Disease Prediction System using Machine Learning

A Project Report

Submitted by

Vaibhav Raheja (C056)
Viraj Shah (C064)
Mayank Shetty (C075)
Purav Patel (C098)

Under the Guidance of

Prof. Manisha Tiwari

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY (Integrated)
COMPUTER ENGINEERING
At

MUKESH PATEL SCHOOL OF TECHNOLOGY MANAGEMENT AND ENGINEERING

September 2021

Table of Contents	Page No.
Chapter-1: Introduction **i. Machine Learning in Disease Prediction/Diagnosis** **ii. An Overview of the Expert System**	**1**
Chapter-2: Literature Review **i. Review of Literature** **ii. Proposed System** **iii. Proposed Algorithms**	**2-12**
Chapter-3 **i. Architecture Diagram** **ii. Plan of Action** **iii. References**	**13-15**

Chapter – 1

Introduction

Machine Learning in Disease Prediction/Diagnosis

Machine Learning is a branch of Artificial Intelligence (AI), where the main objective is to give the computer the ability to learn from a provided set of data. The structure of the data is understood, after which the data is fit into models. These models can be successfully utilized by people for any given application where machine learning is required.

Despite being a field within computer science, it radically differs from the traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers for calculations or problem-solving, whereas machine learning algorithms allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range.

This approach in machine learning facilitates computers in building models from sample data, in order to automate decision-making processes based on data inputs.

Our project will be harnessing the potential of machine learning, in which a model will be trained in identifying various diseases included in our scope, where the output will be Boolean values.

An Overview of Expert Systems in Medicine

Expert systems in medicine are defined as systems with the ability to capture and store expert knowledge, facts, and reasoning techniques to assist doctors in diagnosing a patient's condition.

These systems attempt to mimic a doctor's expertise by applying several computational methods to help in decision support and problem solving, by coming up with reasoned conclusions for a patient's illness or condition.

Our project will incorporate the core elements of an expert system by supporting medical experts with their claims, precursing their diagnosis of a chronic disease in their patients using trained machine learning models.

Chapter – 2

2.1. Literature Review

2.1.1 Papers on machine learning for Chronic Kidney Disease Detection

Sr No.	Paper Name	Authors	Field of Research	Algorithm	Dataset	Paper Summary	Reference
1	Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease	- Rahul Gupta - Nidhi Koli - Niharika Mahor - N. Tejashri	Chronic kidney disease, Prediction, Machine Learning, Decision Tree, Random Forest and Logistic Regression.	- Decision Tree - Random Forest - Logistic Regression	UCI Chronic Kidney Disease Dataset	- The paper evaluates the performance of Decision Tree, Random Forest, and Logistic Regression on the preprocessed and filtered Chronic Kidney Disease Dataset. - Accuracy = Decision Tree (98.48%), Random Forest (94.16%), Logistic Regression (99.24%). - The authors conclude the paper by stating that Logistic Regression provides highest accuracy and recall, while Decision Tree provides highest precision.	International Conference for Emerging Technology (INCET)
2	Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease	- F. M. Javed Mehedi Shamrat - Pronab Ghosh - Mahbubul Hasan Sadek - Md. Aslam Kazi - Shahana Shultana	Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Accuracy.	- Decision Tree - Random Forest - Logistic Regression - K-Nearest Neighbors (KNN)	UCI Chronic Kidney Disease Dataset	- The paper compares the accurate prediction rates of Chronic Kidney Disease using Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN) over the presented dataset. - Accuracy = Decision Tree (97.91%), Random Forest (100%), Logistic Regression (100%), KNN (95.82%). - Random Forest takes the most time to predict, and has the best rating in Receiver Operating Characteristic (ROC) curve.	International Conference for Innovation in Technology (INOCON)
3	A Machine Learning Methodology for Diagnosing Chronic Kidney Disease	- Jiongming Qin - Lin Chen - Yuhua Liu - Chuanjun Liu - Changhao Feng - Bin Chen	Chronic kidney disease, machine learning, KNN imputation, integrated model.	- Random Forest - Logistic Regression - K-Nearest Neighbors (KNN) Imputation - Integrated Model (Random Forest + Logistic Regression)	UCI Chronic Kidney Disease Dataset	- The paper proposes an integrated model, which combines Logistic Regression and Random Forest, to accurately diagnose Chronic Kidney Disease. - Accuracy = Logistic Regression (98.95%), Random Forest (99.75%), Integrated Model (99.83%). - The authors conclude by stating that the integrated model has the highest accuracy and can be further perfected by the increase of size and quality of the data, as it	Institute of Electrical and Electronics Engineers (IEEE)

