

FRACTAL DIMENSION FEATURE BASED BREAST TUMOR CLASSIFICATION USING DATA MINING

A.Krishnaveni¹, M.S Irfan Ahmed²

Assistant professor¹, Professor² in Department of Computer Science 1&2
Thassim Beevi Abdul Kader College for Women, Kilakarai- 623517,
Ramanathapuram, Tamilnadu, India
E-mail: drakrishnaveni@gmail.com¹,directortbakc.rir@gmail.com²

Abstract

Cancer is a disease of Deoxyribonucleic Acid (DNA). A breast tumor is a malignant growth that creates from breast tissue. It is an illness wherein dangerous (malignant growth) cells structure in the tissues of the breast. Whenever cancer is analyzed as harmless, specialists will typically let it be instead of eliminating it. Imagine a future where users can predict when any disease might occur, treat any disorder in real-time, and even prevent diseases from ever happening. This is the future of medicine, one where we will be able to create a precise roadmap for a disease-free life. As an independent, non-profit researcher focused on genomics research has a century of experience, and the foresighted vision to make this exciting future possible. In this paper, Predicting a person's tumor (normal, benign, or malignant) based upon his or her tumor features like fractal dimension, symmetry, concave points, concavity, compactness, smoothness, area, texture, perimeter, and radius are calculated. Machine learning classification techniques and data mining algorithms can be used on these types of problem statements.

Keywords: DNA, ACS, Malignant, Disorder, Researcher, Fractal dimension

I INTRODUCTION

The Indian Council of Medical Research (ICMR) measures that there will be a twelve percent ascend in disease cases in India in the following five years. The most widely recognized types of disease influencing individuals in India are inseparable malignant growth, cervical disease, and oral malignant growth. One in each ten new malignant growth analyses in India yearly, is a bosom disease, according to reports and it causes the most number of disease-related passings in ladies. Inseparable disease happens in milk-delivering channels of ladies and other bosom cells, which partition quickly and structure bumps. They may ultimately spread to lymph hubs or somewhere else. Early determination, medical procedure, prescriptions, and radiation can effectively treat the illness. [1]. Protecting human beings from cancer is needed to improve our country in a healthy way. All of us know about health. Important cells in the human body have functioned in a well-defined manner is necessary. If not functioning means some cells are broken/damaged. These are called unhealthy cells/cancer cells. Abandoned intensification of the breast area severely causes breast cancer. So screening the mammogram of every human being is necessary to control/reduce breast cancer.

In the initial stage of cancer have been some symptoms of precaution like breast pain, emotional stress, a deposit of micro classification, increasing the breast density, etc. The mammography test is the primary and also referred by many doctors for scanning cancer in the initial stage. Areas like machine learning and medical image processing are always ready to give rapid

algorithms for computing breast growth for analysis. Medical images containing confidential information about patients' disease types like whether the mammogram input image has cancer or not being secured. Input images having some disturbances like blurred, low contrast, unwanted information like background, patient id, label, etc. So pre-processing is required to remove unnecessary data in the acquired image.

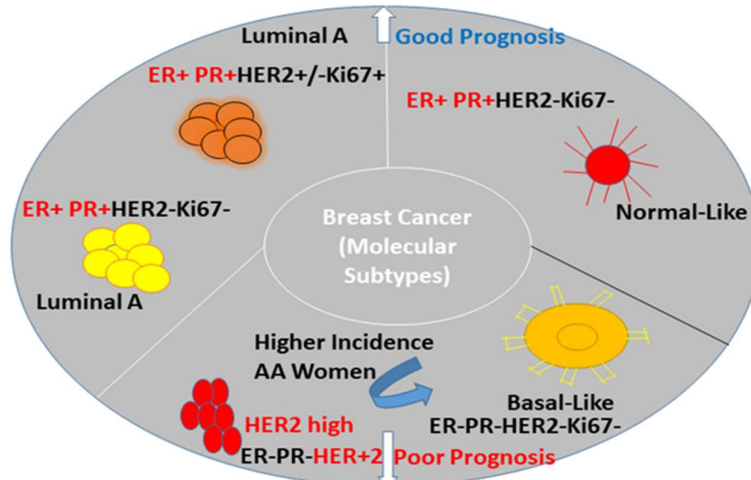


Figure 1: Breast Malignancy image [2].

