**Noname manuscript No.**
(will be inserted by the editor)

# Disease prediction from various symptoms using machine learning

**Rinkal Keniya · Aman Khakharia · Vruddhi Shah · Vrushabh Gada · Ruchi Manjalkar · Tirth Thaker · Mahesh Warang · Ninad Mehendale** *

Received: date / Accepted: date

**Abstract** Accurate and on-time analysis of any health-related problem is important for the prevention and treatment of the illness. The traditional way of diagnosis may not be sufficient in the case of a serious ailment. Developing a medical diagnosis system based on machine learning (ML) algorithms for prediction of any disease can help in a more accurate diagnosis than the conventional method. We have designed a disease prediction system using multiple ML algorithms. The dataset used had more than 230 diseases for processing. Based on the symptoms, age, and gender of an individual, the diagnosis system gives the output as the disease that the individual might be suffering from. The weighted KNN algorithm gave the best results as compared to the other algorithms. The accuracy of the weighted KNN algorithm for the prediction was 93.5 %. Our diagnosis model can act as a doctor for the early diagnosis of a disease to ensure the treatment can take place on time and lives can be saved.

**Keywords** Disease prediction · machine learning · symptoms

## 1 Introduction

Medicine and healthcare are some of the most crucial parts of the economy and human life. There is a tremendous amount of change in the world we are living in now and the world that existed a few weeks back. Everything has turned gruesome and divergent. In this situation, where everything has turned virtual, the doctors and nurses are putting up maximum efforts to save people's lives even if they have to danger their own. There are also some remote villages which lack medical facilities. Virtual doctors are board-certified doctors who choose to practice online via video and phone appointments, rather than in-person appointments but this is not possible in the case of emergency. Machines are always considered better than humans as, without any human error, they can perform tasks more efficiently and with a consistent level of accuracy. A disease predictor can be called a virtual doctor, which can predict the disease of any patient without any human error. Also, in conditions like COVID-19 and EBOLA, a disease predictor can be a blessing as it can identify a human's disease without any physical contact. Some models of virtual doctors do exist, but they do not comprise the required level of accuracy as all the parameters required are not being considered. The primary goal was to develop numerous models to define which one of them provides the most accurate predictions. While ML projects vary in scale and complexity, their general structure is the same. Several rule-based techniques were drawn from machine learning to recall the development and deployment of the predictive model. Several models were initiated by using various machine learning (ML) algorithms that collected raw data and then bifurcated it according to gender, age group, and symptoms. The data-set was then processed in several ML models like Fine, Medium and Coarse Decision trees, Gaussian Naïve Bayes, Kernel Naïve Bayes, Fine, Medium and Coarse KNN, Weighted KNN, Subspace KNN, and RUSBoosted trees. According to ML models, the accuracy varied. While processing the data, the input parameters data-set was supplied to every model, and the disease was received as an output with

* Corresponding author
N. Mehendale
B-412, K. J. Somaiya College of Engineering, Mumbai, India
Tel.: +91-9820805405
E-mail: ninad@somaiya.edu

