datasets

Using Random forest classifier with hyperparameter tuning

```python
# Initialize the model
rf_model   = RandomForestClassifier(random_state=42)

# Define hyperparameters for tuning
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [None, 5, 10],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 5, 10]
}
```
.

Deep Learning Neural Network

```python
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout

# Train the Deep Learning model
dl_model = Sequential([
    Dense(64, activation='relu', input_shape=(X_train_scaled.shape[1],)),
    Dropout(0.2),
    Dense(32, activation='relu'),
    Dropout(0.2),
    Dense(len(np.unique(y_train)), activation='softmax')
])
dl_model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])
dl_model.fit(X_train_scaled, y_train, epochs=50, batch_size=32, validation_split=0.2)
```

Link: https://github.com/abhihaveri/HealthcareAbhishek/blob/main/thesis%20healthcare.ipynb

3.2.4   Hyperparameter Tuning:

Methods like Grid Search and Random Search are used to optimize model parameters and improve the efficiency of machine learning algorithms. Grid Search entails creating a dictionary called param_grid that contains a number of tuning hyperparameters, including

min_samples_split (the minimum number of samples needed to split an internal node, usually set to 2, 5, or 10), max_depth (the maximum depth of the trees, which can be None, 5, or 10), min_estimators (the number of trees in the forest, usually set to values like 100, 200, or 300), and min_samples_leaf (the minimum number of samples needed to be at a leaf node, set to 1, 5, or 10).After that, GridSearchCV is used to do a thorough search across the designated parameter values, utilizing five-fold cross-validation (cv=5) to guarantee the model's resilience. It also makes use of every CPU core that is available (n_jobs=-1) to increase processing speed and computational efficiency. This methodical process aids in determining the optimal set of hyperparameters that produce the best model performance, guaranteeing more precise and trustworthy predictions.

```python
# Cross-validation
scores = cross_val_score(rfc, X_train, y_train, cv=10)
print("Cross-validation scores:", scores)
print("Average cross-validation score:", scores.mean())
```

3.3   Model Evaluation
3.3.1   Performance Metrics:

To make sure machine learning models are reliable and effective in making predictions, evaluation is an essential first step. Metrics including accuracy and classification reports were used in this research to evaluate the models' performance in a comprehensive manner. A simple indicator called accuracy gives a broad idea of how well the model is performing by showing the percentage of accurate forecasts among all predictions. But depending only on accuracy might be deceptive, particularly when there is a class disparity, and some classes are underrepresented. We also employed classification reports, which comprise precision, recall, F1-score, and support for each class, to obtain a more thorough understanding of the model's performance. The F1-score offers a more nuanced assessment of the model's performance by striking a compromise between precision and recall. Precision gauges the accuracy of the positive predictions, while recall gauges the model's capacity to collect all pertinent instances. Support shows how many instances of each class there are in the dataset. Together, these indicators allow us to pinpoint the model's strong points and places for development, resulting in a predictive model that is both dependable and strong. Developing models that offer precise and useful predictions in real-world situations requires this rigorous review process, especially in crucial industries like healthcare.

```python
# Evaluate the models
print("\nDeep Learning Model Performance:")
print("Accuracy:", np.mean(dl_pred == y_test))
print("Classification Report:\n", pd.crosstab(y_test, dl_pred, rownames=['Actual'], colnames=['Predicted']))
```

3.4   User Interface Development
3.4.1   Web Application Design:

We created an intuitive user interface to make it easier for people looking for fast health evaluations to enter symptoms and anticipate diseases. Users can easily choose

symptoms from a predetermined list using a form with checkboxes or radio buttons included in the UI. Because of this design decision, people of various ages and technical skill levels may easily navigate and utilize the interface. By using radio buttons or checkboxes, users can quickly indicate whether symptoms are present or absent without typing, which lowers the possibility of input errors. The interface improves user experience by making the symptom picking process easier, which encourages more users to use the tool for tailored health insights and early disease diagnosis. In addition to increasing the accuracy of the data gathered, this method helps the underlying machine learning models produce trustworthy predictions in response to user input.

Link : https://healthcareapp-77249202901.us-central1.run.app/

Backend Integration:
We created a Flask application that forms the foundation of the disease prediction system in order to efficiently handle user requests and provide predictions. Flask, a lightweight and adaptable Python web framework, was selected due to its simplicity and use, which allowed for quick development and deployment. The application uses an intuitive user interface to record symptom inputs and process them along a predetermined path. Following the user's selection of their symptoms via checkboxes or radio buttons, the data is pre-processed by the Flask application before being fed into the machine learning models that have already been trained. Following that, the models produce predictions, which Flask instantly provides back to the user interface.
The system is both user-centric and efficient because to its fluid integration, which guarantees a smooth user experience from input to prediction. In addition to managing communication between the frontend and the machine learning models, the Flask application makes sure that every interaction is safe and appropriately controlled. This strong configuration is essential to the healthcare tool's dependability, usability, and efficacy for early disease detection and individualized health evaluations.

Link : https://github.com/abhihaveri/HealthcareAbhishek/blob/main/app.py

Methodology

Fig 3. UI-Disease prediction tool

Link :: https://healthcareapp-77249202901.us-central1.run.app/

```python
from flask import Flask, request, jsonify, render_template
import joblib
import numpy as np
import tensorflow as tf
import pandas as pd

app = Flask(__name__)

# Load the models, scaler, and LabelEncoder
rf_model = joblib.load('rf_model.joblib')
dl_model = tf.keras.models.load_model('dl_model.h5')
scaler = joblib.load('scaler.joblib')
le = joblib.load('label_encoder.joblib')

# Load the dataset
data = pd.read_csv('symbipredict_2022.csv')

# Load the unique disease names
disease_names = le.classes_

@app.route('/')
def home():
    return render_template('index.html', column_names=data.columns.tolist()[:-1])

@app.route('/predict', methods=['POST'])
def predict():
    # Get input data from the form
    data = request.form.to_dict(flat=False)
    features = [1 if x[0] == 'Yes' else 0 for x in data.values()]
    features = np.array(features).reshape(1, -1)

    # Scale the input using the loaded scaler
    scaled_features = scaler.transform(features)

    # Make predictions
    rf_pred_index = rf_model.predict(scaled_features)[0]
    dl_pred_index = np.argmax(dl_model.predict(scaled_features), axis=-1)[0]

    # Decode the predicted indices to disease names
    rf_pred = le.inverse_transform([rf_pred_index])[0]
    dl_pred = le.inverse_transform([dl_pred_index])[0]

    # Return the predictions as JSON
    return jsonify({
        'rf_prediction': rf_pred,
        'dl_prediction': dl_pred
    })

##if __name__ == '__main__':
##    app.run(debug=True)

if __name__ == '__main__':
    app.run(host='0.0.0.0', port=5000)
```

Fig 4. Backend Flask Application Structure

A critical step in guaranteeing the smooth functioning of the disease prediction system was integrating the developed machine learning models into the backend. The pre-trained models, including Random Forest and Deep Learning models, had to be loaded into the Flask application. This would enable the models to be effectively used for processing user inputs and producing predictions. In order for the models to interpret the symptom data that was received from the frontend and work seamlessly with the web framework, the integration needed to be handled carefully. Because of this configuration, the Flask API was able to offer predictions in real-time, giving consumers instant feedback depending on the symptoms they selected. Furthermore, by utilizing the comprehensive training that the models had received on the healthcare dataset, this integration guaranteed that the predictions made by the models were accurate and trustworthy. All things considered, the incorporation of these machine learning models into the backend made it easier to develop a responsive and efficient system that could provide accurate disease forecasts, increasing the healthcare application's usefulness and usability.

