Diversity Enhancement Random Forest Model for the Risk Identification of Disease Deterioration

Jun-Feng Peng (pengjunf@mail2.sysu.edu.cn)
Guangdong University of Education https://orcid.org/0000-0002-3039-0308
Xing-Ji Chen
Hezhou University
Xiao-Xin Li
Hezhou University
Mi Zhou
Third Affiliated Hospital of Sun Yat-Sen University
Jun Xu
Guangdong University of Education
Xiong-Yong Zhu
Guangdong University of Education
Jia-Yuan Chen
Guangzhou Medical University Second Affiliated Hospital
Kai-Qiang Zou
Guangdong University of Education
Zhan Zhang
Sun Yat-Sen University
Guo-Ming Chen
Guangdong University of Education

Research article

Keywords: random forest, medical decision-making support, logarithmic loss function, greedy stepwise backward search

Posted Date: December 8th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-120643/v1

Diversity enhancement random forest model for the risk identification of disease deterioration

Jun-Feng Peng[1*], Xing-Ji Chen[4],Xiao-Xin Li[4,] ,Mi Zhou[2], Jun Xu[1], Xiong-Yong Zhu[1], Jia-Yuan Chen[3],Kai-Qiang Zou[1], Zhan Zhang[5] ,Guo-Ming Chen[1]

[1] Department of Computer Science, Guangdong University of Education, Guangzhou 510303, Guangdong, China.

[2] The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong, China.

[3] The Second Affiliated Hospital, Guangzhou Medical University, Guangzhou, Guangdong, China.

[4] Hezhou University, School of Artifical Intelligence (Modern Industry College), Hezhou, Guangxi, China.

[5] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, Guangdong, China.

* Corresponding Author: Jun-Feng Peng

E-mail: pengjunf@mail2.sysu.edu.cn ; Tel: +86-020-36967761.

Abstract

Background

Random forest (RF) is a powerful ensemble algorithm for medical decision-making supporting (MDS). However the requirement of higher accuracy and smaller ensemble size remain significant burdens for the current RF, particularly for the risk

identification of disease deterioration. To achieve the goal of higher accuracy and smaller ensemble size for the risk identification of disease deterioration, a diversity enhancement random forest (DERF) model is proposed.

Methods

We explored the idea of integrating trees that are accurate and diverse to build the DERF model. First, we calculated the accuracy of the out of bag data to select the best K trees. Then, we assessed the diversity of these trees using logarithmic loss functions on the validation data set. Further, we utilized the greedy stepwise backward search to increase the diversity of the random forest. Finally, public bench mark data sets on disease deterioration from KEEL and real data sets from tertiary hospitals in the last three years were used to assess the performance of the proposed DERF model and compared it with the existing model.

Results

Experiments show that the proposed model can improve the prediction performance and reduce the ensemble size of random forest model. Compared with the existing model random forest, the extreme random tree and the ensemble of optimal tree, our proposed DERF model obtains a higher predictive accuracy and a smaller ensemble size.

Conclusion

It reveals that the proposed DERF could reduce the size of the ensemble and achieve good classification results in the risk identification of disease deterioration

Keywords random forest, medical decision-making support, logarithmic loss function, greedy stepwise backward search.

Background

Random forest (RF) algorithm was first proposed by Tin Kam Ho at 1995 [1]. Leo Breiman refined the ensemble methods with random selections of features at each node to build a forest in 2001 [2]. Delgado et al. verified the classification performance of 179 classification algorithms on 121 UCI standard data sets, and the experimental results showed that the random forest algorithm had the best cl**assi**fication performance [3]. Random forest is a commonly used machine learning model, which is widely used in various fields because of its high prediction accuracy, robustness and fast calculation. Random forest has also shown great potential for the disease aided decision support, recently.

In the aspect of direct application of random forest to auxiliary diagnosis, Chowdhury et al. (2018) employed random forest to diagnose retinal abnormalities and got 93.58% accuracy, which demonstrated the great potential of random forest to assist in clinical decision-making [4]. Geetha et al. (2019) utilized random forest for the cervical cancer diagnosis based on the oversampling technique and principal component analysis [5]. Random forest were also widely used to assist in the diagnosis of osteoarthritis, diabetes, chronic kidney disease, bone marrow disease and other diseases [6-9].

In terms of auxiliary diagnosis through model improvement, Guo et al. (2020)

proposed a recursion enhanced random forest by finding the key features of the prediction of cardiovascular diseases with an improved linear model (RFRF-ILM) to detect heart disease [10]. Fawagreh et al (2020) proposed to use a resource-efficient fast prediction model using data clustering to improve the speed and accuracy of the method [11]. Wang et al (2020) proposed the Random Forest (RF)-based rule extraction (IRFRE) method to derive accurate and interpretable classification rules from a decision tree ensemble for breast cancer diagnosis [12].

However, good predictive performance and high interpretation remain the main measure of the effectiveness of the model in the medical diagnosis scene. The improvements of the individual trees and their diversity in the random forest are the main way to achieve goal of the good predictive performance. There has been a significant work done on the problem of minimizing this number to reduce computational cost without decreasing prediction accuracy as the number of trees in random forest is often very large [13-16]. Zhang and Wang (2009) utilized both the tree similarity method and the prediction method to reduce the size of the forest. They called this method the “By similarity method” [17]. Coskun et al. applied the ensemble pruning techniques to reduce the computational cost without relevant loss of performance and applied it to the problem of an early detection of glaucoma, a severe eye disease with low prevalence [18]. Khan et al. employed the smallest individual prediction error, unexplained variance and Brier score [19] to evaluate the individual accuracy and diversity for the optimal trees selection [20].

Inspired by [20], we proposed a diversity enhancement random forest model using diversity enhancement to improve the identification rate of disease deterioration risk and reduce the size of the random forest. Based on the above discussion, our paper aims to select the best trees, in terms of individual strength i.e. accuracy and diversity, from a large ensemble grown by random forest. The results from the new method are compared with those of random forest, Extreme Tree and the ensemble of optimal tree on the bench mark data sets. The main contributions of this paper are summarized as follows:

- An improved random forest model with higher predictive accuracy and a smaller ensemble size is proposed to identify the diseases deterioration risk.
- Both public bench mark data sets from KEEL and data sets from the tertiary hospitals in the last three years are used to evaluate the proposed model.

The rest of the paper is organized as follows. The proposed diversity enhancement random forest (DERF), the underlying algorithm and some other related approaches are given in Sect. 2, experiments and results based on benchmark and simulated data sets are given in Sect. 3. Finally, Sect. 4 gives the conclusion of the paper.

Methods

Diversity enhancement is an effective way to optimize random forest by selecting good but different decision trees. The principle of our proposed DERF is described as follows. First, random forest is optimized by OOB data error rate. Then, the logistic loss function is introduced to evaluate the diversity of OOB optimized random forest.