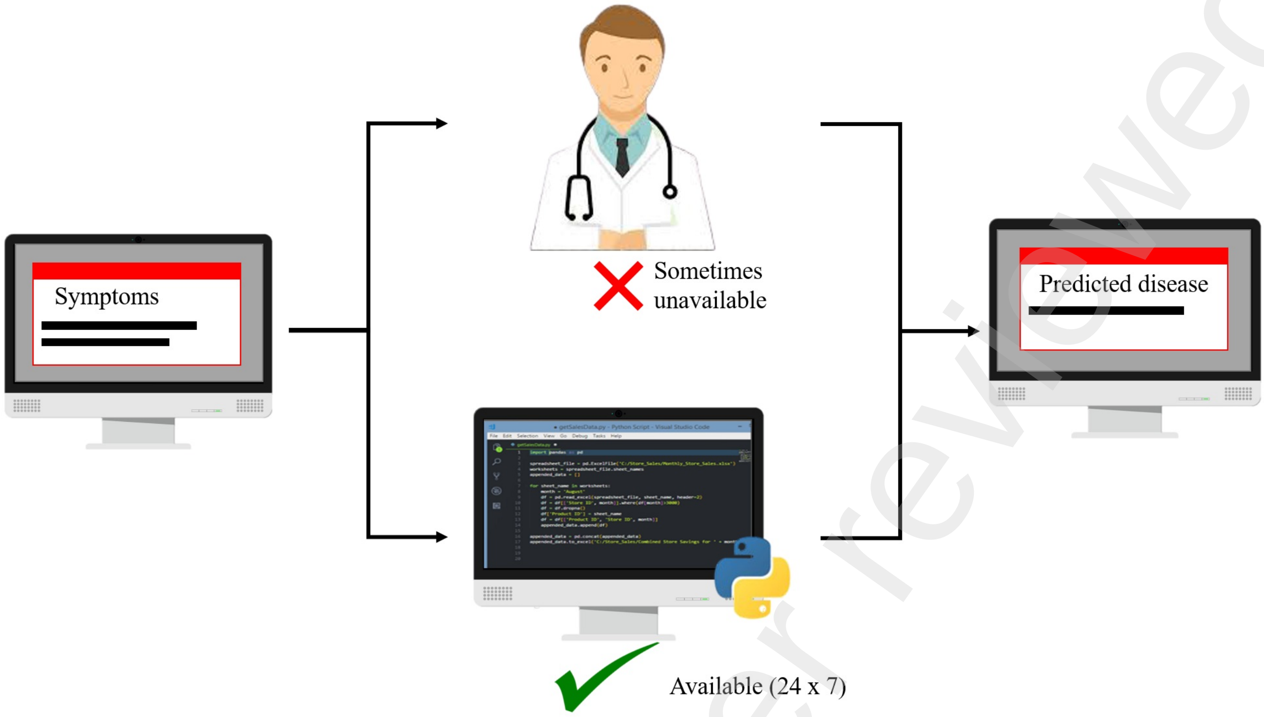

**Fig. 1** Proposed system for disease prediction. The doctor may not be available always when needed. But, in the modern time scenario, according to necessity one can always use this prediction system anytime. The symptoms of an individual along with the age and gender can be given to the ML model to further process. After preliminary processing of the data, the ML model uses the current input, trains and tests the algorithm resulting in the predicted disease.