In Figure 1, breast cancer and its molecular subtypes are explained. Estrogen (E+) Receptor-Positive, Progesterone (P+) Receptor-Positive, and HER-2(H+) Receptor Amplified are the basic three forms of breast cancer. Aromatase inhibitors, Tamoxifen, Herceptin, Trastuzumab, and Monoclonal Antibodies are prescribed treatments by Mark Adams. Triple-Negative Breast Cancer (TNBC) are (E-), (P-), and (H-) and recommended treatment is chemotherapy.

Medical images containing confidential information about patients' disease types like whether the mammogram input image has cancer or not being secured. Input images having some disturbances like blurred, low contrast, unwanted information like background, patient id, label, etc. So pre-processing is required to remove unnecessary data in the acquired image.

II LITERATURE REVIEW

Manish Charan et al. portrayed bosom disease as the second most normal malignant growth and the main source of malignant growth-related passing in ladies all over the planet [2]. The point of Habib Dhahri et al. denoted a hereditary user interface design strategy for selecting the appropriate elements with the help of machine learning [3]. Adam B Nover et al., assessed existing and arising innovations utilized for bosom malignant growth screening and discovery to recognize regions for potential improvement [4]. Omaer Faruq Goni et al., a review has recommended a profound neural organization with highlight choice methods to anticipate bosom disease. The examination of the recommended system is measured by exactness, precision, review, explicitness, awareness, f measure, and MCC [5]. Tanishk Thomas et al. expects to give a similar report by applying different AI calculations, for example, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, Decision Tree, K-implies, and Artificial Neural Networks on Wisconsin Diagnostic dataset to foresee bosom malignant growth at a beginning phase [6]. D. Sandeep et al., Lal Hussain et al., Dana Bazazeh et al., make a correlation of different AI calculations which considered and resolved the breast tumor [7-9].

Mamatha Sai Yarabarla et al., utilize the new advances in the improvement of CAD frameworks and related procedures. The backbone of the venture is to foresee whether or not the individual is having bosom malignant growth [10]. Nur Syahmi Ismail et al., contrasted bosom malignant growth identification and two model organizations of profound learning strategies. The general cycle includes picture preprocessing, order, and execution assessment [11]. Naresh Khuriwal et al. applied profound learning innovation to determine bosom disease to have an exactness of 99.67% [12]. Aditi Kajala et al. presents a short outline of the bosom disease conclusion utilizing AI calculations used to build the proficiency and adequacy of foreseeing malignant growth [13]. Shubham Sharma et al. introduced a correlation between the generally well-known AI calculations and strategies usually utilized for bosom disease expectation [14].

In medical research, correct predictions are needed to identify whether the person has the disease's tumor or not. Even though all the latest technologies came and were utilized for predictions of tumors but the mortality rate of a tumor in the breast is not down. To resolve this context the introduced work is needed to identify the most significant parameter involved in sample selection and classification of tumor class A or B is observed.

III MATERIALS AND METHODS

The Wisconsin Breast Cancer Dataset (WBCD) was downloaded from the UCI machine learning repository [16]. For the classification of breast cancer, real attributes in multivariate form have no missing values from five sixty-nine instances and thirty-two attributes. Based on the physiology element analysis some fluid samples were taken from the patient's breast for the creation of this dataset. Some important attributes are patient ID and category that belongs to benign or malignant. Tumor features like fractal dimension, symmetry, concave points, concavity, compactness, smoothness, area, texture, perimeter, and radius are calculated. From this feature list, the most significant feature identification is the primary task of our proposed work. Then the main aim is the classification of breast image whether the dead cells of tissues of the breast (malignant) or unrestrained growth of the cells but not affected seriously (benign) are identified.

