RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 03 March
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.017

Multi Disease Classification System Based on Symptoms using The Blended Approach

Swathi Buragadda[1], V P Siva Kalyani Pendum [2], Dulla Krishna Kavya[2], Shaik Shaheda Khanam[2]

[1]Sr.Assistant Professor, Department of Computer Science and Engineering,, Lakireddy Balireddy College of Engineering, Affiliated to JNTUK, Kakinada, Mylavaram, India

[2]Department of Computer Science and Engineering, Lakireddy Balireddy College of Engineering, Affiliated to JNTUK, Kakinada, Mylavaram, India

Email: buragaddaswathi@gmail.com

Article History

Received: 3 February 2023
Accepted: 14 March 2023

Keywords:

Blending Model;
Embedded Approach;
Optimizers;
Saturation Points;
Bagging and Boosting

Abstract

In today's world, everyone is preoccupied with work and other activities, leaving little time to visit doctors about illnesses that may appear to be minor at first but develop into life-threatening conditions as time passes. As a result, the proposed model accesses a public repository that maintains numerous symptoms and their possible diseases as a matrix for early disease prediction and prevention. Symptoms are received from the user and fed into the embedded blending algorithm to estimate the type of disease. The patient's records are collected from the several hospitals and the resulting massive volume of data, which results in inefficient prediction model using the machine learning approaches. Since the proposed model is a combined approach of training mechanism, it can reduce the number of accessing records in every step. Traditional approaches like bagging and boosting construct more number of decision trees because of the vast amount of data. This results in the utilization of more number of resources and sometimes CPU enters into saturation state. The proposed system solves this problem by using optimized parameters for tree construction and reduces the memory and resource utilizations.

1. INTRODUCTION:

Machine Learning is defined as a subset of artificial intelligence that is primarily concerned with the development of algorithms that allow a computer to learn on its own from data and previous experiences (Silpa et al.). When a machine learning system receives new data, it predicts the outcome using the prediction models it has built using historical data. The amount of data influences the accuracy of predicted output because a massive volume of data helps to develop a more accurate model that forecasts the output more accurately. The algorithm for machine learning is explained in the figure 1.

1.1. Blending Model

Ensemble learning algorithms are used for better performances in machine learning in this blending is one of the models (Tran et al.) . Blending is the extension of the stacking method. This model combines thousands of prediction models which are described with stacking models. Here the meta-models are fitted using multiple predictions with a holdout validate datasets (Loddo, Buttau, and Ruberto). To predict the models the functions need to be included for training related models and performing predictions for generating new data. Logistic and linear regression techniques use the blending

