Feature based analysis of endometriosis using machine learning

Visalaxi Sankaravadivel[1], Sudalaimuthu Thalavaipillai[1], Surya Rajeswar[2], Pon Ramalingam[2]
[1]School of Computing Science, Hindustan Institute of Technology and Science, Chennai, India
[2]School of Management, Hindustan Institute of Technology and Science, Chennai, India

Article Info

Article history:

Received Aug 19, 2022
Revised Oct 29, 2022
Accepted Nov 2, 2022

Keywords:

Decision tree
Endometriosis
Entropy
Gini index
Laparoscopic surgery
Mass-like

ABSTRACT

Machine learning is a cutting-edge technology used for predicting and diagnosing various diseases. Various machine learning algorithm facilitates the prediction. The decision tree belongs to learning algorithm that performs both classification and prediction. The decision tree constructs the tree-like to evaluate the best features. The decision tree performs well in the prediction of various diseases. Endometriosis is a recurrence disease that creates an emotional impact in women. Endometriosis is a lump-like structure that appears at several locations in reproductive organs of women. The diagnosis of endometriosis was predicted through scanning procedures and laparoscopic procedures. The symptoms identified from laparoscopic surgery were used as the features for predicting the severity of endometriosis. The symptoms include mass-like structure, tissue size, variation in tissue colour, and blockages in fallopian tubes. The decision tree analyze the features of endometriosis by using two criteria such as entropy and Gini index. The entropy and Gini index construct the tree by identifying the size of tissue as major influencing attributes. The Gini index outperforms well with training accuracy of 84.08% and test accuracy of 84.85.

Corresponding Author:

Visalaxi Sankaravadivel
School of Computing Science, Hindustan Institute of Technology and Science
OMR, Padur, Chennai, Tamilnadu
Email: sakthi6visa@gmail.com

1. INTRODUCTION

Endometriosis is a challenging problem for female of fertile group. The endometriosis phases vary from person to person based on the location and severity of occurrences. The endometrium layer inside the uterus shed out for every menstruation. If the layer spread across multiple locations leads to endometriosis. The phases are classified as: i) endometriosis inside the uterus, ii) ovarian endometriosis, iii) peritoneum endometriosis, and iv) deep infiltrating endometriosis. Endometriosis was identified through scanning procedures. The most accurate position of endometriosis was diagnosed through the standard laparoscopic operating procedure.

The severity of endometriosis affects women both physically and mentally. The external factors identified for predicting endometriosis were severe abdominal pain, dysmenorrhea, dyspareunia, abnormal uterine bleeding, and breast tenderness. These external factors play a vital role in the prediction that leads to the scanning and laparoscopic procedure. The internal factors identified through the laparoscopic procedure emphasize the severity of endometriosis. The internal factors include adnexal mass, tissue-like structure, and changes in tissue color [1].

Machine learning algorithms played a predominant role in diagnosing various types of diseases. The types of learning algorithms includes random forest, decision tree, logistic regression, support vector machine, logistic, and linear regression [2]. These algorithms predicts various types of disease. The decision tree is a learning algorithm that analyze various features of the problem by constructing a tree. The tree consists of a root node at the top followed by intermediate leaf nodes and final layer consist of decision nodes. The decision tree identifies the features that was most suitable for analysis [3].

In the decision tree, a concept known as information gain was used for splitting the nodes. There are two approaches used for implementing information gain [4]. They are: i) entropy and ii) Gini index. Information gain identifies the best features that are suitable for classification.

2. RELATED STUDIES

The decision tree was used to forecast the possibilities of cough through the features of fractional exhaled nitric oxide, expiratory flow, and eosinophils. Among all attributes, eosinophils have a major impact on cough [5]. Breast cancer was predicted using logistic regression and decision tree algorithm. Various features including hormonal therapy, tumor size, histological grade, and diagnosis. Were considered for evaluation. The decision tree outperforms well in forceasting the survival rate of cancer when compared with logistic regression and decision tree (C5.0) yields higher accuracy of 86.9% in predicting the breast cancer [6]. Birth defects was analyzed using decision tree algorithms. Various factors including family hereditary, hypertension, diabetes, and nephropathy are considered. Decision tree (C5.0) and C4.5 was used for evaluation. C4.5 algorithm outperforms well in 9.33 seconds and 94.15% of accuracy [7].

