DEPARTMENT OF TECHNOLOGY
SHIVAJI UNIVERSITY, KOLHAPUR
Academic Year 2023-24

A

MAJOR PROJECT SYNOPSIS ON

"MULTIPLE DISEASE DETECTION"

Submitted To:

(Computer Science and Technology)

Department of Technology,

Shivaji University, Kolhapur

Submitted By:	PRN
Mr. Aditya Jairam Nimbalkar	2020076403
Ms. Vasudha Anil Dake	2020076412
Ms. Vaishnavi Premnath Kamble	2020074632
Mr. Tejas Suresh Shingan	2020076435

Under the Guidance of:

Mrs. A. A. Manjrekar

ABSTRACT

Machine learning in detection multiple diseases, revolutionizing healthcare by enabling early diagnosis and personalized treatments. This synopsis explores the role of machine learning algorithms in leveraging vast patient datasets to uncover subtle disease indicators. While this approach holds immense promise, it faces challenges, including data diversity, algorithmic bias, and ethical considerations. The evolving technology and growing data availability are expected to enhance machine learning's accuracy and ultimately improve patient outcomes. The introduction emphasizes the classification aspect of machine learning and its applicability in healthcare analytics. It highlights the pressing need for predictive analytics in healthcare due to the shortage of medical infrastructure and healthcare professionals. The synopsis further introduces a medical test web application designed to predict diseases like lung cancer, heart disease, and diabetes, aiming to make disease detection accessible to the masses. This work aims to shed light on the burgeoning field of machine learning in disease diagnosis, its approaches, and associated challenges.

Keywords: Disease detection, Supervised Learning, Deep learning, Machine Learning, Lung cancer, Heart disease, Diabetes disease.

INDEX

1. INTRODUCTION

In our contemporary world, the most precious possession we hold is our health. Yet, the timely and accurate detection of diseases such as lung cancer, diabetes, and heart-related conditions remains a daunting challenge. Often, individuals remain unaware of their health status until symptoms escalate, resulting in delayed treatment and escalating healthcare expenses. In response to these pressing issues, our project, "Multiple Disease Detection using Machine Learning," emerges as a ground-breaking solution poised to usher in a new era of healthcare.

At present, our healthcare system relies predominantly on periodic check-ups and labour-intensive manual screening procedures to identify these life-altering diseases. These conventional methods are time-consuming, financially burdensome, and not always adept at capturing diseases in their nascent stages when intervention is most effective.

Our project endeavours to redefine disease detection by harnessing the incredible potential of machine learning. By deploying advanced algorithms and sophisticated data analysis techniques, we can rapidly and accurately process vast volumes of medical data. This empowers us to pinpoint potential health issues at their very inception, often long before outward symptoms manifest, and with unparalleled precision. The implications of our project are nothing short of monumental. Early detection of lung cancer, diabetes, and heart-related conditions means that individuals can receive prompt medical attention, significantly enhancing their prospects of successful treatment while alleviating the strain on healthcare systems. Moreover, it holds the promise of substantial cost savings for both individuals and healthcare providers.

Within the framework of this project, we will delve deeply into the capacities of machine learning models to scrutinize diverse health parameters, medical histories, and genetic information. By providing a user-friendly, dependable tool for disease detection, we aim to empower individuals to take proactive measures to safeguard their health. Simultaneously, our endeavour aligns with the imperative of creating a healthcare system that is more efficient, cost-effective, and accessible to all. Furthermore, our project is not just confined to the realm of diagnosis; it extends to proactive disease prevention and management. By analysing vast datasets, we can identify trends and risk factors, enabling healthcare professionals to develop tailored preventive strategies and personalized treatment plans. This proactive approach can lead to a reduction in the overall prevalence of these diseases, making a significant impact on public health.

"Multiple Disease Detection using Machine Learning" is poised to be a transformative force that transcends the present state of disease detection. By harnessing cutting-edge technology, we aspire to save lives, mitigate healthcare expenses, and elevate the overall well-being of individuals within our society. This project represents a beacon of hope for a healthier and more equitable future for all, emphasizing not only early diagnosis but also proactive prevention and personalized care.

2. LITERATURE REVIEW

1. **Multi Disease Prediction System using Random Forest Algorithm in Healthcare System [R. Shanthakumari; C. Nalini, S. Vinothkumar, E.M. Roopadevi, B. Govindaraj March 2022]**
   They represents the critical importance of dynamic healthcare systems in the face of increasing disease prevalence and population growth. It emphasizes the need for predictive models to aid medical practitioners in diagnosing multiple diseases accurately. The focus on Heart disease, Diabetes, and kidney disease using basic health parameters is commendable. The mention of Python pickling for model creation and performance evaluation using various datasets adds depth to the approach

2. **Leveraging Machine Learning Methods for Multiple Disease Prediction using Python ML Libraries and Flask API [Srikanth Narayanan, N. M. Balamurugan, Maithili K, P. Bini Palas May 2022]**
   This research paper addresses the challenge of disease prediction in Machine Learning within the Computer Science domain. Unlike most existing models focused on individual diseases, this paper introduces a system capable of predicting multiple diseases, including Malaria, Alzheimer's Disease, Tuberculosis, and Pancreatic Cancer. The system employs a combination of CNN, Random Forest, and Logistic Regression algorithms for analysis and prediction. Implementation is facilitated through a Flask API and various Python ML libraries, such as TensorFlow, Scikit Learn, and Pandas. The paper aims to offer a quick and efficient disease prediction tool for medical professionals.

