World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:12, No:2, 2018

Comparative Evaluation of Accuracy of Selected Machine Learning Classification Techniques for Diagnosis of Cancer: A Data Mining Approach

Rajvir Kaur, Jeewani Anupama Ginige

***Abstract*—**With recent trends in Big Data and advancements in Information and Communication Technologies, the healthcare industry is at the stage of its transition from clinician oriented to technology oriented. Many people around the world die of cancer because the diagnosis of disease was not done at an early stage. Nowadays, the computational methods in the form of Machine Learning (ML) are used to develop automated decision support systems that can diagnose cancer with high confidence in a timely manner. This paper aims to carry out the comparative evaluation of a selected set of ML classifiers on two existing datasets: breast cancer and cervical cancer. The ML classifiers compared in this study are *Decision Tree (DT)*, *Support Vector Machine (SVM)*, *k-Nearest Neighbor (k-NN)*, *Logistic Regression*, *Ensemble (Bagged Tree)* and *Artificial Neural Networks (ANN)*. The evaluation is carried out based on standard evaluation metrics Precision (P), Recall (R), F1-score and Accuracy. The experimental results based on the evaluation metrics show that *ANN* showed the highest-level accuracy (99.4%) when tested with breast cancer dataset. On the other hand, when these ML classifiers are tested with the cervical cancer dataset, *Ensemble (Bagged Tree)* technique gave better accuracy (93.1%) in comparison to other classifiers.

***Keywords*—**Artificial neural networks, breast cancer, cancer dataset, classifiers, cervical cancer, F-score, logistic regression, machine learning, precision, recall, support vector machine.

I. Introduction

WITH advancements in Information and Communication Technologies, there has been surge in amount of data collected from various sources such as social media, corporate databases, financial and healthcare data. With this, policy makers and corporate are putting huge amount of money to get insight of valuable and hidden knowledge present in the data. In the same direction, Electronic Health (eHealth) is one such growing area which leverages computational methods for diagnosis of various diseases [1]. eHealth includes a range of components including early diagnosis of diseases, medical imaging, Internet of Things (IoT) for health, wearable technologies, electronic storage of data, telemedicine and robotic surgery [2].

In developing countries, where medical facilities are remote and number of medical personnel are in shortage, automated systems for early diagnosis of diseases plays vital role in healthy wellbeing of wide spread of population. Currently, Machine Learning (ML) and statistical methods are growing in popularity as a tool to process medical data in order to provide insight of diagnosis at an early stage. Various medical studies reported that cancer is one of the widely spread diseases which cost many lives every year [3]. Major cancer types that are highly prone include breast cancer, lung cancer, cervical cancer and skin cancer [4]- [6]. Cancer is a leading cause of death, accounting for 8.8 million worldwide in 2015 and the most common types of cancers are: lung cancer (1.69 million deaths), Liver cancer (788,000 deaths), colorectal (774,000 deaths), stomach cancer (754,000 deaths) and breast cancer (571,000 deaths) [7]. According to World Health Organisation (WHO) [8], in the field of health care, Australia is ranked 32 out of 190 countries. Due to increase in population, ageing, modern lifestyle and formation of new diseases have presented many new challenges. So, health organisations and state governments of every nation is trying to set procedures and plans to manage medical resources and infrastructure in order to give better living and services to all their residents and citizens.

Research studies also show that if cancer is detected in its early stage, then it can be cured [9]- [11]. Therefore, automated systems can save millions of lives making it possible to diagnose cancer at its early stage. Apart from this, manual diagnosis by health practitioners is subjective, tedious and provides high chances of error. Therefore, automated tools based on computational methods can serve as second eye to the diagnosis, providing higher confidence level and reducing chances of error.

In this context, this paper evaluates the accuracy of a selected set of ML techniques, using two pre-existing datasets on breast and cervical cancer. The remainder of this paper is organised as follows: Section II describes the literature review, which also highlights the steps associated with the standard data mining pipeline. Followed by the literature review, in Section III the methodology used in this research is presented, using the steps involved in standard data mining pipeline. In Section IV, experimental results are presented followed by comparison of our findings with other similar works. Lastly, the paper is concluded highlighting the challenges, limitations and possible future directions of this work.

