

Original Research

Received 03 June 2026

Revised 18 July 2026

Accepted 01 Aug 2026

Natural Resources for Human
Health 2026, Vol. 6 (15s): 1156–1166

KEYWORDS:

Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), XGBoost (XGB), Gradient Boosting (GB), Gaussian Naïve Bayes (GNB), AdaBoost (AB), Machine Learning (ML), Artificial Intelligence (AI), Deep Learning (DL).

Machine Learning Based Prediction of Diabetes Heart Liver Kidney Breast Cancer and Parkinsons Disease

Ramdas Pandurang Bagawade¹, Dr. Thirupurasundari D.R²

¹ Bharath Institute of Higher Education and Research, India (Chennai)

² Bharath Institute of Higher Education and Research, India (Chennai)

ABSTRACT: Chronic diseases like diabetes and heart, liver, and kidney diseases, as well as Parkinson's and breast cancer, contribute to an increasing burden to the global healthcare industry. Thus, there is a need for sophisticated predictive systems that can proactively and accurately diagnose these conditions. This paper explores how machine learning can assist in predicting high disease risk based on available demographic and clinical data and patient histories. Key techniques such as Support Vector Machines (SVM), Logistic Regression (LR), and Decision Trees (DT), as well as Random Forest (RF), an ensemble technique, will be used to analyze and classify data. To retain the model's interpretability, important predictors will be identified using feature selection. The models will be evaluated based on accuracy, precision, recall, and F1 score to recommend a suitable option for each disease type. It will be shown that machine learning will reshape the diagnosis and prognosis of diseases, building a foundation for timely and individualized medical interventions. Data-driven healthcare as an approach to optimize patient care and alleviate patient care demands will be demonstrated.

1. INTRODUCTION

The last few decades have seen an unprecedented growth in chronic conditions which poses a serious challenge to health care systems the world over. Some of the examples of these are the different types of Diabetes, Heart, Liver, Kidney, Breast Cancer, and even Parkinson's disease. The goal is always to lessen the burden of these diseases on the patients and the systems. For that early detection of conditions is the most critical aspect. There have been efforts to design better diagnostics and prognostics, however, traditional diagnostics are still heavily reliant on human skills and are often, time, and resource inefficient, as well as being subjective and error prone. Machine Learning, and especially the latest developments in the field, has been background for solutions in data driven, automated, and adaptive prediction systems for robust and complex patient data analysis and processing.

The focus of this paper is to facilitate the prediction of the aforementioned chronic conditions using tailored patient data (demographic, behavioral, clinical, and genetic) by using a number of machine learning techniques and classifiers (SVM, LR, and Ensemble techniques). The aim is to enhance and fine tune prediction modeling and to identify important variables in the process. The objective is to demonstrate the potential of machine learning in transforming the health care industry for early measures and personalized care. These predictive models will also help in optimizing limited health care resources and will help improve patients' well-being. Diabetes will eventually cause many related worries including damage to the heart and blood vessels, eyes, kidneys, feet, and nerves [25]. These issues shoot from the harm inflicted by persistently elevated blood sugar levels on blood vessels and nerves across the whole body.

2. METHODS

The aim of this method is to use several machine learning algorithms, such as LR, SVM, KNN, RF, AB, GB and DT, to predict diabetes, heart, liver, kidney, breast cancer and Parkinson chronic diseases. The method is organized into the following steps:

A) Data Collection and Pre-processing

The Patient data, comprising demographics, clinical metrics, lifestyle factors, and medical history, is collected from relevant sources. Then missing values are imputed, and categorical variables are encoded using techniques such as one-hot encoding or label encoding. After that feature scaling methods like Min-Max scaling or Standardization are applied to ensure uniformity in numerical variables.

B) Feature Selection

Feature selection techniques are applied in order to obtain the most important features for given base model for given dataset. It is observed that the different features are selected for different base model for same data set.

C) Model Development

We implemented LR, SVM, KNN, RF, AB, GB and DT in Python to develop models to predict chronic diseases. Kaggle provided the training datasets for Diabetes, Heart, Liver, Kidney, Breast Cancer, and Parkinson's Disease [25] to [33]. We deployed all models.

D) Model Training and Validation:

In this research, we adopted stratified sampling to divide the datasets. As a result, 70% of the data was utilized to train the models and the remaining 30% was utilized for field tests. This ensured that all classes of the dataset were equally represented in both sets. We then validated the model's robustness by assessing the level of overfitting using k-fold cross-validation. Tuning the hyper parameters of the model employed grid and random search methods for optimization.

E) Performance Evaluation:

Model predictive performance will be assessed using the standard metrics of accuracy, precision, and recall, as well as the F1-score and the AUC-ROC. Finally, a comparison of algorithms will be conducted to identify the strengths and weaknesses of each of the algorithms, and the most suited one for the disease in question will be used.

