Multiclass Classification in Machine Learning Algorithms for Disease Prediction

Pallavi Tiwari(✉), Deepak Upadhyay, Bhaskar Pant, and Noor Mohd

Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India
{Pallavitiwari_20061041.cse,deepakupadhyay.ece}@geu.ac.in

Abstract. This paper proposes multiclass classification using different symptoms of patients into 40 different classes. This paper also represents the comparative study of the performance of four different Machine Learning models on the test symptoms data of the patients and suggests the most effient model to classify into 40 classes. Random Forest, Support Vector Machine (SVM), Naive Bayes and Decision tree are used for building the model. The performance of the algorithms is being analyzed on the parameter like accuracy, precision, and F1-score. The results reveal that Random Forest and Decision Tree are more accurate than other machine learning algorithms.

Keywords: Machine learning · Support Vector Machine · Decision Tree · Multiclass classification

1 Introduction

Education and healthcare are two sectors where machine learning is being employed. Machine learning has become more popular as a result of advances in processing power and the availability of datasets on open-source repositories. In healthcare, machine learning is utilized in a wide range of settings. There is a lot of data in the healthcare business that can assist uncover patterns and forecast future outcomes. In healthcare, machine learning is utilized to tackle a variety of issues [1–3]. There is no one-size-fits-all approach to determining the severity of cardiac disease [4]. Machine learning models may be built using the dataset and individual patient data to predict outcomes. As a consequence of the data entered, the forecast result will be unique to that individual. Type-2 diabetes can be avoid-ed by controlling one's weight, diet, and other lifestyle factors [5]. There is no specific therapy for coronavirus. This year's coronavirus came from China. This disease is being treated with a variety of methods, but there aren't any clearly defined measures to follow. Human cognition is the goal of artificial intelligence (AI). A paradigm change in healthcare is being brought about by the rising availability of healthcare data and the rapid advancement of analytics tools [1]. In recent years, several models for automated detection of illnesses such as cancer, COVID-19, and diabetes [2] have been established as of late, researchers have been building smartphone applications that use machine learning models to diagnose diseases in real time. It's even possible to

A. K. Luhach et al. (Eds.): ICAICR 2021, CCIS 1575, pp. 102–111, 2022.
https://doi.org/10.1007/978-3-031-09469-9_9

create smartphone applications that can assess a person's likelihood of contracting an illness and then suggest a diagnosis based on their current health status [6]. In spite of this, early diagnosis remains an ill- posed challenge Many academics have recently begun employing deep-learning models to get much better results than machine learning models [2, 3]. Using machine learning algorithms, this study predicts an individual's risk of coronavirus, heart disease, and type 2 diabetes. People are required to enter their personal in- formation into a mobile application and then submit the information. The danger is forecasted within a few seconds of real-time analysis. Firebase is a cloud-based mobile application that serves as a real-time database. A database stores the model's training parameters, allowing for real-time prediction. Furthermore, the user is informed of the model's accuracy. An additional feature is real-time sharing of news articles from reputable sources. The app also includes a link to the source of the news. One of the most pressing concerns of human civilization is healthcare, which affects the quality of life for everyone.

It is a key component of the system [7]. The healthcare industry, on the other hand, is incredibly diverse, widely distributed, and disjointed. Providing optimal medical care necessitates access to relevant patient information, which is rarely available when it is needed [8]. In addition, the wide variation in the order of diagnostic tests suggests the necessity of a sufficient and appropriate collection of tests [14] expanded this claim by suggesting that the significant differences found in the request for general practice pathology arise primarily from individual variations in clinical practice and are therefore likely to improve through more transparent and better-informed decision-making for physicians [8]. Many heterogeneous factors, such as demographics, medical history, drug allergies, biomarkers or genetic markers may be found in medical data. Each one provides a distinct perspective on the patient's state. In addition, as previously indicated, statistical features varied significantly among the sources. When evaluating such data, researchers and practitioners confront two major challenges: the curse of dimensionality (the number of dimensions as well as the number of samples rises exponentially in the space of features) and the heterogeneity of function sources and statistical attributes. As a result of these factors, individuals have been unable to receive the care they need due to delays and errors in their disease diagnosis. This necessitates the development of methods that can diagnose the disease in its earliest stages as well as assist physicians in determining treatment decisions [14]. To deal with all of this information in the medical, computer, and statistical domains, it is imperative that new methodologies be developed to model illness prediction and diagnosis [12]. Modern machine learning (ML) provides a wealth of useful tools for analysing large amounts of data. Because of this, its technology may already be used to analyse medical data. Furthermore, a wide range of medical diagnostic work has been done on small-specialized diagnostic issues [9] where the early ML applications have been identified. Stable people and those with Parkinson's disease may be distinguished using ML classifiers, making it an important tool in clinical diagnosis.