						currently cannot diagnose the severity of CKD because only two categories of data samples exist in the dataset (ckd and no ckd).	

2.1.2 Papers on machine learning for Diabetes Detection

Sr No.	Paper Name	Author	Field of Research	Algorithm	Dataset	Paper Summary	Reference
1	Diabetes Disease Prediction Using Data Mining	-Deeraj Shetty, -Kishor Rit, -Sohail Shaikh, -Nikita Patil	Diabetes, Prediction, Naïve Bayes, KNN.	- Naive Bayes Algorithm - K-Nearest Neighbors	Pima Indians Diabetes Database	-The Paper recommends using a larger dataset for better prediction -The larger dataset has proved to increase the accuracy of the algorithm, therefore working on some more attributes to better diagnose diabetes.	International Conference on Innovations in Information, Embedded and Communication Systems
2	Diabetes Prediction using Machine Learning Algorithms	-Aishwarya Mujumdara -Dr. Vaidehi Vb	Diabetes Mellitus, Big Data Analytics, Healthcare Machine Learning,	- Logistic Regression - LDA - Random Forest - Extra Trees Classifier	Self created Dataset based on Pima Indians Diabetes Database	-In this study, various machine learning algorithms are applied on the dataset and the classification has been done using various algorithms of which Logistic Regression gives highest accuracy of 96%. - Application of pipeline gave Ada Boost classifier as best model with accuracy of 98.8%.	International Conference On Recent Trends In Advanced Computing
3	Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm	-Samrat Kumar Dey -Ashraf Hossain -Md. Mahbubur Rahman	Diabetes, SVM, ANN, Naïve Bayes, Min Max Scaling	- ANN - Support Vector Machine - Naive Bayes Algorithm - KNN	Pima Indians Diabetes Database	- From different machine learning algorithms Artificial Neural Network (ANN) provide us highest accuracy with Min Max Scaling Method on Indian Pima Dataset.	International Conference of Computer and Information Technology
4	A comparison of machine learning algorithms for diabetes prediction	-Jobeda Jamal Khanam, -Simon Y. Foo	Machine learning, Data Mining, K-fold Cross Validation,	- Decision Tree - K-Nearest Neighbors - Random Forest - Naive Bayes - A/B - Linear Regression - Support Vector	Pima Indians Diabetes Database	- The paper evaluates the performance of various algorithms and measures to get a high accuracy model for Diabetes prediction. All models give an average accuracy of 70%. -LR and SVM have 77-78% accuracy in both K-fold and train-test split -Among all the proposed	The Korean Institute of Communications and Information Sciences (KICS).

				Machine		models, the NN with two hidden layers is considered the most efficient and promising for analyzing diabetes with an accuracy rate of approximately 86%	

