Early Disease Prediction

Early Disease Prediction by giving symptoms using Machine Learning

|  |  |  |
| --- | --- | --- |
| Vikash Kumar  *CSE (NSUT)*  *New Delhi*  vikash.kumar.ug22@nsut.ac.in | Gaurav Gahlot  *CSE (NSUT)*  *New Delhi*  gaurav.gahlot.ug22@nsut.ac.in | Akshaj Gupta  *CSE (NSUT)*  *New Delhi*  akshaj.gupta.ug22@nsut.ac.in |

Dr Gaurav Singal

# Netaji Subhas University Of Technology

*New Delhi*

*Abstract*—Disease prediction using machine learning provides a promising solution for early diagnosis and preventive healthcare. This project focuses on creating a predictive model that analyzes patient information, such as symptoms and lifestyle habits, to estimate the risk of specific diseases.

I. INTRODUCTION

When someone falls ill, they often need to visit a doctor, which can be both expensive and time-consuming. Additionally, for individuals in remote areas with limited access to medical facilities, timely diagnosis becomes challenging, increasing the risk of severe health complications. An automated system capable of addressing this issue could save time and money while providing patients with quicker and more accessible diagnoses. In India, a large part of the population is vulnerable to common illnesses such as viral infections, flu, cough, and cold. Unfortunately, these mild symptoms are often ignored, even though they could indicate the early stages of more severe conditions. This underscores the critical need for early detection and prediction to reduce risks and protect public health. This project uses machine learning to create a fast, accurate, and easy-to-use disease prediction system. It helps people make better health decisions while reducing pressure on healthcare services, supporting early detection and healthier lifestyles.

Contributions:

* This project can be used to predict the disease the user might be suffering from.
* However , this project is not limited to just predicting the disease based on user input.
* There exist various kinds of symptoms , some of which might be difficult for the user to identify themselves.
* One of those symptoms is Anxiety. The user might not be fully aware of suffering from anxiety.
* This project solves this problem by itself identifying the presence of anxiety in the user.

II. LITERATURE/RELATED WORK

# A. Maintaining the Integrity of the Specifications

Many studies have examined how machine learning can be used in healthcare, especially for disease prediction. Algorithms like Decision Trees, Random Forests, Support Vector Machines, and Neural Networks have shown success in diagnosing diseases based on patient symptoms. Tools like chatbots and mobile apps are also being used to offer remote consultations and basic diagnoses. However, these systems often struggle with accuracy, real-time predictions, and ease of use. Moreover, the lack of sufficient and consistent datasets for linking symptoms to diseases makes predictions less reliable.

III. METHODOLOGY

# A. Source of Datasets

* The datasets used in this project are sourced from Kaggle, a reliable platform offering a wide range of publicly available datasets.
* Dataset used for Disease prediction dataset of roughly 5000 entries which consist of 41 unique diseases.
* Lung cancer dataset consist of around 400 entries.

# B. Preprocessing and Transformation

* The first step involves data transformation and preprocessing, ensuring the datasets are clean, structured, and optimized for efficient model training.
* Techniques include handling missing values, feature engineering, scaling, and encoding categorical data.

*C. Disease Prediction System:*

* The Dataset contains various columns . The first column contains the name of the disease and the corresponding columns contain the symptoms found for the corresponding disease .
* The dataset is converted by forming unique rows for each symptom existing in the dataset. These columns contain Boolean values which indicate the presence and absence of the symptom for every disease.

Objective:

This module allows users to input the symptoms they are experiencing.

Based on the input, the system predicts the most probable disease the user might have.

Algorithm: XGBoost (Extreme Gradient Boosting) is employed for disease prediction. It is chosen due to its high accuracy, scalability, and ability to handle structured datasets effectively.

*D. Lung Cancer Prediction System* Objective:

* This module is designed to assess the likelihood of lung cancer based on various user-provided parameters related to their current health status and lifestyle.
* The parameters also include user images which help in better understanding of the user for the prediction.
* The input images make the project more fun and interactive. This also helps in making the project more effective by maximizing the use of machine learning .
* The input images make the project more fun and interactive. This also helps in making the project more effective by maximizing the use of machine learning .
* The detection of anxiety , a required input parameter which the user might not be aware of themselves , is automated.