F) Interpretation and Deployment:

Based on the results of the models, healthcare personnel make informed decisions. The selected models are uploaded to the interface designed for use in the clinic.

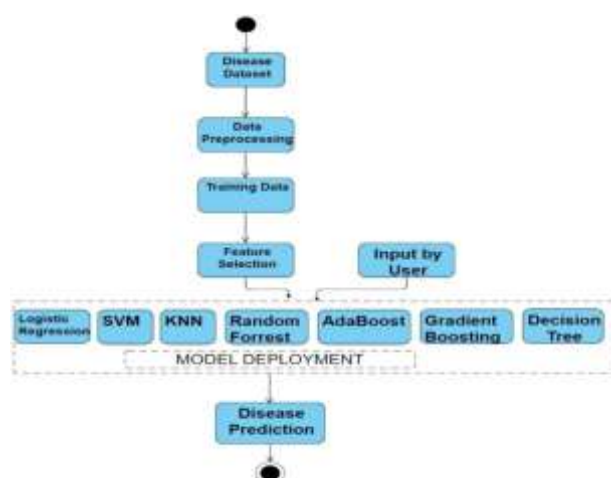


Fig 1: Architecture Diagram

3. RESULTS AND DISCUSSION

The analysis focused on seven machine learning techniques like LR, RF, SVM, DT, GB, AB, and KNN. The goal was to predict seven diseases including Diabetes, Heart, Liver, Kidney, Breast Cancer, and Parkinson's Disease.

- Random Forest achieved perfect scores in predicting Kidney and Heart diseases, and consistently high results across the rest of the diseases. Strong and robust results suggest the efficiency of Random Forest in the multi-disease prediction task.
- Gradient Boosting achieved 100% for predicting Kidney diseases and scores above 96% for rest of the diseases.
- With respect to Kidney diseases, KNN achieved 100% precision. KNN also achieved a score of 94.03% for predicting Liver diseases. Adaboost also achieved scores above 89% for all the diseases.
- With respect to Heart diseases, SVM achieved a 66% score, and therefore has a problem with class imbalance.
- Decision Tree achieved a score of 98% with respect to Kidney diseases but only 73% with respect to predicting Diabetes.
- Logistic Regression falls behind all models with respect to predicting Heart diseases. Logistic Regression achieves one of the highest scores for Heart diseases but falls behind all models with respect to predicting Heart diseases.

Overall Observation:

Among individual classifiers, Random Forest and Gradient Boosting, as types of ensemble learning, achieved the best results for each disease. These models provide healthcare practitioners a higher level of confidence in making timely and accurate predictions regarding illnesses. The chart confirms the superiority of ensemble approaches for stable and precise results in multi-class diagnostic systems.

DIABETES: DATASET 1 [26]:

Table 1. Results obtained for Diabetes Dieses (using Dataset [26])

Sr.No.	Model	Ac	Pr	Re	F1
1	LR	0.964227	0.881657	0.689815	0.774026
2	RF	0.976151	0.975904	0.750000	0.848168
3	SVM	0.964638	0.971014	0.620370	0.757062
4	DT	0.955181	0.727660	0.791667	0.758315
5	GB	0.974918	0.969697	0.740741	0.839895
6	AB	0.974095	0.947368	0.750000	0.837209
7	KNN	0.961760	0.907285	0.634259	0.746594

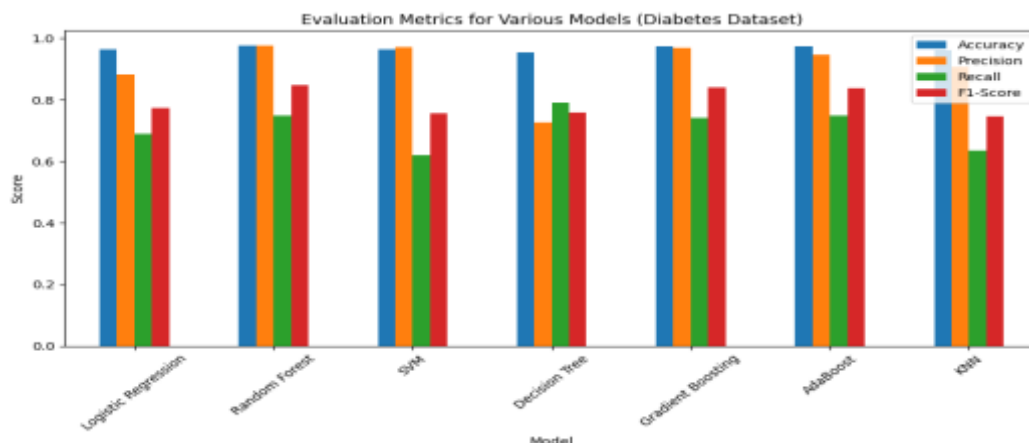


Fig 2: Evaluation Metrics for Various Models using Diabetes Dises (using Dataset [26])