Out of Bag (OOB) optimization

Out of Bag (OOB) optimization is utilized as the first round optimization of random forest, can not only reduce the ensemble size, but also improve the prediction performance. The principle of this method is described as follows. The random forest model utilizes Bootstrap sampling to select training samples for each decision tree. Assuming that the size of the training sample is m, Bootstrap sampling method randomly selects m samples from the training sample. The probability of each sample being selected is $\frac{1}{m}$, which means that the probability not being selected is $1-\frac{1}{m}$. If the sampling is repeated n times, then the probability of one sample will not be selected can be expressed as :

$$q = (1-\frac{1}{m})^m \tag{1}$$

When the training sample size m is large enough, the limit of q can be represented as:

$$q = \lim_{m\to\infty}(1-\frac{1}{m})^m \approx \frac{1}{e} \approx 0.368 = 36.8\% \tag{2}$$

This means that 36.8% of the training samples will not be used to build the corresponding decision tree. This unused sample is called Out-of-Bag (OOB) data. Each decision tree in the random forest has a unique corresponding OOB data that is equivalent in function to test data. For each decision tree of random forest, OOB data plays the same role as test data. OOB data can be exploited for the Out-of-bag Estimate, which has been proved to be an effective method to measure the generalization of the random forest [21]. The Out-of-bag Estimate is defined as

follows：

$$e^{oob} = \frac{1}{|EXD|}(\sum_{i=1}^{N}[g(x_i) \neq y_i] \tag{3}$$

Where, $g(x_i)$ indicates the i-th basic classifier, $|EXD|$ represents the size of OOB data, y_i is the true category of input instance x_i. OOB optimization employs the Out-of-bag Estimate to improve the prediction of the random forest.

Logarithmic loss based greedy stepwise backward search optimization

Logarithmic loss function

Logarithmic loss function, also called log-likelihood Loss or cross-entropy loss，is defined on the basis of probability estimation. It is often used to evaluate the probability output of a classifier. It quantifies the accuracy of the classifier by punishing the error classification. Minimizing the logarithmic loss is basically equivalent to maximizing the accuracy of the classifier. The calculation formula for the logarithmic loss function is defined below：

$$\mathrm{L(Y,P(Y|X)=-logP(Y|X)=-}\frac{1}{\mathrm{N}}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(\mathrm{p}_{ij}) \tag{4}$$

Where, Y is the output label, X is the input variables, L is the loss function, N is the input sample size, M is the number of possible categories, and p_{ij} is a two-value indicator, indicating whether category j is the real category of input instance x_i . p_{ij} denotes the probability that the input instance x_i belongs to category j of the classifier.

For binary classification, the formula （4） can be reduced to （5）.

$$-\frac{1}{\mathrm{N}}(y_i \log p_i + (1-y)\log(1-p_i)) \quad (5)$$

Where, y_i is the true category of input instance x_i, p_i is the probability that the prediction input instance x_i belongs to category 1. The logarithmic loss of all samples represents the average of the logarithmic loss per sample. The classifier is perfect if the logarithmic loss is 0.

Random step size greedy backward search

Considering that it is the NP complete problem to find the optimal trees in the forest, we use heuristic algorithm to search for the optimal trees to build a relatively good forest. Starting from the full M trees MT generated by out of bag optimization, we rank the MT by logarithmic loss values in reverse order and then remove a part of the trees from the set of trees set MT each time to make evaluation value is optimal after the removal of trees. The backward search ceases if the performance of the trees generated by backward selection as expected. Otherwise, random forest is made up of the top half of the M trees by default.

The diversity enhancement random forest algorithm

The frame of algorithm DERF is as follows:

1. Take T bootstrap samples from the given portion of the training data TR =(S1,S2).

2. Grow classification trees on all the bootstrap samples using random forest method.

3. Rank the trees in descending order with respect to their out of bag (OOB) errors on OOB data set. Choose the last M trees with the lowest individual OOB errors.

4. Calculate the logarithmic loss of each tree on the subset S2,

5. Sort the M trees with the logarithmic loss values in reverse order.

6. Select the best tress with the K smallest logarithmic loss values to build the final random forest from M trees using the greedy stepwise backward search.

The proposed DERF method is evaluated on the 5 disease data set. 90% of the total data (TR) is used for training data while the remaining 10% is used for test data. 90% of TR data (S1) is utilized to generate a certain number independent classification and regression trees using bootstrap methods along with randomly selecting a certain number features for splitting the nodes of the trees. The remaining 10% TR data (S2) is employed to check the diversity of the trees using the logarithmic loss function. Further, the greedy stepwise backward search method is applied to improve the performance of the random forest.

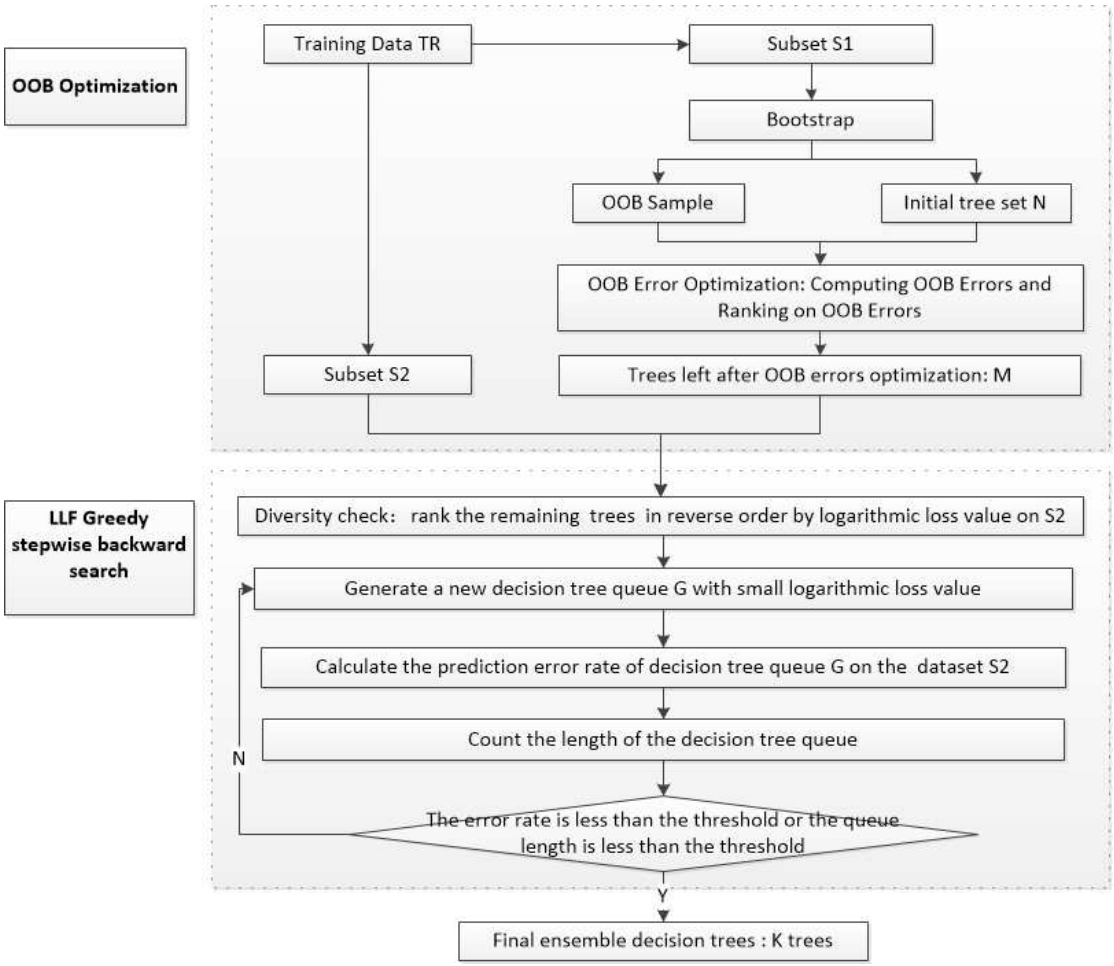

Fig. 1 The frame of algorithm DERF.