2 Methodology

The Methodology diagram as shown in Fig. 1 consists of a dataset which is the Disease Symptom Prediction dataset. It involves the next step of data preprocessing, following

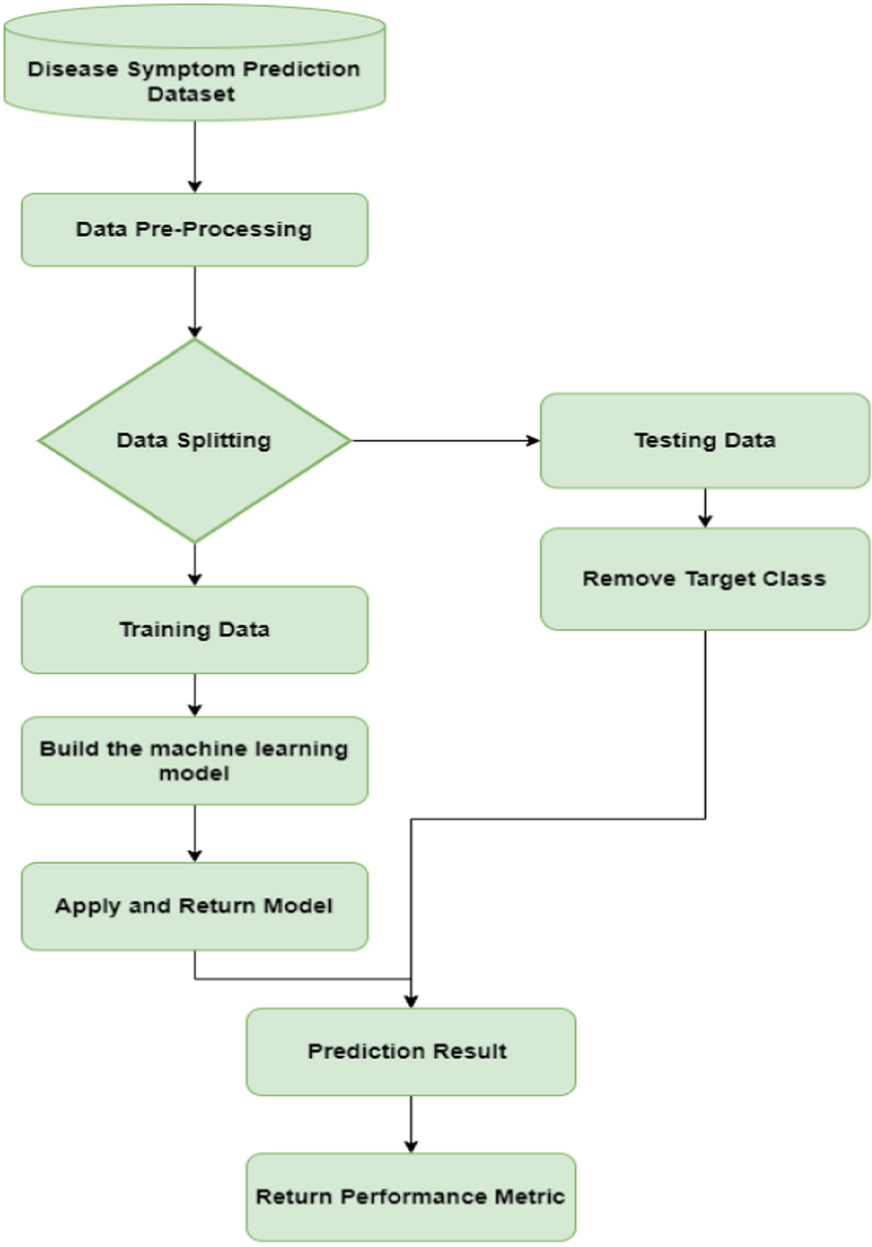

Fig. 1. Methodology diagram

this step the preprocessed dataset is then divided into two parts which are training data and testing data. In the case of training data, the next step is to build different machine learning models and then apply the training data to train the data from these models and

return to the step of the prediction result. On the other hand, from testing data, the next step is to remove the target class and test the output and return to the prediction result. From prediction results, this paper analyzes the performance of each model. Discussion of these steps is done in the following steps.

