Wasit Journal for Pure Science

Journal Homepage: https://wjps.uowasit.edu.iq/index.php/wjps/index
e-ISSN: 2790-5241 p-ISSN: 2790-5233

Medical Datasets Training and Enhancing Disease Prediction Accuracy

Enas Fadhil Abdullah[1], Asraa Mounaf Almousawy[2]*, Fatima Mohamad Abbas[3], Tabarak Akram Aliwi[4], Nour Alayoun Mahdi Hassan[5]

[1,2,3,4,5]Department of Computer Science, Collage of Education for Girls, University of Kufa, IRAQ

*Corresponding Author: Asraa Mounaf Almousawy

DOI: https://doi.org/10.31185/wjps.441
Received 02 June 2024; Accepted 21 July 2024; Available online 30 September 2024

ABSTRACT: Data mining is an effective method that uses sophisticated tools and strategies to sift through massive databases in search of useful patterns. Its usefulness extends to many fields, including medicine. We used AdaBoost, Logistic Regression (LR), K-Nearest Neighbors (KNN), and Random Forest (RF) as predictive models in our investigation (ADaB). Evaluating utilizing methods like cross-validation and random sampling allowed us to concentrate on improving accuracy. Medical datasets were used to construct the following: Heart Disease (HD) dataset, Breast Cancer Wisconsin (BCW) dataset, and Covid-19 dataset. We set out to improve the accuracy of disease predictions and push the boundaries of medical data analysis. After completing the assessment procedure with the training data, the performance measurements showed that the highest accuracy was achieved with LR 83% for HD, KNN 97%, and LR 98% for Covid-19.

Keywords: Keywords: Data mining, medical dataset, Covid-19, and classification.

1. INTRODUCTION

In their 2022 study, Shanbehzadeh et al. set out to create a CDSS that takes lifestyle risk factors into account to enhance the efficacy of machine learning models used for breast cancer diagnosis. Random Forest was determined to be the most successful model among the many machine-learning algorithms that were studied [1]. Among five ML models tested by Zhang and Li (2022) for breast cancer prediction, Logistic Regression (LR) performed best in terms of classification accuracy. Nevertheless, they did note that the dataset they utilized had certain limitations [2]. By comparing various machine learning algorithms, Patidar et al. discovered that their proposed method had the most accurate results for predicting the occurrence of heart disease. Because it uses one-hot encoding, which better represents the input, Random Forest outperformed other algorithms [3]. The results obtained by Fawad Masood et al. (2022) while using the classification approach independently were superior to those obtained when combining it with FSFA (Feature Selection and Feature Aggregation). They discovered that the categorization method produced better experimental accuracy and was more efficient in obtaining results. Because analysis takes more time with higher dimensions, they also stressed the importance of dimensional reduction [4]. In 2020, researchers Ibomoiye Domor M et al. suggested and tested an ensemble approach with the Cleveland and Framingham datasets. Classification accuracy ratings of 91% and 93% on the two test sets, respectively, were shown in the evaluation [5].

Results were better when the classification method was used alone, rather than in conjunction with FSFA, according to Fawad Masood et al. When it came to retrieving results, they found that the categorization procedure was more efficient and accurate [6]. Researchers Ravichandran, B.D., and Keikhosrokiani, P. found that a combination of neuro-fuzzy and neural network approaches improved performance. The ANFIS-DNN model is an innovative strategy for better categorization of COVID-19 misinformation [7].

In this research has been applied classification models are K-nearest neighbor (KNN), logistic regression (LR), random forest (RF), and AdaBoost methods. Three medical databases were used, the first and the second were downloaded from the orange environment, and the third was collected manually and linked to the orange environment.

2. CLASSIFICATION TECHNIQUES

A training set of cases that have already been classified and a clear definition of the classes are the defining features of a classification task. The goal is to create a model that can be used to categorize data that is not yet classified. The term "classification" refers to the steps used to identify and differentiate between different types of data or ideas. Random Forest, k-nearest neighbour, Ada Boots, and many more classification methods also exist. [9]

2.1Random Forest

In the same way that decision trees and Random Forests (RF) are ensemble classifiers, they can both address regression and classification problems. It generates several random trees using the bootstrap method on the training dataset, bagging on samples, a voting scheme, and feature selection at each decision split, all of which increase efficiency and improve predictive power. It outperforms decision trees in the vast majority of cases. One application of the random subspace method is the selection of a random subset of features [10].

2.2 K-Nearest Neighbors

No machine learning technique is more basic than the k-nearest Neighbors (KNN) algorithm. The basic premise is that things that are physically close together will inevitably share some traits. You can forecast the features of an object's nearest neighbor based on your knowledge of its distinguishing qualities. KNN is a variant of the closest neighbor method that has been fine-tuned. The core principle is that as long as 'k' neighbors can agree on a classification for a new instance, it can be categorized. Here, 'k' is a positive integer, typically a modest number. [11]

2.3 The technique boosting (Ada boost)

Making a powerful strong classifier out of a group of smaller, less effective classifiers. Freund and Schapiro presented AdaBoost, the first usable boosting method, in 1996. Boosting has risen to prominence as a top classification tool in computer vision, because of its excellent generalizability, rapid performance, and minimal implementation complexity. [12]

2.4 Logistic Regression

Model One of the most popular approaches to data mining and binary data classification in particular is logistic regression (LR). [13]

Logistic regression is a popular statistical method for analyzing data with binary and proportional answer formats among academics and statisticians.

One kind of supervised machine learning technique is the logistic regression model, which is used for classification purposes. Using a set of independent variables, it estimates discrete values, such as yes/no, true/false, or 0/1 [14]. To put it simply, it uses a logit function to fit data and determine the likelihood of an event occurring.

