ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Multi Disease Detection Using Machine Learning

Dr. N. Sri Hari[1], Associate Professor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
P. Vanaja[2], **M. Ajay Kumar**[3], **M.D.V.S. Akash**[4], **K. Sivaiah**[5]
[2,3,4,5] UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh
E-Mail ID: nshari2011@gmail.com',vanajapulivarthi555@gmail.com[2]**,**
ajaymondithoka123@gmail.com[3]**, aashum657@gmail.com** [4]**,**
sivakurri7095@gmail.com [5]

DOI:10.48047/IJFANS/V11/I12/176

Abstract

Machine learning algorithms are widely used to detect the occurrence of diseases in patients. Many of the machine learning algorithms in use today are only concerned with detecting a single disease. No system exists that can detect multiple diseases in a single application, saving patients' time. Even though there are some systems that can detect a variety of diseases, the models' accuracy varies, which has a significant impact on patients' quality of life. Building a multi-disease detection system is preferable to deploying numerous applications for each disease when a health organization wants to use machine learning models. It also saves capital. We are using the Ensembling Voting Classifier (Boosting) in Logistic Regression, SVM, Naïve Bayes, and KNN algorithms, and Random Forest which is high-efficient and gives the best accuracy. We are detecting three diseases namely Thyroid, PCOS, and Liver. When a user submits their medical information, we will be able to detect which of three diseases they are likely to be suffering from.

Keywords: Ensembling, KNN, Logistic Regression, Machine Learning, Multi-disease Detection, Naïve Bayes, Random Forest, SVM, Voting Classifier.

1. Introduction

A major field of medical study is disease detection. The ability to identify diseases early on makes them easier to treat with certain medications. This makes disease detection extremely important. Disease detection is gaining attention in the medical field because it reduces equipment costs, works as a support system for existing equipment, and has the potential to identify diseases at an early level. Based on user data multiple diseases are detected using muti-disease detection.

Using machine learning (ML), multi-disease detection requires developing algorithms that can evaluate a lot of patient data at once and spot patterns that could be indicative of multiple diseases. Many of the existing systems can detect whether a patient is having a

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

particular disease or not, for instance, they unify multiple diseases under a single user interface. For detecting heart disease they have to go to the heart disease page and enter the input fields for the result. We are detecting three diseases. There will be a single web page, in which there will be features of all three diseases and based on the data entered by the user they will be able to detect more than one disease at one time. Recently, there has been growing interest in disease detection. The detection can be done using a numerical dataset using machine learning algorithms.

SVM, KNN, Random Forest, Decision Tree and Logistic Regression are the popularly used machine learning algorithms for disease detection.

1.1 Thyroid Disease

The Thyroid gland is a vascular gland and one of the most important organs of the human body. One of the most crucial structures in the human body is the thyroid gland, a vascular gland. Two chemicals secreted by this gland aid in regulating the body's metabolism. It influences the rate of metabolism and protein synthesis. There are two distinct types of thyroid disorders: hyperthyroidism, which results in the production of too much thyroid hormone, and hypothyroidism, which results in the production of too little thyroid hormone. We are detecting whether a person is having hypothyroidism or not. Levothyroxine (T4) and triiodothyronine (T3) are two active hormones secreted by the thyroid. The internal temperature is regulated by these hormones. Iodine is regarded as the thyroid glands primary structuralcomponent. TSH(Thyroid Stimulating Hormone) is produced by pituitary glands present in the brain, a fall in the tsh level increases causes hypothyroidism. There are several types of Hypothyroidism such as primary, secondary, tertiary, and congenital.

1.2 PCOS

Menstrual cycles that are infrequent, irregular, or prolonged and frequently have elevated levels of the male hormone androgen are symptoms of polycystic ovary syndrome. The ovaries may stop regularly releasing eggs and instead develop a large number of tiny fluid-filled sacs known as follicles. Researchers say that every 1 out of 10 women are suffering from PCOS. The condition once detected cannot be cured but treatment can help relieve its effects. Some of the symptoms of PCOS are skin darkening, Acne, infertility, Excess hair growth on the face, and Dark patches on the skin. The types of PCOS include inflammatory, insulin-resistant, post-pill, and Hidden-cause. PCOS may have elevated levels of C-reactive protein (CRP), which is very dangerous.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

