

Optimizing random forest classifier with Jenesis-index on an imbalanced dataset

Joylin Zeffora¹, Shobarani²

¹Department of Computer Science, Dr. M. G. R. Educational and Research Institute, Chennai, India

²Department of Computer Science and Engineering, Dr. M. G. R. Educational and Research Institute, Chennai, India

Article Info

Article history:

Received Nov 12, 2021

Revised Jan 20, 2022

Accepted Feb 5, 2022

Keywords:

Gini coefficient

Gini index

Myocardial infarctions

Random forest

ABSTRACT

Random forest is an ensemble algorithm for machine learning. In decision trees, the splitting criteria is built on the prediction of the nodal points and formation of rules by Gini index and Information Gain. Gini index is a measure of inequality. Gini index does not take into consideration the structural changes in the dataset, and inaccurate data can distort the validity of the gini-coefficient. For data with the same feature but different outcomes, the gini-coefficient remained the same. The proposed method for attribute selection measure takes into consideration that there may be structural changes in the dataset overtime and it adapts to such expected changes and maintain the accuracy of the algorithm avoiding under-fitting and over-fitting. A dataset on myocardial infarctions was taken for the study and the results were promising.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Shobarani

Department of Computer Science and Engineering, Dr. M. G. R. Educational and Research Institute
Maduravoyal, Chennai-600095, Tamilnadu, India

Email: shobarani.cse@drmgrdu.ac.in

1. INTRODUCTION

Data analysis and machine learning have become essential components of modern scientific methodologies, enabling automated techniques for predicting a phenomenon based on prior observations, uncovering underlying patterns in data, and providing insights into the problem. Random forest is one of the widely used in ensemble algorithm for machine learning. The splitting criteria in random forest is obtained predominantly by Gini index (GI) or information gain (IG), as Gini index has an edge over Information Gain it is widely used. The newly proposed Jenesis index will overcome the lacuna in Gini index. The accuracy enhanced in Jenesis index over Gini index is studied on the dataset of myocardial infarctions in this paper.

The complexity in real life problems is to test and relate from different data mining techniques and recognize the pattern with multiple techniques. The data base discussed in this study is about the myocardial infarction commonly known as heart attack. Among the various symptoms, the predominant symptoms are chest pain, discomfort in shoulder numbness, palpitation. These symptoms can be identified by changes on an electrocardiogram (ECG), change in ST segment, pathological Q waves, the heart wall motion change or autopsy. Predicting the severity of this complication using this data base is the need of the hour to avoid fatality. The mortality rate is always proportional to the acuteness of myocardial infarction so, it is a quintessential problem to be addressed in today's world. In India around 54.5 million people are prone to cardio vascular disease. It is prevalent more in developed countries due to their poor diet and stress. The MI has a varying effect, patients with acute disorders are vulnerable to frequent illnesses or even fatality. Even experienced physicians cannot foresee the complications from the get-go, hence predicting the complications is necessary to prevent this disease.

2. PROPOSED METHOD – JENESIS INDEX

In this proposed method, the demerits of both GI and IG are emended with Jenesis Index. Let A_j^i denote the i^{th} record (out of total n sample records of the test set) of j^{th} sample attribute (out of total m attributes of the test set) which are converted into numerical values. Let S_R be the number of sample rows and S_C be the number of sample columns. Let n_j^i denotes the number of elements of a class (denoted by C) of the i^{th} row's j^{th} element in S_C . T_j^i denote the number of n_j^i 's in the outcome column showing YES or 1. The data contained in the table maybe ordinal/nominal or real. Therefore Algorithm 1 is applied to columns that have ordinal or nominal values and Algorithm 2 is applied to columns that have real values.

Algorithm 1. Algorithm for columns with ordinal/nominal values

Input: Train set

Output: Probable node with Jenesis index

1. Find all the unique values from the features
2. Calculate the ratios of 0's and 1's in the target column using
 1. $tn(0) = \text{total } 0's / \text{total target length}$
 2. $tn(1) = \text{total } 1's / \text{total target length}$
3. For each feature value
 - a. Find number of occurrences(n)
 - b. find the number of 0's and 1's corresponding target column (t).
 - c. Calculate the ratio of target occurrence (v) using the function $v(0) = t(0) / n$ and $v(1) = t(1) / n$.
 - d. Find the summation(u) of $v(0)$ and $v(1)$ for all unique features.
 - e. Calculate the ratio of $v(0)$ and $v(1)$ with $p(0) = v(0) / u$ and $p(1) = v(1) / u$
 - f. Calculate probability of 0 and 1 using
 - (i) $probability(0) = p(0) * tn(0)$
 - (ii) $probability(1) = p(1) * tn(1)$
 - g. Calculate the final probability of the feature value with $probability(0) + probability(1)$.
4. The feature value with highest probability will be the split value for the feature and corresponding probability will be split probability for the feature.
5. The feature with highest probability will be the split feature and corresponding probability will be split probability for the data set.