2.1.3 Papers on machine learning for Heart Disease Detection

Sr No.	Paper Name	Author	Field of Research	Algorithm	Dataset	Paper Summary	Reference
1	Implementation of Machine Learning Model to Predict Heart Failure Disease	- Fahd Saleh Alotaibi	Machine learning model, heart failure diagnosis, KNN method.	- Decision Tree - Naïve Bayes - Random Forest - Support Vector Machine - Logistic Regression	UCI Heart Disease Dataset	- The paper aims to improve the heart failure (HF) prediction using the UCI heart disease dataset which is available on the internet. Multiple machine learning approaches were used to predict HF. - Accuracy = Decision Tree (93.19%), Random Forest (89.14%), Logistic Regression (87.36%), SVM (92.30%) & Naïve Bayes (87.27%). - In comparison this study showed significant improvement and higher accuracy than previous work.	International Journal of Advanced Computer Science and Applications (IJACSA)
2	Heart Disease Prediction using Hybrid machine Learning Model	- Dr. M. Kavitha - G. Gnaneswar1 - R. Dinesh - Y. Rohith Sai - R. Sai Suraj	Cleveland Heart Disease Database, Hybrid algorithm, Machine learning	- Decision Tree - Random Forest - Hybrid (Decision Tree + Random Forest)	Cleveland Heart Disease Dataset	- The interface is designed to get the user's input parameter to predict heart disease, for which the authors used a hybrid model of Decision Tree and Random Forest. - Accuracy = Decision Tree (79%), Random Forest (81%) & Hybrid Model (88.7%).	Institute of Electrical and Electronics Engineers (IEEE)

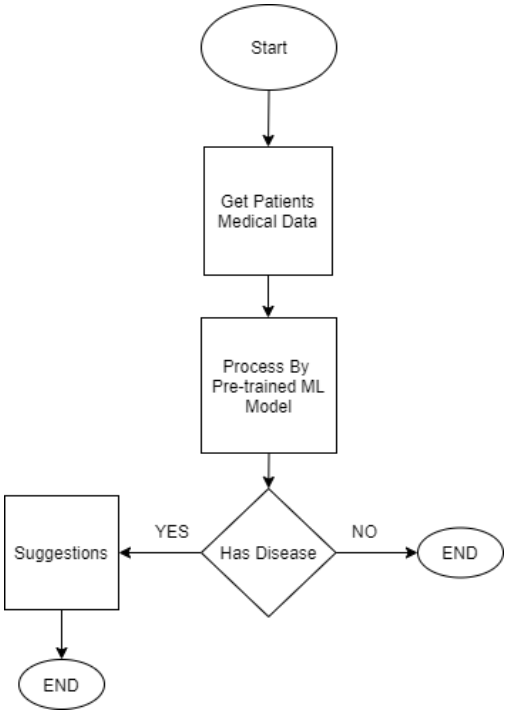
3	Heart Disease Diagnosis using Extreme Learning Based Neural Networks	- Muhammad Fathurachman - Umi Kalsum - Noviyanti Safitri - Chandra Prasetyo Utomo	Heart Disease, Extreme Learning Machine, Medical Diagnosis.	- Artificial Neural Networks - Support Vector Machine - Decision Tree - Extreme Learning Machine	- Cleveland Heart Disease Dataset - Hungarian Institute of Cardiology Dataset - University Hospital Zurich Dataset - Medical Center Long Beach Dataset (All available on UCI Machine Learning Repository)	- The authors divided the experiments into five parts. For each part, we used a different training and testing dataset which had been divided using K-Fold Cross Validation. Then we set the hidden node of our ELM model for predicting heart disease - Accuracy = Decision Tree (75%), ELM (80%), SVM (68%), BP ANN (77%) - From the experimental results and analysis, the authors concluded that the performance of the ELM algorithm tends to be better when compared with SVM, DT and BP ANN. With an average of 83% accuracy, 88% sensitivity and 82% specificity, the application of ELM algorithm can be an alternative solution to help clinicians predict heart disease.	International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)

2.1.4 Papers on machine learning for Pneumonia:

Sr No.	Paper Name	Author	Field of Research	Algorithm	Dataset	Page Summary	Reference
1	Pneumonia Detection Using CNN based Feature Extraction	-Dimpy Varshni -Rahul Nijhawan -Kartik Thakral -Ankush Mittal -Lucky Agarwal	DensetNet, Deep Convolutional Neural Networks, SVM, Transfer Learning, Random Forest, Naive Bayes, K-nearest neighbors, Feature extraction.	-CNN -SVM	-The dataset used is ChestX-ray14 released by Wang et al. (2017) also publicly available on the Kaggle	- The authors have used a customized model which is a combination of CNN based feature extraction and supervised classifier algorithm. - The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors. The C parameter trades off correct classification of training examples against maximization of the decision function's margin. The Area Under the ROC curve (AUC) is an aggregated metric that evaluates how well a logistic regression model classifies positive and negative outcomes at all possible cutoffs. It can range from 0.5 to 1, and the larger it is the better. -	Institute of Electrical and Electronics Engineers (IEEE)