Results

For assessing the performance of DERF, 7 data sets are introduced. 5 disease binary classification data sets are downloaded from KEEL (Knowledge Extraction based on Evolutionary Learning). To further verify the effectiveness of the model, another 2 data sets about chronic obstructive pulmonary disease (AECOPD) and chronic respiratory disease (CRD) from tertiary hospitals in the last three years are collected. A brief summary of the data sets is given in Table 1. n, d denotes the number and the features of the sample, respectively. In terms of model comparison, random forest

(RF), extreme random tree (ET) and optimal tree ensemble (OTE by Khan 2020) are introduced.

Data set	n	d	source
mammographic	961	5	https://sci2s.ugr.es/keel/dataset.php?cod=86
Spectfheart	267	46	https://sci2s.ugr.es/keel/dataset.php?cod=185
hepatitis	80	20	https://sci2s.ugr.es/keel/dataset.php?cod=100
saheart	462	9	https://sci2s.ugr.es/keel/dataset.php?cod=184
heart	270	13	https://sci2s.ugr.es/keel/dataset.php?cod=99
AECOPD	408	9	Third Affiliated Hospital, Sun Yat-sen University
CRD	730	11	The Second Affiliated Hospital of Guangzhou Medical University

Table 1 Data set for disease classification with total number of observation n, number of feature d, and data source.

First, we investigate the factors that influence classification accuracy. The factors include the optimal total number T, the percentage M of best trees and the number of features d. For the sake of simplicity, we first probe the optimal total number T of trees grown before the selection process. Then, we explore the ratio M of best trees checked by the OOB errors. Next, we search the number of features d for the node splitting. Finally, we depicted the ensemble size reduction effect of DERF. We implement the DERF classifier on the development platform of Python 3.6.

Considering stability and algorithmic efficiency, large values are recommended for the size of the initial set under the available computation resources and a value of T $\geq$500 is expected to work well in general. Figure 2 illustrates the effect of the number

of trees in the initial set on the misclassification accuracy for the data sets given using DERF.

One important parameter of our method is the number M of best trees selected at the first phase for the final ensemble. Various values of M reveal different behaviour of the method. We considered the effect of M = (10%, 20%, …, 80%) of the total T trees on the method for both regression and classification as shown in Fig. 3. It was clear from Fig. 3 that the highest accuracy was obtained by using the small portion,30–60%, of the total trees that are individually strong which is further reduced in the second phase. This may significantly decreased the 40-70% storage costs of the ensemble while without loos the accuracy. On the other hand, having a small number of trees (less than 30%) can reduce storage costs of the resulting ensemble but also decrease the overall prediction accuracy of the ensemble.

The effect of various numbers of features selected were also investigated at random for splitting the nodes of the trees on the classification accuracy in the cases of classification for the data sets. The graph is shown in Fig. 4. The only reason that random forest is considered as an improvement over bagging is the inclusion of additional randomness by randomly selecting a subset of features for splitting the nodes of the tree. The effect of this randomness can be seen in Fig. 4 where different values of d results in different classification accuracy for the data sets. For example in the case of heart data, selecting a higher value of d adversely affects the performance. For some data sets, Specheart for example, selecting large d results in better performance.

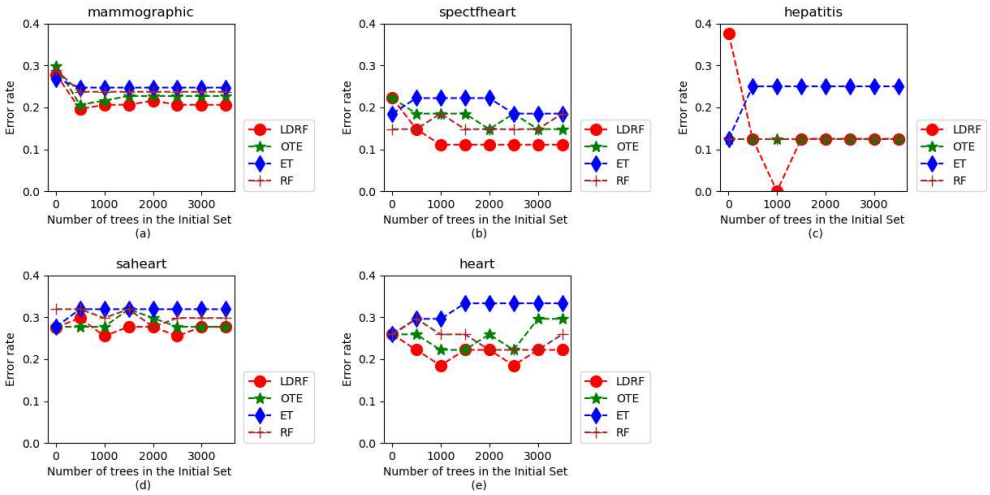

Fig. 2 The effect of the number of trees in the initial set on classification accuracy for the data sets given using LFRF, OTE, ET and RF.

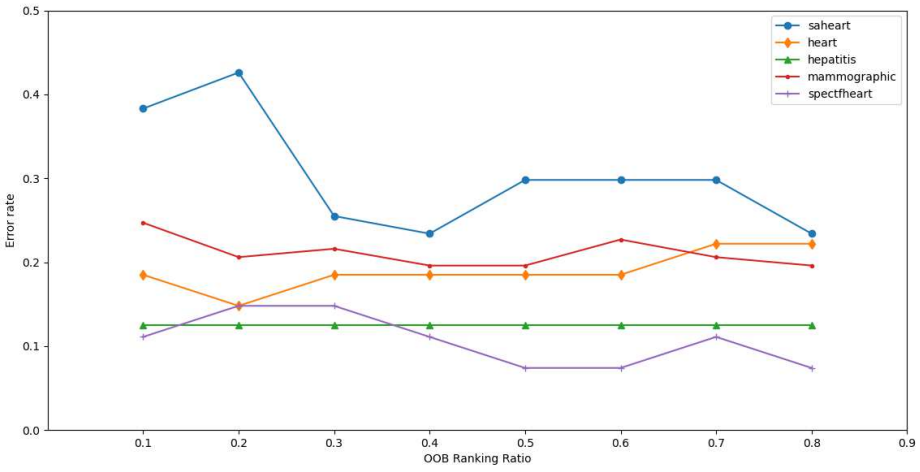

Fig. 3 Effect of M on the OOB classification accuracy, of the data sets shown using DERF. The value of M in percentage is on the x-axis and classification accuracy on the y-axis.

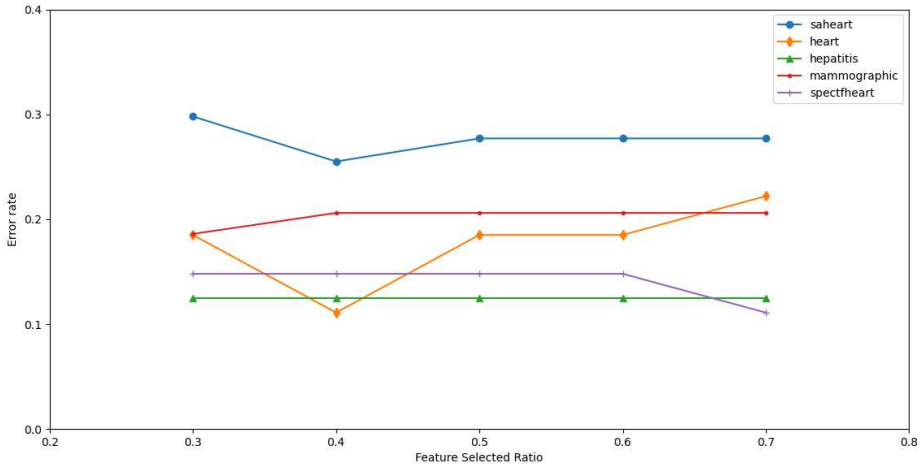

Fig. 4 Effect of the number of features (on x-axis) selected at random for splitting the nodes of the trees on classification accuracy for the data sets shown using DERF.

