International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 04 | Apr 2025 www.irjet.net p-ISSN: 2395-0072

EVALUATING THE EFFECTIVENESS OF MACHINE LEARNING MODELS IN DISEASE PREDICTION

Punithavathi Arikrishnan[1] and Dr. Padmapriya Arumugam[2]

[1] Research Scholar, Department of Computer Science, Alagappa University, Karaikudi, Sivaganga, Tamil Nadu, 630003, India.
[2] Professor & Head, Department of Computer Science, Alagappa University, Karaikudi, Sivaganga, Tamil Nadu, 630003, India.

ABSTRACT- *The ability to accurately predict health care data outcomes is vital for successful disease management. The purpose of this work is to undertake a complete evaluation of several supervised machine learning algorithms for predicting health outcomes using four distinct datasets: Polycystic Ovary Syndrome (PCOS), Breast Cancer, Heart Disease, and Lung Cancer. The research includes preprocessing the data, collecting the most significant features, and training seven prominent machine learning models: Random Forest, K-Nearest Neighbors (KNN), Gaussian Naive Bayes, Logistic Regression, CatBoost, XGBoost, and LightGBM. The models were tested using important performance metrics such as accuracy, precision, recall, and F1score. Additionally, the training and prediction times of the models were also considered to assess their practical applicability. The results showed that among all the models, Logistic Regression consistently outperformed the other models across the datasets, achieving the highest average accuracy, precision, recall, and F1score. The model also exhibited relatively low training and prediction times, making it a promising choice for real-world healthcare applications. Other top performing models included Random Forest and XGBoost, which also exhibited strong predictive capabilities. The findings of this study provide valuable insights into the comparative strengths and weaknesses of various ML algorithms for predicting health outcomes. The results can inform the selection of appropriate models for specific healthcare tasks, ultimately contributing to the development of more accurate and efficient decision support systems in the healthcare area*.

Keywords: Machine Learning, PCOS, Heart disease, Breast cancer, Lung cancer, Logistic Regression

1. INTRODUCTION

Machine learning, which is a branch of artificial intelligence, is developed to train algorithms on patterns in previous data so that it can predict the future and make decisions on its own without being explicitly programmed. Its application in healthcare has completely revolutionized various approaches toward the prediction of diseases by different researchers and practitioners, although developing predictive tools that will be most useful for automation of diagnosis procedures. Thereby, the use of machine learning in the prediction of diseases ultimately has the potential to ease the burden on health-care practitioners through increased accuracy in the diagnoses, reduced costs, and timely interventions Precise prediction of the course of diseases has now become one of the most important tasks. The potential of insights from data within the healthcare sector opens new ways for innovation; hence, this has an increasing interest in utilizing machine learning for various purposes. This work focused on the various machine learning algorithms used for predicting the outcome of four major diseases: Polycystic Ovary Syndrome, breast cancer, heart disease, and lung cancer. The chosen four have been selected because these represent a wide spectrum of disorders that differ in their complexity, prevalence, and importance to public health. For instance, PCOS is one of the most common endocrine disorders among reproductive women, but its sophisticated pathophysiology and numerous symptoms have always made its treatment ignored. Meanwhile, breast cancer has not stopped being a common kind of cancer in all parts of the world, the survival rate being largely affected by the time of its detection. Heart diseases and lung cancers have historically been in the foreground, both being leading causes of morbidity and mortality. Since machine learning algorithms can handle huge and complex data with different data types, such as numerical, categorical, and unstructured data, they are found to be useful in disease prediction. General uses of datasets in healthcare include EHR, clinical measures, imaging data, genetic data, and patient demographics. It will be quite difficult to ensure that the derived patterns in this data are meaningful in order to predict the right state or stage of the disease. Machine learning algorithms generally learn a relation between input features and predefined outcomes by training the models over the labeled data. Data preprocessing and feature selection are the two major process that need to be done before machine learning models applied. One of the significant tasks in machine learning is the method of choosing the features known as Feature Selection, in a dataset that is most relevant and useful for making correct predictions. It is, therefore, very

important to find a subset of the most informative features to establish reliable prediction models. After the preprocessing and feature selection, these datasets were used to train machine learning models. During training, the models learn the clinical characteristics in the dataset that govern a variant disease outcome. All the algorithms learned have a different mode of learning. Within these years, many algorithms of machine learning have been suggested and enhanced. This paper introduces the results for seven models that are Logistic Regression, CatBoost, XGBoostRF, Random Forest, KNearest Neighbors, Gaussian Naive Bayes, and LightGBM. In this paper, accuracy, precision, recall, and F1score are the important metrices to provide the effective prediction of disease outcomes by comparing these models used four different disease datasets to train. These performances were evaluated using the test datasets not being exposed in any training process. In addition, in this paper, training and prediction times shall also be considered in an effort to appraise applicability in clinical environments where making timely decisions is critical. The results of this study provide valuable insights on the relative advantages and disadvantages of various machine learning models with respect to disease prediction.