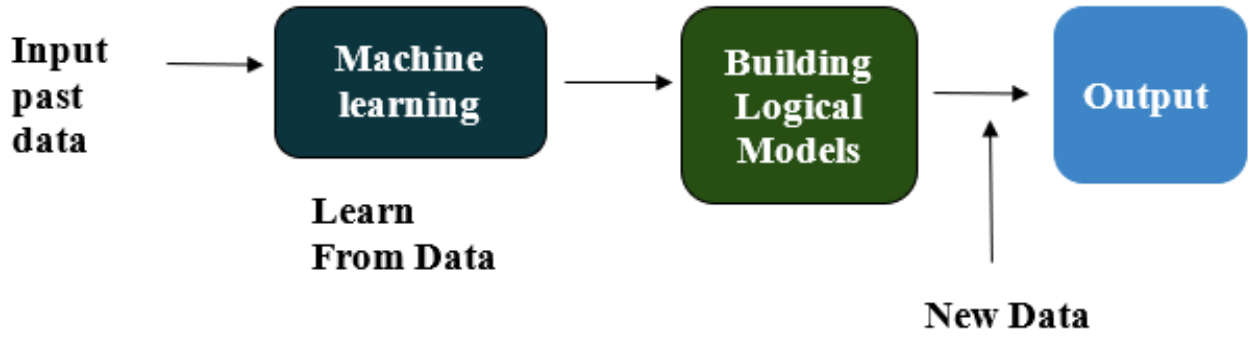

FIGURE 1. Block Diagram of Machine Learning

models mainly for the prediction of true or false values or ones or zero. The process generally creates a base mode that can be anything that can classify or regression issue-solving techniques. Now recall is performed for fitting the training dataset to the base model. The meta-model is fitted in predictions generated by every base model in the holdout dataset. Initially generate the list of models & fit them in the training datasets. These provide the predictions to provide the inputs for the blending methods for the final output. Hence this method is mainly for increasing accuracy.

2. LITERATURE SURVEY:

Many methods has good efficiency to deal with single disease identification and predict them Akkem Yaganteeswarudu ("2020 5th International Conference on Communication and Electronics Systems (ICCES)") introduced multi identification decease technique using python library files and machine learning techniques. In python the flask API has tensflow is used and in machine learning the five method is used for better performances. In python the picking method is used for saving the model attention. The responsibility of the method is to mortality ratio decreased in prediction of diseases based on the patient conditions. The dataset contains diabetes data was collected from the Pima Indian in Germany hospital. In the ML method considered a diabetes images related to retina are used. For this tensorflow CNN is used for designing the model for test and train sets. The process starts with preprocessing, selecting the models, behaviour is identified using picking model in python, designed the Flask API. Now decide the disease type where a person contain disease will find if patient is diabetes or not. If yes pickle files are loaded if not identify the retino diseased person. Follow this process for heart and cancer patients identification. Now if a person is not diseased then decide the patient is healthy feature and check the corresponding python file and return. Finally, the process end with asking user about the other health related to predict. Retrieving multiple diseases for every single patient can reduce the cost of analysis. Hence the proposed method has achieved 91% of accuracy.

K. Arumugam et al (Arumugam et al.) wanted to predict the multiple diseases related to a single patient using machine learning techniques. So, three different are chosen SVM, Naïve Bayes, and DT among these DT has achieved high efficiency and low error rate. The author has chosen diabetes patients affected with heart attacks. As discussed the three approaches has been used let's know about naïve bayes is dependent upon the conditions where if the certain conditions works then only the next process is followed. Independent variable is the assumption where as every variable need to be computed. In the SVM approach achieves the issues in huge prediction of kernel learning. The performances of the data is high at linear and vice versa process with good scaling process compared to many other classifiers. Margining the +ve and -ve error rate are main classification in this method. The kernel function is used for non-linear techniques with five different methods for mapping huge dimensional space to discart non-separable issues. The DT is a tree structure data holds root and leaf nodes which are connected with links or also called as branch. Based on conditions the tree is designed. The DT holds most complex method the C4.5 for mining which is related to profit ratio. Based on

three concepts the method is declared as the best i.e. missing values handling, less memory, features continuation, and functioning. Hence the DT method has achieved 90% of accuracy.

Indukuri Mohit et al (Mohit et al.) consider the issue related to patient disease which is neglected because of frequent check-ups. The author works on heart cancer patients having diabetes in their report too. Here the ML methods are used for the prediction of disease from patient reports and their previous addresses to the hospitals. Three approaches have been used i.e. LR, SVM & KNN these approaches also include three different disease detections. The team of authors has developed a web-based testing app for verification of the approach. The collection of data was from three different areas heart, Pima Indians, & Wisconsin considered from the Kaggle & UCI. Now the data need to be cleaned, removed, etc. in the pre-processing stage which includes different methods in it. Then data need to be partitioned for training and testing purposes which are applied to predicting the diseases using the ML methods. Each method holds the k-values for prediction and calculates the distance metrics. Finally, the data collects the best accuracy from the metrics. Hence the LR has proven the best accuracy in every dataset with 94.55%.

