Journal of Science and Technology
ISSN: 2456-5660 Volume 6, Issue 4, July-August 2021

www.jst.org.in *DOI: https://doi.org/10.46243/jst.2021.v6.i4.pp25-31*

Classification of Various Diseases Using Machine Learning And Deep Learning Algorithms

Mohammed Rauf Ali Khan[1], Mohammed Muhib[1], Mir Mustafa Ali Khan[1], Mirza Musthafa Baig[2]

[1]Student, Department of Computer Science And Engineering, Deccan College Of Engineering And Technology, India.

[2]Assistant Professor, Department of Computer Science And Engineering, Deccan College Of Engineering And Technology, India.

Corresponding Author: raufkhan342@gmail.com

To Cite this Article

Mohammed Rauf Ali Khan, Mohammed Muhib, Mir Mustafa Ali Khan, Mirza Musthafa Baig, "Classification Of Various Diseases Using Machine And Deep Learning Algorithms", Journal of Science and Technology, Vol. 06, Issue 04,July-August 2021, pp34-40.

Article Info

Received: 16-02-2021 Revised: 24-05-2021 Accepted: 04-06-2021 Published: 07-06-2021

Abstract: *The latest trending market has many medical helps available, but there is inadequacy of an application or a website where we can have several machine learning paradigms implemented to predict diseases.*

This approach will have your disease predicted on based of several existing datasets using many of the machine learning algorithms. It can further also be improved with additional of speech modules. The datasets can be .csv or .xlsx or database files. It has a symptom inputting module where the user can enter the information of how he is suffering. The input is parsed and on basis of the keywords found, another panel related to those keywords will appear and take the health update in a more precise format. After submitting it, the disease is predicted and a possibility will be recommended. The project will be a combination of lung pneumonia detection system, chronic heart disease detection, diabetes risk prediction, lung pneumonia, brain tumor, malaria and other detections in a detailed manner to use more parameters thereby increasing the accuracy. Various ML algorithms like Convolutional Neural Networks, Random Forest Classification, Decision Tree and Support Vector Machines, SVM have been used to generate highest possible accuracy. CNN was used to classify Chest X-Ray images and gave 97.03% of accuracy. The pre-existing VGG-16 was used by add-up of the brain tumor prediction dataset and it was combined with the Canny Edge Detection Algorithm to generate an accuracy of 96.32%. Later, a hybrid ML algorithm was designed to classify heart and diabetic risk. It was developed as a Stacking Hybrid Classifier that has SVM on Level-0 and RFC on Level-1 of the stack. It gave cross-validated (boosting) accuracies after a10-fold CV as 91.66% for diabetes risk prediction and around 100% for heart risk prediction.

Keywords*: Disease Classification, Machine Learning, Deep Learning, VGG-16, CNN, Stacking Classifier, Support Vector Machines, Random Forest Classification.*

I. Introduction

In the times of on-going pandemic, the need of medical expertise has seen a great demand. Also, due to the trend in digitalization and automation, the advancements of artificial intelligence and machine learning, people and medical experts prefer the use of expert systems. These systems reduce the work load of medical workers in a period of such high-demand. Disease prediction using such methodologies can be more efficient, accurate and quick.

Availability of various machine and deep learning techniques has paved the way for designing such classifiers that classify diseases and give accurate predictions. Various diseases and their standard data are to be collected. Processing such large amount of data is to be done properly and efficiently. Then deploying appropriate models for prediction is another crucial and critical task as it is matter of one's health concern.

II. Material And Methods

A literature survey was taken up to understand the existing material and methodologies. Dhiraj Dahiwade[1] proposed a general disease prediction model to predict few diseases. In his approach, he explained the use of CNN and the associated accuracy of 84.5% that was comparably low when it is to health concerns.

Sunanda Das, Niyaz Rahman and Nishat Nayla[5] published on three types of brain tumor. The classes here were glioma, meningioma and pituitary tumor. This proposal gave an accuracy of 94.39% using CNN with Adam Optimization.