2. BACKGROUND STUDY

Asif et al. [5] cardiovascular disease prediction research has explored into a various machine learning techniques to enhance diagnosis accuracy. Previous research used algorithms such as J48, KNearest Neighbor, and Random Forest, yielding results ranging from 56.76% to 87% accuracy. Ensemble methods, such as hard voting classifiers, have showed great promise, reaching up to 90% accuracy in some circumstances. This study contributes by testing twelve machine learning algorithms, which resulted in 92% accuracy using ensemble voting classifiers, highlighting machine learning's potential to improve predictive accuracy in cardiovascular disease. Ali et al. [4] discussed about cardiac disease prediction using a numerus supervised machine learning technique. Using a Kaggle dataset, assessed the performance of techniques . The study discovered that the RF algorithm outperformed with highest in metrices, making it a viable tool for early stage of heart disease prediction. The work enhances the ability of machine learning to improve clinical decision making and reduce misdiagnosis rates. Sawhney et al. [16] conducted a correlation of various artificial intelligence models for the early detection of chronic kidney disease (CKD). The study compares the models in terms of random forests , support vector machines and deep neural networks , concentrating on accuracy, specificity , and sensitivity. The DNN model outperformed all of the other models, with the highest accuracy. This study highlights the ability of AI in improving early detection of CKD. Khanam and Foo[12]did an analogy of various machine learning methods for diabetes prediction. Using the PIMA Indian Diabetes dataset, this research investigated techniques such as support vector machine (SVM), logistic regression (LR), knearest neighbors (KNN), decision tree (DT), and random forests (RF). The study discovered that the RF algorithm surpassed others in performance metrics. ALAM SUHA [3] investigated the prediction of Polycystic Ovary Syndrome (PCOS) using machine learning approaches based on patient symptom data and ovarian ultrasound images. This work assessed the various algorithms for predicting PCOS, including support vector machines (SVM) and convolutional neural networks (CNN). This study emphasized the CNN model's superior performance in reliably detecting PCOS from ultrasound pictures. This study combining machine learning with medical imaging to improve diagnosis accuracy and patient outcomes. Suha and Islam[19] did a comprehensive study of computer aided strategies for detecting polycystic ovary syndrome (PCOS). The study compared various machine learning algorithms, such as support vector machines (SVM), random forests (RF), and convolutional neural networks (CNN), concentrating on accuracy and efficiency. The analysis highlighted CNNs' potential for identifying PCOS from ultrasound images, as well as the importance of high-quality data and effective feature selection algorithms. Iftikhar et al. [11] analyzed the performance of various machine learning models for predicting chronic renal disease. The study utilized a dataset from the Buner district in Khyber Pakhtunkhwa, Pakistan. The SVM with the Laplace kernel function outperforms the other models based on performance metrices. Akkaya et al. [2] conducted a study on heart disease prediction by various machine learning models. The research evaluated algorithms such as logistic regression, support vector machine, knearest neighbors, decision tree, and random forests. The study found that the RF algorithm achieved the outperformed accuracy among the models. Hassan et al. [9] conducted a study on predicting chronic kidney disease (CKD) using machine learning techniques on patients' clinical data. The study evaluated various algorithms, including neural networks (NN), bagging tree models (BTM). The research pinpointed the Exemplary performance of the RF with high accuracy. Ahmad et al. [1] an analysis on the optimum medical diagnosis of heart disease using machine learning techniques, with and without sequential feature selection. The research evaluated algorithms such as linear discriminant analysis (LDA), RF, G B C , DT, SVM, and KNN. The analysis found that RF and DT with sequential feature selection achieved the highest accuracy, demonstrating the importance of feature selection in enhancing model performance. Chotrani [6] conducted an in depth study on various machine learning models for predicting disease. The study evaluated ml algorithms across multiple datasets. The research highlighted the superior performance of ensemble methods, particularly random forests (RF). Gupta and Gupta[8] conducted comparison on deep learning methods for predicting

breast cancer survivability. The study evaluated models such as artificial neural networks (ANN), Restricted Boltzmann Machines (RBM), Deep Auto encoders, and Convolutional Neural Networks (CNN). The RBM model achieved the highest accuracy, followed by Deep Autoencoders and CNN. Uddin et al.[20] did a comparative performance analysis of various Knearest neighbour (KNN) algorithm variants for disease prediction. The study evaluated different KNN variants, including , adaptive KNN, locally adaptive KNN, classic KNN, kmeans clustering KNN, mutual KNN, fuzzy KNN, Hassanat KNN, and generalized mean distance KNN. The research found that the Hassanat KNN variant achieved the highest average accuracy (83.62%), followed by the ensemble KNN (82.34%). Sharma and Mishra[17] conducted a performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. The research utilized three feature selection methods: correlation-based selection, information gain-based selection, and sequential feature selection. The ensemble-based Max Voting Classifier, combining the top three performing models, achieved an accuracy of 99.41%. Debal and Sitote[7] discussed the prediction of chronic kidney disease (CKD) using machine learning techniques. The study evaluated models such as random forests (RF), support vector machines (SVM), and decision trees (DT), focusing on both binary and multiclass classification for CKD stages. The research highlights the superior performance of the RF model, particularly when combined with recursive feature elimination and cross validation, achieving the highest accuracy among the tested models in terms of accuracy, precision, f1 score, recall, specificity and sensitivity. Ramesh et al. [14] conducted a study on predictive analysis of heart diseases using various machine learning approaches. The research evaluated algorithms such as Naive Bayes, support vector machine (SVM), logistic regression (LR), decision tree (DT), random forests (RF), and knearest neighbors (KNN). The study found that KNN with eight neighbors demonstrated superior performance in terms of precision, accuracy, effectiveness and sensitivity. Srikanth [18] explored the chronic kidney disease (CKD) prediction using various machine learning algorithms. The research highlighted the superior performance of the RF algorithm in terms of accuracy, f1 score, precision, recall, Jaccard score and log loss. Ibrahim and Abdulazeez [10]]this paper reviewed the seven machine learning algorithms for diagnosis from the medical database. It includes Decision Tree, Logistic Regression, Knearest neighbor, K means clustering, Naive Bayes and Random Forest. This review discovered that many algorithms showed good accuracy for predicting such as Decision Tree, KNN, Random Forest and SVM. Rawal [15]]explored the Breast cancer Prediction using four machine learning algorithms such as logistic regression LR, SVM, KNN, and RF. This research work involves seven phases in terms of Preprocessing data, Data Preparation, Feature Selection, Feature Projection, Feature scaling, Model selection and Prediction. The research highlighted the SVM model had best performance in terms of efficiency and effectiveness based on recall and accuracy. Li et al. [13] discussed different machine learning classifiers, including Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), KNearest Neighbor (KNN), Decision Tree (DT), and Artificial Neural Network (ANN) for heart disease prediction. The researcher utilized the feature selection algorithms like Minimal Redundancy Maximal Relevance (MRMR), Relief, Local Learning Based Feature Selection (LLBFS) , and Least Absolute Shrinkage Selection Operator (LASSO). Additionally, propose a novel feature selection algorithm Fast Conditional Mutual Information (FCMIM) . Key contributions of this research are, addressing feature selection challenges using both standard and novel algorithms. Demonstrating that the proposed FCMIMSVM combination outperformed than other models, achieving a classification accuracy of 92.37%. Highlighting the importance of selecting relevant features for increasing model accuracy and computational efficiency. The work concluded that the FCMIMSVM model is effective for heart disease diagnosis.

