JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 18, No. 1, March 2024, pp. 57-68

A Novel Enhanced Random Forest for Medical Data Classification using Correlation Pearson and Best Number of Trees

Ilhem Tarchoune*, Akila Djebbar, Hayet Farida Merouani, and Harfi Rania
Department of Computer Science, LRI Laboratory, SRF equip, Badji Mokhtar University, Annaba, Algeria
tarchouneilhem@gmail.com, aki_djebbar@yahoo.fr, hayet_merouani@yahoo.fr, harfirania@gmail.com

Abstract
Random forests (RF) is a successful ensemble prediction technique that uses majority voting or a combination-based average. However, each tree in an RF may have a different contribution to the treatment of a certain instance. The objectives of this study were to produce accurate decision trees and to determine the best trees between them with an optimal combination search. In this paper, we proposed three solutions for the prediction of medical data: the first solution optimizes a random forest model using a similarity measure, the second optimizes the RF using feature selection, and finally a simultaneous selection approach to similarity measures based on RFs. We demonstrated that the prediction performance and classification rate of the RF implementation on eleven databases can be further improved by the learning methods applied. Our experiments also showed that the improvement gives better results than the classical method; the results showed that the optimized RF model avoids some limitations of the original RF model. The results obtained in our proposed models are satisfactory and encouraging with an average accuracy of 95% for standard RF, 100% for RF_Similarity, 93% for RF_FS, and 100% for RF_FS_Similarity.

Category: Smart and Intelligent Computing

Keywords: Random Forest; Decision tree; Feature selection; Similarity measure; Classification; Medical database.

I. INTRODUCTION

Artificial intelligence (AI) is a set of techniques designed to approximate and imitate human reasoning. Thanks to its effectiveness and the relevance of its results, AI has seen tremendous increase in its areas of use and applications. Among its applications is decision support, which mainly uses data mining techniques. This highly effective concept is now widely used to extract knowledge from data for optimal decision-making. To do this, it uses, among other approaches, classification methods based on probabilities and differential statistics. Data mining is of great importance in the medical field. It is an excellent field of experimentation for testing and evaluating the various AI paradigms. A number of intelligent systems have been designed for various purposes in the medical field.

Data mining is the search for relevant information to aid decision-making and forecasting. It uses statistical and artificial learning techniques, taking into account the specific nature of large datasets. Machine learning consists of designing high-performance classification systems based on a set of examples representative of a population of data. One type of machine learning is supervised learning, which automatically produces rules from a labeled training database. This technique aimed to predict the class of new data observed, using other classification models

Open Access http://dx.doi.org/10.5626/JCSE.2024.18.1.57 http://jcse.kiise.org

Received 29 December 2023; **Accepted** 14 March 2024
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

(classifiers) such as decision trees, Bayesian networks, neural networks, and k-nearest neighbors.

There are now a large number of methods capable of automatically generating sets of classifiers: bagging, boosting, random subspaces, the random forest (RF) method, etc. This method is a bagging method with improved hyperparameters [1]. It is based on the combination of elementary classifiers of the decision tree type. Individually, these classifiers have interesting properties to exploit, but they are particularly unstable. The specificity of the trees used in RFs is that their induction is perturbed by a random factor, to generate diversity in the ensemble. It was on the basis of these two elements—using decision trees as elementary classifiers and involving randomness in their induction—that RF formalism was introduced. The RF is a representative of integrated learning [2], due to its advantages—few parameters, a strong anti-noise capacity, it is widely used in the medical field: breast cancer [3–7], heart disease [8–10], diabetes [11–13], chronic kidney cancer [14], and other diseases. Despite the success of RF methods, several works have proposed improvements. Individually, each classifier gives poor predictions. They have proposed strengthening each classifier without sacrificing the variety between them and reducing the variance without sacrificing strength. For other works, the improvement of each classifier is insufficient. The difficulty of using the RF algorithm led us to propose changing the standard method by other techniques to integrate the performances of several individual learning algorithms. The feature selection technique is one of the most important data processing strategies, which has been widely used in machine learning [15] and data mining [16]. Feature selection facilitates the use of predictive models for clinicians by reducing the workload associated with data collection.

This work aims to implement the standard RF algorithm on eleven databases and to propose three variants of the latter: the RF algorithm with similarity measure (RF_Similarity), the RF algorithm with attribute selection (RF_FS), and the RF algorithm with attribute selection and similarity measure (RF_FS_Similarity) in order to improve classification performance.

Section II summarizes some of the work proposed in the literature; Section III presents the proposed approach based on the RF algorithm with three variants; Section IV then summarizes the experimentation and analysis of the results. Finally, in Section V, we present a general conclusion.

II. RELATED WORKS

A study on the use of RF in the classification of medical data was carried out by a proposed hybrid neural network model to determine the predicted weights of related genes and RF to analyze the key genes of heart failure [17]. They evaluated the effectiveness of the classification in three datasets, with successful construction and verification of a new heart failure diagnostic model was analyzed using a hybrid model in public datasets.

[18] developed a new evolutionary RF (CERF) cluster analysis method, applying the correlation method to detect relationships between brain regions and genes. They then applied the CERF method to extract the most discriminating features in Alzheimer's disease. The results show that they were able to identify Alzheimer's patients effectively.

[19] analyzed the correlation between features to obtain normal embryonic development. Next, they applied six machine learning algorithms whose RF was predicted to be the most accurate.

[20] developed a model to build a system for pregnant women at high risk of preterm birth before cervical cerclage. First, they built a data balancing technique called the Synthetic Minority Oversampling Technique (SMOTE). Then four classification models were used to build the prediction model. The results showed that the RF outperformed the other classifiers.

[21] presented a feature-based RF approach for automatic detection of postictal generalized electroencephalogram suppression (PGES). The results showed that the proposed approach achieved increasingly better performance.

