Comparative Analysis of Data Mining Algorithms for Cancer Gene Expression Data

Preeti Thareja, Rajender Singh Chhillar
Department of Computer Science and Applications
Maharshi Dayanand University, Rohtak, India

***Abstract*—Cancer is amongst the most challenging disorders to diagnose nowadays, and experts are still struggling to detect it on early stage. Gene selection is significant for identifying cancer-causing different parameters. The two deadliest cancers namely, colorectal cancer and breast malignant, is found in male and female, respectively. This study aims at predicting the cancer at an early stage with the help of cancer bioinformatics. According to the complexity of illness metabolic rates, signaling, and interaction, cancer bioinformatics is among strategies to focus bioinformatics technologies like data mining in cancer detection. The goal of the proposed study is to make a comparison between support vector machine, random forest, decision tree, artificial neural network, and logistic regression for the prediction of cancer malignant gene expression data. For analyzing data against algorithms, WEKA is used. The findings show that smart computational data mining techniques could be used to detect cancer recurrence in patients. Finally, the strategies that yielded the best results were identified.**

Keywords—Colorectal cancer; breast cancer; bioinformatics; data mining; WEKA; machine learning

I. Introduction

Non - communicable diseases (NCDs) responsible for 71 percent of all fatalities worldwide. Non-communicable diseases (NCDs) are illnesses which are never spread by pathogens. They are long-term illnesses with a sluggish course that are caused by a mix of biological, physiologic, ecological, and behavioral variables. Malignant is a non-communicable disorder in wherein some tissues grow out of control and extend to many other areas of the organism. Malignant is another name used for cancer that can begin practically at any place in the trillions of cells that make up the human body. Cancer is among the top contributors of death rate in India, accounting for 63 percent of all fatalities (9 percent). According to the National Cancer Registry Programme Report 2020, males will have a tumor incidence of 679,421 in 2020 and 763,575 in 2025, while women will have a tumor incidence of 712,758 in 2020 and 806,218 in 2025 [1]. As per research, oral, lung, and colorectal are the most frequent malignancies among men while breast and cervix uteri malignancies are most frequent amongst women. Cancer researchers require access to selected data from multiple sources in order to make advances. In medicine, data analysis has a remarkable ability to uncover hidden patterns in disease prediction [2,3].

As an emerging technique, cancer bioinformatics is one of the most important and valuable ways to facilities biochemical engineering for medical advancements, as well as improving the outcomes of cancer victims. Bioinformatics is focused on building an infrastructure to assist researchers in storing, analyzing, integrating, accessing, and visualizing large biological datasets and supporting information [4]. Bioinformatics is a computing platform that focuses on extracting information from biology content. It entails the creation of analysis tools and techniques to obtain, store, retrieve, manipulate, model databases, visualization, and estimation. As a result, custom analytics tools have become extremely important in bioinformatics, and they help to speed up the research process [5]. Sequencing and annotating an individual's entire collection of DNAs, for instance, are two common tasks in biotechnology. Led to the creation of machine learning techniques, bioinformatics models had to be manually configured, which is extremely challenging for problems like proteomics. Massive amounts of health data are gathered and made accessible to medical researchers because of the usage of computers employing automated technologies. As a matter of fact, Knowledge Discovery in Databases, which involves machine learning techniques, has now become a successful learning tool for health investigators to locate and manipulate correlations among a huge set of samples allowing them to foresee disease outcomes using specific instances stored in databases [6].

This study discusses recent research methodologies as well as an examination of predictive approaches, with a focus on classifying co-regulated genes according to their biological function. The work, basically, aims to find the foremost learning models for predicting cancer malignant gene expression data through study of related research work. Further, this study finds out the outperforming learning model by comparing them on certain performance metrics. The work has been divided into several sections. Section 2 talks about the recent advances in the field of cancer bioinformatics and most commonly used techniques for predicting cancer data. Section 3 presents the methodology employed in this research to analyze various algorithms on cancer datasets, as well as the measures used for evaluating their performances. Section 4 discusses the findings, which is proceeded by the conclusion in Section 5.