3. RESEARCH METHODOLOGY

This section outlined the research steps, as illustrated in this figure 1 . It includes the steps we took for data collection, preprocessing, feature selection, data modeling, evaluating the models and data visualization. Each step is designed to ensure accurate and reliability of this research. Following subsection provides the detailed description of each step.

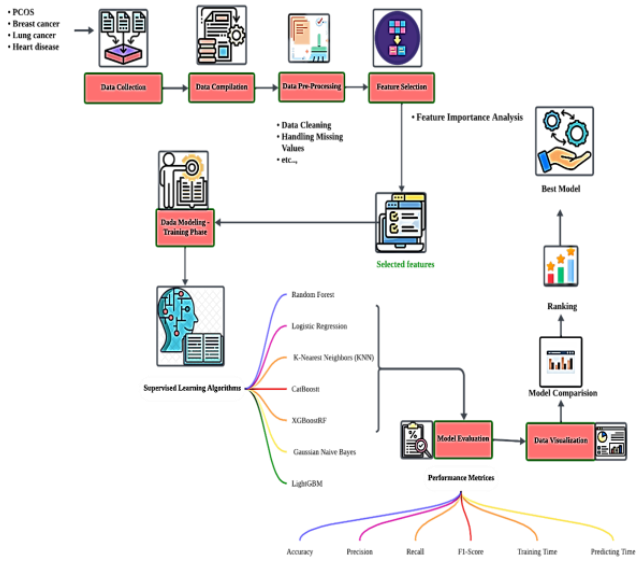

Figure 1: *Block Diagram Illustrating the Evaluation of Machine Learning Models for Disease Prediction*

3.1 Data Collection

In this section, gather the four different dataset which are used to disease prediction related to from the Kaggle repository. The four comprehensive datasets with specific medical conditions includes:

1. PCOS with and without infertility Dataset: It contains 541 rows about the patients data includes clinical measurements and PCOS status whether it is yes or no. Here there are two different datasets relevant to PCOS,
    - One is PCOS with Infertility dataset contains collection of data on patients with PCOS and their associated clinical measurements.
    - Another one is PCOS without Infertility dataset: This dataset includes similar clinical measurements but for patients without infertility.

    Finally, the two datasets are merged based on 'Patient File No.' field, and irrelevant columns are dropped to form a comprehensive dataset.
2. Breast Cancer Dataset: It has 4024 rows and 16 columns about clinical measurements related to breast cancer (BC) diagnosis.
3. Heart disease Dataset: This dataset consists of 303 rows and 14 columns filled with patient data, including age, sex, cholesterol levels, and various clinical characteristics linked to heart health.
4. Lung Cancer Dataset: This dataset provides the overview of potential contributors to cancer risk, involving 1000 rows and 26 columns.

3.2 Data Preprocessing

The initial step of data preprocessing involves thorough data cleaning to prepare the datasets for analysis. It includes:

- Handling Missing Values: Missing values were handled using median imputation, a robust method against outliers.
- Column Renaming: Column names are stripped of leading and trailing spaces to ensure consistency.
- Drop the Features: This removing the irrelevant features and ensures that no duplication existed in the datasets
- Data Transformation: Categorical features were transformed using one-hot encoding, converting them into binary (0 and 1) variables to facilitate machine learning.

3.3 Feature Selection

This section involves feature importance analysis for selecting the top-ranking features, as visualized in the figures below. Here, a Random Forest classifier (RF) is used to evaluate the importance of the various features in datasets (PCOS, BC, HD, LC). Feature Importance scores are computed, allowing for the selection of top-performing features that contribute effectively to predictive accuracy. This selection aims to reduce the dimensionality and enhance computational efficiency. Finally, a bar plot is generated to visualize the importance of the top features (see Figures 2, 3, 4, and 5).