3. RELATIONAL DATABASES

Figure (1) shows the different kinds of databases. A database system, also known as a database management system (DBMS), is comprised of a database—a collection of linked data—and a suite of applications designed to manage and retrieve this data. The software programs incorporate features that allow for the definition of database structures, data storage, access to data in a concurrent, shared, or distributed manner, and the assurance of data consistency and security in the face of system crashes or unwanted access attempts. The building blocks of a relational database are the tables, which are given distinct names. Tuples are often stored in tables that have a collection of columns or fields that represent attributes (records or rows). A relational table stores objects as tuples, which are characterized by a set of attribute values and identifiable by a unique key. [15]

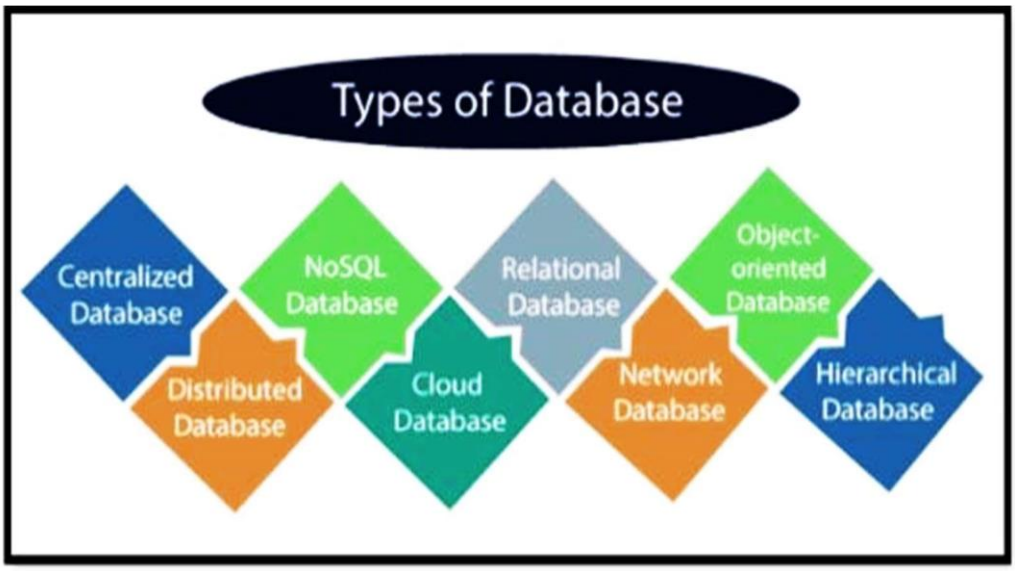

Figure 1. Types of databases

3.1 Entity-Relationship Diagram

One way to depict the interconnections and logic between various parts of a system is with an entity-relationship diagram or ERD. An ERD is a plan for building the physical data structures and a bird's-eye view of the system. The initial stage in creating an ERD is to compile a list of all the entities that were identified in the systems analysis phase and think about the types of links that connect them. You may now demonstrate the relationships between things using a simplified way. There are a variety of approaches to drawing ERDs, but one common practice is to use rectangles for entities and diamonds for relationships. In a typical top-down and left-to-right layout, the Entity rectangles are labelled with singular nouns and the Relation diamonds with verbs. [16]

3.2 Splitting of Dataset

Although there are many approaches to data splitting, they generally fall into one of three categories:

-To guarantee good generalization and to prevent over-training, conventional methodologies often employ cross-validation techniques. One part of the dataset T is utilized for training, while the other part is left out and used to evaluate the final model's performance. This is the main notion.

- Choosing a subset of samples at random, keeping them (holding out) as a validation set, and training with the rest of the samples. It is common practice to do this procedure multiple times, with the average performance on validation sets serving as the final estimate of the model's performance.

4. Prediction Measures (Metrics)

There are four sets of categorized data in a confusion matrix; TP and TN indicate proper categorization of cases, whereas FP and FN indicate wrong classification. This allows one to summarize the results of a classification method. Accuracy, recall, precision, and F-1score were the four metrics utilized to evaluate the expected sentiment (Acc.) Among the many measures used to evaluate the efficacy of information retrieval tasks, the two most well-known are recall and precision.

The equations below demonstrate precision, recall, accuracy, and F1score. [17]

$$Pricision = \frac{\mathrm{Tp}}{\mathrm{Tp} + \mathrm{Fp}} \qquad (1)$$

$$Recall = \frac{\mathrm{Tp}}{\mathrm{Tp} + \mathrm{Fn}} \qquad (2)$$

$$Accuracy = \frac{(\mathrm{Tp} + \mathrm{Tn})}{(\mathrm{Tp} + \mathrm{Fn} + \mathrm{Fp} + \mathrm{Fn})} \qquad (3)$$

$$F1 = 2 * \left(\frac{(\mathrm{Precision} * \mathrm{Recall})}{((\mathrm{Precision} * \mathrm{Recall})}\right) \qquad (4)$$

Interpretation of the symbols (TP, FN, FP, and TN) in equations (1,2,3) as follow:

1. True Positive (TP): If the instance is positive and it is classified as positive.
2. False Negative (FN): If the instance is positive, but it is classified as negative.

3. True Negative (TN): If the instance is negative and it is classified as negative.
4. False Positive (FP): If the instance is negative, but it is classified as positive.

5. MATERIALS AND METHOD

The methodology of our work includes three stages, the first stage is represented by preprocessing, the second stage is represented by the classification model (RF, KNN, LR and Ada boots) and the finally stage is evaluation as shown in the block diagram Figure (2).

