

Machine Learning Based Web Application to Detect Heart and Lung Disease

Mohamadi Ghousiya Kousar
Dept. Computer Science and
Engineering
Vijaya Vittala Institute of Technology
Kawalkhan2023@gmail.com

G S Roja
Dept. Computer Science and
Engineering
Vijaya Vittala Institute of Technology
gsroja124@gmail.com

Sumantha
Dept. Computer Science and
Engineering
Vijaya Vittala Institute of Technology
sumanth08sk@gmail.com

Maala B U
Dept. Computer Science and
Engineering
Vijaya Vittala Institute of Technology
maalabu1794@gmail.com

Shruthi R
Dept. Computer Science and
Engineering
Vijaya Vittala Institute of Technology
sr9208818@gmail.com

Abstract: This study presents Smart health diagnostics, an online tool that uses cutting -edge machine learning to identify heart and lung conditions early on, The platform allows physicians to upload medical photos for quick and accurate analysis, integrating seamlessly with workflows in the healthcare industry, Driven by cutting-edge convolutional and recurrent neural networks, the system is highly proficient at identifying nuanced patterns suggestive of diverse pulmonary and cardiovascular ailments. Real-time risk assessment, dynamic display of illness development, and explainable AI are important elements that give clinicians insights into the results of diagnostic tests.

Keywords: cardiovascular disease, clinical examination, x-rays, CT scans or ecgs, machine learning model, detection of heart disease percentage, cholesterol levels.

1.INTRODUCTION: Heart disease, also known as cardiovascular disease, includes a variety of heart problems and is the leading cause of death in the United States. However, there are numerous prevention and treatment strategies for different types of heart disease. Heart disease comes in many forms, including congenital heart defects that are present at birth and those that develop throughout life. The most common form is coronary artery disease, in which plaque gradually builds up in the heart's arteries, leading to complications such as angina, heart attack, heart

failure, and arrhythmias. Other forms can affect the heart valves or heart muscle (cardiomyopathy). The causes of heart disease vary and may include lifestyle factors, genetics, infections, medications, and other illnesses. Factors such as age, gender, family history, race/ethnicity, and lifestyle contribute to the likelihood of heart disease. Conditions such as high blood pressure, high cholesterol, diabetes, obesity, and certain autoimmune diseases also increase risk. Symptoms vary depending on the specific type of heart disease, but some people may not experience any symptoms at first. Complications such as a heart attack can be the first sign of underlying heart disease. Diagnosis of heart disease: The treatment plan depends on the type and severity of your heart disease and whether a person has other health conditions. Options include heart-healthy lifestyle changes, medications, surgery, and participation in cardiac rehabilitation. Reducing your risk of heart disease includes making heart-healthy lifestyle changes, managing co- occurring medical conditions, and actively promoting your overall health. Lung disease encompasses any disruption in the normal functioning of the lungs.

2.LITERATURE REVIEW: COPD is a global public health challenge due to its high prevalence and long-term effects on related disabilities and mortality. Only 12% of individuals with chronic airflow limitations had a previous spirometry-defined COPD diagnosis during the recent screening of 57779 participants in China