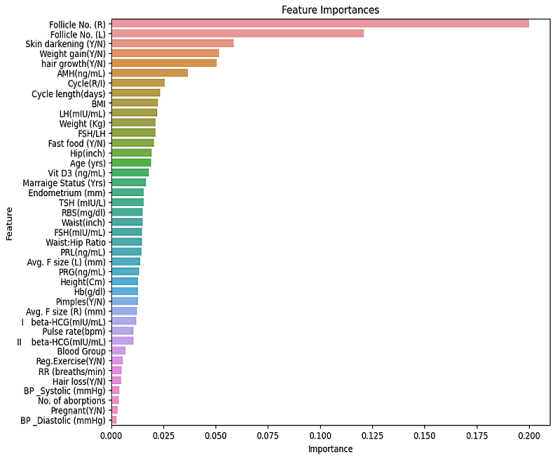

Figure 2: *Feature importance analysis for PCOS dataset*

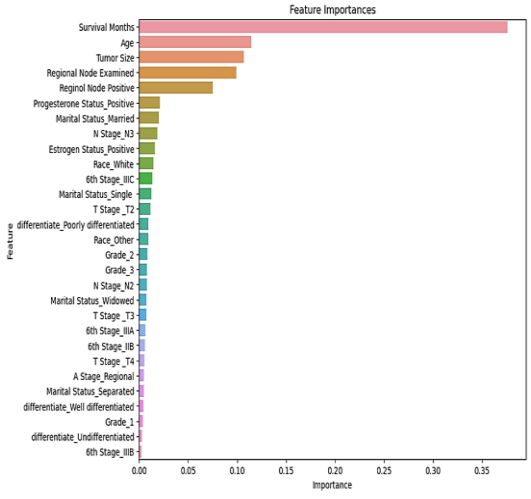

Figure 3: *Feature importance analysis for Breast Cancer Dataset*

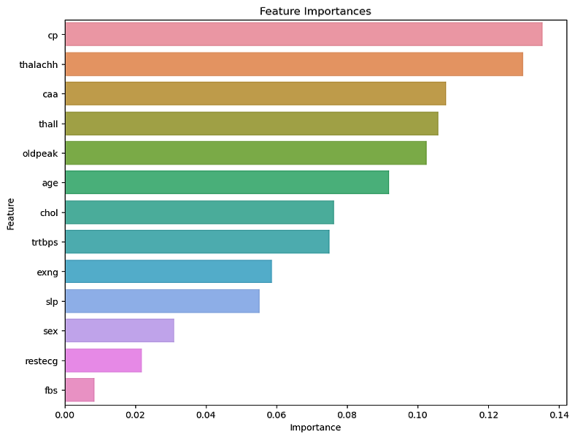

Figure 4 :*Feature importance analysis for Heart Disease Dataset*

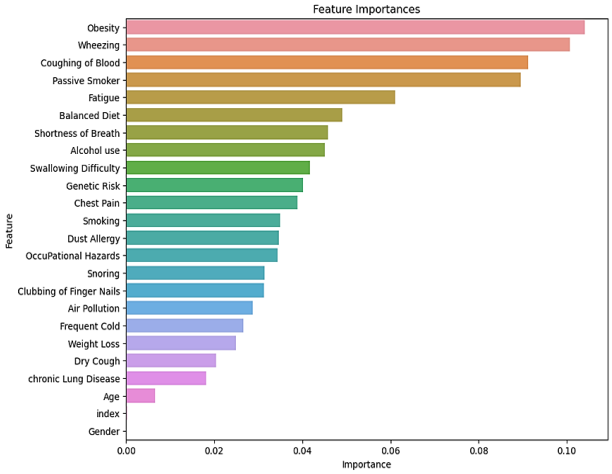

Figure 5: *Feature importance analysis for Lung Cancer Dataset*

3.4 Data Modeling

In the data modeling section, Multiple machine learning models are trained with labeled data in the dataset under supervised learning. Seven ML models are Logistic Regression, Gaussian Naive Bayes, Random Forest, K nearest Neighbor, CatBoost, XGBoostRF,Support Vector Machine, and LightGBM. Random Forest Here's a more detailed explanation of each machine-learning model:

1. **Random Forest**: In order to improve accuracy and manage overfitting, this ensemble learning model constructs several decision trees during the training phase and combines their output. A random subset of the data is used to train each tree, and a majority vote is used to determine the final classification prediction. This technique contributes to the model's increased robustness.
2. **Logistic Regression**:It is a statistical technique for binary classification that calculates the likelihood that an input belongs into a particular class. It squeezes a linear equation's output between 0 and 1 using the logistic function. Despite its name, it is a common choice for many classification problems since it is a linear model that is easy to implement and interpret.
3. **Gaussian Naive Bayes**: The Bayes theorem is applied by this probabilistic classifier under the presumption that characteristics are independent of class label. It makes the assumption that each feature's continuous values are distributed using a Gaussian distribution. This model is renowned for its simplicity and speed, and it works especially well with high-dimensional datasets.
4. **K Nearest Neighbor (KNN)**: A non-parametric learning method for regression and classification that is instance-based. A data point is classified by KNN according to the feature space's K nearest neighbors' majority class.
5. **Support Vector Machine (SVM)**: One supervised learning model that works well for both regression and classification applications is SVM. By determining which hyperplane in the feature space best divides the classes, it optimizes the margin between the nearest points of the classes. With a range of kernel functions, SVM can handle nonlinear data and performs well in high-dimensional domains.
6. **CatBoost**: An approach for gradient boosting that handles categorical information without requiring a lot of preprocessing. Ordered boosting is used to improve accuracy and decrease overfitting.
7. **XGBoostRF**: A gradient-boosted decision tree solution intended for efficiency and speed. The advantages of XGBoost and Random Forest are combined in XGBoostRF. Because of its efficacy and scalability, it is frequently utilized in real-world applications and machine learning contests.
8. **LightGBM**: Tree-based learning methods in a gradient boosting framework. LightGBM can handle massive datasets with less memory utilization because of its highly efficient and scalable design. LightGBM is renowned for its quickness and effectiveness in jobs involving both regression and classification.

The datasets are split, or train-test split, prior to the training phase. Each dataset's data was divided into two categories: training data (70 percent) and testing data (30 percent). Every machine learning model trained using training data.

3.5 Model Evaluation

The performance metrics were evaluated using the test datasets not being explored in any training phase in terms of accuracy, precision, F1-score, and recall (see Figures 6, 7, 8, and 9). Additionally, training time and predicting time of each were recorded.

- **Accuracy**: the proportion of cases that were accurately predicted to all instances.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- **Precision**: The quality of the positive class predictions is indicated by the ratio of true positive predictions to all predicted positives.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

- **Recall**: the proportion of actual positives to true positive predictions, highlighting the model's capacity to identify every relevant cases.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

- **F1 Score**: A balanced metric that takes consideration of both false positives and false negatives is the harmonic mean of precision and recall.

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision}+\text{recall}} \quad (4)$$

	Model	Accuracy	Precision	Recall	F1 Score	Training Time	Prediction Time
0	Random Forest	0.895706	0.891304	0.773585	0.828283	0.240647	0.016001
1	KNN	0.791411	0.731707	0.566038	0.638298	0.000000	0.015648
2	Gaussian Naive Bayes	0.852761	0.773585	0.773585	0.773585	0.000000	0.015605
3	Logistic Regression	0.907975	0.895833	0.811321	0.851485	0.140610	0.000000
4	CatBoost	0.889571	0.872340	0.773585	0.820000	4.890288	0.000000
5	XGBoost RF	0.889571	0.857143	0.792453	0.823529	0.093739	0.015623
6	LightGBM	0.889571	0.843137	0.811321	0.826923	0.093743	0.015622

Figure 6: *Model Evaluation for PCOS Dataset*

	Model	Accuracy	Precision	Recall	F1 Score	Training Time	Prediction Time
0	Random Forest	0.915563	0.910716	0.915563	0.907059	0.499919	0.031250
1	KNN	0.891556	0.880830	0.891556	0.882464	0.015648	0.062492
2	Gaussian Naive Bayes	0.828642	0.838777	0.828642	0.833305	0.000000	0.000000
3	Logistic Regression	0.899007	0.889772	0.899007	0.887629	0.281225	0.000000
4	CatBoost	0.908940	0.902164	0.908940	0.900120	5.937091	0.000000
5	XGBoost RF	0.914735	0.910032	0.914735	0.905644	0.249975	0.015624
6	LightGBM	0.902318	0.894033	0.902318	0.894652	0.156240	0.015623