3. **Disease Prediction Using Machine Learning Techniques [Roop Chandrika Mallela, Reddy Lakshmi Bhavani, B. Ankayarkanni June 2021]**
   This study focuses on leveraging machine learning to predict common diseases based on real symptoms. Researchers collected data on prevalent diseases and used various machine learning algorithms, along with text processing techniques like tokenization, to create a model for accurate disease prediction. The potential applications in the healthcare industry include pre-emptive disease detection, quicker diagnosis, and the ability to review patients' medical histories, all contributing to improved overall health management.

4. **Automated Disease Prediction Using Machine Learning Technology [Neetu Mittal, Hemangi Sharma August 2023]**

   This paper explores the critical issue of inaccurate medical diagnoses, a prevalent cause of medical mishaps. It underscores the challenges faced by healthcare professionals in swiftly diagnosing diseases and assessing symptoms, often due to time constraints and data overload. The paper proposes the application of supervised machine learning (ML) algorithms to overcome these obstacles. By leveraging ML to analyse extensive datasets and uncover hidden patterns, it offers a promising avenue for more precise and personalized diagnoses, potentially outperforming current diagnostic methods. The study evaluates various ML models, including decision trees and Naive Bayes, to identify the most effective algorithm for predicting diseases based on user-provided symptoms.

5. **Multiple Heart Diseases Prediction using Logistic Regression with Ensemble and Hyper Parameter tuning Techniques [Sateesh Ambesange, A. Vijayalaxmi, S. Sridevi, Venkateswaran, B. S. Yashoda] (July 2020)**

   In this study, the researchers addressed the rising prevalence of heart failure and its status as a significant human disease by employing Machine Learning (ML) techniques. They utilized the UCI dataset to train a predictive model for cardiac disease. The researchers first tackled data-related challenges such as imbalance, skewness/kurtosis, and feature relationships through statistical methods. They applied power transformations to normalize skewed features and employed the Turkey Fence technique to eliminate outliers. Feature selection was performed using various methods, including Extra Trees Classifier. Hyperparameter tuning of logistic regression was carried out with random search and grid search. Evaluation metrics like confusion matrix, accuracy score, PRC, and ROC were used to gauge model performance. Among seven models, the one employing power transformation, Kernel PCA, and Gridsearch achieved a remarkable 100% accuracy rate.

6. **Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design [M. Raihan, Saikat Mondal, Arun More, Md. Omar Faruqe Sagor, Gopal Sikder, Mahbub Arab Majumder, Mohammad Abdullah Al Manjur, Kushal Ghosh] (December 2016)**

   In their study, they devised a user-friendly smartphone-based approach to assess the risk of ischemic heart disease (IHD or heart attack). They developed an Android prototype program by

analysing clinical data from 787 IHD patients, considering factors like hypertension, diabetes, abnormal cholesterol, smoking, family history, obesity, stress, and clinical symptoms. Employing data mining techniques, we generated a risk score categorized as low, medium, or high. Our analysis revealed significant associations between cardiac events and the low/high and medium/high risk categories (p-values of 0.0001). Our goal is to offer an accessible means for identifying IHD risk and promoting early cardiac assessment to prevent sudden fatalities, reducing current limitations in risk assessment tools.

7. **Analysis of Different Machine Learning Algorithms Used for Identification of Lung Cancer Disease [Sameeka Saini, Ankit Maithani, Diksha Dhiman, Amita Bisht] (September 2021)**

   This review highlights the critical role of early cancer detection and the impact of rising cancer incidence on mortality rates. The study emphasizes the significance of machine learning and artificial intelligence in cancer research. It discusses the integration of biomedical image processing and knowledge detection techniques, particularly focusing on lung cancer. Various machine learning methods, such as Multinomial Naive Bayes, Logistic Regression, Random Forest, Ridge Classifier, and SGD Classifier, were applied to a lung cancer dataset to assess accuracy, sensitivity, specificity, F1 score, and precision. The primary objective of the research is the analysis of early-stage lung cancer for improved patient outcomes.

8. **Lung Cancer Disease Diagnosis Using Machine Learning Approach [Swati Mukherjee, S. U. Bohra] (December 2020)**

   The analysis and study of lung disorders have been a captivating research area for medical professionals over time. Efforts have been made to create a diagnosis system to reduce the risk to human life by detecting malignant growth early. One approach is using neural network models, particularly deep learning, to identify malignant cells in medical images. An AI and deep neural network-based framework for lung cancer detection has been established, employing supervised learning and CNN classification. This framework involves various steps such as image acquisition, preprocessing, augmentation, segmentation, feature extraction, and neural network identification. Overall, machine learning techniques hold significant potential for improving decision support in cost-effective lung cancer treatment.

9. **Lung Cancer Detection and Classification using CT Scan Image Processing [Nusraat Nawreen, Umma Hany, Tahmina Islam] (July 2021)**

   This study addresses the pressing issue of lung cancer, the second most prevalent cancer after breast cancer, characterized by lower survival rates. Early detection is vital for better prognosis, and the preferred method for this is computed tomography (CT). However, analyzing CT images manually is time-consuming and error-prone. Therefore, the study introduces a novel approach for lung cancer classification and detection through CT image processing. It involves image preprocessing, lung tumor segmentation, and finally, employing a support vector machine (SVM) classifier to categorize lung nodules into benign and malignant cases. The proposed method demonstrates promising accuracy in diagnosing lung cancer nodules and assessing their severity levels.