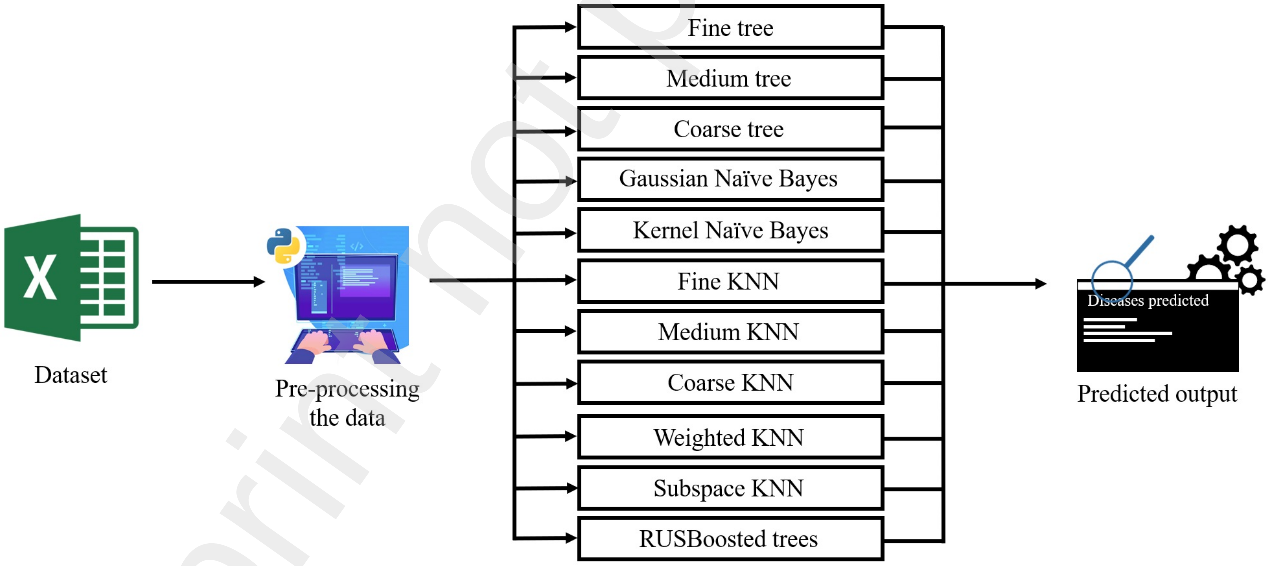

**Fig. 2** Proposed system flow diagram during training. The dataset consisting of gender, symptoms, and age of an individual was preprocessed and fed as an input to different ML algorithms for the prediction of the disease. The different ML models used were Fine, Medium and Coarse Decision trees, Gaussian Naïve Bayes, Kernel Naïve Bayes, Fine, Medium and Coarse KNN, Weighted KNN, Subspace KNN, and RUSBoosted trees. The outcome of the models is the disease as per the symptoms, age, and gender is given to the processing model.