						Results = SVM (rbf kernel), C = 3.5, gamma = 2e-05 & AUC = 0.7904	
2	Pneumonia Detection using CNN with Implementation in Python	-Muhammad Ardi	Computer science undergraduate student of Universitas Gadjah Mada	-CNN	-Chest X-Ray Images (Pneumonia)	- This paper successfully detected pneumonia using CNN - The model is able to predict pneumonia caused by bacteria pretty well since 232 out of 242 samples are classified correctly -Which gives a accuracy of 95.86%	Pneumonia Detection using CNN with Implementation in Python
3	Detection of Pneumonia using ML & DL in Python	-A Sharma, -M Negi, -A Goyal, -R Jain, -P Nagrath	Neural network, confusion matrix, keras, recall, hyper parameters	-CNN	-Chest X-Ray Images (Pneumonia)	-Detection of diseases with the assistance of computers from various Machine and Deep learning techniques are very beneficial in such places where there is shortage of people who are skilled in techniques like radiology -Validation Accuracy 0.8410 -Validation Loss 0.8395 -Accuracy 0.9806 -Loss 0.0742	IOP Conference Series: Materials Science and Engineering

2.2. Proposed System



2.2.1. Get Patients Medical Data

This state deals with collection of the input parameters for the pre-trained model. Depending on the chronic disease selected a specific set of biological parameters of the patients are collected and fed to the pre-trained machine learning model.

2.2.2. Process By Pre-Trained Machine Learning Model

This stage contains a pre-trained model for each chronic disease considered in our project. Appropriate features will be selected for each chronic disease, which will help in further classification of a patient having or not having the chronic disease.

2.2.3. Suggestions

This stage is the output of the pre-trained model for the disease. In the event of a positive result, appropriate suggestions will be provided to the doctors, as there is a high probability of the patient having a chronic disease. If the result is negative, the patient has a low probability of the patient having a chronic disease.

2.3. Proposed Algorithms

After the literature survey we have proposed the following algorithmic model to detect the given chronic disease:

2.3.1. Algorithm for machine learning for Chronic Kidney Disease Detection

- **Logistic Regression:** Logistic Regression is a statistical machine learning technique used for binary classification of categorical dependent variables, by using a set containing independent variables. Assuming the probability of a particular class, Logistic Regression creates a regression model that distinguishes between several samples using the sigmoid function, where one or more variables are used to determine the expected outcome in probabilistic values which lie between 0 and 1. Logistic Regression has previously been implemented to detect chronic kidney disease in patients, and from referenced papers it was found to have a detection accuracy of 98.95% [7], 99.24% [8], and 100% [9] respectively on the UCI dataset, making it a high-accuracy machine learning model for implementation of the expert system.

- **Random Forest:** Random Forest is a supervised learning classification technique and is often considered as an ensemble machine learning method, as it is used for classification, regression, and probability. Random Forest can find missing values from many datasets, and it can provide a more accurate value by creating a forest of decision trees during the learning phase, where the number of trees indicate the robustness of the forest as well as the accuracy of the algorithm. For training characteristics and then random sub characteristics for sampling nodes, random sampling is done. The algorithm aggregates multiple decisions to make a single decision by consolidating multiple forests on different subsets of a dataset and averaging the results to enhance the performance of the dataset's detection accuracy. It can be combined with multiple classifiers to solve a complex problem and improve the overall accuracy of the machine.

  Step 1: Randomly select 'p' features from the total 'q' features, where $p << q$.

  Step 2: Calculate node 'd' using the method best split point from the selected 'p' features.

  Step 3: Split the node into daughter nodes using the best split.

  Step 4: Repeat the above steps till a number '1' is reached.

  Step 5: A forest of 'n' number of trees is built by applying the above steps 'n' number of times.