The decision tree was also used for analyzing the various strategies for endometriosis. Three levels of the evaluation were performed using various factors for predicting endometriosis. “Deep dyspareunia, cyclic defecation pain, cyclic urinary signs” were considered for evaluation. The decision tree uses second-line and third-line evaluation for endometriosis diagnosis as deep lesions [8]. The factors influencing heart disease were predicted using the decision tree. Among multiple attributes, Thalassemia, type of Chest pain, and major vessels color were identified as the best features and the model yields an accuracy of 85% [9]. The decision tree helps in evaluating the features by invoking entropy and the Gini index. The decision tree predicts the various types of diseases including cough, thyroid, cancer, heart problems, birth defects, and endometriosis. C5.0 algorithm works well in predicting the breast cancer, C4.5 outperforms well for predicting birth defects. Internal factors play a vital role in predicting the category of endometriosis. The internal factors were identified as the attributes in endometriosis prediction. It includes the size of the tissues, tissue color, mass identified, and blockages in fallopian tubes. These factors were identified by the retrospective study provided by gynecologist and radiologist [10]. The size of the lesion varies from 1 mm to 6mm, the color of the tissue exist in red, dark brown, and black colour. In several cases, adnexal mass was identified along with blockages in fallopian tubes. A total of 600 records were considered for execution. The list of features are as follows: Size of tissue [1], Color of tissue [2], Blockages in fallopian tubes [3], and Adnexal mass [4].

3. FEATURE ANALYSIS OF ENDOMETRIOSIS USING DECISION TREE

The steps involved in decision tree analysis of endometriosis classification as illustrated in the Figure 1.

a) Dataset holding endometriosis influencing internal factors.
b) Splitting of training and testing datasets.
c) Ordinal encoding of training and testing datasets.
d) Decision tree evaluation using Gini and entropy
e) Performance evaluation.

3.1. Data pre-processing

The dataset holds around 600 records of 5 attributes. The independent and dependent attributes were identified. The dataset was divided as training set holds 402 records and testing data holds 198 records containing both independent and dependent attributes. The split data was transformed using encoders. Two encoders were frequently used. They are: i) one-hot encoding and ii) ordinal encoding. The ordinal encoding technique [11] was adopted for the given dataset as it contains categorical values. The ordinal encoding was applied to the independent attributes of both the training and testing datasets.

3.2. Decision tree analysis

The decision tree is a learning technique that performs both classification and predictions. Here the tree is organized as leaf nodes, root nodes, and decision nodes. Root nodes appear at the top, leaf nodes appear at the middle and decision nodes appear at the bottom layer. The decision tree [12] identifies the features that are helpful for classification and prediction. The two major terms help in analyzing the features. They are: i) Gini impurity and ii) entropy.