II. Literature Review

Cancer is a disease which is caused by the uncontrolled growth of cells. This abnormal growth of cells invades nearby

Rajvir Kaur is with the School of Computing, Engineering and Mathematics at Western Sydney University, Australia, New South Wales (e-mail: 18531738@student.westernsydney.edu.au).

Jeewani Anupama Ginige is also with the School of Computing, Engineering and Mathematics at Western Sydney University, Australia, New South Wales (e-mail: j.Ginige@westernsydney.edu.au).

tissues and sometimes blocks blood vessels and lymphatic parts of the body [12]. Major body parts where cancer generally occurs include breast, lung, stomach, skin and liver. According to the American Cancer Society [13], breast cancer is the second leading cause of deaths among women today, followed by cervical cancer (or the cancer of cervix) [14]. In many developing countries, where due to lack of healthcare infrastructure and shortage of doctors, prognosis is very poor because of which many die from cancer although cancer could be treated with early diagnosis and treatment [15]. Prognosis and diagnosis of disease are important issue as it can save millions of lives [16]. Since manual processing could not be possible, machine learning techniques became prominent in medicine and healthcare, providing an alternate method for early diagnosis.

There are several methods of handling and processing of data but in this research the standard data mining pipeline for the machine learning techniques is used. The steps associated with standard data mining pipeline are described below:

STEP I:Data Acquisition: The very first step in every data mining research is to acquire the relevant data. Various sources of datasets include public repositories, various competition websites as well as data collected by private organizations and hospitals.

STEP II:Data Preprocessing: After acquiring data, data preprocessing is the foremost important step when we prepare data for mining as it may contain any incomplete or incorrect information which can lead to inaccurate results. Therefore, data preprocessing is very important step inorder to fill the missing values and information. There are several methods to handle and manage the missing data at its initial stage and choosing the right method depends upon the type of problem, for example:

- When class label is defined but attributes missing.
- If data is in the form of text, there can be any spelling mistakes.
- If data is the form of images or video, then there can be noise, occlusion or unclear information.

Therefore, if these types of problems are not solved properly at the beginning, then it can lead to poor performance of the system. So, there are number of ways to tackle with these problems such as:

- Ignore the entire row from analysis when class label or attributes are missing.
- Make use of global constant values like ”unknown”, ”N/A”, ”NIL” in order to fill the missing value because it does not make sense to predict the missing value.
- To replace the missing value with mean or median value.
- Making use of data mining algorithms such as regression, decision tree, clustering algorithms.

STEP III:Feature Extraction: After data pre-processing, feature extraction is done. The task of feature extraction is to reduce the dimensionality. When the input data is too large then the data is transformed to reduced set of features called feature vectors. The process of transforming data into feature vectors is called feature extraction. There are various machine learning algorithms used for feature extraction such as Principle Component Analysis (PCA) [17], robust PCA [18], kernel PCA [19], Independent Component Analysis (ICA) and Histogram of Oriented Gradients (HOG) [20].

STEP IV: Classification: After feature extraction, feature vectors are given as an input to the classifier to categorize the data and identify to which set of categories the input belongs to. Various classifiers used for classification includes Backpropagation Neural Network (BPNN), Hidden Markov Model (HMM) and so on. Following are the classifiers used for classification in this research:

1) *Support Vector Machine (SVM)*: SVM is one of the most important machine learning algorithms introduced by Vapnik in 1963. This method has many applications in medical/healthcare areas, pattern recognition, text classification and many more. SVM is supervised learning technique used for classification and regression analysis to maximize predictive accuracy. SVM uses hypothesis space of a linear function which is trained from optimization theory which implements a learning bias that is derived from statistical learning theory. The main goal of using SVM is that it separates the data with hyper planes and extends to non-linear boundaries using kernel trick.
2) *Decision Tree*: Decision Tree is a classifier which contains nodes, branches and leafs. The first node on the tree is called root node and each node is connected with one or more nodes using branches. In decision tree, each internal node represents a test and each branch represents the result of the test data and each leaf node represents a class label.
3) *k-Nearest Neighbor (kNN)*: kNN is a non-parametric method introduced by Fix and Hodges in 1951 for pattern classification. kNN is distance based classifier in which distance is used to classify test data based on labels of its neighbours which are selected from training data. It is important to choose an appropriate value of k (represents the number of neighbours) because if the value of k is very small then the classifier can be very sensitive to noise and on the other hand, if value of k is very large, then the neighbourhood may have too many points from other classes as well. So, the best way to choose the value of k depends upon the nature of data and it is important to choose odd numbers to avoid any ties.
4) *Logistic Regression*: Logistic Regression is supervised learning algorithm developed by David Cox in 1958 used for predicting the outputs of many possible outcomes. LR estimates the logistic function to get value between 0 and 1 to know the risk factor. The logistic regression is given by equation 1.

$$f(z) = \frac{1}{1+e^{-z}} \tag{1}$$

STEP V: Hyperparameter Optimisation: Machine Learning algorithms aims at finding a mapping function f that minimises a loss function $\mathcal{L}$ through optimization of training

criterion with respect to algorithm parameters θ. These set of parameters that model (or algorithm) has learnt are called model (or algorithm) parameters [21]. Apart from these model parameters, training algorithm might have its own parameters that need to be tweaked (or played with) to get best out of it. The set of external parameters that need to be controlled are called *hyperparameters*. Hyperparameters are important because they directly control the behavior of the training algorithm and have a significant effect on the performance of the model being trained. The final result of hyperparameter optimization is the winning set of model parameters that have produced the highest output in terms of classification accuracy. Various hyperparameter optimization techniques [22] include grid search, random search [21], spectral approach [23] and bayesian optimization [24].

STEP VI: Evaluation: After classification, comparative analysis is done based on various evaluation parameters by calculating Precision, Recall, F-score and Accuracy. All these evaluation metrics are calculated based on the confusion matrix as shown below.

		Actual value		
		p	**n**	**total**
Predicted outcome	**p′**	True Positive (TP)	False Positive (FP)	P′
	n′	False Negative (FN)	True Negative (TN)	N′
	total	P	N	

1) *Precision (P)*: Precision defines the fraction of correct positive observation.

$$Precision(P) = \frac{TP}{TP+FP} \tag{2}$$

2) *Recall (R)*: Recall defines the ratio of correctly predicted true observation.

$$Recall(R) = \frac{TP}{TP+FN} \tag{3}$$

3) *F-score*: F-score is weighted average of precision and recall.

$$F-score(F) = \frac{TP}{TP+FN} \tag{4}$$

when $\beta = 1$, then it is called as $F_1 score$ which is defined as given in equation 5

$$F_1 - score(F_1) = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

4) *Accuracy*: Accuracy defines the ratio of correctly predicted observations to the total observations in the dataset and is defined as given in equation 6.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{6}$$

Reference [14] compared performance of statistical methods with machine learning methods on classifying gene expression data. The experimental results in various studies [25]-[28] showed that ML algorithms outperformed classical statistical methods. In [29], three prominent ML algorithms namely, Artificial Neural Networks (ANN), Support Vector Machine (SVM) [30] and Fuzzy Logic (FL) are compared to classify benign and malignant state of breast cancer based on histopathology images collected from the Center for Bio-Image Informatics, University of California, Santa Barbara, USA. The experimental results based on evaluation metrics in terms of sensitivity, specificity and accuracy showed that fuzzy logic outperforms other classical classifiers due to its ability to reason over non-crisp decision boundaries. Moreover, fuzzy logic human like reasoning and provides privilege to tweak various thresholds depending upon the situation. Similar experiments were done by [12] and found that fuzzy logic has strength in managing classification problems where decision boundaries between classes are ambiguous.