10. **Machine Learning-based Diabetes Prediction: A Cross-Country Perspective [Sadia Afrin Shampa, Md. Saiful Islam, Ayatun Nesa] (June 2023)**

    This study delves into the challenging task of predicting diabetes, a chronic condition with significant health implications. With diabetes datasets often plagued by data issues like outliers and missing values, the study employs various machine learning (ML) models to analyze data from Bangladesh, India, and Germany. Notably, boosting ML techniques such as AdaBoost, CatBoost, Gradient Boost, and XGBoost excel in forecasting diabetes in the Bangladesh dataset. Additionally, foundational models like Random Forests and Decision Trees exhibit satisfactory performance. Early detection of diabetes, as underscored by this research, can be a game-changer for mitigating its risks and severity. Leveraging ML algorithms and available data, this study advances our understanding of diabetes prediction and underscores the efficacy of boosting algorithms. Its findings have implications for improving patient outcomes and public health through early detection and control measures.

3. PROBLEM STATEMENT

Develop a machine learning system that can detect three common diseases - lung cancer, diabetes, and heart-related conditions - by analysing personal information and relevant medical data. This system should provide accurate and timely diagnoses to assist healthcare professionals in early detection and treatment planning.

4. CHOICE OF THE TOPIC WITH REASONING

The topic of "Multiple Disease Detection using Machine Learning" is about using computer technology to help find illnesses in people early on. Sometimes, it's hard for people to see a doctor because they live far away or hospitals are too busy. Machine learning can look at lots of health information and pictures to find signs of diseases. This can be a big help, especially in places where it's hard to see a doctor quickly.

Using machine learning for disease detection can also make healthcare better for everyone. It can tell us when we might be sick even before we feel really bad. It can also make sure doctors and nurses have more time to help patients because the computer can do some of the checking. Overall, this idea can make healthcare more available, faster, and better for everyone, no matter where they are.

5. OBJECTIVES

1. To Develop a machine learning model to detect lung cancer, diabetes, and heart-related conditions early, even before symptoms appear.
2. To Analyze data to identify trends and risk factors, helping healthcare professionals develop personalized prevention and treatment plans.
3. To Strive for a healthier and more equitable future by making early diagnosis, prevention, and personalized care accessible to all.

6. OUTLINE OF PROPOSED SYSTEM

6.1 Data Collection:

a. Our focus is on prevalent diseases identified by the World Health Organization, which can be challenging and time-consuming to detect. The data comprises actual parameter values from authentic medical tests, making it distinct.

b. We are utilizing real-time data from surveys conducted by various organizations and dataset providers such as UCI Machine Learning Repository, the World Bank, Kaggle, etc. The data gathered is specifically intended for constructing machine learning models and is permitted for educational purposes.

c. To protect privacy, we have anonymized the data by removing personal information. We retained data
parameters essential for constructing machine learning models.

d. Dataset does not include socioeconomic or lifestyle information. It exclusively contains idiosyncratic data parameters necessary for predicting outcomes.

6.2 Data Pre-processing:

6.2.1 Handling Missing Values:

- Deletion: Eliminate rows or columns with a high percentage of missing values if they don't contain critical information.
- Imputation: Mean/Median Imputation: Replace missing numerical values with the mean or median of non- missing values in that column.
- Mode Imputation: Replace missing categorical values with the mode (most frequent category) of the column.
- Forward Fill or Backward Fill: For time-series data, use the value from the previous or next time step to fill missing values.
- Predictive Modelling: Predict missing values using machine learning algorithms based on other features.
- Interpolation: For time-series data, interpolate missing values based on adjacent values.
- Flagging: Create a binary flag column to indicate whether a value is missing, which aids in analysis.

6.2.2.1 Outliers:

- Visual Inspection: Plot data using box plots, scatter plots, or histograms to identify outliers.

- Z-Score or Standard Deviation: Remove data points that fall outside a specified number of standard deviations from the mean.
- IQR (Interquartile Range) Method: Identify outliers as data points below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR and remove them.
- Winsorizing: Cap extreme values by replacing them with a specified percentile value (e.g., 99th percentile)

6.2.2.2 Handling Noisy Data:

- Smoothing: Apply techniques like moving averages or exponential smoothing to reduce noise in time-series data.
- Binning or Discretization: Group data into bins or categories to reduce noise and improve manageability.
- Filtering: Use filters like median filters or Gaussian filters to remove noise from signals or images.
- Data Transformation: Apply mathematical transformations (e.g., logarithm, square root) to reduce the impact of noise.
- Outlier Removal: As outliers can introduce noise, their removal can help reduce overall noise.

6.3 Feature Selection/Engineering:

a. Our goal is to predict diseases based on test parameters.
b. We achieve highly accurate predictions through extensive training and model selection. For instance, we can predict lung cancer levels solely from a chest CT photo, without the need for a doctor's consultation.
c. To assess models, we employ additional modules to measure their efficiency and accuracy. Submodules from scikit-learn provide various metrics like Root Mean Squared Error (RMSE),Confusion Matrix, Mean Absolute Percentage Error (MAPE), Precision, Recall, F1-Score, Accuracy Score, etc.

