Research Article

International Journal of Advanced Computer Research, Vol 12(61)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2021.1152067

Improving medical diagnostics with machine learning: a study on data classification algorithms

Abhishek Kumar* **and Sujeet Gautam**
Department of Computer Science, Patel College of Science and Technology, Bhopal, Madhya Pradesh

Received: 12-April-2022; Revised: 05-July-2022; Accepted: 10-July-2022

Abstract
This paper investigates the effectiveness of the logistic regression (LR) and random forest (RF) algorithms for classifying breast cancer using the Breast Cancer Wisconsin Dataset, consisting of 699 instances and 10 attributes. After pre-processing the data and performing feature extraction to retain relevant information, the dataset is split into training, validation, and test portions to evaluate the LR and RF algorithms. The LR algorithm achieves an accuracy level ranging from 96% to 97% across different split ratios, and its error rate decreases with larger training sets. The RF algorithm achieves an accuracy level ranging from 96% to 98% across different split ratios. The results indicate that both algorithms are effective for classifying the data, and the figures highlight the impact of different split ratios on accuracy and error rate. Proper selection of the split ratio is essential for obtaining reliable results.

Keywords
LR, RF, Machine learning, Data selection.

1.Introduction
In the present scenario, data is being gathered, processed, and extracted in huge amounts across various fields such as healthcare, education, and business enterprises [1]. It is crucial to acquire this data in a meaningful way to enable effective data mining (DM) and machine learning algorithms for different purposes [2−5].

DM algorithms enable the extraction of meaningful insights from vast datasets. Examples of DM algorithms include association rule mining, clustering, and classification [6, 7]. Data clustering is a crucial operation required in the arrangement of data, wherein data is arranged in similar groups based on their properties. Widely used clustering algorithms include K-means and fuzzy c-means algorithms [8−10]. However, in complex scenarios, clustering alone may fail, and evolutionary algorithms such as ant colony optimization, particle swarm optimization, teaching learning-based optimization, and Cuckoo Search can be useful with soft computing techniques [11, 12]. Machine learning algorithms are also found to be useful in the extraction and categorization of data.

Examples of machine learning algorithms include support vector machine (SVM), logistic regression (LR), random forest (RF) naïve Bayes (NB), and K-nearest neighbor (KNN) [13−18].

Knowledge extraction is a vital aspect of data mining, where useful patterns or significance in raw data are identified [19]. This process of discovering useful patterns or significance in raw data is called knowledge discovery in databases. it provides cleaning of conflicting data, and data mining provides pattern, classification, visualization, and rule separation.

The objective of this paper is to apply machine learning algorithms like LR and RF on breast cancer data and compare the classification performances considering different performance measures like precision, recall, F1-score, sensitivity etc.

2.Literature review
The related work analysis from the previous paper has been performed in this section.

In 2001, Kononenko [20] presented a comparison of state-of-the-art systems representative of various categories of machine learning that are applicable to

*Author for correspondence

medical diagnosis. They discussed their utilities by considering two studies and found that the classifiers' decisions were beneficial for analyzing data intelligently. The discussed machine learning approaches included Naïve Bayes, neural network, and symbolic learning. Their second approach described the uses of machine learning in verifying unexplored phenomena required for complementary medicines.

In 2019, Kamra et al. [21] provided a brief assessment of the literature based on various machine learning and data mining approaches used in medical diagnosis assistance systems. They found that smart healthcare applications linked to the developed systems would grow in popularity.

In 2020, Xiang et al. [22] performed a review based on two criteria: unsupervised learning and semi-supervised learning. They further studied the combination of these learning methods and offered a summary of available research on techniques for building effective machine learning models for medical applications. They discussed overviews of medical big data, healthcare management systems, and existing machine learning algorithms for e-health.

In 2020, Juddoo and George [23] addressed how machine learning can be a powerful method for enhancing data quality, particularly in the context of big data, by identifying poor quality data related to insufficiency and inaccuracy. They collected data on EHR Products Used for Meaningful Use Attestation and used RapidMiner Studio to link the dataset as a local data repository. The statistics feature highlighted concerns with completeness in terms of missing values. They discussed the use of Bayesian isotonic regression, lp-norm regularization (SRSp), and cluster-based best match scanning (CBMS).

In 2020, Leung et al. [24] proposed a solution for processing and interpreting COVID-19 epidemiological data that combines data analytics and machine learning. The program uses OLAP and taxonomy to generalize some attributes for analysis and effectively aids users in better comprehending data on COVID-19 confirmed instances. Although this tool is designed specifically for machine learning and analytics of large epidemiological datasets, it can also be used for machine learning and analytics of large datasets in various other practical applications and services.

In 2021, Jayatilake and Ganegoda [25] discussed various machine learning methodologies and algorithms used in the healthcare industry for decision-making, as well as the use of machine learning in healthcare applications. They explored the knowledge that neural network-based deep learning techniques have excelled in the area of computational biology, aided by the powerful processing of contemporary sophisticated computers, and are widely used due to their high predicting accuracy and dependability.

In 2021, Tchito et al. [26] conducted research on machine learning methods for biomedical classification using the Spark programming model, with a focus on handling large datasets for biomedical image classification problems. They developed a workflow that includes all necessary steps for classifying biological images and found that SVM performs well for medium-sized datasets while deep learning is better suited for larger datasets. They suggested using Spark as the foundation for the workflow.

Also in 2021, Chahar [27] discussed the benefits of using analytical and computational techniques to improve healthcare informatics, with a focus on data mining, evolutionary algorithms, and machine learning methodologies. The study is useful for improving the internal and external components of the decision support system for better performance.