Random Forest has been chosen as it comparatively takes less time to train and has also been used to detect chronic kidney disease in patients from our referenced papers, where it was found to have a detection accuracy of 94.16% [8], 99.75% [7], and 100% [9] respectively on the UCI dataset, making it a suitable high-accuracy machine learning model for the implementation of the expert system. It can be combined with a Logistic Regression classifier to form an Integrated Model, which offers an average detection accuracy of 99.83%, which is better than the average detection accuracies of both the algorithms combined [7].

2.3.2. Algorithm for machine learning for Diabetes Detection

- **K-Nearest Neighbour:** KNN algorithm assumes the similarity between the new input and then puts the new case in the most fitting category. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. After referencing the papers on 'A comparison of machine learning algorithms for diabetes prediction' [2], KNN receives the highest accuracy of 79% in detection of diabetes on PIMA dataset. Hence it can be used for a highly accurate ML model to detect diabetes in patients as well as allow an opportunity for improving the accuracy of the model.

- **Logistic Regression:** Logistic regression, despite its name, is a classification model rather than a regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. After referencing the papers on 'A comparison of machine learning algorithms for diabetes prediction' [2], LR receives an accuracy of 78% in detection of diabetes on PIMA dataset. Hence it can be used for a highly accurate ML model to detect diabetes in patients. Preprocessing the data using K-means and clustering can also increase the accuracy to 96% as referenced from 'Diabetes Prediction using Machine Learning Algorithms' [3] making it a suitable model for implementation of the expert system.

2.3.3. Algorithm for machine learning for Heart Disease Detection

- **Decision Tree:** It's a tree-like classification model, which builds a structure consisting of branches and nodes on the bases of evidence collected for each attribute during the model learning phase. The decision tree's branches and nodes connect according to the number of entities described in the dataset. The forwarding process uses the number of values dedicated for each attribute. Furthermore, following the rules described on each branch and node it reached the decision for each transaction. Finally, according to the decision node the class label will be assigned to the record. This procedure is iterative and repeats till each transaction has a class category. Therefore, this algorithm converts the attributes into branches and nodes, and selects one of the attributes as a decision node.

  Step 1: Importing the libraries

  Step 2: Importing the dataset

  Step 3: Splitting the dataset into the Training set and Test set

  Step 4: Training the Decision Tree Regression model on the training set

  Step 5: Predicting the Results

  Step 6: Comparing the Real Values with Predicted Values

  Step 7: Visualising the Decision Tree Regression Results

  We are going to use decision tree for Heart disease detection, as the outputs of the referenced papers had good accuracy, while being easy to read and interpret without requiring any statistical knowledge. It takes less effort to prepare the data, and once the variables have been created, the data cleaning process is minimal.

- **Random Forest Regression:** Random Forest Regression aggregates multiple decisions to make a single decision. For training characteristics and then random sub characteristics for sampling nodes, random sampling is done. It combines multiple classifiers to solve a complex problem and improve the machine's accuracy. During the learning phase, this model first generates multiple random trees called a forest. It is a classifier that consolidates multiple forests on different subsets

of a dataset and averages the results to enhance the performance of the dataset's detection accuracy.

Step 1: Importing the libraries

Step 2: Importing the dataset

Step 3: Splitting the dataset into the Training set and Test set

Step 4: Training the Random Forest Regression model on the training set

Step 5: Predicting the Results

Step 6: Comparing the Real Values with Predicted Values

Step 7: Visualising the Random Forest Results

Random Forest Regression was chosen as from the referenced papers it comparatively takes less time to train and along with that it provides high accuracy which increases the efficiency of the model.

2.3.4. Algorithm for machine learning for Pneumonia

- **Convolutional neural network:** The system learns to perform feature extraction in a convolutional neural network. The core concept of CNN is to use convolution of image and filters to generate invariant features which are passed onto the next layer. The characteristics of the following layer are combined with different filters to produce more invariant and abstract features. This is done till the final feature / output (say face of X) is achieved. Convolutional neural network consists of several building elements such as convolution layers, pooling layers and fully-connected layers, and is meant to learn the spatial hierarchies of features by means of a background algorithm.