[22] developed a hybrid feature selection approach to rank medical data according to the importance of each disease, and then applied the RF only to highly ranked features. This method showed better performance.

[23] have proposed a weighted Pearson correlation based improved random forest classification (WPC-IRFC), technique to select relevant features and an improved RF classification to improve prediction performance. [24] used RFs and other methods to predict Alzheimer's disease. The algorithms on three datasets with different numbers of features were tested, and the results showed that the RF achieved the best accuracy. [25] provided individualized prediction of conversion from mild cognitive impairment (MCI) to Alzheimer's disease using a balanced RF model based on clinical data. Experimental results show the effectiveness of RF in predicting MCI conversion.

[26] Proposed a three-way selection RF algorithm based on the entropy of the decision boundary which is defined by the importance of the attribute. In this study, they validated the proposed algorithm on six datasets. The results show that the proposed algorithm has a significant improvement effect on multi-class data.

RF remains one of the most widely used ensemble algorithms in data mining, achieving well-documented levels of accuracy and processing speed. In recent years, they have been integrated into other learning techniques to improve performance. For example, [27] integrated decision tree algorithms (C4.5, RepTree, LMT) and standard RFs into the case-based reasoning (CBR) system to compare the performance of the algorithms and model

the retrieval phase of the CBR system. The simulation results confirm the hybridization performance of the CBR system and RFs. The algorithms were tested on four medical databases. [28] Integrated modified RFs in the retrieval phase of the CBR system, and used the RF algorithm in three different ways: standard random forest (CRF), RF with selection of the most important attributes (RF-FS) and weighted random forest (WRF). They tested the three algorithms on eleven medical databases, the results show the efficiency of the proposed algorithms to model the recollection phase. [29] presented a systematic review of CBR and RFs in the medical domain, and showed the performance of each method as well as the performance of the hybridization of these methods.

In this section, we have shown that researchers have applied standard RFs as a classifier or as a selection technique. They have applied RF with these parameters based on entropy which is defined as the importance of the attribute. In this work, we have proposed a hybrid model based on the improvement of RFs through a similarity measure applied only to highly ranked features. The aim of this study was to target the earliest prediction of disease and avoid the limitations of the original RF model, as studies indicate that early diagnosis is essential to delay the development of disease.

III. PROPOSED ARCHITECTURE

As shown in Fig. 1, the methodology in this study involved three main stages: pre-processing the data, developing the model, and evaluating the results.

A. Datasets

The datasets used in the experiment are detailed in Table 1. We used eleven medical databases collected from the machine learning repository of the University of California at Irvine (UCI), from Kaggle, and a database from Algeria.

B. Pre-processing

We began by pre-processing the data, which is a very important step in the system. In this part, we transformed medical databases from csv files and transformed them into a Python program. First, we removed missing values because the dataset contained a small number of entries with missing values.

Given that the RF classifier model cannot fit alphanumeric data, we applied a label encoder to all the

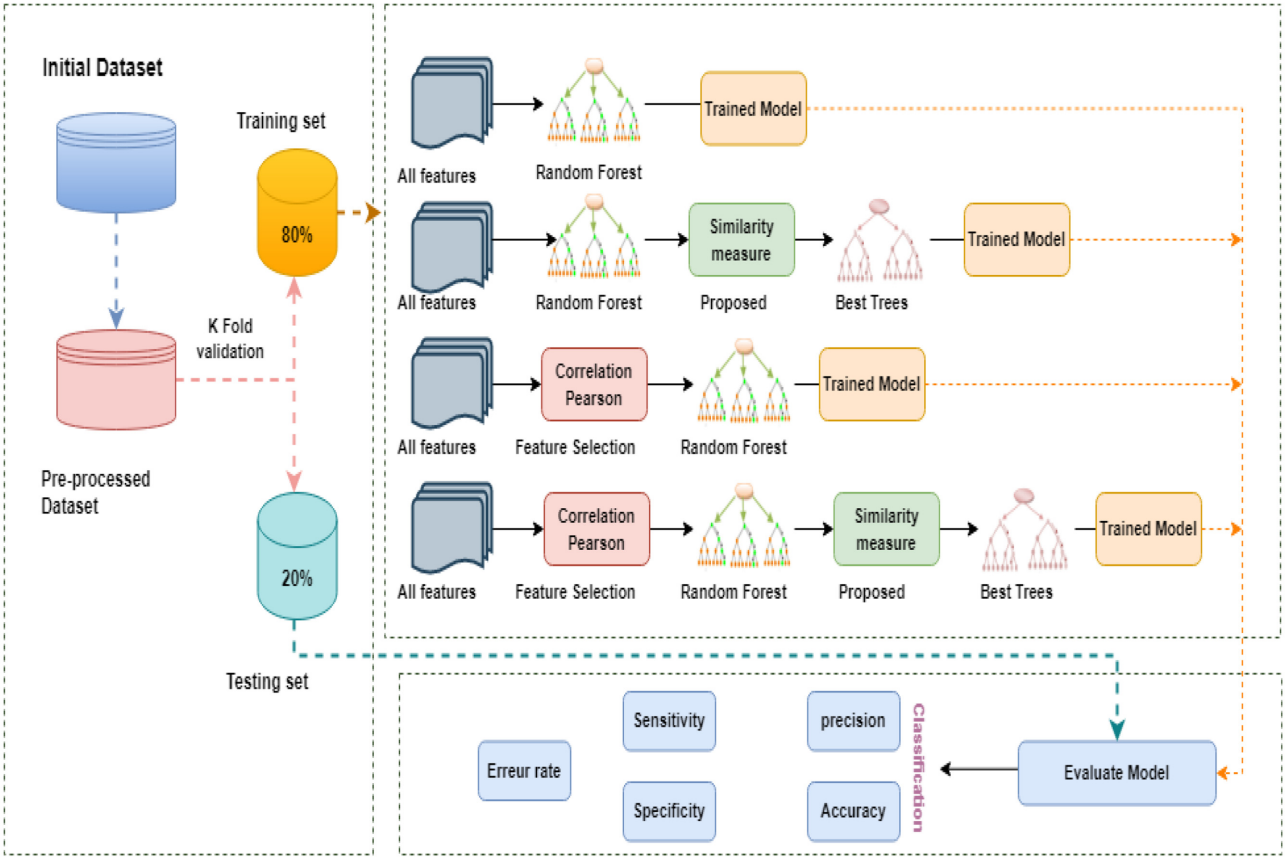


Fig. 1. Overall architecture of the proposed approach.