DIABETES:DATASET 2 (Indian Dataset) [27]:

Table2. Results obtained for Diabetes Dises (using Indian Dataset [27])

Sr.No.	Model	Ac	Pr	Re	F1
1	LR	0.753247	0.649123	0.672727	0.660714
2	RF	0.720779	0.603448	0.636364	0.619469
3	SVM	0.733766	0.645833	0.563636	0.601942
4	DT	0.753247	0.644068	0.690909	0.666667
5	GB	0.740260	0.627119	0.672727	0.649123
6	AB	0.733766	0.625000	0.636364	0.630631
7	KNN	0.694805	0.583333	0.509091	0.543689

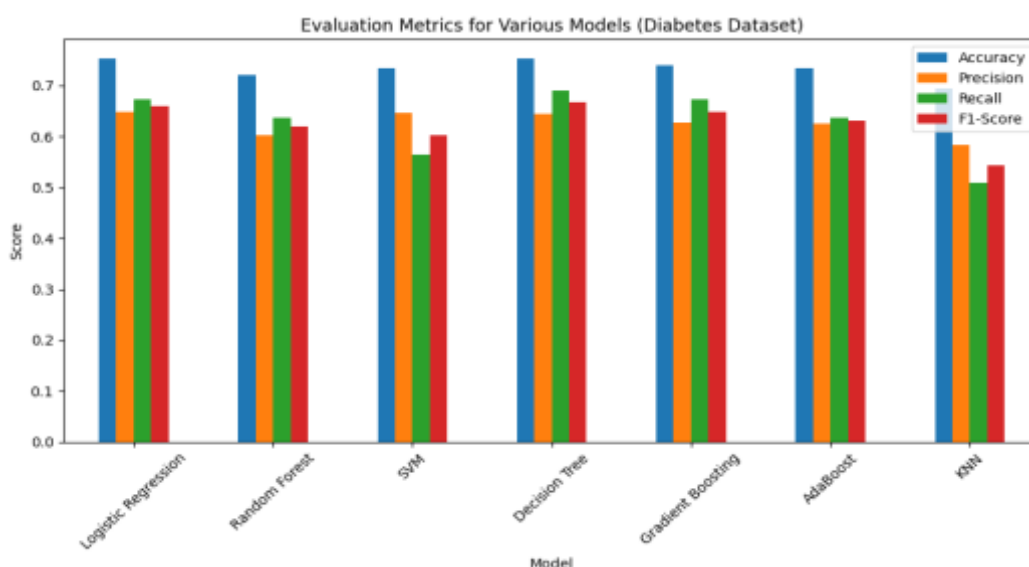


Fig 3: Evaluation Metrics for Various Models using Diabetes Dises (using Indian Dataset [27])

HEART :DATASET 1[28]:

Table 3. Results obtained for Heart Dieses (using Dataset [28])

Sr.No.	Model	Ac	Pr	Re	F1	AUC
1	LR	0.795122	0.756303	0.873786	0.810811	0.876832
2	RF	0.985366	1.000000	0.970874	0.985222	1.000000
3	SVM	0.682927	0.661017	0.757282	0.705882	0.706691
4	DT	0.985366	1.000000	0.970874	0.985222	0.985437
5	GB	0.931707	0.915888	0.951456	0.933333	0.980678
6	AB	0.878049	0.890000	0.864078	0.876847	0.948506
7	KNN	0.731707	0.730769	0.737864	0.734300	0.860461

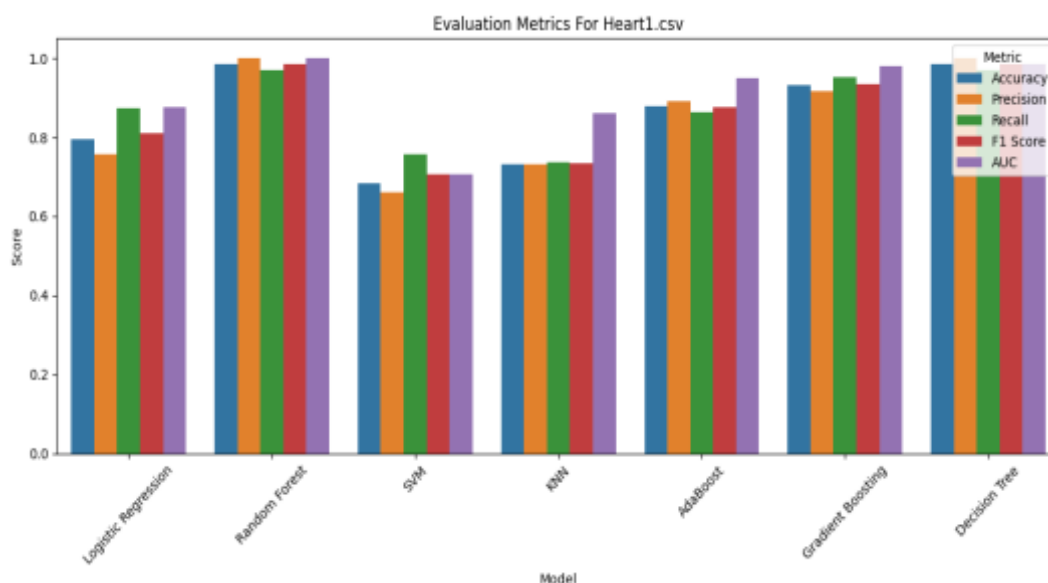


Fig 4: Evaluation Metrics for Various Models using Heart Dieses (using Dataset [28])