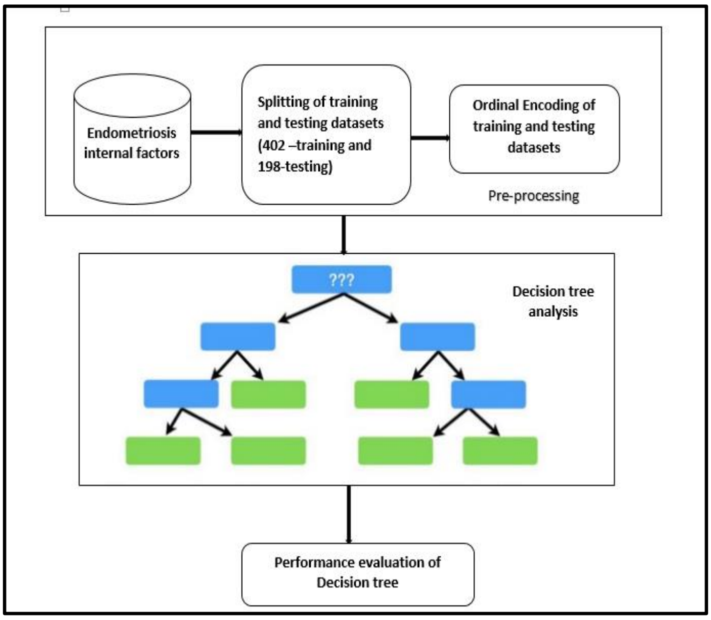

Endometriosis internal factors
Splitting of training and testing datasets (402 –training and 198-testing)
Ordinal Encoding of training and testing datasets
Pre-processing
???
Decision tree analysis
Performance evaluation of Decision tree

Figure 1. Steps involved in analyzing the features using decision tree

Entropy [13] is a term used in the decision tree for splitting trees into smaller subsets. Entropy was used to identify the best feature in the given datasets. The formula for calculating entropy was as follows:

$$Entropy(E) = -\sum_{i=1}^{k} p(i).log_2.\ \mathrm{p(i))} \quad (1)$$

here p(i) represents the probability of value and probability always falls in the range of 0 to 1.

Gini index [14] is a term used for constructing a decision tree. The Gini index determines how well the decision tree was constructed. Similar to entropy, Gini impurity identifies the feature that suits well for classification. The Gini value lies between 0 to 0.5. The formula for calculating the Gini index was as follows:

$$Gini = 1 - \sum_{i=1}^{n} (pi)^2 \quad (2)$$

where p(i) is the probability that falls between 0 to 0.5.

```
Pseudocode:
Start
Read data=: Internal factors of endometriosis
X, Y: = Independent and dependent attributes
Split Training Data (training_X, training_Y)
      Testing Data (testing_X, tesingt_Y)
Perform Encoding: train_X = encoder.fit (training_X)
                  test_X = encoder.fit (testing_X)
Model1:= Gini. Fit (training_X, training_Y)
Model2: = Entrophy.Fit (training_X, training_Y)
Construct Confusion matrix (testing_Y, pred_Gini_Y)
Construct Confusion matrix (testing_Y, pred_Entrophy_Y)
Visualize receiver operating characteristic (ROC) and area under the curve (AUC) (Gini and Entrophy)
```

3.2.1. Evaluation metrics

The dataset was implemented using the decision tree model by selecting the appropriate features using entropy and Gini index. The implemented data was evaluated using several metrics including specificity, sensitivity, precision, accuracy, and F1 score through a classification matrix.

- Precision [15] is the proportion of true and untrue positive values made by the model.

$$precsion = \frac{True\ values}{Overall\ True\ values} \quad (3)$$

- Recall (Sensitivity) [16] is the proportion of identified real positive values to the whole positive values.

$$recall = \frac{True\ positive}{Overall\ positive\ values} \tag{4}$$

- Specificity [17] is defined as the model can predict the accurate negative values for the classification performed.

$$specificity = \frac{Accurate\ negative}{Accurate\ negative + Inaccurate\ postive} \tag{5}$$

- F1 score [18] is the weighted mean of precision and sensitivity.

$$F1 = \frac{2*precision*sensitivity}{precision+sensitivity} \tag{6}$$

- AUC-ROC Curve

The area under the curve – Receiver operating characteristics [19] is a graph to identify the capability of a model to differentiate between two classes. AUC was plotted across true predicate values and false predicate values.

4. RESULTS AND DISCUSSIONS

The identified dataset was spitted as training and testing sets randomly. Ordinal entropy [20] was implemented to both training dataset and test dataset. Now the entropy and Gini impurity were implemented on the training and testing dataset. The highly influencing features of endometriosis were identified by using entropy and the Gini index. The Figure 2 illustrates the decision tree constructed using entropy. In figure X[0] represents Mass, X[1] represents blockages in fallopian tubes, X[2] represents tissue colour, and X[3] represents the size of the tissue. The features identified using entropy were the size of the tissue, blockage in fallopian tubes, and mass. Among all features, tissue size was identified as a predominant attribute in classifying the endometriosis with an entropy value greater than 0.75. The Figure 3 illustrates the decision tree constructed using the Gini index. In figure X[0] represents Mass, X[1] represents blockages in fallopian tubes, X[2] represents tissue colour, and X[3] represents the size of tissue [21].