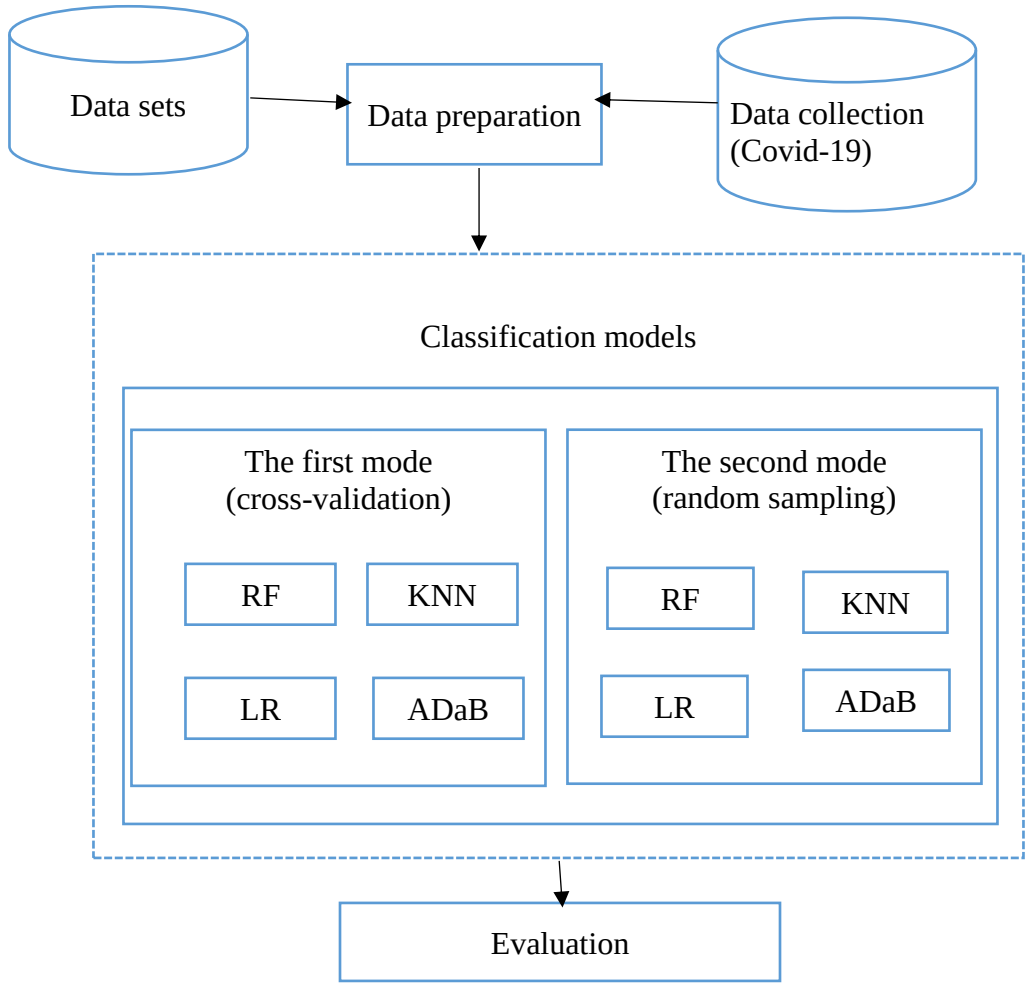

Figure 2. Block diagram of method

5.1 Description of datasets

Our work on three data sets are heart disease dataset (HD_dataset) , the breast cancer Wisconsin dataset (BCW_ dataset) and the Covid_19 database (COV_ database). HD_dataset consist of 14 features and 10 features for BCW_ dataset. Table (1) description for some features (F1, F2, and…. F10).

Table 1. Description of features

Data sets	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
HD_dataset	Age	Gender	Diameter narrowing	Chast pain	Rest SBP	Cholesterol	Fasting blood suger> 120	Rest ECG	Max HR	Thal
BCW_ dataset	Type	Clump thickness	Unif-Cell-Size	Marginal - Adhesion	Single-Cell-Size	Bare-Nuclei	Bland - Chromatine	Normal-Nucleoli	Mitoses	Unif-Cell-Shape

Manually building COVID-19 database, data preparation stage that summarization by steps:

Steps1: Create ERD diagram for each dataset.

Steps2: Create tables using Query design.

Step3: Insert data to tables.

Step4: Build database using form design

Outputs of the preprocessing stage are tables and relationships, tables consist of four tables (patient, region, symptoms and vaccines). The relationship between the patient's schedule and vaccinations, as well as the relationship between the patient's schedule and the region, is represented by one-to-many, and the relationship between the patient's schedule and symptoms is represented by one-to-one, as shown in the following Figure (3).