Step1: Download WBCD Dataset from UCI Machine-learning repository

Step2: Data Discretization

Step3: Ten-fold cross-validation

Step4: Classification by different classifiers

Step5: Output data is either in Benign or in the malignant category.

The above-mentioned five steps are clearly explained in the following Figure 2.

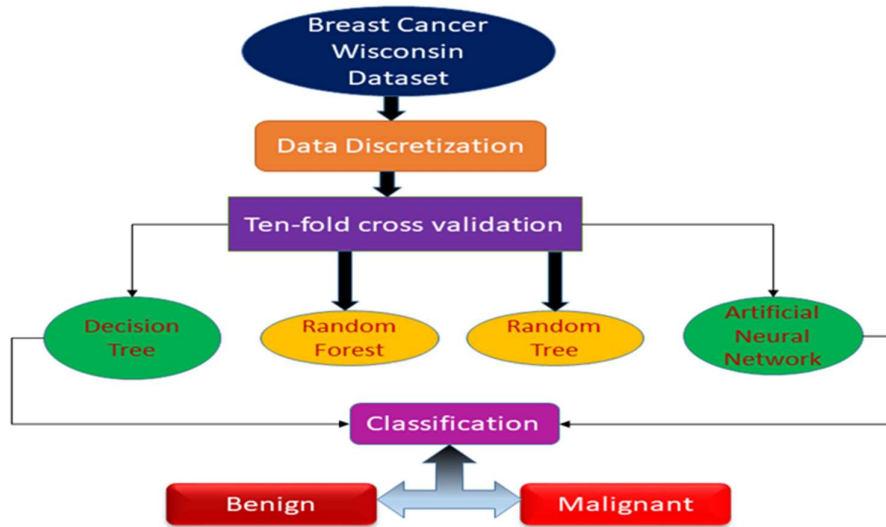


Figure 2: Framework for Classification

Early-stage tumor prediction is mainly observed and picturized in above Figure 2.

Classification:

In this research work, two classes are used for the prediction of breast tumors. CLASS A is used to represent the tumor which is not a serious case but is a benign strategy. CLASS B is involved in the prediction of malignant causes of breast cancer.

3.1 MODELS USED FOR CLASSIFICATION

Selecting the appropriate model for forecasting breast cancer in an earlier stage is a difficult process. The large volume of data set needs dimensionality reduction and type of the model like supervised or unsupervised. In supervised learning (classification) target variable has an either-or choice like in breast cancer, cancer, or non-cancer. Based on this observation and also obtained dataset having the defenseless variable CLASS A/ CLASS B. Therefore, the Decision table, Random forest, Random tree, and artificial neural network models are utilized and explained in the following.

1) Decision Table (DT)

A Decision Table (DT) is only a plain portrayal of all conditions and activities. Choice Trees are constantly utilized while the handling rationale is extremely convoluted and includes numerous circumstances. The primary parts utilized for the arrangement of the Data Table are Conditions Stubs, Action Stubs, and rules.

2) Random Forest (RForest)

It assembles choice trees on various examples and takes their greater part vote in favor of characterization and normal if there should arise an occurrence of relapse.

3) Random Tree (RT)

Random trees arbitrarily select the nodes and decide how to classify the given input randomly. It gives the solution strategy faster and also incorporates the details from the random forest.

4) Multi-Layer Perceptron (ANN)

Multi-facet perceptron characterizes the most perplexing engineering of counterfeit neural organizations. It is considerably shaped by different layers of the perceptron.

IV PERFORMANCE EVALUATION

The accuracy of introduced classifiers using ten-fold cross-validation with the performance of the presented model by Correlation Coefficient (CF), Despicable Complete Inaccuracy (DCI), Source Despicable Square off Miscalculation (DSM), Source Relation Square off Fault (RSF).