Algorithm:

* The initial step involves using a popular machine learning technique known as SMOTE.
* SMOTE (Synthetic Minority Oversampling Technique) is a method in machine learning that generates synthetic samples to balance imbalanced datasets by oversampling the minority class.
* The dataset contains very few entries of the “No” class and applying SMOTE helps counter the imbalance
* The Support Vector Machine (SVM) classifier is utilized for lung cancer prediction. It is selected for its ability to handle binary classification problems efficiently and deliver robust results, even with high-dimensional datasets.

# E. Hardware, Software, Prototype

* User Interface User Interfaces: Users can input the required parameters and simply click submit to view the prediction results. Users can provide symptoms to the system and receive a predicted disease.
* Hardware Interfaces: The application is compatible with personal computers, laptops, smartphones, and similar devices. It does not require any specialized hardware to function.
* Software Interfaces: VS code

Python

Jupyter notebook

Front end: StreamLit

Back end: Django python framework

IV. RESULT ANALYSIS

# A. Disease prediction

The disease prediction model determines the likelihood of the disease the user may be experiencing. The XGBoost algorithm is utilized to deliver accurate and efficient predictions.

XGBoost (Extreme Gradient Boosting) is a powerful and efficient machine learning library in Python, designed for scalable and accurate prediction tasks.