Table 1. Description of the 11 bases used

Database	Size	N° attribute	Class	Reference
Heart	303	14	Bi class	UCI
Lung cancer	59	7	Bi class	Kaggle
Hepatitis	155	20	Bi class	UCI
Breast cancer	100	28	Multi class	Algeria
Alzheimer	354	15	Multi class	UCI
EEG-EYE-STATE	14,980	15	Bi class	Kaggle
Transfusion	748	5	Bi class	Kaggle
Dermatology	366	35	Multi class	UCI
Prostate cancer	100	10	Bi class	UCI
Haberman	306	4	Bi class	UCI
Diabetes	768	9	Bi class	UCI

Data from the University of California Irvine Repository (https://archive.ics.uci.edu/datasets), Kaggle (https://www.kaggle.com/datasets), and the study of Refai et al., "Maintenance of a Bayesian network: application using medical diagnosis," Evolving Systems, vol. 7, pp. 187-196, 2016, https://doi.org/10.1007/s12530-016-9146-8.

databases that contain it, which we built to give them numeric classes. The label encoder is used to replace alphanumeric data with numeric classes while keeping the same data structure.

C. Classification Algorithms Used

For this work we implemented four different models. First, we applied standard RFs to all databases, and then we proposed a hybrid approach based on RFs and a similarity measure. We then integrated feature selection techniques to optimize the databases and obtain more accurate results. Finally, we combined similarity measures and feature selection techniques into a more accurate model.

1) Random Forest

The RF is a set-based automatic learning algorithm proposed by Breiman [1], An RF is a meta-estimator that fits a number of decision tree classifiers to various subsamples of the dataset and uses the mean to improve predictive accuracy and control over-fitting. The classifier needs the parameter n_estimator which defines the number of decision trees generated in our classifier model. In our work, we have used n_estimator =100, aiming to generate a large number of decisions.

2) Random Forest with proposed Similarity Measure

Recall that we used an RF classifier with n=100 (n: decision trees) in order to optimize our results. We calculate the accuracy using one tree at a time (100 iterations to run through all the trees) for each fold, after sorting these trees by a descending sort against the calculated accuracy (Algorithm 1).

Algorithm 1. Proposed similarity measure

Input: RF, Nb trees
Output: RF', Nb best trees
-Trained model with Nb trees.
-Calculate the precision of each tree
-Select a set of trees that have better precision with precision > seuil and remove the trees that have precision < seuil (Nb best tree).
-Trained model RF' with Nb best trees.

Algorithm 2 implements the second model with RFs and the similarity measure.

Algorithm 2. Proposed random forest with similarity measure (RF_Similarity)

1. Define number of K fold
2. Set model RF to best_model from RF
3. Create K Fold (kf) instance with a number of splits k and X data
4. For data X.train, X.test, y.train, y.test in kf
   a. Predict value (pv) for X.test using RF model
   b. Create a confusion matrix with y.test and pv
   c. Calculate performance measures

3) Random Forest with Features Selection

Algorithm 3 implements the third model with RFs and the feature technique.

Bivariate Pearson correlation produces a sample correlation coefficient, r, which measures the strength and direction of linear relationships between pairs of continuous variables. By extension, Pearson correlation assesses whether there is statistical evidence of a linear

relationship between the same pairs of variables in the population, represented by a population correlation coefficient, ρ (rho). Pearson correlation is a parametric measure.

Algorithm 3. Proposed Random forest with features selection (RF_FS)

1. Apply the correlation function to our data X and set the result to corr_columns
2. Define the number of K fold
3. Create K Fold (kf) instance with a number of splits k and X data
4. For data X.train, X.test, y.train, y.test in kf
   a. Remove corr_columns from X.train and y.train
   b. RF model fit with X.train and y.train
   c. Predict value (pv) for X.test using RF model
   d. Create a confusion matrix with y.test and pv
   e. Calculate performance measures
   f. If the current precision is the best, save the current model best_model

4) Random Forest with Features Selection and Similarity Measure

In this model, we have combined similarity measures (Section III-C-2) and feature selection techniques with RFs (Algorithm 4).

Algorithm 4. Proposed Random forest with features selection (correlation) and similarity measure (RF_FS_ Similarity)

1. Apply correlation function to our data X and set result to corr_columns
2. Define the number of K fold
3. Set model RF to best_model from Algo2
4. Create K Fold (kf) instance with a number of splits k and X data
5. For data X.train, X.test, y.train, y.test in kf
   a. Remove corr_columns from X.train and y.train
   b. Predict value (pv) for X.test using RF model
   c. Create a confusion matrix with y.test and pv
   d. Calculate performance measures

5) Validation Technique: K-Fold Cross-Validation

Cross-validation is mainly used in applied machine learning to estimate the competence of a machine learning model on unseen data. That is, we use a limited sample to estimate how the model should perform when used to make predictions on data not used when training the model. In our work, we used k=10 for all the databases. For each k (fold) we do:

Model fit.
- Prediction of test data using our fitted model (test phase).
- Generation of the confusion matrix.
- Performance measurement calculation.
- Choosing the right trees.

