WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Disease Prediction Using Machine Learning

M. Jayamanmadha Rao[1], P. Deexit Kumar[2], P. Akhil Kumar Patnaik[2], V. Suraj[2], G. Sai Manohar[2].

[1]Professor, Department of Electronics and Communication Engineering, Aditya Institute of Technology and Management College, Tekkali, Srikakulam, Andhra Pradesh, India, 532201.

[2]Student, Department of Electronics and Communication Engineering, Aditya Institute of Technology and Management College, Tekkali, Srikakulam, Andhra Pradesh, India, 532201.

Abstract:

In modern times, individuals encounter a range of illnesses as a result of environmental factors and their lifestyle choices, underscoring the significance of early disease detection.. So our project mainly deals with this major problem. which involves, when any person is having any health problem or effected with some disease or some abnormal things in their body, they used to visit a doctor. The problem is that many people cannot understand the kind of problems that they are having based on the symptoms. The diseases which are predicted in this research are heart disease, lung disease, liver disease and monkey pox disease. Our proposed work implements a system that predicts different diseases based on the symptoms by using machine learning algorithms like Random Forest (RF), Decision Tree (DT), K-Nearest Neighbour (KNN), Support Vector Machine (SVM). The predicted output suggests specialized doctor to the patient.

Keywords: CNN, Random Forest (RF), Decision Tree (DT), K-Nearest Neighbour (KNN), Support Vector Machine (SVM).

Introduction:

Now a days the various human beings are affecting with sicknesses and they visit doctors to get therapy of the illness. The problem is that many people affected with some diseases. Disease prediction can identify patients at risk of disease or health conditions. Due to the recent advancement of tools and techniques for data analytics, disease risk prediction can leverage large amounts of semantic information, such as demographics, clinical diagnosis and measurements, health behaviours, laboratory results, prescriptions and care utilization. In this regard, electronic health data can be a potential choice for developing disease prediction models. A significant number of such disease prediction models have been proposed in the literature over time utilizing large-scale electronic health databases, different methods, and healthcare variables. Here we are mainly focusing on some of the diseases like Heart disease, lungs disease, Monkey pox, etc. To get prediction we develop a model which helps in predicting the disease without causing any error. We are utilizing various models to anticipate a variety of diseases that are now prevalent in the world.

However, we only use machine learning techniques like Random Forest (RF), Decision Tree (DT), K-Nearest Neighbour (KNN), Support Vector Machine (SVM) for the majority of the numerical data. After these models have been trained, we use ensemble learning, which combines all the methods into one model before being trained.

The image of the monkeypox image dataset which had minor scars on its body, the model used that image as input to determine whether or not the subject had the monkey pox. And the Fig 2 shows the image belonged lung condition called pneumonia, which can only be diagnosed through x- ray images of the lungs. As a result, the model's input is an x-ray image, and it uses that to make predictions about the outcome. To access the model, we are building a webpage in which the user may submit their data and forecast if they were suffering that particular sickness or not.

Background Motivation:

With the use of machine learning and deep learning, the categorization that is currently in place was far better. In the sphere of medicine, these algorithms were helpful in making predictions. They are even capable of being trained on any kind of data using that specific model. And in order to do so, there are a number of data cleaning strategies available that are assisting in the removal of undesirable data from the dataset. The best challenge was using the machine learning algorithms themselves to forecast the data and remove the missing values from the dataset.

Training Model:

1.Decision Tree:

Decision Tree is a type of Supervised learning method that is applicable for solving both Regression and Classification problems. However, it is more commonly utilized for resolving Classification issues. This classifier has a tree-like structure where the features of a dataset are represented by internal nodes, decision rules are represented by branches, and outcomes are represented by leaf nodes. Decision tree is a predictive modelling method commonly

utilized in statistics, data mining, and machine learning, wherein an algorithmic technique is employed to divide a dataset based on various conditions. Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Classification And Regression Tree (CART) is general term for this. A Decision tree comprises two types of nodes: the Decision Node and the Leaf Node. Decision nodes enable decision-making and have multiple branches, while Leaf nodes serve as outputs resulting from those decisions and do not possess any additional branches The decision tree algorithm commences at the root node of the tree to forecast the class of a provided dataset. By matching the attribute values of the root with those of the dataset, the algorithm follows a branch and shifts to the next node. This process repeats for each subsequent node, as the algorithm evaluates the attribute values of each sub-node to continue traversing down the tree until it arrives at the terminal leaf node.

