Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

Multi-Disease Detection using Hybrid Machine Learning

Subhadeep Chakraborty[1*]

[1]Developer and Consultant in Artificial Intelligence, Tech Learning, Kolkata, India

DOI: 10.36347/sjet.2022.v10i10.002 | **Received:** 01.09.2022 | **Accepted:** 05.10.2022 | **Published:** 12.10.2022

***Corresponding author:** Subhadeep Chakraborty
Developer and Consultant in Artificial Intelligence, Tech Learning, Kolkata, India

Abstract | **Original Research Article**

Machine Learning has a significant application in the detection of disease because of the automated process. Using machine learning models, the detection of disease can be done with higher effectiveness and with less error which may be seen in the context of computations made by humans. In this research, the detection of multiple diseases has been done with the application of machine learning. In this research context, three data have been selected namely Heart Disease Data (from UCI Repository), Liver Disease Data (from Kaggle Repository) and Diabetes Data (from Kaggle Repository). To detect disease, four state-of-the-art classifiers have been applied along with the proposed hybrid model. By applying those classifiers or machine learning models, the detection of three diseases has been done along with the comparison of performances. In that comparison, it has been observed that the proposed hybrid model has performed the best to detect all three types of disease. In the detection of heart disease, the proposed model has achieved 96.7% accuracy, for liver disease, the accuracy has reached 97.42% and for diabetes disease detection, the proposed model has acquired 97.39% accuracy. These performances of the proposed hybrid model have also been seen to be higher compared to the existing approaches for the detection of similar diseases.

Keywords: Machine Learning, Artificial Intelligence, Classification, Disease Detection, Feature Selection.

INTRODUCTION

Human disease detection is one of the important aspects of medical science. Different types of diseases can be seen in humans with a variety of symptoms. A disease may be seen with different types of symptoms for different persons which makes the process of detection challenging. Thus, the identification of symptoms is one of the critical issues so that the disease can be detected with the highest effectiveness.

For a long age, the detection of disease was the inevitable and inseparable part of healthcare which has been done with manual efforts. In such cases, the most significant issue that arises is the computation or detection imperfectness due to human error. Gradually, with the development of statistics and computations, algorithms have been developed that act as implicit programming and can perform some specific tasks. Based on those ideas, evolution has been done for machine learning which can be applied for prediction learning purposes. It is a special kind of learning process which facilitates the detection of data labels by emphasizing the data features. In this context and in the language of machine learning, the predictable data features are called the Target feature and the features based on which the prediction is done are called Predictor features. In the domain of disease detection, the symptoms and the patient's details are taken as the predictor features based upon which the detection of disease can be done.

In machine learning, several algorithms are there which belong to different families. However, not all algorithms are suitable to detect all kinds of diseases. So, assigning the models for the detection of disease based on the symptoms is also a matter of challenge. In general fact, state-of-the-art models are used to predict the disease in one direction by applying a uniform logic. For example, the decision tree applied the tree-based method; the linear model applied the sigmoid rule etc. So, there may be lagging in the identification process as not all data may be suitable for the unified process. Thus, the application of hybrid models plays an important role by accumulating heterogenous state-of-the-art models so that the detection of the disease can be done perfectly. In this present research, the hybrid model will be applied to propose to detect different kinds of diseases.

Existing Works

Heart Disease Detection

Atallah & Al-Mousa (2019) have applied machine learning models to detect heart disease. To execute the research, they selected the heart disease dataset from the UCI repository. They have reviewed existing methods and approaches and have detected that the reasons for heart disease may vary by a person regarding the symptoms. This is one of the important reasons why doctors use to prescribe different types of medicine for the same heart disease to different patients because of the change in the symptoms. So, they have emphasized symptom-based detection. They applied different machine learning models and finally observed that the majority of the voting process detected heart disease with the highest accuracy with an accuracy rate of 90%. So, finally, they have proposed the model for the detection of heart disease with the majority voting ensemble model. Kumari & Mehta (2021) have applied machine learning to detect heart disease. While reviewing the efficiency of the machine learning model for the detection of diseases, especially heart diseases, they have observed that not all machine learning models work well for this purpose. In some of the cases, they have observed very low accuracy in detection by the models of machine learning which are generally known as weak learners. So, they have proposed a new model for the detection of heart disease by improving the efficiency of the weak learner with the application of Adaptive Boosting and Voting ensemble models. With the application of ensemble models, they have obtained 84.7% accuracy in detecting heart disease with no or very less model over fitting.