Another study to be major considered here was of Priyanka Sonar and K. Jayamalini[2] on the diabetic risk. They had achieved an accuracy of 85% using the simple decision tree algorithm to predict diabetes danger to a human being.

Different disease symptom data sets were collected from popular repositories. These include the Heart Risk Prediction dataset, Diabetes dataset, Brain MRI image dataset and the Chest X-Ray images dataset. A simple algorithm comparative study was made to check the existing accuracies and were found to be less or almost same as the ones in literature survey. This study included the validation of heart and diabetes models using the basic classification algorithms like Logistic Regression, Support Vector Machines, K-Nearest Neighbours and Naïve Bayes algorithms.

Then introductory concepts of deep learning were studied and the algorithms like CNN and RNN for deep learning were used to classify the chest x-ray and brain tumor data samples. Deep learning needs high CPU and GPU consumption.

Methodology

A Stacking Classifier is a hybrid classifier that has more than one classification algorithm implemented in it. A number of classification algorithms are put together on various levels starting from level-0 to level-(n-1) if 'n' algorithms are to be stacked.

To enhance the accuracy of predicting heart and diabetic risk, a stacking classifier was developed. Support Vector Machines was implemented on level-0 and Random Forest Classification was implemented on level-1 of the stacking hybrid algorithm. It generated very high accuracy than before for both the diseases when cross fold validation with 10 folds was used. A lot of improvement was noticed in the diabetes model.

The CNN algorithm was implemented for the lung image classification that was configured to have an image of (148,148,32) to (1) via feature mapping.

The brain MRI dataset was first employed with the Canny Edge Detection Algorithm, then it was trained using the Visual Geometry Group 16 (VGG-16) that was configured to have an image of (150,150,3) to (1) via canny edge detection and feature mapping. VGG-16 is also an algorithm that was built on the CNN but was trained using around 50 million pre-existing data.

Statistical analysis

To analyze the accuracies of proposed algorithms and methodologies, we used train-test split and 10-fold cross validation techniques. The train-test split used 33% of the dataset as the test part and 10-fold method generated 10 different accuracies for 10 folds and the highest accuracy was chosen.

Here, to deploy the model having the highest accuracy amongst the ten as mentioned, we used the ensemble techniques of bagging and boosting here. By using the boosting technique we deploy only the model that generated highest accuracy among the remaining nine.

III. Results

The Hybrid Machine Learning algorithm that had SVM on level-0 and RFC on level-1 was used to train the models for prediction of the heart and diabetes occurrence. When taken on par with the existing work of Galla Siva Sai Bindhika, Munaga Meghana, Manchuri Sathvika Reddy[3] for heart risk prediction model, the Stacking Classifier produced accuracy around 100%. It is very much acceptable than the existing work as it had accuracies of 92% using the linear model, 80.33% using the decision tree and 86.89% using support vector machines algorithms.

The algorithm was also seen to be working well when compared against the referenced diabetes work[2]. It generated a boosting accuracy of 91.66% as opposed to the previous ones that were 85% using decision tree induction, 77.3 using the support vector machines and 77% using the naïve bayes algorithm.

Table No 1 : Shows 10-fold Cross Validated Accuracies for Heart Dataset Test

Fold Number(CV)	Accuracy (%)	BOOSTING ACCURACY	BAGGING ACCURACY	SELECTED MODEL
1	80.0	100%	75.5%	The model with cv=7, the seventh cross validated model was deployed.
2	80.0			
3	60.0			
4	60.0			
5	40.0			
6	60.0			
7	100			
8	75.0			
9	100			
10	100			