2. Random Forest:

Random Forest is a well-known technique in machine learning that falls under the category of supervised learning. Its applicability extends to various ML tasks, such as Classification and Regression. The method operates based on the principle of ensemble learning, which involves combining multiple classifiers to tackle complex problems and improve the model's accuracy.

As the name implies, "Random Forest is a classifier that contains a number of decision trees on various subsets of the provided dataset and takes the average to enhance the predictive accuracy of that dataset." Random Forest Classifier is so named because it is composed of multiple decision trees that are trained on different subsets of the provided dataset, and then the predictions from each tree are combined to make a final prediction. This approach reduces overfitting and increases accuracy compared to using a single decision tree. Additionally, there are numerous parameters that can be adjusted to customize the Random Forest Classifier for a given model.

3. K Nearest Neighbours:

K-Nearest Neighbour is a supervised machine learning technique that can be used for classification and regression. Here the K represents the number of points that are to be taken as consideration. It predicts the values of new data points using 'feature similarity,' which implies that the new data point will be assigned a value depending on how closely it resembles the points in the training set. In order to avoid either overfitting or underfitting, different values of k must be considered when defining it. Larger values of k may result in high bias and low variance, while lower values of k may have high variance but low bias.

4. Support Vector Machine:

Support vector machine works by making a straight line between two classes. That means all of the data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category. The straight line can be called as the decision boundary and also this decision boundary may not be the straight line all the time. The SVM algorithm is implemented with a kernel that converts an input data space into the desired shape. SVM employs a technique known as the kernel trick, in which the kernel transforms a low dimensional input space into a higher-dimensional space.

5. Convolutional Neural Network:

The convolution neural network is composed of a number of layers, including the convolution layer, the pooling layer, the linked layer, and the unpooling layer. The convolution layer is the fundamental building component of CNN, and it contains several kernels. Each neuron serves as a kernel. Convolution may be used by many kernels/filters to accomplish picture operations such as edge detection, blur, and sharpen. Images are

divided into small blocks in the convolution kernel, and features are extracted from each little block. To interact with pictures, the kernel multiplies their elements by the corresponding elements of the receptive region using a predefined set of weights. When the picture is too huge, the pooling layer seeks to decrease the number of parameters and lower the danger of overfitting. It also reduces computational strain and memory utilization. Spatial pooling is sometimes known as sub-sampling or down-sampling. Each map's size is lowered yet critical information is retained. CNNs fully connected (FC) layer makes use of high-level information from convolution or pooling layers.

Based on the dataset, the fully connected layer divides the input picture into distinct categories. SoftMax is mostly utilized as an activation function for classification in a fully connected layer, and the number of layers included in the network model is not rigorously restricted. During the pooling process, we build a matrix that records the position of the largest value, and the unpool operation inserts the pooled value in the original location, with the remaining members set to zero. Unpooling captures example-specific structures by returning to picture space from the original places with strong activations. As a consequence, the detailed structure is effectively rebuilt.

Results:

These are the results that we gathered from our machine and deep learning models and are tabulated below.

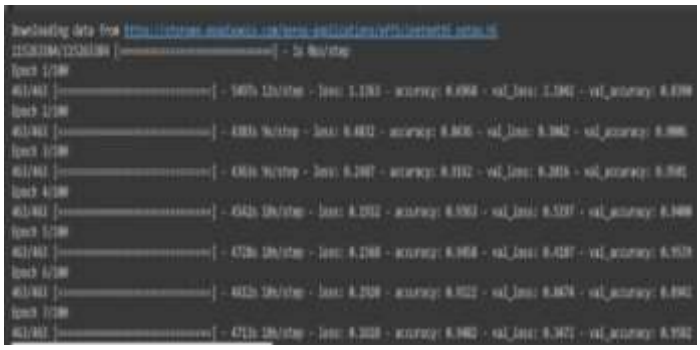
HEART DISEASE RESULT:

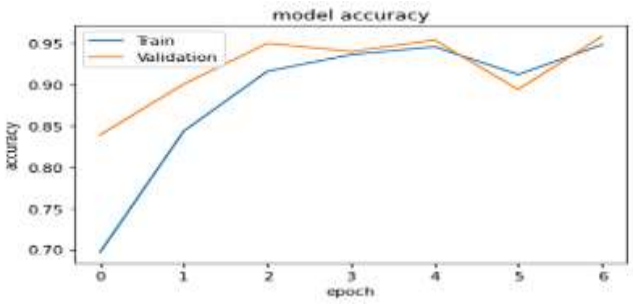
	name	accuracy	f1_score
0	SVC	0.847877	0.780311
1	Logistic Regression	0.850236	0.787910
2	KNN Classifier	0.830189	0.788238
3	Random Forest Classifier	0.839623	0.782030
4	Decision Tree Classifier	0.742925	0.746277
5	Naive bayes	0.799528	0.780886

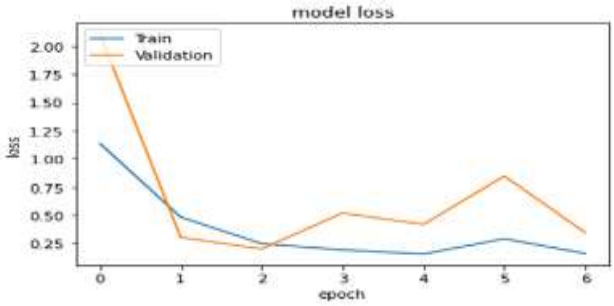
(Person suffering from heart disease)

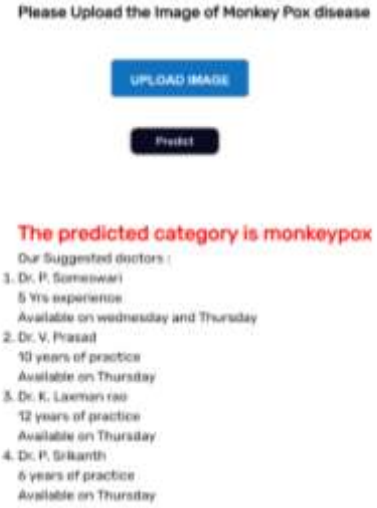
From the table, it shows the results of the heart disease and it was evident from the table that several models had been produced, and we had used some of these base models to create a hybrid method employing ensemble learning. This specific algorithm was called random forest.

MONKEY POX RESULT :





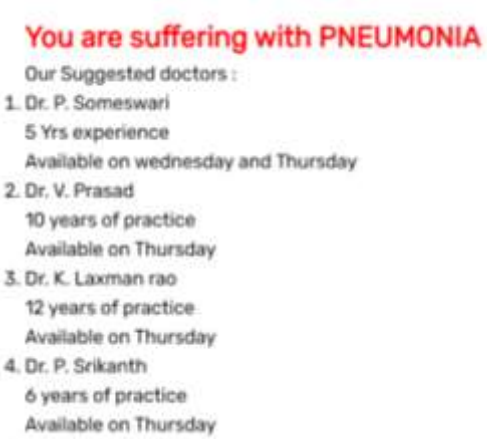




(Person suffering from monkey pox)

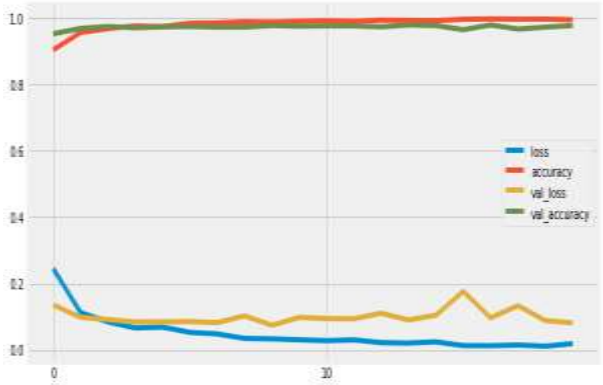
According to the table efficient net b3 and efficient net b5 are used to identify the monkey pox sickness. Of these algorithms, we learned that the efficient method outperformed the others. Because the efficient net algorithms are already trained in most of the datasets and it can detect more correctly.

LUNGS DISEASE RESULT :



(Person suffering from pneumonia)

(Normal)



From the table the implementation of the lung illness using the CNN algorithm and transfer learning led to the discovery that the CNN performed better than the transfer learning. We use the tuning strategy, which trains the model from the beginning and saves the result for later, to improve the performance of transfer learning.