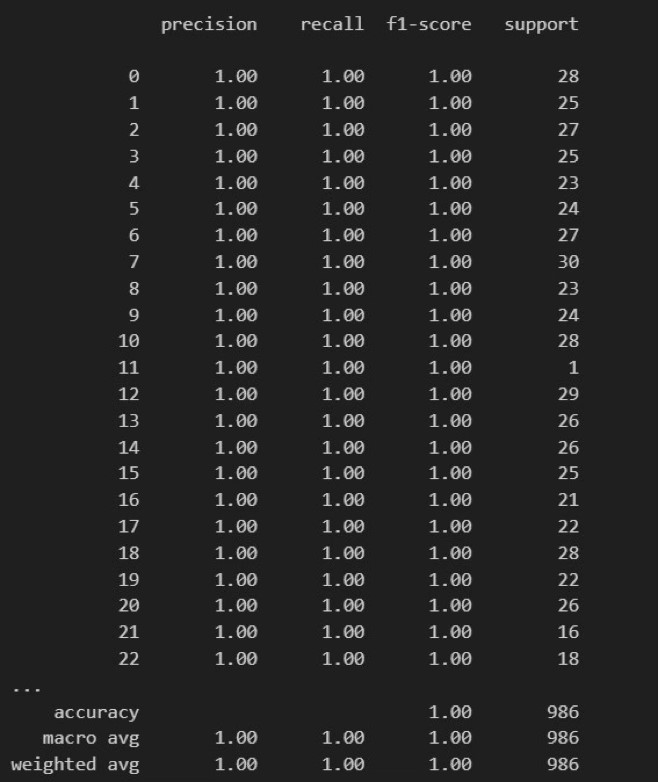


Fig. 1. Performance Metrics

# B. Lung cancer Prediction

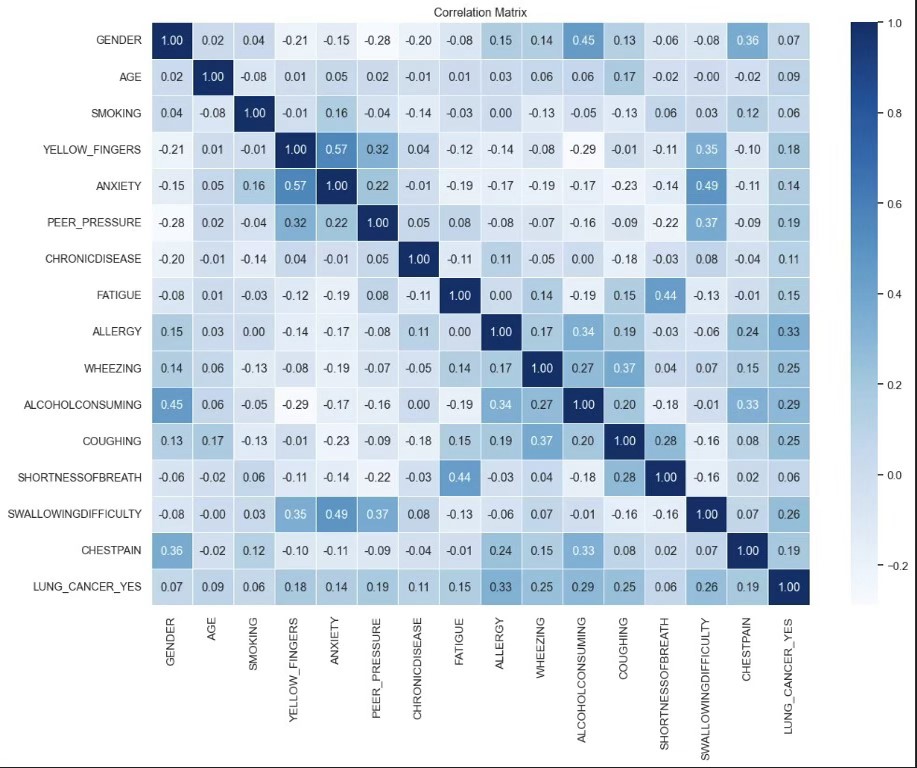


Fig. 2. Correlation Matrix

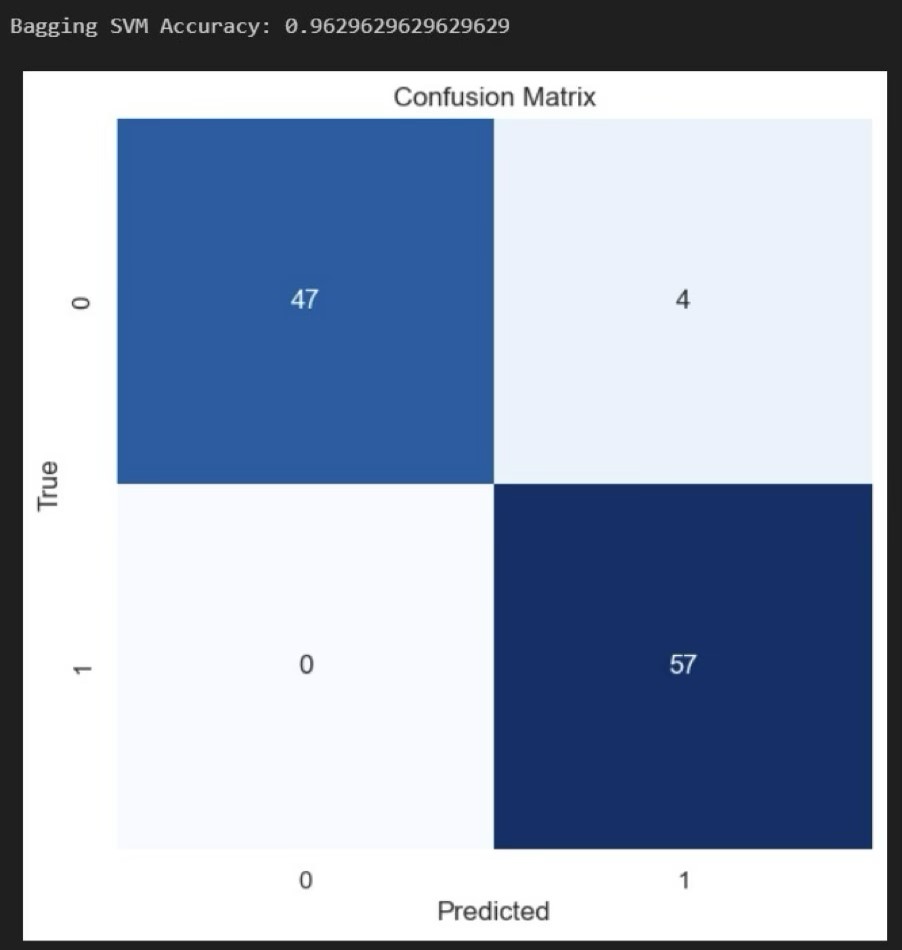


Fig. 3. Confusion matrix

Lung Cancer model Efficiently Predicts the Possibility of lung cancer with an accuracy of 0.96.

V. CONCLUSION AND FUTURE WORK

The main purpose of this project is to reduce the error in prediction In multiple diseases prediction system, a user can analyze more than one disease on a single website.