Algorithm 2. Algorithm for columns with real values

Input: Train set

Output: Probable node with Jenesis index

1. Pick all the unique values from the features
2. Calculate the ratios of 0's and 1's in the target column using
 - a. $tn(0) = \text{total } 0's / \text{total target length}$
 - b. $tn(1) = \text{total } 1's / \text{total target length}$
3. For each feature value
 - a. Find number of occurrences(n)
 - b. find the number of 0's and 1's corresponding target column $t(0)$ and $t(1)$.
 - c. Calculate the final probability of the feature value with $((t(0) / n) * tn(0) + (t(1) / n) * tn(1)) * 100$.
4. The feature value with highest probability will be the split value for the feature and corresponding probability will be split probability for the feature.
5. The feature with highest probability will be the split feature and corresponding probability will be split probability for the data set.

The following architectural diagram as shown in Figure 1 is a diagrammatic representation of Jenesis algorithm. The dataset was split into five folds with four folds for training and validation set and one fold for test set. The mean accuracy was calculated based on the scores computed for each fold. A confusion matrix is used to identify the number of true positives, true negatives, false positives and false negatives which is shown in Table 1.

The confusion matrix Table 2 gave more insight into the accuracy of the predicted results. The results obtained from ORF-Jenesis were compared with the results obtained with RF-Gini. The confusion matrix of the RF-Gini and ORF-Jenesis as shown in Table 3 are observed in the analysis of myocardial infarctions. The aim was to focus on predicting true positives and true negatives and minimizing false negatives and false positives.