Figure 5 reveals the relationship between the size (number) of the initial decision trees and the size (number) of the final ensemble decision trees in the DERF. It reveals that DERF has the effect of ensemble size reduction. If the final ensemble size of DERF is equal to the initial decision size, these points will match the linear function $y = x$. Similarly, If the final ensemble size of DERF is half of the initial decision size, these points will match the linear function $y = \frac{1}{2}x$.We can infer from Figure 5, on KEEL data set, the final ensemble size of DERF model is less than 1/2 of the size of the initial decision tree, which means that DERF model reduces the ensemble size by half. Experiments on the KEEL data set show that the DERF model can reduce the ensemble size of random forest without compromising the accuracy of predictions.

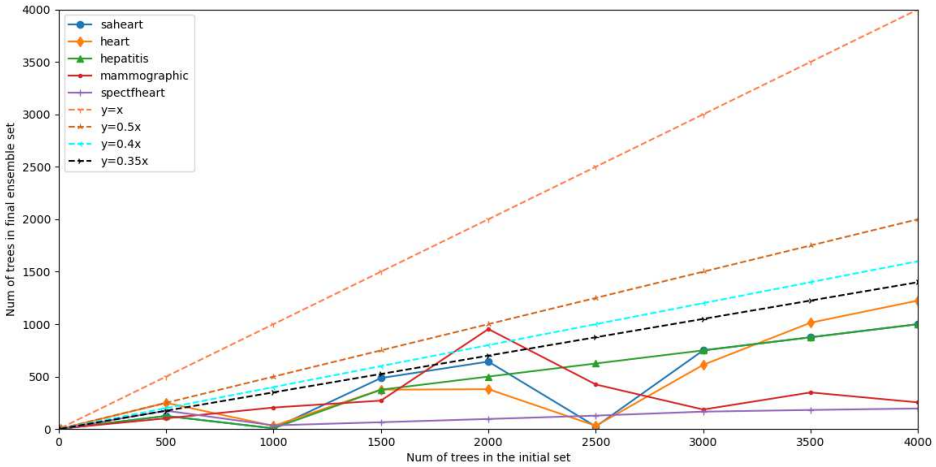

Fig. 5 Ensemble size reduction of DERF. Number of trees in initial set (on x-axis) selected for the generation of random forest. Number of trees in final ensemble set (on y-axis) indicated the final ensemble size of random forest.

Table 2-4 shows that the classification accuracy and final ensemble size produced by the proposed DERF on the KEEL data sets. Overall, we found that the proposed method DERF performed better the other methods on the 3 data sets (saheart, mammographic and spectfheart), and was comparable to the other methods on the 2 data set (heart and hepatitis). The numbers in parentheses represent trees in a random forest. The smaller the number means the smaller ensemble size of the random forest. We discovered that the proposed DERF model used smaller ensemble size to achieve better performance than OTE, ET and RF in general. The result of the best performing method for the corresponding data set was shown in bold.

Table 2 Classification accuracy of RF, ET, OTE and LFRF (number of trees in the initial set: 100).

Data set	n	d	RF	ET	OTE	LFRF
saheart	462	9	0.681(100)	0.702(100)	**0.723(25)**	0.702(25)
heart	270	13	0.741(100)	0.741(100)	0.778(25)	**0.815(25)**
hepatitis	80	20	0.75(100)	0.75(100)	**0.875(25)**	**0.875(25)**
mammographic	961	5	0.763(100)	0.753(100)	0.773(25)	**0.804(20)**
spectfheart	267	46	0.852(100)	0.778(100)	0.889(25)	**0.926(22)**

Table 3 Classification accuracy of RF, ET, OTE and LFRF ((number of trees in the initial set: 1500).

Data set	n	d	RF	ET	OTE	DERF
saheart	462	9	0.702(1500)	0.681(1500)	0.681(375)	**0.723(486)**
heart	270	13	0.741(1500)	0.667(1500)	**0.778(375)**	**0.778(375)**
hepatitis	80	20	0.875(1500)	0.75(1500)	**0.875(375)**	**0.875(375)**
mammographic	961	5	0.763(1500)	0.753(1500)	0.773(375)	**0.794(271)**
spectfheart	267	46	0.815(1500)	0.778(1500)	0.815(375)	**0.889(65)**

Table 4 Classification accuracy of RF, ET, OTE and LFRF((number of trees in the initial set: 2500).

Data set	n	d	RF	ET	OTE	DERF
saheart	462	9	0.702(2500)	0.681(2500)	0.725(625)	**0.745(25)**
heart	270	13	0.815(2500)	0.667(2500)	0.778(625)	**0.815(35)**
hepatitis	80	20	0.875(2500)	0.75(2500)	**0.875(625)**	**0.875(625)**
mammographic	961	5	0.763(2500)	0.753(2500)	0.773(625)	**0.794(425)**
spectfheart	267	46	0.852(2500)	0.815 (2500)	0.815(625)	**0.889(128)**

In addition to the experiments on the KEEL data set, we have also evaluated our DERF on the acute exacerbation of chronic obstructive pulmonary disease (AECOPD) and chronic respiratory disease (CRD) data set we collected, as shown in Figure 6 and Figure 7. Figure 6 illustrates the effect of the number of trees in the initial set on misclassification accuracy for the data sets (AECOPD and CRD) given using LFRF, OTE, ET and RF. It can be found that the DERF algorithm has the best prediction accuracy. DERF obtains the best prediction accuracy when the number of trees is 0-500. While DERF achieves the same predictive performance as the RF model when the number of decision trees exceeds 500. Figure 7 demonstrates the relationship between the size (number) of the initial decision trees and the size (number) of the final ensemble decision trees in the DERF on the AECOPD and CRD data sets. Figures 6 and Figures 7 show that DERF is better than OTE, the ET and RF models which suggests that DERF achieve better performance at the disease risk monitoring and emergency care peak visits prediction.

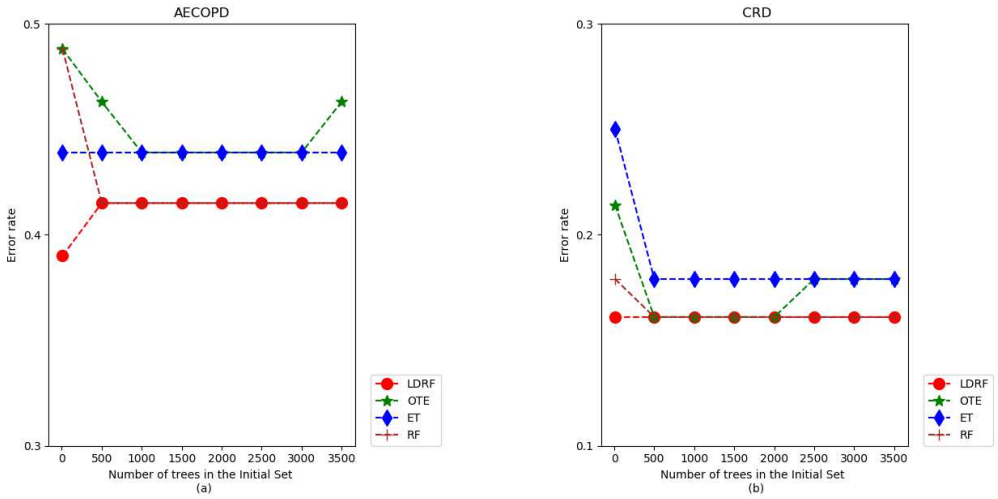

Fig. 6 The effect of the number of trees in the initial set on classification accuracy for the data sets given using LFRF, OTE, ET and RF.