The features identified using Gini index was size of tissue, blockage in fallopian tubes, mass, and tissue colour. Among all features, tissue size was identified as predominant attribute in classifying the endometriosis with Gini value of 0.42, Gini index for tube blockage was 0.236. The endometriosis influencing factors was analyzed by constructing decision tree algorithm. The performance of decision tree model was assesed by several metrics: i) precision, ii) recall, iii) specificity, iv) F1 score, and v) accuracy [22], [23]. Two factors including entropy and Gini index were considered for analysing the features of endometriosis. Among two factors Gini index outperforms well in terms of various metrics. The confusion matrix obtained was illustrated in Figure 4 for entropy as Figure 4(a) and Gini index as Figure 4(b)

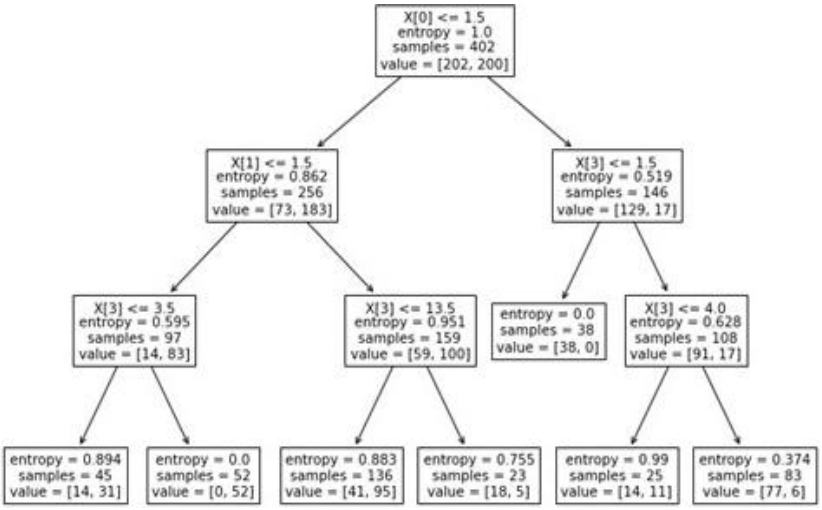


Figure 2. Construction of decision tree via entropy

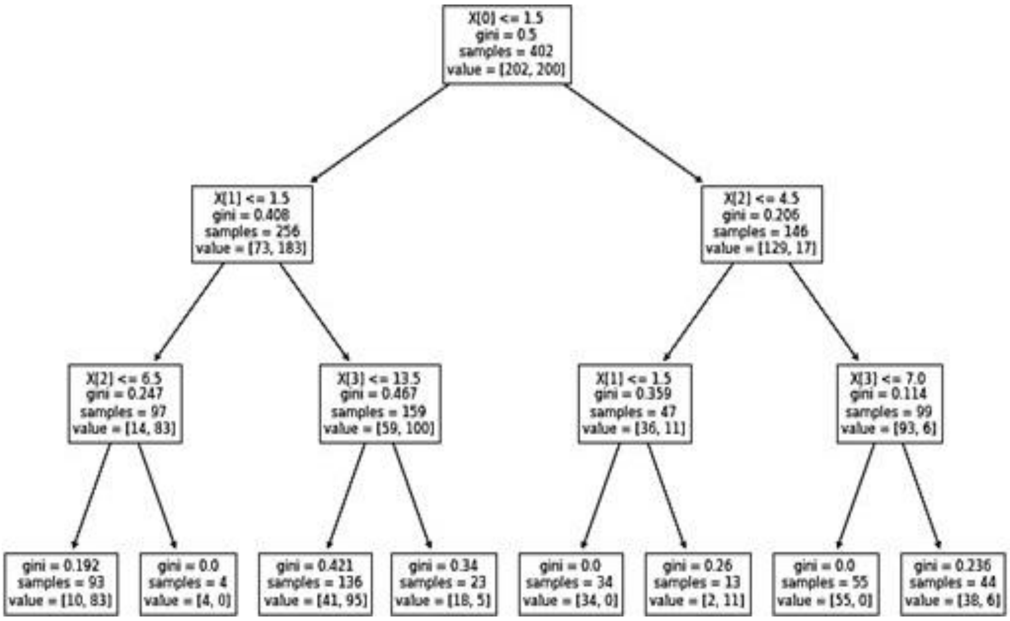
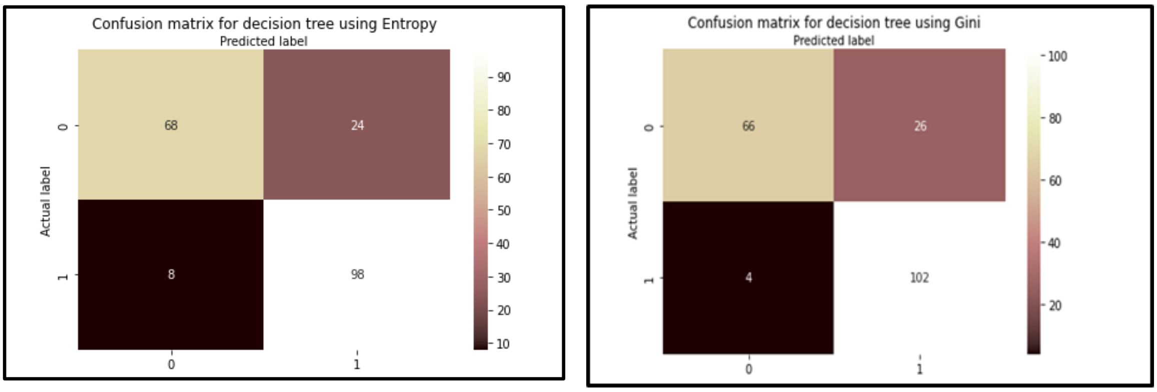


Figure 3. Construction of decision tree via Gini index



(a) (b)

Figure 4. Confusion matrix (a) confusion matrix using entropy and (b) confusion matrix using Gini index

The predicted and actual values for entropy was 68, 98, 8 and 24 respectively. Similarly, for the Gini index the true positive was 66, the true negative was 102, the false positive was 4, and the false negative was 26. Based on the classification matrix other metrics were evaluated and illustrated in the Table 1 and Figure 5.

The precision, recall, specificity, and F1 score for entropy were 89.47, 73.91, 92.45, and 80.94 respectively. Similarly for the Gini index, the precision was 94.2, the recall was 71.73. Specificity was 96.22 and the F1 Score was 81.44. Gini index outperforms well in terms of various metrics when compared to entropy. The next metric accuracy was evaluated [24]. The accuracy was computed for training data and test data. The executed model obtained training accuracy was 80.85% and testing accuracy was 83.84% for entropy. For the Gini index the model obtained the accuracy for training and test data was 84.08% and 84.85% respectively and the comparison was illustrated in Figure 6 and Table 2.

Table 1. Performance metrics comparison of entropy and Gini

	Precision	Recall	Specificity	F1
Entropy	89.47	73.91	92.45	80.94
Gini	94.2	71.73	96.22	81.44

