Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 22,January 2023

Human Diseases Detection Using Ensemble Machine Learning and Deep Learning Techniques

[1]Shankaranarayanan R, [2]Shruthhi K,

[3]SujithM, [4]Chithrakumar T
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore, India

Abstract—In recent times, India encounters huge varieties of diseases and detections which plays a most important role in the medical field. Early-stageidentification of diseases is high priority and necessary due to the practice of sedentary lifestyle. This study aims at appropriate and timely prediction of diseases and thereby to classify the disease based upon accuracy which would determine the person's state of health. In order to increase the efficiency of disease prediction process, which revives amidst the surroundings, the practice of Custom Ensemble Learning using Pipeline, K-Nearest Neighbor (KNN), Logistic Regression, Convolutional Neural Networks (CNN) and Principal Component Analysis is bought into usage. The data of the patients were pre-processed., compartmented and analyzed to predict which patient requires immediate treatment and is provided on priority basis which provides an immediate impact in the surroundings frequently. Ensemble Machine Learning's model is one of the wide variety of classifiers used along with K Nearest Neighbor, Nearest Mean Classifier, Random Forest, Logistic Regression.Major Medical records given a chance to find out which person requires more vital assistance. Machine Learning was used to pre-process the dataset to validate the model. To attain the high ratio of the performance of the model, both Machine learning and Deep Learning using the automatic feature engineering mechanism with multiple layers to predict the accuracy of the disease of the patients, thus creating results which are more reliable. The objective of the project model is to build a system model to splitpatient's records ,and to ensure if it is malignant or benign thereby to checkwhich affects the patients' health and to reduce the cost of the entire processes altogether eradicating redundancy produced during medical tests as well.

Keywords: Machine Learning, K Nearest Neighbors, Nearest Mean Classifier, Mean Feature Voting Classifier, KDtree KNN, Random Forest, Custom Ensemble

I. INTRODUCTION

For decades, medical diagnostics has been undergoinga series of changes that are closely associated with the advancement of technology. At present, with the advent of rapid innovations in the area of Machine learning entwined in the healthcare industry, the processes for disease prediction with the usage of high-end technology has its own progressive nature. Today, the methodology of input assimilation of patient from the external environment is subjective to the steady development of integrated algorithms and deployment of the same. The field of Machine learning and deep learning etc. are all acknowledged as part of the current backdrop of medical diagnostics. Thus, the detection of human diseases could be referred to as a constantly evolving process from the conventional process till the intelligent way of accurate disease detection with the aid of Machine learning and Deep learning techniques. Patient medical records are captious in the healthcare field for predicting variety of disease and determining the huge impact. Organizing and keeping track of all history of the patient becoming tedious and priority for treatments which requires more significance.

To administer the above problem, this model focusses on providing better solution by using Ensemble Machine Learning algorithms, KNN, Logistic Regression, Random Forest, XGBoost etc., to increase the performance and optimization of classification and prediction of patient data and produce the results based on the prediction of various diseases.

The healthcare industry has become big business. The healthcare industry produces large amounts of healthcare data every day that can be used to extract information to predict diseases that a patient may experience in the future while using treatment

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 22,January 2023

history and healthcare data. This information hidden in health data is later used to make affective decisions for the patient's health. In addition, this area needs to be improved through the use of meaningful data in healthcare. The rapid adoption of electronic health records has created a wealth of new patient data that represents a gold mine for a better understanding of human health. The main challenge is to extract the information from this data as the amount is very large so some machine learning techniques can be used.

During the manual procedure, the doctor's first step is to ensure if it is malignant or benign thereby to check which affects the patients' healthand to evaluate the blood values and perform a biopsy of the specific part. Patients may or may not have collective ailments that are not identified when the procedure is performed concurrently. To increase efficiency in predicting the types of diseases people suffer from and finding the multitude of diseases that are re-emerging in the environment, Custom Ensemble Learning is used with Pipeline, KNearest Neighbor (KNN), Logistic Regression, Convolutional Neural Networks (CNN) and Principal Component uses analysis. In addition, the expected outcome and scope of this project is that if the disease can be predicted, patients can be treated early, which can reduce life risk and save patients' lives, and the cost of disease treatment can be partially reduced through early detection.

In order to rectify the challenges and issues raised during the automatic detection of lung nodules, such as the false positives that are automated in the systems which may be present as a resultant of false procedure leading to wrong findings. This may end up in fatal treatment of the patient.Here,Precise segmentation acts as a solution which would facilitate in the accurate extraction method for the extraction of the nodules from the lung which would thereby improves the performance of the diagnostic approach. [1].