Finally, Mustafa and Azghadi [28] reviewed AutoML technology's tools and methods, particularly in the healthcare sector. They discussed the unique challenges of processing medical notes and reviewed relevant ML research for clinical notes. They concluded that significant research issues and challenges must be resolved to establish an AutoML platform for clinical notes.

In their 2021 study, Aldahiri et al. [29] demonstrated how popular machine learning algorithms have been utilized in the healthcare industry for prediction and classification tasks. They provided a comprehensive overview of current ML techniques and their application in IoT medical data. The authors also noted that several ML prediction algorithms have significant drawbacks, which must be considered when selecting the best strategy to forecast critical healthcare data based on the type of IoT dataset. According to their research, the KNN method is frequently used for classification and prediction tasks, but it may take some time to forecast results in

real-time applications. To improve prediction performance, some researchers have suggested integrating recurrent neural networks (RNN) with long short-term memory neural networks (LSTM).

In their 2021 paper, Vokinger et al. [30] discussed the potential for bias at various stages of the development process in the healthcare industry, including data collection and preparation, model construction, model review, and post-authorization deployment in clinical practice. They found that several methods, such as transparency regarding the chosen training datasets, mathematical de-biasing techniques, machine learning interpretability, and post-authorization monitoring, can help to reduce the risk of bias.

In 2021, Rafi and Shubair [31] attempted to provide a comprehensive overview of recent publications related to the prognosis of diabetes, Parkinson's disease, heart disease, and breast cancer. Their primary objective was to highlight the use of optimization methods in machine learning to develop effective prediction systems. They provided examples of classification of optimization, deep learning, and machine learning approaches. Additionally, they aimed to develop small solutions for intelligent decision-making systems utilizing optimization methods, and for this purpose, they analyzed the need for large datasets.

In 2021, Sun et al. [32] utilized data from the research group of Svetlana Ulianova to conduct an experiment aimed at predicting heart disease using various computational methods such as random forest, support vector machine, and logistic regression. They used the correlation coefficient method of feature extraction and applied five-fold cross-validation. The results indicated that support vector machine performed better than random forest and logistic regression with an ROC curve of 78.84%. The study highlighted the potential benefits of using machine learning techniques for disease prediction.

In 2022, Dhinakaran et al. [33] proposed a remote monitoring system that continuously monitors patients and alerts medical professionals when necessary, using cloud computing and other machine learning methods. The study provided an overview of remote patient monitoring systems, their components, and their benefits. The researchers specifically addressed wireless body area network (WBAN) issues in remote patient monitoring and proposed machine learning based healthcare system solutions for remote patient monitoring.

In 2022, Elyan et al. [34] examined at the most recent developments in computer vision technology as they were used in the field of medicine. They talked about the main issues with intelligent data-driven medical applications and Computer Vision. In order to solve complicated vision problems, such as medical image classification, shape and object recognition from images, and medical segmentation, they first critically analysed the body of literature in the computer vision domain. Secondly, they provided a thorough analysis of the numerous issues that are thought to be obstacles to advancing the study, creation, and use of intelligent computer vision approaches in actual medical settings and hospitals.

In 2022, Hinterwimmer et al. [35] identified the requirements for the efficient implementation of this unique technique by examining which predictions are already practicable using machine learning models in knee arthroplasty. To locate machine learning applications for knee arthroplasty, a thorough search of PubMed, the Medline database, and the Cochrane Library was carried out. Their search strategy identified 225 articles and out of which 19 articles were considered. Further, they considered methodological assessment: a modified Coleman Methodology Score (mCMS). They found the AUC median of 0.76 and the median mCMS was 65 (interquartile range, 40-80).

In 2022, Zhang et al. [36] presented advancements and difficulties that define machine learning for healthcare from a data-centric perspective. They explored the usage of more modern transformer models for handling larger datasets and boosting the modelling of clinical language, as well as deep generative models like Generative adversarial network (GAN) and federated learning as techniques to augment datasets for greater model performance. They discussed about the issues with the deployment of machine learning that are data-focused. They stressed the necessity to account for natural data shifts that can impair model performance as well as the importance of rapidly delivering data to machine learning models for timely clinical predictions.

In 2022, Severn et al. [37] proposed the framework on TCGA-GBM dataset available publicly and which is obtained from The Cancer Imaging Archive. They discussed a unique pipeline for explainable machine learning imaging that employs Shapley values and

radiomics data as tools to explain result predictions using intricate prediction models created using well-defined predictors from medical imaging. They find it beneficial to use the approach of model-agnostic explainer. It was identified as limitation of their work that some more relevant metrics were required for the quality explanation.

In 2022, Zhu et al. [38] considered 343 artificial intelligence/machine learning devices enabled with medical facility.. Cardiovascular (12.0%) and radiology (70.3%) medical specialist panels evaluated most of the devices. Since the middle of 2010, the rise of these devices has rapidly increased. Most devices (95.0%) were approved for sale under the 510(k) premarket notification method, and 69.4% of them were software-based medical devices (SaMD). The most frequent uses of the 241 radiology-related gadgets were for diagnostic support (48.5%) and picture reconstruction (14.1%). 20.5% of the 117 radiology-related diagnostic aids for breast lesion assessment and 14.5% for echocardiography cardiac function assessment were created. The most frequent uses of the 41 cardiology-related devices were hemodynamic and vital sign monitoring (26.8%) and arrhythmia diagnosis based on electrocardiography (46.3%). Their study was restricted to FDA-approved devices, which decreased the generalizability of the findings. Some of the recent analysis is shown in *Table 1*.