Reference [31] did experimental study on breast cancer dataset based on SVM with three different kernel functions namely, linear kernel, Multi-Layer Perceptron (MLP) kernel and Radial Basis Function (RBF) kernel. Authors in [31] found that SVM with RBF kernel provides highest classification accuracy of 98.32%. Similar study was carried out by [32] to see the effect of various kernels on SVM classifiers performance. Pap smear test images of cervical cancer are taken and after pre-processing, segmentation and feature extraction, classification is done to find accuracy. Experimental results showed that out of linear, quadratic, Gaussian and multi-class polynomial kernels, polynomial kernel outperforms all other kernels giving highest classification accuracy. In order to provide second opinion to doctors, authors in [33] proposed Automatic Lesion Detection System (ALDS) for detecting melanoma (a kind of skin cancer) based on images. Experiments were performed on database of Dermatology Service of Hospital, Pedro Hispano, Portugal. First, pre-processing is done to improve image quality, after this segmentation is done jointly by watershed algorithm and active contours. Then, shape features are combined to form feature vector. Finally, classifiers are trained on the features extracted and classification accuracy is calculated. Comparative analysis of two classifiers based on ANN and SVM showed that ANN gives higher accuracy compared to SVM. Various experiments showed that ML algorithms are data hungry and they need huge amount of annotated data. Feature selection and tuning parameters of machine learning algorithms properly plays vital role in the performance of machine learning algorithms. [34] showed that selecting relevant features and wisely tuning SVM parameters improves the classification accuracy.

III. Proposed Methodology

The methodology of this research mimics the standard data mining pipeline. The steps in the standard data mining pipeline are applied as described below:

TABLE I
VARIOUS ATTRIBUTES OF BREAST CANCER DATASET

S.No.	Attributes	Domain
1	Clump thickness	1-10
2	Uniformity of cell size	1-10
3	Uniformity of cell shape	1-10
4	Marginal adhesion	1-10
5	Single epithelial cell size	1-10
6	Bare nuclei	1-10
7	Bland Chromatin	1-10
8	Normal nucleoli	1-10
9	Mitosis	1-10
10	Class	0-Benign 1-Malignant

TABLE II
OPTIMAL HYPERPARAMETERS FOR THE PROPOSED ANN CLASSIFIER

Parameter	Optimal Value
Learning rate (η)	0.01
Momentum (m)	0.1
No. of epochs (n_{epoch})	43
No. of hidden layers (n_{hidden})	1
No. of neurons in hidden layer ($h_{neurons}$)	10

STEP I:**Data Acquisition:** The dataset taken in this study is available online.

1) *Breast Cancer Dataset*: Breast cancer Wisconsin (original) dataset [35] has been taken from UCI Machine Learning repository[1], which contains 9 attributes and 699 number of instances, out of which 241(34.5%) have malignant tumors and other 458 (65.5%) were diagnosed as benign.
2) *Cervical Cancer Dataset*: Cervical cancer dataset has been taken from Kaggle [2] which consists of various machine learning datasets. Cervical cancer dataset consists of 1428 samples having 714 benign and 714 malignant samples. Each sample has 29 attributes.

STEP II:**Data Preprocessing:** During data preprocessing, the dataset consists of 16 missing values for bare nuclei attribute but class label is defined. For missing value, k-nearest neighbor classifier is used. As there are 9 attributes and each attributes domain ranges between 1-10, during data analysis, if the sum of 9 attributes falls below 40 out of 90, then it belongs to class 0 benign otherwise it belongs to class 1 malignant if sum value is more than 40.

STEP III:**Feature Extraction:** In this experimental study, the provided datasets have features that have range numerically from $1-10$ as given in Table I. The effect of various features on the overall classification is also analysed.

STEP IV:**Classification:** Though, there are lot many classifiers in machine learning but the most prominent ones includes Support Vector Machines, Decision Tree, kNN, Logistic Regression, Ensembled (bagged tree) and ANN. In this study, comparative analysis of these classifiers is done on cancer datasets.

STEP V:**Hyperparameter Optimisation:** Hyperparameter optimization(or tuning) plays an important role in ML. Hyperparameters are another set of parameters, apart from model actual parameters (weights and bias), that need to be properly tuned to get model's best performance. For optimization of neural network, the most common hyperparameters include *number of hidden layers*(n_{hidden}), *learning rate* (η), *momentum* (m), *number of epochs* (n_{epoch}), *neurons in hidden layers* ($h_{neurons}$). Similarly, for SVM hyperparameters include *epsilon* (ε), *kernel* and *soft margin constant* (C). In this study, we choose the hyperparameters from large set of options and the best model is selected providing highest classification accuracy.