Identifying different patients' conditions based on the input provided by them is the main responsibility chosen by Anuj Kumar et al (Kumar and Pathak) using machine learning approaches. Four kinds of methods that are related to supervised learning techniques are used for the prediction of diseased content. The responsibility regarding the approach has provided timely diagnoses based on the condition of the patient. Here the author considered five symptoms related to the patient and those are applied to the ML methods for predicting the stage or condition. The process of the proposed method is by collecting data from different sites. The complete dataset is in binary format so, it's easy to identify by the system. This data is divided randomly and split into train and test sets for further processing. As known four methods are performed on the trained data and then send the learned data to a model. This model is a combination of the ML methods and tested models. From these metrics, the performances need to be evaluated and the diseases are identified based on patients' symptoms. The performance of the method is high in Naïve Bayes with 95.21% of accuracy. Even though the remaining methods also achieved high performances with only a change of .10% of differences.

A patient may have multiple diseases with different conditions so, Rudra A. Godse et al (Godse et al.) has collected the majority of the information from the related doctors. The author has built websites that are user-friendly and can be used for any kind of situation for a patient. To design this method python and ML methods have been chosen. The structure of the approach is the collection of data from different hospitals and then pre-processing them accordingly, for training and test purpose. This data will be processed for ML approaches and create the best model from those including the tested dataset. For every section, the tested data is pre-processed because of new user details. As known an app is built the infrastructure of the process is the user & application where initially login to the application and enter the symptoms related to the user. Based on the given data the database gets activated and performs the matching of the symptoms and returns it to the screen. If any other different disease is identified then a searching process is seen in the ML model and returns the probability of diseases through the list format. In addition to this, a user can answer any of the sub-queries. Hence the performances of the data are good and efficient.

3. PROPOSED METHODOLOGY:

The proposed model uses the blending model for detection of diseases based on the symptoms. Blending is the extension of the stacking which uses the training data for the classifier and sends the output variables for the classifier and results are obtained. Here there are two levels ground level that is considered as zero and the followed layers are considered as the 1,2,3.....etc. The data is used in multiple ways financial, emotion recognition, and security in computers. For decision making in financial sector for predicting the failures related to the business cruciality. It has good ability to realize the out stocks for marketing manipulations. Second, many industries are dependent on the reviews and performances of their company search. This is the best recognition for the speech-based content. Third, security is the major issue for every one who is using social networks. In this, three methods

TABLE 1. Analysis on Existing Approaches

Author	Algorithm	Merits	Demerits	
Akkem Yagan-teeswarudu	Tensorflow, ML methods	Identifies multiple diseased person condition in low cost.	May be NLP can be used for data reading.	91%
K. Arumugam et al	Naïve Bayes, SVM, Decision Tree.	DT is considered as the best method	Validation technique can be used for better results.	90%
Indukuri Mohit et al	LR, SVM, KNN	Data partition has achieved good performances	An additional pre-propagation technique needs to be initiated.	94.55%
Anuj Kumar et al	DT & RF	Creating a model for combining test & trained data is efficient.	The methods need to be improved by unsupervised technique.	95.21%
Rudra A. Godse et al	KNN, DT, SVM, Naïve Bayes.	Generating medication is the best part.	Always the system need to be trained for every report.	97%

are present malware detection which can classify with codes related to the ML methods. Provides the best results based on the blended ensemble models. Second, intrusion-based detection will help them by reducing the errors with low-rate successions. Third, Denial services main issue in the security of many companies by using blending method can combine the single classifier which can decrease the error completely by detecting and discrimination in every attack. The proposed model uses CAT boosting and Logistic Regression algorithm in blending process.