2.1 Data Collection

The Disease Symptom Prediction Dataset has been used which was available online on the Kaggle. The sample image of the dataset used is shown in Tables 1 and 2.

So, there are 4920 instances in total, containing 40 unique disease and 17 symptoms columns that have been used to predict the disease in this research work.

The dataset used has missing values too, as some of the diseases can be detected using few symptoms and some require a greater number of symptoms to be accurately detected. These missing values are handled through data preprocessing using python.

Table 1. Dataset 1 sample: predicting disease.

	Disease	Symptom_1	…	Symptom_16	Symptom_17
0	Fungal infection	itching	…	NaN	NaN
1	Fungal infection	skin_rash	…	NaN	NaN
2	Fungal infection	itching	…	NaN	NaN
3	Fungal infection	itching	…	NaN	NaN
4	Fungal infection	itching	…	NaN	NaN

Table 2. Dataset 2 sample: symptom severity

	Symptom	Weight
0	itching	1
1	skin_rash	3
2	nodal_skin_eruptions	4
3	continuous_sneezing	4
4	shivering	5

2.2 Data Preprocessing

The missing data in this case comes under MNAR (Missing Not At Random), this means when data are missing, not at random, the missingness is specifically related to what is missing, e.g. a person does not attend a drug test because the person took drugs the night before [16]. In this case, symptoms are missing because there exists only that number

of symptoms for that disease. These are handled carefully in this model using python. Table 3 shows the missing values in the Dataset 1.

With the help of dataset 1 and dataset 2, preprocessing of data through python library is being used and final resultant dataset Table 4 which is clean and preprocessed is used in building the model.

The resultant dataset has been then splitted into training and testing data in where mostly about 80% considered to be training data and remaining 20% to be testing data. For dividing the dataset python inbuild library scikit-learn and function train_test_split() is used.

Table 3. Sum of missing values of each attribute

Disease	0
Symptom_1	0
Symptom_2	0
Symptom_3	0
Symptom_4	348
Symptom_5	1206
Symptom_6	1986
Symptom_7	2652
Symptom_8	2976
Symptom_9	3228
Symptom_10	3408
Symptom_11	3726
Symptom_12	4176
Symptom_13	4416
Symptom_14	4614
Symptom_15	4680
Symptom_16	4728
Symptom_17	4848

Table 4. Dataset after preprocessing

	Symptom_1	Symptom_2	Symptom_3	…	Symptom_17	Disease
0	1	3	4	…	0	Fungal infection

(*continued*)

Table 4. (*continued*)

	Symptom_1	Symptom_2	Symptom_3	…	Symptom_17	Disease
1	3	4	0	…	0	Fungal infection
2	1	4	0	…	0	Fungal infection
3	1	3	0	…	0	Fungal infection
4	1	3	4	…	0	Fungal infection
…	…	…	…	…	…	…
4915	5	3	5	…	4	(vertigo) Paroymsal Positional Verigo
4916	3	2	2	…	0	Acne
4917	6	4	0	…	0	Urinary tract infection
4918	3	3	3	…	2	Psoriasis
4919	3	7	4	…	3	Impetigo

2.3 Building the Model

The model is built using Python.Scikit-learn library to implement the four machine learning algorithm. Scikit-learn is a Python module that integrates a wide range of cutting-edge machine learning methods for supervised and unsupervised issues on a medium-scale[18]. The library imported to build the models are as follows:

1. **from sklearn.naive_bayes import GaussianNB:**
   This Library is used to build the Gaussian Naïve Bayes model and to implement and train this model GaussianNB().fit(x_train, y_train) function is used.
2. **from sklearn import svm:**
   This Library is used to build the Support Vector machine model and to implement and train this model svm.SVC(kernel = 'rbf', gamma = 0.5, C = 0.1).fit(x_train, y_train) function is used.
3. **from sklearn.ensemble import RandomForestClassifier:**
   This Library is used to build the Random Forest model and to implement and train this model RandomForestClassifier().fit(x_train, y_train) function is used.
4. **from sklearn.tree import DecisionTreeClassifier:**
   This Library is used to build the Support Vector machine model and to implement and train this model DecisionTreeClassifier().fit(x_train, y_train) function is used.