II RELATED WORK

There are many ways to get affected by lung cancer. The prolonged symptoms may lead to fatal death. In order to minimize the risk of the affecting symptoms which gives wrong findings obtained in the manual process of CT images, Geometric texture features descriptor (GTFD) and Support Vector Machine based ensemble model designed to classify the nodules of the lung from the dataset. Image Database and dataset used along with the nodules to increase the performance of the ensemble model and attained 99% of accuracy is performed by classification.[1]

A methodology for the overall prediction of the carcinogenic white blood cells from the bone marrow biopsy microscopicimage samplesthat is to be found in an automatic way is carried out with the aid of convolutional neural networks (CNN)... The method for diagnosing the same, which is carried out in a conventional manner through a manual process taken up by a skilled professional proves to be time consuming. The proposed system's motto lies in the eradication of the occurrence in the probability of errors as a result of practiceof manual method through the deployment of deep learning techniques. The overall accuracy of this methodology accounts up to 96.2%. The resultant was obtained as a cause of the application of the Dense convolutional neural network technique.[2]

The exorbitant shootup in theliver-related disorderswhich in huge figures. This system consists of the patient's documents is identified and predicted results of an individual to be present with a liver disease whichisbased on a end-to-end analyzing of the system classified model with five modular approaches of classification. The final outcome results J48 algorithm as the most efficient in the aspect of reflect phenomenon with a feature accuracy rate of 95.04% respectively.[3]

Deep Neural Network is designed with intelligent hybrid methods and achieved good performance compared than ANN on predicting coronary heart disease (CAD). Datasets containing irrelevant and mismatched features are minimized in the network is tuned slightly.The Proposed system showcases the improved performance of Conventional DNN with the maximum accuracy ranges in different hidden layers 90%, 91.83% and 87.80%. It shows clearly that it has achieved 3.33 % of feature selection in Conventional DNN model.[4]

An efficient method of detecting Parkinson's disease using multiple Algorithmic techniques to preprocess the dataset to the state of benign and malignant. Deep learning algorithms were used to detect distortion of the patients and having additional symptom of impaired speech. UCI machine learning is used to gather dataset and VGFR Spectrogram Detector using CNN and ANN, attained accuracy of 88.17% and 89.15% with the proposed model[5].

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 22,January 2023

The main aim of the breast cancer detection and diagnosis by means of FS algorithm REF, SVM. Machine learning based method diagnosis have good efficiency in finding the accuracy of cancer prognosis. The proposed method used 70% and 30% training, testing splits for model validation.
99% accuracy attained through the means of REF-SVM algorithmic combinations.[6]

The exact systematic model is built by the means of various algorithmic techniques such as KNN, SVM, Discriminant analysis. For the identification of diabetes predictions, MATLAB Classifier model and some more ensemble models with logistic regression and fold validations to attain the maximum accuracy of 95%. To get the maximum efficiency of the proposed model, Deep neural networks (DNN) will be used in future.[7]

A Combined approach of both feature selection and feature extraction methodologies is used to extract more variety of information for the identification sof breast related cancer prognosis by means of supervised classifier and unsupervised feature learning strategies such as PCA,AE, ADA algorithms and used gene based classifiers GSE11121.PCA and Autoencoder features is very accurate and far superior than any other algorithms.[8]

A methodology of data mining is primarily used nowadays to predict disease by the means of decision-making processes. Genetic Algorithms (GA) are used to extract the required information, features and minimize the efficiency of the computational works. Multiple classifiers used to predict the accuracy of breast cancer prognosis by using Gene 14 features with the well achieved accuracy of 99.48%.[9]

The approach is well developed with the Ensemble model which is used for accurate prediction of diabetes retinopathy by the combination of multiple algorithms such as Pre-screening, Lesion Specific and anatomical methods to calculate the presence of diseases. Finally attained 90% of accuracy in predicting the diabetes. Retinal image processing is very better compared to other algorithms.[10]

III. PROPOSED SYSTEM

The proposed methodologyisto develop a new model of disease diagnosis using Custom ensemble, K-Nearest Neighbor algorithm (KNN), NaïveBayes (NB), DecisionTree, Convolutional Neural Networks (CNN), Random Forest Classifier infused together for high performance-oriented outcomes in terms of disease diagnosis with quick and accurate results to the maximum possible.