4.1 Correlation Coefficient (CC): It is utilized to quantify how solid a relationship is between two factors. It very well may be helpful in information investigation and displaying to all the more likely comprehends the connections between factors. The factual connection between the two factors is alluded to as their relationship. In the breast cancer dataset, the correlation coefficient is used to scatter the images in visualization. So how many features of the same variable made the difference in the entire data set was identified easily.

4.2 Mean Absolute Error (MAE): An average of all outright errors are called a Despicable Complete Inaccuracy (Mean Absolute Error).

MAE = Average of All outright errors


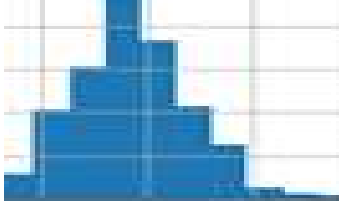

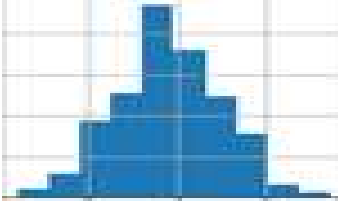
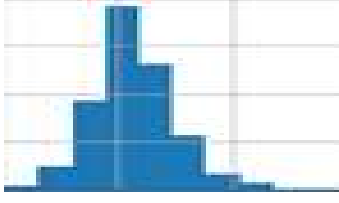
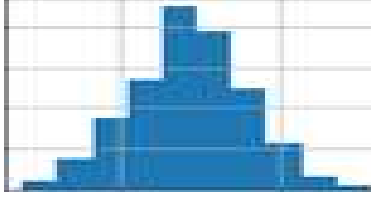
4.3 Root Mean Squared Error (RMSE): The relapse line fits the proportion of the informative elements is called root mean squared error. Root Mean Squared Error can likewise be interpreted as Standard Deviation in the residuals.

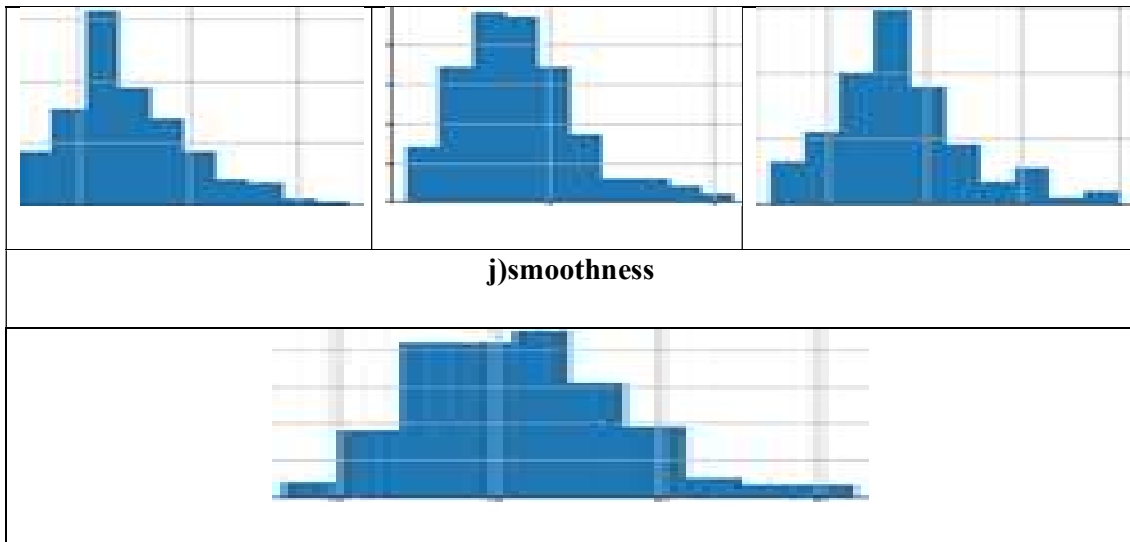
4.4 Relative Absolute Error (RAE): Overall outright blunder between two numeric vectors is known as SDSM. It is also known as a relative absolute error [27].

