Feasibility Study on Data Mining Techniques in Diagnosis of Breast Cancer

Keerthana Rajendran, Manoj Jayabalan, Vinesh Thiruchelvam, and V. Sivakumar

Abstract—**Survivability of patients suffering from breast cancer varies according to the stages. The early detection of breast cancer increase the longevity of patients. However, the number of risk factors involved in the detection exponentially increases with the medical examinations. The need for automated data mining techniques to enable cost-effective and early prediction of cancer is rapidly becoming a trend in healthcare industry. The optimal techniques for prediction and diagnosis differs significantly due to the risk factors. This study reviews article provides a holistic view of the types of data mining techniques used in prediction of breast cancer. On a whole, the computer-aided automatic data mining techniques that are commonly employed in diagnosis and prognosis of chronic diseases include Decision Tree, Na ïve Bayes, Association rule, Multilayer Perceptron (MLP), Random Forest, and Support Vector Machines (SVM), among others. The accuracy and overall performance of the classifiers differ for every dataset and thereby this article attempts to provide a mean to understand the approaches involved in the early prediction.**

Index Terms—**Breast cancer, data mining, early prediction.**

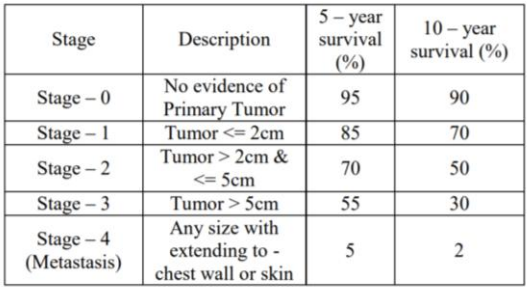
I. Introduction

Healthcare generates enormous amount of patient medical data in various structures and formats, which does not fit into the traditional processing. The big data processing framework and related technologies provide an opportunity to overcome the issues. The involvement of big data analysis in healthcare can curtail the cost of treatment and diagnosis of patients, reduce clinical trials, recognize patients who are prone for re-admission, enable real-time update of patient conditions, and precision medicine [1].

According to the World Health Organization (WHO), cancer is one of the leading causes of death worldwide, and the economic significance of cancer is prominent and increasing [2]. Cancer is defined as unlimited division and growth of substantial number of cells that form tissue masses known as tumor. They damage the normal cells and secrete hormones that modify the normal body functions. There are more than 100 types of cancer that have been classified based on the body parts such as breast, lung, brain, stomach, bone and colon cancers, to name a few common types. Generally, cancer can be grouped as benign (limited tumor growth at only one spot) and malignant (tumor moves to other body parts and damages the healthy tissues). When the tumor spreads to other regions of the body through the circulatory system, they invade and destroy the normal tissues in a process called metastasis [3], [4]. Recurrence and remission of cancer or its spread to other parts of the body will result in chronic cancer which will eventually cause fatality [5].

The most common cancer that occurs among women globally is breast cancer, which is the leading invasive cancer in developing countries. The benign breast lump is non-cancerous while the malignant breast lump is cancerous. The malignant cells can either grow in the breast (non-invasive) or spread to the surrounding tissues (invasive). Invasive ductal cancer is the most well-known type of invasive cancer which accounts to 80% of all types of breast cancer [6]. Usually, the assessments for presence of breast cancer in women include self-examination, mammography, ultrasound, cytology, and core-cut biopsy. Upon detection of breast cancer, this disease can be treated via surgery (mastectomy and lumpectomy), radiation therapy, chemotherapy, and hormone therapy. Table I shows the 5-year and 10-year survivability of breast cancer according to various stages.