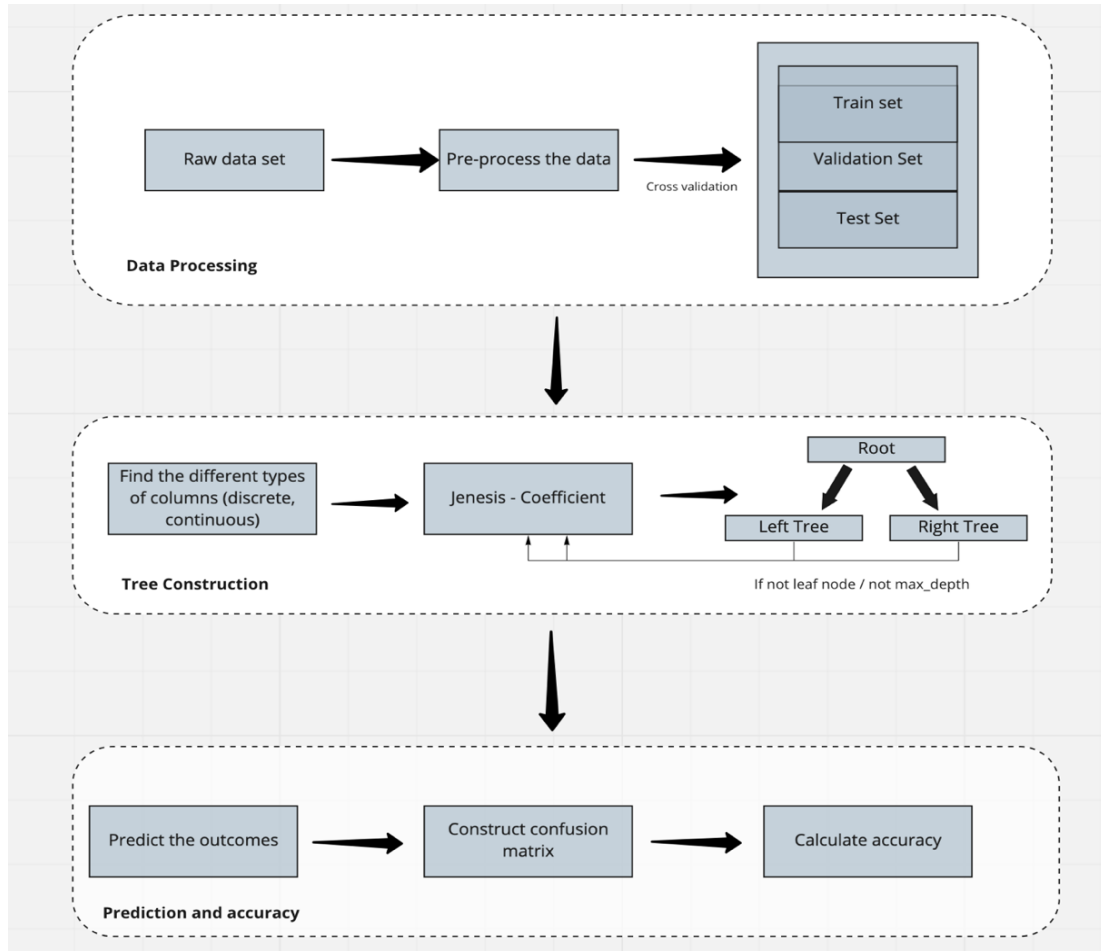


Figure 1. Architecture diagram

Table 1. Confusion matrix

Predicted class	Actual class	
	Positive(1)	Negative(0)
Positive(1)	<i>True Positive</i>	<i>False Positive</i>
Negative(0)	<i>False Negative</i>	<i>True Negatives</i>

Table 2. Confusion matrix for myocardial infarctions

Function	Accuracy	Actual_Total_Positives	Actual_Total_Negatives	True Positives	False Negatives	True Negatives	False Positives
RF-Gini	80.58	65	275	2	63	272	3
ORF-Jenesis	81.47			4	61	273	2

Table 3. Time complexity and space complexity

	No of trees	Time Taken to complete execution	Space Complexity
RF-Gini	1	76	5.3 Mb
	5	383	
	10	730	
	20	1727	
ORF-Jenesis	1	11	8 Mb
	5	45	
	10	101	
	20	185	

3. METHOD

In random forests Breiman, [1] algorithms, decision trees play a pivotal role in deciding the node to split the tree and, in turn, creating decision trees from the test set to predict the percentage of trueness. While these approaches have shown to be a reliable, accurate, and useful tool for a wide range of machine learning problems, such as classification, regression, density estimation, manifold learning, and semi-supervised learning, we still have a lot to learn about them. G Louppe studied the induction of decision trees and the construction of ensembles of randomized trees showing their good computational performance and scalability, along with an in-depth discussion of their implementation details, as contributed within Scikit-Learn. Also, he analysed the interpretability of random forests in the eyes of variable importance measures. The core of our contribution's rests in the theoretical characterization of the mean decrease of impurity variable importance measure and derived some of its properties in the case of multiway totally randomized trees and in asymptotic conditions [2].

Saffari *et al.* [3] combined the ideas from on-line bagging and extremely randomized forests and propose an on-line decision tree growing procedure and also on the temporal weighting scheme for adaptively discarding some trees based on their out-of-bag-error in given time intervals and consequently growing of new trees. Kalidas and Tamil [4] proposed method of AF detection, combining Markov models and random forests, achieves high accuracy across multiple databases and demonstrates comparable or superior performance to several other state-of-the-art algorithms. Kaur *et al.* [5] discuss the usage of random forest classifier to detect atrial fibrillation over a 10-fold cross validation. Gradient boosting is a technique has been used to predict the likelihood of acute myocardial infarction a study by Than *et al.* [6]. Yadav and Pal [7] combined pearson correlation and lasso regularization with random forest to achieve 99% accuracy with the heart disease dataset. Belhadj *et al.* [8] proposed a fuzzy version of gini index to improve the performance of gini index. A diverse range of research studies has been conducted around coronary illness. One such research work is the prediction of heart disease using a neural classifier to predict heart disease by Mathan *et al.* [9]. In order to improve the accuracy of the classifier several combinations of techniques such as fuzzy logic and weighting of decision trees [10].

Jain *et al.* [11] Investigated the joint splitting criteria using two of the most used criterions i.e., Information Gain and Gini index and proposed the data split points when Information Gain is maximum and Gini index is minimum. Kulkarni *et al.* [12] proposed a method to generate the individual decision tree in the random forest using randomly selecting one out of three split measures IG, GI and Gain ratio. Raileanu and Stoffel [13] has done the theoretical comparison of the most popular split criteria namely the GI and IG and have theoretically compared these two criteria.

The general approach in predicting random forest are: Decision trees → Data Set → Training data set → Formation of rules → Test set → Classification → Result.