Liver Disease Detection

Alfisahrin & Mantoro (2016) have used the classifiers of machine learning to detect liver disease. To conduct the research, they selected the dataset known as the Indian Liver Disease dataset from which they have selected the important features. Based on the selected features, the researchers have prepared and finalized the data. They have applied the classifiers such as Naïve Bayes, Decision Tree etc to classify the symptoms and detect liver disease. Finally, they predicted liver disease with the application of a Decision Tree with 92% accuracy. Sivasangari *et al.*, (2020) have emphasized the detection of liver disease by detecting important symptoms. In this context, they have used the feature selection process through correlation and obtained the final features. After the selection of features and preparing the adat for detection, they have applied machine learning models such as K-Neighbors, Logistic regression, Support Vector Machine etc. They have finally proposed a Support Vector Machine to detect liver disease detection as they have obtained an accuracy of 92%.

Diabetes Disease Detection

According to Dunbray *et al.*, (2021), diabetes mellitus is one of the most prominent health problems at the current time. It is one of the critical and chronic diseases for which the pancreas stops producing insulin hormones that raise the amount of glucose in the blood. This raises the possibility of diabetes as the cells cannot absorb sugar anymore as the amount of insulin is below the expected level. Cholesterol is another factor that usually generates a higher amount of glucose in the blood. So, to retain the body free from the possibility of diabetes, the amount of glucose should be less in the blood, the amount of insulin needs to be normal in the blood and the amount of bad cholesterol needs to be less in the blood. They have gained ideas from the review of the existing research. So, they have understood the fact to detect diabetes, the emphasis on the symptoms is important. In this context, they have selected the diabetes dataset and applied machine learning classifiers to detect diabetes. However, they have not got significant accuracy in classification. Finally, they have prepared and proposed the Voting or Ensemble classifiers with Grid Search evaluation to detect diabetes and performed the detection using 93% accuracy. Kumari *et al.*, (2021) have observed that there are different symptom variations present for which the diabetes disease is seen in different patients. In the study of the existing approaches and models, they have seen that different researchers have applied different models of machine learning to detect different kinds of diseases. So, it is evident that not all models are good to detect all kinds of diseases. As they have aimed to detect diabetes disease, they have selected the dataset from UCI Machine Learning Repository. After selecting the data, they applied machine learning models such as Naïve Bayes, Random Forest etc. to detect diabetes disease. By comparing the model performances, they have finally proposed Random Forest to detect diabetes as they have obtained the highest accuracy (74%).

PROPOSED METHODOLOGY

The process to detect multiple diseases will be done based on the selected machine learning classifiers along with the proposed hybrid model. The steps of the methodology and the design of the hybrid model will be discussed in this section.

Proposed Model

The proposed model for Multi-Disease Detection is shown below:

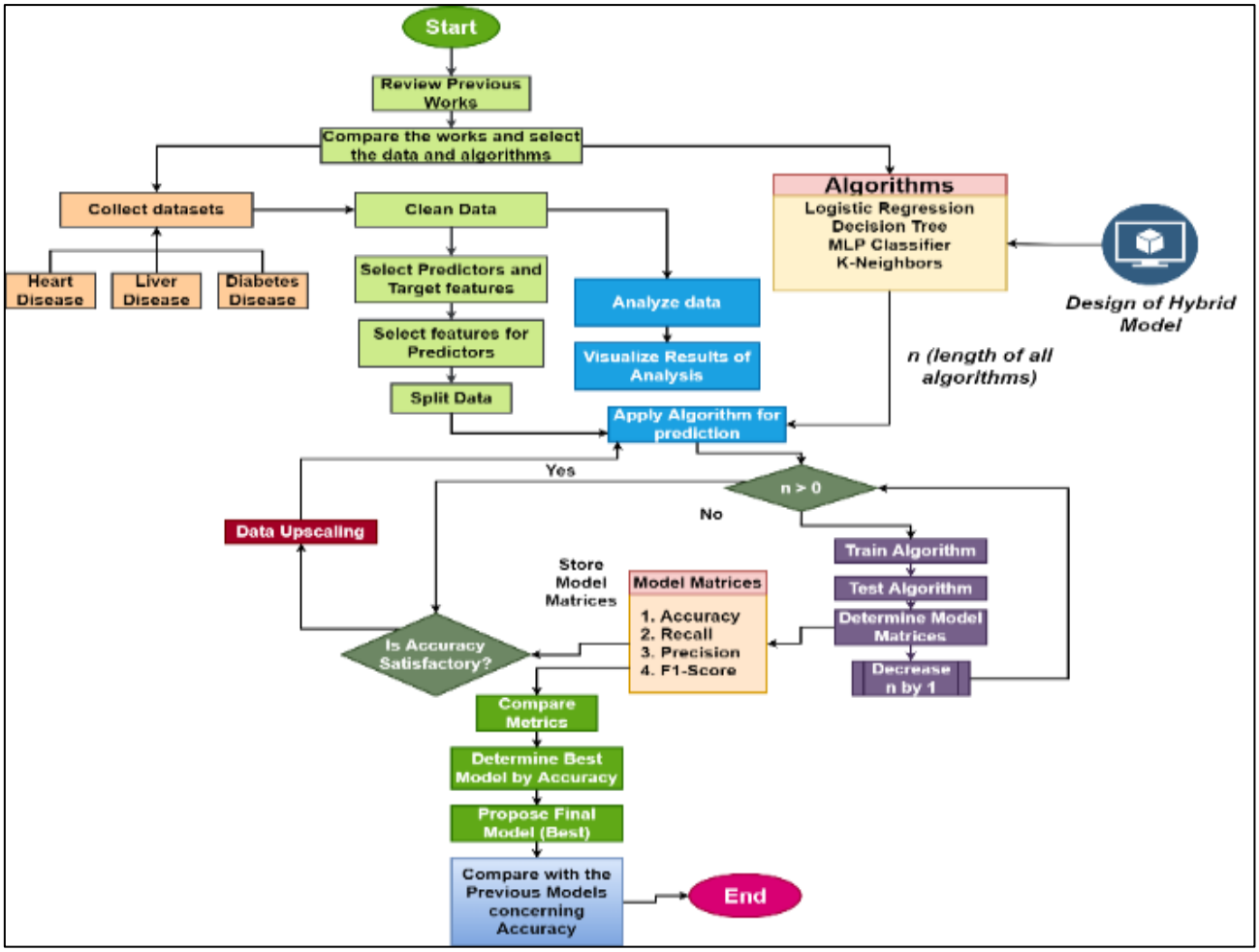

Figure 1: Proposed Model for Multi-Disease Detection

In this process, three data for three different diseases namely Heart disease, liver Disease and Diabetes Disease will be selected. The selected classifiers and the proposed hybrid model will be applied to those data for the detection of diseases based on symptoms.