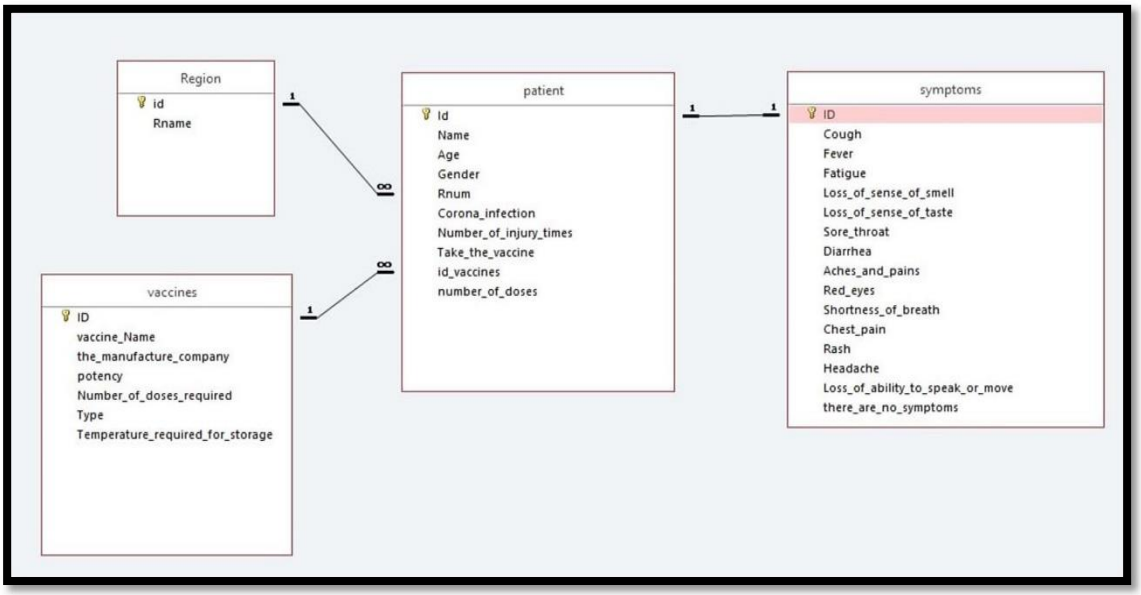

Figure 3. Sample of creating COVID-19 database

5.2 hd_datasets results

Table (2) and Figure (4) show the result of HD_datasets using cross-validation to four model methods (KNN, RF, LR, AdaBoost).

Table 2. Performance Measures of HD-Datasets

Model	AUC	CA	F1	Precision	Recall
KNN	0.684651	0.633803	0.623435	0.63285166	0.633803
Random Forest	0.882657	0.816901	0.817226	0.81845155	0.816901
Logistic Regression	0.892272	0.802817	0.802376	0.80250819	0.802817
AdaBoost	0.758146	0.755869	0.756462	0.76038456	0.755869

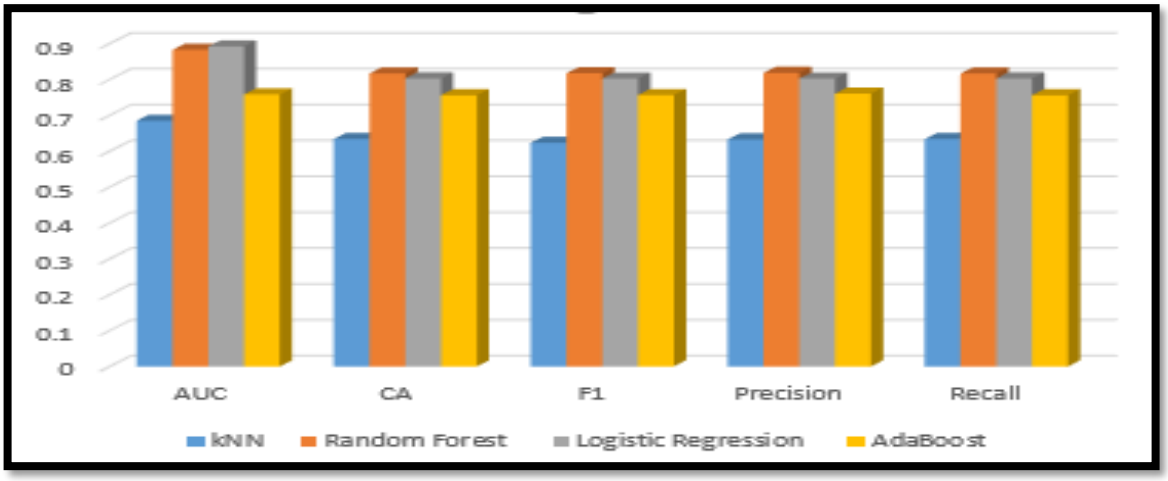

Figure 4. Accuracy of Classification Methods

Conclusion from Figure (4) and Table (2) the best classification accuracy is82% for RF, 80% LR, 76% AdaBoost and 63% for KNN.

Table (3) and Figure (5) show the result of HD_datasets using randomsampling (50% train data and 50% test data) to four model methods.

Table 3. HD-Datasets Performance Measures

Model	AUC	CA	F1	Precision	Recall
KNN	0.65579	0.630093	0.628243	0.62814459	0.630093
Random Forest	0.859269	0.784486	0.783799	0.78418426	0.784486
Logistic Regression	0.89066	0.818318	0.817998	0.81808622	0.818318
AdaBoost	0.744192	0.746729	0.746668	0.74661681	0.746729