HEART :DATASET 2: (Indian Dataset) [29]

Table 4. Results for Heart Dieses Dataset [29]

Sr.No.	Model	Ac	Pr	Re	F1	AUC
1	LR	0.728333	0.755906	0.655290	0.702011	0.794255
2	RF	0.683333	0.684588	0.651877	0.667832	0.750592
3	SVM	0.720000	0.775330	0.600683	0.676923	0.785928
4	DT	0.585000	0.579710	0.546075	0.562390	0.584112
5	GB	0.740000	0.754647	0.692833	0.722420	0.805689
6	AB	0.730000	0.771784	0.634812	0.696629	0.795155
7	KNN	0.693333	0.695341	0.662116	0.678322	0.713822

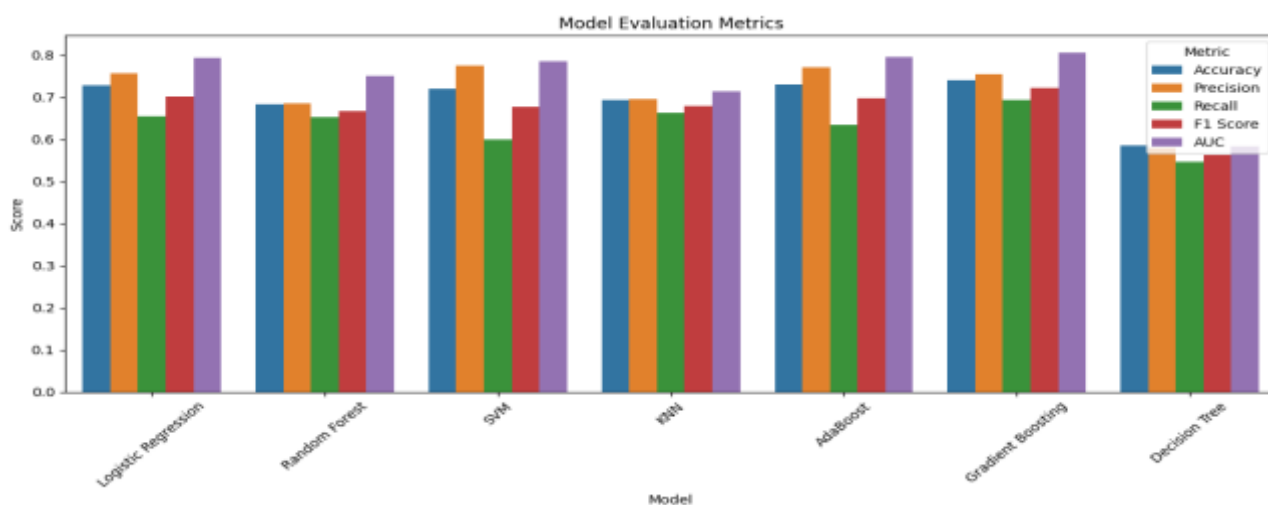


Fig 5: Evaluation Metrics for Various Models using Heart Dieses (using Dataset [29])

HEART: DATASET 3:

Table 5. Results obtained for Heart Dieses (using Dataset [30])

Sr.No.	Model	Ac	Pr	Re	F1	AUC
1	LR	0.783333	0.833333	0.60	0.697674	0.814857
2	RF	0.716667	0.750000	0.48	0.585366	0.838857
3	SVM	0.583333	0.000000	0.00	0.000000	0.522286
4	DT	0.633333	0.588235	0.40	0.476190	0.600000
5	GB	0.750000	0.777778	0.56	0.651163	0.884571
6	AB	0.750000	0.750000	0.60	0.666667	0.841143
7	KNN	0.533333	0.285714	0.08	0.125000	0.439429