The model is well trained and tested on the dataset so as to attain the best accuracy outcomes. The model can be designed to detect the Disease which finds out whether it's benign or malignant through appropriate check of entire body's parameters.Machine learning algorithms models are designed for detect the disease stages and predicts the stages in a stepwise manner. The inclusion of Custom ensemble learning techniquesthus enabling the improvement and boost up of the overall performance of the models developed.

A. Methodology

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 1, Paper 22,January 2023

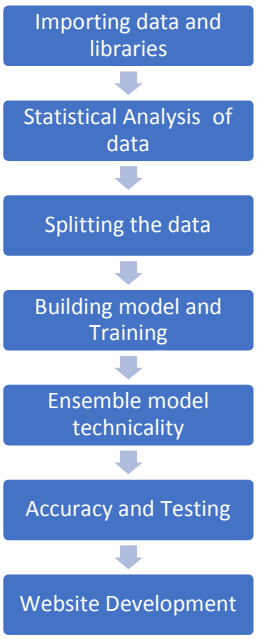

Fig.1. Flowchart of the Entire working system

B.Ensemble Algorithm with pipeline

```
models = []
models.append(('LR', LogisticRegression(random_state = 12345)))
models.append(('KNN', KNeighborsClassifier()))
models.append(('RF', RandomForestClassifier(cust_random_state = 12345)))
models.append(('SVM',SVC(gamma='auto', cust_random_state = 12345)))
models.append(('XGB', GradientBoostingClassifier(cust_random_state = 12345)))
models.append(("LightGBM", LGBMClassifier(cust_random_state = 12345)))
# evaluate each model in turn
results = []
names = []
```

IV. RESULTS

The model built using custom ensemble learning algorithms and Deep Learning algorithms paves a reliable and thus proves to be an efficient methodology for timely prediction of the diseases with the complete essence of accuracy, classification, and the deployment of the same resulting asan added featuristic aspect, thus making valuable contribution to the medical diagnostics industry.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 22,January 2023