Table 1 Methodological analysis based on the results

S. No.	Author and year	Dataset	Method	Result	Advantage	Limitation
1	Kobashi et al. 2016 [39]	Recruited 52 osteoarthritis (OA) at Hyogo College of Medicine at Japan	Linear model, Generalized linear model with and without optimized and neural network	Achieved the correlation coefficient of 79%, 78%,79%, 79% and root mean square error of 3.43, 3.63, 3.44, 3.58 in linear model, Generalized linear model with and without optimized and neural network.	It is helpful in the prediction of post-operative knee kinematics.	They are a need to apply other features also.
2	Lu et al. 2017 [40]	Considered 987 dataset from Chung Shan Medical University Hospital Tumor Registry	Multivariate adaptive regression splines, RF, Extreme learning machine, SVM, C5.0	The C5.0 method is the best one for predicting ovarian cancer recurrence.	Ensemble learning is better to provide the classification accuracy than normal parts of machine learning.	There is a need to apply C5.0 method in hybrid scheme.
3	Pitoglou et al. 2018 [41]	A total of 131,872 hospitalisation records from 2000 to 2017 were used.	Support Vector Machine, LR, Gaussian NB, Deep Multilayer Neural Network and KNN	Achieved Highest AUROC of 78.5% and MCC of 0.57 in KNN whereas Gaussian NB gives the lowest AUROC of 70.8% and 0.431.	It lowers the obstacle to implementation.	There may be biasness in collected data as it is collected from only one hospital.
4	Reamaroon et al. 2018 [42]	401 patient cases among which 48 were positive and 353 were negative suffering from moderate hypoxia.	Applied LR, SVM (with class-weighted and uncertain labels), random forest, machine learning algorithms	Attained accuracy of 72.63%, 74.34%, 74.92%, 78.04%, 81.57% from LR, RF, SVM with class-weighted and support vector machine with uncertain labels.	It showed the Increased performance.	There is a need of reformulation in context to label uncertainty and privileged Information. Small sample size.
5	Liu et al. 2019 [43]	Autistic spectrum disorder screening, Breast cancer, coronary artery disease, Diabetic Retinopathy Debrecen, Fertility, Immunotheraphy, Chronic Kidney,	Applied Schlesinger-Kozinec (SK) algorithm based on Scaled Convex Hull-Based.	100% accuracy attained by KSK-S for Autistic spectrum disorder, coronary artery disease and Chronic kidney dataset. KSK achieved the highest accuracy of 99.85% in Chronic	KSK-S obtained the better performance in the prediction.	Need to apply KSK-S for the problems of multiclass.

S. No.	Author and year	Dataset	Method	Result	Advantage	Limitation
		Parkinson's Disease Detection and Spect Heart		Kidney dataset. LibSVM obtained the 100% accuracy for Chronic Kidney dataset.		
6	Chang et al. 2019 [44]	A Beijing cardiovascular hospital's hypertension database. They collected the 1357 cases.	Decision Tree, SVM, XGBoost, RF	XGBoost outperformed than decision Tree, RF, and SVM. XGBoost achieved the accuracy of 94.36%, F1 measure of 87.5% and AUC of 92.7%	Lower the cost. Increase the effectiveness of treatment and diagnostics.	Small sample size. To achieve a greater predictive effect, the feature selection approach needs to be improved.
7	Khushi et al. 2021 [45]	Two datasets were considered as: The Prostate, Lung, Colorectal, and Ovarian (PLCO) and National Lung Screening Trial (NLST)	LR, linear support vector regression (SVC), and RF	PLCO dataset: Obtained AUC of 71.34% in Logistic regression by using SMOTEENN method, AUC of 86.84% in random forest by using SMOTETomek and AUC of 71.26% in Linear SVC by using SMOTEENN. **NLST dataset:** Obtained AUC of 65.50% in Logistic regression, 88% in random forest and 65.50% in Linear SVC by using SMOTETomek method.	Reduce the damage brought on by misdiagnosis.	Need to combine imbalance approaches with more diversified classifiers to improve model prediction.
8	Yang et al. 2021 [46]	Collected the medical records 1256 with 21 variables of Second Affiliated Hos- pital of Shanxi Medical University.	Applied machine learning models: weighted RF, LR, SVM, and weighted SVM	Achieved higher G-means of 82%, AUC of 82% and F-measure of 46%,	The issue of data imbalance in liver cirrhosis with Hepatic encephalopathy (HE) was resolved.	The study is restricted to HE that emerged during that time. HE cannot be predicted while the disease is still in progress. It had some missing data. The data might not accurately reflect the entire HE population. The model needs to be explored further on other datasets.
9	Bharti et al. 2021 [47]	Used the four datasets of Hungary, leveland, Long Beach V and Switzerland,	Used Deep Learning, Logistic regression, support vector machine, Decision tree, K-nearest neighbours, random forest,	Attained accuracy of 83.3%, 84.8%, 83.2%, 80.3%, 82.3%, 94.2% Specificity of 82.3%, 77.7%, 78.7%, 78.7%, 78.9%, 83.1% Sensitivity of 86.3%, 85.0%, 78.2%, 78.2%, 78.5%, 82.3% in Logistic regression, K-nearest neighbour, support vector machine,	Less computational time.	The dataset needs to be normalised. Small data size.