Biau and Scornet emphasised on replacing mathematical forces for driving the algorithm, with special attention given to the selection of parameters, the resampling mechanism, and variable importance measures [14]. Information gain is another technique that is predominantly used in classifiers, but they are seldom used because it is complicated and the results are biased in unbalanced trees, Antonin Leroux suggests a method to improve the prediction accuracy using IG. [15].

The application of different patterns of heart disease by data mining techniques was studied by Kirmani and Ansarullah [16]. Lempitsky *et al.* [17] solved problem with random forests, which are discriminative classifiers developed lately in the machine learning field, allows for accurate delineations of the full 3D volume in a matter of seconds (on a CPU) or even in real-time (on a GPU). Yosefian *et al.* [18] determined the applicability of saturated tree (ST), pruned tree (PT), and RSF. Methods Khened *et al.* [19] proposed a fully automatic method for segmentation of left ventricle, right ventricle and myocardium from cardiac magnetic resonance (MR) images using densely connected fully convolutional neural network. dense convolutional neural network (DenseNet) facilitates multi-path flow for gradients between layers during training by back-propagation and feature propagation using random forest. The development of methods for precise quantification is critical for improving myocardial infarction patient diagnosis and therapy was studied by Allen *et al.* [20]. Mansoor *et al.* [21] found the using logistic regression and random forest, design and evaluate prediction models for all-cause in-hospital mortality in women hospitalised with STEMI, and compare the performance and validity of the different models. The efficacy of contemporary machine learning algorithms in individualised risk prediction for patients undergoing elective heart valve surgery was examined. Correct anticipation of this risk allows for the improved counselling of patients and avoidance of possible complications. We therefore investigated the benefit of modern machine learning methods in personalized risk prediction for patients undergoing elective heart valve surgery Bodenhofer *et al.* [22]. Asadi *et al.* [23] proposed method's effectiveness is investigated by comparing its performance over six heart datasets with individual and ensemble classifiers. The results suggest that the proposed method with the (near) optimal number of classifiers outperforms the random forest algorithm with different classifiers. The

dataset on myocardial infarctions is obtained from the UCI repositories which as 1700 rows and 124 columns containing test and lab reports of patients' data [24]. Zahibi *et al.* [25] proposed a method where the characteristics from the ECG signals' time, frequency, time-frequency domains, and phase space reconstruction are used. A random forest classifier is employed in the final stage to categorize the selected characteristics into one of the four aforementioned ECG classifications.

3.1. Random forest

Random forest algorithm is a predominantly used supervised machine learning algorithm that gives accurate predictions. The core idea behind random forest implementation is dividing the training set into sub-samples of data and constructing multiple trees. Gini index or information gain are the techniques that are used as attribute selection measures which in turn determines the splitting of the nodes while constructing the tree. The leaf nodes are then analysed using bootstrap aggregation or commonly called bagging to predict the outcome of a specific tuple. The primary concern of any algorithm is primarily based on two concepts:

- The versatility in accommodating data with various factors will lead to trueness in prediction.
- Enhancing the accuracy of the result of an algorithm.

3.1.1. Random forest: existing attribute selection measure:

The nodal points of the decision trees and the formation of rules are done through primarily by GINI index or information gain. The main features of Gini index and information gain are:

Gini index:

- (a) Gini index = $1 - \sum_i^n p_i^2$.
- (b) If Gini index is zero then the attributes are spread across equally. If Gini index= 1 then is pertaining to only one attribute.
- (c) In Decision tree if the value is less than 0.5, we accept and proceed further for next classification.

Merits of Gini index:

- (i) It is used for large partitions.
- (ii) It is useful in inequality measures.
- (iii) It is easy to implement.

Demerits of Gini index:

- (i) Not compatible for more distinct values.
- (ii) The measure will give different results when applied to different sets.

Example:

Information Gain:

Information gain is used for smaller partitions and more distinct values and comparatively hard to implement than Gini index.

- (i) Entropy = $-\sum_i^n p_i * \log_2 p_i$
- (ii) Information Gain (Target, Predictor) = Entropy (Target)- Entropy (Predictor).
- (iii) Choose the largest information gain as split to proceed to the next step.

Merits of Information Gain:

- (i) It is used for more distinct values.
- (ii) It is good measure for deciding the relevance of the attributes.

Demerits of Information Gain:

- (i) Not compatible for large partitions.
- (ii) It is hard to implement.

