Letters in High Energy Physics
ISSN: 2632-2714
Volume 2024

Investigation of MIHP Codes Over Free Space Optical Communication

Mohamed Ismail A M[1], Sadham Hussain K[2], Sheik Alaudin A[3]

[1],[2]Assistant Professor ,[3]Final Year M.E. VLSI Design

Department of Electronics and Communication Engineering, Al-Ameen Engineering College (Autonomous), Erode – 638 104, Tamilnadu, India

Abstract

Introduction: Breast cancer has become the greatest frequent cancer among worldwide. Machine learning techniques contribute much to cancer prognosis.

Objectives: The prime focus of the work is to enhance the prognosis of breast cancer at an earlier stage using an ensemble of machine learning classifiers.

Methods: Next generation genetic sequences of homo sapiens, BRCA1 and BRCA2 from National Centre for Biotechnology Information were derived for prediction of breast cancer. The proposed ensembled classifiers by hard voting and soft voting, combined models like Decision Tree technique, SVM algorithm, LR statistical model, Linear Discriminant analysis model, Naive Bayes classifier and k-nearest neighbours' algorithm.

Results: Five ensembled models from 6 machine learning classifiers were concatenated for the prediction purpose. Classification accuracy of ensemble hard voting and soft voting classifiers were evaluated statistically. Soft voting classifier for model 1(DT & SVM) and model 2(DT, SVM &LR) achieved greatest value for classification performance metrics.

Conclusion: Among all ensembled models, model 1 as well as model 2 achieved maximum classification precision of 94%.

***Keywords*:** Breast cancer, Voting classifier, Machine learning models, Classification performance.

1. Introduction

Breast cancer has become the first common cancer surpassing lung cancer among women as per the latest statistics. This cancer forms either in breast ducts or breast lobules. An important cause for this malignancy is due to major changes in *breast cancer* gene 1 or gene 2 [1]. It is also characterized by mutations, chromosomal structure variations and copy number alterations [2]. A total of 2.3 million people is affected by breast cancer in the world and 684,996 deaths are estimated. Global cancer cases are expected to be 28 million in 2040 which is 47% more than the number of breast cancer cases in 2020 [3].

Many ML methods are used in breast cancer prognosis these days, which resulted in outstanding efficiency [4]. From complex datasets, these machine learning techniques can determine and detect patterns and relationships between them and predict future consequences of a cancer type simultaneously. Some classification models outperform certain other models in the classification accuracy for the purpose of breast cancer prediction [5]. According to major research, machine learning techniques enhance the cancer prediction precision (15–25%), mortality and recurrence [6]. Statistical measures to predict future result is implemented in several medical sectors especially cancer sector. This was characterized by both supervised as well as unsupervised classifiers with a ratio of 85:15 [7].

Hence, we proposed an ensembled classification approach which improve breast cancer prediction. An ensembled technique improve classification model performance and can achieve better performance than any individual model used in the ensemble technique. This method combined the predictions from many classification models. The voting classifier is a meta-classifier in which classification is done by majority voting [8]. In order to classify breast cancer, the ensemble approach proposed in the research use hard and soft voting classifiers. Decision tree technique, Support-Vector Machine algorithm, LR statistical algorithm, LDA classifier, Naive Bayes classifier and k-nearest

neighbours' algorithm were ensembled and outperformed all the corresponding base classifiers. Classification model performance of various combinations of above-mentioned algorithms were assessed using classification performance metrics.

2. Related work

Plethora of research in prognosis of breast cancer by machine learning classifiers have caught attention these years. The precise cause of breast cancer is still uncertain, but researchers identified some of the potential risk factors such as gender, permanent change in genetic sequence, age and life behaviour [9-10]. Machine Learning and Artificial Intelligence has an important role in detection of cancer but studies on cancer prediction have low implication [11-14]. Improved ML techniques are better in classification accuracy, even though the volume and complexity of the data is more [15].