II. Related Research Work

This section will provide an overview of many cancers gene expression data-related research publications from a variety of databases, including IEEE Xplore, Google Scholar, Scopus, and Springer. Mostly, the publications are from year 2019-2021. This will aid in the discovery of techniques that have recently been used in the cancer detection. A list of prominent algorithms and performance metrics used in publications are produced in Tables I and II, respectively. These techniques will also be used to analyze performance.

Keerthika et al. [7] used information-mining strategy for proposing the cancer prediction model. This algorithm helps to find the amount of breast malignant that will occur in the near future. The main objective of this strategy is to safeguard users while also making it cheaper for them to use. Physical injury prevention and diagnosis will be aided by a prediction model. This discovery aids in detecting a person's risk of cancer at such an initial phase of treatment.

Changhee et al. [8] developed a better prognosis model based on machine learning named Survival Quilts. This model is being developed on the 10-year data of US prostate cancer patients to predict their mortality rate. Survival Quilts was compared with 9 prognosis models that are in clinical use and it showed a better decision curve.

T. Jayasankar et al. [9] used OGHO for optimal feature selection and kernel SVM (Support Vector Machine) in conjunction with gray wolf optimization algorithm to predict the breast cancer on the Wisconsin Breast Cancer dataset from UCI.

Heydari et al. [10] took a survey of leading data mining used for cancer detection in early phases. The author compared the top techniques of data mining and listed their advantages and disadvantages.

Byra et al. [11] proposed 2 CNN (convolutional neural network) techniques for breast malignant prediction. This method combines transfer learning with pre-trained CNN to produce excellent results of prediction.

M.A.Fahami et al. [12] clustered the colon cancer patients into 2 important categories and as a result they found out top 20 genes that are effective in both the categories.

Alireza et al. [13] applied novel and traditional data mining methods viz, linear vector quantization (LVQ) neural network (NN), multi-layer perceptron (MLP), Bayesian NN, Decision Tree (DT-C5.0), kernel principal component analysis with support vector machine (KPCA-SVM), and random forest (RF). The author clearly demonstrates the impact of machine learning technology on breast cancer recur classification. In compared to other approaches, the C5.0 and the KPCA-SVM have demonstrated to perform better in terms of accuracy. C5.0, on the other hand, had the finest sensitivity result.

Mostafa et al. [14] examined the results of different classification algorithms for identifying Colorectal Cancer (CRC), namely, J-48, Bayesian NN, RF, and MLP. All methods were found to be suitable and capable of providing reasonable results. J-48, on the other hand, performed best across the board.

Yanke et al. [15] mined 7 colorectal cancer related datasets using their new technique that combined NB (Naïve Bayes), RF and DT. After analysis the final result are optimized using an appropriate optimization technique. It was found that the proposed algorithm is superior than the SVM. This helps in better discovery of the genuine possibility of colorectal cancer target genes, and provides suggestions for its medical trials and promoting gene extraction.

Md. Rejaul et al. [16] created a technique for detecting the danger of stomach cancer beforehand. To acquire the feature score in a range of 0 to 1, the authors employed 5 distinct features extraction strategies along with ranker algorithms. The average rating was then used to provide a one exact score of each attribute. Then, apply predictive apriori algorithm to find the data's hidden pattern. The experiment had 300 patients, 150 of them were sick and the remaining 150 were not. Out of the 32 risk variables, they found 18 major risk factors for stomach cancer.

Ahmed et al. [17] proposed a Radial Basis Function NN (RBFNN) for diagnosing chronic diseases like breast cancer. The author has also compared his proposed method with other state of the art methods and found out that proposed method accuracy is the highest among all. Also, the author compared his method with learning classifier like RF, SVM, NB, ANN and many others. The result showed that his method is more accurate than other predefined classifiers.

Hooda et al. [18] employed a prediction model Bagoost to predict the breast cancer risk. The framework showed the accuracy of around 98%. The author states that it has better accuracy as compared to SVM, RF and adaboost.

Shanjida et al. [19] compared NB, k-nearest neighbor (kNN) and J48 data mining techniques on nine different types of cancer datasets. It was found that all the three algorithms were performing well but kNN outperforms the other two by a difference of around 0.4 percent in accuracy.