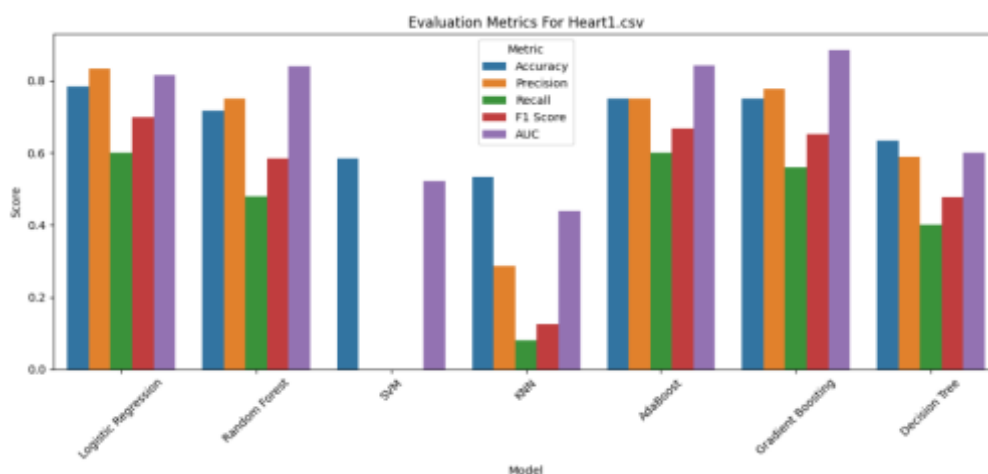


Fig 6: Evaluation Metrics for Various Models using Heart Dieses (using Dataset [30])

Kidney [31]

Table 6. Results obtained for Kidney Dieses (using Dataset [31])

Sr.No.	Model	Ac	Pr	Re	F1	AUC
1	LR	0.9625	0.962264	0.980769	0.971429	0.975962
2	RF	1.0000	1.0000	1.0000	1.0000	1.0000
3	SVM	0.9375	1.000000	0.903846	0.949495	0.989698
4	DT	0.9625	0.980392	0.961538	0.970874	0.962912
5	GB	0.9875	1.000000	0.980769	0.990291	1.000000
6	AB	0.9875	0.981132	1.000000	0.990476	1.000000
7	KNN	0.9375	1.000000	0.903846	0.949495	0.978022

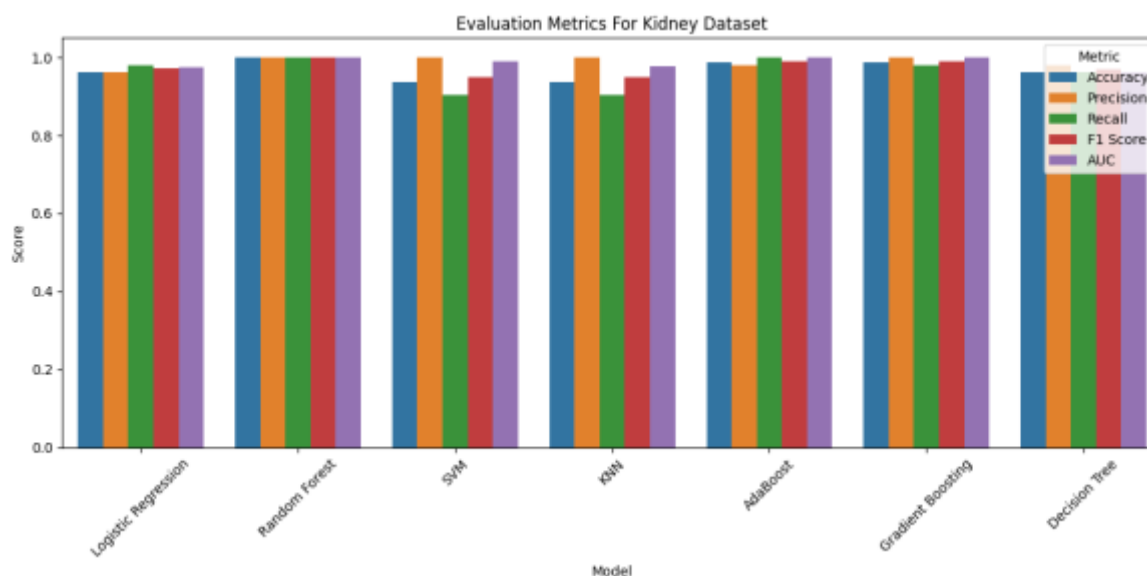


Fig 7: Evaluation Metrics for Various Models using Kidney Dieses (using Dataset [31])

Parkinson's [32]

Table7. Results obtained for Parkinson's Dieses (using Dataset [32])

Sr.No.	Model	Ac	Pr	Re	F1	AUC
1	LR	0.897436	0.888889	1.00000	0.941176	0.839286
2	RF	0.923077	0.939394	0.96875	0.953846	0.948661
3	SVM	0.846154	0.861111	0.96875	0.911765	0.678571
4	DT	0.897436	0.937500	0.93750	0.937500	0.825893
5	GB	0.923077	0.939394	0.96875	0.953846	0.928571
6	AB	0.871795	0.909091	0.93750	0.923077	0.866071
7	KNN	0.820513	0.878788	0.90625	0.892308	0.698661