4. RESULTS AND DISCUSSION

The results obtained from the model showed that RF-Jenesis index performed better in comparison with RF-Gini. The dataset was first trained with 1000 rows and 100 columns. Most healthcare datasets contain more negatives than positives and the dataset in question is no different. It has 275 instances of negative instances and 65 occurrences of positive instances. The accuracy achieved by ORF-Jenesis is calculated as 81.47% and the accuracy of RF-Gini is 80.58%. The target column in the dataset contains more instances of 0s therefore the number of prediction of negatives is higher than the number of prediction of positives. From the total number of negative instances (275) present in the target column of the actual dataset 273 instances were correctly predicted as true negative and 2 were incorrectly predicted as false positive by ORF-Jenesis, whereas RF-Gini could correctly predict 272 instances as true negative. From the total number of positive instances (65) present in the target column of the actual dataset 4 instances were correctly predicted as positive and 61 instances were incorrectly predicted as false negatives by ORF-Jenesis and 63 were incorrectly predicted as false negatives by RF-Gini. From the confusion matrix f-measure, sensitivity

and specificity can be calculated. The accuracy of an algorithm depends on not just the prediction of the positives but on the prediction of negatives as well. ORF- Jenesis predicts better than RF-Gini on the positive scale as well as the negative scale.

- a) Time and space complexity: Performance of an algorithm is evaluated by means of time and space complexity. Time complexity and space complexity are shown in Table 3 after analysis of the dataset on myocardial infarctions
- b) Limitations of the proposed model: The result of the prediction purely depends on the balance of the dataset. The f-measure cannot be considered as a measure of accuracy in a dataset where the ratio of positives to negatives is low. The dataset used contains more instances of 0s than 1s therefore f-measure and sensitivity is very minimal.

5. CONCLUSION

The classification problems the target attribute contains 0 s and 1 s. In classification problems that involve medical data the target attribute classifies where a patient has a disease or not. In a general sense it is not sufficient to specify whether a patient has a disease or not, in cases of cancer it is imperative to postulate the degree of infection. Therefore, it is inadequate to classify the target as just 0 and 1. The values in the attributes could be modified to predict a real value which would signify the degree of illness.




REFERENCES

- [1] L. Breiman, "Random forest" *Springer Machine Learning*, vol. 45, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.
- [2] G Louppe, *Understanding random forests: From theory to practice*, arXiv preprint arXiv:1407.7502, 2014.
- [3] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof, "On-line random forests," *IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, 2009*, pp. 1393-1400, doi: 10.1109/ICCVW.2009.5457447.
- [4] V. Kalidas, and L. S. Tamil, "Detection of atrial fibrillation using discrete-state Markov models and random forests," *Journal of Biomedical Informatics*, vol. 115, p. 103386, 2021.
- [5] M. Kaur, "An approach for sentiment analysis using gini index with random forest classification," *International Conference on Computational Vision and Bio Inspired Computing*. Springer, Cham, 2019, pp. 541-554, doi: 10.1007/978-3-030-37218-7_62.
- [6] M. P. Than, *et al.*, "Machine learning to predict the likelihood of acute myocardial infarction," *Circulation*, vol. 140, no. 11, pp. 899-909, 2019, doi: 10.1161/CIRCULATIONAHA.119.041980.
- [7] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," *International Journal of Pharmaceutical Research*, vol. 12, no. 4, 2020, doi: 10.31838/ijpr/2020.12.04.013.
- [8] B. Belhadj, F. Kaabi, and M. Bouanani, "Fuzzy version of Gini's index," *Social Indicators Research*, pp.1-9, 2021.
- [9] K. Mathan, P. M. Kumar, P. Panchatcharam, G. Manogaran, and R. Varadharajan, "A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease," *Design automation for embedded systems*, vol. 22, no. 3, pp. 225-242, 2018, doi: 10.1007/s10617-018-9205-4.
- [10] P. Bonissone, J. M. Cadenas, M. C. Garrido, and R. A. Diaz-Valladares, "A fuzzy random forest: Fundamental for design and construction," *Proceedings of the 12th International Conference on Information Processing an Management of Uncertainty in Knowledge-Based Systems (IPMU'08)*, 2008.
- [11] V. Jain, A. Phophalia, and J. S. Bhatt, "Investigation of a joint splitting criteria for decision tree classifier use of information gain and Gini index," *Proceedings of TENCON 2018 - 2018 IEEE Region 10 Conference*, Oct. 2018, doi: 10.1109/TENCON.2018.8650485.
- [12] V. Y. Kulkarni, P. K. Sinha, and M. C. Petare, "Weighted hybrid decision tree model for random forest classifier," *Journal of The Institution of Engineers (India): Series B*, vol. 97, no. 2, pp. 209-217, 2016, doi: 10.1007/s40031-014-0176-y.
- [13] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Annals of Mathematics and Artificial Intelligence*, vol. 41, no. 1, pp. 77-93, 2004, doi: 10.1023/B:AMAI.0000018580.96245.c6.
- [14] G. Biau, and E. Scomet "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197-227, 2016, doi: 10.1007/s11749-016-0481-7.
- [15] A. Leroux, M. Boussard, and R. D s, "Information gain ratio correction: Improving prediction with more balanced decision tree splits," arXiv preprint arXiv:1801.08310, 2018.
- [16] M. M. Kirmani and S. I. Ansarullah, "Prediction of heart disease using decision tree a data mining technique," *International Journal of Computer Science and Network*, vol. 5, no. 6, pp. 885-892, 2016.
- [17] V. Lempitsky, M. Verhoeck, J. A. Noble, and A. Blake, "Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography," *Springer Berlin Heidelberg*, pp. 447-456, 2009, doi: 10.1007/978-3-642-01932-6_48.
- [18] I. Y. Ehsan, M. Farkhani, and M. R. Baneshi, "Application of random forest survival models to increase generalizability of decision trees: a case study in acute myocardial infarction," *Computational and Mathematical Methods in Medicine*, Article ID 576413, 2015, doi: 10.1155/2015/576413.
- [19] M. Khened, V. Alex, and G. Krishnamurthi, "Densely connected fully convolutional network for short-axis cardiac cine MR image segmentation and heart diagnosis using random forest," *Springer International Publishing*, pp. 140-151, 2018, doi: 10.1007/978-3-319-75541-0_15.
- [20] J. Allen, E. Zacur, E. Dall'Armellina, P. Lamata, and V. Grau, "Myocardial infarction detection from left ventricular shapes using a random forest," *Springer International Publishing*, pp. 18-189, 2016, doi: 10.1007/978-3-319-28712-6_20.
- [21] H. Mansoor, I. Y. Elgendy, R. Segal, A. A. Bavry, and J. Bian, "Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: A machine learning approach," *Heart & Lung*, vol. 46, no. 6, pp. 405-411, 2017, doi: 10.1016/j.hrtlng.2017.09.003.