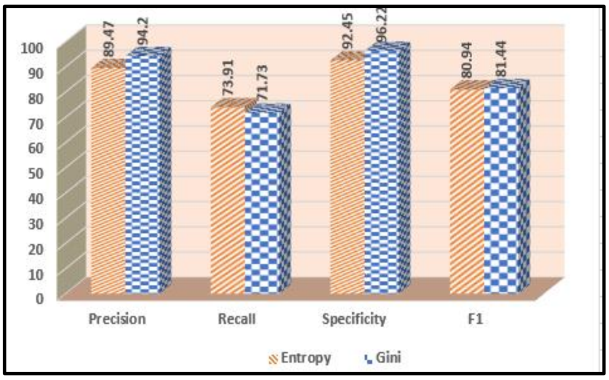


Figure 5. Performance metrics comparison of entropy and Gini

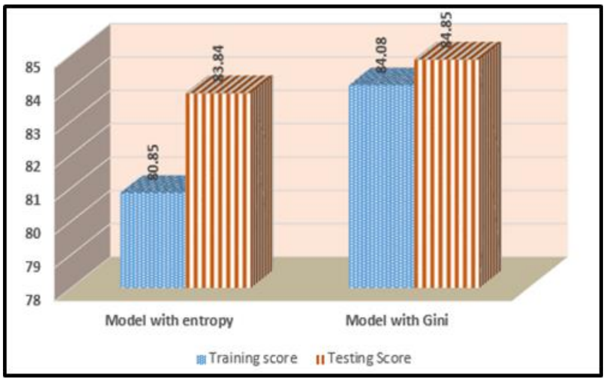


Figure 6. Accuracy score for entropy and Gini

Table 2. Accuracy score for entropy and Gini

	Model with entropy	Model with Gini
Training score	80.85	84.08
Testing Score	83.84	84.85

The other metric area under the curve [25] was evaluated for the given model in terms for both entropy and Gini index. AUC was constructed by plotting sensitivity against specificity. The AUC values for the prediction of endometriosis obtained for entropy were 0.87 and 0.89 as the AUC value for the Gini index as illustrated in the Figure 7.

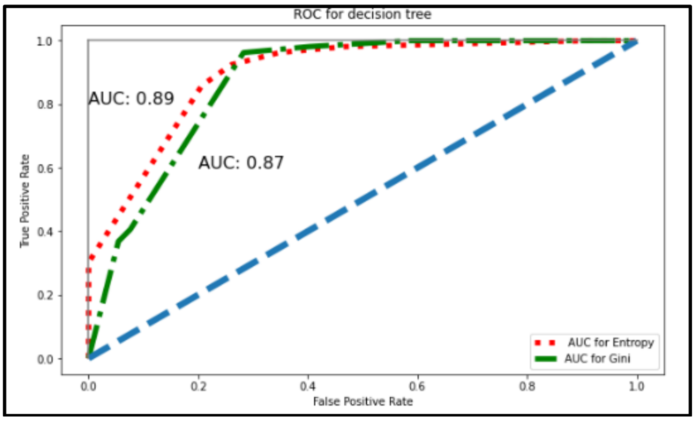


Figure 7. Area under curve for entropy and Gini

5. CONCLUSION

Endometriosis nowadays considered as a pretty common disease affecting 15% of women's global population. The impact of endometriosis affected women are more vigorous. From the laparoscopic surgery, several symptoms were identified and including mass, extra tissue size, extra tissue colour, and blockages in fallopian tubes. The decision tree algorithm evaluates the best features using the Gini index and entropy. The best features includes size of tissue, mass was more accurately predicted using Gini index with an accuracy of 84.85%, precision of 94.2%, recall of 71.73%, specificity of 92.45%, and F1 score of 81.44% respectively. The area under the curve obtained for entropy was 0.89 and the Gini index was 0.87 respectively. Gini index performs well in identifying the most suitable features of endometriosis.

REFERENCES

[1] A. Leibetseder, K. Schoeffmann, J. Keckstein, and S. Keckstein, "Endometriosis detection and localization in laparoscopic gynecology," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6191–6215, 2022, doi: 10.1007/s11042-021-11730-1.

[2] A. Javeed, S. U. Khan, L. Ali, S. Ali, Y. Imrana, and A. Rahman, "Machine learning-based automated diagnostic systems developed for heart failure prediction using different types of data modalities: a systematic review and future directions," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–30, 2022, doi: 10.1155/2022/9288452.

[3] T. Sudalaimuthu, "Endometrium phase prediction using k-means clustering through the link of diagnosis and procedure," *2021 8th International Conference on Signal Processing and Integrated Networks* 2021, pp. 1178-1181, doi: 10.1109/SPIN52536.2021.9566041.