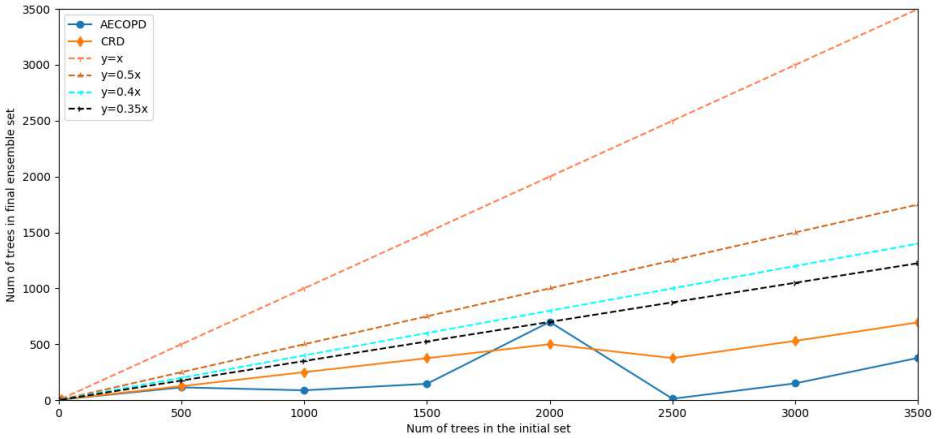

Fig. 7 Ensemble size reduction of DERF on the AECOPD and CRD data sets. Number of trees in initial set (on x-axis) selected for the generation of random forest. Number of trees in final ensemble set (on y-axis) indicated the final ensemble size of random forest.

Discussion

In this research, we propose an improved lightweight and diversity random forest model DERF with good performance and small ensemble size. Our model outperforms than the OTE, ET and RF models on the KEEL data set. Compare with the related studies by Khan (2020), we introduce the logarithmic loss function to measure the diversity of the random forest, then sort the trees in reverse order, and finally utilize the greedy stepwise backward search method to find the best ensemble tree size.

Khan (2020) recommended that a small portion, 1–10%, of the total trees that are individually strong which is further reduced in the second phase obtained the highest accuracy. While our results showed that the highest accuracy was obtained by using the small portion, 30–60%, of the total trees. This can be seen in Fig. 3 in the cases of Saheart, Heart，Hepatitis, Mammographic, Spectfheart data sets where the best results are obtained by using the small portion, 30–60%, of the total trees at the first phase. Different from the method Khan (2020) proposed, we emphasize the use of the greedy stepwise backward search method to find the highest accuracy. If the portion of the total trees left for the second phase optimization is too small, the scope of the search of DERF would become very narrow. This might be due to the reason that in such cases the possibility of searching good and diversity trees is high if the portion is medium-sized.

There are a few limitations to this study. In this study, we mainly focus on the disease classification problems, as these problems are critical clinical issues. However, the

regression problems about disease contribute to the critical clinical issues, such as bone disease. Regression problems of disease could be extended in the future work. The high accuracy and low integration scale of the model are realized by diversity search. This results in a longer time for the model training. Therefore, it is of great significance to speed up the generation of DERF model.

Conclusions

In order to improve the prediction performance and reduce the ensemble size of random forest model in view of the high prediction accuracy of medical scenarios, an improved diversity enhancement random forest algorithm has been proposed. Compared with the existing RF, ET and OTE model, the proposed DERF model obtains a higher predictive accuracy for the risk assessment of disease progression and a smaller ensemble size. This result shows that the proposed EDRF model may potentially support clinic physicians in diagnoses of the deterioration and death risk in patients.

Availability of data and materials

The datasets used during the current study are available from the corresponding author on reasonable request.

Abbreviations

RF: Random forest

ET: Extreme random tree

OTE: Ensemble of optimal tree

DERF: Diversity enhancement random forest

OOB: Out of Bag

MDS: Medical decision-making supporting

KEEL: Knowledge Extraction based on Evolutionary Learning

References

[1] Ho TK, Random decision Forest, Proceedings of the Third, International Conference on IEEE Computer Society, 1995.

[2] Breiman L, Raymon A, Random Forest Machine Learning, journal of clinical microbiology,2001,2:199-228.

[3] Fernandez-Delgado M,Cernadas E,Barro S,et al.Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?[J].Journal of Machine Learning Research,2014,15:3133-3181.

[4] Chowdhury A R,Chatterjee T,Banerjee S.A Random Forest classifier-based approach in the detection of abnormalities in the retina[J].Med Biol Eng Comput,2018,57(1):193–203.

[5] Geetha R,Sivasubramanian S,Kaliappan M,et al.Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier[J].J Med Syst,2019,43(9):286

[6] Aprilliani U,Rustam Z.Osteoarthritis Disease Prediction Based on Random

Forest[C].//International Conference on Advanced Computer Science and Information Systems (ICACSIS).Indonesia:Yogyakarta, 2018,237-240.

[7] VijiyaKumar K,Lavanya B,Nirmala I,et al.Random Forest Algorithm for the Prediction of Diabetes[C].// IEEE International Conference on System,Computation,Automation and Networking (ICSCAN).India:Pondicherry,2019,1-5.

[8] Raju N V G,Lakshmi K P,Praharshitha K G,et al.Prediction of chronic kidney disease (CKD) using Data Science[C].// International Conference on Intelligent Computing and Control Systems (ICCS),India:Madurai,2019,642-647.

[9] Cui H,Wang Y,Li G,et al.Exploration of Cervical Myelopathy Location From Somatosensory Evoked Potentials Using Random Forest Classification[J].IEEE Transactions on Neural Systems and Rehabilitation Engineering,2019,27(11):2254-2262.

[10] Guo C , Zhang J , Liu Y , et al. Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform[J]. IEEE Access, 2020, 8:59247-59256.

[11] Fawagreh K , Gaber M M . Resource-efficient fast prediction in healthcare data analytics: A pruned Random Forest regression approach[J]. Computing, 2020:1-12.

[12] Wang S , Wang Y , Wang D , et al. An improved random forest-based rule extraction method for breast cancer diagnosis[J]. Applied Soft Computing, 2020, 86:105941.

[13] Bernard S, Heutte L, Adam S (2009) On the selection of decision trees in random

Forest. In: International joint conference on neural networks, IEEE, pp 302–307

[14] Meinshausen N (2010) Node harvest. Ann Appl Stat 4(4):2049–2072

[15] Oshiro T, Perez P, Baranauskas J (2012) How many trees in a random forest? Machine Learning and Data Mining in Pattern Recognition, pp 154–168

[16] Latinne P, Debeir O, Decaestecker C (2001a) Limiting the number of trees in random Forest. In: Multiple Classifier Systems: Second International Workshop, MCS 2001 Cambridge, UK, July 2-4, 2001 Proceedings, Springer Science & Business Media, vol 2, p 178.

[17] Zhang H, Wang M (2009) Search for the smallest random forest. Stat Interface 2(3):381–388

[18] Coskun I , Colkesen Y , Altay H , et al. Ensemble pruning for glaucoma detection in an unbalanced data set[J]. Methods of Information in Medicine, 2016, 55(6):557-563.