TABLE I: Breast Cancer Survivability Rate [6]



Stage	Description	5 – year survival (%)	10 – year survival (%)
Stage – 0	No evidence of Primary Tumor	95	90
Stage – 1	Tumor <= 2cm	85	70
Stage – 2	Tumor > 2cm & <= 5cm	70	50
Stage – 3	Tumor > 5cm	55	30
Stage – 4 (Metastasis)	Any size with extending to - chest wall or skin	5	2

As the chances of survival differs largely by breast cancer stages, the earliest diagnosis will improve the rate of survival greatly. Women who were diagnosed at the early, non-invasive stage will have better chances of survival than those diagnosed at the later invasive stages. It is crucial for clinicians to diagnose the women who have breast cancer accurately and prevent false positive results. Therefore, The purpose of this study is to review the predictive models proposed for breast cancer. Further, the study will explore in detail to understand the current trends used to diagnose and predict the diseases. Previous data mining techniques implemented on some commonly available open source data such as Wisconsin Breast Cancer (WBC) dataset, Surveillance, Epidemiology, and End Results (SEER) dataset, and other openly available or real world datasets will be discussed.

Manuscript received November 20, 2018; revised April 19, 2019.

The authors are with the Asia Pacific University of Technology & Innovation, Kuala Lumpur 57000, Malaysia (e-mail: keer.abhitham@gmail.com, manoj@apu.edu.my, dr.vinesh@apu.edu.my, dr.sivakumar@apu.edu.my).

doi: 10.18178/ijmlc.2019.9.3.806

II. Data mining Techniques in Breast Cancer Diagnosis

This section will elaborate in detail on the current data mining techniques that are applied on a number of popular open-source breast cancer datasets. Each of these datasets carries different breast cancer related parameters and application of distinct data mining techniques.

The approaches employed on the Wisconsin Breast Cancer (WBC) dataset, which consists of the Wisconsin Breast Cancer Diagnosis (WBCD) dataset, and Wisconsin Breast Cancer Prognosis (WBCP) dataset will be highlighted. Moreover, the existing studies on the Surveillance, Epidemiology, and End Results (SEER) dataset will be discussed. Further, past works done on other breast cancer datasets such as those obtained from hospitals or from other openly available breast cancer databases will be elaborated.

A. WBC Datasets

Two studies have shown interest in building ensemble classifier (Random Forest) for the breast cancer diagnosis and prognosis [7], [8]. The significant amount of features utilized in the breast cancer diagnosis possess a major challenge to predict the cancer accurately. Therefore, the researchers have shown interest in selecting the features that are relevant for model development. The feature selection based on Bayesian probability and feature impurity were employed with backward elimination method [7]. This approach of constructing random forest classifier using feature selection produced better accuracies on the WBCD dataset with 99.82% and on the WBCP dataset with 99.7%.

The natural selection of the features to improve the accuracy of model may lead to constrained and unconstrained optimization problems. The Genetic Algorithm (GA) solves the optimization problems through evolving the population to get an optimum solution. One study portrayed that Rotation Forest with GA-based feature selection yielded the highest classification accuracy (99.48%) compared to Decision Tree, Bayesian Network, Logistic Regression, Random Forest (RF), SVM, Multilayer Perceptron (MLP), and Radial Basis Function Networks (RBFN) [9].

Further one study focused on the use of single SVM classifiers based on kernel functions and SVM ensembles using bagging and boosting [10]. The results shows that for small-scale dataset, GA with linear SVM ensembles by bagging at ROC 0.98 and GA with RBF SVM ensembles by boosting at accuracy 98.28%. For large-scale dataset, GA with RBF SVM ensembles using boosting proved to be a better prediction model than the rest of the classifiers with accuracy of 99.41% and ROC 0.875.

The feature selection using Pearson correlation coefficient and principal component analysis, and data discretization were employed on the WDBC dataset. The comparative study on different classifiers namely, Na ïve Bayes, SVM, and ensemble classifiers were implemented on the processed dataset. Na ïve Bayes yields the optimum accuracy of 97.39% on classifying the breast cancer with time complexity of 0.1020 milliseconds [11]. Further, study have shown improvement using the Sequential Minimal Optimization (SMO) to overcome the quadratic programming problem arises during the SVM training [12]. Therefore, providing flexibility to handle larger datasets and achieved 96.2%.

