**Disease Prediction Research Paper**

Vince Francis Dsouza

MITU20BTCS0342  
MIT ADT UNIVERSITY  
LONI KALBHOR  
+91 9767851524

[vincedsouza02@gmail.com](mailto:vincedsouza02@gmail.com)

Abhishek Tekavade

MITU20BTCS0008  
MIT ADT UNIVERSITY  
LONI KALBHOR  
+91 9922839147

[abhishektekavade@gmail.com](mailto:abhishektekavade@gmail.com)

Aaditya Hariharan Nair

MITU20BTCS0003

MIT ADT UNIVERSITY  
LONI KALBHOR  
+91 7757984061

[aadityanair1803@gmail.com](mailto:aadityanair1803@gmail.com)

Tanuj Sinha

MITU20BTCS0314  
MIT ADT UNIVERSITY  
LONI KALBHOR

+91 8484040035

[tanujsinha2002@gmail.com](mailto:tanujsinha2002@gmail.com)

3rd E-mail

**ABSTRACT**

Machine Learning and its approaches are generally helpful for healthcare and biomedical sectors for predicting the disease. For trivial symptoms, the difficulty is to meet the doctors at any time in the hospital. Big Data provides essential data regarding the diseases on the basis of the patient's symptoms. For several medical organizations, disease prediction is important for making the best feasible health care decisions. Conversely, the conventional medical care model offers input as structured that requires more accurate and consistent prediction. This project is planned to develop the multi-disease prediction using the various machine learning concept.

Machine Learning Approach for Identifying Disease Prediction Using Machine Learning is based on prediction modeling that predicts disease of the patients according to the symptoms provided by the users as an input / output to the system.Here, the different datasets pertain to "Diabetes, lung cancer, liver and heart disease", from the benchmark UCI repository is gathered for conducting the experiment also few datasets are taken from Kaggle.

Here we will use the concept of supervised Machine Learning in which implementation will be done by applying Decision Tree, Random Forest and few other algorithms which will help in early prediction of diseases accurately and better patients care. The results ensured that the system would be functional and user oriented for patients for timely diagnoses of diseases in a patient.

**Keywords**

UCI;Disease;biomedical sectors;machine learning;kaggle;

# INTRODUCTION

Now-a-days, people face various diseases due to the environmental condition and their living habits. Especially due to Covid, one never knows what problem he can encounter. So, the prediction of disease at earlier stage becomes important task. Disease prediction using patient treatment history and health data by applying data mining and machine learning techniques is ongoing struggle for the past decades. The correct prediction of disease is the most challenging task. The recent success of deep learning in disparate areas of machine learning has driven a shift towards machine learning models that can learn rich, hierarchical representations of raw data with little pre-processing and produce more accurate results.

With the help of disease data, Machine Learning finds hidden pattern information in the huge amount of data, which in turn helps the patient to identify the problem. The main focus is on to use machine learning in healthcare to supplement patient care for better results. Machine learning has made it easier to identify different diseases and diagnosis correctly. Predictive analysis with the help of efficient multiple machine learning algorithms helps to predict the disease more correctly and help treat patients. Machine learning in healthcare aids the humans to process huge and complex medical datasets and then analyze them into clinical insights. This then can further be used by physicians in providing medical care. Hence machine learning when implemented in healthcare can leads to increased patient satisfaction.

# PROBLEM STATEMENT

In day to day life, it is difficult for one person to go to a doctor and get a check of diseases as it takes quite a lot amount of money as well as time. In our developing and technology dependent life we totally rely on gadgets. So, there should be a way, with whose help a person can at least check whether he has a particular disease or not. Using tech like machine learning in predicting the diseases using symptoms or concerned medical data is a need of the hour.

# RELATED WORK

## Existing work

* + 1. **Heart Disease**

Marimuthu et al. [16] aimed to predict heart diseases using supervised ML techniques. The authors structured the attributes of data as gender, age, chest pain, gender, target and slope [16]. The applied ML algorithms that were deployed are DT, KNN, LR and NB. As per analysis, the LR algorithm gave a high accuracy of 86.89%, which deemed to be the most effective compared to the other mentioned algorithms.

* + 1. **Lung cancer**

A technique k-Nearest-Neighbors was developed by Maleki et al [39], for which a genetic algorithm was used to efficiently pick features, to reduce the dimensions of the dataset and to improve the speed of the classifier. The experimental approach is used to determine the best value for k to increase the precision of the proposed algorithm. Use of the proposed solution to the database for lung cancer shows 100% accuracy.

* + 1. **Liver Disease**

A. Gulia et al. [11] in their proposed work researchers have done classification of the liver patient data using the algorithms like Bayesian Network, Support Vector Machine, J48, Multi-Layer Perceptron and Random Forest. The data from the UCI repository which is afforded by Center of Machine Learning and Intelligent Systems has used. After completion of their three-phase analysis, the Random Forest Algorithm is the best one with an accuracy of 71.87% has been concluded.