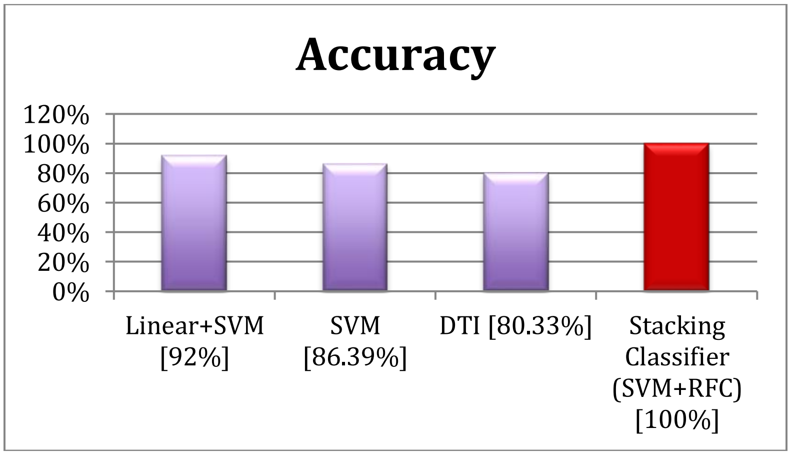

Fig 1: Comparative Study of Heart Model

In the comparison shown above the bar with color red shows the proposed stacked algorithm and its accuracy in contrast to existing work.

Table No 2 : Shows 10-fold Cross Validated Accuracies for Diabetes Dataset Test

Fold Number(CV)	Accuracy (%)	BOOSTING ACCURACY	BAGGING ACCURACY	SELECTED MODEL
1	66.66	91.66%	76.95%	The model with cv=6, the sixth cross validated model was deployed.
2	66.66			
3	71.66			
4	75.0			
5	75.0			
6	91.66			
7	72.72			
8	85.0			
9	83.33			
10	81.81			

The following is a visualization to show the comparative study of the proposed algorithm, shown as a red colored bar in the graph with contrast to the existing algorithms.

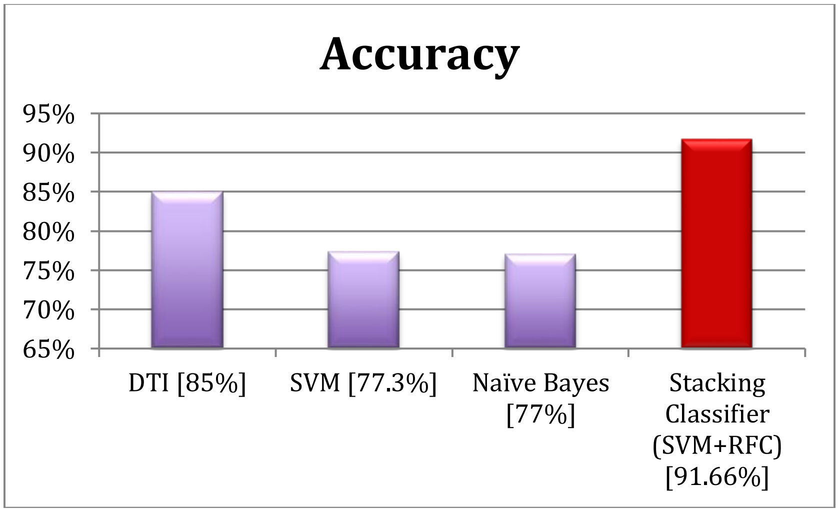

Fig 2: Comparative Study of Diabetes Model

The general disease prediction given in the reference "Designing Disease Prediction Model Using Machine Learning Approach"[1] by Dhiraj Dahiwade, had explained the normal accuracy of 84.5% on an average to be obtained after using Convolutional Neural Networks. But on comparing the proposed stacked algorithm with the explained ones, we have determined accuracies on an average to be 90% and above for most of the general diseases.