TABLE I. LIST OF ALGORITHMS USED IN RELATED RESEARCH WORK

Data Mining Techniques	**References**	**Count**
SVM	[9], [13], [15], [16], [18]	5
ANN	[11], [13], [14], [16]	4
NB	[13]-[16]	4
RF	[13]-[16], [18], [20]	6
DT	[7], [13]-[16], [19]-[21]	8
kNN	[19]-[21]	3
Bayesian NN	[13], [14]	2
LVQ-NN	[13]	1

TABLE II. LIST OF PERFORMANCE METRICS USED IN RELATED RESEARCH WORK

Performance Metrics	**References**	**Count**
Accuracy	[10]-[13], [17]-[19], [21]	8
Sensitivity	[11]-[14], [19]-[21]	7
Specificity	[11]-[14], [19]-[21]	7
ROC	[11], [14], [15], [18]	4
AUC	[11], [13], [15], [18]	4
F-Measure	[13], [14], [18], [19]	4
F1-Score	[20]	1
Decision Curve	[8]	1

Ray et al. [20] explored different classification methods (Gaussian NB, kNN, DT, RF) for detecting breast cancer involving both numeric and image datasets. It was found that the accuracy of RF was better in both numeric and image datasets.

Harikumar et al. [21] presented model that uses two machine learning (ML) techniques to categorize Breast Cancer (BC), viz. DT and kNN algorithms. Following feature selection with principal component analysis (PCA), these two techniques are tested on the BC dataset. The typical performance measures like accuracy, specificity, sensitivity, precision and Matthew's correlation are used to compare them. The findings show that the kNN classifier outperforms the DT classifier in the BC classification.

III. Methodology

The mathematical aspects of the problem are presented in this section of the paper. It all begins by outlining the key features of each of the data mining algorithms applied, with an emphasis on the description of the adjustable hyper-parameters. Fig. 1 highlights the methodology of this paper.

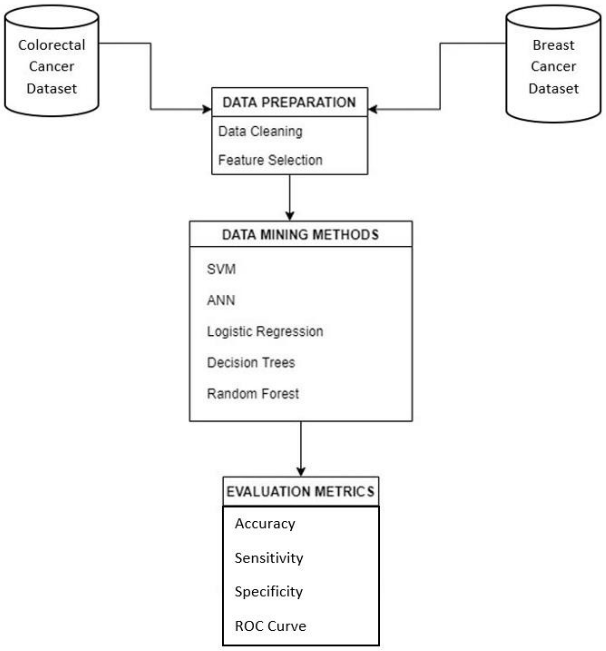


Fig. 1. Methodology for Cancer Gene Expression Data Prediction

A. Data Understanding and Preparation

The information was gathered from a public genomic database Gene Expression Omnibus (GEO) [22]. Two gene expressions namely, GSE45827 (Breast Cancer) and GSE41328(Colorectal Cancer). These are microarray datasets. Microarray data analysis is one of the most significant advances in statistical data and biology in the recent two decades. Microarray data may be examined using a number of methods and technologies. This section outlines a typical strategy for processing microarray data with Weka. The two most common cancers, Colorectal and Breast, will be studied here. The dataset for colorectal cancer contains 22284 attributes to be classified into four classes adenoma, carcinoma, metastasis and normal. The dataset for breast cancer contains 54676 attributes to be classified into six classes basal, HER, cell line, normal, luminal A and luminal B. Table III lists some properties of datasets.