* + 1. **Diabetes**

Aiswaryaet al. [2] aims to discover solutions to detect the diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms. The research hopes to propose a faster and more efficient method of identifying the disease that will help in well-timed cure of the patients. Using PIMA dataset and cross validation approach the study concluded that J48 algorithm gives an accuracy rate of 74.8% while the naïve Bayes gives an accuracy of 79.5% by using 70:30 split.

## SOLUTION PROVIDED BY THE PROJECT

## This module deals with the filtering of data. Separating the data. Identifying the feature variables (input) and the Labels(output). Feature imputation: filling missing values .Feature encoding (turning values into numbers). One hot encoding: Turn all unique values into lists of 0’s and 1’. Label Encoding: Turn labels into distinct numerical values.

# METHODOLOGY AND GOAL

**4.1 Identification of the problem**

This Module deals with the identification of the type of the problem, whether it is a classification problem or a regression problem etc

**4.2 Dataset preparation and preprocessing**

This module deals with the filtering of data. Separating the data. Identifying the feature variables (input) and the Labels(output). Feature imputation: filling missing values .Feature encoding (turning values into numbers). One hot encoding: Turn all unique values into lists of 0’s and 1’. Label Encoding: Turn labels into distinct numerical values.

**4.3 Dataset Splitting**

This module deals with the Splitting the data into the Training set (usually 70-80% of data): Model learns on this. Validation set (usually 10-15% of data): Model hyperparameters are tuned on this. Test set (usually 10-15% of data). The Final test is done on this set.

**4.4 Modeling**

Choosing an algorithm, whether to choose Supervised algorithms – Linear Regression, Logistic Regression, Random forests, etc. Overfit the model, reducing the overfitting with regularization. Ensembling and Hyperparameter Tuning – run a bunch of experiments with different settings and see which works best.

**4.5 Analysis/evalution**

Comparing the model on various evaluation metrics such as Classification- Accuracy, Precision, Recall, F1, Confusion matrix, Mean average precision (object detection), Regression – Mean Squared Error(MSE), Mean Absolute Error (MAE) and Coefficient of Determination or sometimes also known as Goodness of fit( R2)

**4.6 Model Deployment**

Deploying the model for the use of Public, In our case deploying, the model on a website with Flask as back-end framework.

# Conclusion:

The project presented the technique of predicting the disease based on the symptoms, age, and gender of an individual patient. Different Machine learning algorithms were used to carry out the project such as the RandomForest, Decision Tree and Logistic Regression. Almost all the ML models gave good accuracy values. As some models were dependent on the parameters, they couldn’t predict the disease and the accuracy percentage was quite low. Once the disease is predicted, we could easily manage the medicine resources required for the treatment. This model would help in lowering the cost required in dealing with the disease and would also improve the recovery process.

We have also created a GUI for better interaction with the system by users which is very easy to operate .This paper shows that Machine Learning algorithm can be used to predict the disease easily with different parameters and models. To conclude, our system is helpful to those people who are always worrying about their health and they need to know what happens with their body. Our main motto to develop this system is to know them for their health. Especially, people who are suffering from mental illness like depression, anxiety. They can come out of these problems and can live their daily lives easily. Besides, our system provides better accuracy of disease prediction according to symptoms of the user, and also it will provide motivational thoughts and images. In the end, we can say that our system has no boundary of the user because everyone can use this system.

# REFERENCES

1. 1. A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, “Prediction of heart disease using machine learning,” in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275–1278.
2. 2. Y. Hasija, N. Garg, and S. Sourav, “Automated detection of dermatological disorders through image-processing and machine learning,” in 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 1047–1051
3. 3. S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1– 16, 2019
4. 4. Shrestha,Ranjit.,& Chatterjee,Jyotir. Moy. (2019).Heart Disease Prediction System Using Machine Learning . LBEF Research Journal of Science Technology and Management , 1(2), 115-132
5. 5. Godse, Rudra A.,Gunjal,Smita S., JagtapKaran A .,Mahamuni ,Neha S., Wankhade, Prof. Suchita. (2019). Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively.International Journal of Advance Research in Computer and Communication Engineering, 8(12), 50-52
6. 6. Anitha ,Dr.S.,& Sridevi,Dr.N. (2019). Heart Disease Prediction Using Data Mining Techniques.Journal of analysis and Computation ,13(2) , 48-55.
7. 7. Bindhika,Galla Siva Sai., Meghana,Munaga., ReddyManchuriSathvika. , N Rajalakshmi. (2020). Heart Disease Prediction Using Machine Learning Techniques. International Research Journal of Engineering and Technology, 7(4) , 5272-5276.
8. 8. Pingale,Kedar., Surwase, Sushant., Kulkarni,Vaibhav.,Sarage ,Saurabh., Karve, Prof. Abhijeet .(2019). Disease Prediction using Machine Learning.International Research Journal of Engineering and Technology, 6(12) , 2810-2813.
9. 9. Chauhan Raj H., NaikDaksh N. ,Halpati,Rinal A., Patel,Sagarkumar J. , &PrajapatiMr. A.D. (2020). Disease Prediction using Machine Learning.International Research Journal of Engineering and Technology, 7(5) , 2000-2002.

(Note: The paper should not exceed 4 pages.)