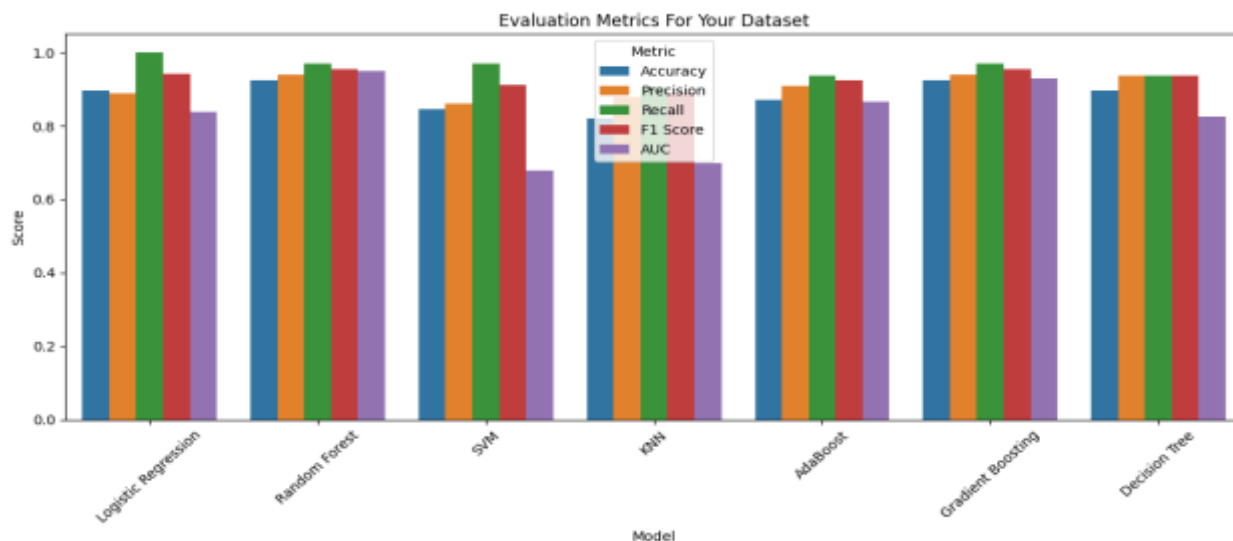


Fig 8: Evaluation Metrics for Various Models using Parkinson's Diseses (using Dataset [32])

BREAST CANCER [33]:

Table 8. Results obtained for Breast Cancer (using Dataset [33])

Sr.No.	Model	Ac	Pr	Re	F1	AUC
1	LR	0.929825	0.956522	0.929577	0.942857	0.988208
2	RF	0.947368	0.985075	0.929577	0.956522	0.993940
3	SVM	0.921053	0.887500	1.000000	0.940397	0.969538
4	DT	0.894737	0.940299	0.887324	0.913043	0.897150
5	GB	0.947368	0.957746	0.957746	0.957746	0.990501
6	AB	0.938596	0.984848	0.915493	0.948905	0.989519
7	KNN	0.938596	0.921053	0.985915	0.952381	0.951523

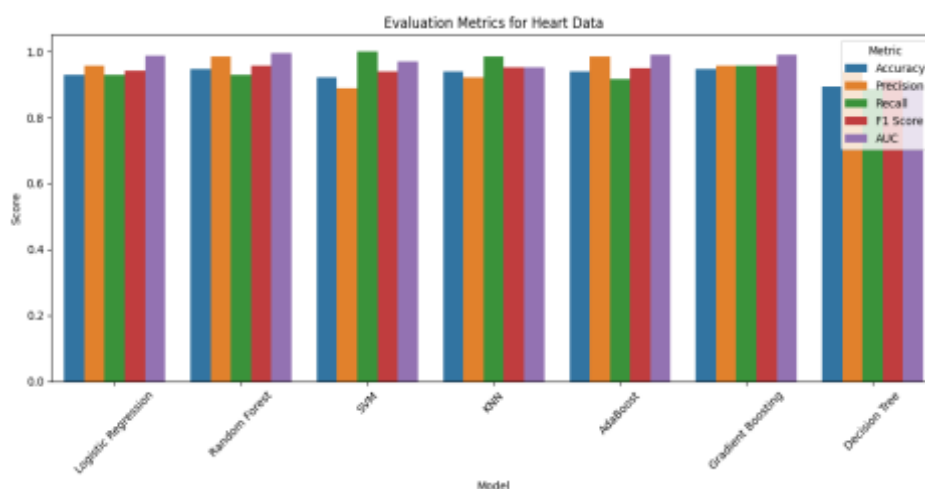


Fig 9: Evaluation Metrics for Various Models using Breast Cancer (using Dataset [33])