[2,3]. The coronavirus disease 2019 pandemic has severely restricted spirometry tests, which are the primary tool used for COPD case-finding and severity evaluation. CT-based imaging could lead to improvements in COPD detection and evaluation. Recent advances in deep learning (DL) based artificial intelligence (AI) have enabled the direct interpretation of medical images without relying on specific radiographic features of interest. However, objective CT analysis requires prior knowledge of the anatomical and physiological implications of diseases likely to be associated with certain clinical outcomes. This study recruited 1393 participants from outpatient, inpatient and physical examination center settings of 4 large hospitals in China to mimic the 'real world' scenario. [2] COPD is the third most common global cause of death and is a main cause of chronic morbidity. Early detection of exacerbations, achieved with the support of telehealth systems and interventions aimed to promote self-management, could decrease the negative effects and costs associated with COPD patients. A COPD self-management intervention has been defined by an International Expert Group consensus as a 'structured but personalized and often Multi-component, with goals of motivating, engaging and supporting the patients to positively adapt their health behavior and develop skills to better manage their diseases.' Home telehealth systems are being designed and developed to help patients with chronic disorders, carers and health and social professionals. This study involved 16 COPD patients in the Pulmonology and Allergy Department of Puerta del Mar University Hospital, Cadiz, Spain.[4] The patients had been diagnosed with COPD by Spirometry and had a history of two or more exacerbations requiring treatment with antibiotics or oral corticosteroids. They used a sensor device and a base station to record their respiratory sounds daily for 6 months. A DTF classifier was designed using features extracted from the respiratory sounds daily recorded at home with a respiratory sensor. The detection accuracy was 78.0% and 75.8% of the detected episodes were reported exacerbations, whereas 87.5% were unreported events[4] Heart disease is a major death factor due to insufficient knowledge and experience of medical practitioners.

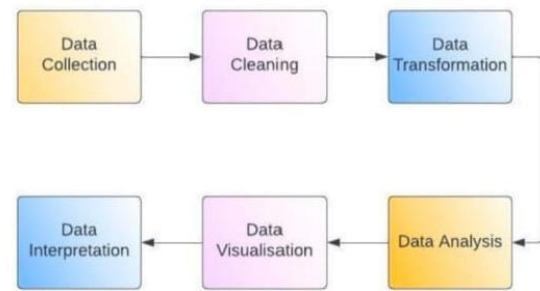
Data mining techniques such as Machine Learning (ML) and Data Mining (DM) are effective and significant in the medical industry. This research aims to examine various risk

Parameters highlighted in the investigation of Heart disease and to discover multiple techniques for the identification and prediction of heart disease. CNN, LSTM is proposed for heart disease identification and prediction which yields more improvised output than the rest of the prevailing techniques. Comparative performance in datasets concerning medicine is adopted in predicting heart disease techniques in comparison to other ML approaches [5,6,7] Heart disease is caused by blockage in coronary arteries, which leads to reduced nutrition to the heart muscle. The current research work incorporates predictive mining to build a predictive and diagnosis system concerning heart diseases. A Naive Bayesian (NB) algorithm is proposed for heart disease identification and prediction, and an ED&P approach is adopted for building a Data Mining (DM) approach using Machine Learning (ML) techniques to achieve prior detection and prevention of malignancy of heart disease [6]. The present research work has the following stages: assembles online datasets from the UCI medical dataset, then is the process of classification which involves training and testing, and collection of data [6] Data scientists have developed several models to help diagnose coronary heart disease, such as neural networks, Naive Bayes classifiers, and associative classification. To develop a heart disease classifier, a data mining algorithm was built and a Support Vector Machine (SVM), Artificial Neural Network (ANN), and a Decision Tree (DT) were used. K-folds validation and confusion matrices were used to evaluate the consistency, sensitivity, and specificity of the data. Ensemble techniques increased accuracy by an average of 7.26%. This research examined the effectiveness of machine learning in predicting coronary heart disease. Data analytics revealed patterns in the data and important features for binary logistic classification, but the models had a capped accuracy of 75%. The ensemble techniques of bagging, boosting, and stacking were effective in raising the accuracy of the

base models. The Gradient Boosting approach was the most powerful, providing an accuracy of 73.89% [5]. The K-Folds Cross Validation showed consistency in the accuracy produced with models having a low standard deviation ranging from 0.3%-0.6%. Stacking, involving heterozygous models, was the most effective ensemble method, producing an accuracy of 75.1%. However, the precision was high for each model, with an average of 76.1%.