6) Evaluation Metrics Performance

The performance measures used to compare algorithms are accuracy, classification rate, sensitivity, specificity, and error rate. All performance measures are defined below.

Precision: The proportion of the number of instances correctly predicted as positive in relation to the total number of cases indicated as positive.

$$Precision = \frac{TP}{(TP + FP)}.$$

Accuracy: This is the number of well-ranked examples, in absolute terms, and then as a percentage of the total number of examples.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}.$$

Sensitivity: This is the ability to give a positive result when the disease is present.

$$Sensitivity = \frac{TP}{(TP + FN)}.$$

Specificity: This is the ability to give a negative result when the disease is absent.

$$Specificity = \frac{TN}{(TN + FP)}.$$

The error rate: Is the proportion of misclassified individuals. It estimates the probability of misclassifying an individual taken randomly from the population.

$$The\ error\ rate = \frac{(FP + FN)}{(TP + TN + FP + FN)}.$$

IV. EXPERIMENTAL AND RESULTS

In this section, we evaluate the multi-class classification of medical databases using the RF approach by comparing the results obtained by curves between RF_Standard, RF_Similarity, RF_FS, and RF_FS_Similarity for each database.

In our experiments, we used a standard RF and three other improved forests. We use the following eleven databases for each experiment: breast cancer, dermatology, EEG-Eye-State, Haberman, Diabetes, heart, hepatitis, lung cancer, Alzheimer's disease, prostate cancer, and transfusion. For the number of trees in the forest, we use 100 decision trees (nbtree).

For 100 trees we built four different RFs, where the first is a standard RF with the k cross-validation technique to divide our data into 10 different tests. The second forest is improved by adding a similarity measure. The third forest is improved by adding attribute selection

to the first standard method. The fourth forest is improved by adding a similarity measure to the third attribute selection method. We run the program ten times for each forest in order to compare the performance of the four methods on the eleven databases.

By comparing the results of the standard RF containing 100 decision trees for each database with each of the four methods, we note that the improvements provide better results than the standard method for all databases.

In Table 2 and Figs 2–3, we compare the accuracy performance of the four different RFs. We present the accuracy results for each evaluation criterion. We note a

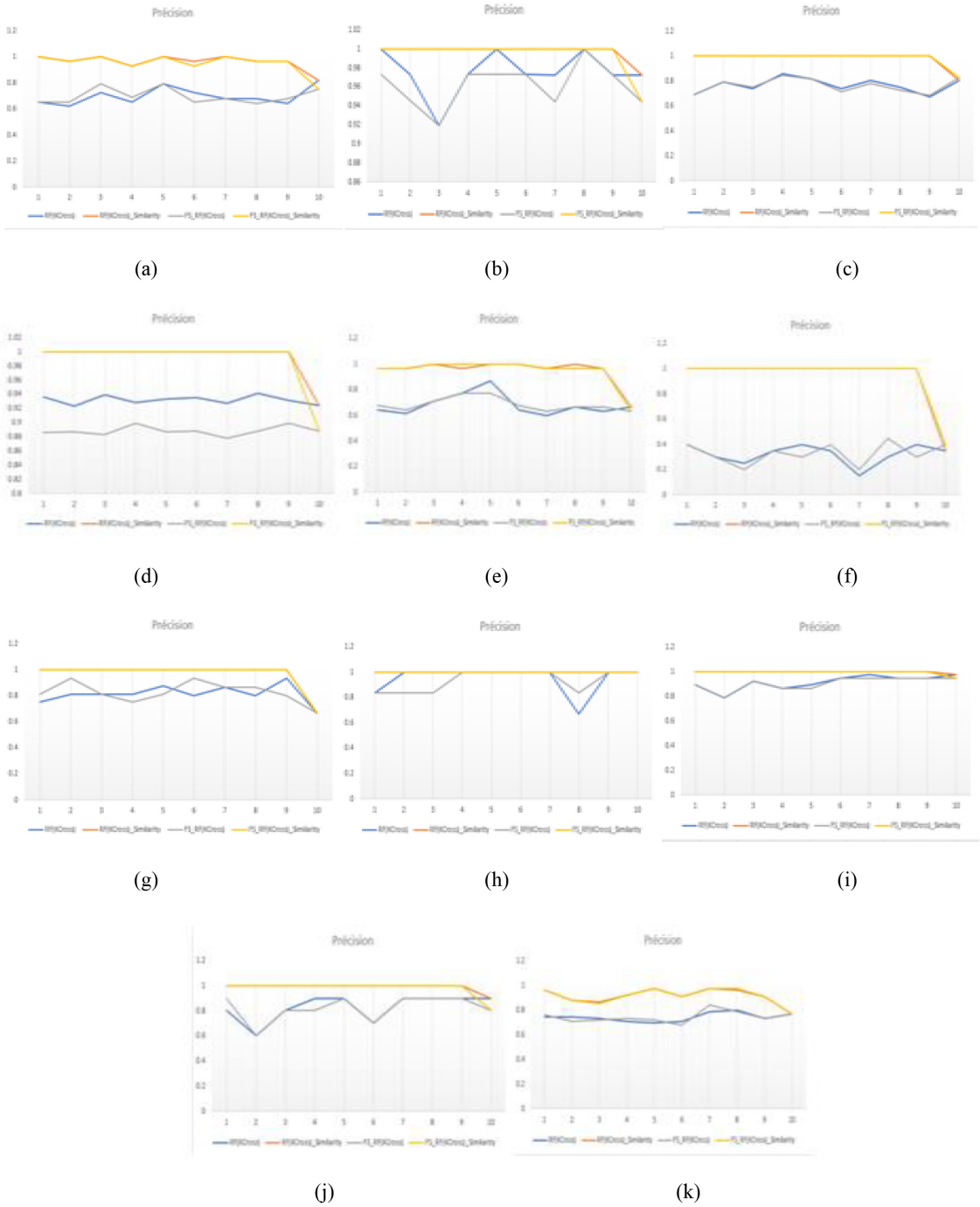


Fig. 2. Comparative accuracy curves obtained by the 4 classification methods on the 11 medical databases: (a) breast cancer, (b) dermatology, (c) diabetes, (d) EEG-Eye-State, (e) Haberman, (f) Heart, (g) hepatitis, (h) lung cancer, (i) Alzheimer, (j) prostate cancer, and (k) transfusion.