Distinction between supervised and unsupervised classifier is done based on the data types and structures. ANN is proved as simplification of LR classifiers when compared with Logistic statistic regression [16]. k-NN classifier was used to forecast the survivability of BC patients and efficiency of the classifier was evaluated by measurement slope [17-18]. For achieving accurate and faultless breast cancer detection, Support Vector Machine method evaluate instances in grouping and outlier discovery. [19]. Genetic Algorithm along with a web oriented Gradient Boosting algorithm is used in detection of breast cancer [20]. The random forest classifier is used in the early diagnosis of breast cancer and the efficiency is calculated by evaluating accuracy, ROC curve and F-measure [21]. The breast tumour patterns mostly occurred were considered as restrictions in cancer detection which made the base classifiers using an AdaBoost regulation and therefore resolved classification errors [22].

Ensemble classification algorithms such as k-NN, ANN, NB, J48, zeroR, simple cart, cv parameter selection and filtered classifier acquired an accuracy of 77.01% [23]. By combining boosting ANNs (BANN) and SVM, a new breast cancer diagnosis method proved a classification accuracy of 100% which concentrated on few ensembled method [24]. A multi layered ensemble classifier is combined with base classifiers (i.e., BayesNet classifier and Naïve Bayes classifier) and classification efficiency is evaluated with classification performance metrics, achieved 98.07% accuracy. By the combination of 12 different support vector machines, accuracy of breast cancer diagnosis was increased by 33.34% and diagnosis accuracy variance was decreased by 97.89%. These ensembled classifiers performed better than the corresponding individual base classifiers.

3. System Description

The prime purpose of this research is to find an optimal ensembled classification model for improving the accuracy of breast cancer prediction. Rather than depending on individual base classifiers, an improved voting ensemble of 6 classifiers are implemented. To categorize the genetic sequences into normal, BRCA1 and BRCA2, ensemble voting machine learning methods are used. Different combinations of various ensemble of classifiers such as DT technique, SVM algorithm, LR model, LDA classifier, Naive Bayes classifier and k-nearest neighbours' algorithm are concatenated for the prediction purpose. Classification accuracy of these ensembled hard and soft voting classifiers are evaluated by classification performance metrics. The flow diagram showing the proposed ensembled method for breast cancer prediction is depicted in Fig 1.

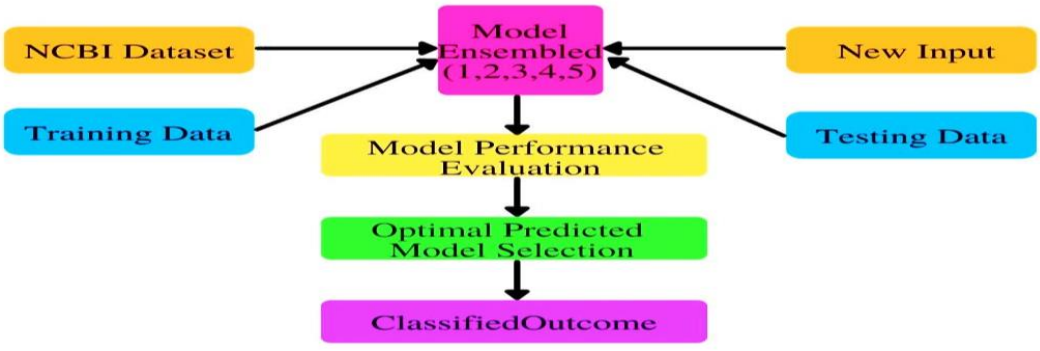

Fig 1. Ensembled Method for Breast Cancer Prediction

3.1 Data Extraction

Genetic data of homo sapiens, breast cancer gene1 and breast cancer gene 2 as class 0,1 and 2 were retrieved as. fasta file from NCBI for classification. The number of both benign and malignant genetic sequences extracted for the breast cancer prediction were 1580. The Fasta formatted sequence has '>' sign as first, description of sequence as second, followed by a gap and the corresponding genome. A variable holds the derived sequences for ensembled classification purpose.

