# HUMAN DISEASE PREDICTION SYSTEM: HARNESSING THE POWER OF MULTIPLE MACHINE LEARNING ALGORITHMS

**Jyoti Anand[1*], Shatabdi Basu[2], Shivam Choudhary[3], Subhodeep Nayak[4], Salman Hossein Peada[5]**

**Abstract:**
Prior prediction of various diseases such as cancer, diabetes, heart, lungs, etc. using Machine Learning (ML) algorithms have become a big boon in recent times. Lavish lifestyle and environmental pollution lead to occur fatal diseases viz. heart attack, cancer, asthma, etc. in the human body and may cause premature death. With the help of an ML predicting model, it's easy to identify the disease without going to the hospital physically. ML is the subset of artificial intelligence, which helps to develop the intelligence ability in a system. In this work, we are providing a graphical interface or, web interface to the users, where they can feed their physiological symptoms and predict the diseases. Four supervised ML algorithms i.e., Random Forest, Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) are applied to the dataset to find the accuracy. The dataset is taken from Kaggle, consisting of 18 different features to predict multiple diseases like asthma, chicken pox, dengue, thyroid, etc. Results show that Random Forest performs well among others with 98% of accuracy. The accuracy scores of Logistic Regression (LR), Support Vector Machine (SVM) & K- Nearest Neighbor (KNN) are 91%, 90% & 82% respectively.

**Keywords:** Disease Prediction, Random Forest, SVM, KNN, Linear Regression, Machine Learning

[1*, 3, 4, 5]University of Engineering & Management, Jaipur
[2]Manipal University, Jaipur
[1] Email: anandjyoti136@gmail.com
[2] Email: shatabdi.bs@gmail.com
[3] Email: shivamkumar11062@gmail.com
[4] Email: subhadipnayak56@gmail.com
[5] Email: peadasalman1819@gmail.com

[*]**Corresponding Author**: Jyoti Anand
[*]University of Engineering & Management, Jaipur

**DOI:** 10.48047/ecb/2022.11.12.38

## I. Introduction

Today we are living in a modern era of technology, and the majority of people own smartphones in addition to other digital gadgets. These gadgets produce an enormous quantity of data that is further used to create a prediction model. Prior to now, predictions have been made using statistical models, but this is inadequate for large amounts of data [1]. These models couldn't deal with data categorization, missing values, and large data points. Machine Learning (ML) has become a big boon to performing all these tasks efficiently. It's a subset of Artificial Intelligence (AI), which also helps to develop the intelligence in machines. A number of statistical, probabilistic, and optimization techniques are used by ML algorithms to learn from prior knowledge or experience and find meaningful patterns in huge, unstructured, and complicated datasets [2]. ML has a range of applications such as pattern recognition, image & text categorization, disease prediction, recommendation systems, etc. As we know, the good health of mankind is a priority concern now [3]. Because so many fatal diseases viz. heart attack, cancer, asthma, etc. cause premature death. Any patient who is sick must schedule an expensive and time-consuming appointment with a doctor. It becomes a challenging task for the patient to pay a visit to the doctors or hospitals and know about their sickness. Therefore, in this paper, we are creating a prediction model to identify the different diseases based on their symptoms. Additionally, it saves time and money. Patients can simply put their symptoms like runny nose, headache, body ache, blood pressure, temperature, etc., and diagnose their disease.

Building a prediction model require ML algorithms, which are broadly categorized into three types; supervised, unsupervised &

reinforcement. Supervised ML learns from the examples and uses a labeled training dataset to train the algorithm [4]. It further categorizes the unlabeled dataset into similar groups. Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Regression Tree (RT), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Bayesian Models come under supervised ML. Next, unsupervised ML learns patterns from unlabeled data [5]. The algorithms like Principal Component Analysis (PCA), k-means, and factor analysis are examples of unsupervised learning. Third, reinforcement learning aims to explore different possibilities until it finds the correct response. In contrast to supervised learning, reinforcement learning does not require precise input and output sets or explicitly defined suboptimal actions. To build our disease prediction model, we are using four supervised ML algorithms viz. RF, LR, SVM & KNN. Initially, we used a dataset from Kaggle. Afterward, the data is cleaned and trained. Section II facilitates an overview of related work done by various authors. All studied research papers are also summarized in Table I. Then, a brief discussion of the proposed prediction model is given in section III. The methodology adopted in this paper has been detailed in section IV. The outcomes of the simulation of the Python-based algorithms RF, LR, SVM, and KNN are compiled in the results section V.