6.4 Model Selection:

a. By understanding the problem statement and identifying the type of supervised machine learning technique needed, we select different algorithms for specific diseases based on accuracy.
b. The choice of algorithm depends on its suitability and efficiency for the given dataset.
c. Instead of integrating multiple algorithms to improve accuracy, we opt for distinct algorithms for different diseases.

6.5 Model Training:

a. **Training Set:** The training set, typically comprising 60-80% of the total data, trains the machine learning model. It allows the model to learn patterns and relationships in the data, adjusting its parameters for predictions. The training set should represent the overall data distribution and encompass diverse instances for robust generalization.
b. **Validation Set:** This smaller subset (around 10-20% of the data) fine-tunes model hyper parameters, selects the best-performing model, and prevents over fitting. Independence from the training set is crucial to avoid bias or data leakage.
c. **Test Set:** This distinct subset (similar in size to the validation set) evaluates the model's generalization to unseen data and real-world performance. It should not be used during model development to ensure unbiased evaluation.

6.6 Model Evaluation:

a. Evaluation Metrics for Disease Prediction: Specific evaluation metrics are chosen based on the disease prediction task. Common metrics include accuracy, precision, recall, F1-score, ROC-AUC, and area under the precision-recall curve (PR AUC). The choice depends on the clinical significance of false positives and false negatives.
b. Determining Appropriate Threshold Values: Threshold values for classification are selected based on trade-offs between sensitivity (recall) and specificity (true negative rate). Techniques like ROC curve analysis and precision-recall curve analysis guide threshold selection.
c. Challenges in Model Evaluation in Healthcare: Unique challenges in healthcare model evaluation include data privacy, class imbalance, clinical interpretability, ethical considerations, real-world generalization, and validation with limited data.

6.7 Ensemble Techniques:

a. Bagging (Bootstrap Aggregating): Bagging involves training multiple instances of the same base model on different bootstrap samples of the training data. Their predictions are aggregated through averaging or voting, reducing over fitting and improving stability.
b. Stacking (Stacked Generalization): Stacking combines diverse base models by training a meta-model on their predictions. It excels when base models have complementary strengths and weaknesses, capturing complex data relationships.

6.8 Interpretability:

a. Feature Importance Analysis: Techniques like feature importance scores, SHAP, or LIME highlight the impact of features on predictions, aiding in understanding model decisions.
b. Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) Plots: Visualize how predicted outcomes change with varying individual features while holding others constant.
c. Rule-Based Models: Create interpretable decision trees or rule lists to transparently outline decision-making processes.

6.9 Deployment:

a. Choose between cloud or on-premises deployment based on scalability, cost, and data control.
b. Implement data encryption, access control, and secure APIs to protect data and comply with regulations like HIPAA/GDPR.
c. Ensure compliance with data privacy regulations, minimize data usage, and maintain audit trails.

6.10 Continuous Monitoring and Improvement:

a. Continuously monitor healthcare data sources, automate data pre-processing, and assess model performance.
b. Automate data pre-processing steps and employ data drift detection to identify data distribution changes.
c. Schedule regular model assessments and employ automated testing frameworks.
d. Retrain or update the model based on the rate of change in disease patterns and data availability.
e. Consider transfer learning to reduce the data required for retraining.
f. Implement model versioning to track changes during updates.

7. METHODOLOGY

The Methodology Comprising:

a) Methods of Data Collection

- **Data Sources:** Collect data from various sources, including hospitals, clinics, research studies, and publicly available datasets like those from UCI Machine Learning Repository, World Bank data, and Kaggle. Ensure data is relevant to the diseases under consideration and comes from reputable sources.
- **Survey Data:** Utilize real-time survey data conducted by different organizations. Ensure that the data is collected systematically, following established survey methodologies. Survey data can provide valuable insights into patient characteristics and medical history.
- **Ethical Data Handling:** Adhere to ethical guidelines and data protection regulations. Anonymize data by removing personal information, prioritizing patient privacy and confidentiality.
- **Data Exclusion:** Exclude socioeconomic and lifestyle information from the dataset, focusing solely on medical parameters relevant to disease prediction.

b) Probable Methods of Data Analysis

- **Exploratory Data Analysis (EDA):** Begin with EDA to understand the dataset's characteristics. Use statistical and visualization techniques to identify patterns, outliers, and correlations among variables. EDA helps inform feature selection and preprocessing decisions.
- **Feature Engineering:** Create new features, if necessary, based on domain knowledge or insights from EDA. Feature engineering can enhance model performance by capturing meaningful information from the data.
- **Machine Learning Algorithms:** Apply machine learning algorithms suitable for the specific disease prediction tasks. Depending on the nature of the data (structured or unstructured), consider algorithms like logistic regression, decision trees, random forests, support vector machines, or deep learning techniques.
- **Cross-Validation:** Implement cross-validation techniques (e.g., k-fold cross-validation) to assess model performance and robustness. Cross-validation helps estimate how well the model will generalize to unseen data.

- **Model Hyperparameter Tuning:** Fine-tune model hyperparameters using techniques like grid search or random search. Optimizing hyperparameters can improve model accuracy and generalization.
- **Ensemble Methods:** Consider ensemble techniques such as bagging and stacking to enhance prediction accuracy and model stability. Ensemble methods combine multiple models to make collective predictions.
- **Evaluation Metrics:** Evaluate model performance using appropriate classification metrics for each disease prediction task, including accuracy, precision, recall, F1-score, ROC-AUC, and PR AUC. Select metrics that align with clinical priorities and the specific disease's significance.
- **Threshold Determination:** Choose optimal threshold values for classification by considering the trade-off between sensitivity and specificity. Threshold selection is critical in healthcare applications.