3.METHODOLOGY

Collection of data: As we know, data is an essential thing in machine learning algorithms. Without data, there can be no precedence in the process. Data can be taken from a pre- made dataset or collected through the survey. The data used has to be of good quality for a better and more efficient output, a good data can be recognized from the quality and authenticity of its source. 2) Preparation of the data: Data preparation means putting together different data together for a more precise idea of what is required. Data has to be cleaned, removing unwanted data which can interrupt the final output or is not necessary for the algorithm, visualization data is also essential to understand the relationship between various variables and data Finally, the data is prepared by dividing it into the training set and testing set. Selection of model Many machine learning models are in the market, but the different models have different outcomes. Various machine learning models are used for speech/ voice recognition, image recognition prediction and many more. We need to choose the best one to get the required output. The cleaned data is sent to the model for processing according to the pattern of the algorithm of that particular model. The predictions are made, the more training is given to the model, the prediction of the model gets better and better, and the outcome also becomes more efficient.



Evaluation of model: Once the model is trained, it's time to see whether it performs. We provide the model with test data which serves the patterns and gives out the result based on which it was trained with training data. Once training testing is done, we need an accurate output to change or tune the parameters better. Making predictions When the process is done, we can use any data to determine the output from that particular model. The forecast may vary with different sets of data, as we have discussed in the early stages regarding the quality and quantity of data.

4.ALGORITHMS

Previously it has been used convolutional neural networks by using this the amount of outcome we are expecting it is not supporting so that we have used the logistic regression algorithm for the effective results.

Logistic regression is a supervised machine learning algorithm primarily employed for binary classification. It utilizes a logistic function, also known as a sigmoid function, to calculate a probability value between 0 and 1. This approach is especially useful when dealing with two classes, such as Class 0 and Class 1.

Before we have used the CNN,KNN, and the linear regression algorithms but we didn't get the expected outcome so we have used the logistic regression algorithm.

In logistic regression, the logistic function's output is compared to a threshold value, typically 0.5. If the value surpasses this threshold, the input is assigned to Class 1; otherwise, it belongs to Class 0. This

classification process distinguishes it from linear regression, where the output is continuous.



Logistic Regression Model

COMPARISON WITH LINEAR REGRESSION

Logistic regression is an extension of linear regression, but it serves a different purpose. While linear regression predicts continuous values, logistic regression predicts the probability that an instance belongs to a particular class. The discrete or categorical outcome falls between 0 and 1

As the linear regression and logistic regression will vary in giving the resulted outcomes as it is very important where it can take to another point. The project status will improve as the outcomes will give the expected results in a proper way so the logistic regression is best for our project so that we have chose it.

STEP BY STEP PROCEDURE OF COLLECTING DATASET

1.Collection of Data;

```
[ ] #Loading the csv data to a pandas DataFrame
heart_data=pd.read_csv("data.csv")

#Print first 5 rows of the dataset
heart_data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

We initiated our study by acquiring pertinent health data, focusing on respiratory information. The dataset, encapsulating diverse health parameters, was gathered for analysis and development of disease detection models.

MODEL TRAINING

We will train different types of models like linear regression, k nearest neighbour ,CNN, Random forest. After training the models we will select the best one which will support our project and we will train that model in the way where it should be able to collect the data which is required to detect the heart disease for the demo purpose and also for the real time examples and from taking the patient data we will detect the percentage.

```
[ ] model_0=LogisticRegression()
model_0.fit(X_train, Y_train)
model_1=KNeighborsRegressor()
model_1.fit(X_train, Y_train)
model_2=LinearRegression()
model_2.fit(X_train, Y_train)
model_3=Ridge()
model_3.fit(X_train, Y_train)
model_4 = Lasso()
model_4.fit(X_train , Y_train)
model_5 = SVC()
model_5.fit(X_train , Y_train)
model_6 = GaussianNB()
model_6.fit(X_train , Y_train)
model_7 = RandomForestClassifier()
model_7.fit(X_train , Y_train)
```

DATA QUALITY CHECK

```
# Check for missing values
print("\nMissing Values:")
print(lung_data_processed.isnull().sum())
```