## II. Literature Review

A group of tools created by the AI community are included in the survey article [4]. Such tools offer the chance for a better decision-making process and are highly helpful for the examination of such issues. SVM, KNN and Confusion Matrix learning algorithms are used for predicting multiple diseases (especially three diseases i.e., heart, diabetes & Parkinson's) on a web application [6]. For this, users have to perform a login or signup process on the web application and input his\her physical symptoms. The model verifies all provided symptoms and offers a diagnosis but only about one disease. Mohit et. al. proposed a web platform to predict three multiple diseases viz. breast cancer, heart, and diabetes [7]. Models are trained using a lightweight framework such as Flask API. Three machine learning algorithms i.e., Logistic Regression, KNN, and SVM are utilized to check the accuracy. Results show that Logistic Regression gave more accuracy as compared to the other two. In [8], authors predict heart disease, especially for diabetic patients. Diabetes is a chronic illness that develops when the pancreas either produces insufficient amounts of insulin or when the body improperly uses the insulin that is produced. Here, the decision tree model beats the Naïve Bayes and SVM learning models in terms of accuracy and error. Multiple supervised learning algorithms like SVM, KNN, Decision Tree, Naïve Bayes, and Random Forest are used to analyze the huge medical data in [9]. It helps to give a feasible, robust, and reliable system to diagnose the disease on time. With the use of a Python built-in module called Streamlit, authors attempted to diagnose the ailment at its earliest stages of development [10]. Different parameters viz. pulse rate, cholesterol, blood pressure, heart rate, etc. are taken to train the model and predict the diseases. Ture et. al. collected the datasets from Kaggle to predict three fatal diseases such as heart, kidney, and diabetes at an early stage [11]. Data is then appropriately cleansed and evaluated to find additional accuracy. The dataset is split in an 80:20 ratio for training and testing purposes respectively. Multiple learning models are applied, but random forest gave the best result of all. Furthermore, the Pickle library and Flask framework of python are used to dump the data and create interactive web applications. Dhomse Kanchan B and others first applied the Principle Component Analysis (PCA) to minimize the attributes in the dataset [12]. Then SVM, Naïve Bayes and DT are applied to find the possibility of cardiovascular disease and diabetes in the patients. The WEKA is a data mining technique, used here for accuracy purposes. Paper [13] made use of various learning algorithms for early prediction of any diseases with more accuracy. Authors are particularly focused on predicting heart disease on clinical data of patients [14]. Four attributes demographic, behavioural, and medical history along with the current situation are adopted from the dataset. The work given in [15] predicts the risks of heart attack with numerous learning models. The dataset is consisting of 14 attributes with 303 records taken from Kaggle. The Exploratory Data Analysis (EDA) process is employed for data cleaning. It considers only nominal data because the model works with the same. Numerous classification techniques are utilized to build an accurate model to predict diabetes, heart & liver types of diseases [16]. For this, it trains the UCI dataset and then applied KNN, XGBoost, and RF machine learning algorithms on them. In order to predict prior information about disease Multinomial Naïve Bayes (MNV), LR and DT ML algorithms are used in [17]. Each tag's likelihood is evaluated by MNV for each sample, and the tag with the peak likelihood is produced. Decision Tree is quite simple to read and expressive for smaller and learned discrete values respectively. With the use of TensorFlow, Flask API, ML algorithms, the authors built a prediction model in [18]. Python pickling and un-pickling save the model behavior and load the pickle file respectively.

**Table I:** Summary of Review Papers

| Authors' Name & Year | Predicted Diseases | Used ML Algorithms | Accuracy (in %) |
|---|---|---|---|
| Fatima et. al., 2017 [4] | Hear, Diabetes, Lever, Dengue & Hepatitis | SVM, NB, FT, RS, FFNN | 94.60, 95, 97.10, 100, 98 |
| Pattar et. al., 2022 [6] | Heart, Diabetes & Parkinson | SVM, KNN & Confusion Matrix | Numerical Analysis isn't given. |
| Mohit et. al., 2021 [7] | Heart, Diabetes & Breast Cancer | SVM, Logistic Regress & KNN | 83.84, 77.60, 94.55 resp. |
| Arumugam et. al., 2021 [8] | Heart for diabetic persons | Naïve Bayes, SVM, Decision Tree | 76, 87, 90 resp. |
| Ture et. al., 2023 [11] | Heart, Kidney & Diabetes | SVM, LR, RE, DT, & XG Boost | Separate numerical analysis is carried out for each disease |
| Kanchan et. al., 2016 [12] | Diabetes & Heart | SVM, Naïve Bayes & DT | 45, 52, 55 |
| Kumar et. al., 2021 [13] | Multiple common diseases | DT, RF, Naïve Bayes & KNN | 95.12, 95.11, 95.21, 95.12 |
| Saim et. al., 2022 [14] | Heart | Logistic Regression, KNN & SVM | 65.5, 82.45, 86.64 |
| Shanbhag et. al., 2021 [15] | Risk of heart attack | KNN, DT, RF, LR & SVM | 84.2, 70.33, 79, 85, 85.7 |
| Singh et. al., 2022 [16] | Heart, Lever & Diabetes | KNN, RF & XGBoost | 85, 77, 89 |
| Patil et. al., 2022 [17] | Cold, Malaria, Typhoid | Multinomial NB, DT, LR | 92, 97, 89 |
| Yaganteeswarudu, 2020 [18] | Diabetes, Heart & Cancer | LR, RF, SVM | 92, 95, 96 |