Fig. 2. Final application running in the local cluster machine

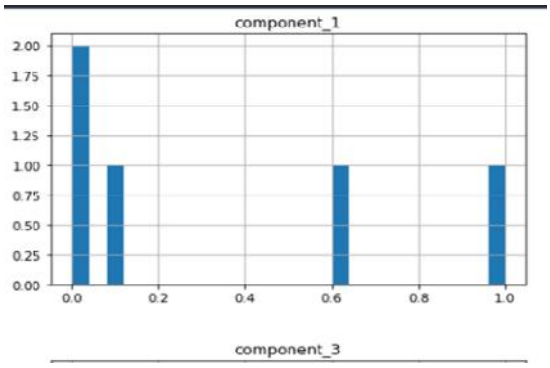

Fig 3 PCA Clustered components Plot

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 22,January 2023

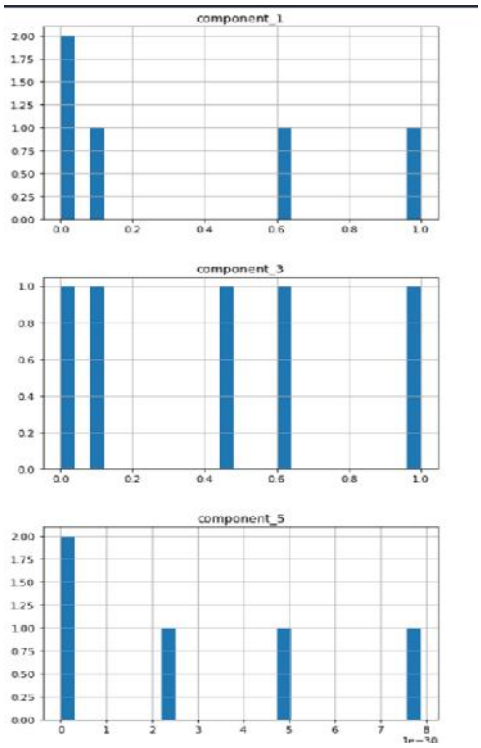

Fig 4. PCA Clustered components Plot

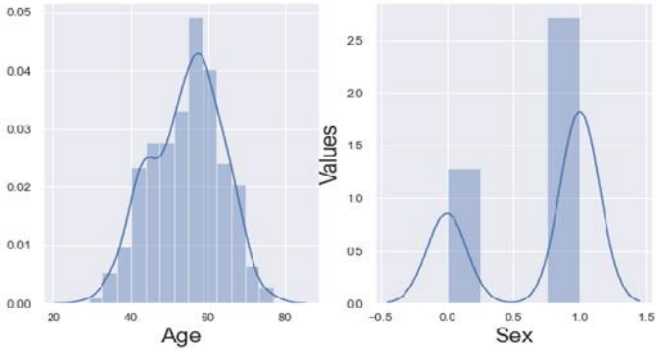

Fig 5. Histogram Comparison parameters taken

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 1, Paper 22,January 2023

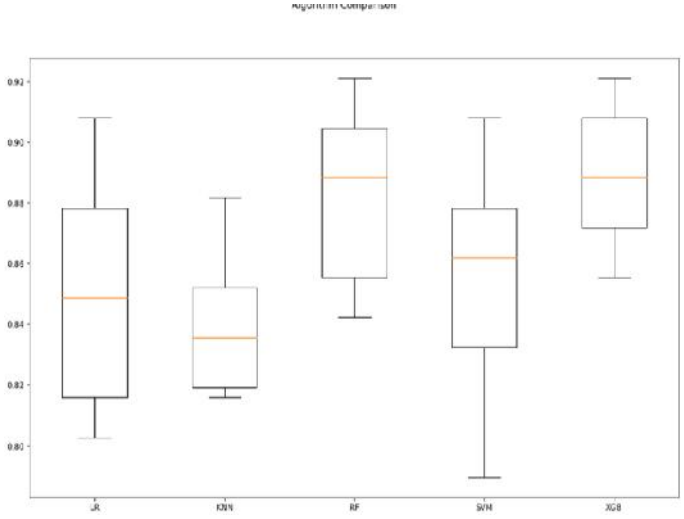

Fig 6.Algorithmic comparison and custom ensemble machine learning pipeline accuracy

[89]

	component_1	component_2	component_3	component_4	component_5
0	-55.022196	-11.260290	21.522459	4.864215	3.671235e-15
1	-45.676487	50.656851	1.033617	-6.806620	3.671235e-15
2	61.110557	49.012626	16.527997	1.295948	3.671235e-15
3	43.592004	34.244521	-13.054435	7.002067	3.671235e-15
4	116.216436	-24.628457	7.025356	-3.824533	3.671235e-15

Fig 7.Components classfications for custom ensemble model

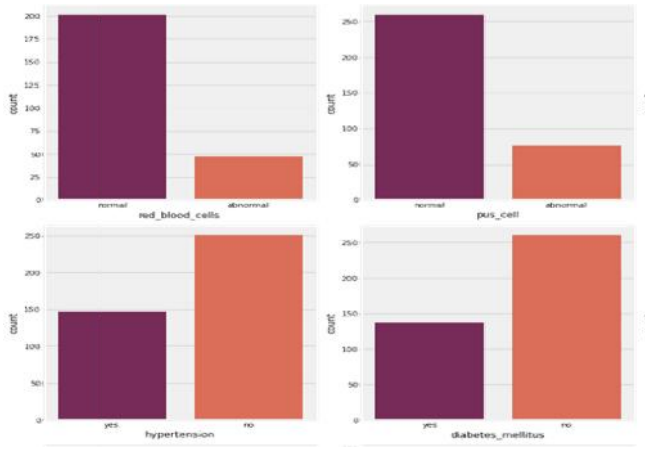

Fig 8. Plots for the attributes presents in the dataset.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 22,January 2023