Table 2. Comparing the precision of all methods used for 11 medical databases

Dataset	Precision (%)			
	RF_Standard	RF_Similarity	RF_FS	RF_FS_Similarity
Breast cancer	69.94	96.12	69.91	95.06
Dermatology	97.55	99.72	96.18	99.44
Diabetes	76.68	98.03	76.30	98.29
EEG-Eye-State	93.20	99.25	88.84	98.88
Haberman	68.25	95.03	68.58	94.69
Heart	33.00	94.00	33.00	94.00
Hepatitis	81.29	96.67	82.58	96.67
Lung cancer	95.00	100	93.33	100
Alzheimer	91.46	99.73	90.65	99.46
Prostate cancer	83.00	99.00	82.00	98.00
Transfusion	74.20	91.16	74.47	91.16

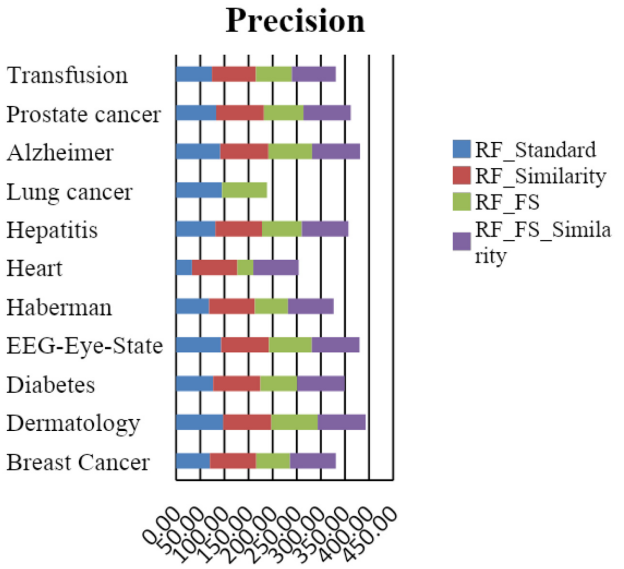


Fig. 3. Comparative histogram of the pricision results obtained by the 4 classification methods on the databases used.

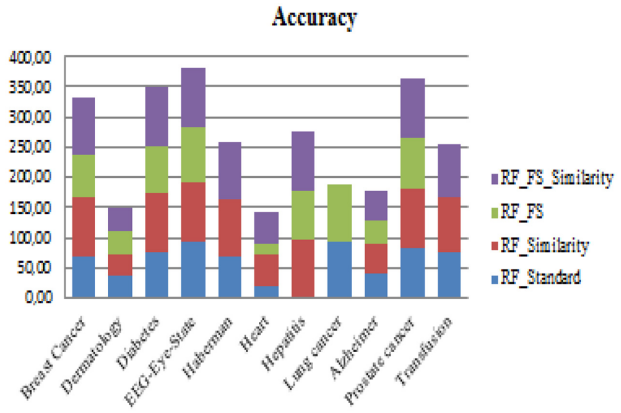


Fig. 4. Comparative histogram of accuracy results obtained by the 4 classification methods on the databases used.

stabilization of the accuracy rates for 100 decision trees for the standard method and the attribute selection method without the similarity measure. We also note that the two classifications with the similarity measure provide better results most of the time than the other two.

In Table 3 and Fig. 4, we compare the accuracy performances of the four different RFs. We present the results of the classification rates for each evaluation criterion, and we note a stabilization of the classification rates for 100 decision trees for the standard method and the method of selecting attributes without a similarity measure. We also note that the two classifications with the similarity measure provide better results most of the time than the other two.

In this work, we have implemented our database in four phases: the first phase is designed for the standard RF classification with the *k*-fold cross-validation technique, and the second phase is for the RF classification with similarity measure with which we have improved our system by noising the model of the best classification rate. In this method, only the decision trees with higher classification rates are retained, which improves the prediction performance compared to the standard method.

Subsequently, we applied the third method of RFs with attribute selection, where the Pearson correlation feature selection technique aims to select important features and ignore irrelevant features to increase the speed of training due to the reduction of inappropriate features and solve the problem of high dimensionality. It is therefore possible to apply it to different types of data features.

Lastly, we further improved our system by implementing

Table 3. Results of the accuracy obtained by the 4 classification methods on 11 medical databases

Dataset	Accuracy (%)			
	RF_Standard	RF_Similarity	RF_FS	RF_FS_Similarity
Breast cancer	69.94	96.12	69.91	95.06
Dermatology	36.11	36.11	38.89	38.89
Diabetes	76.68	98.03	76.30	98.29
EEG-Eye-State	93.20	99.25	88.84	98.88
Haberman	68.25	95.03	68.58	94.69
Heart	20.00	51.00	20.00	51.00
Hepatitis	81.29	96.67	82.58	96.67
Lung cancer	95.00	100	93.33	100
Alzheimer	41.02	48.76	40.21	48.49
Prostate cancer	83.00	99.00	82.00	98.00
Transfusion	74.20	91.16	74.46	91.16

the fourth RF method with simultaneous attribute selection and similarity measurement, and carried out a comparative study between them based on the same performance measures applied to choose the best method.