Figure 7: *Model Evaluation for BC Dataset*

	Model	Accuracy	Precision	Recall	F1 Score	Training Time	Prediction Time
0	Random Forest	0.802198	0.802198	0.802198	0.802198	0.234357	0.015625
1	KNN	0.659341	0.657769	0.659341	0.657834	0.015624	0.000000
2	Gaussian Naive Bayes	0.824176	0.832115	0.824176	0.824558	0.015625	0.000000
3	Logistic Regression	0.824176	0.824176	0.824176	0.824176	0.078132	0.000000
4	CatBoost	0.813187	0.813747	0.813187	0.813369	3.162531	0.000000
5	XGBoost RF	0.802198	0.803646	0.802198	0.802534	0.062490	0.000000
6	LightGBM	0.736264	0.743820	0.736264	0.736838	0.046869	0.000000

Figure 8: *Model Evaluation for HD Dataset*

	Model	Accuracy	Precision	Recall	F1 Score	Training Time	Prediction Time
0	Random Forest	1.000000	1.000000	1.000000	1.000000	0.234340	0.015620
1	KNN	1.000000	1.000000	1.000000	1.000000	0.000000	0.015630
2	Gaussian Naive Bayes	0.916667	0.919779	0.916667	0.916365	0.000000	0.000000
3	Logistic Regression	1.000000	1.000000	1.000000	1.000000	0.078115	0.015623
4	CatBoost	1.000000	1.000000	1.000000	1.000000	2.406076	0.000000
5	XGBoost RF	1.000000	1.000000	1.000000	1.000000	0.156233	0.000000
6	LightGBM	1.000000	1.000000	1.000000	1.000000	0.171873	0.015614

Figure 9: *Model Evaluation for LC Dataset*

3.6 Data Visualization

Utilizing libraries such as Seaborn and Matplotlib, bar plots were generated to depict model comparison (MC) in terms of performance metrics (Accuracy, Precision, F1 Score, and Recall) of models effectively (see Figures 10, 11, 12, 13, 14, 15, 16, 17, and 18).