TABLE III. Properties of Datasets

Properties	**Colorectal Cancer Dataset**	**Breast Cancer Dataset**
Number of Attributes	22284	54676
Number of Instances	55	151
Missing Values	No	No
Attribute Data Type	Numeric	Numeric
Target Attribute	Class	Class

B. Data Mining Methods

This section discusses a brief introduction of the selected data mining techniques to be applied on the selected dataset. The motive behind selecting these techniques is that they are widely used methods for analyzing bioinformatics dataset in state-of-the-art techniques.

1) SVM: SVM stands for Support Vector Machine and is a guided technique in data mining that may be used for regression and classification. SVMs are based on the concept of determining the optimal decision boundary for dividing a sample into two groups. SVM method discovers the points from both classes that are nearest to the boundary, which are called as support vectors. The separation seen between boundary and the support vectors is termed as margin. The main objective is to increase this margin. The optimum hyperplane is the one for which the margin is the greatest. As a result, SVM seeks to create a decision boundary with as much split into two different classes as feasible. The SVM method is depicted in Fig. 2.

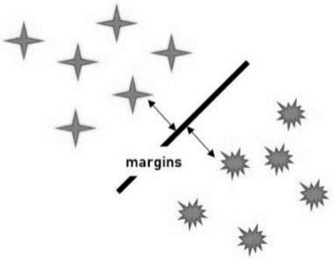


Fig. 2. SVM Method.

2) ANN: An Artificial Neural Network (ANN) is a system built on neurons that is motivated by biology neuron for the creation of artificial brains. It is built to evaluate and interpret data in the same way as beings do. Since more information becomes accessible, the algorithm may self-learn and provide superior outcomes. The inputs will be pushed into an aggregate of layers by an algorithm. A loss function must be used to measure the network's efficiency. The network may use the loss function to figure out which direction it wants to enforce to acquire the knowledge, as shown in Fig. 3. With the aid of an optimization, the net intends to promote its knowledge. For predictive analytics, ANN is hardly applied. The reason behind this is that ANN tend to over-fit the correlation in most instances. In most situations, ANN is employed when something that happened in the past is replicated almost identically in the same way.

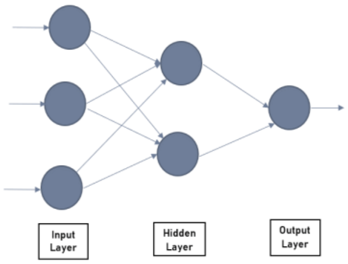


Fig. 3. ANN Method.

3) Logistic regression: Logistic Regression (LR) is analytical tool that is guided. This is a parametric regression models, meaning they employ numerical methods to make forecasts. The categorization issues are solved using logistic regression. The output is discrete value. The linear parameters are fitted to the sigmoid curve using logistic regression, as depicted in Fig. 4. Maximum likelihood estimation is the technique used to devise the loss function.

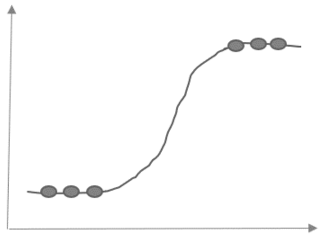


Fig. 4. LR Method.

4) Decision trees: A decision tree (DT) is a flowchart that aids in the decision-making process or displays statistical probabilities. A probable option, consequence, or response is represented by each node of the DT, as shown in Fig. 5. The tree's farthest branch reflects the outcomes of a particular choice route. DTs are a non-parametric ensemble learning approach. The aim is to understand basic rule base from extracted features to construct a system that anticipates the performance of the model. DTs are a prominent technique in neural networks and are widely used in business analytics to identify the best method for achieving a goal.

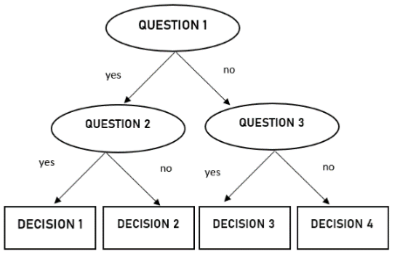


Fig. 5. DT Method.