3.2 Voting Ensemble Classification Model Formation

In the proposed ensembled method for benign and malignant sequences, 6 classifiers like DT technique, SVM algorithm, LR model, LDA model, Naive Bayes classifier and k-nearest neighbours' algorithm are combined in incremental order. Classifiers are ensembled to form five different models such as:

1. Model 1 - DT technique and SVM algorithm
2. Model 2 - DT technique, SVM algorithm and LR model
3. Model 3 - DT technique, SVM algorithm, LR model and LDA model
4. Model 4 - DT technique, SVM algorithm, LR model and LDA model and k-NN algorithm
5. Model 5 - DT technique, SVM algorithm, LR model, LDA model, k-NN algorithm and Naive Bayes classifier

3.3 Hard Voting Ensemble for Classification

Ensemble classifier or voting classifier is a metaclassifier for combining similar or different concept wise classifiers for disease diagnosis with plurality or majority voting. Hard voting and soft voting are two approaches to the plurality vote prediction for classification. Hard voting involves prediction of class with the most votes from models. Hence, predict the class label ŵ, according to Eq. (1), using majority voting of each classifier C_i such that

$$\hat{w} = \text{mode}\{C_1(z), C_2(z), \ldots\ldots, C_n(z)\} \qquad (1)$$

Hard voting is done for all 5 ensembles of classifiers and the classification accuracy is calculated. Performance of ensembled hard voting classifiers for breast cancer prediction was done with classification performance metrics.

3.4 Soft Voting Ensemble for Classification

Based on the **probabilities** of all the predictions made by various classifiers, soft voting classifier categorize input data. Prediction of the class is made by the highest probability averaged by each classifier. Therefore, predict the class labels based on the predicted probabilities q, according to Eq. (2) for classifier such that

$$\hat{w} = \arg\max_i \sum_{j=1}^{n} m_j q_{ij} \qquad (2)$$

where ŵ = class label and $m_j$is the weight that can be assigned to j^{th} classifier. The algorithm for the implementation of ensemble hard voting and soft voting classifiers for breast cancer prediction is shown below.

Algorithm for Implementation of Ensemble Hard & Soft Voting Classifiers

```
For i=2 to 6
    For j=1 to i
        x ←classify (model(j), training_set)
      Estimator ←x
    End
    #Hard score evaluation
    Hard _vote ←voting_classifier(estimator,
                                  hard)
     Fit (hard_vote (x_t,y_t)
Y_pre      predicate (x_v)
     Display (accuracy_score(y_v,y_pre))
     Display (confusion matrix(y_v,y_pre))
     Display (classification_report (y_v,y_pre))
    #Soft score evaluation
Soft_votevoting_classifier(estimator,
                                  soft)
     Fit (soft_vote (x_t,y_t)
Y_pre      predicate (x_v)
     Display (accuracy_score(y_v,y_pre))
     Display (confusion matrix(y_v,y_pre))
     Display (classification_report (y_v,y_pre))
End
```

3.5 Performance Evaluation Metrics

The efficiency of the ensembled classifiers is calculated by finding the performance metrics for both hard voting and soft voting classifiers. If both actual class value and predicted class value is 1, it is called true positive. If the value of both actual class and predicted class is 0, then it is true negative. False negative and false positive appear in the confusion matrix when the actual class and predicted class negate each other. Classification performance metrics are defined by the below Eq. (3) – Eq. (5):

$$\text{Precisions} = \text{True Positives}/\text{True Positives} + \text{False Positives} \quad (3)$$

$$\text{Recalls} = \text{True Positives}/\text{True Positives} + \text{False Negatives} \quad (4)$$

$$\text{F1\_Score} = (2 * \text{Precisions} * \text{Recalls})/(\text{Precisions} + \text{Recalls}) \quad (5)$$