3.1. Logistic Regression

Logistic Regression is one of the best machine learning methods which is implemented based on the supervised learning approach. Based on the independent variables a set of dependents is obtained with the predictions. The values are retrieved with zero's or one's format these values are known as probability. The differences between logistic and linear are similar. This mainly solves the regression issues but not the classification. Whereas linear only solves the classification issues. Here the graph is shown in the form of "S" and only predicts two values. The values are designed based on weights. And if the condition is the major set of connecting the variables. Here it uses sequences and discrete-valued datasets. The predicted models are converted as regression in these techniques, it is generally used for sampling. Sigmoid is the function used in the LR which is used for the mapping of the predicted variable to probability. The threshold values are used related to 0 and 1.

3.2. CatBoost Architecture:

Catboosting is part of the gradient boosting technique responsible for dealing with machine learning methods. The method can automatically handle the features according to the categories. This can many resolves the issue related to the classifications and regressions. This method was recently introduced by Russia in the engineering department. This technique is famous for designing strong learning values from weak variables or learners. Catboostind is used in ranking, recommended systems, etc. It was designed based on the tree structure which was mainly used in decision trees. One hot encoding is one of the handling techniques which is used for categorical data. Based on the targets the data is categorized and tackled with the group of features. To calculate the target statistic can handle four features hold out, greedy, ordered, and leave one. This method mainly uses the ordered target values. This method can be used in two coding techniques Python and R. Parameter tuning is applied to the usage of already available parameters for reducing time.

4. RESULTS & DISCUSSION:

Any machine learning algorithm performance is computed based on the following metrics

Precision is an actual prediction which is divided based on the total predictions created by the method. Suppose the data has predicted seven persons related heart disease out of 10 but, with prediction, it will

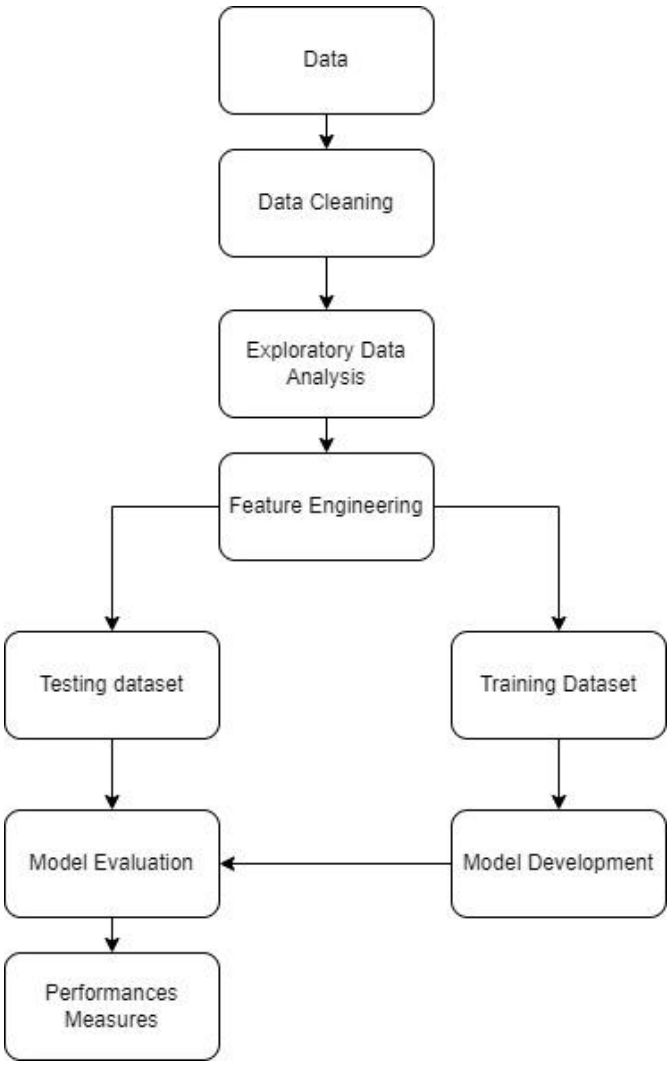

FIGURE 2. Block Diagram for Proposed Model

only acquire only three this value will be divided by total predictions, and output is derived. The recall is dependent upon the true-false values and positive negatives. The calculation is treated based on the total number of acquired true positives divided by the values depending on the total tp and fn. Accuracy is defined by the complete number of correctly classified values which is divided by the complete number of classifications. Compared to fp, fn only fp is most required to address. It is just used for the classification of values based on if or not. Here it calculates the tpr and fpr values. F1-Score is a weight combination of precision and recall. It considers two types of values fp and fn. This method is higher than accuracy which is not class is not distributed unproperly.