From these experiments, we can say that the four methods gave good results.

Tables 4 and 5 show the sensitivities and specificities obtained by implementing each algorithm. In this comparison between the proposed methods (Fig. 5), the two methods with similar blades obtained better performances by up to 100%.

Table 6 shows the error rates obtained by the proposed methods on all the databases. We note that the two methods using the similarity measure obtained the best error rate up to 0 on the dermatology and lung cancer databases using the RF_FS_Similarity method and on the lung cancer and Alzheimer databases using the RF_Similarity method.

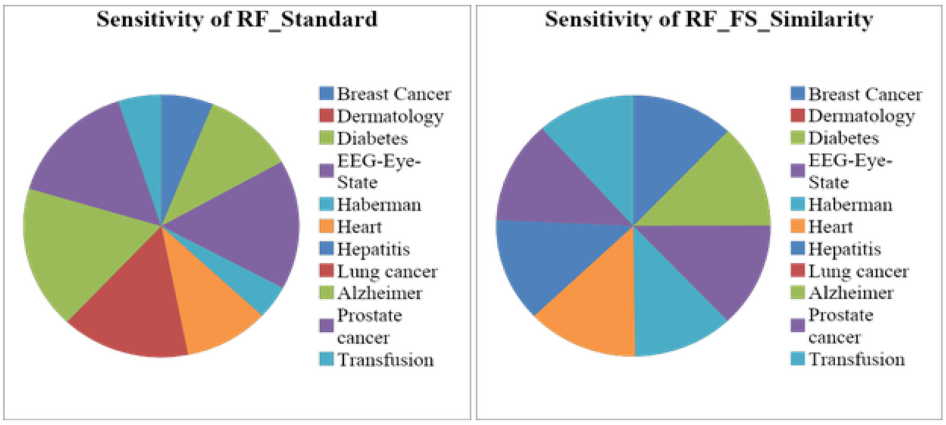
RF is a representative of integrated learning [2] due to its advantages, despite the efficiency of RFs obtained in different works; several works seek to continuously improve it. A standard RF model was applied in [20], but they only achieved a classification rate of 86%. Furthermore [19] and [21] used standard RFs with a feature selection technique (correlation): however, the classification rate of the obtained models is 97%, and 94%, respectively. Another model performed by [18] was used as an improved RF with the Pearson technique in a single database; they

Table 4. Sensitivity results obtained by the 4 classification methods on 11 medical databases

Dataset	Sensitivity (%)			
	RF_Standard	RF_Similarity	RF_FS	RF_FS_Similarity
Breast cancer	34.28	92.33	48.73	92.33
Dermatology	98.00	98.00	100	100
Diabetes	60.77	96.67	69.13	98.18
EEG-Eye-State	89.84	98.78	89.80	99.11
Haberman	23.34	88.02	34.84	90.40
Heart	56.00	96.00	58.50	100
Hepatitis	91.77	98.18	85.46	98.00
Lung cancer	85.00	100	84.17	100
Alzheimer	99.33	100	90.45	99.33
Prostate cancer	86.29	100	87.40	96.67
Transfusion	28.66	69.92	42.48	88.66

Table 5. Specificity results obtained by the 4 classification methods on 11 medical databases

Dataset	Specificity (%)			
	RF_Standard	RF_Similarity	RF_FS	RF_FS_Similarity
Breast cancer	85.57	97.64	75.56	96.49
Dermatology	100	100	100	100
Diabetes	85.15	98.78	79.67	98.33
EEG-Eye-State	95.94	99.63	88.11	98.71
Haberman	84.36	97.00	75.68	96.83
Heart	55.19	95.00	45.83	95.00
Hepatitis	43.17	93.00	71.67	93.33
Lung cancer	98.00	100	96.33	100
Alzheimer	33.33	100	30.00	9.00
Prostate cancer	70.00	98.33	65.24	100
Transfusion	88.13	97.90	80.28	91.56



(a) (b)

Fig. 5. Comparative histogram of the sensitivity results obtained by the two classification methods (standard RF and improved RF) on the databases used.

Table 6. Results of the error rates obtained by the 4 classification methods on 11 medical databases

Dataset	Error rate			
	RF_Standard	RF_Similarity	RF_FS	RF_FS_Similarity
Breast cancer	0.30	0.03	0.30	0.04
Dermatology	0.02	0.02	0	0
Diabetes	0.19	0.19	0.17	0.17
EEG-Eye-State	0.06	0.007	0.11	0.01
Haberman	0.31	0.04	0.31	0.05
Heart	0.17	0.02	0.19	0.01
Hepatitis	0.33	0.33	0.33	0.33
Lung cancer	0.05	0.0	0.06	0.0
Alzheimer	0.03	0.0	0.04	0.002
Prostate cancer	0.17	0.01	0.18	0.02
Transfusion	0.25	0.08	0.25	0.08

Table 7. Performance comparison between the proposed model and other research models

Methods	Techniques used	Used database	Accuracy (%)
Rawashdeh et al. [20]	SMOTE + Standard RF	Preterm birth	86
Liu et al. [19]	Correlation + Standard RF	Fetal heart rate	97
Li et al. [21]	Correlation + Standard RF	Postictal generalized electroencephalogram	94
Bi et al. [18]	Pearson + CERF	Alzheimer	86
Our approach		11 bases medicals	
RF_Standard	Standard RF		Max 95
RF_Similarity	Standard RF + Similarity measure		Max 100
RF_FS	Standard RF + Pearson correlation		Max 93
RF_FS_Similarity	Standard RF + Pearson correlation + Similarity measure		Max 100

achieved only an 86% classification rate. Therefore, the RF model established in this study with these three variants has a better predictive effect on 11 medical databases; details of the techniques used in the performance of our model compared to other studies are provided in Table 7.