## III. Predictive algorithms

### 3.1 Logistic Regression

It's an S-shaped curve, developed for statistical functions. In a categorical dependent variable, the output is predicted via logistic regression. The result should be a distinct or categorical value, providing probabilistic values ranging from 0 to 1 instead of precise values within that range. It can take the form of True or False, 0 or 1, or Yes or No. It is employed to address classification-related issues. In logistic regression, instead of fitting a regression line, we utilize an "S" shaped logistic function to predict two possible outcomes (0 or 1). The curve of the logistic function represents various possibilities, such as determining if cells are malignant or not, or if a mouse is obese based on its weight. Mathematically it's obtained by the use of equation (1).

$$\log \frac{y}{1-y} = y \qquad (1)$$

Here, y is the simple straight-line equation which can be either 0 or 1 for zero and infinity respectively.

### 3.2 Support Vector Machine

Two approaches i.e., classification and regression are used by SVM. This technique employs coordinates to plot data in n-dimensional space. SVMs come in the form of both linear and nonlinear varieties. We employ the linear SVM classifier in our research because we are working with linearly separable data. It is obtained using equation (2).

$$y = wx + b \qquad (2)$$

Here, 'w' is the weighted factor, and 'b' is the bias.

### 3.3 K-Nearest Neighbor (KNN)

In KNN, 'K' is the total number of neighbors required to classify the dataset. KNN is computationally expensive because it requires numerous iterations to get the highest level of accuracy. It predicts the output based on the labeled data. Even with big and noisy training data, the algorithm still performs well. The dataset is split into train and test datasets by the algorithm. To create and train models, we use the training dataset. The model created predicts the test data. It's obtained from the Euclidean Distance formula, given in equation (3).

$$ED = \sqrt{\sum_{i=1}^{n} (\text{xi} - \text{yi})\text{\^{}}2} \qquad (3)$$

### 3.4 Random Forest

An extremely popular supervised machine learning approach for classification and regression problems is the Random Forest Method. It becomes increasingly accurate and capable of addressing problems as it contains more trees. RF is consisting of numerous decision trees on various subsets of a given dataset to improve the predictability of its results further. Ensemble learning is employed to tackle a challenging problem and improve the model's performance. Ensemble signifies the blending of various models. RF uses Gini Index (GI) to find the accuracy, expressed in equation (4).

$$GI = 1 - \sum_{i=1}^{n} (Pi)\text{\^{}}2 \qquad (4)$$

Here, P is the probability of positive class and negative class.

## IV. Methodology

### 4.1 Datasets

For disease prediction, we have taken the dataset from Kaggle. The dataset consists of 18 features with 4920 numbers of rows. After data collection, it must be cleaned to remove the duplicate data from the data set. Data cleaning is done in two steps, first for finding the duplicate values and the others for null values. For finding duplicate values ‘data.duplicate().sum()’ function is used in python. Here, duplicate() finds the duplicate value and sum() gives the total number of duplicate values. Then drop_duplicates() is applied to eliminate all the duplicate values and keep only the unique values. After removing duplicate values there are 304 rows and 18 features in the dataset. Get_dummies() python function converts all categorical data to the numerical data to get more accuracy.

### 4.2 Workflow

We are providing an interface, aiming to predict multiple diseases. Figure 1 shows the workflow of our prediction model. Four algorithms i.e., RF, LR, SVM, and KNN are utilized in the prediction model to forecast the disease. At first, data is collected, then goes for pre-processing followed by five steps. These five steps are feature selection, model selection, model training, model evaluation, and model optimization. The data has been pre-processed before visualization. Data is divided into training data and testing data after data pre-processing is complete. The next phase involves applying algorithms to find the disease accurately. The output will then be stored in a pickle file for each ailment. The Flask API framework is used to display the model's output on the web.