Two studies applied Particle Swarm Optimization for improving the feature selection and modeling using Decision Tree (C4.5) to improve the accuracy of early detection and Na ïve Bayes for early recurrence prediction [13], [14]. Another study have shown improvement in the SVM has highest specificity, accuracy (97%) and precision (97%) but RF has greater probability of discriminating between benign and malignant tumors with ROC of 99.9% [15].

In development of a Computer-Aided Diagnosis (CAD) system for breast cancer detection using deep belief networks and back propagation neural network [16]. Wisconsin Breast Cancer database was used to construct a pre-trained back propagation neural network with unsupervised phase deep belief networks, which achieved a greater classification accuracy compared to other classifiers with only one supervised phase. The model produced 99.68% accuracy with 99.47% specificity and 100% sensitivity.

One study have shown Expectation Maximization (EM) algorithm to cluster the Wisconsin Diagnostic Breast Cancer (WDBC) and mammographic mass that consist of benign and malignant instances. The dimensionality of the breast cancer risk factors were reduced using the Principal Component Analysis (PCA). The Classification and Regression Trees (CART) was employed on all the clusters formed from the datasets to classify breast cancer [17]. Further, fuzzy-based rules were applied for accurate prediction of the disease.

Another study utilized Fuzzy C-Means and Gustafson Kessel to obtain membership values from the medical data. The values were considered as additional informative features to improve the classification process [18].

B. SEER Breast Cancer Datasets

One study have shown interest to classify patients based on their breast cancer stages, as either carcinoma in situ or malignant [19]. The outcome of the model showed that the C4.5 algorithm yielded an accuracy of 93%. Another study, compared three machine learning techniques which are SVM, artificial neural network, and semi-supervised learning methods to allow the prognosis of breast cancer survivability using the SEER dataset [20]. As the dataset was large, class balancing was done on the positive and negative classes by randomly selecting 25,000 records from each class. Then, 5-fold cross validation was applied. The results showed that the best performance was achieved from the semi-supervised learning model where the accuracy was 71% and sensitivity was 76%.

The survivability rate of the patients differs due to varying factors. In [21] proposed a model to predict the 5-year survivability of breast cancer using SEER dataset. They employed logistic regression and decision tree methods and the data were divided using 10-fold cross validation method. The authors reported that logistic regression outperformed decision tree with better ROC curve (0.829) and g-mean (0.403). However, building a model on several factors might yield biased results. Therefore, one study consider Self-Organizing Map (SOM) and density based spatial clustering to create patient cohort. The importance of features in each cohorts was selected using the information gain and the model was created using MLP [22]. Another study, have built a predictive model for 5-year survivability of breast cancer on imbalanced data [23]. The class imbalance problem on the dataset were addressed by applying Borderline-Synthetic Minority Oversampling Technique (Borderline-SMOTE) and Density-based Synthetic Oversampling (DSO) methods. Combination of two feature selection methods, namely Particle Swarm Optimization (PSO) and Correlation-based Feature Selection (CFS) were applied to determine the key predictive variables. The predictive models were constructed using decision tree, Bayesian Network, and Logistic Regression. The authors reported that the hybrid approach using DSO + PSO_CFS + C4.5 yielded the highest efficiency with accuracy of 94.33%, sensitivity of 0.930 and AUC of 0.939.