5) Random forest: Random forest (RF) is a supervised learning algorithm that is commonly used to for classification and regression. It creates tree structure from several samples, using "majority vote" for classification and "average" for regression. RF collects data at random, creates a tree structure, and averages the results. It does not rely on any formulae. To create a RF, three important steps are there to follow. Firstly, randomly select slice of the whole dataset set for training particular trees separately. This is known as Bootstrapping or Sampling with Replacement. If these particular trees lack in connection, this RF Ensemble Learning performs well. Secondly, picking arbitrary features to examine at every node to accomplish connection. Lastly, Hundreds of times these steps are repeated to create a huge forest with a diverse range of trees. This variation is what distinguishes a RF from a single DT, as can be seen from Fig. 6.

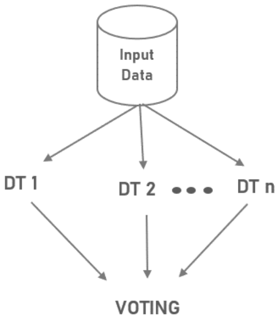


Fig. 6. RF Method.

C. Performance Metrics

In terms of performance, each prediction could be one of four kinds: True Positive, True Negative, False Positive, or False Negative. These entries are a part of confusion matrix which tells how well the classifier has performed. True Positive test is expected to be positive like saying the person is predicted to get sick but the label is really positive in terms of saying the person will get the sickness in actual. True Negative test is anticipated to be negative like saying the person is not predicted to get sick and the label is also predicted to be negative in terms of saying the person will not get the sickness in actual. False Positive occurs when a test is anticipated to be positive like saying the person is predicted to get sick) but the label is really negative in terms of saying the person will not get sick in actual. The test is “falsely” forecasted as positive in this scenario. False Negative test is expected to be negative like saying the person is predicted to not get sick but the label is really positive in terms of saying the person will get sick in actual. The test is “falsely” forecasted as negative in this situation. These values will help in determining the model’s accuracy, sensitivity, specificity, and receiver operating characteristic curve.

1) Accuracy: The number of properly categorised points (forecasts) divided by the total range of forecasts is Accuracy. Its value varies from 0 to 1. Accuracy is basically measured as shown in equation 1.

$$Acc = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \qquad (1)$$

2) Sensitivity: Among all of the algorithms, the one with the higher sensitivity should be picked. Sensitivity determines what percentage of true positives was accurately recognised as depicted in equation 2.

$$Sensitivity = \frac{True\ Positive}{True\ Negative + False\ Negative} \quad (2)$$

3) Specificity: The goal of specificity is to determine what percentage of real negatives were properly recognised. How many of the genuine negative cases were identified as such is the job performed by specificity. Specificity is calculated using equation 3.

$$Specificity = \frac{True\ Negative}{False\ Positive + True\ Negative} \quad (3)$$

4) Receiver operating characteristic: The trade-off between sensitivity and specificity is depicted by the ROC curve. Classifiers with curves that are closer to the top-left side perform better, as can be seen in Fig. 7. A random classifier is anticipated to give values that are falling diagonally mostly as baseline. The test becomes less accurate when the curve approaches the ROC space's 45º diagonally.

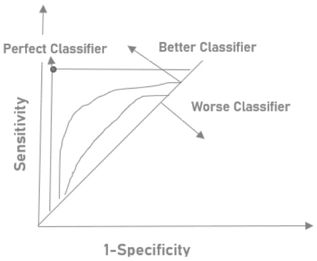


Fig. 7. ROC Curve.

D. K-fold Cross Validation

When there isn't enough data to utilize other more efficient approaches like the three - way division of training, validating and testing, then cross-validation is commonly used in deep learning to improve prediction performance. Initially dataset is scrambled such that the sequence of the inputs and outputs is totally arbitrary. This step is performed to ensure that our inputs are not skewed in any manner. The dataset then is divided into k equal portions. In this analysis, stratified 3-fold cross-validation is used. When cross-validation is used with the stratified sampling approach, the training and test sets contain the same fraction of the interested feature as the original dataset. When this is done with the class label, the cross-validation score is a close estimate of the generalization error.