S. No.	Author and year	Dataset	Method	Result	Advantage	Limitation
				random forest, decision tree and deep learning		
10	Ram and Vishwakarma 2021 [48]	PIMA Indian dataset	RF, LR, SVM, KNN, Gradient Boosting Machine, NB	Achieved highest accuracy of 84.7% in Logistic Regression with nine features.	Not Found	It cannot determine which kind of diabetes a person has.
11	Khan et al. 2022 [49]	UCI: Adult database having 2,206 pieces data with a size of 5.5 MB. It contained 15 attributes	Incognito algorithm	The proposed model's privacy is superior to existing models.	Accelerating execution time. Improved data quality and the services.	incorrect search results
12	Urban et al. 2022 [50]	381 patients suffered from acute heart failure	Implemented K-medoids algorithm for creating the clusters	six patient phenotypes with AHF having p=0.002 of one year mortality	Consistent and heterogeneous clusters.	Restricted clinical parameters. Collected data had missing values. Small data size. Lack of external validation.
13	Lee et al. 2022 [51]	Data from 685,225 blood culture tests were retrieved from Seoul, Republic of Korea's Sinchon and Gangnam Severance Hospitals, which are connected with Yonsei University.	Gradient boosting algorithms, random forest, multi-layer perceptron was Applied.	Achieved AUROC of 76.2% for 12-hour data and 75.3% for 24-hour data in multi-layer perceptron.	Might enhance the clinical prognosis in actual clinical practise.	Use of historical data from electronic health records. So, need to compare it with real data. Data taken from numerous different centres must undergo external validation.
14	Ahmad et al. 2022 [52]	Heart Disease UCI Kaggle Dataset, Cleveland, Hungary, Switzerland, and Long Beach V	Extreme Gradient Boosting classifier without and with GridSearchCV, Support vector machine, K-nearest neighbour, Logistic regression	Achieved higher training and testing accuracy of 99.03% and 100% in Extreme Gradient Boosting Classifier with GridSearchCV.	NA	There is a need to apply GridSearchCV with a Gradient Boosting Classifier.
15	Dong et al. 2022 [53]	The electronic medical records (EMR) database available at the People's Liberation Army (PLA) General Hospital, North China. The data was collected from 816 T2DM patients (585 men)	They considered Seven machine learning techniques: decision tree, support vector machine, logistic regression, light gradient boosting machine [LightGBM], extreme gradient boosting, artificial neural network, adaptive boosting.	The highest AUC was for the LightGBM model 0.815.	In contrast to data collected from clinical trials, they used RWD generated from EMR, which is probably more representative of the heterogeneous T2DM patient group. They discovered risk factors that have not previously been linked to a higher incidence of Diabetic kidney disease (DKD).	The collected data have potential bias.
16	Tanioka et al. 2022 [54]	They considered 51 of the 930 patients in the development and	They applied logistic regression, k-nearest neighbors (k-NN)	K-NN achieved the highest AUC of 0.790, sensitivity of	The model was developed using computerized tomography images	To obtain more robust quality, the development and validation cohorts

S. No.	Author and year	Dataset	Method	Result	Advantage	Limitation
		71 of the 212 patients in the validation cohort.	algorithm, XGBoost, support vector machines (SVMs) and random forests	0.846, specificity of 0.733 and accuracy of 0.775.	from multiple vendors and multiple centres.	need more patients. computerized tomography results were assessed by people.

3.Methods

The paper describes the use of two algorithms, LR and RF, for the classification of a disease using the Breast Cancer Wisconsin Dataset, which has 699 instances and 10 attributes. The purpose of the research is to investigate the effectiveness of these algorithms for medical data classification.

The first step in the experimentation process is data pre-processing, which involves transforming the raw data into a structured format that is free of errors and inconsistencies. This step is necessary to ensure that the data is suitable for further processing.

The next step is feature extraction, which is the process of reducing the amount of data while retaining the most relevant information. This step is essential in reducing the number of resources required for the analysis and speeding up the machine learning task, such as classification.

To test the effectiveness of the LR and RF algorithms, the dataset is split into three portions: training data, validation data, and test data. The first portion, training data, is used to fit the model, while the second portion, validation data, is used to check the unbiased evaluation of the model. The final portion, test data, is used for the final evaluation of the model.

The step-wise algorithm for LR is as follows:

- Data preparation: Collect the data and prepare it for analysis by cleaning, transforming, and organizing it in a suitable format.
- Select the predictor variables: Choose the predictor variables that will be used to predict the outcome variable. The selection process can be based on domain knowledge or statistical techniques such as correlation analysis.
- Define the outcome variable: Choose the outcome variable that will be predicted by the predictor variables. This variable should be categorical or binary.
- Create a training and testing dataset: Split the data into a training dataset and a testing dataset. The training dataset will be used to train the model, while the testing dataset will be used to evaluate the model's performance.
- Fit the logistic regression model: Use the training dataset to fit the logistic regression model. This involves estimating the model coefficients using maximum likelihood estimation.
- Evaluate the model: Use the testing dataset to evaluate the performance of the logistic regression model.
- Iterate and improve the model: If the model performance is not satisfactory, repeat steps 2-6 by adding or removing predictor variables or using different modeling techniques until the desired performance is achieved.
- Use the model for prediction: Once the logistic regression model is developed and validated, it can be used to predict the outcome variable for new data.

The step-wise algorithm for RF is as follows:

Step 1: Initialization: Initialize the number of decision trees to be constructed and the cycle criteria for termination.

Step 2: Random Sampling: Select k data points randomly from the training set.

Step 3: Tree Construction: Gather the selected data points from the associated tree and build the decision tree.

Step 4: Feature Selection: Consider a random number of features for the decision tree construction. This step ensures that each decision tree is constructed using a different set of features, which helps in reducing the correlation among the trees.

Step 5: Repeat: Steps 2 to 4 will be repeated until the cycle criteria for termination are met. This ensures that multiple decision trees are constructed using different data points and features.