TABLE II: SUMMARY OF BREAST CANCER DATASET

	Reference	Data mining technique	Performance measure	Scope of study
Breast cancer (WBC dataset)	[7]	Random forest	Accuracy = 99.82% (WBCD dataset), 99.7% (WBCP dataset)	Predictive model for breast cancer diagnosis and prognosis
	[12]	SMO (SVM)	Accuracy = 96.2%	Diagnostic model for breast cancer
	[9]	Rotation Forest	Accuracy = 99.48%	Breast tumor classification model
	[16]	Deep belief networks and back propagation neural network	Accuracy = 99.68%	CAD system for breast cancer diagnosis
	[15]	SVM	Accuracy = 97%	Detection and diagnosis of breast cancer
	[11]	Na ïve Bayes	Accuracy = 97.3978%	Classification of breast cancer stages
	[13]	Decision Tree (C4.5) and Particle Swarm Optimization	Accuracy = 96.49%	Early cancer prediction.
	[18]	Fuzzy C-Mean, Gustafson Kessel, and Support Vector Machine	Accuracy = 99.06%	To aid the process of data analysis and clinical decision
	[14]	Particle Swarm Optimization, Na ïve Bayes	Accuracy = 81.3%	Early recurrence prediction
	[8]	Random Forest	Accuracy = 98%	Early detection of breast cancer
Breast cancer (SEER dataset)	[19]	Decision Tree C4.5	Accuracy = 93%	Prediction of breast cancer stages
	[20]	Semi-supervised learning	Accuracy = 71%	Prognosis of breast cancer survivability
	[21]	Logistic regression	ROC curve = 0.829	Predictive model for 5-year survivability of breast cancer
	[23]	Decision Tree C4.5	Accuracy = 94.33%, ROC curve = 0.939	Predictive model for 5-year survivability of breast cancer
	[24]	Priority based decision tree	Accuracy = 98.51%, ROC curve = 0.989	Classification of breast cancer
	[22]	Self-Organizing Map, Density based spatial clustering, Neural Network	Accuracy = 78% to 90%	To predict survivability rate of different patient cohort.
	[23]	Density-based Synthetic Oversampling, Particle Swarm Optimization, Correlation-based Feature Selection and Decision Tree (C4.5)	Accuracy = 94.33%, Sensitivity = 0.930, G-mean = -0.939, and AUC = 0.939	Predictive model for 5-year survivability of breast cancer
Breast cancer (other datasets)	[25]	Weighed LS-SVM	ROC curve = 0.8465	Prognostic model in breast cancer therapy
	[26]	SVM and FFBP neural network	ROC curve = 0.7775 (SVM on micro calcification data), 0.9440 (FFBP neural network on masses data)	Classification of breast cancer
	[27]	k-NN	ROC curve = 0.604	5-year risk score model
	[29]	Decision Tree C4.5	Accuracy = 99%	Classification of breast cancer
	[28]	k-NN	Accuracy = 81%, ROC curve = 0.78	5-year survival prediction model for breast cancer
	[31]	Random Forest	Accuracy = 75.8% (dataset 1), 78.3% (dataset 2)	Classification and diagnosis of breast cancer
	[30]	Decision Tree, Neural Network	DT (Accuracy = 81.62%, Specificity = 79.80%, and Sensitivity = 89.49%) NN (Accuracy = 81.62%, Specificity = 89.99%, and Sensitivity = 90.80%))	Early cancer prediction.
	[32]	Support Vector Machine based Ensemble Learning	WBC (Accuracy = 97.10%, Specificity = 97.23%, and Sensitivity = 97.11%) SEER (Accuracy = 76.42%, Specificity = 72.80%, and Sensitivity = 80.02%)	To reduce diagnosis variance and increase accuracy

The priority based decision tree method on the SEER breast cancer dataset was employed to reduce the feature and improve computational time [24]. Data reduction was done using information gain based feature selection after which Decision Tree and priority based decision tree algorithms were implemented. The attributes were prioritized by user for the splitting of decision tree node using the latter algorithm. The results displayed that priority based decision tree classifier is a better model to classify the types of breast cancer in terms of lesser time complexity, greater accuracy (98.51%) and ROC (0.989).

C. Other Breast Cancer Datasets

One study have shown integration of the clinical and microarray data to enable an enhanced prognostic model in breast cancer therapy [25]. The five datasets were acquired from the Integrated Tumor Transcriptome Array and Clinical Data Analysis (ITTACA) warehouse where each dataset was transformed into a kernel matrix and an integration framework was generated. The authors proposed weighted Least Square-Support Vector Machines (LS-SVM) classifier to perform breast cancer prediction. The results showed that the weighed LS-SVM model established an optimized single framework to curb the problems of prohibitive diagnosis cost and variations in classifications due to heterogenous datasets with AUC value of 0.8465.