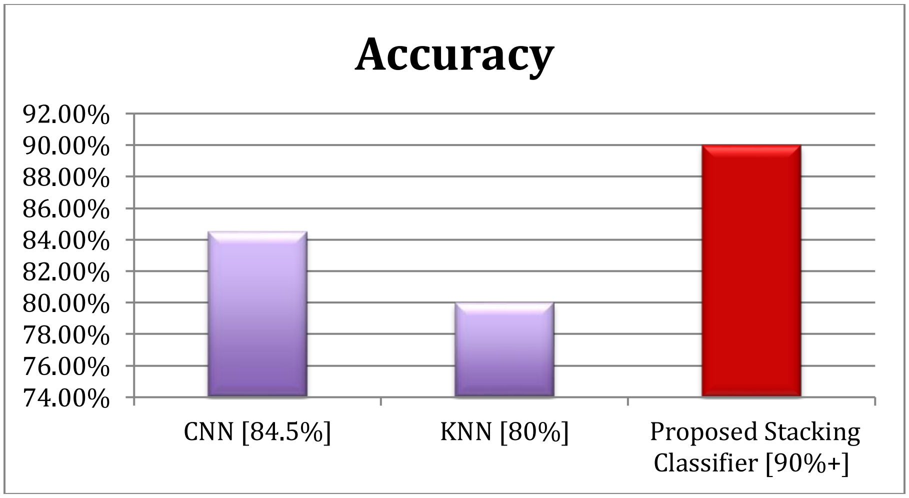

Fig 3: Comparative Study of General Disease Classification Models

The brain classifier model is a deep learning model to classify images of brain MRI samples. VGG-16 that was built upon the CNN algorithm was used with RMSProp Optimizer and Canny Edge Detection for classifying the input image to be of one among the three tumor classes we assigned during training.

The layers in VGG-16 model were configured with many convolutional layers, max pooling and dense layers. A bar graph that represents the comparison of the existing CNN model with Adam Optimization and the proposed algorithm VGG-16 with RMSProp Optimization is also shown in fig-4.

Table no 3 : Shows Configuration Of VGG-16 Brain Model

Model: “Model”		
Layer <type>	Output Shape	Param #
input_1 <InputLayer>	[<None, 150, 150, 3>]	0
block1_conv1 <Conv2D>	<None, 150, 150, 64>	1792
block1_conv2<Conv2D>	<None, 150, 150, 64>	36928
block1_pool <MaxPooling2D>	<None, 75, 75, 64>	0
block2_conv1 <Conv2D>	<None, 75, 75, 128>	73856
block2_conv2<Conv2D>	<None, 75, 75, 128>	147584
block2_pool <MaxPooling2D>	<None, 37, 37, 128>	0
block3_conv1 <Conv2D>	<None, 37, 37, 256>	295168
block3_conv2<Conv2D>	<None, 37, 37, 256>	590080
block3_conv3<Conv2D>	<None, 37, 37, 256>	590080
block3_pool <MaxPooling2D>	<None, 18, 18, 256>	0
block4_conv1 <Conv2D>	<None, 18, 18, 512>	1180160
block4_conv2 <Conv2D>	<None, 18, 18, 512>	2359808
block4_conv3<Conv2D>	<None, 18, 18, 512>	2359808
block4_pool <MaxPooling2D>	<None, 9, 9, 512>	0
block5_conv1 <Conv2D>	<None, 9, 9, 512>	2359808
block5_conv2 <Conv2D>	<None, 9, 9, 512>	2359808
block5_conv3 <Conv2D>	<None, 9, 9, 512>	2359808
block5_pool <MaxPooling2D>	<None, 4, 4, 512>	0
flatten <Flatten>	<None, 8192>	0
dense <Dense>	<None, 512>	4194816
dropout <Dropout>	<None, 512>	0
dense_1 <Dense>	<None, 1>	513
Total params: 18,91,017 Trainable params: 4,195,329 Non-trainable params: 14,714,688		