Step 6: Prediction: To predict the class of a new data point, the algorithm uses all the decision trees to vote for the final class. The class that gets the highest number of votes is assigned as the final prediction.

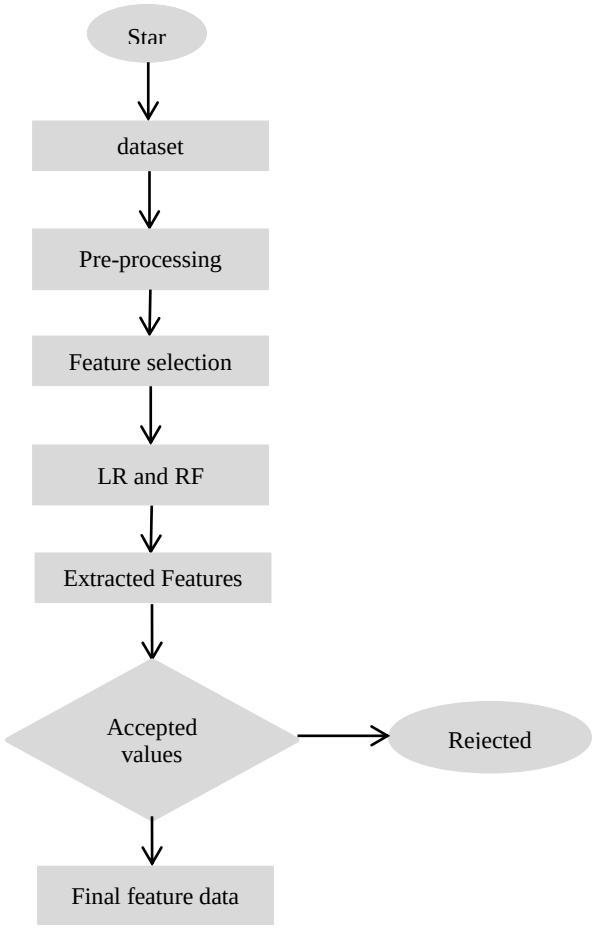

Figure 1 Working procedure of the complete approach

4.Results and discussion

Figure 2 displays the accuracy comparison of the LR algorithm across different split ratios. The accuracy level refers to the percentage of correct predictions made by the model. It shows that the accuracy level varies between 96% to 97% across different split ratios, indicating that the LR algorithm is effective in classifying the data. On the other hand, *Figure 3* displays the error rate comparison of the LR algorithm based on different split ratios (70-30 (S1); 75-25 (S2) and 80-20 (S3)). The error rate refers to the percentage of incorrect predictions made by the model. Root mean squared error (RMSE) and mean absolute error (MAE), were considered for error rates. Both are metrics used to measure the difference between predicted and actual values in a regression analysis. The figure shows that the error rate decreases as the split ratio increases, indicating that the LR algorithm performs better with a larger training set. Overall, the result suggests that the LR algorithm is effective in classifying the data with an accuracy level of 96% to 97%. The figures demonstrate the impact of different split ratios on the accuracy and error rate of the algorithm, highlighting the importance of selecting an appropriate split ratio for training and testing the model. *Figure 4* displays the accuracy comparison of the RF algorithm across different split ratios. The accuracy level refers to the percentage of correct predictions made by the model. It shows that the accuracy level varies between 96% to 98% across different split ratios, indicating that the RF algorithm is effective in classifying the data.

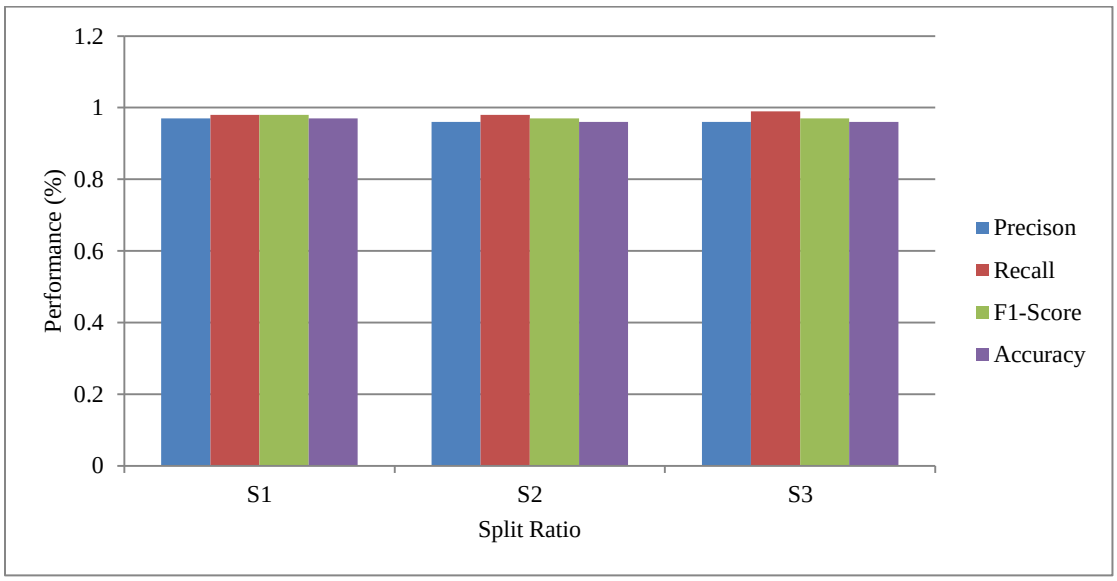

Figure 2 LR based result comparison considering precision, recall, F1-score, and accuracy