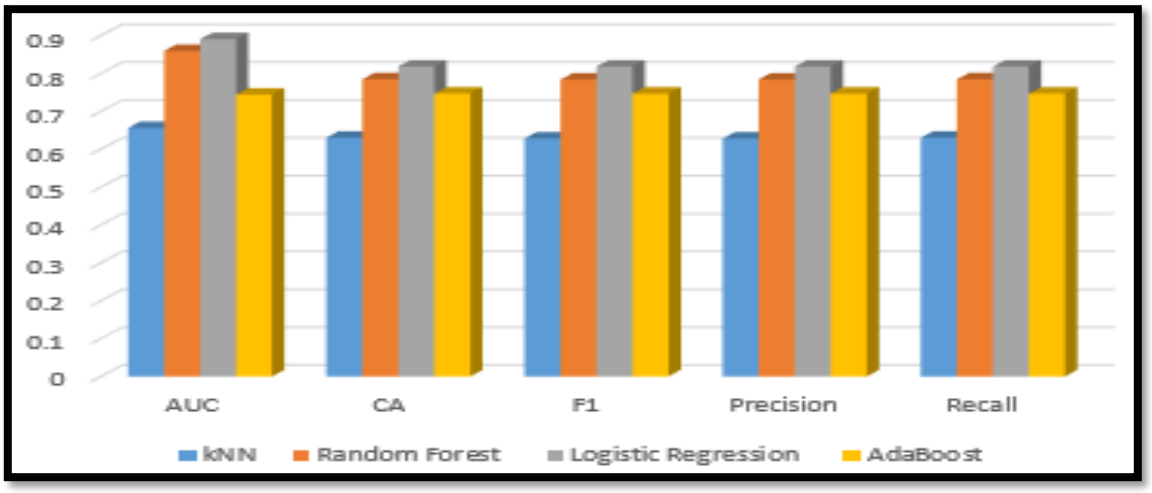

Figure 5. Classification Methods Accuracy

Conclusion From Figure (5) and Table (3) the best classification accuracy is82% for LR, 78% RF, 75% AdaBoost and 63% for KNN.

Table (4) and Figure (6) show the result of HD_datasets using random sampling(80% train data and 20% test data) to four model methods.

Table 4. Performance Measures of HD_Datasets

Model	AUC	CA	F1	Precision	Recall
KNN	0.653465	0.612791	0.60883	0.60957116	0.612791
Random Forest	0.860162	0.769767	0.769963	0.77029264	0.769767
Logistic Regression	0.895494	0.82907	0.828715	0.82887031	0.82907
AdaBoost	0.738783	0.739535	0.739932	0.74084085	0.739535

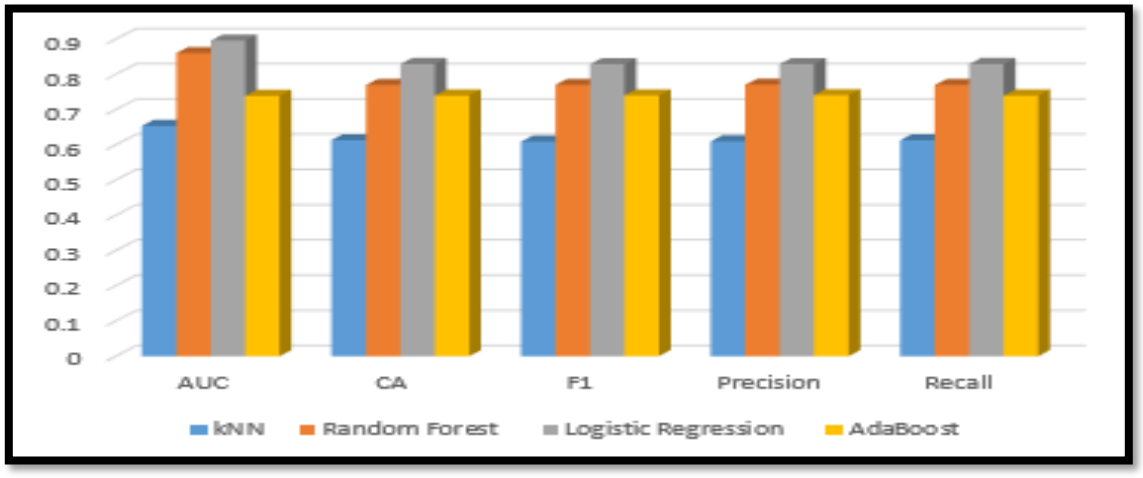

Figure 6. Accuracy of Classification Methods

Conclusion from Figure (6) and Table (4) the best classification accuracy is83% for LR, 80% RF,74% AdaBoost and 61% for KNN.

5.3 bcw_ datasets results

Table (5) and Figure (7) show the result of BCW_datasets using cross‐validation to four model methods.

Table 5. BCW_Datasets Performance Measures

Model	AUC	CA	F1	Precision	Recall
KNN	0.984892	0.97286	0.972841	0.97283075	0.97286
Random Forest	0.988722	0.966597	0.966732	0.967194157	0.966597
Logistic Regression	0.992686	0.962422	0.962368	0.962350587	0.962422
AdaBoost	0.940259	0.947808	0.947694	0.94766275	0.947808

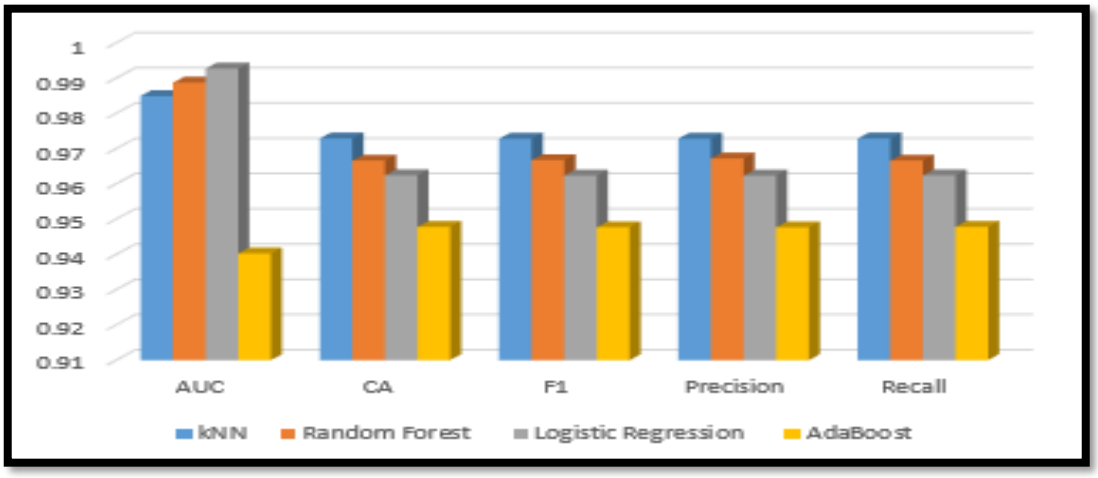

Figure 7. Classification Method Accuracy

The best accuracy of classification methods is 97% KNN, 97% RF, 96%LR,and 95 %AdaBoast respectively.