1.3 Liver Disease

Any condition that negatively impacts the liver and reduces its function is referred to as liver disease. The liver is a vital organ that serves many purposes in the body, including detoxifying the blood of toxins, generating bile to aid in digestion, storing vitamins and minerals, and controlling blood sugar levels. If these functions are not performed by the liver, it causes dysfunction and leads to death. As the symptoms of liver disease cannot be visible until the condition becomes chronic, it is challenging and daunting for medical health professionals to identify liver disease at its early stages. The traditional liver disease detecting systems like MRI, CT scans can cause enormous side effects. So, we can use Machine learning algorithms to detect whether a person is having a liver disease or not. The types of liver disease include Non-alcoholic fatty liver disease (NAFLD), Alcoholic liver disease, Hepatitis, Cirrhosis, and Liver cancer.

SVM, KNN, Logistic Regression, Naive Bayes, and Random Forest are frequently used techniques for detecting these three illnesses. In addition to applying the same algorithms, we are additionally using ensemble Voting Classifiers, which uses these five as the base models and output the majority class.

2. Literature Survey

In the study [7], several machine learning algorithms were used to detect the diseases of cardiac and Type-2 diabetes and obtained an accuracy of 80% and 88%. They have used Fuzzy KNN for cardiac disease and a hybrid model consisting of SVM, Naive Bayes, and decision trees with radial basis function kernels in SVM for type-2 diabetes prediction. In the study [11] they detected diabetes using Decision Tree, Naïve Bayes, and SVM algorithms, achieving accuracies of 85%, 77%, and 77.3%. Additionally, they used SVM, Decision Tree, Linear Regression, and KNN algorithms with 83%, 79%, 78%, and 87% accuracy for heart disease. Finally, using the SVM, Decision Tree, and Random Forest algorithms, they were able to identify the liver disease with accuracy levels of 95%, 87%, and 92% respectively, further they are planning to include more diseases. Furthermore, in work [12] authors created a website that incorporates different functionalities such as a doctor, patient, admin, and hospital. The authors employed four machine learning models - decision tree, linear regression, k-neighbor, and SVM, and selected the SVM algorithm, which got the highest accuracy, however, they could only predict one disease at a time. In work [9], the authors developed a system to detect Diabetes and Breast cancer using Logistic Regression and achieved an accuracy of 77.60% and 94.55%. On the other hand, KNN had the highest accuracy for predicting heart disease, with a score of 83.83%. A multi-disease prediction system is built using CNN and Random Forest Classifier in the study

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

[10]. Breast cancer (98.25%), diabetes (98.25%), heart (85.25%), kidney (99%), and liver (78%) diseases are accurately predicted by random forest. Pneumonia and malaria obtained an accuracy of 95% and 96% using CNN, where the users upload their images. In the study [3], an application is built to detect, which disease the patient is suffering from, they have considered 95 features, and took 5 symptoms from the user with the help of a Decision Tree, Random Forest, and Naive Bayes model, they were able to classify, which disease the patient is suffering from out of 41 diseases, and obtained an accuracy of 95% for all the three models. The authors can detect heart disease, diabetes, and breast cancer using Logistic Regression, SVM, and AdaBoost classifiers and obtained accuracy levels of 87%, 85%, and 98% in the study [1]. Further, they want to enhance their work using different feature selection models. Finally, in the study [15], they are predicting the common cold, malaria, and Typhoid using Multinomial Naive Bayes, Logistic Regression, and Decision tree and the obtained accuracies are 92%, 98%, and 97% respectively. We have chosen 5 models and an ensemble voting classifier, which is very efficient.

2.1 Problem Identification

Traditional methods for identifying the disease take an immense amount of time and cost, also high-equipments like MRI, and CT-Scan can cause side effects to the patients, so it is better to take support from machine learning algorithms. After reading through numerous study papers, we discovered that different feature selection techniques are used in addition to different machine learning algorithms to increase accuracy. Feature selection methods have to be carefully selected, this is the most crucial part. The current systems can either detect just one disease in a single application or multiple diseases in a single application. Multiple disease detection is common, where the user is able to detect the disease at one website. For instance, a single website contains multiple diseases such as diabetes and heart disease, but it does not find what disease the patient has based on the features given, instead, the user has to click on the heart disease and check whether he/she is having that particular disease or not. To overcome this we have proposed a solution so that based on the features given by the user the disease will be detected.