[4] C. A. Ratanamahatana and D. Gunopulos, "Feature selection for the naive bayesian classifier using decision trees," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 475–487, 2003, doi: 10.1080/713827175.

[5] A. Renjini, M. S. Swapna, V. Raj, K. S. Kumar, and S. Sankararaman, "Complex network-based pertussis and croup cough analysis: A machine learning approach," *Physica D: Nonlinear Phenomena*, vol. 433, p. 133184, 2022, doi: 10.1016/j.physd.2022.133184.

[6] S. Narayan and J. Gobal, "Optimal decision tree fuzzy rule-based classifier for heart disease prediction using improved cuckoo search algorithm," *International Journal of Business Intelligence and Data Mining*, vol. 15, no. 4, p. 408, 2019, doi: 10.5812/ijcm.9176.

[7] R. Lusk *et al.*, "Exploratory analysis of machine learning approaches for surveillance of zika-associated birth defects," *Birth Defects Research*, vol. 112, no. 18, pp. 1450–1460, 2020, doi: 10.1002/bdr2.1767.

[8] N. Bourdel, P. Chauvet, and M. Canis, "Stratégies diagnostiques dans l'endométriose, RPC endométriose CNGOF-has," *Gynécologie Obstétrique Fertilité & Sénologie;* vol. 46, no. 3, pp. 209–213, 2018, doi: 10.1016/j.gofs.2018.02.008.

[9] V. Sabarinathan and V. Sugumaran, "Diagnosis of heart disease using decision tree," *International Journal of Research in Computer Applications & Information Technology*, vol. 2, no. 6, pp. 74-79, 2014, doi: 10.2174/1573405614666180322141259.

[10] V. Sankaravadivel and S. Thalavaipillai, "ymptoms based endometriosis prediction using machine learning," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 10, no. 6, pp. 3102-3109, 2021, doi: 10.11591/eei.v10i6.3254.

[11] H. Zhang *et al.*, "Working memory for spatial sequences: Developmental and evolutionary factors in encoding ordinal and relational structures," *The Journal of Neuroscience*, vol. 42, no. 5, pp. 850–864, 2021, doi: 10.1523/JNEUROSCI.0603-21.2021.

[12] M. Esteve, J. Aparicio, A. Rabasa, and J. J. Rodriguez-Sala, "Efficiency analysis trees: A new methodology for estimating production frontiers through decision trees," *Expert Systems with Applications*, vol. 162, p. 113783, 2020, doi: 10.1016/j.eswa.2020.113783.

[13] O. Popova, Y. Shevtsov, B. Popov, V. Karandey, V. Klyuchko, and A. Gerashchenko, "Entropy and algorithm of the decision tree for approximated natural intelligence," *Advances in Intelligent Systems and Computing*, pp. 310–321, 2018, doi: 10.1007/978-3-319-94229-2_30.

[14] E. Furman, Y. Kye, and J. Su, "Computing the gini index: A note," *Economics Letters*, vol. 185, p. 108753, 2019, doi: 10.7759/cureus.17636.

[15] A. Gupta, A. Anand, and Y. Hasija, "Recall-based machine learning approach for early detection of cervical cancer," *2021 6th International Conference for Convergence in Technology (I2CT)*, 2021, doi: 10.1109/I2CT51068.2021.9418099.

[16] E. K. Makowski *et al.*, "Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space," *Nature Communications*, vol. 13, no. 1, 2022, doi: 10.1038/s41467-022-31457-3.

[17] D. Parekh and V. Dahiya, "Predicting breast cancer using machine learning classifiers and enhancing the output by combining the predictions to generate optimal F1-score," *Biomedical and Biotechnology Research Journal (BBRJ)*, vol. 5, no. 3, p. 339, 2021, doi: 10.4103/bbrj.bbrj_131_21.

[18] A. Rachakonda and A. Bhatnagar, "ARatio: Extending area under the ROC curve for probabilistic labels," *Pattern Recognition Letters*, vol. 150, pp. 265-271, 2021, ISSN 0167-8655, doi: 10.1016/j.patrec.2021.06.023.