8. SYSTEM ARCHITECTURE

8.1 Hardware Requirement:

- Processor: i5 to i9 from Intel and ryzen 5 to ryzen 7000 from AMD Ryzen
- RAM : Min 8GB
- Hard Disk: Min 60GB.

8.2 Software Requirements:

- **Python:** Python is the primary programming language for machine learning and data analysis. You'll need Python installed on your system, preferably version 3.7 or higher.

- **Integrated Development Environment (IDE):**
  - Used Jupyter Notebook, Visual Studio Code (with Python extensions).
  - These IDEs provide an interactive and organized environment for writing code, running experiments, and documenting your project.
  - Jupyter Notebook is particularly useful for creating and sharing documents that combine code, visualizations, and explanations.

- **Python Libraries and Packages:**

  You'll need various Python libraries and packages for data pre-processing, model development, and evaluation. Some essential libraries include:

  - NumPy: For numerical operations and array handling.
  - Pandas: For data manipulation and analysis.
  - Scikit-Learn: For machine learning algorithms and tools.
  - Matplotlib and Seaborn: For data visualization.
  - TensorFlow or PyTorch: For deep learning models.
  - Scipy: For scientific and statistical functions.
  - Streamlit: For creating web-based interfaces (if applicable).

- Termcolor: Filling Colors in Graphs

➢ **Database Management System (DBMS):**

Used Depending on our project, you might require a DBMS like PostgreSQL or MySQL for data storage, retrieval, and management. This is especially relevant if you're dealing with large-scale healthcare datasets.

➢ **Cloud Computing Services:**

Depending on the scale of your project and the need for computational resources, consider cloud services Google Cloud Platform (GCP) for scalable computing power and data storage.

Choice of operating system (Windows, macOS, or Linux) is flexible, as most machine learning tools and libraries are cross-platform. However, Linux distributions (e.g., Ubuntu) are popular among data scientists and offer advantages in terms of performance and package management.

➢ **Operating System:**

Choice of operating system (Windows, macOS, or Linux) is flexible, as most machine learning tools and libraries are cross-platform , we used Windows here. However, Linux distributions (e.g., Ubuntu) are popular among data scientists and offer advantages in terms of performance and package management.

8.3 Tools:

- Jupyter Notebook: An interactive and web-based tool for creating and sharing documents containing live code, visualizations, and narrative text. It's useful for data exploration and documenting your work.
- Git and GitHub: Git is a version control system, and GitHub is a platform for hosting and collaborating on Git repositories. These tools help manage code changes, collaboration, and project history.

- Docker: A containerization platform that simplifies the packaging and deployment of machine learning models and their dependencies, ensuring reproducibility across different environments.

- Streamlit: A Python library for creating web applications with minimal effort. It's valuable for building user-friendly interfaces to showcase your disease prediction models.

- Database Management Systems (DBMS): Tools like PostgreSQL, MySQL, or SQLite can be used for efficient data storage and retrieval, especially when handling large healthcare datasets.

- Continuous Integration/Continuous Deployment (CI/CD)Tools: Tools like Jenkins or Travis CI can help automate testing, building, and deployment processes, ensuring the project remains robust and up-to-date

8.4 Block diagram:

8.4.1 Diabetic Prediction

Attributes of Diabetics disease:-

a. Pregnancies: - Frequency of pregnancy.
b. Glucose: - Concentration of plasma glucose.
c. Blood pressure: - Diastolic blood pressure (mm Hg).
d. Skin thickness: - Triceps skinfold thickness (mm).
e. Insulin: - Two hours serum insulin (mu U/ml).
f. BMI: - Body mass index.
g. Diabetics pedigree Function: - Pedigree- function for diabetes.
h. Age: - Age (log (years)).
i. Outcome.