3. Methodology

Data pre-processing is used to clean, format, and organize raw data for machine learning models.

The proposed system is developed with four phases: Data Acquisition, Data Cleaning, Attribute subset selection, and Feature Scaling.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

3.1 Data Acquisition

Data acquisition is a crucial step in machine learning as it involves the process of collecting and preparing data for use in ML models. The quality and quantity of the data used to train the ML models directly impact the accuracy and performance of the resulting model. We acquired the disease dataset from Kaggle, which is useful for detection. All three thyroid, PCOS, and liver disease datasets are Indian datasets. Thyroid disease contains 29 features, Pcos contains 44 features, and liver disease contains 11 features. These are further reduced to small numbers using the feature selection methods.

3.2 Data Cleaning

Data cleaning, which fills in missing values, smoothes noisy data, resolves inconsistencies, and removes outliers, is often done as part of data pre-processing.

3.3 Attribute Subset Selection

If the Attribute selection method is not done it might lead to high dimensional data, which are difficult to train due to underfitting/overfitting problems. Only consider attributes that add more value to model training and discard those that don't.

We attempted dimensionality reduction techniques like PCA, but they decreased our accuracy, so we dropped them. Additionally, the top feature's highest entropy using the mutual info classif method is 0.2, which is very low. If the value is 0.5 or higher, we can choose that way.

After doing some research, we have found that the common feature selection method for all the six classification methods we are using is RFE(Recursive Feature Elimination), We have also tried the ANOVA method which is suitable for our dataset as it is numerical and the output is categorical.

RFE;-Recursive Feature Elimination (RFE) is a feature selection method used in machine learning. It involves recursively removing attributes (or features) from the dataset and building a model with the remaining attributes. The process is repeated until the desired number of features is obtained. RFE uses a scoring function to evaluate the importance of each attribute and selects the ones with the highest scores. We have selected nine best features using this method.

ANOVA-Analysis of Variance (ANOVA) is a statistical method used to compare the means of two or more groups of data. It tests the hypothesis that the means are equal using the F-test.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

3.4 Feature Scaling

It is a process to standardize the independent variables of a dataset within a specific range. We have used standardscaler to transfer the data into the same range. To fit the data into the same columns we have used the Standardscaler() method. Further, we used transform() along with the assigned object to transform the data and standardize it.

3.5 Models used

3.5.1 Support Vector Machine

The SVM machine learning algorithm is utilized for classification and regression tasks, determining an optimal decision boundary to divide the data while maximizing the margin between the boundary and the closest data points.

3.5.2 KNN

The KNN machine learning technique is utilized for classification and regression tasks. It allocates the input to the class that is most frequent among its k closest neighbours.

3.5.3 Logistic Regression

The logistic regression machine learning algorithm is applied to classification tasks and creates a model of the likelihood of an event happening based on input features.

3.5.4 Random Forest

The Random Forest machine learning technique is utilized for classification and regression tasks. It generates several decision trees and combines their predictions.

3.5.5 Naive Bayes

The Naive Bayes machine learning technique is applied to classification tasks and evaluates the probability of a class, considering the input features by employing Bayes' theorem.

3.5.6 Ensemble (Voting Classifier)

The ensemble learning voting classifier machine learning approach involves merging the predictions of multiple individual classifiers to arrive at a final decision. The ultimate prediction is determined through a majority vote or weighted average of the individual predictions. [17-25]

3.5.7 Framework

We have used Django to build our web application, where users enter the value to detect the disease.

4. Implementation

To implement this project we have selected python as the programming language, which contains several in-built libraries and modules. The most practical and reliable Python library for machine learning is Skearn (Skit-Learn). It gives users access to a variety of effective machine learning tools. The pseudo code as follows:

Input: Numerical Input

Output: Categorical Output (Detecting whether person having any disease)

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Begin

```
#load the dataset
data=load_data()
#Train, test ,and split the data
train_data,test_data=split_data(data)
#preprocess the data
train_data =preprocess(train_data)
test_data =preprocess(test_data)
#Define all the five models
m1 = LogisticRegression()
m2 = RandomForestClassifier()
m3= KNeighborsClassifier(n=3)
m4 = svm.SVC(kernel='linear')
m5 = GaussianNB()
#fit all the models using fit method
model.fit(train_data)
#DefineVotingClassifier
estimators=[m1,m2,m3,m4,m5]
m6=VotingClassifier(estimators)
#evaluate models
accuracy=model.evaluate(test_data)
Performance=Confusion_matrix(Predicted, Actual)
```