Further the researchers have shown the integration of Portuguese breast cancer database containing masses and micro calcification datasets to employ ensemble feature selection method for breast cancer classification [26]. Upon 10-fold cross validation method, feature ranking was done using several feature selection methods including the proposed ensemble method named RMean. The authors applied feed forward back propagation (FFBP) Neural Network, SVM, Linear Discriminant Analysis (LDA), k-NN, and Naïve Bayes classification methods on the datasets. The results displayed that the classification performances were better with the RMean method. The AUC scores were 0.7775 for SVM employed on micro calcification data and 0.9440 for FFBP Neural Network employed on masses data.

A 5-year risk score model for French women with breast cancer where the dataset was obtained from a large epidemiological cohort studies in France [27]. The authors employed k-NN algorithm on the imbalanced data and performed exhaustive search to generate the best possible combinations of attributes based on the restrictions set by the domain expert. The results highlighted that the k-NN model with combination of four attributes yielded the best risk score with AUC of 0.604.

Similarly, [28] explored on developing prediction models for breast cancer datasets with missing categorical values imputed using unknown to predict 5-year survival rate. The prediction models were constructed using k-NN, Logistic Regression, Decision Tree, and SVM. The results highlighted that k-NN achieved the best prediction model with greater than 81% accuracy and ROC value of 0.78.

The decision tree provides the intuitive approach to understand the contributing features of breast cancer. Therefore, another study employed decision tree algorithms namely, C4.5, Alternating Decision Tree, Classification and Regression Trees (CART), and Best First Tree classifiers on breast cancer dataset [29]. The dataset was obtained from a diagnostic center hospital in India, which consisted of breast cancer images and related attributes. The classifiers were tested using 10-fold cross validation and percentage split methods. The authors reported that C4.5 has the highest accuracy of 99% compared to other algorithms. In [30], the data was taken from Breast Cancer Research Center and the author applied Decision Tree and Neural Network to predict the early diagnosis of cancer.

A decision support system with the employment of data mining approach which could assist oncologists to classify and diagnose breast cancer [31]. Two Portuguese-based binary class mammography datasets were used in this study and key features in the images were then selected using feature extraction. Classification techniques such as SVM, k-NN, Decision Tree, Random Forest, and Naïve Bayes were implemented. The results showed that Random Forest had 75.8% and 78.3% accuracies for the two datasets respectively for masses and micro calcification recognitions. Nevertheless, Naïve Bayes proved to be a better classifier of breast cancer masses characterization with 83.1% accuracy.

One study explored the data repository at the Houston Methodist Hospital for breast cancer patients who have biopsy reports and Breast Imaging Reporting and Data System (BI-RADS) category 5 mammogram results [33]. The authors then established natural language processing (NLP) software algorithms to computationally derive mammographic features and pathological information from the text-based mammogram and biopsy reports. The NLP analysis was done using processing steps including tokenization, which used a Bayesian model, stemming, vector-space modeling, and calculation of similarity using Jaccard similarity coefficient. The findings showed that NLP tool could be used to differentiate breast cancer subtypes based on mammographic features, which reduces cost and time of manual analysis of breast cancer reports.

III. Discussion

The past research works done on breast cancer, will be briefed as means to provide information on the data mining techniques that have been explored for diagnosis and prediction of cancer. Table II provide the summary of the previous studies done on breast cancer diagnosis. Across he tables, the commonly used data mining techniques were SVM, Decision Tree, Random Forest and k-NN. These classification techniques are well-known in other fields of study as well besides cancer diagnosis. Most of the past works done on cancer were focused on developing prediction models for survivability at different time intervals, classification of the cancer stage or type as well as a computerized system for early diagnosis of cancer.