LIVER DISEASE RESULT :

```
knn Score:
 95.11
knn Test Score:
 90.97
Accuracy:
 0.909653695773599
[[14414   990]
 [  951  5129]]
              precision    recall  f1-score   support

         1.0       0.94      0.94      0.94     15404
         2.0       0.84      0.84      0.84      6080

    accuracy                           0.91     21484
   macro avg       0.89      0.89      0.89     21484
weighted avg       0.91      0.91      0.91     21484
```

```
Random Forest Score:
 99.99
Random Forest Test Score:
 99.4
Accuracy:
 0.9939955315583691
[[15372    32]
 [   97  5983]]
              precision    recall  f1-score   support

         1.0       0.99      1.00      1.00     15404
         2.0       0.99      0.98      0.99      6080

    accuracy                           0.99     21484
   macro avg       0.99      0.99      0.99     21484
weighted avg       0.99      0.99      0.99     21484
```

You are suffering with Liver disease

Our Suggested doctors :

1. Dr. P. Someswari
   5 Yrs experience
   Available on Wednesday and Thursday
2. Dr. P. Akshay
   3 years of experience
   Available on Monday and Saturday
3. Dr. P. Yaswanth
   Fresher
   Available on Tuesday and Friday
4. Dr. P. Krishna vamsi
   Fresher
   Available on Wednesday and Thursday

(Person suffering from liver disease)

From the liver disease result table clearly stated that numerous algorithms were used, and from those algorithms we employed the ensemble approach, which is max voting from these algorithms.

Conclusion:

We wrapped up our talk by exploring numerous academic papers showcasing the efficacy of various models and technologies in diagnosing Heart, Lung, Monkeypox, Vector Borne, and Malaria. These studies employed diverse machine learning algorithms such as support vector machine, naive bayes, K-nearest neighbour, random forest, decision tree, and deep learning techniques like convolution neural network (CNN). To ensure the latest information, we focused on recently published papers until June 2020, accessible through renowned databases such as Science Direct, IEEE Xplore, Elsevier, and ResearchGate. Although a few studies relied on their own data, the majority utilized datasets from reliable sources including Cleveland, the UCI collection, and Lister Hill National Center for Biomedical Communications.

References:

[1]. Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning algorithm in e-healthcare. IEEE Access, 8, 107562-107582.

[2]. Kumar, N. K., Sindhu, G. S., Prashanthi, D. K., & Sulthana, A. S. (2020, March). Analysis and prediction of vascular disease using machine learning classifier. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 15-21). IEEE.

[3]. Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452- 457). IEEE.

[4]. Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine Learning Technology-Based Heart Disease Detection Models. Journal of Healthcare Engineering, 2022.

[5]. Louridi, N., Amar, M., & El Ouahidi, B. (2019, October). Identification of cardiovascular diseases using machine learning. In 2019 7th mediterranean congress of telecommunications (CMT) (pp. 1-6). IEEE.

[6]. Sharma, S., & Parmar, M. (2020). Heart diseases prediction using deep learning neural network model. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(3), 124-137.

[7]. Kieu, S. T. H., Bade, A., Hijazi, M. H. A., & Kolivand, H. (2020). A survey of deep learning for lung disease detection on medical images: state-of-the-art, taxonomy, issues and future directions. Journal of imaging, 6(12), 131.

[8]. Tripathi, S., Shetty, S., Jain, S., & Sharma, V. (2021). Lung disease detection using deep learning. Int. J. Innov. Technol. Exploring Eng., 10(8).

[9]. Tariq, Z., Shah, S. K., & Lee, Y. (2019, November). Lung disease classification using deep convolutional neural network. In 2019 IEEE international conference on bioinformatics and biomedicine (BIBM) (pp. 732-735). IEEE.

[10]. Pham, L., Phan, H., Palaniappan, R., Mertins, A., & McLoughlin, I. (2021). Cnn-moe based framework for classification of respiratory anomalies and lung disease detection. IEEE journal of biomedical and health informatics, 25(8), 2938-2947.

[11]. Olugboja, A., & Wang, Z. (2017, July). Malaria parasite detection using different machine learning classifier. In *2017 International Conference on Machine Learning and Cybernetics (ICMLC)* (Vol. 1, pp. 246-250). IEEE