Design of Hybrid Model

The hybrid model is a special type of ensemble model in machine learning. In the generalized ensemble models like the random forest, etc. the decisions are made by one kind of classifier which is the decision tree. However, for Adaptive Boosting, the estimator can be changed as per the requirement but that is single. The hybrid model is a structure of ensemble learning where multiple heterogeneous machine learning classifiers can be accumulated to design the model. The structure of the hybrid model is shown below:

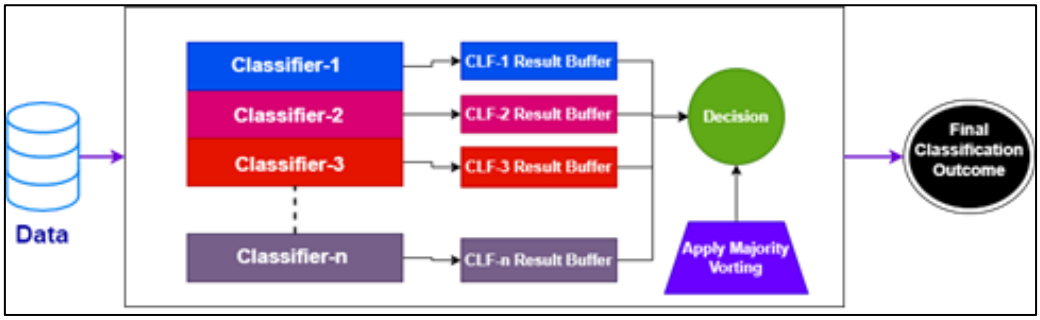

Figure 2: Structure of Hybrid Model

Based on this structure, the hybrid model has been designed with the implications of the Decision Tree and K-Neighbor classifier. The structure of the proposed model is shown below:

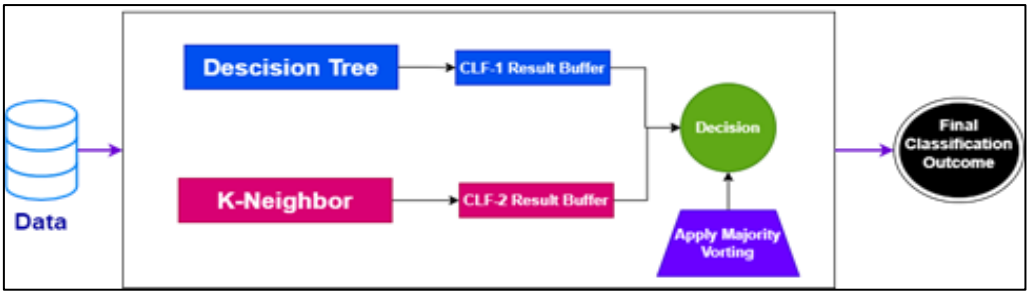

Figure 3: Proposed Hybrid Model

Selection of State-of-the-Art Algorithms

Logistic Regression

This model belongs to the Linear Family of machine learning that incorporates the sigmoid rule to classify the data. In the sigmoid rule, the labels of the data are classified as either class 0 or class 1. However, by changing the penalty, multiclass can also be applied here.

Decision Tree

It is a tree-based model of machine learning that either work with Gini impurity or Entropy. Here, Gini impurity has been applied to the design and application of the model. It starts splitting from the root node and the final decision can be achieved at the leaf node.

MLP Classifier

This classifier has a neural network structure that classifies the data with help of the linear neuron layers. The size of the hidden layer can be modified as per the requirement which has been set here as default (100 layers).

K-Nearest Neighbor

This is known as the lazy model in machine learning which classifies the data based on the neighbour values. The number of neighbours can be varied as per the necessity which has been set to default here.

Performance Measure

Accuracy

The accuracy can be measured using the following formula:

$$Acc = \frac{Correct\ Prediction}{Total\ Test\ Instances} \text{ --} \quad (1)$$

Precision

The precision can be measured using the following formula:

$$Prec = \frac{Correct\ Prediction}{Total\ Test\ Instances\ Predicted\ in\ a\ Class} \text{ --} \quad (2)$$

Recall

The recall can be measured using the following formula:

$$Recall = \frac{Correct\ Prediction}{Total\ Test\ Instances\ Present\ in\ a\ Class} \text{ --} \quad (3)$$

F1-Score

The f1-score can be measured using the following formula:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \text{ --} \quad (4)$$

Data Selection

Heart Disease Data

The data for heart disease has been collected from UCI (UCI, 1988). The database contains heart disease records for three regions namely Cleveland, Hungary and Switzerland each of which contains 303 records. All three data will be merged to form a single data which will be used in this research for the detection of heart disease. Out of all the observations, 492 people were healthy and the rest 417 people are affected with heart disease.