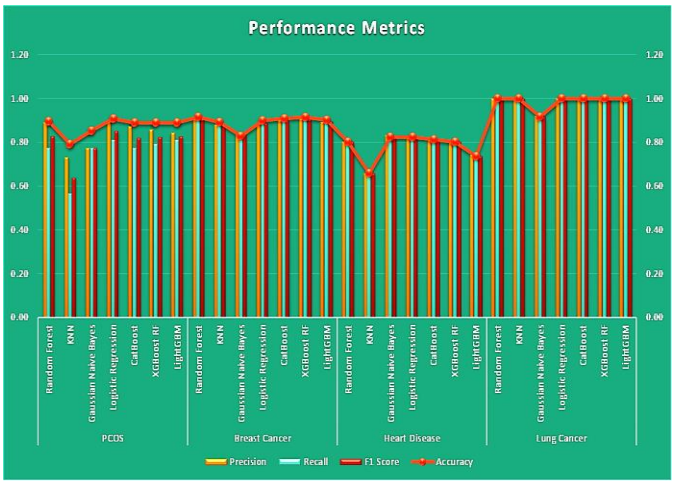

Figure 10: *Overall Performance Analysis of ML models*

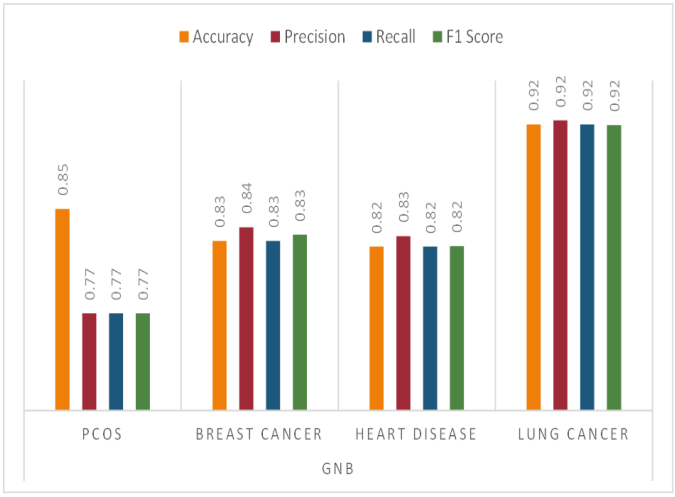

Figure 13: *Performance Analysis of GNB for 4 datasets*

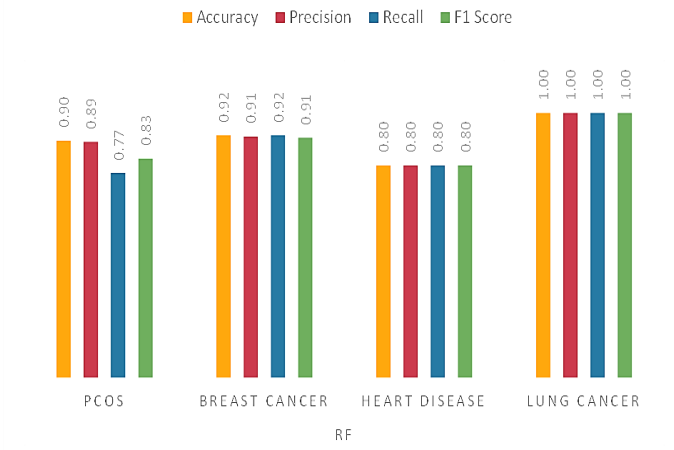

Figure 11: *Performance Analysis of RF for 4 datasets*

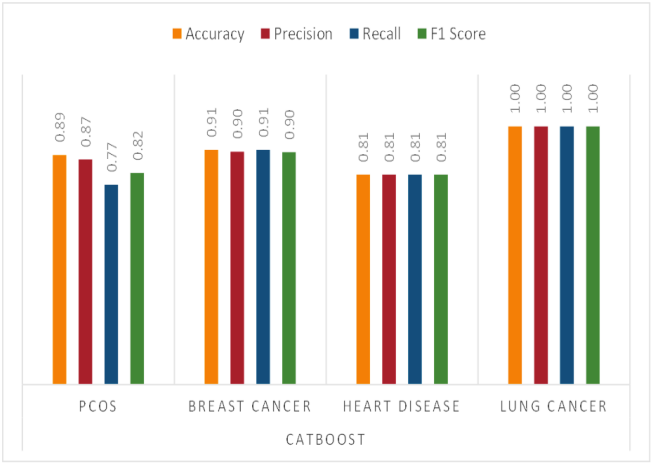

Figure 14: *Performance Analysis of CB for 4 datasets*

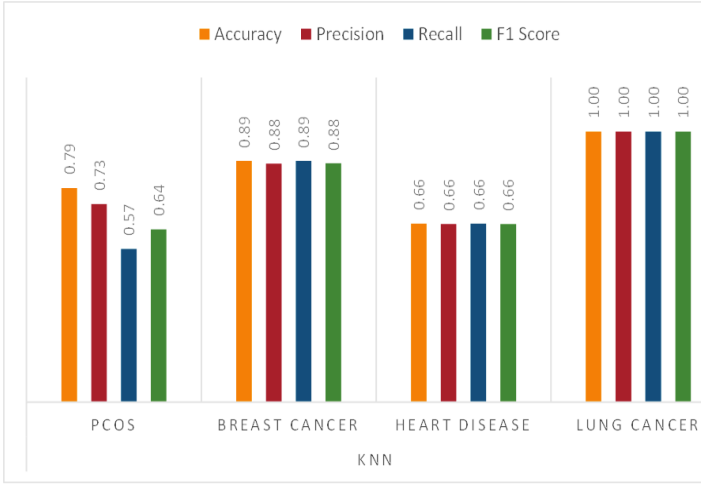

Figure 12: *Performance Analysis of KNN for 4 datasets*

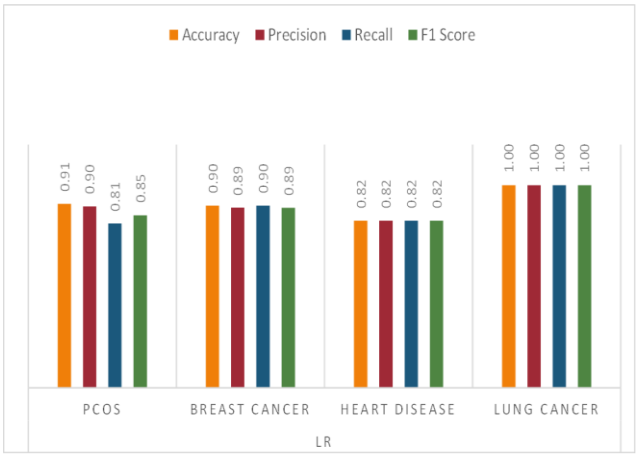

Figure 15: *Performance Analysis of LR for 4 datasets*

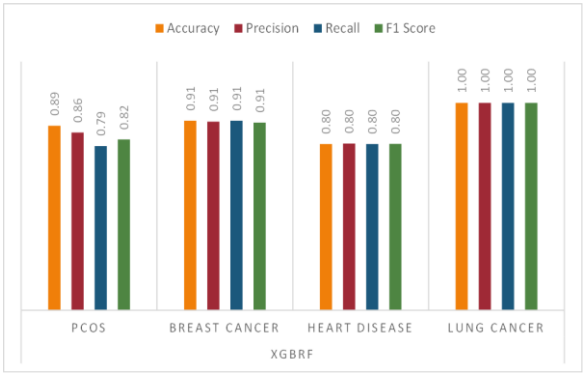

Figure 16: *Performance Analysis of XGBRF for 4 datasets*

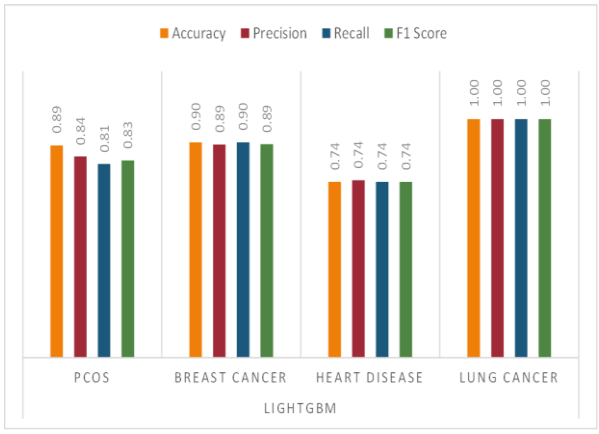

Figure 17: *Performance Analysis of LGBM for 4 datasets*

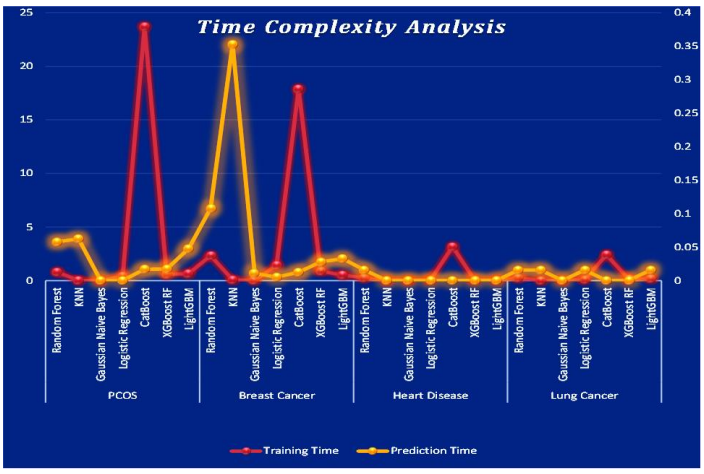

Figure 18: *Time Complexity Analysis of 4 datasets*

In figure 18, the primary axis is used to plot the training time, and at the same time secondary axis for plot the prediction time for better visualization. As shown in figures 19,20,21,and 22, Another bar plot was generated to visualize the rank of each model, that lower ranks indicate better performance .