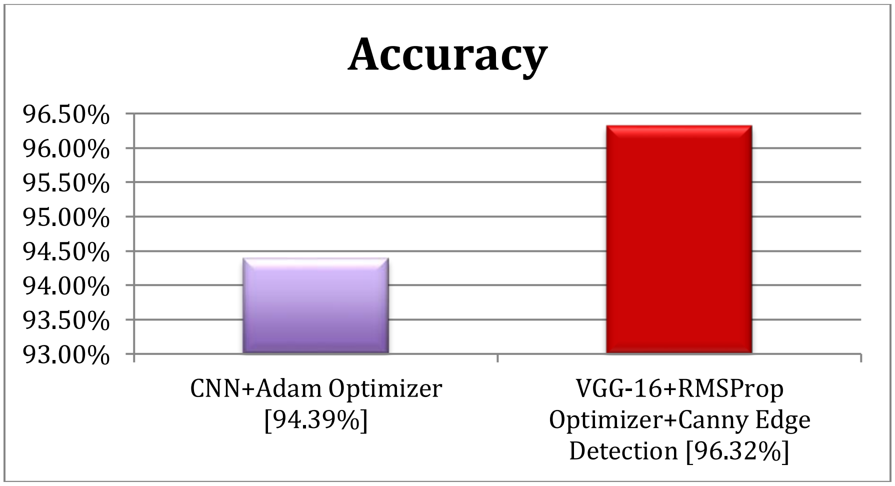

Fig 4: Comparative Study of Brain Classification Models

The Lung Classifier was deployed for Chest X-Ray image classification to detect signs of pneumonia/ COVID-19 using the CNN algorithm. It generated an accuracy of 97.03% and was configured to be-

Table no 4 : Shows Configuration Of CNN Lung Model

Model: "Sequential"		
Layer <type>	Output Shape	Param #
conv2d <Conv2D>	<None, 148, 148, 32>	896
activation <Activation>	<None, 148, 148, 32>	0
max_pooling2d <MaxPooling2D>	<None, 74, 74, 32>	0
conv2d_1 <Conv2D>	<None, 72, 72, 32>	9248
activation_1<Activation>	<None, 72, 72, 32>	0
max_pooling2d_1<MaxPooling2D>	<None, 36, 36, 32>	0
conv2d_2<Conv2D>	<None, 34, 34, 64>	18496
activation_2<Activation>	<None, 34, 34, 64>	0
max_pooling2d_2<MaxPooling2D>	<None, 17, 17, 64>	0
flatten <Flatten>	<None, 18496>	0
dense <Dense>	<None, 64>	1183808
activation_3 <Activation>	<None, 64>	0
dropout <Dropout>	<None, 64>	0
dense_1 <Dense>	<None, 1>	65
activation_4<Activation>	<None, 1>	0
Total params: 1,212,513 Trainable params: 1,212,513 Non-trainable params: 0		

As a contribution to the research work of J. Somasekar, P. Pavan Kumar Visulaization, Avinash Sharma, G. Ramesh, "Machine Learning and Image Analysis Applications in the Fight against COVID-19 Pandemic: Datasets, Research Directions, Challenges and Opportunities"[4], we have deployed a deep learning image classification model using CNN that is of acceptable accuracy.

IV. Discussion

Medical Disease diagnosis, prediction and classification have become a critical issue these days due to the running season of the pandemic. Also it is very important that the prediction is correct. Dhiraj Dahiwade[1] made a possibility of this prediction to be only up to 84.5%. The author was not specific about the disease as well.

A specific approach to diabetes risk diagnosis was given by Priyanka Sonar[2], that had generated 85% of accuracy using the decision tree algorithm, had to be enhanced for more précised predictions. There was no special mention like scaling or correlations as such.

Another research was about the brain tumor detection from the given MRI images. Sunanda, Nishat and Riaz[5] proposed the CNN with Adam optimizer that had no help towards more accurate detection of tumors with smaller impact on the MRI copy. Canny Edge Detection made the dataset very simple for the model to train. Also, VGG-16 being the highly popular model that was trained using 50 million images belonging to 5000 classes was more effective towards this approach.