Liver Disease Data

This dataset has been selected from Kaggle and contains 583 records of a patient who have liver disease and are healthy (Ramana & Venkateswarlu, 2012). Out of 568 instances, 416 people are in healthy conditions and the rest 167 are affected with liver disease.

Diabetes Disease Data

The diabetes dataset also has been collected from Kaggle (Kaggle, 2016). This dataset contains 768 total observations of the patient out of which 500 are non-diabetic and 268 are diabetic.

Disease Detection and Result

The detection of diseases has been done with the application of the selected classifiers and the implication of the proposed model. The result of disease detection will be discussed in this section.

Feature Selection

The important features of the data (all three data) have been selected with the application of correlation. The features with correlation value either highest than 0.1 or higher than -0.1 have been taken as the final features. The selected features of all three data are shown below:

Table 1: Selected Features

Heart Disease		Liver Disease		Diabetes	
Feature	**Corr**	**Feature**	**Corr**	**Feature**	**Corr**
age	0.223	age	0.219	Pregnancies	0.204
cp	0.414	tot_bilirubin	0.229	Glucose	0.474
trestbps	0.151	direct_bilirubin	0.25	Insulin	0.149
restecg	0.169	tot_proteins	0.206	BMI	0.261
thalach	-0.41	albumin	0.163	Diabetes Pedigree Function	0.174
exang	0.432	ag_ratio	0.132	Age	0.244
oldpeak	0.425	sgot	-0.172	----	----
slope	0.339	alkphos	-0.152	-----	----
ca	0.46	-----		-----	-----

Data Segmentation

The final data has been prepared with the selected features. Next, the final data has been split into 80-20 ratio which implies that the training data will contain 80% of the overall data instances and the test data contains 20% of the overall data.

Disease Detection

Result of Logistic Regression

The detection results of Logistic Regression are shown below:

Table 2: Detection Results of Logistic Regression

	Accuracy	**Precision**	**Recall**	**F1-Score**
Heart	80.7692	80.78	80.77	80.75
Liver	73.4286	70.99	73.43	70.89
Diabetes	77.4403	77.85	77.44	76.06

Result of Decision Tree

The detection results of the Decision Tree are shown below:

Table 3: Detection Results of Decision Tree

	Accuracy	**Precision**	**Recall**	**F1-Score**
Heart	78.5714	79.74	78.57	78.26
Liver	76.2857	74.76	76.29	74.71
Diabetes	76.5727	76.91	76.57	75.07

Result of MLP Classifier

The detection results of the MLP Classifier are shown below:

Table 4: Detection Results of MLP Classifier

	Accuracy	**Precision**	**Recall**	**F1-Score**
Heart	82.4176	82.42	82.42	82.41
Liver	76.8571	75.82	76.86	76.08
Diabetes	75.0542	74.86	75.05	74.94

Result of K-Nearest Neighbor

The detection results of the K-Nearest Neighbor are shown below:

Table 5: Detection Results of K-Nearest Neighbor

	Accuracy	**Precision**	**Recall**	**F1-Score**
Heart	68.6813	68.68	68.68	68.63
Liver	82.5714	82.19	82.57	82.32
Diabetes	85.4664	85.5	85.47	85.16

Result of Proposed Model

The detection results of the Proposed Model are shown below:

Table 6: Detection Results of Proposed Model

	Accuracy	**Precision**	**Recall**	**F1-Score**
Heart	96.7033	96.7	96.7	96.7
Liver	97.4286	97.44	97.43	97.43
Diabetes	97.397	97.39	97.4	97.39

Performance Comparison

After detecting the diseases, the comparison of performances has been done concerning the selected metrics for all three data. The comparison of performances for Heart Disease data is shown below:

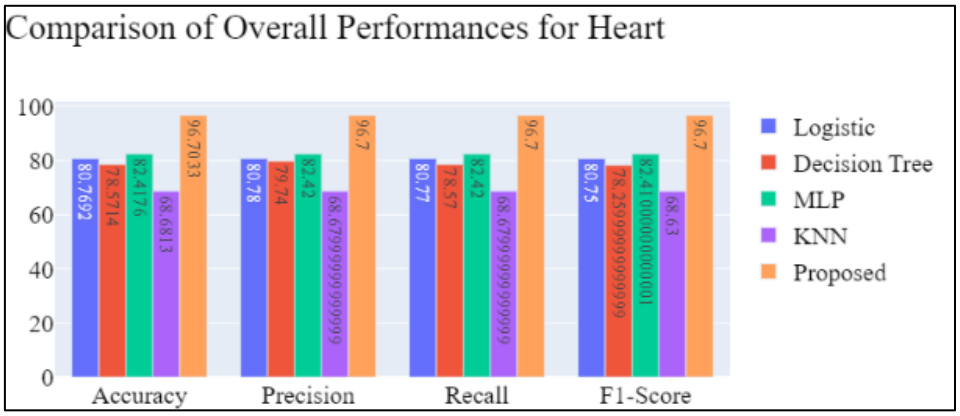

Figure 4: Comparison of Performances for Heart Disease Detection

The comparison of performances for Liver Disease data is shown below:

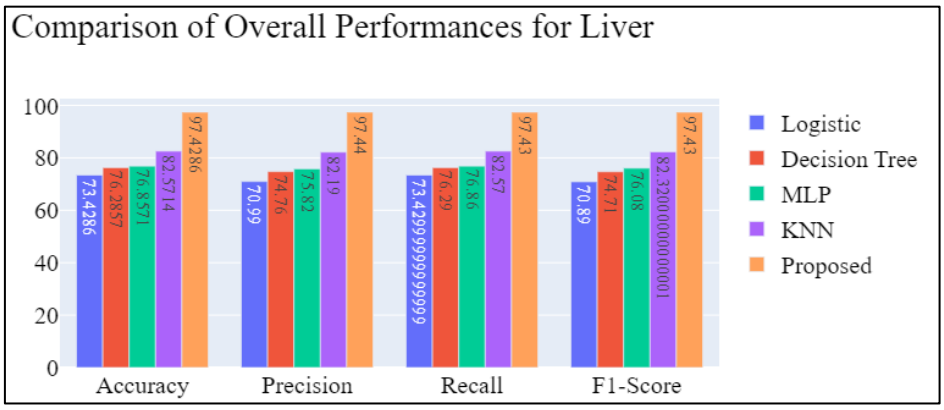

Figure 5: Comparison of Performances for Liver Disease Detection

The comparison of performances for Diabetes Disease data is shown below:

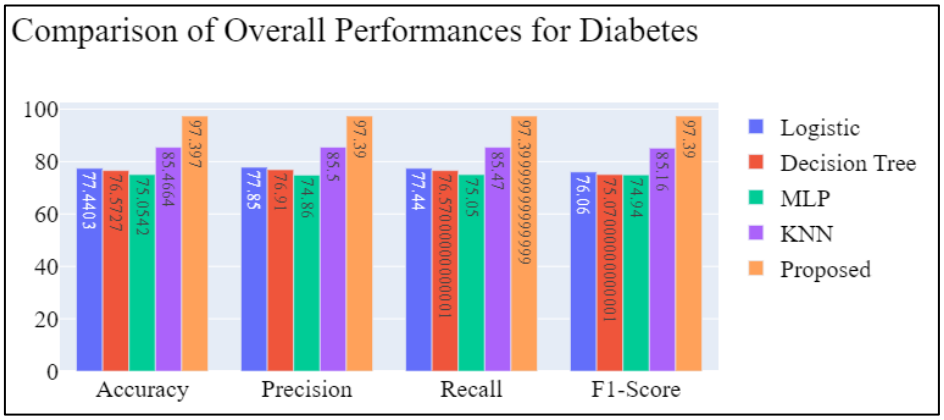

Figure 6: Comparison of Performances for Diabetes Disease Detection

From the overall comparison of the disease detection, it can be seen that the proposed model has performed the best compared to the other selected models. So, the proposed model can be said to be the best model out of the selected classifiers to detect all three types of disease. The overall accuracy comparison of the proposed model is shown below:

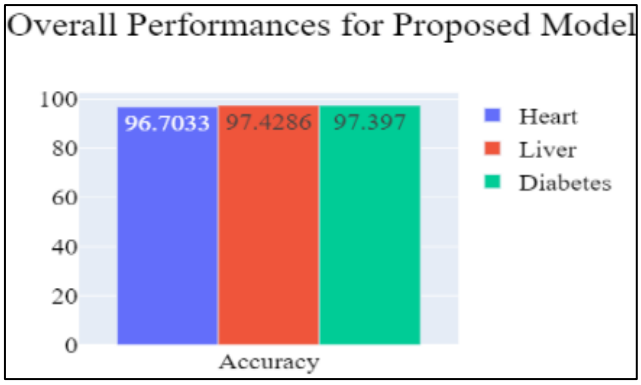

Figure 7: Overall Accuracy of Proposed Model

Additionally, the effectiveness of the proposed model has been compared with the existing research. It has been observed that the proposed model is outperforming the existing approaches. The comparison is shown below:

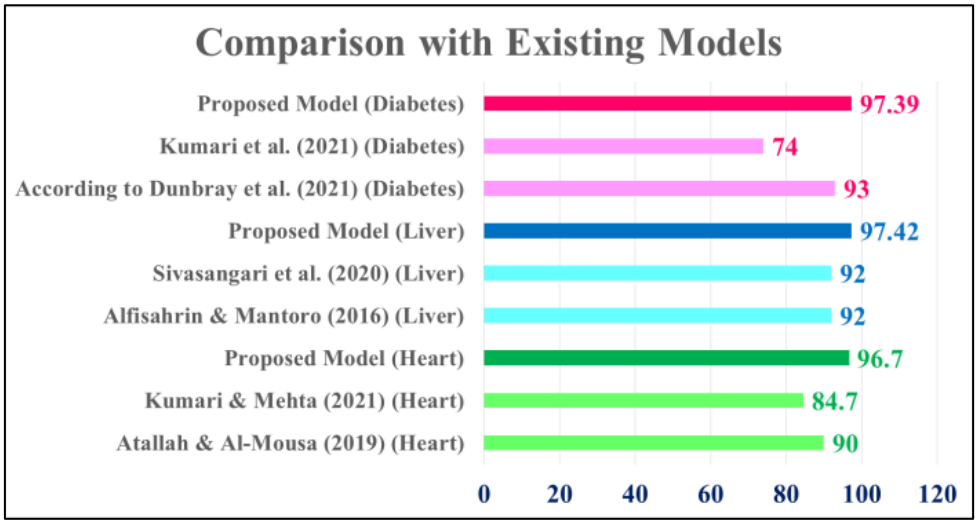

Figure 8: Comparison with Existing Research

CONCLUSION

Detection of the disease has been commenced in this research with the application of machine learning. In this scenario of research, the hybrid model has been designed which has been applied to three datasets for heart disease, liver disease and diabetes disease. As the detection of disease should be robust and should be done equally for all kinds of diseases, the achievement has been attained in this research. All the data have been processed and prepared similarly and with the help of the proposed hybrid model, the detection of all three diseases has been done with the highest accuracy compared to the selected models. Additionally, the accuracies of detecting all three diseases have been seen to be highest compared to the existing approaches.