2.4 Performance Metric of Model

Following performance metric of the four models have been considered for the analysis and comparison. Table 5 shows the comparison and analysis of the four machine learning algorithms in the tabular form.

1. **Accuracy**
   Accuracy of model is defined as total prediction that are predicted correctly divided by the total predictions done by the model.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\textit{True Positive} + \textit{True Negative} + \textit{False Positive} + \textit{False negative}}$$

2. **Precision**
   Precision is defined as positive predictions that are predicted correctly divided by the total positive prediction.

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

3. **Recall**
   Recall is defined as how many of the returned predictions that predicted that it belongs to certain class were actually predicted that they belong to that class.

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative}}$$

4. **F1-Score**
   F1-Score is defined as the harmonic mean of Precision and Recall.

$$\text{F1-Score} = 2 * (\frac{\textit{Precision} * \textit{Recall}}{\textit{Precision} + \textit{Recall}})$$

The performance metric of the models are calculated using the library sklearn.metrics. Since multiclass classification is where there exist more than two classes and they are solved by further diving the problem into series of binary classes. And since the many binary classes so will be that many values of precision, recall and f1_score and here comes the method of averaging which will output one precision value of that model. There are three types of averaging that can be done macro, micro and weighted. In this paper the metric is calculated using the average type "weighted" as the classes are imbalanced and it is best technique to used it for.

3 Result Analysis as Per Given Algorithms

In this paper, four machine learning models are used, SVM, Random Forest, Decision Tree and Gaussian Naïve Bayes for the prediction of disease based on the symptoms using disease symptom prediction, Kaggle Dataset. The performance of all models was evaluated based on four parameters, accuracy, precision, F1_score and recall. The performance result of the models is shown in Table 5 and Fig. 2 respectively.

Table 5. Performance metric of models

Model	Accuracy	Recall	Precision	F1_score
Multiclass SVM	0.92	0.99	0.99	0.92
Random Forest	0.99	0.99	0.99	0.99
Decision Tree	0.99	0.99	0.99	0.99
Naïve Bayes	0.86	0.86	0.87	0.85

Decision tree and Random Forest performed well in comparison to Gaussian Naive Bayes and SVM. Model. Most accurate algorithm is Decision tree and Random Forest with accuracy 99%, followed by SVM with 92% and then Naive Bayes that has the least accuracy of 86%. When it comes to Precision all algorithm has Precision 99% except Naïve bayes which has Precision 87%.

This study demonstrates how ML Predictive models can be created, verified, and used to diagnose various diseases quickly. The study also demonstrates the critical significance of supervised machine learning algorithms in the prediction and diagnosis of diseases, which can help alleviate the enormous load on healthcare systems in most countries throughout the world.

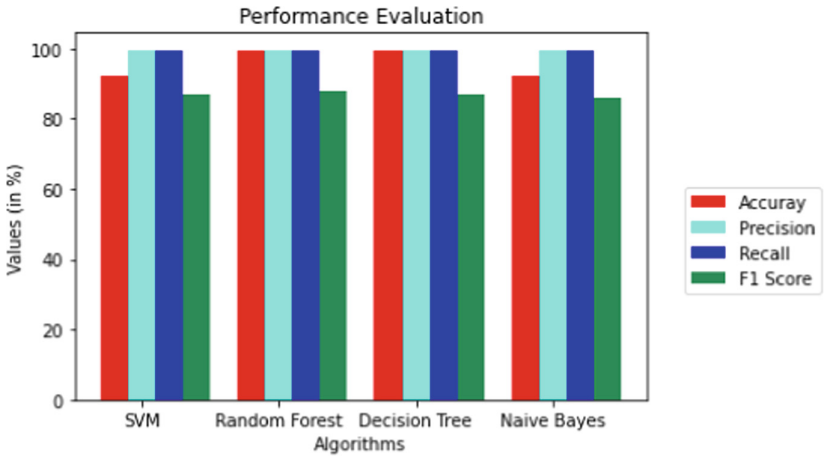

Fig. 2. Performance evaluation of algorithms

4 Conclusion

In this paper attempt has been made to analyze and compare various machine learning models based on multiclass classification of various diseases based on their symptoms. The goal of this study was to see how well algorithms perform when dealing with