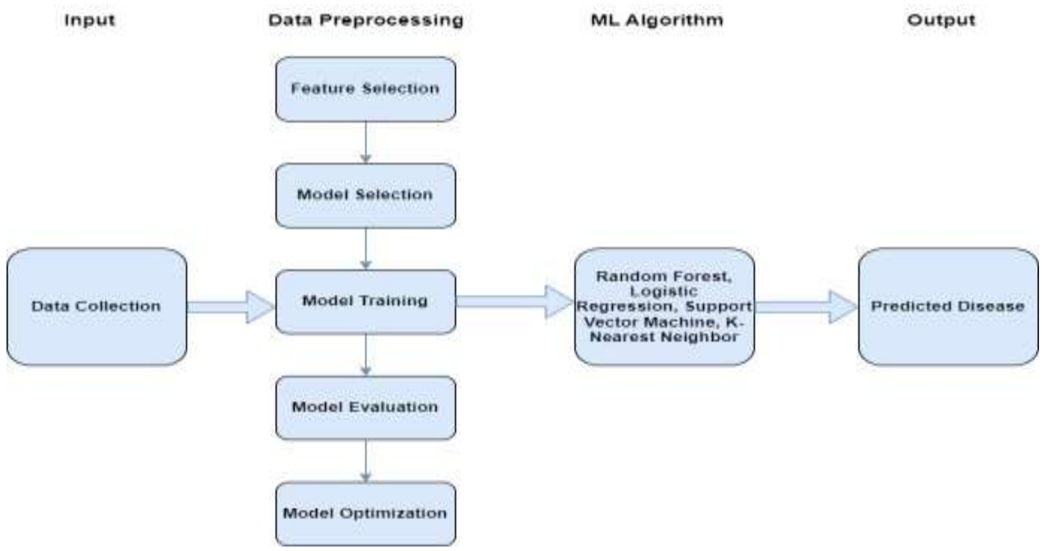


**Figure 1:** Block Diagram of Disease Prediction Model

## V. Results and discussions

As discussed above, four machine learning algorithms are utilized to predict diseases. Figure 2 shows the result of the accuracy score of all four learning models. The most reasonable performance metric is the accuracy score, which is just a ratio of properly expected observations to all observations. It can also be determined in terms of positives (mean 1) and negatives (means 0) for binary classification. Random Forest outperforms all other algorithms in terms of accuracy. Figure 3 is the result of test accuracy. Generally, test accuracy refers to the "validation accuracy," or the accuracy you determine using a data set that was not used for training but was instead used for validating (or "testing") throughout the training process. It is calculated by determining the accuracy score between $y_{test}$ and predictions on $x_{test}$ data. Again Random Forest works well here as compared to other three algorithms. Figure 4 shows a snapshot of the user interface which accepts symptoms from the user and provides a predicted disease name.

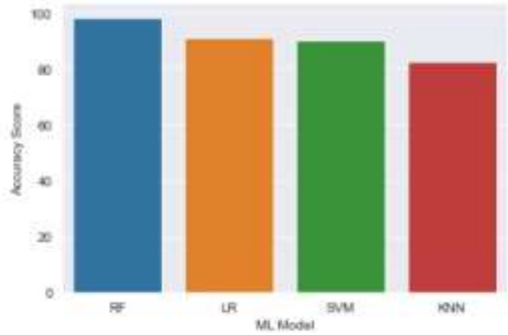


**Figure 2:** Accuracy Score of RF, LR, SVM & KNN

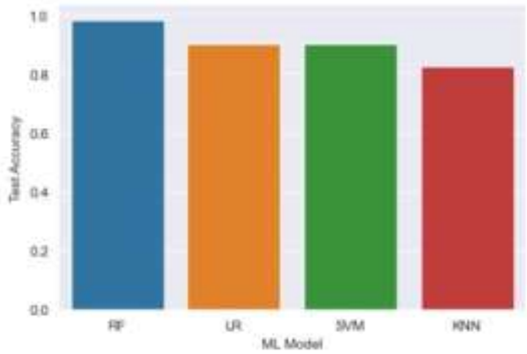


**Figure 3:** Test Accuracy of RF, LR, SVM & KNN


**Figure 4:** User interface of the system

Moreover, the obtained values of train accuracy, test accuracy and accuracy scores of all four models are summarized in Table 2. Before applying the learning algorithms to the dataset, it needs to be split into two parts. One part is for training the data and the other is for testing the data. Trained data is the first dataset that we use to teach a machine learning program to recognize patterns. It is the combined factor of the predictor's training data ($x_{train}$) and the prediction's training data ($x_{test}$).While testing or validation data are used to assess the accuracy of the model, a combination of the predictor's testing data ($y_{train}$) and prediction's testing data ($y_{test}$). The accuracy score determines how accurately a model predicts its output.