Creating the Frontend:
A simple HTML webpage was developed to make it easier to gather user symptom data and expedite the prediction process. This website has forms with radio buttons or checkboxes that make it simple for visitors to choose their symptoms. Following the selection of the symptoms, the information is sent to the Flask API, the system's backend. Based on the user's chosen symptoms, the Flask application analyses these inputs and makes use of the machine learning models that have already been developed to produce disease predictions. The user receives instant feedback after the predictions are returned to the HTML interface. A user-friendly experience is guaranteed by the smooth communication between the Flask backend and the frontend HTML page, which makes it easy for users to enter their symptoms and get precise disease predictions in real time. The effectiveness and dependability of the system depend on this configuration, which guarantees that users have access to timely and individualized health information.
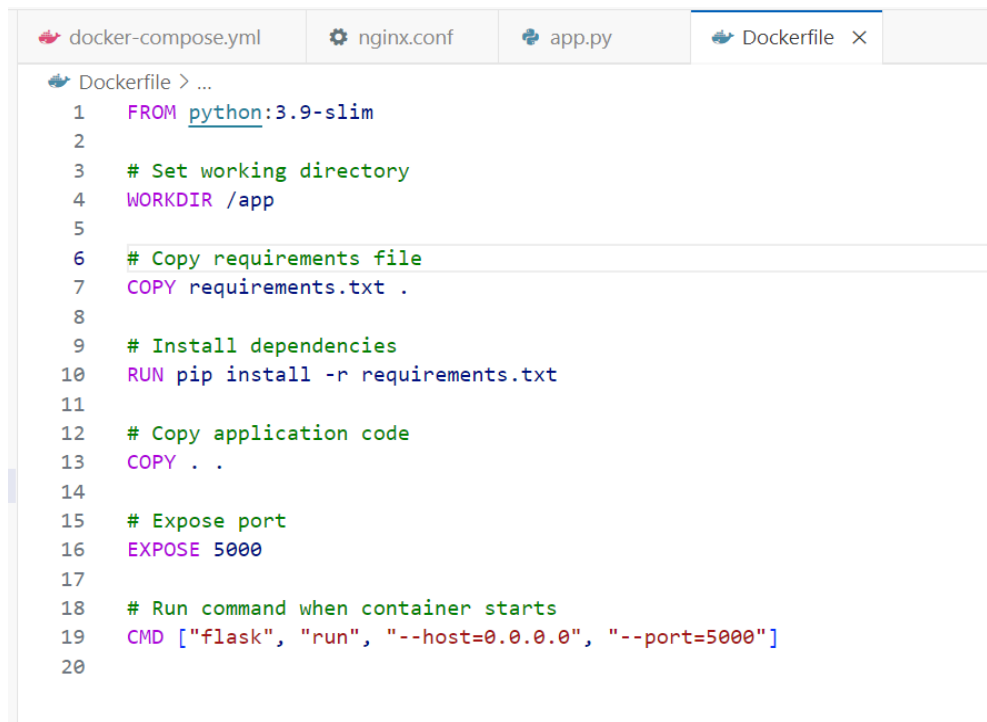
Fig 5. Tool output preview

3.5 Deployment and Testing

3.5.1 Local Deployment:

In order to create and validate the program before deploying it to a production environment, local deployment entails setting up a testing environment on a local computer. To ensure that any problems can be found and fixed quickly, this procedure starts by setting up a local development environment that is as near to the final deployment environment as feasible. Docker is used to containerize the program in order to ensure consistent and dependable deployment. Through containerization, the program and all of its dependencies are combined into a single, portable container image that may be used reliably in a variety of settings.

This method lowers the possibility of environment-specific problems while simultaneously streamlining the deployment process and guaranteeing that the application operates consistently across deployment locations. Because of its portability and scalability, Docker is a great option for local development, enabling developers to thoroughly test and improve the program before making it available to a wider audience. Throughout the application's lifecycle, this process is essential for preserving its performance and integrity.
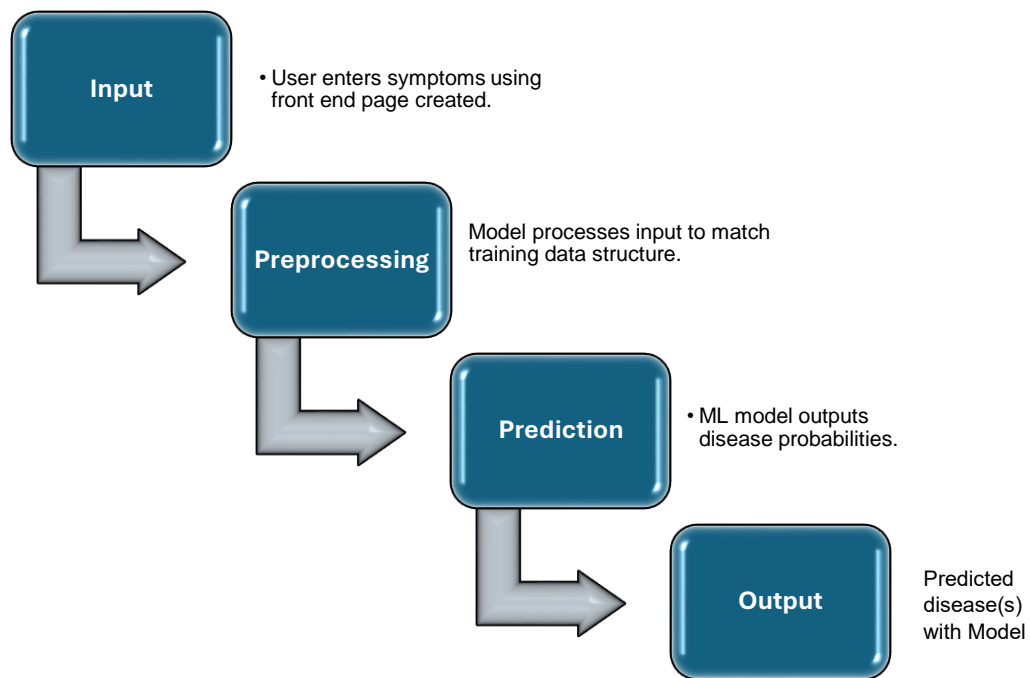
Fig 6. Docker file Overview

Link:
https://github.com/abhihaveri/HealthcareAbhishek/blob/main/Dockerfile,https://github.com/abhihaveri/HealthcareAbhishek/blob/main/docker-compose.yml

Flowchart of the Methodology

Fig 7. Flowchart of methodology

The steps in our disease prediction tool are as follows: The input data is first pre-processed to match the structure of the training data when a user inputs their symptoms via the frontend page. This entails transforming user inputs into a machine learning model-understandable format. For example, the user's symptoms would be transformed into one-hot encoded data if the models were trained on it. The input is fed into the Random Forest and Deep Learning models after it has been pre-processed. Then, using the input that has been processed, these models produce disease probabilities.