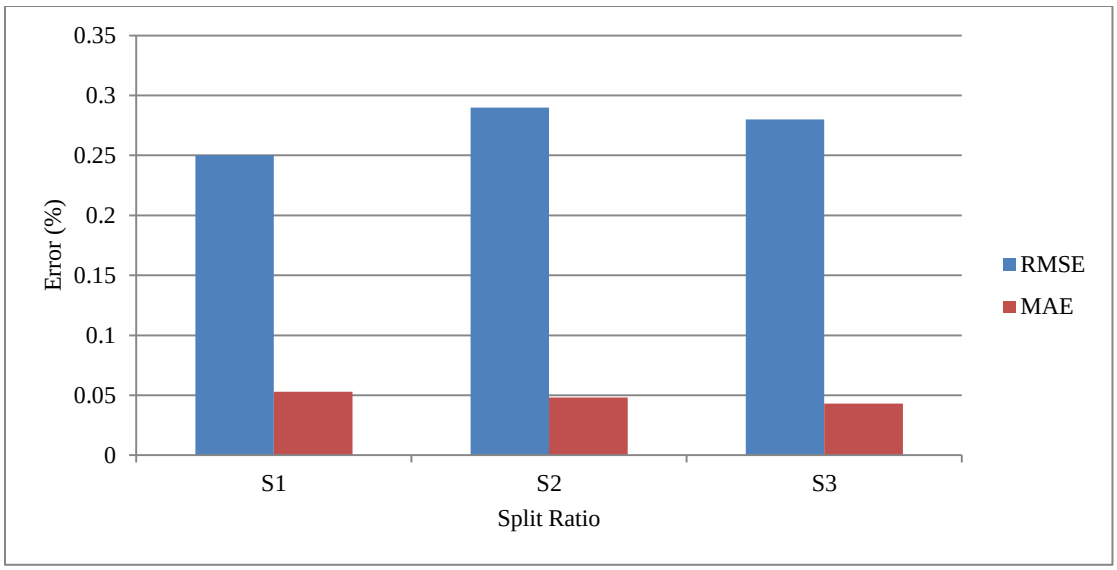

Figure 3 LR based error rate comparison

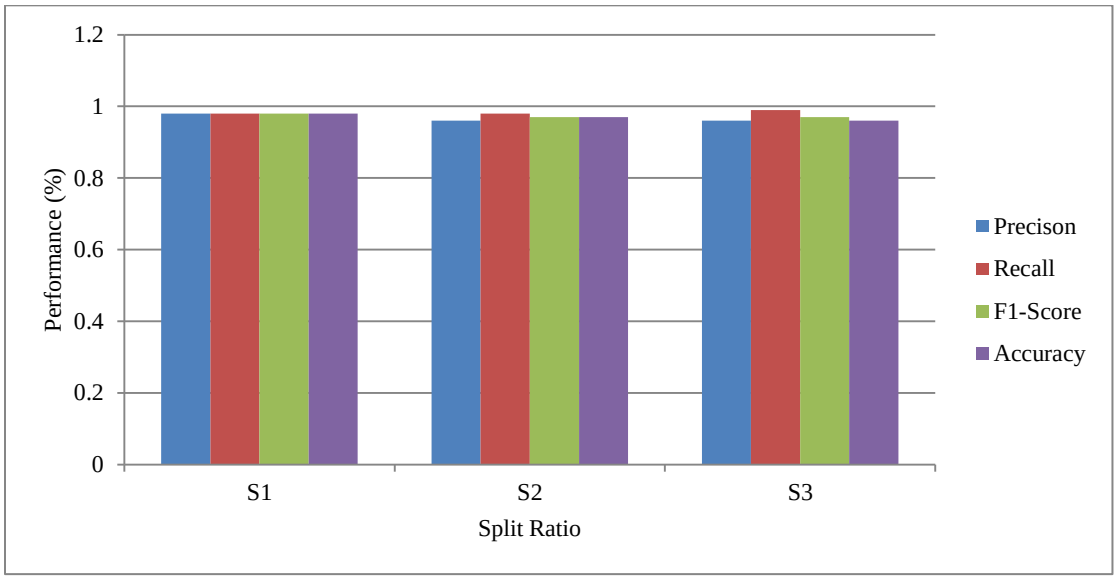

Figure 4 RF based result comparison considering precision, recall, F1-score, and accuracy

5.Conclusion

This paper investigated the effectiveness of two machine learning algorithms, LR and RF, for classifying a disease using the Breast Cancer Wisconsin Dataset. The results indicate that both algorithms are effective in classifying the data, with the LR algorithm achieving an accuracy level of 96% to 97%, and the RF algorithm achieving an accuracy level of 96% to 98%. The figures also demonstrate the impact of different split ratios on the accuracy and error rate of the algorithms, highlighting the importance of selecting an appropriate split ratio for training and testing the model. The study emphasizes the importance of data pre-processing and feature extraction in preparing the data for machine learning tasks. Overall, the findings suggest that machine learning algorithms can be effective in medical data classification, with the potential to contribute to the development of more accurate and efficient diagnostic tools.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Abideen ZU, Mazhar T, Razzaq A, Haq I, Ullah I, Alasmary H, et al. Analysis of enrollment criteria in secondary schools using machine learning and data mining approach. Electronics. 2023; 12(3):1-25.

[2] Suiçmez Ç, Yılmaz C, Kahraman HT, Cengiz E, Suiçmez A. Prediction of hepatitis C disease with different machine learning and data mining technique. In smart applications with advanced machine learning and human-centred problem design 2023(pp. 375-98). Cham: Springer International Publishing.