```
Missing Values:
Age          0
Gender       0
Air Pollution 0
Alcohol use  0
Dust Allergy 0
OccuPational Hazards 0
Genetic Risk 0
chronic Lung Disease 0
Balanced Diet 0
Obesity      0
Smoking      0
Passive Smoker 0
Chest Pain   0
Coughing of Blood 0
Fatigue      0
Weight Loss  0
Shortness of Breath 0
Wheezing     0
Swallowing Difficulty 0
Clubbing of Finger Nails 0
Frequent Cold 0
Dry Cough    0
Snoring      0
Level        0
dtype: int64
```

These are some data quality checks we usually do for taking the parameters as people will say these many are required yes we will accept but only doing CT scans ECGs we can't able to detect the disease only one test cant able to prove that disease is present or not and also became very costly so our web application will help in all the problems by taking these many data quality checks and these parameters are very important to check the presence of heart disease in the percentage level

5.RESULT

disorder test these are the parameters that are taken for the detection of the disease.

This is the final output of the project where we enter the data of the patient and detection the presence of heart disease. These are the parameters required for the detection if the disease

Here's the explanation about all the parameters firstly the chest pain it consists of three types those are typical angina, Atypical angina, Nonanginal pain, Asymptomatic pain and next exercise induced angina if yes it will denote 1 if it is not there it will give 0 and then st depression segment it means float part of the ECG between s and t waves and then st depression segment means downward shift below the baseline and major blood vessels if 0 major vessel is blocked if 1 one vessel is blocked and if 2 two vessels are blocked and if 3 three vessels are blocked so the higher the number the more serious the heart problem is because more vessels are affected and then thalassemia blood

These are the other parameters to detect the lung diseases. As it contains is that lungs affected by the air pollution is its affected it will denote as 1 if not denotes as 0 and next if the person is alcoholic or not and is that person affected by dust allergy and is the person is suffering from the dry cough or the person as the habit of snoring or not if all the parameters are there it will denotes as 1 if not it denotes as 0.if these are all present lung disease is **detected**.

6.CONCLUSION

The implementation of machine learning models for the prediction of lung and heart diseases has shown promising results. The evaluation metrics provide a comprehensive view of the models' performance.

The successful development of our machine learning based web application marks a significant step toward accessible early-stage detection of both heart and lung diseases.

In conclusion, this project illustrates the real-world impact of integrating machine learning with healthcare.

While the application is not intended to replace clinical diagnosis, it serves as an effective decision- support system.

7.REFERENCES

Researches from Charles Darwin university has published the AI model for diagnosing lung diseases from ultrasound videos with 94.1% accuracy in 2024

Zeynep Hilal published the research paper with the title as heart disease detection using vision based transformer model ECG image with 93.4% accuracy in 2023

M.S.G Prasad published a paper titled as An AI model for heart disease using machine learning

With 92.1% accuracy in 2022.

Matthew Bonomo¹, Michael G Hermsen¹, Samuel Kaskovich¹, Maximilian J Hemmrich¹, Juan C Rojas, Kyle A Carey, Laura Ruth Venable, Matthew M Churpek, Valerie G Press (2022)- “Using Machine Learning to Predict Likelihood and Cause of Readmission After Hospitalisation for Chronic Obstructive Pulmonary Disease Exacerbation”

“Detection and Staging of Chronic Obstructive Pulmonary Disease Using a Computed Tomography based Weakly Supervised Deep Learning Approach”- 2021.

“Comparison of machine learning algorithms for the identification of acute exacerbations in chronic obstructive pulmonary disease” by Chenshuo Wang Xianxiang Chen , Lidong Du , Qingyuan Zhan ,Ting Yang , Zhen Fang.(2019)

“An artificial intelligence approach to early predict symptom-based exacerbations of COPD” by Miguel Angel Fernandez-Granero, Daniel Sanchez-Morillo & Antonio Leon- Jimenez -2018.