REFERENCES

- Cıhan, P., & Coşkun, H. (2021). Performance Comparison of Machine Learning Models for Diabetes Prediction. *29th Signal Processing and Communications Applications Conference (SIU)*, pp. 1-5.
- Costea, N. E., Moisi, E. V., & Popescu, D. E. (2021). Comparison of Machine Learning Algorithms for Prediction of Diabetes. *16th International Conference on Engineering of Modern Electric Systems (EMES)*, pp. 1-5.
- Dunbray, N., Rane, R., Nimje, S., Katade, J., & Mavale, S. (2021, October). A Novel Prediction Model for Diabetes Detection Using Gridsearch and A Voting Classifier between Lightgbm and KNN. In *2021 2nd Global Conference for Advancement in Technology (GCAT)* (pp. 1-7). IEEE.
- Kaggle. (2016). *Pima Indians Diabetes Database.* [Online] Available at: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database [Accessed 2021].
- Katarya, R., & Jain, S. (2020). Comparison of Different Machine Learning Models for diabetes detection. *IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE)*, pp. 1-4.

- Keya, M. S., Shamsojjaman, M., Hossain, F., Akter, F., Islam, F., & Emon, M. U. (2021, March). Measuring the Heart Attack Possibility using Different Types of Machine Learning Algorithms. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (pp. 74-78). IEEE.
- Kumari, L. R., Shreya, P., & Begum, M. (2021). Machine Learning based Diabetes Detection. *6th International Conference on Communication and Electronics Systems (ICCES)*, pp. 1-5.
- Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2019). A machine learning methodology for diagnosing chronic kidney disease. *IEEE Access*, *8*, 20991-21002.
- Ramana, B. V., & Venkateswarlu, S. P. B. a. N. B. (2012). *ILPD (Indian Liver Patient Dataset) Data Set.* [Online] Available at: https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)
- Sivasangari, A., Reddy, B. J. K., Kiran, A., & Ajitha, P. (2020, October). Diagnosis of Liver Disease using Machine Learning Models. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 627-630). IEEE.
- UCI. (1988). Heart Disease Data Set. [Online] Available at: https://archive.ics.uci.edu/ml/datasets/heart+disease [Accessed 2022].
- Alfisahrin, S. N. N., & Mantoro, T. (2013, December). Data mining techniques for optimization of liver disease classification. In *2013 International Conference on Advanced Computer Science Applications and Technologies* (pp. 379-384). IEEE.
- Sheshadri, H. S., Shree, S. B., & Krishna, M. (2015, August). Diagnosis of Alzheimer's disease employing neuropsychological and classification techniques. In *2015 5th International Conference on IT Convergence and Security (ICITCS)* (pp. 1-6). IEEE.
- Hasija, Y., Garg, N., & Sourav, S. (2017, December). Automated detection of dermatological disorders through image-processing and machine learning. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 1047-1051). IEEE.
- Atallah, R., & Al-Mousa, A. (2019, October). Heart disease detection using machine learning majority voting ensemble method. In *2019 2nd international conference on new trends in computing sciences (ictcs)* (pp. 1-6). IEEE.
- Hamsagayathri, P., & Vigneshwaran, S. (2021, February). Symptoms Based Disease Prediction Using Machine Learning Techniques. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 747-752). IEEE.
- Chauhan, T., Rawat, S., Malik, S., & Singh, P. (2021, March). Supervised and unsupervised machine learning based review on diabetes care. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 581-585). IEEE.
- Abdulhadi, N., & Al-Mousa, A. (2021, July). Diabetes detection using machine learning classification methods. In *2021 International Conference on Information Technology (ICIT)* (pp. 350-354). IEEE.
- Kumari, A., & Mehta, A. K. (2021, August). A Novel Approach for Prediction of Heart Disease using Machine Learning Algorithms. In *2021 Asian Conference on Innovation in Technology (ASIANCON)* (pp. 1-5). IEEE.
- Varshney, N., & Ahuja, S. (2021, October). An Intelligence System for Medicine Recommendation. In *2021 5th International Conference on Information Systems and Computer Networks (ISCON)* (pp. 1-5). IEEE.
- Hussain, A., & Sharma, A. (2022, April). Machine Learning Techniques for Voice-based Early Detection of Parkinson's Disease. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 1436-1439). IEEE.