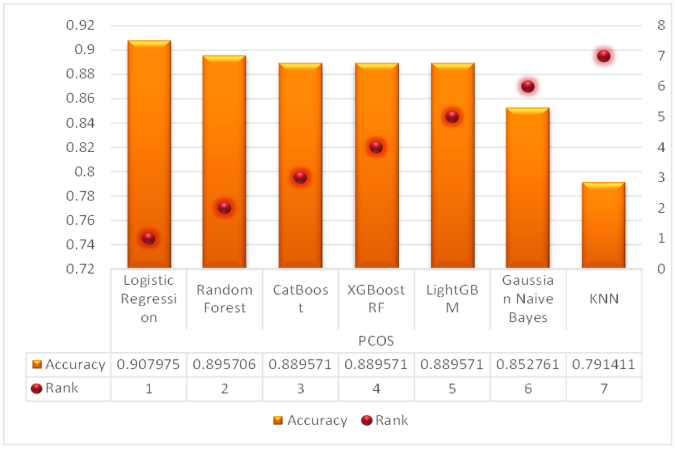

Figure 19: *Ranking of ML models based on performance in PCOS dataset*

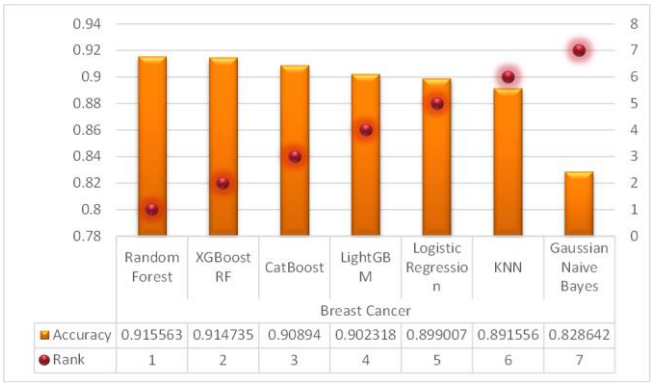

Figure 20: *Ranking of ML models based on performance in BC dataset*

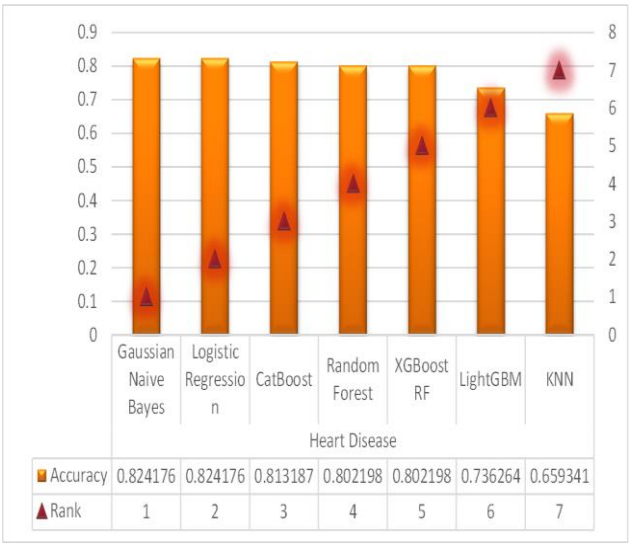

Figure 21: *Ranking of ML models based on performance in HD dataset*

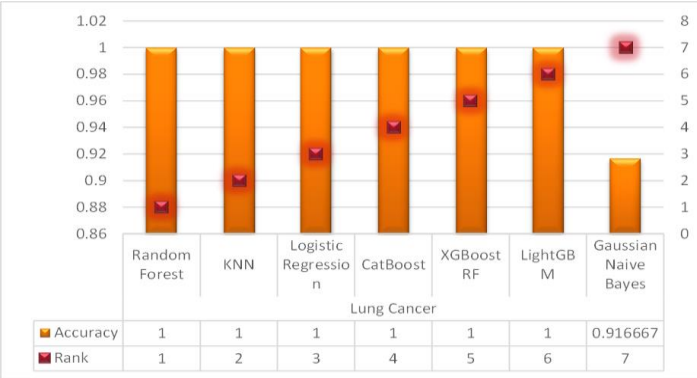

Figure 22: *Ranking of ML models based on performance in LC dataset*

In figures 19, 20,21,22 used secondary axis to plot the rank of ML models for clear visualization.

4 RESULT AND DISCUSSION

This research presents the performance of seven machine learning models on four different medical datasets: PCOS, Breast Cancer (BC), Heart Disease (HD), and Lung Cancer (LC). Additionally, each of the models was analyzed based on its performance metrics. In the case of the PCOS dataset, Logistic Regression reached an accuracy of 90.80%, with the highest F1 Score-85.15%-reflecting appropriate prediction and a balance between false negatives and false positives. Random Forest achieved 89.57% accuracy but exhibited slightly lower recall and F1 scores, indicating reduced effectiveness in identifying positive cases. Other models, KNN, Gaussian Naive Bayes, CatBoost, XGBoost, and LightGBM performed reasonably well, with their accuracy falling between 79.14% and 88.96%.

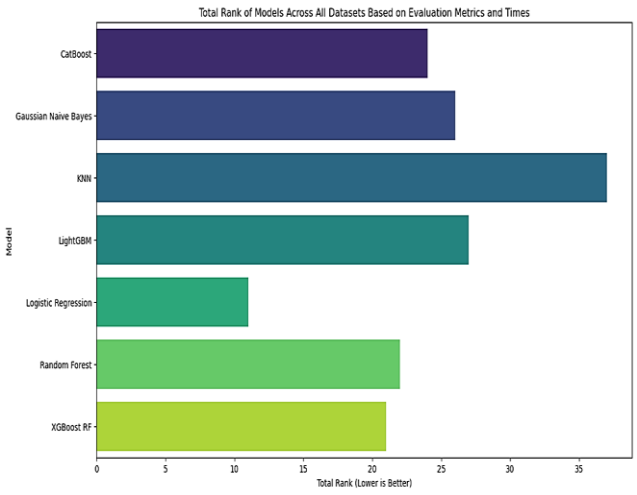

Figure 23: *Overall Ranking of ML models based on performance of four datasets*

RF performed better on the Breast Cancer dataset, having an accuracy of 91.56% and still possessing very good Precision and Recall values. Next came CatBoost, with 90.89%, but the very long training time made it less practical than RF. Logistic Regression also showed good results (89.90%), lagging behind these models. In the heart disease dataset, Gaussian Naive Bayes and Logistic Regression performed best, with each of them achieving an accuracy of 82.42%, while all the metrics were well-balanced. Though competitive, the results from Random Forest, CatBoost, and XGBoost in this problem not show good performance against simpler models, which showed satisfying results with less training time. The Lung Cancer dataset quite an exceptional case where all the metrics were perfect at 100% for the models Random Forest, KNN, CatBoost, and XGBoost. This means that this may be one of the easiest datasets to classify. Gaussian Naive Bayes, while not perfect, performed well with 91.67% accuracy, making it an effective alternative model. Overall,As shown in Figure 23, Logistic Regression was found to be quite consistent across all datasets. Especially for the PCOS dataset (see figure 19), this seemed to perform the best. Random Forest performed remarkably well in both the classification tasks of Breast Cancer and LungCancer, while Gaussian Naive Bayes turned out to be the best model for heart disease prediction. Whereas CatBoost and XGBoost were giving promising results, time and again they were found to be outperformed by simpler models computationally efficiently as well as in terms of accuracy, considering the training and prediction times. It reflects the fact that although complex models could result in high accuracy, simpler models like Logistic Regression and Random Forest yield better efficiency with good performance.