STEP VI:**Evaluation:** The performance of various classifiers is measured based on evaluation metrics namely, Precision, Recall, F1-score and Accuracy. In all the experiments, k-fold cross-validation is used to avoid over-fitting of data. In this study, k=10 is taken for cross-validation and all experiments have been done in MATLAB (R2017a)[3].

[1]http://archive.ics.uci.edu/ml/datasets.html
[2]www.kaggle.com

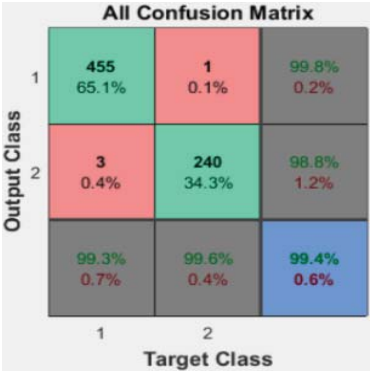

Fig. 1 Confusion matrix for classifying breast cancer dataset using ANN classifier

IV. EXPERIMENTAL RESULTS

In this section presents the experiments that were performed on two cancer datasets, namely breast cancer and cervical cancer. Confusion matrix illustrating number of labels having correct classification vs misclassification is also given.

Performance of various classifiers on breast cancer dataset: Cancer Wisconsin (original) is a standard dataset which contains 9 attributes and 699 number of samples. Each sample represents a particular class, where class 0 represents benign and class 1 represents malignant. An experimental study is done to investigate the performance of various ML classifiers in classifying medical data. Each classifier parameters are run by choosing different values for each hyperparameter and final classifier is set to the one that achieves better results. In our experimental study, ANN classifier outperforms all other classifiers and its final optimal hyperparameters are given in Table II.

From experimental results, as shown in Table III, it is seen that ANN gives higher accuracy (99.4%) in comparison to other classifiers. The results of ANN classifier are given in the form of confusion matrix as shown in Fig. 1. A confusion matrix contains the information about the actual and the predicted classification done by the system. Each column of the matrix represents the actual class and each row represents

[3]MATLAB: The Language of Technical Computing

TABLE III
PERFORMANCE OF VARIOUS CLASSIFIERS ON BREAST CANCER DATASET IN TERMS OF PRECISION (P), RECALL (R), F1-SCORE (F1) AND ACCURACY (ACC.) IN %

S.No.	classifier	P	R	F1	Acc.
1	Decision Tree	0.95	0.93	0.94	93.3
2	SVM (quadratic)	0.98	0.96	0.96	96.3
3	kNN (medium)	0.97	0.97	0.97	96.4
4	Logistic Regression	0.97	0.97	0.97	96.4
5	Ensemble (Bagged Tree)	0.98	0.97	0.97	96.7
6	ANN	0.99	0.99	0.99	99.4

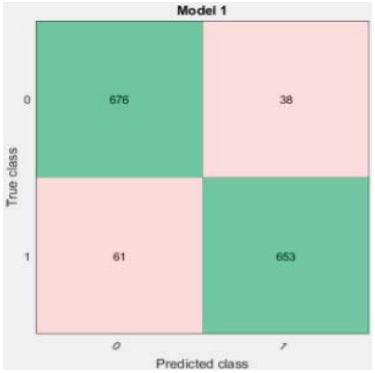

Fig. 2 Confusion matrix for classifying cervical cancer dataset using ensemble bagged classifier

the predicted class. If the actual and the predicted class fall under the same category or class, then the classification done by the system is true or positive. And if the actual and the predicted class fall under the different category or class, then results are false or negative or we can say that system has done misclassification. In confusion matrix, class 1 represents benign and class 2 represents malignant. As it is observed, out of 458 benign tumor samples, 455 samples are correctly classified and 3 samples are misclassified in class 2 and out of 241 malignant tumor samples, 240 samples are correctly classified and only 1 sample is misclassified in class 1. So, the overall accuracy is 99.4%.