E. Software Used

Weka – Waikato Environment for Knowledge Analysis is a machine learning package created by the University of Waikato in New Zealand [23]. The software is built in the Java programming language. It comes with a graphical interface and a variety of visualization tools and techniques for large - scale data processing. Data pre-processing, grouping, categorization, regressing, visualization, and dimensionality reduction are just a few of the common analysis methods that Weka offers. Weka v3.8.5 is used for the experiments on 11th Gen Intel® Core™ i5 @ 2.40 GHz with 8 GB RAM, windows 10 and 64-bit operating system.

IV. Experimental Results and Discussion

This research is intended to classify the selected datasets according to the class labels. The datasets are normalized and processed to reduce the unwanted features. Then, the data is passed through stratified 3-fold cross validation to separate it for training and testing purpose. The comparison of the two datasets, Colorectal and Breast, is done by applying algorithms, namely SVM, ANN, LR, DT and RF. The basis of comparison are the metrics, namely, Accuracy, Sensitivity, Specificity and Execution Time. Table IV shows the results of the performance metrics of Colorectal Cancer dataset and Table V shows the results of the performance metrics of Breast Cancer dataset.

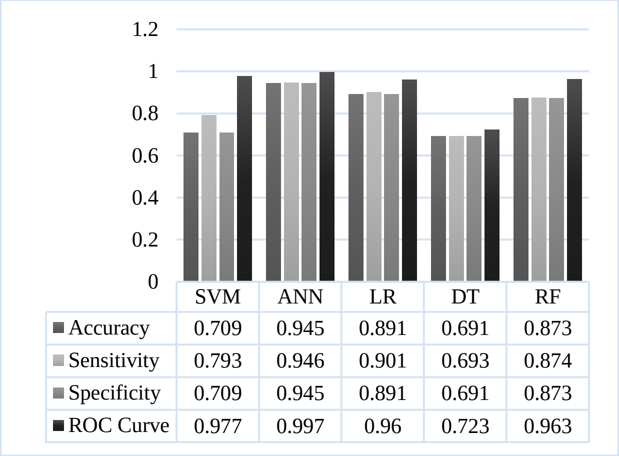
The results for the algorithms with comparison on performance metrics is also depicted through graphs as shown in Fig. 8 and 9, respectively.

TABLE IV. Performance Metrics of Algorithms against Colorectal Cancer Dataset

Algorithms	Performance Metrics			
	Accuracy	**Sensitivity**	**Specificity**	**ROC Curve**
SVM	0.709	0.793	0.709	0.977
ANN	***0.945***	***0.946***	***0.945***	***0.997***
LR	0.891	0.901	0.891	0.960
DT	0.691	0.693	0.691	0.723
RF	0.873	0.874	0.873	0.963

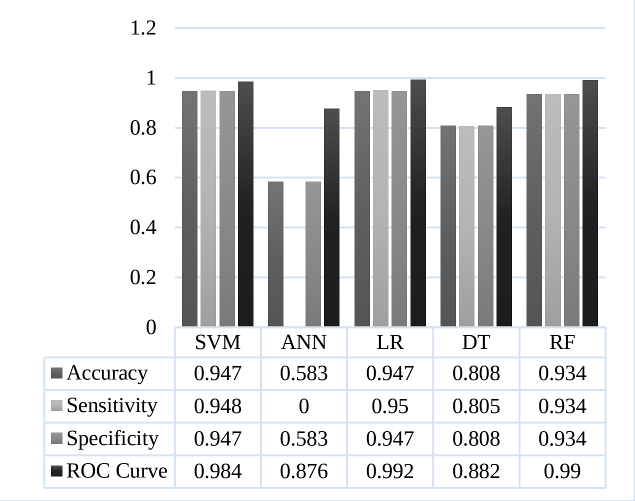
TABLE V. Performance Metrics of Algorithms against Breast Cancer Dataset

Algorithms	Performance Metrics			
	Accuracy	**Sensitivity**	**Specificity**	**ROC Curve**
SVM	***0.947***	0.948	***0.947***	0.984
ANN	0.583	NA	0.583	0.876
LR	***0.947***	***0.950***	***0.947***	***0.992***
DT	0.808	0.805	0.808	0.882
RF	0.934	0.934	0.934	0.990



	SVM	ANN	LR	DT	RF
▪ Accuracy	0.709	0.945	0.891	0.691	0.873
▪ Sensitivity	0.793	0.946	0.901	0.693	0.874
▪ Specificity	0.709	0.945	0.891	0.691	0.873
▪ ROC Curve	0.977	0.997	0.96	0.723	0.963

Fig. 8. Comparison of Performance Metrics against Colorectal Cancer Dataset.