[19] Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78(1):1–3

[20] Khan Z , Gul A , Perperoglou A , et al. Ensemble of optimal trees, random forest and random projection ensemble classification[J]. Advances in Data Analysis and Classification, 2020, 14(1):97-116.

[21] Breiman L.Bagging predictors [J].Machine Learning,1996,24(2):123-140.

Acknowledgments

We would like to acknowledge the Laisen Nie professor who participated in our focus group discussions.

Author information

Affiliations

Department of Computer Science, Guangdong University of Education, Guangzhou 510303, Guangdong, China

JunFeng Peng, Jun Xu, Xiongyong Zhu, Kaiqiang Zou, Guo-Ming Chen

The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong, China

Mi Zhou

The Second Affiliated Hospital, Guangzhou Medical University, Guangzhou, Guangdong, China

Jia-Yuan Chen

Hezhou University, School of Artifical Intelligence (Modern Industry College), Hezhou, Guangxi, China

Xing-Ji Chen,Xiao-Xin Li

School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, Guangdong, China

Zhan Zhang

Contributions

JFP conceived and designed the study, CGM acquired the funding, JFP, XJC, and XXL drafted the manuscript. ZZ, MZ, JYC, KQZ, XYZ and JX critically revised the manuscript. All authors read and approved the final manuscript.

Corresponding author

Correspondence to Junfeng Peng.

Ethics declarations

Ethics approval and consent to participate

Ethics approval has been granted by the Ethics Committee of the Third Affiliated Hospital, Sun Yat-sen University (TAHSYU) Institutional Review Board (IRB), protocol #[2019]-02-334-01. Signed informed consent will be sought from all participants. Confidentiality of information collected will be strictly maintained throughout the study. The trial database will be de-identified, password protected and only accessible by the research team. Only approved study investigators have access to the identifiable data on a need-to-know basis for study purposes (i.e. to conduct questionnaire interviews). The aggregated data and findings of this study will be reported at national and international academic conferences and be disseminated in peer-reviewed journals.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Natural Science Foundation of Guangdong Province [No.2018A0303130169].

Figures

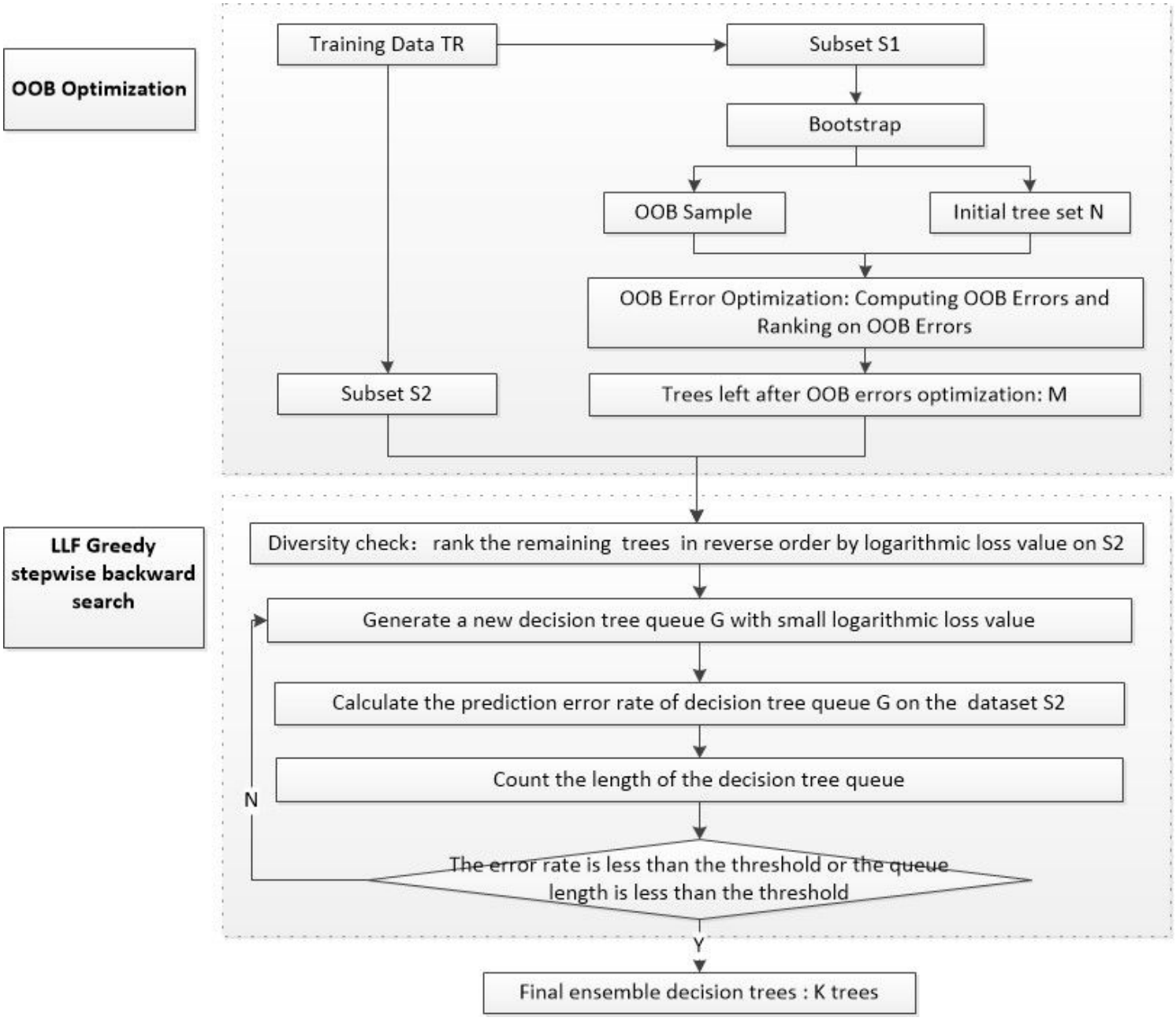

Figure 1

The frame of algorithm DERF.