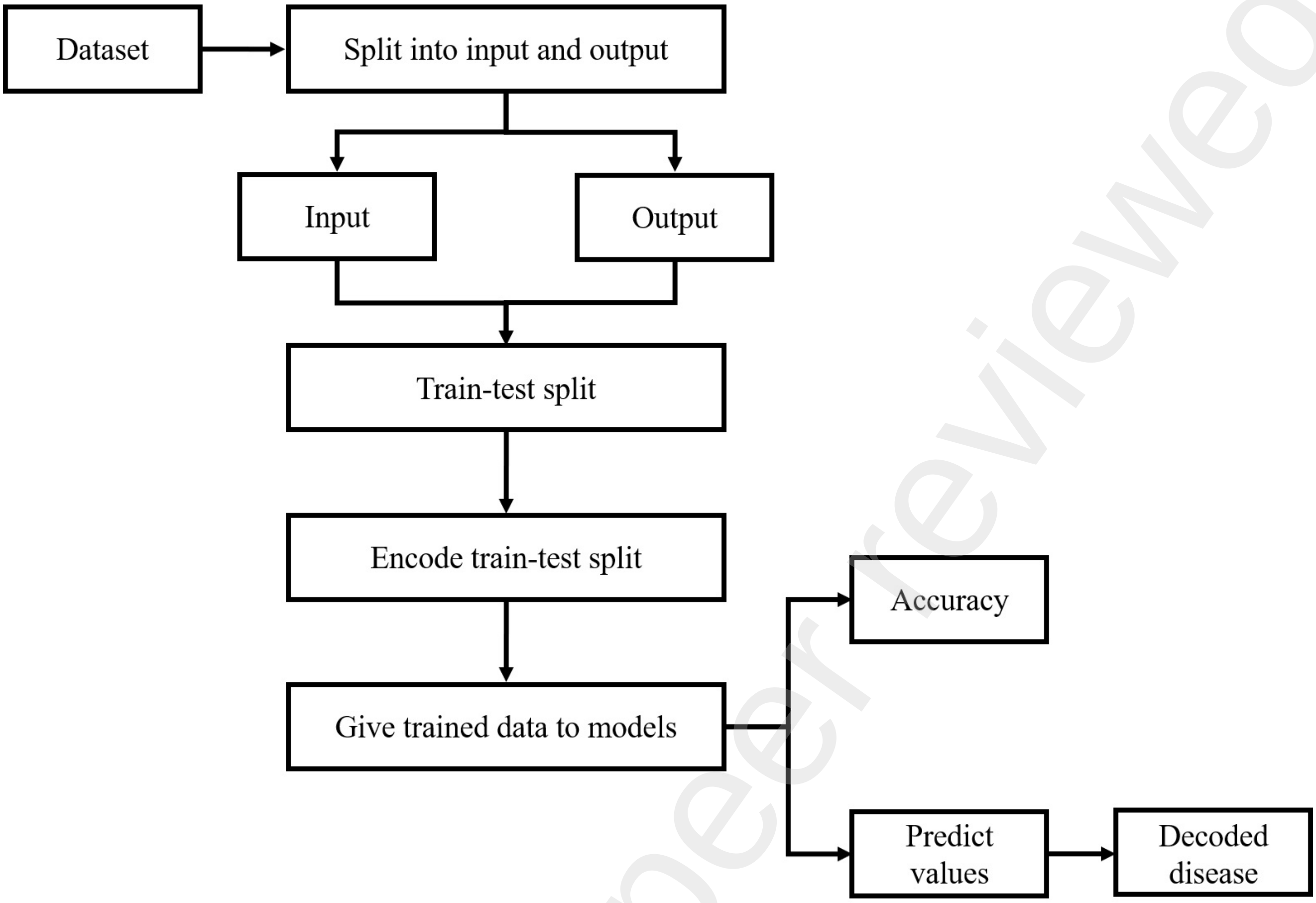

**Fig. 3** Functioning of the ML models. The dataset was split into input consisting of age, gender, and symptoms and the output as the diseases based on the input factors. We randomly split the available data into train and test sets. These sets were then encoded and further trained using different algorithms. After which the algorithms test the training set and predict the values, resulting in the accuracy of different ML algorithms. The predicted values were then decoded to give the output as the disease.

| **Model** | **Accuracy** |
|---|---|
| Fine Tree | 21.8 |
| Medium Tree | 12.3 |
| Coarse Tree | 6.4 |
| Fine KNN | 80.3 |
| Medium KNN | 61.8 |
| Coarse KNN | 5.3 |
| Weighted KNN | 93.5 |
| Gaussian Naïve Bayes | 16.8 |
| Kernel Naïve Bayes | 16.8 |
| Subspace KNN | 73.2 |
| RUSBoosted Trees | 0.5 |

**Table 1** Accuracy values of different ML models

dissimilar accuracy levels. The model with the highest accuracy has been selected.

## 2 Literature review

Numerous research works have been carried out for the prediction of the diseases based on the symptoms shown by an individual using machine learning algorithms. Monto *et al.* [6] designed a statistical model to predict whether a patient had influenza or not. They included 3744 unvaccinated adults and adolescent patients of influenza who had fever and at least 2 other symptoms of influenza. Out of 3744, 2470 were confirmed to have influenza by the laboratory. Based on this data, their

| Method | Model used | Maximum Accuracy (%) |
|---|---|---|
| Mir *et al.* [1] | Naive Bayes, SVM<br>Random Forest and<br>Simple CART | 79.13 |
| Khourdifi *et al.* [2] | KNN | 99.7 |
| Vijayarani *et al.* [3] | SVM | 79.66 |
| Mohan *et al.* [4] | HRFLM | 88.4 |
| Sriram *et al.* [5] | Random forest | 90.26 |
| Our proposed method | Fine, Medium, Coarse, and Weighted KNN;<br>Gaussian Naïve Bayes,<br>Kernel Naïve Bayes;<br>Coarse, Medium, and Fine Decision trees;<br>SubSpace KNN,<br>RUSBoost algorithm | 93.5 |

**Table 2** Comparison of the methodologies reported in existing literature

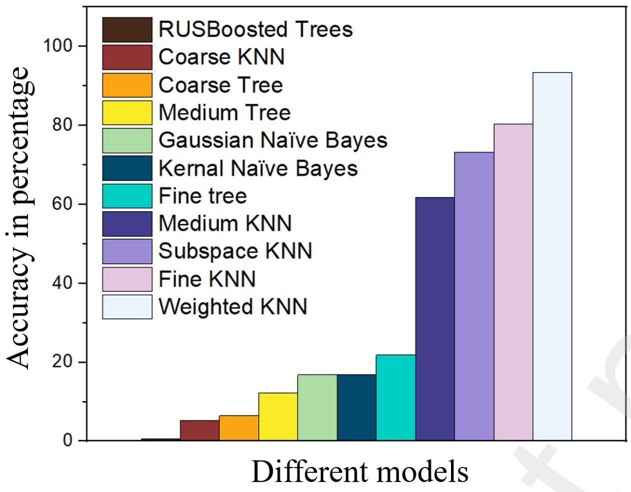

**Fig. 4** Comparison of the accuracy values of the different ML algorithms. The Weighted KNN model gave the highest accuracy as compared to the other ML algorithms. The RUSBoosted trees were the least accurate model. The Fine KNN performed better than the Subspace, Medium, and Coarse KNN models. The least efficient KNN model was coarse KNN. The Gaussian and the Kernel Naïve Bayes algorithm had a comparable accuracy with each other though less than the KNN models. The Fine tree had a higher accuracy than the medium and the coarse decision tree models.