This comparative analysis indicates that the proposed model is comparable to the various classification models present in the relevant literature: however, the implementation of the enhanced RF in a clinical environment may assist physicians in making clinical decisions.

V. CONCLUSION

In this work, we studied the performance of a set model called RF on a classification task related to the medical field. To this end, we first analyzed the work carried out in the field, which highlighted several advantages as well as certain limitations of RFs used in the classification of medical data.

As a result, we found that the classifiers already proposed based on RFs perform well, but can still be improved to bring greater precision to the results.

To create a high-performance application used for classification, we proposed three methods that use several variations of RFs. We used eleven UCI and Kaggle databases to evaluate our model. We first re-implemented the RF-standard using the *k* cross-validation technique, and then proceeded to develop several variants of the same classifier using the attribute selection method and a proposed similarity measure.

We evaluated and tested the performance of each forest in terms of accuracy, sensitivity, specificity, classification rate, and error rate. The results obtained with our four methods are among the best ever obtained for the classification of these databases. The results obtained are highly competitive with other versions of RFs.

In the future, we hope to make further improvements on RFs. We also plan to integrate this application into a medical diagnosis aid system for use in hospitals or by surgeons.

CONFLICT OF INTEREST

The authors have declared that no competing interests exist.

REFERENCES

1. L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5-32, 2001. https://doi.org/10.1023/A:1010933404324
2. L. Gadomer and Z. A. Sosnowski, "Pruning trees in C-fuzzy random forest," *Soft Computing*, vol. 25, pp. 1995-2013, 2021. https://doi.org/10.1007/s00500-020-05270-3
3. C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," *Journal of Biomedical Science and Engineering*, vol. 6, no. 5, pp. 551-560, 2013. http://dx.doi.org/10.4236/jbise.2013.65070
4. F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier," in *Proceedings of 2013 13th International Conference on Intelligent Systems Design and Applications*, 2013, pp. 121-125. https://doi.org/10.1109/ISDA.2013.6920720
5. S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis," *Applied Soft Computing*, vol. 86, article no. 105941, 2020. https://doi.org/10.1016/j.asoc.2019.105941
6. S. Kabiraj, M. Raihan, N. Alvi, M. Afrin, L. Akter, S. A. Sohagi, and E. Podder, "Breast cancer risk prediction using XGBoost and random forest algorithm," in *Proceedings of 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2020, pp. 1-4. https://doi.org/10.1109/ICCCNT49239.2020.9225451
7. Y. Ono and Y. Mitani, "Effect of the random forests with recursive

feature elimination for breast cancer classification," in *Proceedings of 2021 6th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Oita, Japan, 2021, pp. 95-96. https://doi.org/10.1109/ICIIBMS52876.2021.9651590
8. S. M. M. Hasan, M. A. Mamun, M. P. Uddin, and M. A. Hossain, "Comparative analysis of classification approaches for heart disease prediction," in *Proceedings of 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, Rajshahi, Bangladesh, 2018, pp. 1-4. https://doi.org/10.1109/IC4ME2.2018.8465594
9. A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection," *IEEE Access*, vol. 7, pp. 180235-180243, 2019. https://doi.org/10.1109/ACCESS.2019.2952107
10. S. Asadi, S. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," *Journal of Biomedical Informatics*, vol. 115, article no. 103690, 2021. https://doi.org/10.1016/j.jbi.2021.103690
11. S. Benbelkacem and B. Atmani, "Random forests for diabetes diagnosis," in *Proceedings of 2019 International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, 2019, pp. 1-4. https://doi.org/10.1109/ICCISci.2019.8716405
12. A. Hebbar, M. Kumar, and H. A. Sanjay, "DRAP: decision tree and random forest based classification model to predict diabetes," in *Proceedings of 2019 1st International Conference on Advances in Information Technology (ICAIT)*, Chikmagalur, India, 2019, pp. 271-276. https://doi.org/10.1109/ICAIT47043.2019.8987277
13. K. VijiyaKumar, B. Lavanya, I. Nirmala, and S. S. Caroline, "Random forest algorithm for the prediction of diabetes," in *Proceedings of 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India, 2019, pp. 1-5. https://doi.org/10.1109/ICSCAN.2019.8878802
14. M. Pavithra and B. T. Geetha, "Prediction of chronic kidney cancer using RBF support vector machine compared with random forest for better accuracy," in *Proceedings of 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, Chennai, India, 2022, pp. 1-5. https://doi.org/10.1109/ICSES55317.2022.9914342
15. B. S. Wade, S. H. Joshi, B. A. Gutman, and P. M. Thompson, "Machine learning on high dimensional shape data from subcortical brain surfaces: a comparison of feature selection and classification methods," *Pattern Recognition*, vol. 63, pp. 731-739, 2017. https://doi.org/10.1016/j.patcog.2016.09.034
16. L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, "Simultaneous spectral-spatial feature selection and extraction for hyperspectral images," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 16-28, 2018. https://doi.org/10.1109/TCYB.2016.2605044
17. Y. Tian, J. Yang, M. Lan, and T. Zou, "Construction and analysis of a joint diagnosis model of random forest and artificial neural network for heart failure," *Aging*, vol. 12, no. 24, article no. 26221, 2020. https://doi.org/10.18632/aging.202405
18. X. A. Bi, X. Hu, H. Wu, and Y. Wang, "Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2973-2983, 2020. https://doi.org/10.1109/JBHI.2020.2973324
19. L. Liu, Y. Jiao, X. Li, Y. Ouyang, and D. Shi, "Machine learning algorithms to predict early pregnancy loss after in vitro fertilization-embryo transfer with fetal heart rate as a strong predictor," *Computer Methods and Programs in Biomedicine*, vol. 196, article no. 105624, 2020. https://doi.org/10.1016/j.cmpb.2020.105624
20. H. Rawashdeh, S. Awawdeh, F. Shannag, E. Henawi, H. Faris, N. Obeid, and J. Hyett, "Intelligent system based on data mining techniques for prediction of preterm birth for women with cervical cerclage," *Computational Biology and Chemistry*, vol. 85, article no. 107233, 2020. https://doi.org/10.1016/j.compbiolchem.2020.107233
21. X. Li, S. Tao, S. Jamal-Omidi, Y. Huang, S. D. Lhatoo, G. Q. Zhang, and L. Cui, "Detection of postictal generalized electroencephalogram suppression: random forest approach," *JMIR Medical Informatics*, vol. 8, no. 2, article no. e17061, 2020. https://doi.org/10.2196/17061
22. M. Z. Alam, M. S. Rahman, and M. S. Rahman, "A random forest based predictor for medical data classification using feature ranking," *Informatics in Medicine Unlocked*, vol. 15, article no. 100180, 2019. https://doi.org/10.1016/j.imu.2019.100180
23. B. Kalaiselvi and M. Thangamani, "An efficient Pearson correlation based improved random forest classification for protein structure prediction techniques," *Measurement*, vol. 162, article no. 107885, 2020. https://doi.org/10.1016/j.measurement.2020.107885
24. M. Song, H. Jung, S. Lee, D. Kim, and M. Ahn, "Diagnostic classification and biomarker identification of Alzheimer's disease with random forest algorithm," *Brain Sciences*, vol. 11, no. 4, article no. 453, 2021. https://doi.org/10.3390/brainsci11040453
25. M. Velazquez and Y. Lee, "Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects," *PLOS One*, vol. 16, no. 4, article no. e0244773, 2021. https://doi.org/10.1371/journal.pone.0244773
26. C. Zhang, J. Ren, F. Liu, X. Li, and S. Liu, "Three-way selection random forest algorithm based on decision boundary entropy," *Applied Intelligence*, vol. 52, pp. 13384-13397, 2022. https://doi.org/10.1007/s10489-021-03033-7
27. I. Tarchoune, A. Djebbar, and H. F. Merouani, "A hybrid CBR classification model by integrating decision tree and random forest into case retrieval," in *Proceedings of 2021 International Conference on Networking and Advanced Systems (ICNAS)*, Annaba, Algeria, 2021, pp. 1-6. https://doi.org/10.1109/ICNAS53565.2021.9628920
28. I. Tarchoune, A. Djebbar, H. F. Merouani, and D. Hadji, "An improved random forest based on feature selection and feature weighting for case retrieval in CBR systems: application to medical data," *International Journal of Software Innovation*, vol. 10, no. 1, pp. 1-20, 2020. https://doi.org/10.4018/IJSI.293265
29. I. Tarchoune, A. Djebbar, and H. F. Merouani, "A case-based reasoning system-based random forest for classification: a systematic literature review," in *Handbook of Research on Driving Socioeconomic Development with Big Data*. Hershey, PA: IGI Global, 2023, pp. 170-196. https://doi.org/10.4018/978-1-6684-5959-1.ch008