multiclass data. For disease symptom prediction dataset used from the Kaggle, which comprises 4920 instances, and used a test_train split to divide the data into two halves, training and testing datasets. To test the performance, 17 different symptoms are analyzed to predict various diseases using four different algorithms. Finally, after the implementation phase, it was determined that Random Forest and Decision Tree provide the highest level of accuracy in the dataset used, at 99%, while Naive Bayes provides the lowest level of accuracy, at 86%.

References

1. Ali, F., et al.: A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. J. Inf. Fusion **63**, 208–222 (2020)
2. Bakator, M., Radosav, D.: Deep learning and medical diagnosis: a review of literature. J. Multimodal Technol. Interact. **2**, 47 (2018)
3. Ebrahimighahnavieh, M.A., Luo, S., Chiong, R.: Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. J. Comput. Methods Progr. Biomed. **187**, 105242 (2020)
4. Goel, S., Deep, A., Srivastava, S., Tripathi, A.: Comparative analysis of various techniques for heart disease prediction. In: 4th International Conference on Information Systems and Computer Networks, pp. 88–94. IEEE, Mathura (2019)
5. Hong, S., Zhou, Y., Shang, J., Xiao, C., Sun, J.: Opportunities and challenges of deep learning methods for electrocardiogram data: a systematic review. J. Comput. Biol. Med. **122**, 103801 (2020)
6. Hossain, M.E., Khan, A., Moni, M.A., Uddin, S.: Use of electronic health data for disease prediction: a comprehensive literature review. In. IEEE/ACM Trans. Comput. Biol. Bioinform. **18**(2), 745–758 (2021). IEEE
7. Ibrahim, I., Abdulazeez, A.: The role of machine learning algorithms for diagnosing diseases. J. Appl. Sci. Technol. Trends **2**(01), 10–19 (2021)
8. Jo, T., Nho, K., Saykin, A.J.: Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. J. Front. Aging Neurosci. **11**, 220 (2019)
9. Khalid, H., et al.: A comparative systematic literature review on knee bone reports from MRI, x-rays and CT scans using deep learning and machine learning methodologies. J. Diagn. **10**, 518 (2020)
10. Liu, X., et al.: A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. J. Lancet Digit. Health **1**(6), e271–e297 (2019)
11. Nagaraju, M., Chawla, P.: Systematic review of deep learning techniques in plant disease detection. Int. J. Syst. Assur. Eng. Manag. J. PeerJ. Comput. Sci. **7**, e432 (2020)
12. Shafaf, N., Malek, H.: Applications of machine learning approaches in emergency medicine; a review article. J. Arch. Acad. Emerg. Med. **7**(1), 34 (2019)
13. Mohd, N., Singh, A., Bhadauria, H.S.: A novel SVM based IDS for distributed denial of sleep strike in wireless sensor networks. Wirel. Pers. Commun. **111**(3), 1999–2022 (2019). https://doi.org/10.1007/s11277-019-06969-9
14. Solares, J.R.A., et al.: Deep learning for electronic health records: a comparative review of multiple deep neural architectures. J. Biomed. Inform. **101**, 103337 (2020)
15. Kumar, I., Mohd, N., Bhatt, C., Sharma, S.K.: Development of IDS using supervised machine learning. In: Pant, M., Kumar Sharma, T., Arya, R., Sahana, B.C., Zolfagharinia, H. (eds.) Soft Computing: Theories and Applications. AISC, vol. 1154, pp. 565–577. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-4032-5_52

16. All About Missing Data Handling. https://towardsdatascience.com/
17. Mohd, N., Singh, A., Bhadauria, H.S.: Intrusion detection system based on hybrid hierarchical classifiers. Wirel. Pers. Commun. **121**(1), 659–686 (2021). https://doi.org/10.1007/s11277-021-08655-1
18. Scikit-learn: Machine Learning in Python. http://jmlr.org/
19. Sharma, V., Yadav, S., Gupta, M.: Heart disease prediction using machine learning techniques. In: 2nd International Conference on Advances in Computing, Communication Control and Networking, pp. 177–181 (2020)