model gave an accuracy of 79 %. Sreevalli *et al.* [7] used the random forest machine-learning algorithm to predict the disease based on the symptoms. The system resulted in low time consumption and minimal cost for the prediction of diseases. The algorithm resulted in an accuracy of 84.2 %. Various tools were developed by Langbehn *et al.* [8] to detect Alzheimer's disease. Data for 29 adults were used for the training purpose of the ML algorithm. They had developed classification models to detect reliable absolute changes in the scores with the help of SmoteBOOST and wRACOG algorithms. A variety of ML techniques such as artificial neural networks (ANNs), bayesian networks (BNs), support vector machines (SVMs) and decision trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making [9].

Karayilan *et al.* [10] proposed a heart disease prediction system that uses the artificial neural network backpropagation algorithm. 13 clinical features were used as input for the neural network and then the neural network was trained with the backpropagation algorithm to predict absence or presence of heart disease with an accuracy of 95 %. Various machine learning algorithms were streamlined for the effective prediction of a chronic disease outbreak by Chen *et al.* [11]. The data collected for the training purpose was incomplete. To overcome this, a latent factor model was used. A new convolutional neural network-based multimodal disease risk prediction (CNN-MDRP) was structured. The algorithm reached an accuracy of around 94.8 %. Chae *et al.* [12] used 4 different deep learning models namely deep neural networks (DNN), long short term memory (LSTM), ordinary least squares (OLS), an autoregressive integrated moving average (ARIMA) for monitoring 80 infectious diseases in 6 groups. Of all the models used, DNN and LSTM models had a better performance. The DNN model performed better in terms of average performance and the LSTM model gave close predictions when occurrences were large. Haq *et al.* [13] used a database that contained information about patients having any heart disease. They extracted features using three selection algorithms which are relief, minimum redundancy, and maximum relevance (mRMR), and least absolute shrinkage and selection operator which was cross-verified by the K-fold method. The extracted features were sent to 6 different machine learning algorithms and then it was classified based on the presence or absence of heart disease. An effective heart disease prediction system was developed by Mohan *et al.* [4].

They achieved an accuracy level of 88.4 % through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM). Maniruzzaman *et al.* [14] classified the diabetes disease using ML algorithms. Logistic regression (LR) was used to identify the risk factors for diabetes disease. The overall accuracy of the ML-based system was 90.62 %.

## 3 Methodology

From an open-source dataset, an excel sheet was created where we listed down all the symptoms for the respective diseases. After which depending on the diseases, age and gender were specified as a part of the dataset. We listed down around 230 diseases with more than 1000 unique symptoms in all. The symptoms, age, and gender of an individual were used as input to various machine learning algorithms.

### 3.1 K-nearest neighbors (KNN)

The K-nearest neighbors (KNN) algorithm used is a type of supervised machine learning algorithm. It simply calculated the distance of a new data point to all other training data points. The distance can be of Euclidean or Manhattan type. After this, it selects the K-nearest data points, where K can be any integer. Lastly, it assigns the data point to the class to which the majority of K data points belong.

### 3.2 Fine, Medium, and Coarse KNN

We need to assign the integer values of K to find the distance. So, in our fine KNN model, we assigned a low value of K which means it approximately uses only one neighbor for the prediction. Similarly, the medium KNN model uses approximately 10 neighbors and the coarse KNN uses 100 neighbors. Since the neighbors for each, the model differs the accuracy percentages also varied with a wide range. Among all the three models our fine KNN gave us a very high accuracy whereas the coarse KNN resulted in a low prediction value.