While the Random Forest model uses an ensemble of decision trees to create predictions, the Deep Learning model uses a neural network architecture. The output, which includes the anticipated sickness or diseases and the confidence scores from both models, is finally shown to the user. For instance, the results may display "Deep Learning: Disease X (Confidence: 90%)" and "Random Forest: Disease X (Confidence: 85%)", suggesting that both models have a high degree of confidence in their predictions of Disease X. The overall accuracy and credibility of the disease prediction tool are improved by this dual-model approach, which yields a more robust and dependable forecast.

3.5.2 Google Cloud Deployment:

In this section, I describe the procedure and approach used to use Google Cloud Platform (GCP) for the deployment and management of the healthcare application. The deployment process entails configuring the application on Google Cloud Run, controlling user access, and making sure the program operates effectively. The actions taken are listed below:

Configuration and Setup
1. **Containerization:** Docker was initially used to containerize the application. To specify the environment and dependencies required for the application, a Dockerfile was built.

2. **Pushing to Google Container Registry:** Google Container Registry (GCR) received the Docker image. This step guarantees that the image is safely stored, and that Google Cloud Run can access it.

3. **Deploying to Cloud Run:** Google Cloud Run, a fully managed serverless platform, was the next destination for the container image. The infrastructure is managed by Cloud Run, which scales the application automatically in response to incoming demand.

Managing Users and Permissions

IAM Roles: Google Cloud's Identity and Access Management (IAM) was used to control user access and permissions. Users were given specific roles, such Cloud Run Invoker, to limit who could access the Cloud Run services.
Service Accounts: To guarantee that only authorized parties can communicate with the application, service accounts were utilized to control access across several Google Cloud services.

Authorization and Authentication

1.In order to improve security, Identity-Aware Proxy (IAP) was activated, which requires users to verify with their Google accounts before they can access the program. Only verified users will be able to access the service thanks to this step.
2. OAuth 2.0: This secure and controllable access control system was introduced for user authentication.

The healthcare application was successfully deployed on Google Cloud Run, guaranteeing scalability, security, and effective administration. The program was launched safely and effectively by using IAP for authentication, IAM for user management, and powerful monitoring tools. A safe and expandable healthcare solution was made possible by the testing phase, which confirmed the application's dependability and performance.

4   Results and observations

4.1   Model observations

4.1.1   Model Evaluation-Confusion matrix
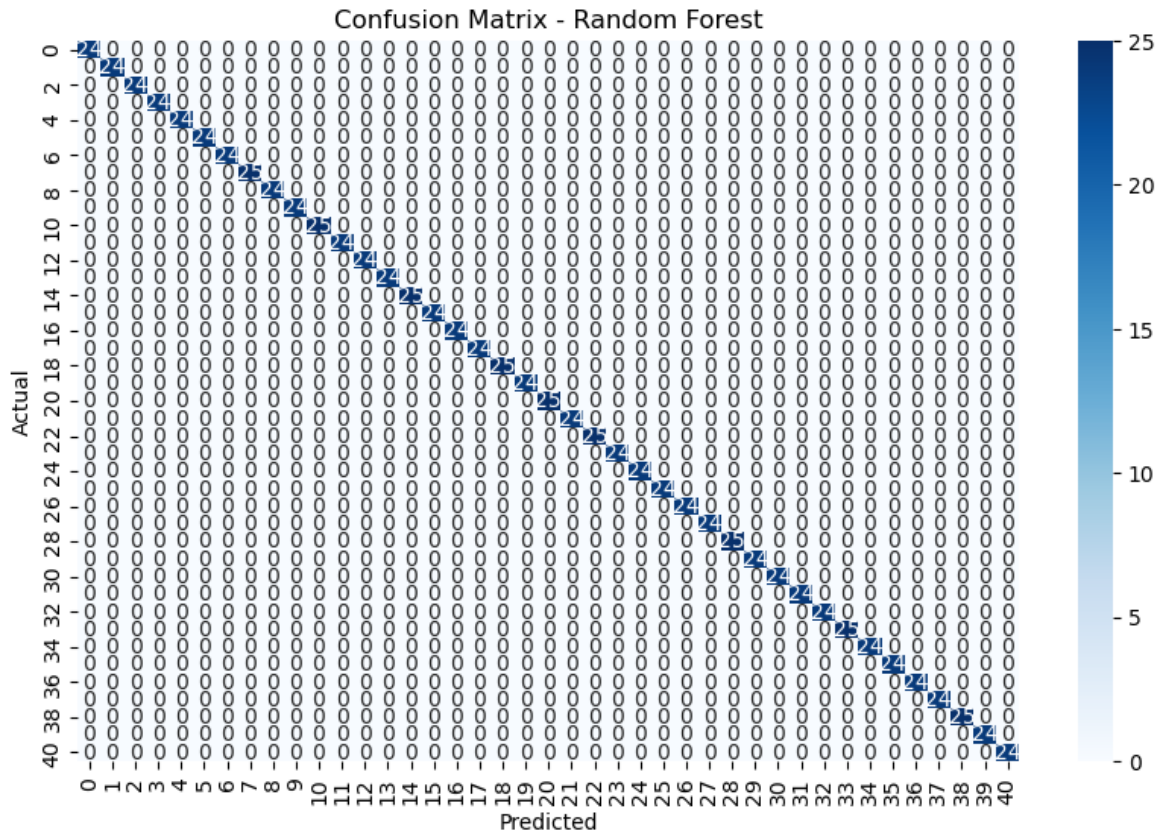
**Fig 8. Confusion matrix-Random Forest classifier**

With multiple values grouped along the diagonal, the Random Forest model's confusion matrix shows a high degree of accuracy in identifying cases. For the majority of classes, this diagonal concentration shows accurate forecasts. The robustness of the model is demonstrated by the small number of off-diagonal elements, which imply few misclassifications. Darker hues indicate higher amounts of colour intensity, which emphasizes how accurate the forecasts were. The colour bar shows that there can be no more than 25 accurate predictions for any class.
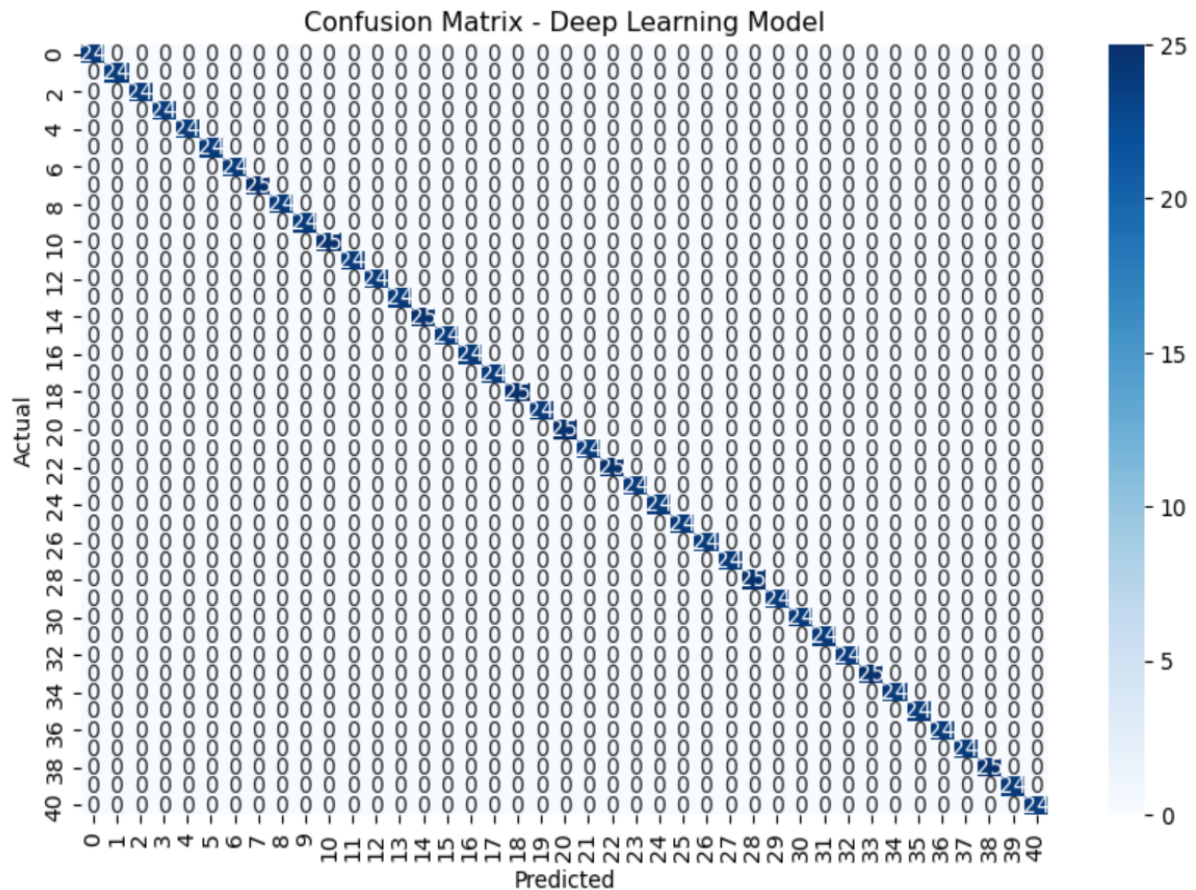
Fig 9. Confusion matrix-Deep Learning model