Looking across the breast cancer studies in Table II, each dataset has a commonly used data mining method. For WBC dataset, most of the researchers have reported successful prediction using SVM technique. Meanwhile, for SEER breast cancer dataset, Decision Tree C4.5 appear to have produced the most accurate prediction model. As for other datasets, k-NN and SVM have been adopted as comparatively better prediction models for breast cancer. One of the reasons behind the variations in the choice of the classifiers could be due to the different attributes found in each of this dataset. As the importance of the variables with regards to breast cancer prediction differs, the structure of the prediction model would also differ. Thus, most of the prediction models are dataset-specific. Even then, the accuracy achieved by the prediction models developed from the same dataset, such as WBC or SEER dataset, does differ. The scope of the study could be the reason behind this where different pre-processing steps will be applied on the datasets. Feature selection, class balancing, and hybrid data mining techniques can lead to the variation in the final prediction model.

One study has proposed SVM based ensemble algorithm to reduce the variation and improve diagnosis accuracy in both WBC and SEER database [32]. However, the class imbalance is one of the significant challenge in the medical dataset, which is not considered.

IV. Conclusion

The involvement of data analytics in healthcare to predict and diagnose the breast cancer are gaining immense interest. The survivability rate of the patients can be improved through early detection and recurrence. This study reviewed the WBC, SEER, and other publically available breast cancer datasets with regard to different data mining techniques applied on them. The accuracy of the model depends on the selection of the pre-processing techniques. The feature selection is the most commonly used techniques for identifying the prominent attributes for building the models. However, the application of the feature selections were performed on the limited set of risk factors leading to reducing the many unknown potential variables. The class balancing are gaining interest to reduce the biasness in the results.

There are several classifiers proposed in the existing studies for improving the accuracy rates. However, the concept of drift is not given enough consideration. Moreover, the application of a unified big data processing framework for breast cancer analysis are expected to be seen in the future.

References

[1] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Heal. Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, Dec. 2014.

[2] P. Ramachandran, N. Girija, T. Bhuvaneswari *et al*., "Early detection and prevention of cancer using data mining techniques," *International Journal of Computer Applications*, vol. 97, no. 13, pp. 975–8887, 2014.

[3] C. Nordqvist, "What you need to know about breast cancer," *Medical News Today*, 2017.

[4] R. J. Oskouei, N. M. Kor, and S. A. Maleki, "Data mining and medical world: Breast cancers' diagnosis, treatment, prognosis and challenges," *Am. J. Cancer Res.*, vol. 7, no. 3, pp. 610–627, 2017.

[5] American Cancer Society, "Managing cancer as a chronic illness," *American Cancer Society*, 2016.

[6] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, "Diagnosis of breast cancer using decision tree data mining technique," *Int. J. Comput. Appl.*, vol. 98, no. 10, pp. 16–24, Jul. 2014.

[7] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," *J. Biomed. Sci. Eng.*, vol. 6, no. 5, pp. 551–560, 2013.

[8] A. Basu, R. Roy, and N. Savitha, "Performance analysis of regression and classification models in the prediction of breast cancer," *Indian J. Sci. Technol.*, vol. 11, no. 3, pp. 1–6, 2018.

[9] A. Subasi, E. Alickovic, and J. Kevric, "Diagnosis of chronic kidney disease by using random forest," vol. 7, no. 1, pp. 589–594, 2017.

[10] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, "SVM and SVM ensembles in breast cancer prediction," *PLoS One*, vol. 12, no. 1, pp. 1–14, 2017.

[11] A. Hazra, S. K. Mandal, and A. Gupta, "Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and ensemble algorithms," *Int. J. Comput. Appl.*, vol. 145, no. 2, pp. 975–8887, 2016.

[12] V. Chaurasia and S. Pal, "A novel approach for breast cancer detection using data mining techniques," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 2, no. 1, pp. 2456–2465, 2014.

[13] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah, "Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis," *J. Phys. Conf. Ser.*, vol. 983, p. 012063, Mar. 2018.