### 3.3 Weighted KNN

It is a modified version of KNN. In KNN we chose an integer parameter K and by using that parameter we found where the major predicted values lied. But if the value of K is too small the algorithm is much more sensitive to the points that are outliers. Also, if the value of K is too large then all the points that are almost very close to the K value are selected. To overcome this issue the weighted KNN gave more weight to the points that were nearest to the K value and the less weight to the points that were farther away. We were able to get the highest accuracy using this model. Also among all the KNN models, this model gave us the best results.

### 3.4 Naïve Bayes

It is a machine learning algorithm for classification problems and is based on Bayes' probability theorem. The primary use of this is to do text classification which involves high dimensional training data sets. We used the Bayes theorem that can be defined as:
$P(h|d) = \frac{P(d|h) \cdot P(h)}{P(d)}$

Where $P(h|d)$ is the probability of hypothesis h given the data d. This is called the posterior probability. $P(d|h)$ is the probability of data d given that the hypothesis h was true. P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h. P(d) is the probability of the data (regardless of the hypothesis).

### 3.5 Gaussian Naïve Bayes

It follows the same procedure as the Naïve Bayes. But for Naïve Bayes we need a categorical dataset and for Gaussian Naïve Bayes we need a dataset that has all the continuous features. Our dataset consisted of continuous features of symptoms, age, and gender so it was mandatory to use this model. The accuracy using this model was not a very high value.

### 3.6 Kernel Naïve Bayes

Our dataset had some numerical attributes such as age so we implied Kernel Naïve Bayes to predict the medicines. The steps followed for this algorithm are similar to the Naïve Bayes. The major benefit of using this algorithm is that it provides estimators that have a nonparametric nature. If there is no prior knowledge that the dataset used is parametric or not this model can give more accurate results. The results given by this model were almost the same as shown by the Gaussian Naïve Bayes.

### 3.7 Decision trees

Decision trees algorithm belongs to the family of supervised learning algorithms. It is used for regression and classification. In the decision tree, for prediction, it uses the method of tree diagram at the top. It contains a root node after which it gets split in the dominant input feature and then it again gets split. These processes continue till all input is placed and at the node, the extreme last node contains the weights on the bases of these weights it classifies the input.

In a coarse tree, the maximum number of splits from each node is 4. Whereas in a Medium tree, the maximum number of splits from each node is 20. In a fine tree, the maximum number of splits from each node is 100.

### 3.8 SubSpace KNN

The SubSpace KNN method is similar to bagging except that the features are randomly sampled, with replacement. Informally, this causes individual learners to not over-focus on features that appear highly predictive/descriptive in the training set, but fail to be as predictive for points outside that set. For this reason, random subspaces are an attractive choice for problems where the number of features is much larger than the number of training points.

### 3.9 RUSBoost algorithm

Our data was required to be trained properly to get proper and good accuracy. RUSBoost algorithm is used for improving the performance of the trained data set acquired from the skewed data set. RUSBoost is a hybrid data sampling/boosting algorithm. This algorithm is one of the methods producing the fastest results amongst the hybrid boosting algorithms.

## 4 Results and Discussion

Different machine learning models were used to examine the prediction of disease for available input dataset. We used 11 different ML models for the prediction. Out of the 11 models we managed to get 50 % or above accuracy for 6 models. As shown in Figure 4, among all the models, we gained the highest accuracy for the Weighted KNN model of 93.5 %. The accuracy is high because the weighted KNN was high since in this model the value of K varied. This value changed according to our dataset i.e. it was small and large for the training set. Due to this variation, it proved to be the most accurate model as compared to the other ML algorithms. We took raw information and distinguished them on the basis of gender, age group, and symptoms. The lowest accurate model was the RUSBoosted Tree of accuracy 0.5 %. Fine tree displayed accuracy of 21.8 %. The medium tree had an accuracy of 12.3 %. The coarse tree had an accuracy of 6.4 %. Gaussian Naïve Bayes had an accuracy of 16.8 %. Kernel Naïve Bayes had an accuracy of 16.8 %. Fine KNN had an accuracy of 80.3 %. Medium KNN displayed an accuracy of 61.8 %. Coarse KNN had an accuracy of 5.3 %. Subspace KNN exhibited an accuracy of 73.2 %.