Table (5) and Figure (7) show the result of BCW_datasets using randomsampling (50% train data and 50% test data) to four model methods.

Table 6. Performance Measures of BCW_Datasets

Model	AUC	CA	F1	Precision	Recall
KNN	0.985804	0.964	0.963951	0.963935662	0.964
Random Forest	0.984484	0.955	0.955117	0.955366369	0.955
Logistic Regression	0.990635	0.960083	0.960044	0.960022149	0.960083
AdaBoost	0.914336	0.92475	0.924539	0.924460583	0.92475

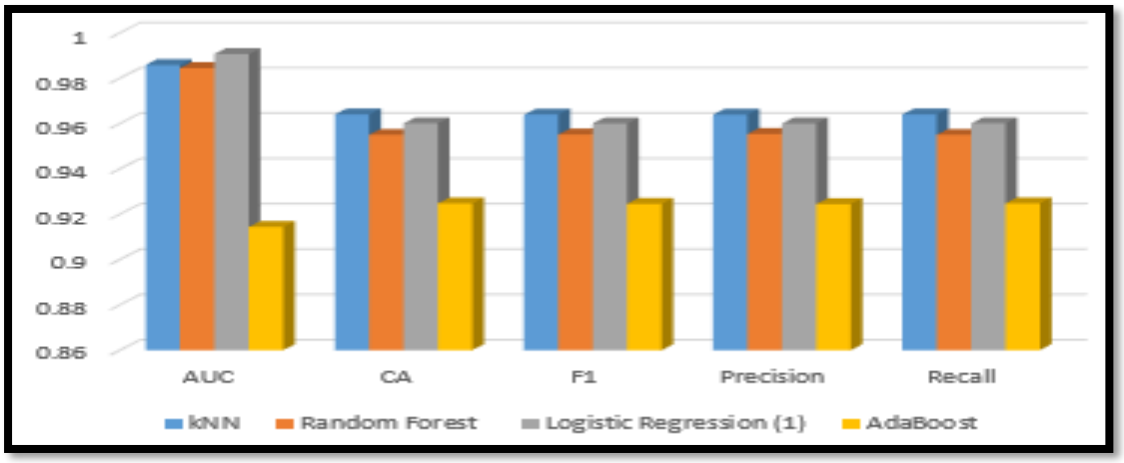

Figure 8. Accuracy of Classification Method

Conclusion From Figure (8) and Table (6) the best classification accuracy is96% for KNN, RF, LR, and 92% for AdaBoost.

Table (7) and Figure (9) show the result of the BCW_datasets using randomsampling (80% train data and 20% test data) to four model methods.

Table 7. BCW_Datasets Performance Measures

Model	AUC	CA	F1	Precision	Recall
KNN	0.984935	0.965625	0.965565	0.965562492	0.965625
Random Forest	0.98215	0.95	0.950132	0.950406528	0.95
Logistic Regression	0.991095	0.964583	0.964595	0.964609735	0.964583
AdaBoost	0.923075	0.930729	0.930669	0.930623186	0.930729

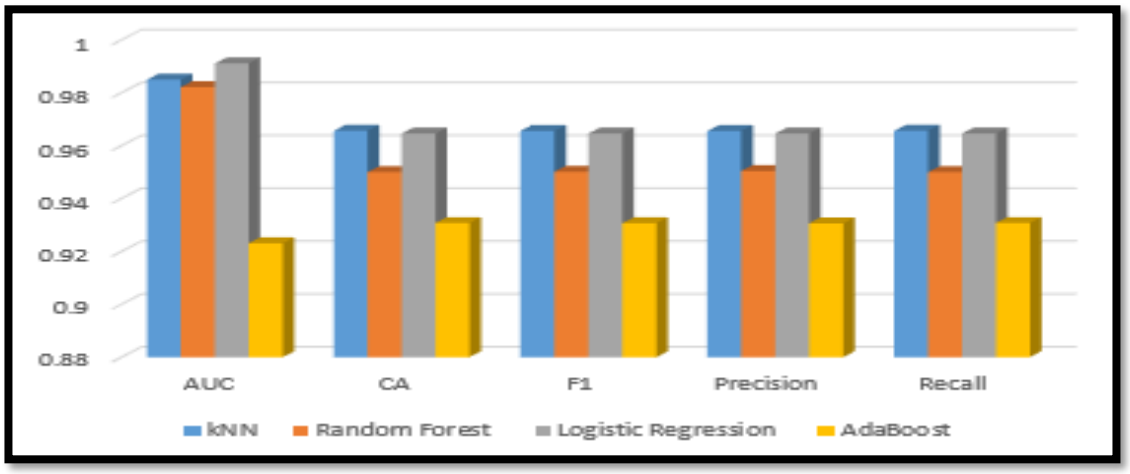

Figure 9. Classification Method Accuracy

Conclusion Figure (9) and Table (7) show the best way to model areKNN 97%, LR 96%, RF 95% and AdaBoost with an accuracy of 93%.