Functions: The user doesn’t need to traverse different places in order to predict whether he/she has a particular disease or not. In multiple diseases prediction system, the user needs to select the names of the symptoms the are facing as input parameters and just click on submit. The machine learning model will be invoked and it would predict the output and display it on the screen.

This project aims at further utilizing the machine learning technology rather than being limited to only making predictions based on the already existing datasets.

This project can identify few parameters for the user and hence they need not enter these manually.

Future work

* The project has further scope of improvement .
* There might be some symptoms which the user might not be able to correctly identify themselves.For these type of parameters , new ways can be introduced to efficiently identify them.

Few ways include:

* Using more machine learning/api’s models to identify more parameters.
* Integrating hardware devices (eg. Apple watch) for better understanding of users condition.
* Try to make the system user-friendly and provide a chatbot for normal queries.

REFERENCES

1. Zhou, S.-M., Fernandez-Gutierrez, F., Kennedy, J., Cooksey, R., Atkinson, M., Denaxas, S., Siebert, S., Dixon, W.G., O’Neill, T.W. and Choy, E., ”Defining disease phenotypes in primary care electronic health records by a machine learning approach: A case study in identifying rheumatoid arthritis”, PloS One, Vol. 11, No. 5, (2016), e0154515. https://doi.org/10.1371/journal.pone.0154515
2. Rathi, M. and Pareek, V., ”Disease prediction tool: An integrated hybrid data mining approach for healthcare”, IRACST- International Journal of Computer Science and Information Technology and Security (IJCSITS), ISSN, (2016), 2249-9555
3. Roberts, M.A. and Abery, B.H., ”A person-centered approach to home and community-based services outcome measurement”, Frontiers in rehabilitation Sciences, Vol. 4, (2023). doi: 10.3389/fresc.2023.1056530
4. Farooqui, M. and Ahmad, D., ”Disease prediction system using support vector machine and multilinear regression”, International Journal of Innovative Research in Computer Science and Technology (IJIRCST) ISSN, (2020), 2347-5552. https://doi.org/10.21276/ijircst.2020.8.4.15
5. Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M. and Mehendale, N., ”Disease prediction from various symptoms using machine learning”, Available at SSRN 3661426, (2020).

https://dx.doi.org/10.2139/ssrn.3661426

1. Pingale, K., Surwase, S., Kulkarni, V., Sarage, S. and Karve, A., ”Disease prediction using machine learning”, International Research Journal of Engineering and Technology (IRJET), Vol. 6, (2019), 831833. doi: 10.1126/science.1065467.
2. Negaresh, F., Kaedi, M. and Zojaji, Z., ”Gender identification of mobile phone users based on internet usage pattern”,International Journal of Engineering, Transactions B: Applications, Vol. 36, No. 2, (2023), 335347. doi: 10.5829/IJE.2023.36.02B.13.
3. Men, L., Ilk, N., Tang, X. and Liu, Y., ”Multi-disease prediction using lstm recurrent neural networks”, Expert Systems with Applications, Vol. 177, (2021), 114905. https://doi.org/10.1016/j.eswa.2021.114905
4. Paul, S., Ranjan, P., Kumar, S. and Kumar, A., ”Disease predictor using random forest classifier”, in 2022 International Conference for Advancement in Technology (ICONAT), IEEE.,(2022),1-4
5. Chhogyal, K. and Nayak, A., ”An empirical study of a simple naive bayes classifier based on ranking functions”, in AI 2016: Advances in Artificial Intelligence: 29th Australasian Joint Conference, Hobart, TAS, Australia, December 5-8, 2016, Proceedings 29, Springer., (2016), 324331.
6. Christianini, N. and Shawe Taylor, J., An introduction to support vector machines, cambridge unv. 2000, Press.
7. Ren, Q., Cheng, H. and Han, H., ”Research on machine learning framework based on random forest algorithm”, in AIP conference proceedings, AIP Publishing LLC. Vol. 1820,(2017),080020
8. Hamidi, H. and Daraee, A., ”Analysis of pre-processing and postprocessing methods and using data mining to diagnose heart diseases”, International Journal of Engineering, Transactions B: Applications, Vol.

29, No. 7, (2016), 921-930.