Performance of various classifiers on cervical cancer dataset: Cervical cancer data is taken from kaggle, which contains 1428 number of samples, out of which 714 samples fall under class 0 (benign) and the other 714 samples fall under class 1 (malignant). So, we can say that the data is equally balanced between two classes. As it observed from Table IV, Ensemble (Bagged Tree) gives better accuracy in comparison to other classifiers. The result of Ensemble (Bagged Tree) classifier is displaced in the form of confusion matrix. As shown in Fig. 2, out 714 benign samples of class 0, 676 samples (cell with green color) are correctly classified and 38 samples are misclassified in class 1. Similarly, out of 714 malignant samples, 653 samples are truly predicted and 61 samples are misclassified. So, the overall accuracy is 93.1%.

V. SIMILAR WORK

In literature, many methods have been proposed to classify medical datasets. The effectiveness of any machine learning algorithm is based on:

- Ease of training the model
- Faster training time
- Requirement of less number of parameters
- High performance in terms of various metrics such as precision, recall, F1-score and accuracy.

TABLE IV
PERFORMANCE OF VARIOUS CLASSIFIERS ON CERVICAL CANCER DATASET IN TERMS OF PRECISION (P), RECALL (R), F1-SCORE (F1) AND ACCURACY (ACC.) IN %

S.No.	classifier	P	R	F1	Acc.
1	Decision Tree	0.88	0.89	0.88	88.9
2	SVM (quadratic)	0.50	0.87	0.63	50.0
3	kNN (medium)	0.84	0.90	0.86	86.7
4	Logistic Regression	0.58	0.48	0.51	56.7
5	Ensemble (Bagged Tree)	0.95	0.91	0.92	93.1
6	ANN	0.93	0.76	0.83	82.4

TABLE V
COMPARING PERFORMANCE OF THE PROPOSED METHODOLOGY WITH BASELINES IN TERMS OF ACCURACY

Author (year)	Accuracy (in %)
[36]	96.84
[37]	94.84
[38]	97.56
[39]	96.99
[31]	98.09
[40]	97.6
[11]	96.71
Our Proposed Methodology	**99.4**

In order to evaluate the effectiveness of our proposed methodology, comparison of various existing methods is done on breast cancer and cervical cancer datasets. As seen from results in Table V, it is clear that the proposed methodology outperforms various methods existing in literature, giving state-of-the-art classification accuracy.

VI. CONCLUSION

Cancer is one of most widely spread disease which takes millions of lives every year around the globe. Breast cancer and cervical cancer are the most common cancer types that occur in women. Manual diagnosis of cancer involves screening which involves many factors, including knowledge of the health practitioner, availability of screening setup as well as availability of medical facility within reach. Nowadays, automated systems based on machine learning algorithms can overcome such challenges, making better lives of individuals. This projects did comparative analysis of various machine learning algorithms, including K-NN, SVM, logistic regression, ANN, Decision Tree and ensemble bagged based classifiers. Experimental results in terms of Precision, Recall, F1-score and accuracy for various classifiers showed that ANN and ensemble based classifiers outperforms various other classifiers. Moreover, the proposed approach gives highest classification accuracy on breast cancer dataset as well as cervical cancer dataset which outperforms various state-of-the-art existing methods in the literature.

A. Challenges and Limitations

There were various challenges faced during this work. First and foremost challenge was non-availability of annotated data in medical domain. Various factors responsible for this include

privacy issues as well as anonymity of subjects. Apart from this, the available datasets were quite small in size. This is an issue because machine learning algorithms cannot be trained on small datasets as they are data hungry. It is worth highlighting that there was lot of bias in the available datasets because of which we have calculated precision, recall and F1-score along with classification accuracy to take account of biases in data.

B. Future Work

The proposed methodology in this work achieved 99.4% accuracy on breast cancer dataset. Though the results are pretty impressive but still there are lot of chances of improvement. First, though machine learning algorithms provide high performance but they do not provide reasoning how decision has been made. Machine learning algorithms acts as a black box wherein plenty of training data is provided and algorithms generalize well to predict the actual class on unseen data. Fuzzy logic provide reasoning like human beings making it possible to know how algorithm has reached to its decision. So, in future Adaptive Neuro-Fuzzy Inference System (ANFIS) could be used for classifying cancer data. Moreover, with availability of high training data, deep learning models can be trained to get better results.