Figure 3 represents the transformed values of the categorical elements in the dataset. During the data cleaning process, the proposed model transforms all the symptoms into the numerical values to quicken the prediction approach.

	Symptom_1	Symptom_2	Symptom_3
0	itching	skin_rash	nodal_skin_eruptions
1	skin_rash	nodal_skin_eruptions	dischromic _patches
2	itching	nodal_skin_eruptions	dischromic _patches
3	itching	skin_rash	dischromic _patches
4	itching	skin_rash	nodal_skin_eruptions
...	...	...	...
4915	vomiting	headache	nausea
4916	skin_rash	pus_filled_pimples	blackheads
4917	burning_micturition	bladder_discomfort	foul_smell_of urine
4918	skin_rash	joint_pain	skin_peeling
4919	skin_rash	high_fever	blister

DATA TRANSFORMATION

	Symptom_1	Symptom_2	Symptom_3	S
0	33	35	35	
1	24	27	18	
2	33	27	18	
3	33	35	18	
4	33	35	35	

FIGURE 3. Data Transformation of Symptoms using Pre-processing Approaches

```
Learning rate set to 0.5
0:      learn: 3.2554198        test: 3.2058839 best: 3.2058839 (0)     total: 3.96s    remaining: 35.7s
1:      learn: 2.4686270        test: 2.4715254 best: 2.4715254 (1)     total: 7.79s    remaining: 31.2s
2:      learn: 2.3620690        test: 2.3630787 best: 2.3630787 (2)     total: 11.6s    remaining: 27.2s
3:      learn: 1.7281014        test: 1.7052492 best: 1.7052492 (3)     total: 15.5s    remaining: 23.3s
4:      learn: 1.0870799        test: 1.0056113 best: 1.0056113 (4)     total: 19.3s    remaining: 19.3s
5:      learn: 0.7326263        test: 0.6157278 best: 0.6157278 (5)     total: 23.1s    remaining: 15.4s
6:      learn: 0.5495497        test: 0.4461493 best: 0.4461493 (6)     total: 26.8s    remaining: 11.5s
7:      learn: 0.4318814        test: 0.3285921 best: 0.3285921 (7)     total: 30.7s    remaining: 7.67s
8:      learn: 0.3865919        test: 0.2929232 best: 0.2929232 (8)     total: 34.5s    remaining: 3.83s
9:      learn: 0.3168940        test: 0.2175025 best: 0.2175025 (9)     total: 38.3s    remaining: 0us

bestTest = 0.2175024899
bestIteration = 9
```

FIGURE 4. Learning Process of CAT Boost Algorithm

Figure 4 represents the feature engineering process on the symptoms using the CAT boost algorithm, which presented that only symptoms are sufficient out of 37. This process helps the memory to save the data with less utilization.

The proposed model uses the multi classification dataset with nearly 42 diseases representing the class labels. Figure 5 represents the one versus one matrix representation of all the possible class labels.