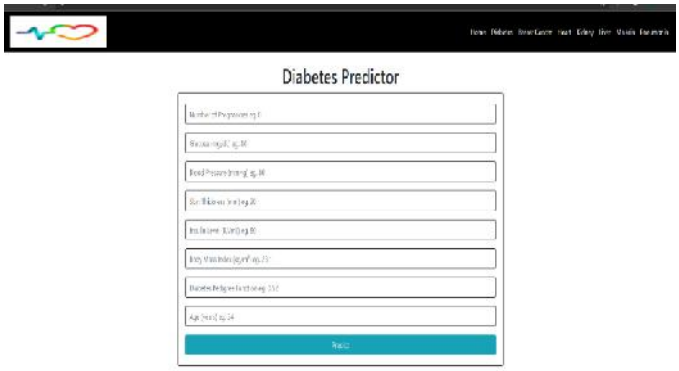

Fig 9. Main page for disease prediction website

Fig 10 Malaria Disease prediction by means of CNN using flask

Fig 11. CNN Predicted output with 96% accuracy

[1]Shankaranarayanan R, [2]Shruthhi K, [3]SujithM, [4]Chithrakumar T

V. CONCLUSION AND FUTURE ENHANCEMENT

This system will be continuously monitoring and reporting the nature and malignancy of the patient's disease, along with the clear outline of the overall well-being of the patient. The blood sample, which will be taken as input for disease prediction will be taken and thereby the vital data of patient's stage wise disease monitoring information will be updated on the database. The details can be viewed and accessed through a completely intelligent functioning of a dynamic website from time to time. Hence a software system to be developed is proposed which acts as a complete aid and supportive aspect in the field of healthcare industry for the efficient diagnosis of diseases including any discomfort too. This project will be very much useful for medical community in making appropriate decisions and timely discovery of presence of any malignant diseases tend to be present in an individual's body. Also, for additional benefit of effective methodology carried out in medical diagnostics, cost reduction and redundancy of various tests can be reduced to a large extent. The system can be expanded to measure these medical terminologies with proper integration of IOT-based input along with proper specifications.

REFERENCES

[1] Syed Muhammad Naqi,Muhammad Sharif,Mussarat Yasmin, “Multistage segmentation model and SVM-ensemble for precise lung nodule detection”,International Journal of Computer Assisted Radiology and Surgery", Vol 13, February 2018.

[2] Kumar, D., Jain, N., Khurana, A., Mittal, S., Satapathy, S. C., Senkerik, R., & Hemanth, J. D. “Automatic Detection of White Blood Cancer from Bone Marrow Microscopic Images Using Convolutional Neural Networks. IEEE Access, 8(Mm), 142521–142531,2020.

[3] Vasan Durai,Suyan Ramesh(1), “Liver disease prediction using Machine learning”, International Journal of Advance Research, Ideas and Innovations in Technology[IJARIIT],Vol 5,Issue 2, 2019.

[4] Liaqat Ali,Atiqur Rahman,Aurangazeb Khan,Mingyi Zhou,Ashir Javeed and Javed Ali Khan, “An Automated Diagnostic System for Heart Disease Prediction Based on χ 2 Statistical Model and Optimally Configured Deep Neural Network”, IEEE Access, Vol 7, April 2019.

[5] Shivangi, Anubhav Johri,Ashish Tripathi, “Parkinson Disease Detection Using Deep Neural Networks”, IEEE, Twelfth International Conference on Contemporary Computing(IC3),2019.

[6] M. Islam, H. Iqbal, R. Haque, and K. Hasan, “Prediction of breast cancer using support vector machine and K-nearest neighbors,” in Proceedings of the IEEE Region 10 Humanitarian Technology Conference (R10-HTC), vol. 23, pp. 1–5, Dhaka, Bangladesh, December 2017.

[7] Al-Zebari, A., & Sengur, A. (2019). Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection.1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019.

[8] D. Zhang, L. Zou, X. Zhou, and F. He, “Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer,” IEEE Access, vol. 6, pp. 28936–28944, 2018.

[9] E. Alickovi ˇ c and A. Subasi, “Breast cancer diagnosis using GA ´ feature selection and rotation forest,” Neural Computing and Applications, Vol. 28, no. 4, pp. 753–763, 2017.

[10] B. Antal and A. Hajdu, “An ensemble-based system for microaneurysm detection and diabetic retinopathy grading,” IEEE Trans. Biomed. Eng., vol. 59, no. 6, pp. 1720–1726, Jun. 2012.

[11] A. U. Haq, J. P. Li, M. H. Memon et al., “Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson’s disease using voice recordings,” IEEE Access, vol. 7, pp. 37718–37734, 2019.

[12] A. Ozcift and A. Gulten, “Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms,” Journal of Computer Methods and Programs in Biomedicine, Vol. 104, pp. 443-451, 2011.

[14] R. E. Schapire and Y. Singer, “Improved boosting algorithms using confidence-rated predictions,” Springer, Vol. 37, Issue 3, pp 297-336, December 1999.

[15] Y. Freund and R. E. Schapire, “Experiments with a new Boosting algorithm,” AT&T Research, Murray Hill, New Jersey, pp. 1-15, January 22, 1996, available http://www.research.att.com/orgs/ssr/people/{yoav,schapire}