5.4 covid-19 datasets results

Table (8) and Figure (10) show the result of the COV _datasets usingcross-validation to four model methods.

Table 8. Performance Measures Of Covid-19 Datasets

Model	AUC	CA	F1	Precision	Recall
KNN	0.641754	0.640244	0.641017	0.642215996	0.640244
Random Forest	0.986976	0.963415	0.963415	0.963414634	0.963415
Logistic Regression	0.988339	0.981707	0.981722	0.981810416	0.981707
AdaBoost	0.968121	0.969512	0.969486	0.969532582	0.969512

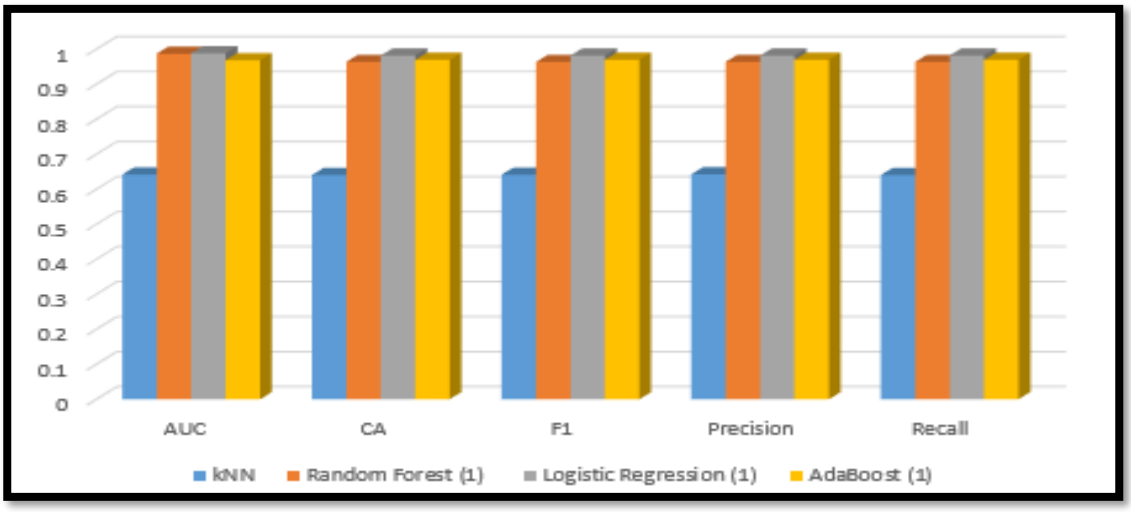

Figure 10. Accuracy of Classification Method

Conclusion in this Figure (10) and Table (8) show the best way to model LR98%, AdaBoost 97%, RF 96% and with accuracy KNN 64%.

Table (9) and Figure (11) show the result of the COV _datasets using randomsampling (50%train data and 50% test data) to four model methods.

Table 9. Covid-19 Performance Measures

Model	AUC	CA	F1	Precision	Recall
KNN	0.6041	0.571951	0.572942	0.574491758	0.571951
Random Forest	0.9763	0.926341	0.926199	0.926408929	0.926341
Logistic Regression	0.982944	0.967561	0.96756	0.967559073	0.967561
AdaBoost	0.961958	0.96439	0.964316	0.96466394	0.96439

Figure 11. Classification Method Accuracy

Conclusion in this Figure (11) and Table (9) show the best way of model LR97%, AdaBoost 96%, RF 93% and with accuracy KNN 57%.

Table (10) and Figure (12) show the result of the COV _datasets using randomsampling (80% train data and 20%test data) to four model methods.

Table 10. Performance Measures of Covid-19 Datasets

Model	AUC	CA	F1	Precision	Recall
KNN	0.65129	0.637879	0.638566	0.639746157	0.637879
Random Forest	0.983803	0.95	0.949953	0.950016347	0.95
Logistic Regression	0.988619	0.984848	0.984859	0.98494186	0.984848
AdaBoost	0.980215	0.980303	0.980306	0.980314458	0.980303

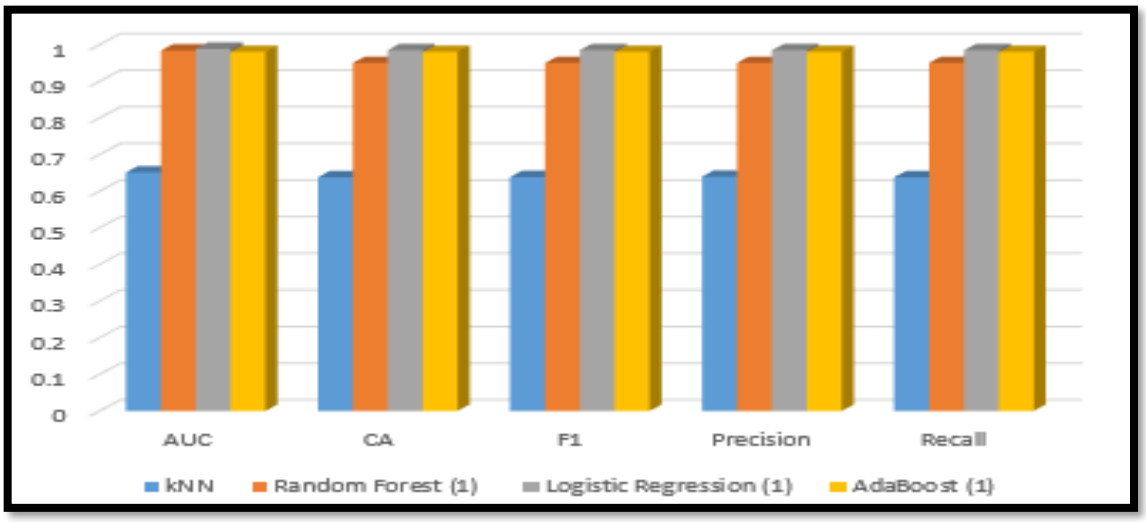

Figure 12. Accuracy of Classification Methods

Conclusion Figure (11) and Table (9) show the best way to model are LR 98%, AdaBoost 98%, RF 96% and with accuracy KNN 64%.