[19] E-Wen Huang *et al.*, "Machine-learning and high-throughput studies for high-entropy materials," *Materials Science and Engineering: R: Reports*, vol. 147, 2022, 100645, ISSN 0927-796X, doi: 10.1016/j.mser.2021.100645.

[20] J. H. Bamber and S. A. Evans, "The value of decision tree analysis in planning anaesthetic care in obstetrics," *International Journal of Obstetric Anesthesia*, vol. 27, pp. 55-61, 2021, doi: 10.1016/j.ijoa.2016.02.007.

[21] E. B. Palad, M. J. F. Burden, C. R. D. Torre, and R. B. C. Uy, "Performance evaluation of decision tree classification algorithms using fraud datasets," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 9, no. 6, pp. 2518-2525, 2022, doi: 10.11591/eei.v9i6.2630.

[22] F. J. M. Shamrat, R. Ranjan, K. Md, A. Y. Hasib, and A. H. Siddique, "Performance evaluation among ID3, C4. 5, and CART decision tree algorithms," *Pervasive Computing and Social Networking: Proceedings of ICPCSN*, vol. 317, 2021, p. 127, doi: 10.1007/978-981-16-5640-8_11.

[23] S. Khatri, D. Arora, and A. Kumar, "Enhancing decision tree classification accuracy through genetically programmed attributes for wart treatment method identification," *Procedia Computer Science*, vol. 132, pp. 1685-1694, 2018, doi: 10.1016/j.procs.2018.05.141.

[24] N. Sahani and T. Ghosh, "GIS-based spatial prediction of recreational trail susceptibility in protected area of Sikkim Himalaya using logistic regression, decision tree and random forest model," *Ecological Informatics*, vol. 64, p. 101352, 2021, doi: 10.1016/j.ecoinf.2021.101352.

[25] A. Myall *et al.*, "Predicting hospital-onset COVID-19 infections using dynamic networks of patient contacts: an observational study," 2021, doi: 10.1101/2021.09.28.21264240.

BIOGRAPHIES OF AUTHORS

Visalaxi Sankaravadivel is pursuing as research scholar in the Department of Computer Science and Engineering at Hindustan Institute of Technology and science, Chennai, India. She has taught in the field of higher education and research for ten years. She earned her Master Degree from SRM Institute of Science and Technology, Chennai, India. She is continuing her research work in analyzing of endometriosis using machine learning and deep learning. She has publications in journals and conferences around the world that are Scopus/SCI indexed. She received Australian Grant Innovation Patent. She can be contacted at email: sakthi6visa@gmail.com.

Sudalaimuthu Thalavaipillai is a Professor in the department of Computer Science at Hindustan Institute of Technology and Science, Chennai, India. From the Hindustan Institute of Technology and Science in Chennai, India, he received his PhD. He is a Specialized Ethical Hacker. He has published in more than 50 reputable international journals. He has won numerous honours throughout his career, including the Top Innovator Award and the Pearson Award for Best Teacher. His research interests include machine learning, grid and cloud computing, and cyber network security. He has lifetime memberships in IEEE, ACM, and CSI. He was awarded innovation patents from Australia, Germany, and India. He can be contacted at email: sudalaimuthut@gmail.com.

Surya Rajeswar Completed B.E in Computer Science and Engineering at Hindustan Institute of Technology and Science. He has completed Masters of Business administration at Hindustan Institute of Technology and Science. He is pursuing PhD in the department of management studies and his area of research includes stock price prediction using machine learning techniques. He has published papers in international conferences. He obtained Australian Innovation patent. He can be contacted at email: suryasurya123@yahoo.co.in.

Pon Ramalingam is working as a Registrar and professor in the Department of Management studies at Hindustan Institute of Technology and Science, Chennai, India. He obtained more than 15 publications in the reputed International Journals. His research areas includes management studies. He is a life time member of CSI, ISTE, he is a member of ACM and IEEE. He can be contacted at email: registrar@hindustanuni.ac.in.