[3] Dubey AK, Gupta U, Jain S. Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data. International Journal on Advanced Science, Engineering and Information Technology. 2018; 8(1):18-29.

[4] Hussin SK, Omar YM, Abdelmageid SM, Marie MI. Traditional machine learning and big data analytics in virtual screening: a comparative study. International Journal of Advanced Computer Research. 2020; 10(47):72-88.

[5] Mumtaz G, Akram S, Iqbal W, Ashraf MU, Almarhabi KA, Alghamdi AM, et al. Classification and prediction of significant cyber incidents (SCI) using data mining and machine learning (DM-ML). IEEE Access. 2023.

[6] Sanjeetha R, Raj A, Saivenu K, Ahmed MI, Sathvik B, Kanavalli A. Detection and mitigation of botnet based DDoS attacks using catboost machine learning algorithm in SDN environment. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(76):445-61.

[7] Saha JK, Patidar K, Kushwah R, Saxena G. Object oriented quality prediction through artificial intelligence and machine learning: a survey. ACCENTS Transactions on Information Security. 2020; 5(17): 1-5.

[8] Dubey AK, Gupta U, Jain S. Computational measure of cancer using data mining and optimization. In sustainable communication networks and application: ICSCN 2019 2020 (pp. 626-32). Springer International Publishing.

[9] Mohammady M. Badland erosion susceptibility mapping using machine learning data mining techniques, Firozkuh watershed, Iran. Natural Hazards. 2023:1-9.

[10] Nemade V, Pathak S, Dubey AK. A systematic literature review of breast cancer diagnosis using machine intelligence techniques. Archives of Computational Methods in Engineering. 2022; 29(6):4401-30.

[11] Ashtiani MN, Raahmei B. News-based intelligent prediction of financial markets using text mining and machine learning: a systematic literature review. Expert Systems with Applications. 2023.

[12] Kannan R, Nandwana P. Accelerated alloy discovery using synthetic data generation and data mining. Scripta Materialia. 2023.

[13] Sher T, Rehman A, Kim D. COVID-19 outbreak prediction by using machine learning algorithms. Computers, Materials and Continua. 2023:1561-74.

[14] Dubey A, Gupta U, Jain S. Medical data clustering and classification using TLBO and machine learning algorithms. Computers, Materials and Continua. 2021; 70(3):4523-43.

[15] Nemade V, Pathak S, Dubey AK, Barhate D. A review and computational analysis of breast cancer using different machine learning techniques. International Journal of Emerging Technology and Advanced Engineering. 2022; 12(3):111-8.

[16] Mahoto NA, Shaikh A, Sulaiman A, Al Reshan MS, Rajab A, Rajab K. A machine learning based data modeling for medical diagnosis. Biomedical Signal Processing and Control. 2023.

[17] Cheng LC, Lu WT, Yeo B. Predicting abnormal trading behavior from internet rumor propagation: a machine learning approach. Financial Innovation. 2023; 9(1).

[18] Chahar R, Dubey AK, Narang SK. A review and meta-analysis of machine intelligence approaches for mental health issues and depression detection. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(83):1279-314.

[19] Ananthi J, Sengottaiyan N, Anbukaruppusamy S, Upreti K, Dubey AK. Forest fire prediction using IoT and deep learning. International Journal of Advanced Technology and Engineering Exploration. 2022; 9(87):246-56.

[20] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine. 2001; 23(1):89-109.

[21] Kamra V, Kumar P, Mohammadian M. Formulation of an elegant diagnostic approach for an intelligent disease recommendation system. In 9th international conference on cloud computing, data science & engineering (Confluence) 2019 (pp. 278-81). IEEE.

[22] Xiang Z, Jinghua C, Tao W. Review of machine learning algorithms for health-care management medical big data systems. In international conference on inventive computation technologies (ICICT) 2020 (pp. 651-4). IEEE.

[23] Juddoo S, George C. A qualitative assessment of machine learning support for detecting data completeness and accuracy issues to improve data analytics in big data for the healthcare industry. In 3rd international conference on emerging trends in electrical, electronic and communications engineering (ELECOM) 2020 (pp. 58-66). IEEE.

[24] Leung CK, Chen Y, Hoi CS, Shang S, Cuzzocrea A. Machine learning and OLAP on big COVID-19 data. In IEEE international conference on big data (Big Data) 2020 (pp. 5118-27). IEEE.

[25] Jayatilake SM, Ganegoda GU. Involvement of machine learning tools in healthcare decision making. Journal of healthcare engineering. 2021:1-20

[26] Tchito Tchapga C, Mih TA, Tchagna Kouanou A, Fozin Fonzin T, Kuetche Fogang P, Mezatio BA, et al. Biomedical image classification in a big data architecture using machine learning algorithms. Journal of Healthcare Engineering. 2021; 2021:1-11.