Wasit Journal for Pure Science

Journal Homepage: https://wjps.uowasit.edu.iq/index.php/wjps/index
e-ISSN: 2790-5241 p-ISSN: 2790-5233

6.Results and Discussion

The analysis of the three medical datasets yields significant results, highlighting the importance of early diagnosis in disease prevention. In the case of HD_datasets, the highest classification accuracies were achieved with Random Forest (RF) using cross-validation (82%), Logistic Regression (LR) with 50% train data (82%), and LR with 80% train data (83%), as illustrated in Figures 4,5 and 6 respectively.

Regarding the BCW_datasets, the summarized results demonstrate impressive accuracies of 97% using K-Nearest Neighbors (KNN) and RF with cross-validation, and 96% using LR, RF, and KNN with random sampling of 50% train data. Additionally, the use of KNN with 80% train data and 20 test data resulted in a commendable accuracy of 97%, as depicted in Figures 7, 8, and 9.

In the case of COVID-19 datasets, the accuracy results emphasize the significance of early diagnosis, with LR using cross-validation achieving a remarkable accuracy of 98%. Similarly, LR using random sampling of 50% train data yielded an accuracy of 97%, while the combined use of LR and AdaBoost with 80% train data and 20 test data resulted in an accuracy of 98%. These findings are elucidated in Figures 10, 11, and 12.

7. Conclusion

In conclusion, our study involved heart disease, and breast cancer and creating databases for Covid-19. We applied various machine learning algorithms and evaluation methods to measure accuracy rates. High accuracies were achieved in predicting breast cancer, heart disease, and Covid-19. The results emphasize the importance of early diagnosis for disease prevention. Notably, Random Forest, Logistic Regression, K-Nearest Neighbors, and AdaBoost showed impressive performance across the different datasets. These findings highlight the potential of machine learning algorithms in supporting accurate and timely medical diagnoses.

REFERENCES

[1] M. Shanbehzadeh, H. Kazemi-Arpanahi, M. B. Ghalibaf, and A. Orooji, "Performance evaluation of machine learning for breast cancer diagnosis: A case study," *Informatics in Medicine Unlocked*, 2022.

[2] Z. Zhang and Z. Li, "Evaluation methods for breast cancer prediction in machine learning field," *SHS Web of Conferences*, vol. 144, p. 03010, 2022.

[3] S. Patidar, D. Kumar, and D. Rukwal, "Comparative analysis of machine learning algorithms for heart disease prediction," 2022.

[4] F. Masood, J. Masood, H. Zahir, K. Driss, N. Mehmood, and H. Farooq, "Novel approach to evaluate classification algorithms and feature selection filter algorithms using medical data," *Journal of Computational and Cognitive Engineering*, vol. 2, no. 1, pp. 57–67, 2022.

[5] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," 2020.

Corresponding author: inasf.alturky@uokufa.edu.iq

[6] F. Masood, J. Masood, H. Zahir, K. Driss, N. Mehmood, and H. Farooq, "Novel approach to evaluate classification algorithms and feature selection filter algorithms using medical data," *Journal of Computational and Cognitive Engineering*, vol. 2, no. 1, pp. 57–67, 2022.

[7] B. D. Ravichandran and P. Keikhosrokiani, "Classification of Covid-19 misinformation on social media based on neuro-fuzzy and neural network: A systematic review," *Neural Computing and Applications*, vol. 35, pp. 699–717, 2023.

[8] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716-80727, 2020.

[9] D. Senthilkumar and S. Paulraj, "Prediction of low-birth-weight infants and its risk factors using data mining techniques," in *Proceedings of the 2015 International Conference on Industrial Engineering and Operations Management*, 2015.

[10] H. S. Khamis, K. W. Cheruiyot, and S. Kimani, "Application of k-nearest neighbour classification in medical data mining," *International Journal of Information and Communication Technology Research*, vol. 4, no. 4, 2014.

[11] A. Vezhnevets and V. Vezhnevets, "Modest AdaBoost-teaching AdaBoost to generalize better," *Graphicon*, vol. 12, no. 5, 2005.

[12] M. Maalouf, "Logistic regression in data analysis: an overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, pp. 282-299, 2011; J. M. Hilbe, *Logistic Regression Models*. Chapman and Hall/CRC, 2009.

[13] J. Han, M. Kamber, and J. Pei, "Mining frequent patterns, associations, and correlations," in *Data Mining: Concepts and Techniques*, 2nd ed., San Francisco, USA: Morgan Kaufmann Publishers, 2006, pp. 227-283.

[14] G. B. Shelly, T. J. Cashman, and H. J. Rosenblatt, *Systems Analysis and Design*, 6th ed., Course Technology, 2006.

[15] Y. Xu and R. Goodacre, "On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," *Journal of Analysis and Testing*, 2021.

[16] S. Garg, "Drug recommendation system based on sentiment analysis of drug reviews using machine learning," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, 2021.

[17] S. Makridakis, "Accuracy measures: theoretical and practical concerns," *International Journal of Forecasting*, vol. 9, no. 4, pp. 527-529, 1993.