End

5. Results & Conclusion

After applying 6 Models on the dataset on the Liver, PCOS, and Thyroid Disease datasets. We have used two feature selection methods namely RFE(Recursive Feature Elimination), and ANOVA. The refined dataset is applied on the six models and the obtained accuracies are shown below in Table1. D1 - Liver Disease, D2 - PCOS, D3 - Thyroid, M1 - Logistic Regression, M2 - Random Forest, M3 - K-Nearest Neighbors , M4 - Support Vector Machine, M5 - Naïve Bayes and M6 – Voting Classifier.

	M1	M2	M3	M4	M5	M6
D1	71.8	99.6	95.4	71.9	28.1	97.5
D2	88.9	87.1	76.0	87.7	85.5	90.1
D3	94.8	99.1	95.9	95.4	55.0	96.6

Table 1. Comparison using accuracy of models

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

From the above table we can see that, in every disease ensembling voting classifier outperformed all the other models. Therefore,Ensembling models are efficient because they combine the predictions of multiple models to improve overall prediction accuracy and reduce the risk of over-fitting. Finally, we have chosen the voting classifier model into consideration to build our web application.

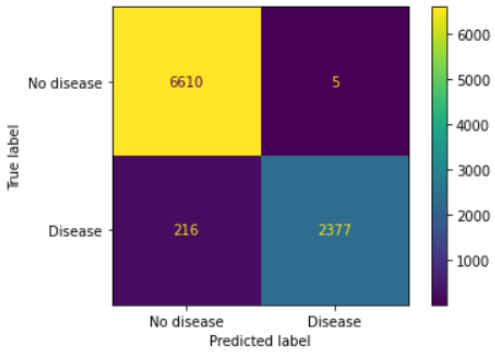

Figure 1. Confusion matrix for Liver Disease Detection

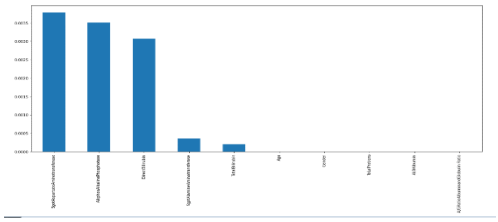

Figure 2. Visualizing Important features for Liver disease using Mutual Info Classif

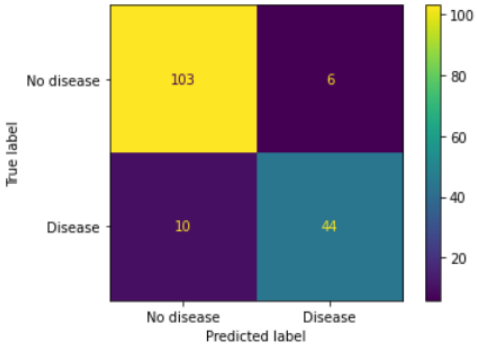

Figure 3. Confusion matrix for Liver PCOS Detection

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

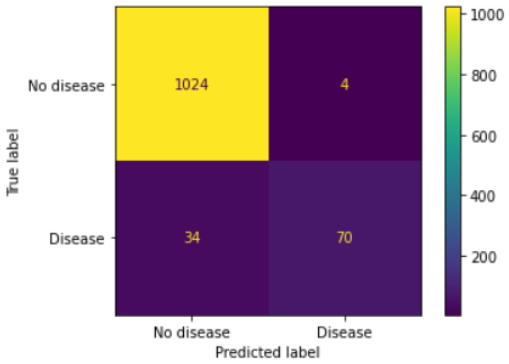

Figure 4. Confusion matrix for Thyroid Detection

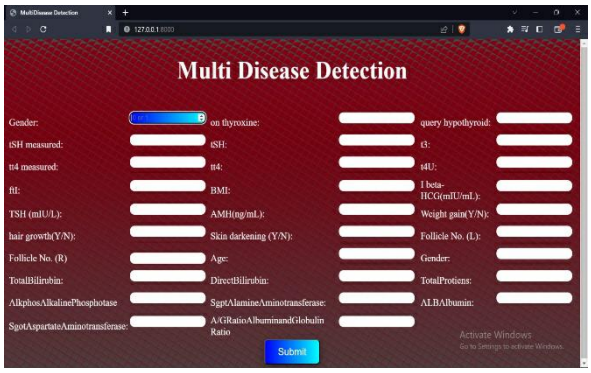

Figure 5. Web application page using django

Figure[5] depicts how our project appears on the website; when the user enters the values, they will be able to determine which of the three diseases they are currently suffering from.