[27] Chahar R. Computational decision support system in healthcare: a review and analysis. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(75):199-220.
[28] Mustafa A, Rahimi Azghadi M. Automated machine learning for healthcare and clinical notes analysis. Computers. 2021; 10(2):1-31.
[29] Aldahiri A, Alrashed B, Hussain W. Trends in using IoT with machine learning in health prediction system. Forecasting. 2021; 3(1):181-206.
[30] Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. Communications Medicine. 2021; 1(1):25.
[31] Rafi TH, Shubair RM, Farhan F, Hoque MZ, Quayyum FM. Recent advances in computer-aided medical diagnosis using machine learning algorithms with optimization techniques. IEEE Access. 2021; 9:137847-68.
[32] Sun W, Zhang P, Wang Z, Li D. Prediction of cardiovascular diseases based on machine learning. ASP Transactions on Internet of Things. 2021; 1(1):30-5.
[33] Dhinakaran M, Phasinam K, Alanya-Beltran J, Srivastava K, Babu DV, Singh SK. A system of remote patients' monitoring and alerting using the machine learning technique. Journal of Food Quality. 2022:1-7.
[34] Elyan E, Vuttipittayamongkol P, Johnston P, Martin K, McPherson K, Jayne C, et al. Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward. Artificial Intelligence Surgery. 2022:1-25.
[35] Hinterwimmer F, Lazic I, Suren C, Hirschmann MT, Pohlig F, Rueckert D, et al. Machine learning in knee arthroplasty: specific data are key-a systematic review. Knee Surgery, Sports Traumatology, Arthroscopy. 2022; 30(2):376-88.
[36] Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nature Biomedical Engineering. 2022:1-6.
[37] Severn C, Suresh K, Görg C, Choi YS, Jain R, Ghosh D. A pipeline for the implementation and visualization of explainable machine learning for medical imaging using radiomics features. Sensors. 2022; 22(14):1-16.
[38] Zhu S, Gilbert M, Chetty I, Siddiqui F. The 2021 landscape of FDA-approved artificial intelligence/machine learning-enabled medical devices: an analysis of the characteristics and intended use. International Journal of Medical Informatics. 2022.
[39] Kobashi S, Hossain B, Nii M, Kambara S, Morooka T, Okuno M, Yoshiya S. Prediction of post-operative implanted knee function using machine learning in clinical big data. In 2016 international conference on machine learning and cybernetics (ICMLC) 2016 (pp. 195-200). IEEE.
[40] Lu YC, Lu CJ, Chang CC, Lin YW. A hybrid of data mining and ensemble learning forecasting for recurrent ovarian cancer. In 2017 international conference on intelligent informatics and biomedical sciences (ICIIBMS) 2017 (pp. 216-6). IEEE.
[41] Pitoglou S, Koumpouros Y, Anastasiou A. Using electronic health records and machine learning to make medical-related predictions from non-medical data. In international conference on machine learning and data engineering (iCMLDE) 2018 (pp. 56-60). IEEE.
[42] Reamaroon N, Sjoding MW, Lin K, Iwashyna TJ, Najarian K. Accounting for label uncertainty in machine learning for detection of acute respiratory distress syndrome. IEEE Journal of Biomedical and Health Informatics. 2018; 23(1):407-15.
[43] Liu Y, Leng Q, Wang S. Learning medical diagnosis via scaled convex hull-based SK algorithm. In 8th data driven control and learning systems conference (DDCLS) 2019 (pp. 377-81). IEEE.
[44] Chang W, Liu Y, Xiao Y, Yuan X, Xu X, Zhang S, Zhou S. A machine-learning-based prediction method for hypertension outcomes based on medical data. Diagnostics. 2019; 9(4):1-21.
[45] Khushi M, Shaukat K, Alam TM, Hameed IA, Uddin S, Luo S, et al. A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access. 2021; 9:109960-75.
[46] Yang H, Li X, Cao H, Cui Y, Luo Y, Liu J, Zhang Y. Using machine learning methods to predict hepatic encephalopathy in cirrhotic patients with unbalanced data. Computer Methods and Programs in Biomedicine. 2021.
[47] Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of heart disease using a combination of machine learning and deep learning. Computational Intelligence and Neuroscience. 2021:1-11.
[48] Ram A, Vishwakarma H. Diabetes prediction using machine learning and data mining methods. In IOP conference series: materials science and engineering 2021 (pp. 1-11). IOP Publishing.
[49] Khan S, Saravanan VN, Lakshmi TJ, Deb N, Othman NA. Privacy protection of healthcare data over social networks using machine learning algorithms. Computational Intelligence and Neuroscience. 2022:1-8.
[50] Urban S, Błaziak M, Jura M, Iwanek G, Zdanowicz A, Guzik M, et al. Novel phenotyping for acute heart failure-unsupervised machine learning-based approach. Biomedicines. 2022; 10(7):1-20.
[51] Lee KH, Dong JJ, Kim S, Kim D, Hyun JH, Chae MH, et al. Prediction of bacteremia based on 12-year medical data using a machine learning approach: effect of medical data by extraction time. Diagnostics. 2022; 12(1):1-13.
[52] Ahmad GN, Fatima H, Ullah S, Saidi AS. Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV. IEEE Access. 2022; 10:80151-73.
[53] Dong Z, Wang Q, Ke Y, Zhang W, Hong Q, Liu C, et al. Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical

records. Journal of Translational Medicine. 2022; 20(1):1-10.
[54] Tanioka S, Yago T, Tanaka K, Ishida F, Kishimoto T, Tsuda K, et al. Machine learning prediction of hematoma expansion in acute intracerebral hemorrhage. Scientific Reports. 2022; 12(1):1-8.

Abhishek Kumar is doing M. tech. in Computer Science , Patel College of Science and Technology, Bhopal (MP) and completed B.E. from Patel College of Science and Technology, Bhopal (MP). His area of interest are Data mining optimization, Machine learning, Artificial intelligence.
Email: Kumar0128pcst@gmail.com

Sujeet Gautam is working as Assistant professor with the department of Computer Science and Engineering, at Patel College of Science and Technology, Bhopal, India. He has completed his MTech. degree with Software System from Samrat Ashok technological institute, Vidisha. He has more than 10 publications in reputed, peer-reviewed national and international journals and conferences. His research areas are Data Mining, Cloud Computing and Artificial Intelligence.
Email: sujeetgautam21@gmail.com