	SVM	ANN	LR	DT	RF
■ Accuracy	0.947	0.583	0.947	0.808	0.934
■ Sensitivity	0.948	0	0.95	0.805	0.934
■ Specificity	0.947	0.583	0.947	0.808	0.934
■ ROC Curve	0.984	0.876	0.992	0.882	0.99

Fig. 9. Comparison of Performance Metrics against Breast Cancer Dataset.

As per the results of applying algorithms on Colorectal dataset, ANN shows the highest accuracy of around 95% with same level of sensitivity and specificity. The lowest level of accuracy, sensitivity and specificity is shown by decision trees. According to the results of applying algorithms on Breast dataset, SVM and logistic regression beats all others in accuracy of around 95%.

V. Conclusion

This work is carried to provide a brief outline of the state of art techniques SVM, ANN, Logistic Regression, Decision Tree and Random Forest applied on datasets for classification. These are applied on two most common problems prevailing in India, namely, Colorectal Cancer and Breast Cancer, in men and women, respectively. The experiment is carried out in Weka and the results are compared on certain metrics like accuracy, sensitivity, specificity and ROC. The experimental test shows an accuracy of 94.5% in colorectal cancer data with ANN outperforms all other algorithms. Similarly, an accuracy of 94.7% is found in breast malignant data with SVM and logistic regression beating all other algorithms. The input dataset has a significant impact on the limitations of an algorithm analysis. Like in breast dataset the ANN model is unable to recollect the results of sensitivity; it shows a question mark for it. This work can be improved by taking more folds in cross validation and also implying hybrid models for better analysis and results. Further, it can be compared for other datasets that include cancer patients of varying types of cancer.

References

[1] P. Mathur et al., “Cancer Statistics, 2020: Report From National Cancer Registry Programme, India,” JCO Glob. Oncol., no. 6, pp. 1063–1075, 2020, doi: 10.1200/go.20.00122.

[2] Aman and R. S. Chhillar, “Disease predictive models for healthcare by using data mining techniques: State of the art,” Int. J. Eng. Trends Technol., vol. 68, no. 10, pp. 52–57, 2020, doi: 10.14445/22315381/IJETT-V68I10P209.

[3] Aman and R. S. Chhillar, “Analyzing Predictive Algorithms in Data Mining for Cardiovascular Disease using WEKA Tool,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 8, p. 2021, Oct. 2021.z.

[4] “Bioinformatics, Big Data, and Cancer,” Cancer Research Infrastructure, 2020.

[5] P. Thareja and R. S. Chhillar, “A review of data mining optimization techniques for bioinformatics applications,” Int. J. Eng. Trends Technol., vol. 68, no. 10, pp. 58–62, 2020, doi: 10.14445/22315381/IJETT-V68I10P210.

[6] V. Krishnaiah, D. Narsimha, and D. Chandra, “Diagnosis of lung cancer prediction system using data mining classification techniques,” Int. J. Comput. Sci. Inf. Technol., vol. 4, no. 1, pp. 39–45, 2013.

[7] J. ; D. S. D. S. S. S. R. V. Keerthika, “Diagnosis of Breast Cancer using Decision Tree Data Mining Technique,” in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 1530–1535.

[8] C. Lee, A. Light, A. Alaa, D. Thurtle, M. van der Schaar, and V. J. Gnanapragasam, “Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database,” Lancet Digit. Heal., vol. 3, no. 3, pp. e158–e165, 2021, doi: 10.1016/S2589-7500(20)30314-9.

[9] T. Jayasankar, N. B. Prakash, and G. R. Hemalakshmi, “Big Data based breast cancer prediction using kernel support vector machine with the Gray Wolf Optimization algorithm,” in Applications of Big Data in Healthcare, Elsevier, 2021, pp. 173–194.