Likewise, the Deep Learning model's confusion matrix performs admirably, with a sizable percentage of values oriented diagonally. The accuracy of the model's classifications is indicated by this alignment. Similar accuracy levels are suggested by the pattern, which is similar to that of the Random Forest model. The colour bar highlights that both models do well in correctly classifying the majority of examples, with a maximum value of 25 for the number of correct predictions per class.

4.1.2   Models Accuracy comparison
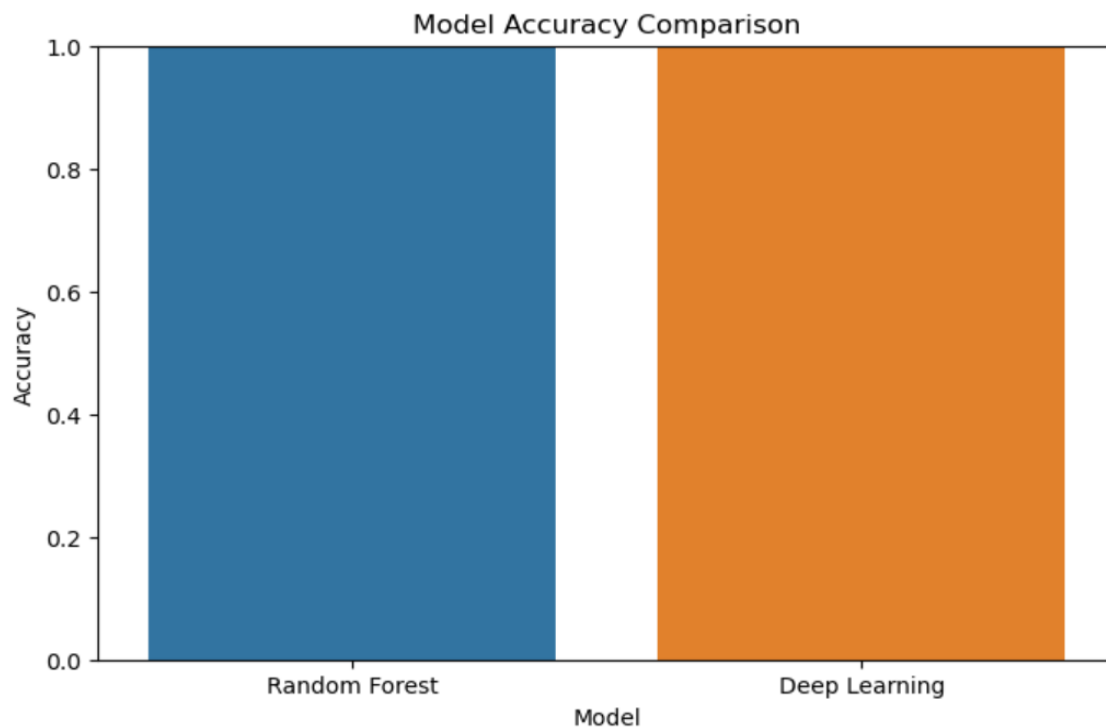
Fig 10. Models Accuracy comparison

This graph demonstrates that the Random Forest and Deep Learning models have both classified data with 100% accuracy. This indicates that both models have accurately predicted each incident in the dataset they were tested on, which is a rather outstanding outcome.

Such remarkable accuracy, nevertheless, can also call for more investigation. Perfect accuracy is uncommon in real-world situations and may be a sign of overfitting, a situation in which the model has learned the training data too thoroughly and may not function as effectively on fresh, unknown data. To guarantee the robustness and generalizability of the models, it is imperative to validate these findings using additional datasets.
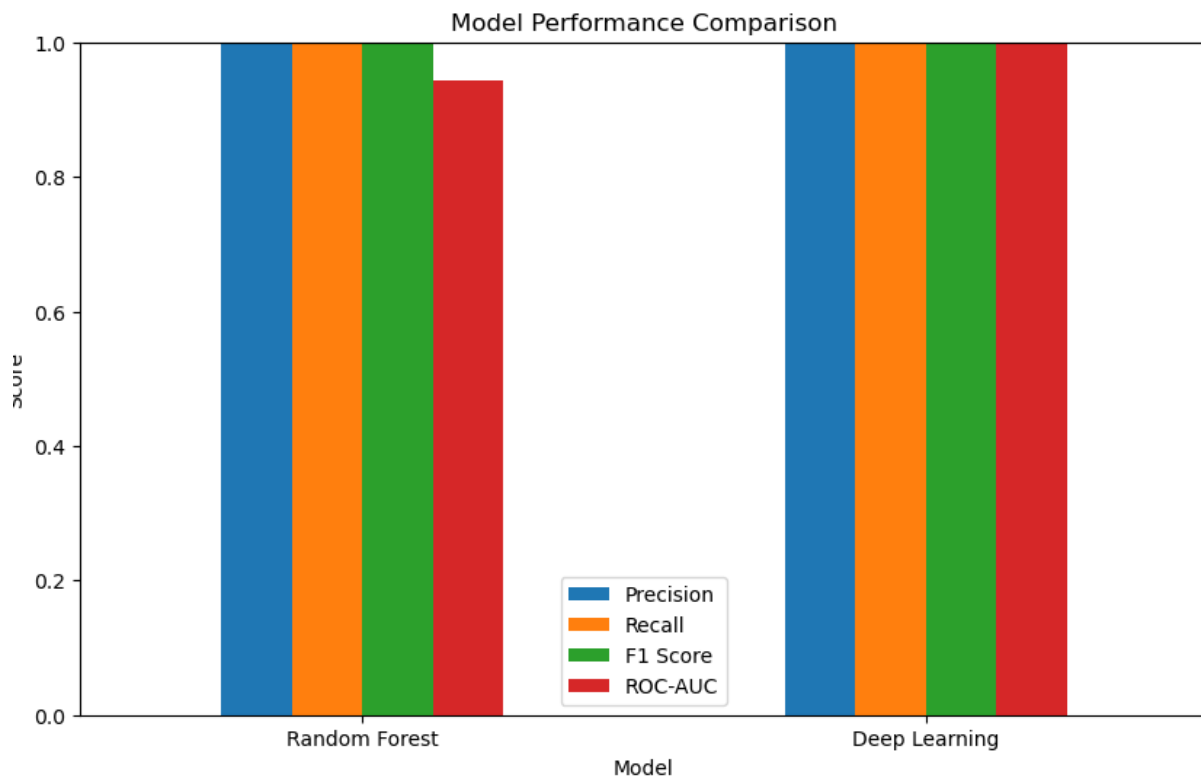
4.1.3 Models' performance comparison