[14] S. B. Sakri, N. B. Abdul Rashid, and Z. M. Zain, "Particle swarm optimization feature selection for breast cancer recurrence prediction," *IEEE Access*, vol. 6, pp. 29637–29647, 2018.

[15] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in *Proc. 2016 5th Int. Conf.Electron. Devices, Syst. Appl. (ICEDSA)*, 2016, pp. 1–4.

[16] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," *Expert Syst. Appl.*, vol. 46, pp. 139–144, 2016.

[17] M. Nilashi, A. P. D. O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," *Telemat. Informatics*, vol. 34, 2017.

[18] M. Abdullah, F. Al-Anzi, and S. Al-Sharhan, "Hybrid multistage fuzzy clustering system for medical data classification," in *Proc. 2018 International Conference on Computing Sciences and Engineering, ICCSE 2018 - Proceedings*, 2018, pp. 1–6.

[19] [19] K. Rajesh and S. Anand, "Analysis of SEER dataset for breast cancer diagnosis using C4.5 classification algorithm," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 1, no. 2, pp. 72–77, 2012.

[20] [20] K. Park, A. Ali, D. Kim, Y. An, M. Kim, and H. Shin, "Robust predictive model for evaluating breast cancer survivability," *Eng. Appl. Artif. Intell.*, vol. 26, pp. 2194–2205, 2013.

[21] K.-M. Wang, B. Makond, W.-L. Wu, K.-J. Wang, and Y. S. Lin, "Optimal data mining method for predicting breast cancer survivability," *Int. J. Innov. Manag. Inf. Prod.*, vol. 4, no. 2, pp. 28–33, 2013.

[22] N. Shukla, M. Hagenbuchner, K. T. Win, and J. Yang, "Breast cancer data analysis for survivability studies and prediction," *Comput. Methods Programs Biomed.*, vol. 155, pp. 199–208, 2018.

[23] S. M. Rostami and M. Ahmadzadeh, "Extracting predictor variables to construct breast cancer survivability model with class imbalance problem," *J. AI Data Min.*, vol. 6, no. 2, pp. 263–276, 2018.

[24] P. Hamsagayathri and P. Sampath, "Priority based decision tree classifier for breast cancer detection," in *Proc. 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2017, pp. 1–6.

[25] M. Thomas, K. De Brabanter, J. A. K. Suykens, and B. De Moor, "Predicting breast cancer using an expression values weighted clinical classifier," *BMC Bioinformatics*, vol. 15, no. 1, pp. 1–11, 2014.

[26] N. Pérez, A. Silva, and I. Ramos, "Ensemble features selection method as tool for breast cancer classification," *Int. J. Image Min.*, vol. 1, no. 2/3, p. 224, 2015.

[27] E. Gauthier, L. Brisson, P. Lenka, and S. Ragusa, "Breast cancer risk score: a data mining approach to improve readability.," *Int. Conf. Data Min.*, pp. 15–21, 2011.

[28] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," *Comput. Biol. Med.*, vol. 59, pp. 125–133, 2015.

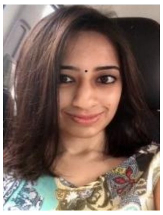
[29] E. Venkatesan and T. Velmurugan, "Performance analysis of decision tree algorithms for breast cancer classification," *Indian J. Sci. Technol.*, vol. 8, no. 12, 2015.

[30] A. Atashi, S. Sohrabi, and A. Dadashi, "Applying two computational classification methods to predict the risk of breast cancer: A comparative study," *Multidiscip. Cancer Investig.*, vol. 02, no. 02, pp. 08-13, Mar. 2018.

[31] J. Diz, G. Marreiros, and A. Freitas, "Applying data mining techniques to improve breast cancer diagnosis," *J. Med. Syst.*, vol. 40, no. 9, p. 203, Aug. 2016.

[32] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, 2018.

[33] T. A. Patel *et al.*, "Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods," *Cancer*, vol. 123, no. 1, pp. 114–121, 2017.