- **Support Vector Machine (SVM):** The objective of the SVM algorithm is to find the optimal line or decision boundary for categorising n-dimensional space to categorize new data points in the correct category. The best decision boundary is referred to as a hyperplane.

Step 1: Import Python libraries

Step 2: Display image of each bee type

Step 3: Image manipulation with rgb2gray

Step 4: Histogram of oriented gradients

Step 5: Create image features and flatten into a single row

Step 6: Loop over images to preprocess

Step 7: Scale feature matrix + PCA

Step 8: Split into train and test sets

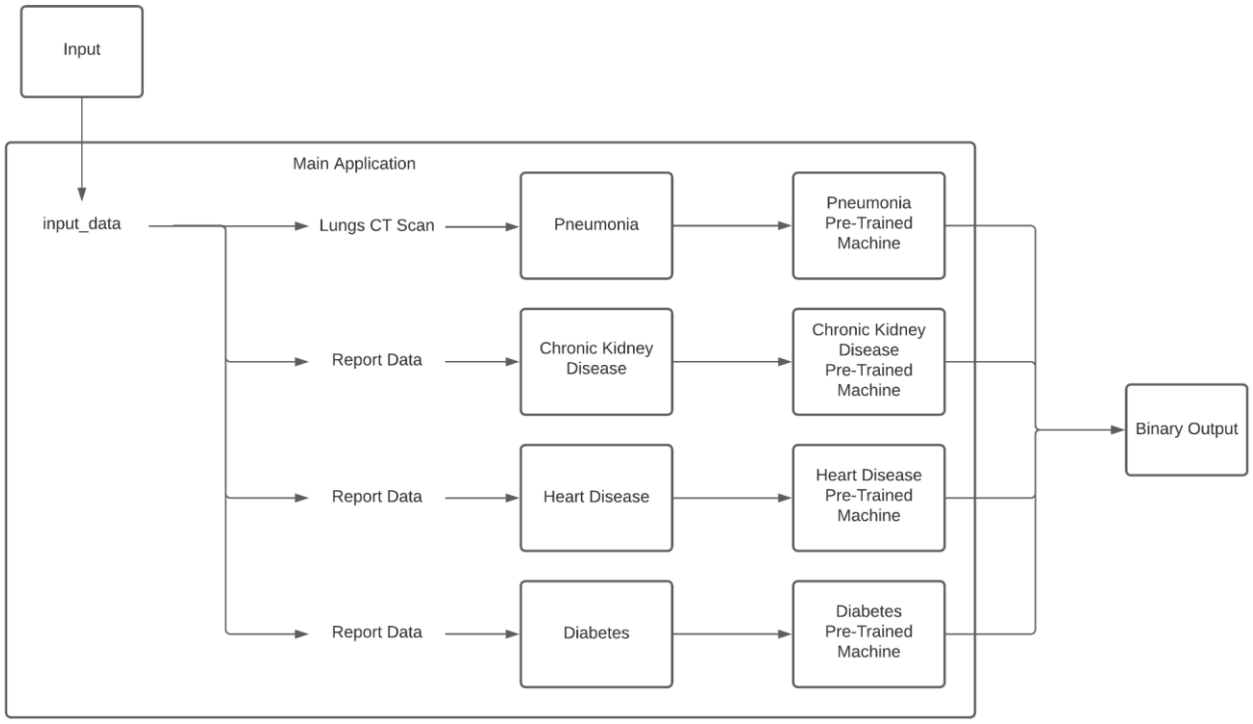
Step 9: Train model

Step 10: Score model

Step 11: ROC curve + AUC

Chapter – 3

3.1 Architecture Diagram

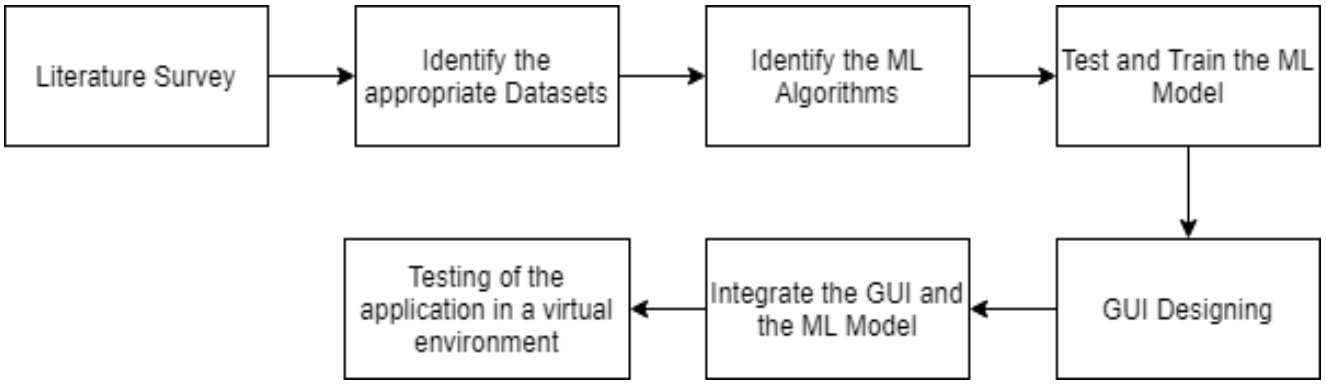


Disease Detection System

The user will first decide which disease machine they want to use and upon selecting the disease, the data is input into Disease Detection Systems' Pre-Trained Machine model. The machine will then give a binary output, whether the patient has the selected disease, or they are healthy.

The Pre-Trained Machines will be trained on a pre-processed dataset. The dataset will go through multiple processing steps before it will be fed to the machine learning models to train them. The data will be checked for any noisy and missing data. K Nearest Neighbor (KNN) will be used for imputing missing values, outlier detection methods will be used to estimate any noise in the data and rapid miner will be used to remove noise in the dataset. The dataset will also be checked for discrepancies, data transformation, discretization and binning techniques will be used.

3.2 Plan of Action



3.3 References

[1] J. J. Khanam, S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction”, *ICT Express (Feb. 2021)*, 2021, doi: 10.1016/J.ICTE.2021.02.004.

[2] F. G. Woldemichael and S. Menaria, "Prediction of Diabetes Using Data Mining Techniques," *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2018, pp. 414-418, doi: 10.1109/ICOEI.2018.8553959.