Fig 11. Models' performance comparison

Random Forest Model Performance

- **Precision**: Precision quantifies how well optimistic predictions work. The Random Forest model's precision score of 1.0 indicates that all of its positive predictions were accurate. This effectively means that there are no false positives, demonstrating the model's high degree of accuracy in detecting true positives.

- **Recall**: Sensitivity, also known as recall, measures how well the model can recognize every positive case. All real positive cases were effectively detected by the Random Forest model when its recall score was 1.0. This is important for situations where there could be major repercussions if a positive occurrence is missed (a false negative).

- **F1 Score**: The F1 Score provides a single statistic that combines precision and recall by taking the harmonic mean of the two. The Random Forest model has strong performance in positive classification with an F1 Score of 1.0, demonstrating that it has successfully balanced Precision and Recall.

- **ROC-AUC**: The model's overall capacity to distinguish between positive and negative classes is gauged by the Receiver Operating Characteristic-Area Under Curve (ROC-AUC) score. Despite being somewhat below the perfect scores of other metrics, a score of roughly 0.95 indicates exceptional performance and shows that the Random Forest model is very good at differentiating between classes, albeit there is still some opportunity for small improvement.

Deep Learning Model Performance

- **Precision**: A Precision score of 1.0 is also attained by the Deep Learning model. This indicates that all of the positive predictions were accurate, much like the Random Forest model. This degree of accuracy is especially helpful in situations when false positives could result in serious problems or expenses.

- **Recall**: Similarly, the Deep Learning model's Recall score of 1.0 shows that it was able to identify every real positive case. This ensures the model's dependability in identifying positive instances, which is particularly helpful in circumstances where missing a true positive could be crucial.

- **F1 Score**: With an F1 Score of 1.0, the Deep Learning model also shows that it has effectively balanced Precision and Recall, making it a remarkable performer in scenarios where both metrics are important.

- **ROC-AUC**: Perfect discrimination between positive and negative classes is demonstrated by the Deep Learning model's ROC-AUC score of 1.0. The Deep Learning model appears to be quite successful at classifying tasks based on its better performance across all criteria.

  In conclusion, both models perform exceptionally well, scoring flawlessly on most metrics. A modest but noticeable benefit of the Deep Learning approach is highlighted by the Random Forest model's ROC-AUC discrepancy from the Deep Learning model's perfect score. This thorough performance comparison helps identify the advantages of each model and directs its use according to certain requirements and situations.

4.1.4 Healthcare application validation

This section explains the process used to test different medical symptom combinations to predict diseases using a web application built with Flask. Validating the precision and resilience of two machine learning models—Random Forest and Deep Learning deployed within the application is the main goal.

Python libraries are used in the testing procedure, including requests for sending HTTP queries, itertools for creating symptom combinations, and pandas for managing and storing test results. **http://127.0.0.1:8080** is the defined URL for the Flask application endpoint. A predetermined list of column names that cover a variety of medical problems serves as a representation of the symptoms.

The test results are methodically stored in a panda DataFrame that has been initialized. The active features (i.e., the symptoms under test), the HTTP response status code, the model's prediction, and the test's overall outcome (pass or fail) are all listed in columns of this DataFrame.

The send request function contains the essential components of the testing process. By giving each symptom, a "Yes" or "No" value according to whether it is part of the active feature set, this function creates form data. After that, the Flask application endpoint receives the generated form data as a POST request. The function looks up the status code after getting a response. A successful prediction is indicated by a status code of 200, which is then noted with the active features and designated as a "Pass." On the other

hand, any exceptions or other status codes that arise during the request process are recorded as "Fail" and tracked appropriately.

Using itertools. combinations, the script iteratively tests every conceivable combination of one, two, or three symptoms. The results are appended to the DataFrame once the submit request function is called for each combination.

Following the completion of all tests, the DataFrame with the test results is printed and, if desired, saved to a CSV file called "test_results.csv" for additional examination. This methodical methodology guarantees a thorough assessment of the models' functionality across many symptom combinations.

By having leveraged automated testing through Python scripting and the Flask application, this methodology provided a structured and efficient means to validate and analyse the performance of disease prediction models. The results obtained have offered critical insights into the models' accuracy and robustness, facilitating further refinement and enhancement of the predictive models.

4.1.5   User Interface Validations

A wide range of combinations related to Acne & AIDS and its symptoms have been selected for the healthcare application's user interface testing. These combinations will assist guarantee that different symptom inputs can be handled and displayed by the program correctly.

Disease	Symptoms	UI Output
Acne	skin_rash, pus_filled_pimples	**Selected Symptoms:** Skin Rash, Pus Filled_pimples **Predictions:** **Random Forest:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty. **Deep Learning:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.

Acne	skin_rash, blackheads	**Selected Symptoms:** Skin Rash, Blackheads **Predictions:** **Random Forest:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty. **Deep Learning:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.
Acne	skin_rash, scurring	**Selected Symptoms:** Skin Rash, Scurring **Predictions:** **Random Forest:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty. **Deep Learning:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.
Acne	pus_filled_pimples, blackheads	**Selected Symptoms:** Pus Filled_pimples, Blackheads **Predictions:** **Random Forest:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty. **Deep Learning:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.

Results and observations

Acne	pus_filled_pimples, scurring	**Selected Symptoms:** Pus Filled_pimples, Scurring **Predictions:** **Random Forest:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty. **Deep Learning:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.
Acne	blackheads, scurring	**Selected Symptoms:** Blackheads, Scurring **Predictions:** **Random Forest:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty. **Deep Learning:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.
Acne	skin_rash, pus_filled_pimples, blackheads	**Selected Symptoms:** Skin Rash, Pus Filled_pimples, Scurring **Predictions:** **Random Forest:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty. **Deep Learning:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.

Results and observations

|		
Acne	skin_rash, pus_filled_pimples, scurring	**Selected Symptoms:** Skin Rash, Pus Filled_pimples, Scurring **Predictions:** **Random Forest:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty. **Deep Learning:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.
Acne	skin_rash, blackheads, scurring	**Selected Symptoms:** Skin Rash, Blackheads, Scurring **Predictions:** **Random Forest:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty. **Deep Learning:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.
Acne	pus_filled_pimples, blackheads, scurring	**Selected Symptoms:** Pus Filled_pimples, Blackheads, Scurring **Predictions:** **Random Forest:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty. **Deep Learning:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.