4.5 Relative Root Squared Error (RRSE): It is very useful to find out the error rate by using the direct pointer method [18].

From the following Table 1, visualization of all the specified attributes results is viewed by the Weka data mining tool after applying the ten-fold cross-validation. The purpose of choosing the ten-fold cross-validation is it returns the desired results in a fine-tuned manner.

Table 1 Visualization results obtained by the weka data mining tool

a) fractal dimension	b) area	c) symmetry
		
d) perimeter	e) texture	f) radius
		
g) compactness	h) concavity	i) concave point



In following Table 2, describes the various machine learning models (DT, RF, RT, ANN) used for classification, and the performance of each classifier is noted with the quality metrics such as CF, MAE, RMSE, RAE, and RRSE. The results show that ANN has a high accuracy rather than other models. Here classification error occurred during the report generated with the multilayer perceptron model like MLP zero. So test dataset faced this type of classification error and sensitivity is used to measure the positive results obtained by the introduced multilayer perceptron classifier.

Table 2 Classification results obtained by DT, RF, RT, and ANN

Metrics/ Models	CF	MAE	RMSE	RAE	RRSE
DT	0.7588	0.0082	0.0118	60.7786	65.1356
RF	0.9157	0.0053	0.0077	39.1832	42.4013
RT	0.7726	0.0085	0.012	63.0825	66.6229
ANN	0.9666	0.0028	0.0047	21.2167	25.92

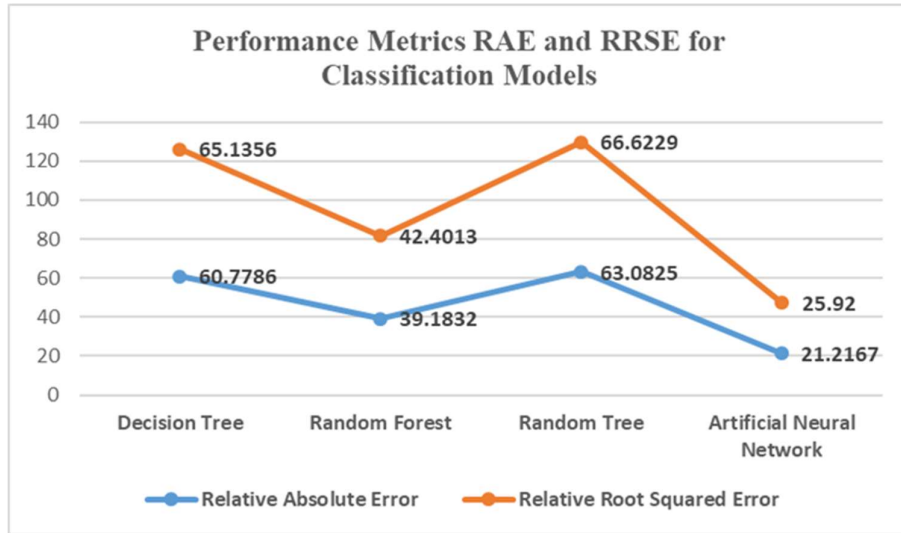


Figure 3: Classification Results

Figure 3, shows that Artificial Neural Network (ANN) has good performance. Then Random Forest has taken a place. Therefore ANN and Random Forest Classifiers give better classification results.

V CONCLUSION

Globe contains many objects which perform certain operations. Human perception is eminent to find out those objects are called images. We can't imagine our world without objects. Image capturing is the primary task for performing several operations. But the captured image should be in a clear manner such as high contrast, high quality, without blurred, without noise are required. Biomedical imaging is necessary to synthesize and analyze the internal regions of the human body without any changes. To give extra stamina to the classification is done by using machine learning models is addressed in this work. In this research work, DT, RF, RT, and Artificial Neural Network (ANN) with Ten-fold cross-validation are applied and accuracy was calculated by using error rates. Correlation Coefficient (CC), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Root Relative Squared Error (RRSE) are used to evaluate the performance of the presented model. Finally, ANN has a minimum error rate and high accuracy than other introduced models. In the future, machine learning models applied to various real-time applications of social media like Facebook, Twitter, and so on.