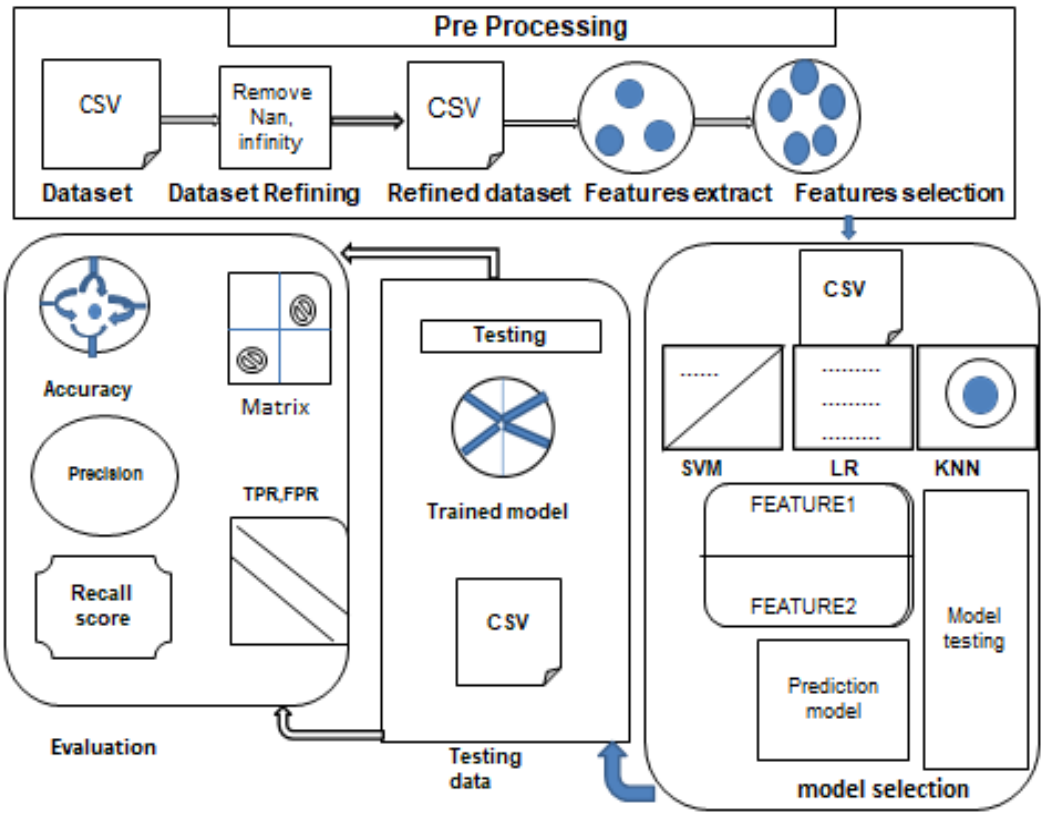

Fig. 8.4.1 Working of ML Model

8.4.2 Heart Disease prediction

Attributes Used for testing:-

a. Age / Resting electro cardiographic
b. Results(Restecg)
c. Sex (Exercise inducedangina)
d. Chest Pain (CP)
e. Blood Pressure (Trestbps)
f. Major vessels (Ca)
g. Fasting Blood Sugar (fbs)
h. Heart Rate (Thalach).