As shown in table 2, we have compared our methodology with the other methodologies reported in the literature. Some of the literature has used the SVM and the KNN model for the prediction of the diseases. We have used 11 different ML models for the prediction of around 230 diseases. We achieved the highest accuracy of 93.5 % which was high as compared to most of the other methodologies reported. The highest accuracy was achieved because of the Weighted KNN model. Khourdifi *et al.* [2] achieved the highest accuracy of 99.7 % using the KNN model for the prediction and classification of heart diseases. Sriram *et al.* [5] used the Random Forest model and achieved an accuracy of 90.26 %. The SVM model proved to have a higher accuracy of 79.13 % as compared to the other methods used by Mir *et al.* [1].

Doctors and medical professionals are always required in case of an emergency. In the current situation of COVID-19, where sufficient facilities and resources are unavailable, our prediction system can prove to be helpful and can be used in the diagnosis of a disease.

## 5 Conclusions

The manuscript presented the technique of predicting the disease based on the symptoms, age, and gender of an individual patient. The Weighted KNN model gave the highest accuracy of 93.5 % for the prediction of diseases using the above-mentioned factors. 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.

**Acknowledgements** Authors would like to thank all doctors and faculties ho helped us in completion of the work.

## Compliance with Ethical Standards

Conflicts of interest

Authors A. Khakharia, R. Keniya, R. Manjalkar, T. Thaker, V. Gada, V. Shah, M. Warang and N. Mehendale, declare that he has no conflict of interest.

Involvement of human participant and animals

This article does not contain any studies with animals or Humans performed by any of the authors. All the necessary permissions were obtained from the Institute Ethical Committee and concerned authorities.

Information about informed consent

No informed consent was required as the studies does not involve any human participant.

Funding information

No funding was involved in the present work.

## References

1. A. Mir, S.N. Dhage, in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (IEEE, 2018), pp. 1–6
2. Y. Khourdifi, M. Bahaj, Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization, Int. J. Intell. Eng. Syst. **12**(1), 242 (2019)
3. S. Vijayarani, S. Dhayanand, Liver disease prediction using svm and naïve bayes algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) **4**(4), 816 (2015)
4. S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access **7**, 81542 (2019)
5. T.V. Sriram, M.V. Rao, G.S. Narayana, D. Kaladhar, T.P.R. Vital, Intelligent parkinson disease prediction using machine learning algorithms, International Journal of Engineering and Innovative Technology (IJEIT) **3**(3), 1568 (2013)
6. A.S. Monto, S. Gravenstein, M. Elliott, M. Colopy, J. Schweinle, Clinical signs and symptoms predicting influenza infection, Archives of internal medicine **160**(21), 3243 (2000)
7. R.D.H.D.P. Sreevalli, K.P.M. Asia, Prediction of diseases using random forest classification algorithm
8. D.R. Langbehn, R.R. Brinkman, D. Falush, J.S. Paulsen, M. Hayden, an International Huntington's Disease Collaborative Group, A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length, Clinical genetics **65**(4), 267 (2004)
9. K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and structural biotechnology journal **13**, 8 (2015)
10. T. Karayılan, Ö. Kılıç, in *2017 International Conference on Computer Science and Engineering (UBMK)* (IEEE, 2017), pp. 719–723
11. M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, Disease prediction by machine learning over big data from healthcare communities, Ieee Access **5**, 8869 (2017)
12. S. Chae, S. Kwon, D. Lee, Predicting infectious disease using deep learning and big data, International journal of environmental research and public health **15**(8), 1596 (2018)
13. A.U. Haq, J.P. Li, M.H. Memon, S. Nazir, R. Sun, A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms, Mobile Information Systems **2018** (2018)
14. M. Maniruzzaman, M.J. Rahman, B. Ahammed, M.M. Abedin, Classification and prediction of diabetes disease using machine learning paradigm, Health Information Science and Systems **8**(1), 7 (2020)