REFERENCES

[1] A. E. K. Sobel, "The move toward electronic health records," *Computer*, vol. 45, no. 11, pp. 22–23, Nov 2012.

[2] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K. S. Kwak, "The internet of things for health care: A comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015.

[3] X. Ma and H. Yu, "Global burden of cancer," in *YALE Journal of Biology and Medicine*, vol. 79, no. 3-4, December 2006, pp. 85–94.

[4] A. I. of Health Welfare (AIHW), "Australian cancer incidence and mortality (acim) books: All cancers combined," in *ACIM books*, February 2017.

[5] J. Ferley, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. Coebergh, H. Comber, D. Forman, and F. Bray, "Cancer incidence and mortality patterns in europe: Estimates for 40 countries in 2012," in *European Journal of Cancer*, vol. 49, February 2013, pp. 1374–1403.

[6] J. A. Lewis and J. Bernstein, *Women's Health: A Relational Perspective across the Life Cycle*, ser. 1. Sudbury, Massachussets: Jones and Bartlett Publishers, 1996, vol. 1.

[7] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015. (Online). Available: http://dx.doi.org/10.3322/caac.21262

[8] C. J. Murray, J. Lauer, A. Tandon, and J. Frank, "Overall health system achievement for 191 countries," *Computer*, vol. 28, 2000.

[9] P. Bhati and M. Singhal, "Early stage detection and classification of melanoma," in *2015 Communication, Control and Intelligent Systems (CCIS)*, Nov 2015, pp. 181–185.

[10] Y. S. Cho, C. L. Chin, and K. C. Wang, "Based on fuzzy linear discriminant analysis for breast cancer mammography analysis," in *2011 International Conference on Technologies and Applications of Artificial Intelligence*, Nov 2011, pp. 57–61.

[11] R. Alyami, J. Alhajjaj, B. Alnajrani, I. Elaalami, A. Alqahtani, N. Aldhafferi, T. O. Owolabi, and S. O. Olatunji, "Investigating the effect of correlation based feature selection on breast cancer diagnosis using artificial neural network and support vector machines," in *2017 International Conference on Informatics, Health Technology (ICIHT)*, Feb 2017, pp. 1–7.

[12] A. Ali, S. M. Shamsuddin, A. L. Ralescu, and S. Visa, "Fuzzy classifier for classification of medical data," in *2011 11th International Conference on Hybrid Intelligent Systems (HIS)*, Dec 2011, pp. 173–178.

[13] A. C. S. ACS, "American cancer society, cancer facts and figures 2017," *Atlanta; American Cancer Society ;2017*, 2017.

[14] B. Shamsaei and C. Gao, "Comparison of some machine learning and statistical algorithms for classification and prediction of human cancer type," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Feb 2016, pp. 296–299.

[15] Y. Y. Leung, C. Q. Chang, Y. S. Hung, and P. C. W. Fung, "Gene selection in microarray data analysis for brain cancer classification," in *2006 IEEE International Workshop on Genomic Signal Processing and Statistics*, May 2006, pp. 99–100.

[16] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8 – 17, 2015. (Online). Available: http://www.sciencedirect.com/science/article/pii/S2001037014000464

[17] K. P. F. R. S., "Liii. on lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901. (Online). Available: http://dx.doi.org/10.1080/14786440109462720

[18] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, Jun. 2011. (Online). Available: http://doi.acm.org/10.1145/1970392.1970395

[19] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Advances in kernel methods," B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: MIT Press, 1999, ch. Kernel Principal Component Analysis, pp. 327–352. (Online). Available: http://dl.acm.org/citation.cfm?id=299094.299113

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, June 2005, pp. 886–893 vol. 1.