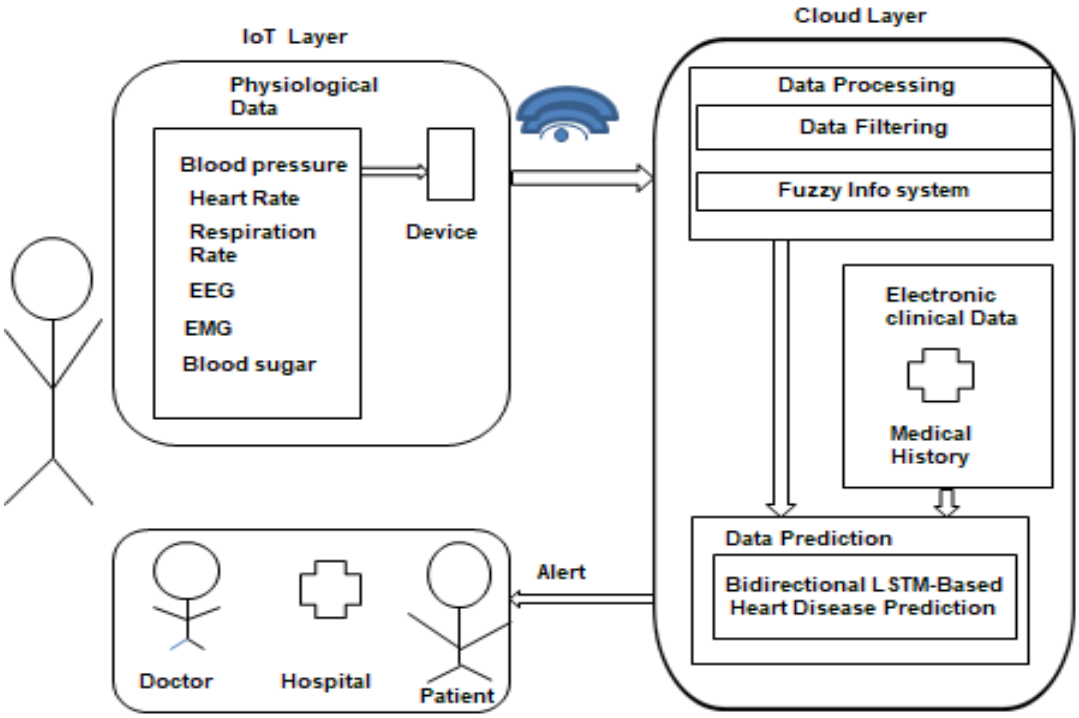

Fig: 8.4.2.1 Working of ML Model

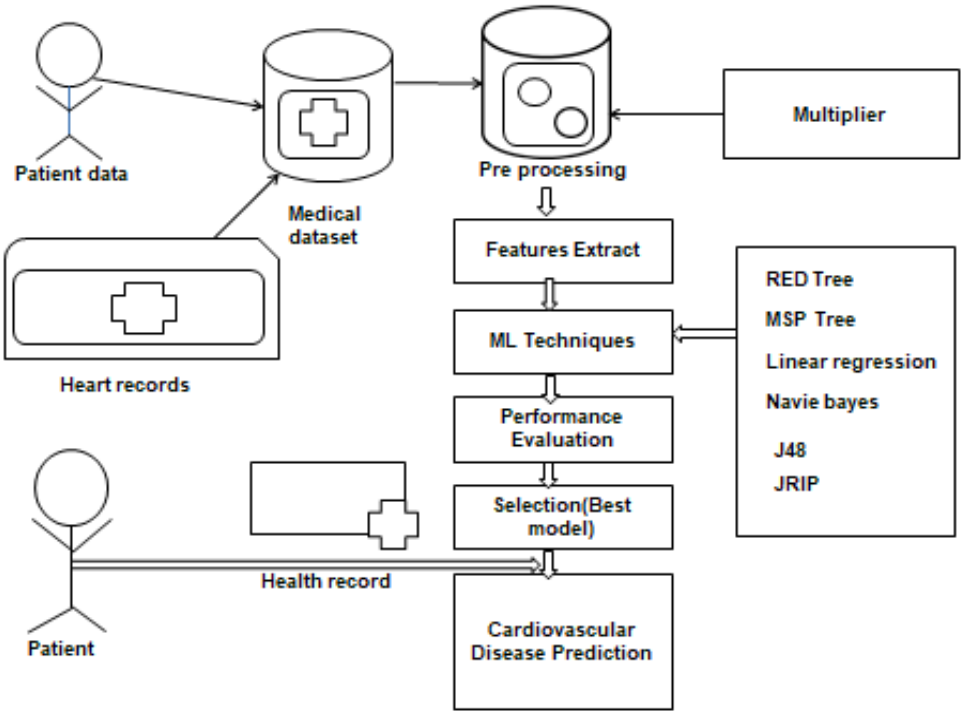

Fig.8.4.2.2 Data details of Heart disease.

8.4.3 Lung Cancer

Dataset Description:

Chest CT scan image in png format (size 256 x 256)

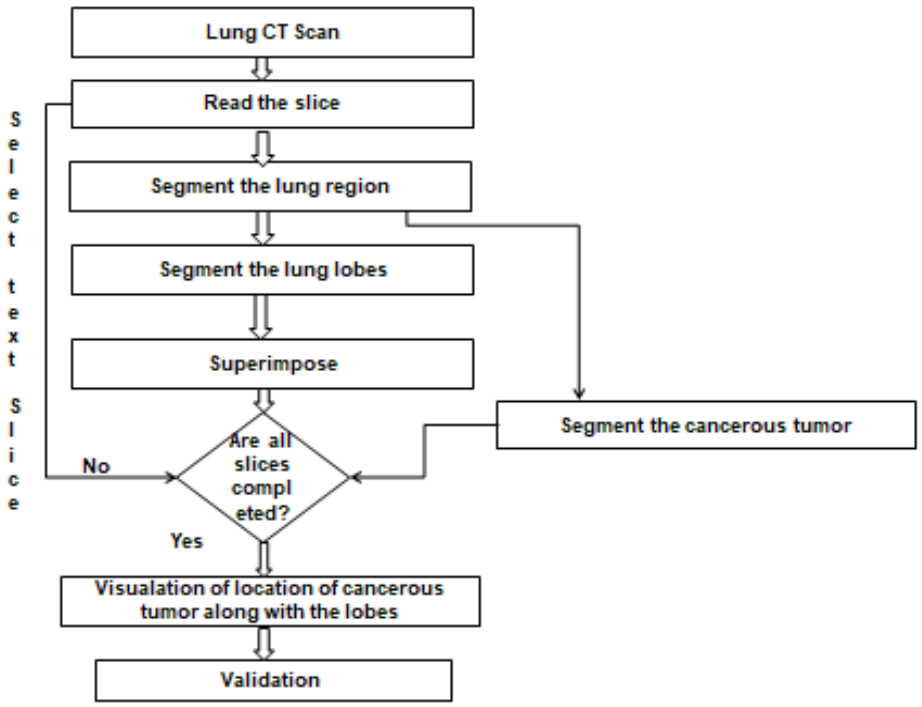

Fig. 8.4.3.1 Proposed lung cancer diagnosis system.