Keerthana Rajendran from Selangor, Malaysia was born on February 11, 1992. She pursued her undergraduate degree in B.Tech in biotechnology in 2011 at SRM University, Chennai, India and earned her first class degree with distinction in 2015. She furthered her postgraduate study in MSc in data science and business analytics in Asia Pacific University, Kuala Lumpur, Malaysia in 2016 where she was awarded a distinction for her degree in 2017.

She worked as a research assistant in University Malaya, Kuala Lumpur for a year carrying out her research on mesenchymal stem cells. She currently works as a junior data scientist in TradingPost International, Kuala Lumpur holding responsibilities to different levels of data management and analytics platforms to provide valuable business insights to the company. She published a research paper entitled "MicroRNA-590-5p Stabilizes Runx2 by Targeting Smad7 During Osteoblast Differentiation" in 2016 for the Journal of Cellular Physiology. She is passionate to continue her research on the application of predictive analytics in various fields of data science and enhance her knowledge in business intelligence tools.

Ms. Rajendran was awarded as an outstanding student in her postgraduate degree and received the best project achievement for her thesis work done in her university.

Manoj Jayabalan is a lecturer in the School of Computing, Asia Pacific University of Technology & Innovation. Manoj obtained his Master of science in software engineering from Staffordshire University, the UK with research area focusing on the database. He also holds a bachelor of engineering in computer science from Anna University, India.

He engaged in research activities focusing in the area of big data, data mining, machine learning, health informatics, and software engineering. His area of expertise is in the data analytics in performing data wrangling, and implementing models. He has supervised many industrial projects, master dissertations and mentored students for National level competitions. He has been invited guest speakers for several talks on big data and conducted many workshops.

Vinesh Thiruchelvam earned his B.S in electrical engineering from the University of Western Michigan, USA. He completed his PhD at University Tun Abdul Razak, Malaysia. He attained his PEng from the Board of Engineers Malaysia (BEM) in 2012, his CEng from Engineering Council, UK in 2011 and is a fellow of the Institute of Mechanical Engineers (IMechE-UK). He has managed international projects in the Maldives, India, Russia, Iran, UAE, Qatar, Saudi Arabia, Oman, Vietnam, Brunei and Malaysia for the property sector, ports, oil & gas and power plant industries. He is currently the dean of the Faculty of Computing, Engineering & Technology at Asia Pacific University (APU).

He has been involved in key education and engineering sectors such as being the chairman of the Engineering Education Technical Division (E2TD) at the Institute of Engineers Malaysia, advisory to IEMASB, member of the Ministry of Education's 'STEM Task Force' directly involved with the development of the MoE's 2015 Education Blueprint, member of the Ministry of Human Resources' BPIC on quality of graduates at the Ministry of Human Resources, member of National Task Force for Big Data Movements with Malaysian Digital Economy Corporation (MDeC), appointed to National Professor Council and Chair of the Centre of Analytics (APCA). His core scholarly research areas are in sustainable development, reliability engineering using smart devices with iot and data analytics with business intelligence.

V. Sivakumar received his Masters and Doctoral degree in Computer Science from Gandhigram Rural Institute, a Central Government University in India. He was the gold medalist of his batch. To add, he has done a Post Graduate Diploma in Applied Operations Research from Annamalai University and M.Phil. degree in Computer Science from Manonmaniam University in India. He has served as Assistant Professor in Gandhigram Rural Institute & VNR Vignana Jyothi Institute of Engineering in India and in the government universities of Ethiopia & Libya covering a total of around 20 years. His areas of research include Medical Imaging, Image Segmentation & Classification and Big Data Analytics. His research caliber includes 87 citation-index and 6 h-index. He has presented his research articles in several International Conferences. He was associated as Project Team Member for INDO-US 21st Century Knowledge Initiatives Awards, 2015 (4th Cycle) titled "Augmenting the curriculum of higher educational institutions with an on-line integrated cognitive-based employability skills assessment system using signal and video analytics", Indo US project awarded to Gandhigram Rural Institute, India in Collaboration with University of North Florida, USA.