[21] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012. (Online). Available: http://dl.acm.org/citation.cfm?id=2188385.2188395

[22] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 2546–2554. (Online). Available: http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf

[23] E. Hazan, A. Klivans, and Y. Yuan, "Hyperparameter optimization: A spectral approach," *CoRR*, vol. abs/1706.00764v2, 2017. (Online). Available: http://arxiv.org/abs/1606.08140

[24] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 2951–2959. (Online). Available: http://dl.acm.org/citation.cfm?id=2999325.2999464

[25] J. Liu, X. Yuan, and B. P. Buckles, "Breast cancer diagnosis using level-set statistics and support vector machines," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2008, pp. 3044–3047.

[26] L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Transactions on Medical Imaging*, vol. 24, no. 3, pp. 371–380, March 2005.

[27] C. Y. Wang, C. G. Wu, Y. C. Liang, and X. C. Guo, "Diagnosis of breast cancer tumor based on ica and ls-svm," in *2006 International Conference on Machine Learning and Cybernetics*, Aug 2006, pp. 2565–2570.

[28] N. Prez, M. A. Guevara, A. Silva, I. Ramos, and J. Loureiro, "Improving the performance of machine learning classifiers for breast cancer diagnosis based on feature selection," in *2014 Federated Conference on Computer Science and Information Systems*, Sept 2014, pp. 209–217.

[29] F. T. Johra and M. M. H. Shuvo, "Detection of breast cancer from histopathology image and classifying benign and malignant state using fuzzy logic," in *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, Sept 2016, pp. 1–5.

[30] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995. (Online). Available: http://dx.doi.org/10.1023/A:1022627411411

[31] Z. Nematzadeh, R. Ibrahim, and A. Selamat, "Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques," in *2015 10th Asian Control Conference (ASCC)*, May 2015, pp. 1–6.

[32] D. Kashyap, A. Somani, J. Shekhar, A. Bhan, M. K. Dutta, R. Burget, and K. Riha, "Cervical cancer detection and classification using independent level sets and multi svms," in *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, June 2016, pp. 523–528.

[33] M. A. Farooq, M. A. M. Azhar, and R. H. Raza, "Automatic lesion detection system (alds) for skin cancer classification using svm and neural classifiers," in *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, Oct 2016, pp. 301–308.

[34] E. Olfati, H. Zarabadipour, and M. A. Shoorehdeli, "Feature subset selection and parameters optimization for support vector machine in breast cancer diagnosis," in *2014 Iranian Conference on Intelligent Systems (ICIS)*, Feb 2014, pp. 1–6.

[35] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Operations Research*, vol. 43, no. 4, pp. 570–577, 1995. (Online). Available: https://doi.org/10.1287/opre.43.4.570

[36] S.Aruna, D. S. Rajagopalan, and L. V.Nandakishore, "Knowledge based analysis of various statistical tools in detecting breast cancer," in *CCSEA 2011, AIRCCJ*, Aug 2011, pp. 37–45.

[37] D. Lavanya and D. K. Rani, "Ensemble decision tree classifier for breast cancer data," in *International Journal of Information Technology Convergence and Services (IJITCS)*, vol. 2, no. 1, February 2012, pp. 17–24.

[38] G.I.Salama, M. B. Abdelhalim, and M. Zeid, "Breast cancer diagnosis on three different datasets usign multi-classifiers," in *International Journal of Computer and Information Technology (IJCIT)*, vol. 1, no. 1, September 2012, pp. 36–43.

[39] A. C. Y and D. O. SivaPrakasam, "The negative impact of missing value imputation in classification of diabetes dataset and solution for improvement," in *IOSR Journal of Computer Engineering (IOSRJCE)*, vol. 7, no. 4, November 2012, pp. 16–23.

[40] H. Jouni, M. Issa, A. Harb, G. Jacquemod, and Y. Leduc, "Neural network architecture for breast cancer detection and classification," in *2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*, Nov 2016, pp. 37–41.

Rajvir Kaur Rajvir Kaur is doing Master of Research at Western Sydney Unversity, Australia. Her research interests include Machine Learning, Natural Language Processing and Clinical Text Processing.

Jeewani Anupama Ginige Jeewani Anupama Ginige is a senior lecturer at School of Computing, Engineering and Mathematics at Western Sydney University, Australia. Her research interest is in the field of Health Informatics, Data Analytics , Clinical Text mining and Machine Learning.