B. Qian, X. Wang, N. Cao, H. Li and Y. Jiang[6] had build the model that could predict Alzheimer disease that can be considered and put to more enhancements using the provided EHR dataset. Ajinkya, Harshal and Nuzhat put all the comparative students in form of bar charts to visualize unseen patient conditions.

To classify liver dataset, Sharmila, Dharuman and Venkatesan[8] proposed two models using Decision Tree Induction and Fuzzy Neural Networks with high accuracy of 91%. More algorithms like KNN and Naïve Bayes can be used to implement it and improve the accuracy up to 100%. Apriori is also useful sometimes for disease diagnosis. This was proposed by the authors Allen Daniel and Satyam Singh[9].

V. Conclusion

A Hybrid Machine Learning Algorithm that combined Support Vector Classification and Random Forest Classification was developed for disease predictions like heart and diabetes. It generated accuracies of 90% and above for diagnosing general diseases. The algorithm is a Stacking ML Model, to be named "A Hybrid ML Algorithm for General Disease Risk Classification".

The best models, the models with highest cross validation accuracies were picked up using boosting ensemble and were deployed for further usage.

For image-based classification, CNN and VGG-16 were used. The brain classifier was 96.32% accurate and the lung classifier was 97.03% accurate. Other algorithms like Canny Edge Detection, F1 Score and Data Mining techniques were also used for data pre-processing.

References

[1]. Dhiraj Dahiwade, Prof. Gajanan Patle, Prof. Ektaa Meshram ,"Designing Disease Prediction Model Using Machine Learning Approach", "Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019), IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4", 2019.

[2]. Priyanka Sonar, Prof. K. JayaMalini, "Diabetes Prediction using different Machine Learning approaches", "Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019), IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4", 2019.

[3]. Galla Siva Sai Bindhika, Munaga Meghana, Manchuri Sathvika Reddy, Rajalakshmi ,"Heart Disease Prediction Using Machine Learning Techniques" ,"International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 p-ISSN: 2395-0072 Volume: 07 Issue: 04 | Apr 2020".

[4]. J. Somasekar, P. Pavan Kumar Visulaization, Avinash Sharma, G. Ramesh, "Machine Learning and Image Analysis Applications in the Fight against COVID-19 Pandemic: Datasets, Research Directions, Challenges and Opportunities" PII:S2214-7853(20)37062-0 DOI: https://doi.org/10.1016/j.matpr.2020.09.352, Reference: MATPR 18192, 2020.

[5]. Sunanda Das, O.F.M. Riaz Rahman Aranya, Nishat Nayla Labiba, "Brain Tumor Classification Using Convolutional Neural Network", "1st International Conference on Advances in Science, Engineering and Robotics Technology 2019 (ICASERT 2019)", 2019.

[6]. B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," *Springer Data Mining Knowl. Discovery,* vol. 29, no. 4, pp. 1070–1093, 2015.

[7]. Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh, "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare," *in IEEE big data analytics and computational intelligence*, Oct 2017 pp.2325.

[8]. S.Leoni Sharmila, C.Dharuman and P.Venkatesan, "Disease Classification Using Machine Learning Algorithms - A Comparative Study", *International Journal of Pure and Applied Mathematics* Volume 114 No. 6 2017, 1-10.

[9]. Allen Daniel Sunny1, Sajal Kulshreshtha, SatyambSingh3, Srinabh, Mr. Mohan Ba, Dr. Sarojadevi H " Disease Diagnosis System By Exploring Machine Learning Algorithms", *International Journal of Innovations in Engineering and Technology (IJIET)* Volume 10 Issue 2 May 2018.

[10]. Heart and Diabetes Symptom Datasets – *www.kaggle.com*

[11]. Lung/ Chest X-Ray Dataset – *www.kaggle.com*

[12]. Brain MRI Scan Dataset – *www.ucirepository.com*