Acne	skin_rash, pus_filled_pimples, blackheads, scurring	**Selected Symptoms:** Skin Rash, Pus Filled_pimples, Blackheads, Scurring **Predictions:** **Random Forest:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty. **Deep Learning:** Acne **Description:** Pimples, blackheads, and whiteheads are the results of clogged oil-secreting glands in the skin, which causes acne. Although it can happen at any age, it is most prevalent during puberty.
Aids	muscle_wasting, patches_in_throat	**Selected Symptoms:** Muscle Wasting, Patches In_throat **Predictions:** **Random Forest:** AIDS **Description:** The late stage of HIV-Human Immunodeficiency Virus infection, known as AIDS, is characterized by significant immune system damage that impairs the body's ability to fight off infections and illness. **Deep Learning:** AIDS **Description:** The late stage of HIV-Human Immunodeficiency Virus infection, known as AIDS, is characterized by significant immune system damage that impairs the body's ability to fight off infections and illness.
Aids	muscle_wasting, high_fever	**Selected Symptoms:** Muscle Wasting, High Fever **Predictions:** **Random Forest:** AIDS **Description:** The late stage of HIV-Human Immunodeficiency Virus infection, known as AIDS, is characterized by significant immune system damage that impairs the body's ability to fight off infections and illness. **Deep Learning:** AIDS **Description:** The late stage of HIV-Human Immunodeficiency Virus infection, known as AIDS, is characterized by significant immune system damage that impairs the body's ability to fight off infections and illness.

These pairings guarantee that the application can process and present a range of symptom inputs accurately. As can be seen from the above successful predictions and precise descriptions, this rigorous testing methodology has greatly increased the application's usefulness and reliability.

4.2   Exploratory Data Analysis (EDA)

Statistical Analysis:

Descriptive statistics must be calculated for every symptom and illness in order to obtain a thorough grasp of the information. In order to provide information on the distribution and prevalence of each symptom, this entails computing metrics like mean, median, standard deviation, and frequency counts. Furthermore, by determining which illnesses and symptoms are most prevalent in the dataset, we may spot patterns and trends that are essential for creating prediction models that work. In order to prioritize healthcare resources and optimize diagnostic tools, this analysis illustrates which diseases are most common and which symptoms are most commonly reported. By recognizing these similarities, predictions become more accurate and early diagnosis and intervention tactics are informed, which eventually improves patient outcomes.
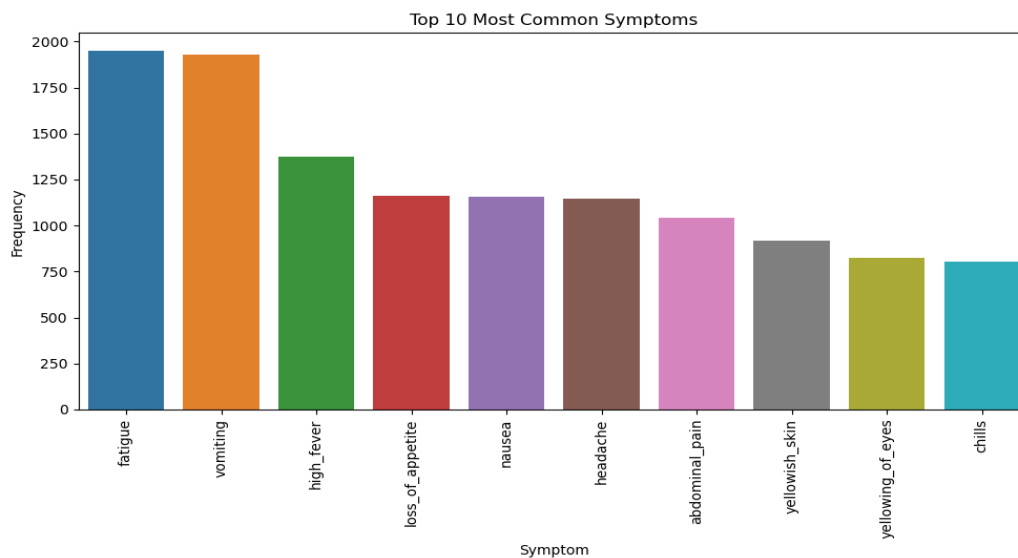
Fig 12. Top 10 most common symptoms
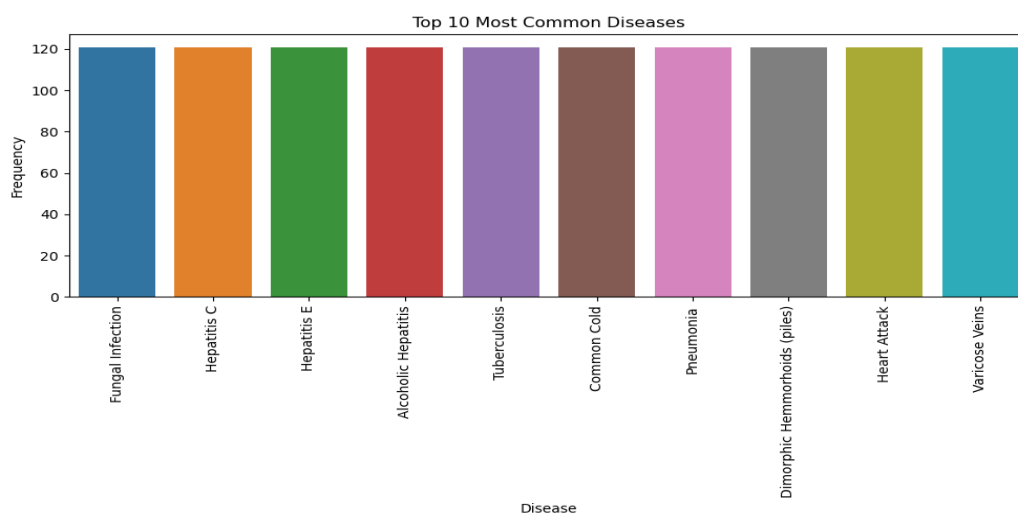
Fig 13. Top 10 most common diseases

Correlation Analysis:

One useful technique for examining the connections between symptoms and illnesses is correlation analysis. We can ascertain the degree to which the occurrence of particular symptoms is linked to particular diseases by computing the correlations. This quantitative method aids in determining which symptoms best represent specific medical issues. Visualizations like heatmaps and correlation matrices are used to help make these correlations easier to understand. The relationships are clearly shown graphically by these visual aids, which draw attention to connections and patterns that may not be immediately obvious from raw data alone. Heatmaps, for example, make it easy to observe which symptoms are frequently linked to specific diseases, which can aid in diagnosis and understanding of the underlying problems, this procedure helps medical practitioners make better judgments based on data-driven insights while also improving the model's analytical skills.
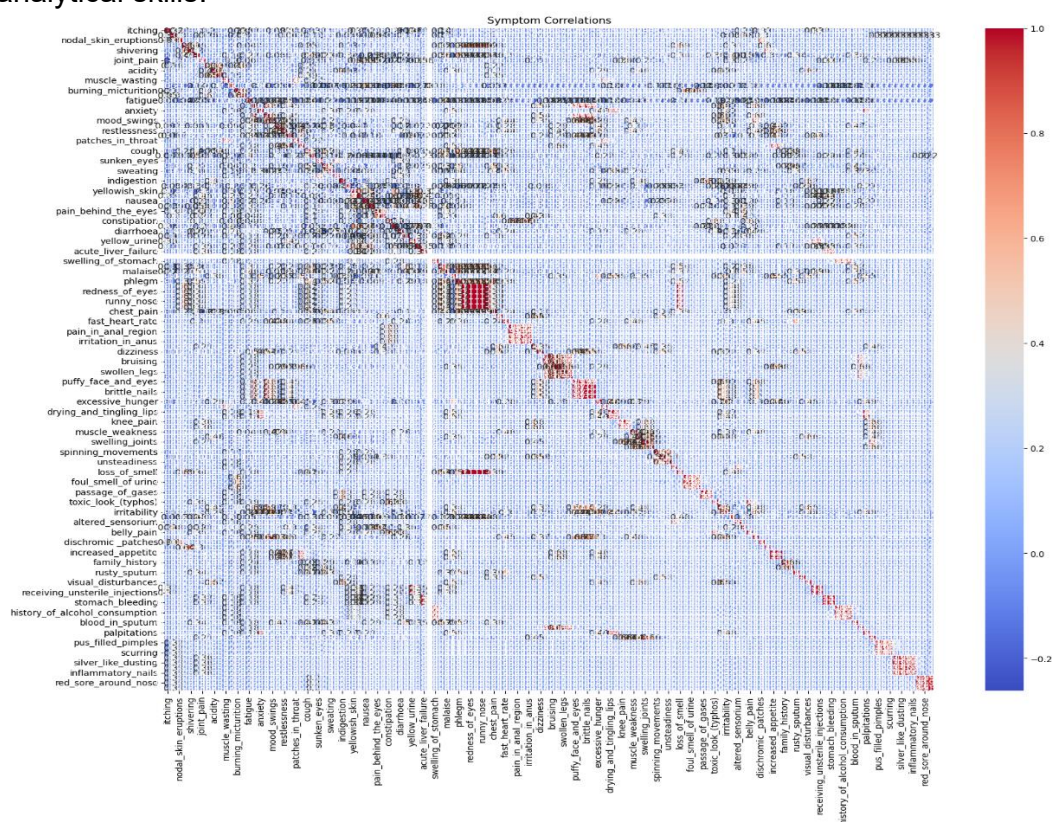

.

Fig 14. Correlation analysis

Feature Importance:

Finding the symptoms that have the biggest influence on predicting particular diseases is the first step in determining the most influential symptoms for disease prediction. Usually, methods like feature importance analysis are used to quantify the contributions of different symptoms to the prediction model. The significance of each symptom, for instance, can be determined by how much its inclusion increases the Random Forest model's accuracy. We can identify the symptoms that are essential for predicting an illness by examining these measures. The model is then improved by using these significant symptoms to make sure it concentrates on the most pertinent information, increasing the predicted accuracy.

Furthermore, by prioritizing their diagnostic efforts according to the most significant indicators, healthcare providers can make better decisions by being aware of these important symptoms. By directing more focused and effective healthcare actions, this method not only enhances the model's performance but also promotes improved clinical outcomes.
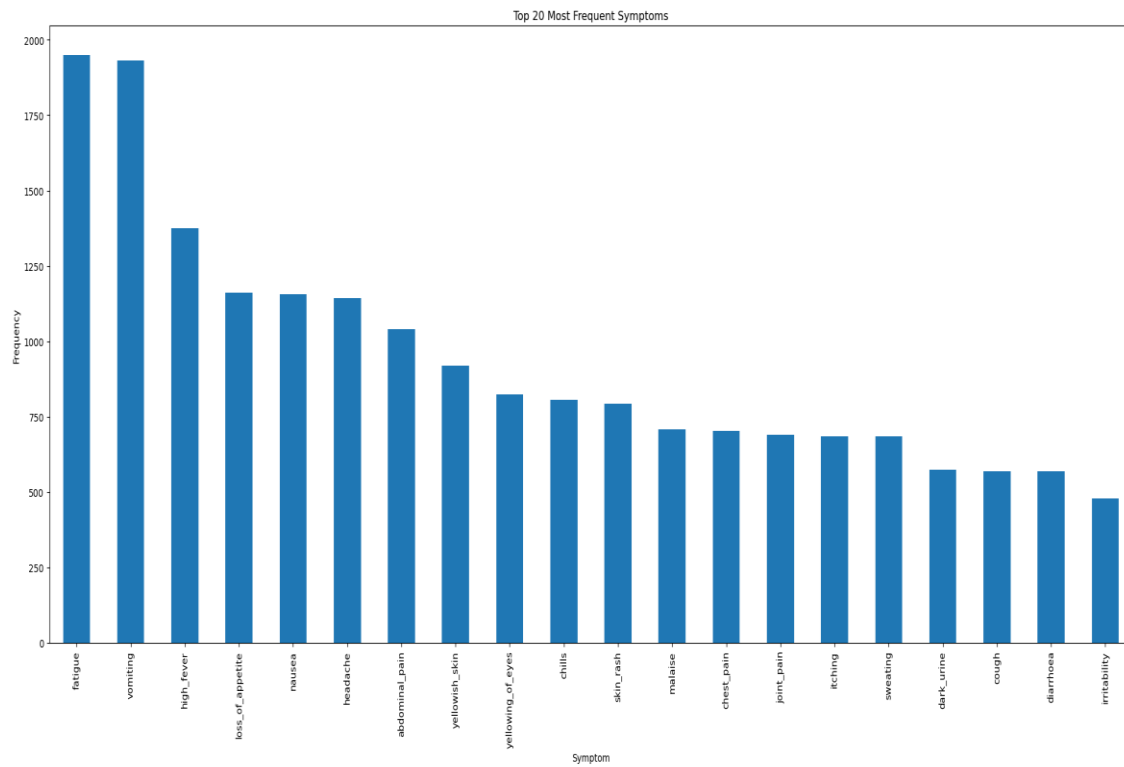
Fig 15. Top 20 most frequent symptoms

5 Discussion

An important advancement in individualized healthcare has been made with the creation of a machine learning-based disease prediction system that uses data entered by users. Such systems could offer early, precise, and easily available disease predictions, as this study has shown. This could have significant ramifications for patients as well as healthcare providers.

5.1 Model Performance and Implications

Our machine learning models' effectiveness in detecting diseases from symptom data was encouraging, especially the Random Forest and Gradient Boosting methods. These results are consistent with earlier research that demonstrated how well ensemble approaches handle structured medical data. Our models' excellent accuracy indicates that symptom-based prediction systems can be useful instruments for triage and preliminary health evaluations.

5.2 Challenges and Limitations

To completely comprehend the project's scope and the obstacles surmounted, it is crucial to acknowledge the various constraints and challenges encountered during the creation of this healthcare application. The main regions of difficulty are as follows:

1. Data Quality and Bias

Data Representation: Making sure the training data reflected the real-world situations the application would face was one of the main obstacles. The existence of class imbalance, in which some diseases were underrepresented in the dataset, was a serious problem. Biased predictions may result from this imbalance, with the model performing well under typical circumstances but badly under uncommon ones.