- [22] U. Bodenhofer, B. Haslinger-Eisterer, A. Minichmayer, G. Hermanutz, and J. Meier, "Machine learning-based risk profile classification of patients undergoing elective heart valve surgery," *European Journal of Cardio-Thoracic Surgery*, 2021, doi: 10.1093/ejcts/ezab219.
- [23] S. Asadi, S. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," *Journal of Biomedical Informatics*, vol. 115, 2021, doi: 10.1016/j.jbi.2021.103690.
- [24] S. E. Golovenkin *et al.*, "Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data," *GigaScience*, vol. 9, no. 11, 2020, doi: 10.1093/gigascience/giaa128.
- [25] M. Zabihi, A. B. Rad, A. K. Katsaggelos, S. Kiranyaz, S. Narkilahti, and M. Gabbouj, "Detection of atrial fibrillation in ECG hand-held devices using a random forest classifier," *Computing in Cardiology (CinC)*. IEEE, 2017, doi: 10.22489/CinC.2017.069-336.

BIOGRAPHIES OF AUTHORS



Joylin Zeffora    is currently working as an Assistant Professor, Department of Computer Science, Loyola College, Chennai-600034, Tamil Nadu, India. She has completed her B.Sc., Computer Science from Women's Christian College, Chennai, India, her M.Sc., Computer Science from Loyola College, Chennai, India, has cleared the National Eligibility Test, India and is pursuing her doctoral degree from M.G.R University, Chennai, India. Her research interests include machine learning, artificial intelligence and IoT. She can be contacted at email: joylinzeffora@gmail.com.



Dr. Shobarani    received her Postgraduate degree in Computer Applications from Bharathidasan University, Tamilnadu, India in 2002 and M.Phil from Annamalai University, Tamilnadu, India in 2004 and the Ph.D. degree from Mother Teresa women's University, Tamilnadu, India in Feb'2014. She is the author/coauthor of over 28 referred research papers. She is currently working as a Professor and Research guide in the Department of Computer Science and Engineering, at Dr. M.G.R. Educational and Research Institute, Chennai, Tamilnadu, India. Her publication and research areas include Image Processing, Machine Learning and Datamining. She can be contacted at email: shobarani.cse@drmgrdu.ac.in.