Ilhem Tarchoune http://orcid.org/0000-0002-4931-6458

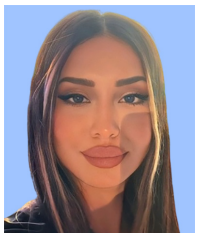
Ilhem Tarchoune is currently pursuing her Ph.D. at the LRI Laboratory at the Department of Computer Sciences, Badji Mokhtar University, Annaba, Algeria. Her current research interests include: case-based reasoning (CBR), random forest (RF), hybrid CBR with RF and computer-aided diagnosis systems.

Akila Djebbar http://orcid.org/0009-0005-4699-8755

Akila Djebbar is currently associate professor (PHD, HDR) at the university of Badji Mokhtar-Annaba, Algeria. She received her engineer in Computer Science in 2002 from Annaba University. She received a Magister in "Distributed Artificial Intelligence" in 2006. She has her Ph.D. (2013) in artificial intelligence. Dr Akila Djebbar is one of reviewer of *International Journal of Business Intelligence and Data Mining* (IJBIDM) and *International Journal of Intelligent Information and Database Systems* (IJIIDS). She is a doctor at computer science department of Badji Mokhtar Annaba University (Algeria) and member of LRI Laboatory. Dr Akila Djebbar has authored more than 50 research papers. Her research interests are in: machine learning, case based reasoning approach, uncertain knowledge, supervised learning, semi-supervised learning, medical image classification, CAD system.

Hayet Farida Merouani http://orcid.org/0000-0001-9530-1663

Hayet Farida Merouani received her engineer degree at the University of Badji Mokhtar, Annaba, Algeria in 1984, followed by a Ph.D. degree from Robert Gordon University, Aberdeen, UK. Currently, she is a full professor at the Badji Mokhtar University, Annaba. She leads the Computer Laboratory at Badji Mokhtar University more than five years. She also leads Research Group of Pattern Recognition, as a National Program Research of Breast Cancer. She also leads several national projects. Prof. Hayet Farida Merouanui has authored more than 100 research papers publication. She is a permanent reviewer member of *Artificial Intelligence in Medicine* (Springer), in Scientific Research and Essays (academic journals), and in *Journal of King Saud University - Computer and Information Sciences* (Springer). Her current works focus on the computer vision, pattern recognition and data mining, medical imaging, CAD system, artificial Intelligence.



Harfi Rania http://orcid.org/0009-0006-1192-0015

Harfi Rania recieved her master's degree in computer science in 2021 at the Department of Computer Sciences, Badji Mokhtar University, Annaba, Algeria University. Her current research interests include: machine learning, random forest, and computer aided diagnosis (CAD).