5. CONCLUSION

This study highlighted various benefits and drawbacks of different machine learning models applied to various medical datasets. Logistic Regression and Random Forest were consistently very versatile, with the performance of Logistic Regression better on the PCOS dataset, whereas Random Forest performed very well on the Breast Cancer and Lung Cancer datasets. Gaussian Naive Bayes shows an ideal balance between performance and computational efficiency on the heart disease dataset. The simple model may be used where speed and efficiency are required, such as Logistic Regression and Gaussian Naive Bayes. Complex models like Random Forest and CatBoost are suitable for scenarios requiring high accuracy, provided there is sufficient time for training. In summary, selecting the best machine learning model for medical prediction depends on balancing accuracy, computational efficiency, and dataset characteristics.

DECLARATIONS

CRediT Author Statement

Conceptualization: Padmapriya Arumugam; Methodology: Punithavathi Arikrishnan; Formal Analysis: Punithavathi Arikrishnan; Validation: Padmapriya Arumugam; Writing - original draft preparation: Padmapriya Arumugam and Punithavathi Arikrishnan; Writing - review and editing: Padmapriya Arumugam and Punithavathi Arikrishnan.

ACKNOWLEDGEMENT

This work was supported by the Alagappa University Research Fund (AURF)-Research Fellowship [vide Letter No. Rc.R2/Ph.D./R20223225/AURF Fellowship/2024, Alagappa University,Karaikudi,India,Date 28 November 2024], and I would like to thank my Supervisor Dr.Padmapriya A for her assistance throughout the research.

Conflicts of Interest

The authors don't have any conflict of interest.

Ethics Approval and Consent to Participate

This study did not involve human participants, and therefore, ethical approval and content are not applicable.

Consent for Publication

Consent to publish has been granted by each author.

Availability of Data

PCOS with and without infertility Dataset, Breast Cancer Dataset,Heart disease Dataset,Lung Cancer Dataset,(Accessed on 8th August 2024).

Abbreviations

- HD: Heart Disease
- PCOS: Polycystic Ovary Syndrome
- LC: Lung Cancer
- BC: Breast Cancer
- RF: Random Forest
- LR: Logistic Regression
- GNB: Gaussian Naive Bayes
- XGBRF: eXtreme Gradient Boosting Random Forest
- KNN: K-Nearest Neighbors
- LightBGM: Light Gradient Boosting Machine

REFERENCES

[1] Ahmad, G.N., et al., 2022. Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection. IEEE Access 10, 23808– 23828.

[2] Akkaya, B., Sener, E., Gursu, C., 2022. A comparative study of heart disease prediction using machine learning techniques, in: 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), IEEE.

[3] ALAM SUHA, S.A.Y.M.A., 2022. Predicting polycystic ovary syndrome through machine learning technique using patients' symptom data and ovary ultrasound images. Diss. Department of Computer Science and Engineering, MIST.

[4] Ali, M.M., et al., 2021. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine 136, 104672.

[5] Asif, M.A., et al., 2021. Performance evaluation and comparative analysis of different machine learning algorithms in predicting cardiovascular disease. Engineering Letters 29.

[6] Chhotrani, A., 2022. Comparative analysis of machine learning models for disease prediction. Journal of Science and Technology 3, 10–20.

[7] Debal, D.A., Sitote, T.M., 2022. Chronic kidney disease prediction using machine learning techniques. Journal of Big Data 9, 109.

[8] Gupta, S., Gupta, M.K., 2022. A comparative analysis of deep learning approaches for predicting breast cancer survivability. Archives of Computational Methods in Engineering 29, 2959–2975.

[9] Hassan, M.M., et al., 2023. A comparative study, prediction and development of chronic kidney disease using machine learning on patients' clinical records. Human-Centric Intelligent Systems 3, 92–104.

[10] Ibrahim, I., Abdulazeez, A., 2021. The role of machine learning algorithms for diagnosing diseases. Journal of Applied Science and Technology Trends 2, 10–19.

[11] Iftikhar, H., et al., 2023. A comparative analysis of machine learning models: a case study in predicting chronic kidney disease. Sustainability 15, 2754.

[12] Khanam, J.J., Foo, S.Y., 2021. A comparison of machine learning algorithms for diabetes prediction. Ict Express 7, 432–439.

[13] Li, J.P., et al., 2020. Heart disease identification method using machine learning classification in e-healthcare. IEEE Access 8, 107562–107582.

[14] Ramesh, T.R., et al., 2022. Predictive analysis of heart diseases with machine learning approaches. Malaysian Journal of Computer Science , 132–148.

[15] Rawal, R., 2020. Breast cancer prediction using machine learning. Journal of Emerging Technologies and Innovative Research (JETIR) 13, 7.

[16] Sawhney, R., et al., 2023. A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. Decision Analytics Journal 6, 100169.

[17] Sharma, A., Mishra, P.K., 2022. Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. International Journal of Information Technology 14, 1949–1960.

[18] Srikanth, V., 2023. Chronic kidney disease prediction using machine learning algorithms , 106–109.

[19] Suha, S.A., Islam, M.N., 2023. A systematic review and future research agenda on detection of polycystic ovary syndrome (pcos) with computer-aided techniques. Heliyon .

[20] Uddin, S., et al., 2022. Comparative performance analysis of k-nearest neighbour (knn) algorithm and its different variants for disease prediction. Scientific Reports 12, 6