6. Limitations and Future Scope

This approach enables the detection of multi-diseases; it acts as a support to the existing systems. Though it cannot be a replacement for the existing systems it acts as a support. Users can reduce the use of harmful diagnosis machines equipment like MRI, and CT-Scan. In the future, we can also add more diseases to this so all the diseases are available in one application, with the advancement of technology new algorithms would be discovered and can be applied to this application. Using different feature selection methods we have selected the features, which are efficient and fewer in number. Moreover, one limitation is that it can detect only if the data is given in numerical format.

7. References

[1] “Application of Machine Learning in Disease Prediction”(Pahulpreet Singh Kohli et al., 2018)

[2] “Disease Prediction using Machine Learning”(Kedar Pingale et al., 2019)

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved,** UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

[3] "Disease Prediction using Machine Learning Algorithms"(Sneha Grampurohit et al., 2020)

[4] "Disease Prediction using Machine Learning"(Raj H. Chauhan et al., 2020)

[5] "Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively"(Rudra A. Godse et al., 2020)

[6] "Disease prediction from various symptoms using machine learning"(RinkalKeniya et al., 2020)

[7] "Multiple disease prediction using Machine learning algorithms"(K. Arumugam et al., 2021)

[8] "Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach"(TalasilaBhanuteja et al., 2021)

[9] "An Approach to detect multiple diseases using machine learning algorithm"(Indukuri Mohit et al., 2021)

[10] "Multi Disease Prediction System"(DivyaMandem et al., 2021)

[11] "Multiple Disease Prediction System"(Ankush Singh et al., 2022)

[12] "Applications of Machine Learning in the field of Medical Care"(Mohammad Khaja SonalSharieff et al., 2022)

[13] "A Novel Approach for Forecasting Disease Using Machine Learning"(Jeevita D et al., 2022)

[14] "Disease Prediction Using Machine Learning Algorithms KNN and CNN"(K. Praveen Kumar et al., 2022)

[15] "Multiple Disease Prognostication Based On Symptoms Using Machine Learning Techniques"(Kajal Patil et al., 2022)

[16] "Disease Prediction using Machine Learning"(Suresh Singh Rajpurohit et al., 2022)

[17] Sri Hari Nallamala, et al., "A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 729 – 732, March 2018.

[18] Sri Hari Nallamala, et.al., "An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 542 – 545, March 2018.

[19] Sri Hari Nallamala, et.al, "Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems", International Journal of Advanced Trends in Computer Science and Engineering, (IJATCSE), ISSN (ONLINE): 2278 – 3091, Vol. 8 No. 2, Page No: 259 – 264, March / April 2019.

[20] Sri Hari Nallamala, et.al, "Breast Cancer Detection using Machine Learning Way", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-2S3, Page No: 1402 – 1405, July 2019.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

[21] Sri Hari Nallamala, et.al, "Pedagogy and Reduction of K-nn Algorithm for Filtering Samples in the Breast Cancer Treatment", International Journal of Scientific and Technology Research, (IJSTR), ISSN: 2277-8616, Vol. 8, Issue 11, Page No: 2168 – 2173, November 2019.

[22] Kolla Bhanu Prakash, Sri Hari Nallamala, et al., "Accurate Hand Gesture Recognition using CNN and RNN Approaches" International Journal of Advanced Trends in Computer Science and Engineering, 9(3), May – June 2020, 3216 – 3222.

[23] Sri Hari Nallamala, et al., "A Review on 'Applications, Early Successes & Challenges of Big Data in Modern Healthcare Management'", Vol.83, May - June 2020 ISSN: 0193-4120 Page No. 11117 – 11121.

[24] Nallamala, S.H., et al., "A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems", IOP Conference Series: Materials Science and Engineering, 2020, 981(2), 022008.

[25] Nallamala, S.H., Mishra, P., Koneru, S.V., "Breast cancer detection using machine learning approaches", International Journal of Recent Technology and Engineering, 2019, 7(5), pp. 478–481.