References

[1] News18.com
 [2] Manish Charan, Ajeet K. Verma, Shahid Hussain, Swati Misri, Sanjay Mishra, Sarmila Majumder, Bhuvanewari Ramaswamy, Dinesh Ahirwar, Ramesh K. Ganju, "Molecular and Cellular Factors Associated with Racial Disparity in Breast Cancer", *Int. J. Mol. Sci.* 2020, 21, 5936; doi:10.3390/ijms21165936
 [3] Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, Mohammed Faisal Nagi, "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms Hindawi Journal of Healthcare Engineering", Volume 2019, Article ID 4253641,11pp. <https://doi.org/10.1155/2019/4253641>

- [4] Adam B. Nover, Shami Jagtap, Waqas Anjum, Hakki Yegingil, Wan Y. Shih, Wei-Heng Shih, Ari D. Brooks, "Modern Breast Cancer Detection: A Technological Review" Hindawi Publishing Corporation, International Journal of Biomedical Imaging Volume 2009, Article ID 902326, 14 pages doi:10.1155/2009/902326
- [5] Omaer Faruq Goni, Fahim Sifnatul Hasnain, Abu Ismail Siddique, Oishi Jyoti, Habibur Rahaman, "Breast Cancer Detection using Deep Neural Network", 2020 23rd International Conference on Computer and Information Technology (ICCIT), 19-21 December 2020 IEEE
- [6] Tanishk Thomas, Nitesh Pradhan, Vijaypal Singh Dhaka, "Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey", Proceedings of the Fifth International Conference on Inventive Computation Technologies (ICICT-2020) IEEE
- [7] D. Sandeep, G.N. Beena Bethel, "Accurate Breast Cancer Detection and Classification by Machine Learning Approach" 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC) DOI: 10.1109/I-SMAC52330.2021.9640710 IEEE
- [8] Dana Bazazeh, Raed Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis", 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA) DOI: 10.1109/ICEDSA.2016.7818560 IEEE
- [9] Lal Hussain, Wajid Aziz, Sharjil Saeed, Saima Rathore, Muhammad Rafique, "Automated Breast Cancer Detection Using Machine Learning Techniques by Extracting Different Feature Extracting Strategies", 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE) IEEE
DOI: 10.1109/TrustCom/BigDataSE.2018.00057
- [10] Mamatha Sai Yarabarla, Lakshmi Kavya Ravi, A. Sivasangari, "Breast Cancer Prediction via Machine Learning", 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), DOI: 10.1109/ICOEI.2019.8862533 IEEE
- [11] Nur Syahmi Ismail, Cheab Sovuthy, "Breast Cancer Detection Based on Deep Learning Technique", 2019 International UNIMAS STEM 12th Engineering Conference (EnCon), DOI: 10.1109/EnCon.2019.8861256 IEEE
- [12] Naresh Khuriwal, Nidhi Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm", 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) DOI: 10.1109/ICACCCN.2018.8748777 IEEE
- [13] Aditi Kajala, V K Jain, "Diagnosis of Breast Cancer using Machine Learning Algorithms-A Review", 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3) DOI: 10.1109/ICONC345789.2020.9117320 IEEE
- [14] Shubham Sharma, Archit Aggarwal, Tanupriya Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms" 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) 10.1109/CTEMS.2018.8769187 IEEE
- [15] William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/] (1992).
- [16] <https://archive.ics.uci.edu/ml/datasets.php>