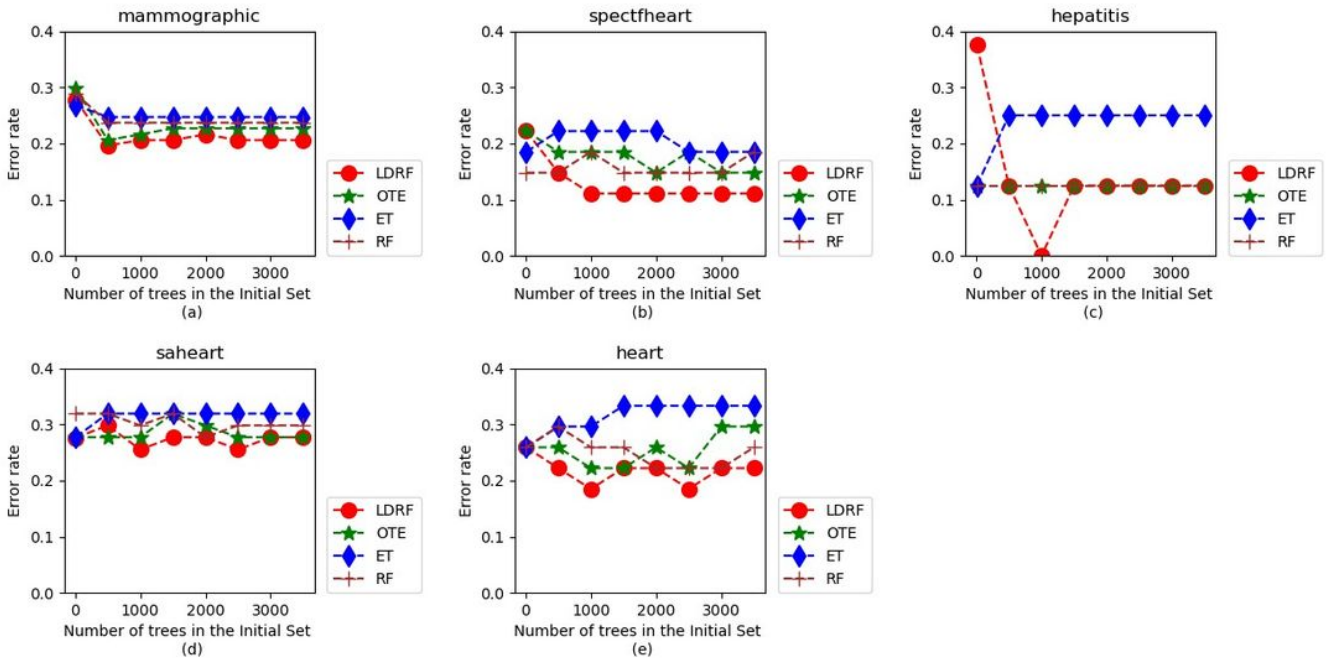

Figure 2

The effect of the number of trees in the initial set on classification accuracy for the data sets given using LFRF, OTE, ET and RF.

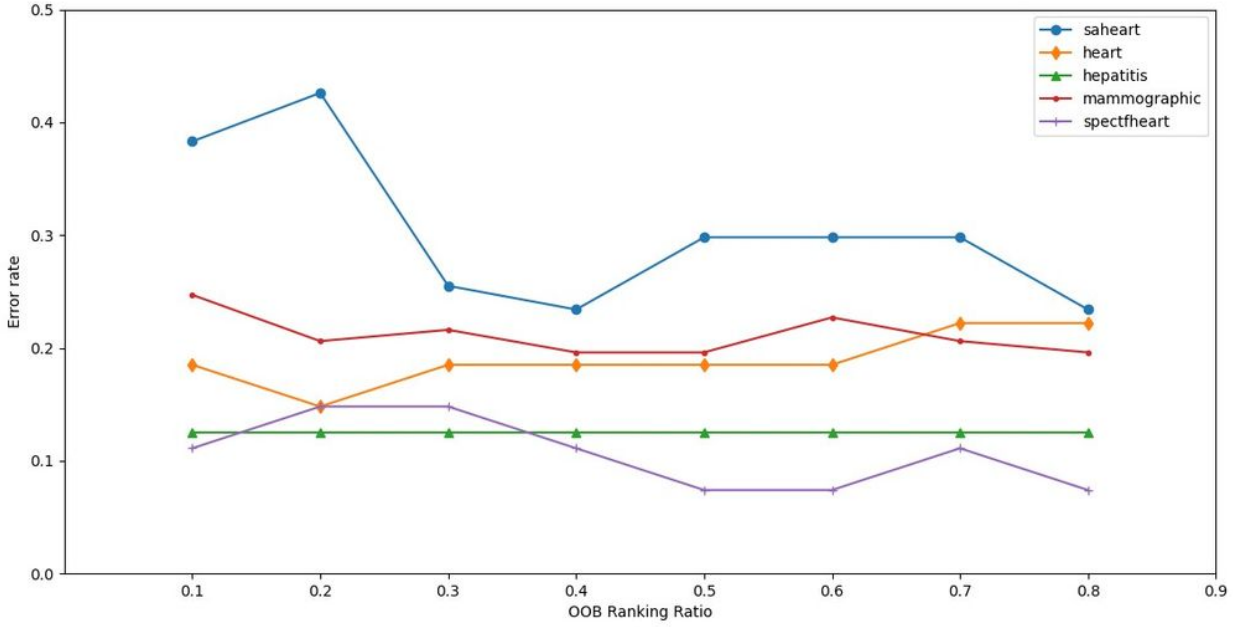

Figure 3

Effect of M on the OOB classification accuracy, of the data sets shown using DERF. The value of M in percentage is on the x-axis and classification accuracy on the y-axis.

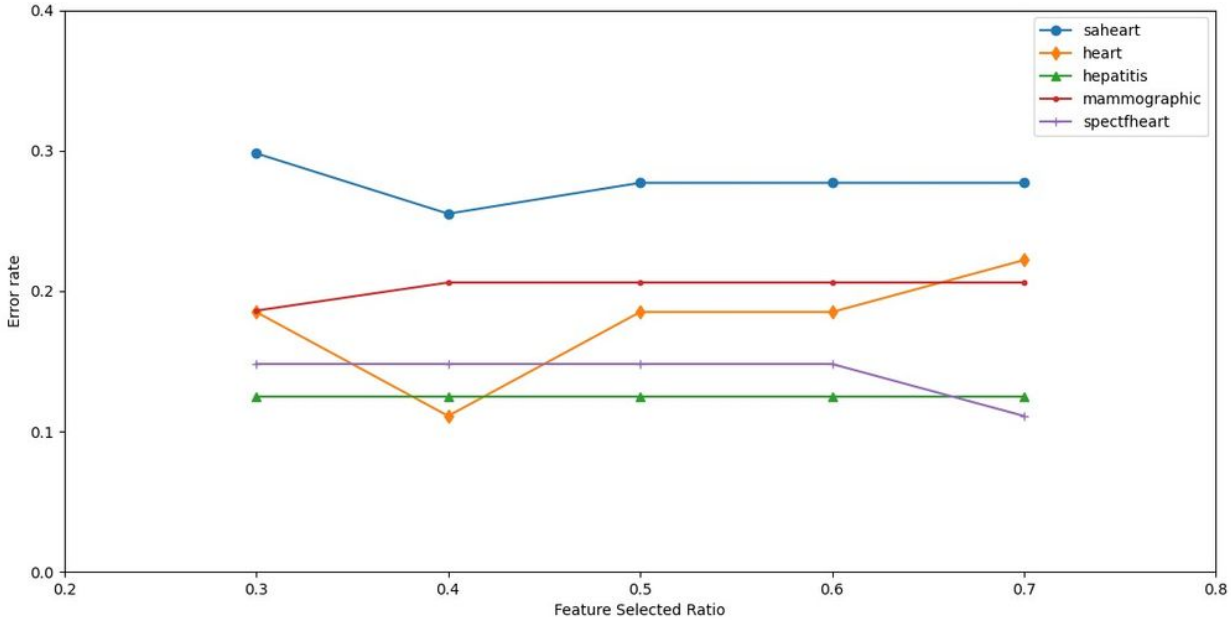

Figure 4

Effect of the number of features (on x-axis) selected at random for splitting the nodes of the trees on classification accuracy for the data sets shown using DERF.