[3] A. Mujumdar, V. Vaidehi, “Diabetes Prediction using Machine Learning Algorithms”, *Procedia Computer Science, Volume 165*, 2019, pp. 292-299, doi: 10.1016/J.PROCS.2020.01.047.

[4] S. K. Dey, A. Hossain and M. M. Rahman, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm," *2018 21st International Conference of Computer and Information Technology (ICCIT)*, 2018, pp. 1-5, doi: 10.1109/ICCITECHN.2018.8631968.

[5] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid Machine Learning Model," *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[6] F. Alotaibi, “Implementation of Machine Learning Model to Predict Heart Failure Disease”, *International Journal of Advanced Computer Science and Applications (IJACSA), 10(6)*, 2019, doi: 10.14569/IJACSA.2019.0100637.

[7] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng and B. Chen, "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease," in *IEEE Access*, vol. 8, pp. 20991-21002, 2020, doi: 10.1109/ACCESS.2019.2963053.

[8] R. Gupta, N. Koli, N. Mahor and N. Tejashri, "Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease," *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154147.

[9] F. M. Javed Mehedi Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi and S. Shultana, "Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease," *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 2020, pp. 1-7, doi: 10.1109/INOCON50539.2020.9298026.

[10] T. Ozturk, M. Talo, E. Yildirim, U. Baloglu, O. Yildirim, U. Acharya, “Automated detection of COVID-19 cases using deep neural networks with X-ray images.”, *Computers in biology and medicine vol. 121 (2020)*, 2020, doi: 10.1016/j.compbiomed.2020.103792.

[11] A. Sharma, M. Negi, A. Goyal, R. Jain and P. Nagrath, “Detection of Pneumonia using ML & DL in Python.”, *IOP Conference Series: Materials Science and Engineering* 1022, 2021, doi: 10.1088/1757-899X/1022/1/012066.