The model to prove its efficiency in terms of accuracy, it compared the proposed output with the models studied in the literature survey single approaches in traditional have almost obtained nearly 90% but the proposed has achieved 98.01% while the ensemble approaches lies between in 95 to 97 %. In the figure 6 X-axis represents the approaches while y-axis denotes accuracy metric.

```
[[13  0  0 ...  0  0  0]
 [ 0 22  0 ...  0  0  0]
 [ 0  1  0 ...  0  0  0]
 ...
 [ 0  0  0 ... 20  0  0]
 [ 0  0  0 ...  2 26  0]
 [ 0  0  0 ...  0  0 15]]
```

FIGURE 5. Confusion Matrix of Blended Model

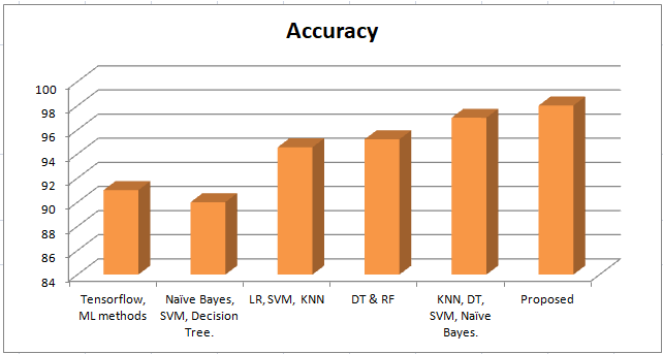

FIGURE 6. Analysis over Accuracy

5. CONCLUSION :

The proposed model has taken the dataset with 42 diseases ranging from normal to severe with maximum of 17 symptoms to identify the disease. In this paper, the model has implemented the blending model in which one algorithm is categorical boosting because all the attributes in the dataset are categorical in nature. Using this algorithm, the model learns the features that are similar with other diseases. This constructs an augmented dataset; this is passed as input to the logistic regression with Gaussian distribution because it has the ability to find the probability of each symptom in finding the disease accurately. When compared to ensemble approach the accuracy of the model is improved by +1.2%. In future work, the model is extended to neural networks with attention model because medical factors are more associated with time series transformations.

References

"2020 5th International Conference on Communication and Electronics Systems (ICCES)". *Proceedings of the 5th International Conference on Communication and Electronics Systems* 2020 (2020): 10–12. 10.1109/icces48766.2020.

Arumugam, K, et al. "Multiple disease prediction using Machine learning algorithms". *Materials Today: Proceedings* (2021). 10.1016/j.matpr.2021.07.361.

Godse, R A, et al. "Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively". *International Journal of Advanced Research in Computer and Communication Engineering* 9 (2020). 10.17148/IJARCCE.2020.9423.

Kumar, A and M A Pathak. "A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Lives". *In Turkish Journal of Computer and Mathematics Education* 4013.6 (2021).

Loddo, Andrea, Sara Buttau, and Cecilia Di Ruberto. "Deep learning based pipelines for Alzheimer's disease diagnosis: A comparative study and a novel deep-ensemble method". *Computers in Biology and Medicine* 141 (2022): 105032–105032. 10.1016/j.compbiomed.2021.105032..

Mohit, Indukuri, et al. "An Approach to detect multiple diseases using machine learning algorithm". *Journal of Physics: Conference Series* 2089.1 (2021): 012009–012009. 10.1088/1742-6596/2089/1/012009.

Silpa, P Sri, et al. "Designing of Augmented Breast Cancer Data using Enhanced Firefly Algorithm". *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)* (2022): 759–767. 10.1109/ICOSEC54921.2022.9951883..

Tran, Samer Nam K, et al. "Evolving Applications of Artificial Intelligence and Machine Learning in Infectious Diseases Testing". *Clinical Chemistry* 68.1 (2022): 125–133. 10.1093/clinchem/hvab239.

Embargo period: The article has no embargo period.

To cite this Article: Buragadda, Swathi, V P Siva Kalyani Pendum , Dulla Krishna Kavya, and Shaik Shaheda Khanam. "**Multi Disease Classification System Based on Symptoms using The Blended Approach**." International Research Journal on Advanced Science Hub 05.03 March (2023): 84–90. http://dx.doi.org/10.47392/irjash.2023.017