Table9. Accuracy of Various Models for Different Dieses

Models	Heart Disease	Diabetes	Kidney	Parkinson's	Breast Cancer	Spine	Liver
LR	0.809756	0.951891	0.9875	0.923077	0.982456	0.811111	0.655172
RF	1	0.970806	0.9875	0.923077	0.97076	0.822222	0.666667
SVM	0.926829	0.957237	0.925	0.923077	0.976608	0.788889	0.649425
KNN	0.863415	0.946957	0.9125	0.923077	0.959064	0.788889	0.66092
GB	0.97561	0.969161	0.975	0.923077	0.959064	0.855556	0.643678
DT	0.985366	0.949424	0.975	0.794872	0.935673	0.8	0.683908
AB	0.863415	0.970395	0.975	0.948718	0.976608	0.822222	0.666667

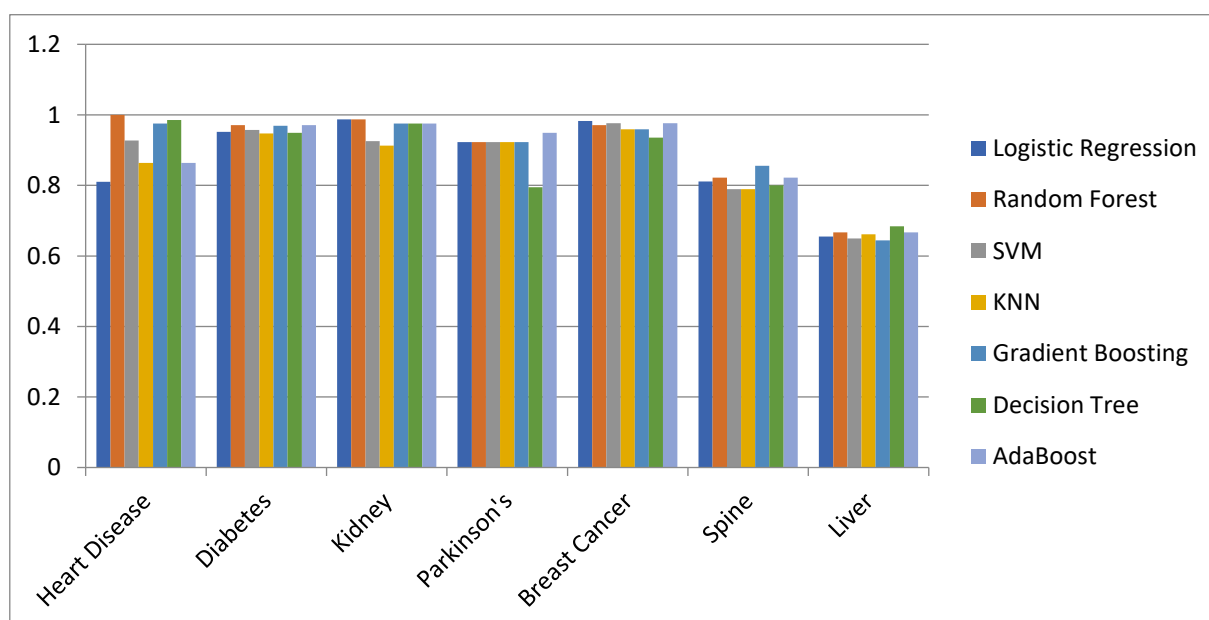


Fig 10: Evaluation Metrics of Accuracy of Various Models for Different Dieses

4. CONCLUSION

Compared to LR, SVM, KNN, GB, and DT, RF is the only technique that achieves 100% accuracy in predicting heart disease. RF also improves the accuracy of diabetes disease prediction at 97%. For the prediction of kidney disease, RF achieves an accuracy of 98%, which is also the best compared to the remaining techniques. The highest accuracy of 92% in the prediction of Parkinson's Disease is achieved by SVM. LR achieves the best accuracy (92%) in the prediction of breast cancer compared to SVM, RF, KNN, GB, AB, and DT. In the prediction of Spin data, the highest accuracy (85%) is achieved by GB. For the prediction of liver data, an accuracy of 85% is achieved by the DT, which is also the best compared to the remaining techniques. The input data of liver and spine diseases contained a greater number of null values as compared to the other datasets, which explains the poor prediction accuracy of the other techniques.

ACKNOWLEDGEMENTS

The authors acknowledge the dataset provided by the <https://www.kaggle.com/>.

AUTHOR CONTRIBUTIONS

Ramdas Pandurang Bagawade: Conceptualization, Investigation, Formal Analysis, Laboratory Analysis, Methodology.

Dr. Thirupurasundari D.R: Supervision, Writing—Review & Editing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Ethical approval was not required for this study.

DATA AVAILABILITY STATEMENT

Data supporting the findings of this study are publically available from the kaggle website.