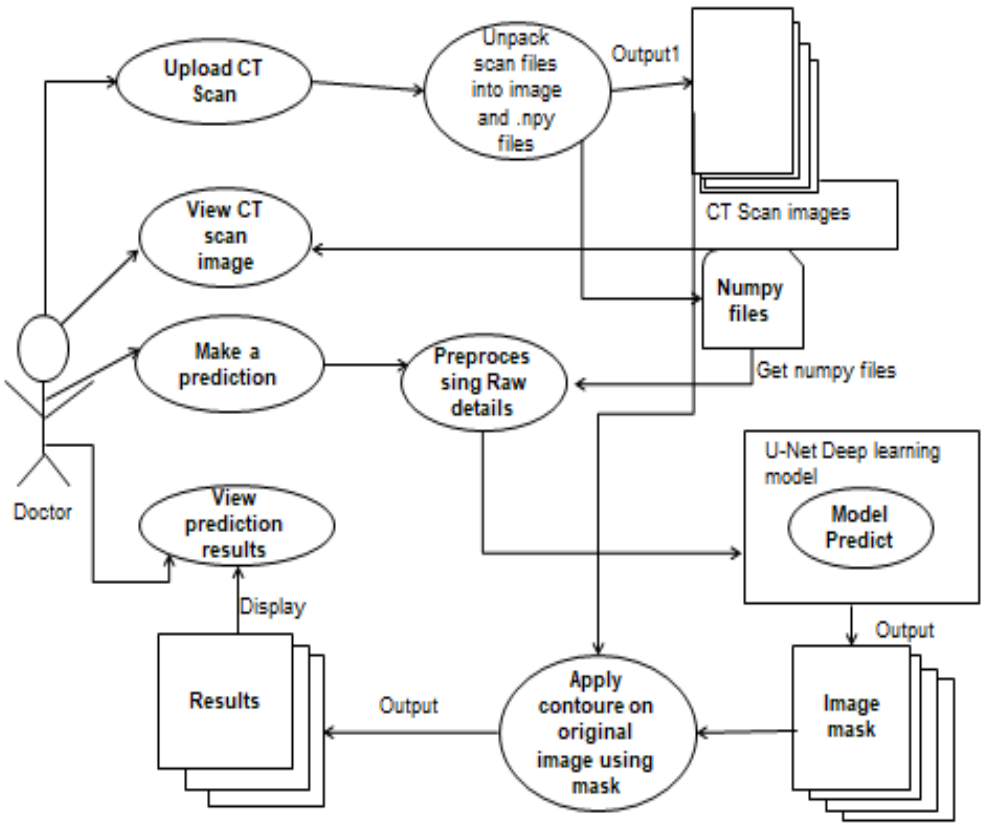

Fig. 8.4.3.2 Use case for testing phase

9. REFERENCES

1. Multi Disease Prediction System using Random Forest Algorithm in Healthcare System *[R. Shanthakumari; C. Nalini, S. Vinothkumar, E.M. Roopadevi, B. Govindaraj March 2022]*
2. Leveraging Machine Learning Methods for Multiple Disease Prediction using Python ML Libraries and Flask API *[Srikanth Narayanan, N. M. Balamurugan, Maithili K, P. Bini Palas May 2022]*
3. Human Disease Prediction based on Symptoms *[Uday K. Kommineni Dhatri Phani Priya Kowthavarapu, Guna Sri Manjunadh Polukonda, Kalyan Chakravarti Yelavarti April 2023]*
4. Symptom Based Health Prediction using Data Mining *[S Vijava Shetty, G A Karthik, M Ashwin July 2019]*
5. Disease Prediction Using Machine Learning Techniques *[Roop Chandrika Mallela, Reddy Lakshmi Bhavani, B. Ankayarkanni June 2021]*
6. Automated Disease Prediction Using Machine Learning Technology *[Neetu Mittal, Hemangi Sharma August 2023]*
7. Multiple Heart Diseases Prediction using Logistic Regression with Ensemble and Hyper Parameter tuning Techniques *[Sateesh Ambesange, A. Vijayalaxmi, S. Sridevi, Venkateswaran, B. S. Yashoda] (July 2020)*
8. Data Science And Its Application In Heart Disease Prediction *[Mohammed Jawwad Ali Junaid, Rajeev Kumar] (June 2020)*
9. Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design *[M. Raihan, Saikat Mondal, Arun More, Md. Omar Faruqe Sagor, Gopal Sikder, Mahbub Arab Majumder, Mohammad Abdullah Al Manjur, Kushal Ghosh] (December 2016)*
10. Analysis of Different Machine Learning Algorithms Used for Identification of Lung Cancer Disease *[Sameeka Saini, Ankit Maithani, Diksha Dhiman, Amita Bisht] (September 2021)*
11. Lung Cancer Disease Diagnosis Using Machine Learning Approach *[Swati Mukherjee, S. U. Bohra] (December 2020)*
12. Lung Cancer Detection and Classification using CT Scan Image Processing *[Nusraat Nawreen, Umma Hany, Tahmina Islam] (July 2021)*
13. Machine Learning-based Diabetes Prediction: A Cross-Country Perspective *[Sadia Afrin Shampa, Md. Saiful Islam, Ayatun Nesa] (June 2023)*

14. An Efficient Diabetes Prediction Model using Machine Learning *[Esther Daniel, Jobin Johnson, Ujjal Amitya Victor, G.V. Aditya, Steve Abraham Sibby] (July 2023)*

15. Towards a mobile solution for predicting illness in Type 1 Diabetes Mellitus: Development of a prediction model for detecting risk of illness in Type 1 Diabetes prior to symptom onset *[Jonas N Lauritzen, Eirik Årsand, Klaske Van Vuurden, Johan Gustav Bellika, Ole K Hejlesen, Gunnar Hartvig-sen] (July 2011)*

10. EXPECTED SCHEDULE

TIME PERIOD	WORK TO BE COMPLETED
July 2023 – August 2023	Search and detailed study of topic.
August 2023 – September 2023	Implementation of Module 1 with unit testing.
September 2023 – October 2023	Implementation of Module 2 with unit testing.
December 2023 – January 2023	Implementation of Module 3 with unit testing.
January 2023 – Ferbruay 2023	Integration of all modules and deployment.
February 2023 – May 2023	Training and adding additional features.

Students:

Mr. Aditya Jairam Nimbalkar

Ms. Vasudha Anil Dake

Ms. Vaishnavi Premnath Kamble

Mr. Tejas Suresh Shingan

Guide:

Mrs. Amrita A. Manjarekar

Assistant Professor,
Computer Science and Techonolgy,
Department of Technology, Shivaji University, Kolhapur.

Class Co-Ordinator:

Mrs. Rupali Dhabarde

Assistant professor,

Department of Technology,

Shivaji University Kolhapur.

Co- Ordinator:

Dr. R. J. Deshmukh

Computer Science and Technology,

Department of Technology,

Shivaji University Kolhapur.