[10] H. Farzad; Rafsanjani, Marjan K, “A Review on Lung Cancer Diagnosis Using Data Mining Algorithms,” Curr. Med. Imaging, vol. Volume 17, no. 1, pp. 16–26, 2021.

[11] M. Byra, K. Dobruch-Sobczak, Z. Klimonda, H. Piotrzkowska-Wroblewska, and J. Litniewski, “Early Prediction of Response to Neoadjuvant Chemotherapy in Breast Cancer Sonography Using Siamese Convolutional Neural Networks,” IEEE J. Biomed. Heal. Informatics, vol. 25, no. 3, pp. 797–805, 2021, doi: 10.1109/JBHI.2020.3008040.

[12] M. A. Fahami, M. Roshanzamir, N. H. Izadi, V. Keyvani, and R. Alizadehsani, “Detection of effective genes in colon cancer: A machine learning approach,” Informatics Med. Unlocked, vol. 24, no. May, p. 100605, 2021, doi: 10.1016/j.imu.2021.100605.

[13] A. Mosayebi, B. Mojaradi, A. B. Naeini, and S. H. K. Hosseini, “Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer,” PLoS One, vol. 15, no. 10 October, pp. 1–23, 2020, doi: 10.1371/journal.pone.0237658.

[14] M. Shanbehzadeh, R. Nopour, and H. Kazemi-Arpanahi, “Comparison of four data mining algorithms for predicting colorectal cancer risk,” J. Adv. Med. Biomed. Res., vol. 29, no. 133, pp. 100–108, 2021, doi: 10.30699/jambs.29.133.100.

[15] Y. Li, F. Zhang, and C. Xing, “Screening of Pathogenic Genes for Colorectal Cancer and Deep Learning in the Diagnosis of Colorectal Cancer,” IEEE Access, vol. 8, pp. 114916–114929, 2020, doi: 10.1109/ACCESS.2020.3003999.

[16] M. Rejaul Islam Royel, M. Ajmanur Jaman, F. Al Masud, A. Ahmed, A. Muyeed, and K. Ahmed, “Machine learning and data mining methods in early detection of stomach cancer risk,” J. Appl. Sci. Eng., vol. 24, no. 1, pp. 1–8, 2021, doi: 10.6180/jase.202102_24(1).0001.

[17] A. H. Osman and H. M. A. Aljahdali, “An Effective of Ensemble Boosting Learning Method for Breast Cancer Virtual Screening Using Neural Network Model,” IEEE Access, vol. 8, pp. 39165–39174, 2020, doi: 10.1109/ACCESS.2020.2976149.

[18] N. Hooda, R. Gupta, and N. R. Gupta, “Prediction of Malignant Breast Cancer Cases using Ensemble Machine Learning: A Case Study of Pesticides Prone Area,” IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 5963, no. c, pp. 1–1, 2020, doi: 10.1109/tcbb.2020.3033214.

[19] S. K. Maliha, R. R. Ema, S. K. Ghosh, H. Ahmed, M. R. J. Mollick, and T. Islam, “Cancer Disease Prediction Using Naive Bayes,K-Nearest Neighbor and J48 algorithm,” 2019 10th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2019, pp. 1–7, 2019, doi: 10.1109/ICCCNT45670.2019.8944686.

[20] R. Ray, A. A. Abdullah, D. K. Mallick, and S. Ranjan Dash, “Classification of Benign and Malignant Breast Cancer using Supervised Machine Learning Algorithms Based on Image and Numeric Datasets,” J. Phys. Conf. Ser., vol. 1372, no. 1, 2019, doi: 10.1088/1742-6596/1372/1/012062.

[21] H. Rajaguru and S. R. Sannasi Chakravarthy, “Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer,” Asian Pacific J. Cancer Prev., vol. 20, no. 12, pp. 3777–3781, 2019, doi: 10.31557/APJCP.2019.20.12.3777.

[22] “GEO Dataset,” https://www.ncbi.nlm.nih.gov/gds/, 2021.

[23] “Weka-Data Mining with Open Source Machine Learning Software in Java,” https://www.cs.waikato.ac.nz/ml/weka/, 2021.