**Table 2:** Train, Test & Accuracy Score of all ML Model

| ML Model | Train Accuracy | Test Accuracy | Accuracy Score |
|---|---|---|---|
| RF | 1.000 | 0.984 | 98.36 |
| LR | 1.000 | 0.902 | 91.30 |
| SVM | 1.000 | 0.902 | 90.16 |
| KNN | 0.967 | 0.826 | 82.60 |

## VI. Conclusion

The primary motive of this paper is to predict the diseases as per the symptoms fed by the patients. A prediction model is created on the basis of already existing medical data. To categorize patient data, a number of general illness prediction systems based on machine learning algorithms are included called random forest, logistic regression, KNN, and support vector machines. Results show that random forest gives the best performance test accuracy as well as accuracy score compared to others. In conclusion, our approach will help people who are constantly concerned about their health. Our goal in creating this system is to increase people's awareness of their health problems and enable them to live healthier lives. In the future, we can utilize more unexplored physiological symptoms of the human body to predict and analyze diseases.

## References

1. A Yaqoob, R. M. Aziz, N. K. Verma, P. Lalwani, A. Makrariya, and P. Kumar, “A review on nature-inspired algorithms for cancer disease prediction and classification,” Mathematics, vol. 11, no. 5, p. 1081, 2023.
2. S. Xie, Z. Yu, and Z. Lv, “Multi-disease prediction based on deep learning: A survey,” CMES-Computer Modeling in Engineering & Sciences, vol. 128, no. 2, 2021.
3. S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” IEEE access, vol. 7, pp. 81 542–81 554, 2019.
4. M. Fatima, M. Pasha, “Survey of machine learning algorithms for disease diagnostic,” Journal of Intelligent Learning Systems and Applications, vol. 9, no. 01, p. 1, 2017.
5. 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.
6. Pattar et. al., “A SURVEY PAPER ON “MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING,” International Journal of Engineering Applied Sciences and Technology, 2022, vol. 7, ISSN No. 2455-2143, pp. 53-56.
7. I. Mohit, K. S. Kumar, U. A. K. Reddy, and B. S. Kumar, “An approach to detect multiple diseases using machine learning algorithm,” in Journal of Physics: Conference Series, vol. 2089, no. 1. IOP Publishing, 2021, p. 012009.
8. K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, “Multiple disease prediction using machine learning algorithms,” Materials Today: Proceedings, vol. 80, pp. 3682–3685, 2023.
9. M. E. Farooqui and D. J. Ahmad, “A detailed review on disease prediction models that uses machine learning,” International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, pp. 2347–5552, 2020.
10. Shaikh et. al., “Multiple Disease Prediction Webapp,” Journal of Emerging Technology & Innovative Research, 2022, Vol. 9, ISSN No. 2349-5162, p. 225-234.
11. Tanmay Ture, Amol Sawant, Rohan Singh, and Chetna Patil. "Multiple Disease Prediction System," International Journal for Research in Applied Science & Engineering Technology (IJRASET), 2023, ISSN No. 2321-9653, pages 1238-1244.
12. B. D. Kanchan and M. M. Kishor, “Study of machine learning algorithms for special disease prediction using principal of component analysis,” in 2016 international conference on global trends in signal processing, information computing and communication (ICGTSPICC). IEEE, 2016, pp. 5–10.
13. A. Kumar and M. A. Pathak, “A machine learning model for early prediction of multiple diseases to cure lives,” Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 6, pp. 4013–4023, 2021.
14. M. M. Saim and H. Ammor, “Comparative study of machine learning algorithms (SVM, logistic regression, and KNN) to predict cardiovascular diseases,” in E3S Web of Conferences, vol. 351. EDP Sciences, 2022, p. 01037.
15. A. A. Shanbhag, C. Shetty, A. Ananth, A. S. Shetty, K. K. Nayak, and B. Rakshitha, “Heart attack probability analysis using machine learning,” in 2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER). IEEE, 2021, pp. 301–306.
16. Singh et. al. “Multiple Disease Prediction System,” International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, pp. 1698- 1701, 2022.
17. K. Patil, S. Pawar, P. Sandhyan, and J. Kundale, “Multiple disease prognostication based on symptoms using machine learning techniques,” in ITM Web of Conferences, vol. 44. EDP Sciences, 2022, p. 03008.

18. A. Yaganteeswarudu, “Multi disease prediction model by using machine learning and flask api,” in 2020 5th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2020, pp. 1242–1246.