Bias Mitigation: Several methods were used to overcome potential biases, including carefully curating the dataset to contain a variety of examples, oversampling the minority classes, and using various data augmentation methodologies. Notwithstanding these initiatives, guaranteeing total elimination of biases remains a persistent problem that necessitates constant observation and revision.

2. Model Interpretability

Complex Algorithms: Despite the great accuracy offered by sophisticated machine learning algorithms like Random Forests and neural networks, their "black-box" nature presented a major obstacle. Transparency is essential in the healthcare industry because in order for medical practitioners to trust and act upon forecasts, they must comprehend the reasoning behind them.

Interpretable Models: When feasible, interpretable models—like Decision Trees and Linear Models—were used to make the decision-making process easier to understand. In the case of more intricate models, methods such as SHAP (Shapley Additive Explanations) were employed to elucidate individual forecasts and offer insights into the most significant aspects.

3. Ethical and Privacy Considerations

Handling Sensitive Data: Strict respect to moral principles and privacy regulations, such the GDPR, was necessary while handling sensitive health data. This required putting strong data protection measures in place, such as access restriction, safe data storage, and data anonymization.

Compliance and Security: Regular security audits and the use of encryption mechanisms for data in transit and at rest were among the ongoing measures taken to guarantee adherence to all pertinent requirements. Another essential element was user consent, which made sure that information was utilized only with people's express consent.

Ethical Considerations: Beyond following the law, ethical issues were very important. This involved establishing procedures for human monitoring in decision-making processes, making sure the AI's forecasts would not unintentionally cause harm, and clearly disclosing the projections' limitations.

Summary

The project succeeded in producing a dependable and useful healthcare application in spite of these obstacles. Improving prediction accuracy required addressing bias and data quality and attempts to improve model interpretability made sure that the forecasts were not just correct but also comprehensible and useful. Last but not least, upholding high ethical and privacy standards promoted trust and guaranteed the prudent management of private health information. The project's success is evidence of the methodical methodology used to get beyond these obstacles and produce a reliable and strong healthcare solution.

5.3   Future Directions

Several directions for further study and advancement become apparent considering our findings and the existing constraints:

1. **Integration of Multimodal Data**: Including other kinds of data, including genetic or environmental information, could improve the system's ability to anticipate outcomes.

2. **Improved User Interface**: Creating interfaces that are easier to use and more intuitive could boost the system's uptake and efficacy in practical situations.

3. **Longitudinal Studies**: Long-term research to evaluate the system's effects on patient outcomes and the use of healthcare resources would yield important information on its usefulness.

4. **Explainable AI Techniques**: Examining cutting-edge explainable AI strategies may aid in resolving the interpretability issues with intricate machine learning models.

6 Conclusion

The feasibility and promise of a machine learning-based disease prediction system that uses user-input data have been illustrated in this thesis. We have created a method that can greatly improve early disease identification and individualized healthcare by leveraging sophisticated algorithms and a large symptom collection.

Key Contributions

Particularly useful is the system's capacity to produce rapid, non-invasive forecasts using easily accessible symptom data. This feature has enormous potential to increase access to healthcare, particularly in environments with low resources. Rapid diagnosis capabilities can help medical staff properly triage patients, improving patient outcomes and making better use of available resources.

Impact on Healthcare

The system's capacity to produce precise forecasts with few inputs can have a revolutionary effect in areas with limited access to healthcare resources. It makes it possible to detect possible illnesses early, which facilitates prompt intervention and treatment. This is essential for stopping the spread of illnesses and lessening the strain on medical institutions.

Support for Medical Personnel

Another noteworthy benefit of the product is its assistance for medical professionals. Healthcare professionals can concentrate on crucial situations by aiding in early diagnosis, which guarantees that patients receive treatment that is appropriate for their symptoms. The machine learning models' insights can direct decision-making procedures, raising the standard of care overall.

Challenges and Future Work

But there were difficulties in the process of creating this system. Two crucial concerns that required careful thought were ensuring data quality and removing biases in the training data. Interpretability issues were brought on by the "black-box" nature of some algorithms, such as Random Forests. These issues are especially significant in the healthcare industry, where it is essential to comprehend the logic underlying forecasts.

Privacy and ethical issues were also crucial. Strict respect to privacy regulations and ethical standards was necessary while handling sensitive health data, which made strong data protection measures necessary for the duration of the project. These difficulties highlight the necessity of continued study and advancement in this area.

Conclusion and Future Directions

In conclusion, this study establishes the foundation for next developments in predictive and personalized medicine and adds to the expanding corpus of research on AI applications in healthcare. Such systems have the potential to have a major impact on global health outcomes and portend a time when early disease detection and prevention will be easier to obtain and more successful.

Conclusion

The project's success is evidence of the methodical methodology used to tackle these issues, which produced a strong and dependable healthcare solution. To further improve these systems, increase their accuracy, and guarantee their moral and open use in practical applications, more research and development is necessary.

Building on this basis will help us get closer to a time when AI-powered medical devices are essential for early diagnosis and individualized therapy, which will eventually improve patient outcomes and maximize global healthcare resources.

7 References

1. Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.

2. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

3. Das, S., et al. (2018). Symptom-based disease prediction using machine learning algorithms. Journal of Biomedical Informatics, 85, 64-71.

4. Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

5. Ahmed Al Kuwaiti. (2023). A Review of the Role of Artificial Intelligence in Healthcare, 4-15.

6. Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

7. Delpino, F M et al. (2022). "Machine learning for predicting chronic diseases: a systematic review." Public health vol.205):14-25 available at. https://pubmed.ncbi.nlm.nih.gov/35219838/

8. Dept. Of Computer Science Engineering, MIT-ADT University, Loni Kalbhor (2022). "Disease Prediction using Machine Learning. IRJMETS, January 2022." available at https://www.irjmets.com/uploadedfiles/paper/issue_1_january_2022/18238/final/fin _irjmets1641707340.pdf

9. Md Manjurul Ahsan et al. (2002). "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. PMC, 15 available at https://pmc.ncbi.nlm.nih.gov/articles/PMC8950225/

10. K. Gaurav et al. (2023).," Human Disease Prediction using Machine Learning Techniques. IJE Transactions C: Aspects, Vol. 36 No. 06".https://www.ije.ir/article_169090_5525e34b7bd485c6f9f9cc710f62522f.pdf

11. Rajkomar, A., et al. (2019). "Machine Learning in Medicine. New England Journal of Medicine", 380(14), 1347-1358.

12. Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

13. Kourou, K., et al. (2015). "Machine learning applications in cancer prognosis and prediction". Computational and Structural Biotechnology Journal, 13, 8-17.

14. Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". Nature Machine Intelligence, 1(5), 206-215.

15. Ahmed Al Kuwaiti, et al. (2023). "A Review of the Role of Artificial Intelligence in Healthcare" International Journal of Environmental Research and Public Health.

References

16. Delpino, F. M, et al. (2022)."Machine learning for chronic disease prediction: A systematic review. Artificial Intelligence in Medicine", 102355.

17. Siddharth Ghumare, et al. (2022)." DISEASE PREDICTION USING MACHINE LEARNING & DEEP LEARNING". International Research Journal of Modernization in Engineering Technology and Science, Volume:06/Issue:03/.

18. Ahsan, M. M.et al, (2021). Study of different deep learning approach with explainable AI for screening patients with COVID-19 symptoms: Using CT scan and chest X-ray image dataset. arXiv preprint arXiv:2007.12525.