REFERENCES

- [1] Nidhi Kosarkar; Pallavi Basuri; Poonam Karamore; Prachi Gaw, "Disease Prediction using Machine Learning", International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22), 2022.
- [2] Sneha Grampurohit and Chetan Sagarnal, "Disease Prediction using Machine Learning Algorithms", International Conference for Emerging Tehnology (INCET), 2020.
- [3] Raj H. Chauhan, Daksh N. Naik, Rinal A. Halpati, Sagarkumar J. Patel and A.D. Prajapati, "Disease Prediction using Machine Learning", International Research Journal of Engineering and Technology, vol. 7, no. 5, pp. 2000-2002, 2020.
- [4] Kedar. Pingale, Sushant. Surwase, Vaibhav. Kulkarni, Saurabh. Sarage and Abhijeet Karve, "Disease Prediction using Machine Learning", International Research Journal of Engineering and Technology, vol. 6, no. 12, pp. 2810-2813, 2019.
- [5] "Disease Prediction in Data Mining Techniques", International Journal of Computer Applications and Information Technology (IJCAIT), 2013.
- [6] "Disease Prediction System using Support Vector Machine and Multi-linear Regression", (IJIRCST), 2020.
- [7] S. Jadhav, R. Kasar, N. Lade, M. Patil and S. Kolte, "Disease Prediction by Machine Learning from Healthcare Communities", International Journal of Scientific Research in Science and Technology, pp. 29-35, 2019.
- [8] D. Dahiwade, G. Patle and E. Meshram, "Designing disease prediction model using machine learning approach", Proceedings of the 3rd International Conference on Computing Methodologies and Communication ICCMC 2019, no. Iccmc, pp. 1211-1215, 2019.
- [9] S. Uddin, A. Khan, M. E. Hossain and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction", BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1-16, 2019.
- [10] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction", 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1-4, 2018.
- [11] K. Pingale, S. Surwase, V. Kulkarni, S. Sarage and A. Karve, Disease Prediction using Machine Learning, 2019.
- [12] Satyabhama Balasubramanian and Balaji Subramanian, "Symptom based disease prediction in medical system by using Kmeans algorithm", International Journal of Advances in Computer Science and Technology, vol. 3.
- [13] Gaurav Shilimkar, "Disease Prediction using Machine Learning", International Journal of Scientific Research in Science and Technology, vol. 8, no. 3, pp. 551-555, 2021.
- [14] "Decision Tree based Health Prediction System", International Journal for Research in Applied Science and Engineering Technology (IJRASET), 2020.
- [15] "Disease Prediction and Hospital Recommendation using Machine Learning Algorithms", International Journal for Research in Applied Science and Engineering Technology (IJRASET), 2020.
- [16] "A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Lives", Turkish Journal of Computer and Mathematics Education, vol. 12, no. 6, pp. 4013-4023, 2021.
- [17] <https://www.kaggle.com/datasets?fileType=csv>

- [18] https://en.wikipedia.org/wiki/Gradient_boosting
- [19] R.P.Bagawade, R Manda, J. Gavith, V. Karanjakar, A. Lohar “Heart Disease Prediction Using Deep Learning”. IRJMETs, Vol. 06, Issue.5, pp: 8186-8191, May 2024.
- [20] R.P.Bagawade, D.R Thirupurasundari, “A Review Data Pre-Processing Techniques in Machine Learning”. In Proceedings of Hinweis Second International Conference , pp: 3937-3946, RTET, (2024) <https://www.scopus.com/pages/publications/105034156803>.
- [21] R.P.Bagawade, D.R Thirupurasundari, “Disease Prediction Using Machine Learning”. JETM , Vol. 74, no.1, pp: 1-9, Dec 2024
- [22] R.P.Bagawade, D.R Thirupurasundari, “A Review of Feature Selection Techniques for Machine Learning Algorithms used for Disease Prediction”. In Proceedings of ICRTSET Fifth International Conference , pp: 330-335, (2025).
- [23] Ramdas Bagawde, Sanket Pehera, Shubham Gadekar, Utkarsh Kulkarni, “Liver, Kidney, Heart and Diabetes Disease Prediction using Machine Learning”. International Journal of Innovative Research in Computer and Communication Engineering , Vol. 13, Issue 03, pp: 2728-2734, March 2025
- [24] R.P.Bagawade, D.R Thirupurasundari, “Ensemble Machine Learning Techniques”. JJETAEE , Vol. 15, Issue.4, pp: 15-23, Apr 2025
- [25] <https://medlineplus.gov/diabetescomplications.html>
- [26] <https://www.kaggle.com/datasets/mathchi/diabetes-data-set>
- [27] <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>
- [28] <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>
- [29] <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>
- [30] <https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset>
- [31] <https://www.kaggle.com/datasets/mansoordaku/ckdisease>
- [32] <https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-set>
- [33] <https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset>
- R.P.Bagawade, D.R Thirupurasundari, “Heart, Liver, Lungs, Diabetes and Kidney Disease Prediction using Machine Learning and Voting Ensemble Technique”, ICSFT 2026 International IEEE Conference 24-25 April 2026, DOI: [10.1109/ICSFT67370.2026.11566293](https://doi.org/10.1109/ICSFT67370.2026.11566293)