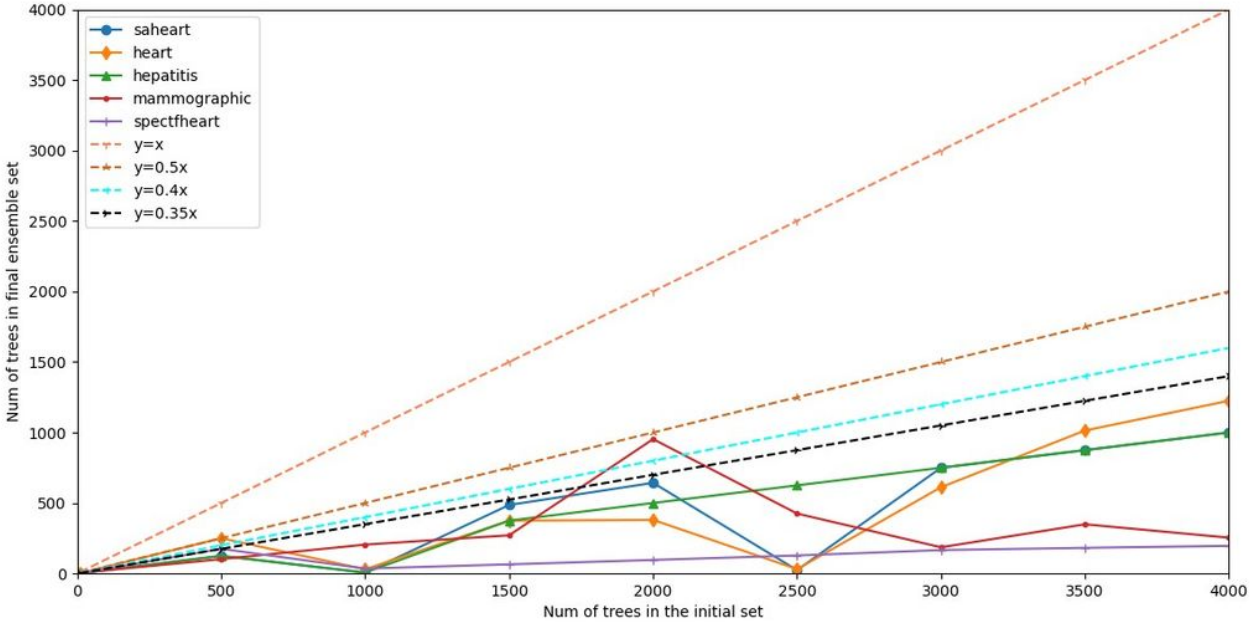

Figure 5

Ensemble size reduction of DERF. Number of trees in initial set (on x-axis) selected for the generation of random forest. Number of trees in final ensemble set (on y-axis) indicated the final ensemble size of

random forest.

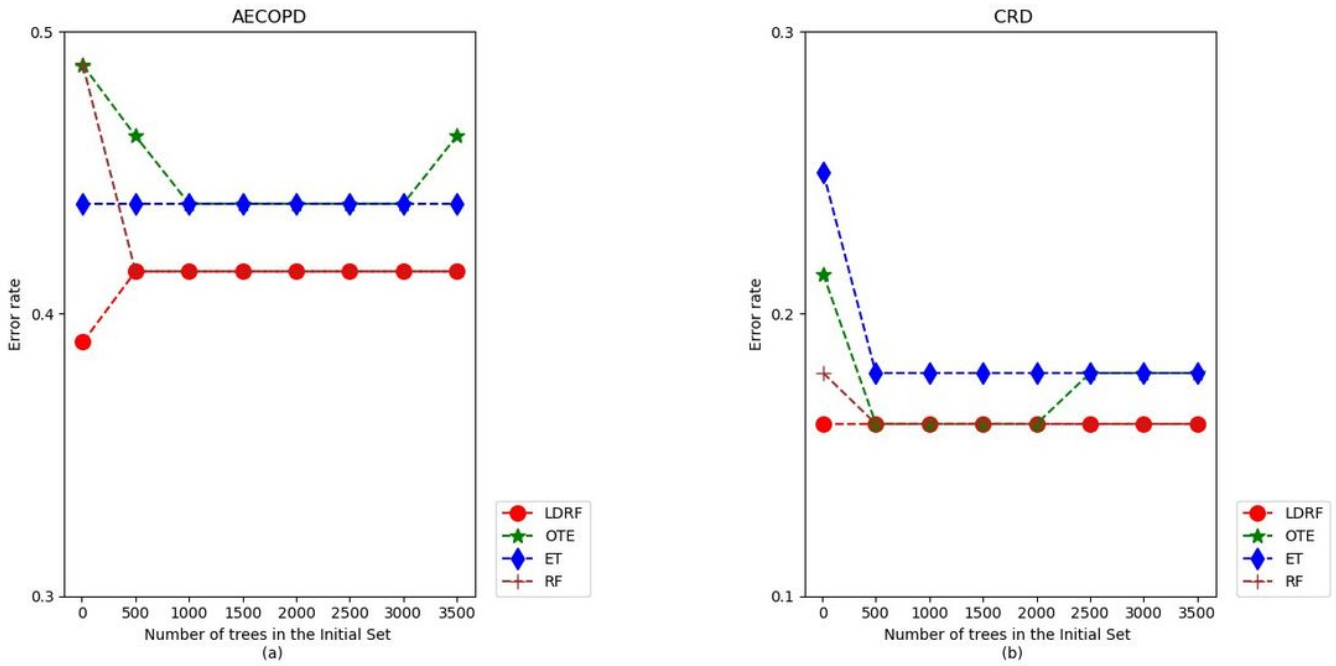

Figure 6

The effect of the number of trees in the initial set on classification accuracy for the data sets given using LFRF, OTE, ET and RF.

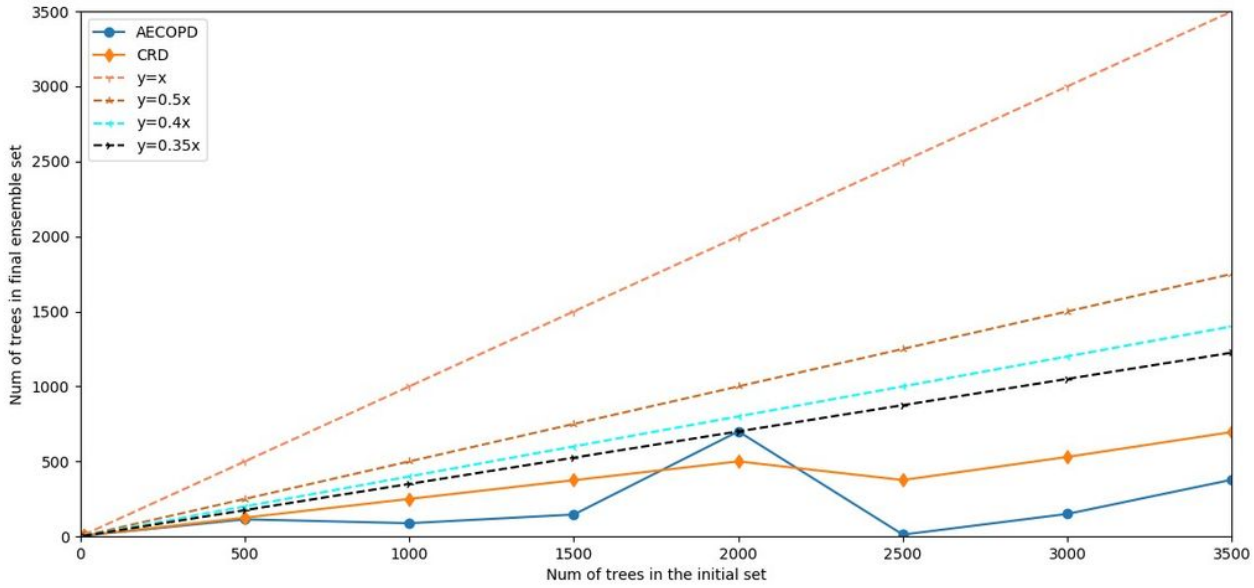

Figure 7

Ensemble size reduction of DERF on the AECOPD and CRD data sets. Number of trees in initial set (on x-axis) selected for the generation of random forest. Number of trees in final ensemble set (on y-axis)

indicated the final ensemble size of random forest.