- [17] Arumugam K, Ramasamy S, Subramani D, “Binary Duck Travel Optimization Algorithm for Feature Selection in Breast Cancer Dataset Problem”, Vol.251, pp-157-167 Springer (2022).
- [18] <https://www.gepsoft.com/gxpt4kb/Chapter10/Section1/SS07.htm>
- [19] Krishnaveni Arumugam, Shankar Ramasamy, Duraisamy Subramani, “Chaotic Duck Traveler Optimization (cDTO) Algorithm for Feature Selection in Breast Cancer Dataset Problem”, Turkish Journal of Computer and Mathematics Education, Vol.12 No.4 (2021), pp-250-262.
- [20] A.Krishnaveni and Shankar, R. and Duraisamy, S., Versatile Duck Traveler Optimization (VDTO) Algorithm Using Triple Segmentation Methods for Mammogram Image Segmentation to Improving Accuracy (March 13, 2021). Available at SSRN: <https://ssrn.com/abstract=3803814> or <http://dx.doi.org/10.2139/ssrn.3803814>
- [21] A.Krishnaveni, R.Shankar, S.Duraisamy, “Duck Cluster Optimization Algorithm with K-Means Clustering for Mammogram Image Segmentation”, Solid State Technology, SCOPUS Indexed ISSN NO: 0038-111X Volume 63, Issue 6, (2020).
- [22] A.Krishnaveni, R.Shankar, S.Duraisamy, “An Efficient Methodology for Breast Tumor Segmentation using Duck Traveler Optimization Algorithm”, PalArch’s Journal of Archaeology of Egypt/Egyptology, Scopus Indexed ISSN NO: 1567- 214X Volume 17, Issue 9, (2020).
- [23] A. Krishnaveni, S.Duraisamy, R.Shankar, T.Latha Maheswari, “Heightened biased optimized duck traveler collaborative feature selection and classification for breast cancer dataset problem”, IJSRC, 3(2), pp-19-24 (2021).
- [24] Krishnaveni A, Shankar R, Duraisamy S, “A Survey on Nature Inspired Computing (NIC): Algorithms and Challenges”, Global journal of computer science and technology: D Neural & Artificial Intelligence volume 19 issue 3 version 1.0 Year (2019).
- [25] A.Krishnaveni , R. Shankar and S. Duraisamy, “A Review on various Image Thresholding Methods for Mammogram Image Segmentation” Compliance Engineering Journal ISSN NO: 0898-3577 Volume 11, Issue 2, (2020).
- [26] Krishnaveni Arumugam, Shankar Ramasamy, Duraisamy Subramani “Improved Duck and Traveler Optimization (IDTO) Algorithm: A Two-way efficient approach for breast tumor segmentation using multilevel thresholding”, European Journal of Molecular & Clinical Medicine ISSN 2515-8260 Volume 7, Issue 10, (2020).
- [27] <https://byjus.com/physics/errors-absolute-error-relative-error/>
- [28] M.S.Irfan Ahmed, A.Krishnaveni, “Hybrid Data Mining Method for Early Detection of Breast Cancer”, Webology, Volume 18, No.6, pp-959-978, (2021).

Authors Profile

Dr. Irfan Ahmed is currently working as the Director-Research, Industry-Institute Relations at Thassim Beevi Abdul Kader College for Women, Kilakarai, Ramnad, Tamilnadu. He received his MCA from Bharathidasan University in 1994. After that, he obtained his Ph.D in Computer Science with specialization in Trusted Networks from Algappa University, Karaikudi. Dr. Irfan’s research expertise includes Trustworthy Computing, Networking, and Data Mining, etc. He has authored 4 books and published more than 58 international and national journals along with several conference publications. Professor Ahmed has 27 years of experience including teaching, research, and industry. He has supervised 12 Ph. D students and currently, 3 more are under his supervision. He is the reviewer of a number of journals all over the globe. He has

given several invited and contributed presentations in multiple national and international conferences and symposia. He is the life member in IST (Indian Society for Technology Education), ACM (Association for Computing Machinery), Associate life member in IACSIT (International Association of Computer Science and Information Technology), a member in IAENG (International Association of Engineers).