Support = Count of real instances in the class for the specified dataset

4. Results and Discussion

Breast cancer prediction was done with genetic content of homo sapiens, BRCA1 and BRCA2 as classes 0,1 and 2, were retrieved as fasta data from NCBI database for machine learning purpose. Different ensembled models were formed and prediction was done with 1580 sequence instances. The proposed ensembled models were formed with 6 classifiers, DT technique, SVM algorithm, LR classifier, LDA classifier, NB classifier and k-NN algorithm. Five models were formed with different combinations of 6 classifiers.

For the classification of breast cancer, genetic sequences were categorized as trained and tested datasets with 80:20 ratio. Five ensembled models from 6 machine learning classifiers were concatenated with trained and tested datasets. The classification execution for disease prediction for 5-ensemble hard voting and soft voting classifiers are represented in confusion matrix which is shown in Table 1.

Table 1. Confusion matrix representation of ensemble classifiers

S.No	No of Classifiers ensembled	Classifiers ensembled	Confusion Matrix	
			Hard Voting	**Soft Voting**
1	2	DT & SVM	[38 0 0] [3 37 1] [0 8 47]	[37 0 1] [0 39 2] [0 5 50]
2	3	DT, SVM & LR	[38 0 0] [3 29 9] [0 3 52]	[38 0 0] [0 36 5] [0 3 52]
3	4	DT, SVM, LR & LDA	[38 0 0] [5 27 9] [0 4 51]	[38 0 0] [3 31 7] [0 3 52]
4	5	DT, SVM, LR, LDA & KNN	[38 0 0] [5 26 10] [0 3 52]	[38 0 0] [3 30 8] [0 3 52]
5	6	DT, SVM, LR, LDA, KNN & NB	[38 0 0] [5 30 6] [0 5 50]	[38 0 0] [3 34 4] [0 9 46]

The entire classes for classification are represented by a confusion matrix as 3 by 3 matrix. The class C0 is denoted by the first row and column, C1 by second and C2 by third respectively. The total testing instances are represented by confusion matrix row sum. First row sum denotes C0 testing instance, second row sum represent C1 testing instance and third denotes C3 testing data. Common instances are applied for all ensembled models. Precisely recognized testing data for the corresponding class

Ci, where i = 0, 1, 2 is represented by the confusion matrix diagonal values. Classification performance metrics was used to evaluate classification accuracy, in prediction of breast cancer for all 5 ensembled hard voting and soft voting classifiers which are represented in Table **2**

Table 2. System Generated Classification Report of Ensembled Models

S. No	No: of classifiers ensembled	Classifiers ensembled	Hard Voting				Soft Voting			
			Class	Precision	Recall	F1-score Support	Class	Precision	Recall	F1-score Support
1	2	DT & SVM	0	0.93	1.00	0.96 38	0	1.00	0.97	0.99 38
			1	0.82	0.90	0.86 41	1	0.89	0.95	0.92 41
			2	0.98	0.85	0.91 55	2	0.94	0.91	0.93 55
			Accuracy			0.91 134	Accuracy			0.94 134
			Macro avg	0.91	0.92	0.91 134	Macro avg	0.94	0.94	0.94 134
			Weighted avg	0.92	0.91	0.91 134	Weighted avg	0.94	0.94	0.94 134
2	3	DT, SVM & LR	Class	Precision	Recall	F1-score Support	Class	Precision	Recall	F1-score Support
			0	0.93	1.00	0.96 38	0	1.00	1.00	1.00 38
			1	0.91	0.71	0.79 41	1	0.92	0.88	0.90 41
			2	0.85	0.95	0.90 55	2	0.91	0.95	0.93 55
			Accuracy			0.89 134	Accuracy			0.94 134
			Macro avg	0.90	0.88	0.88 134	Macro avg	0.95	0.94	0.94 134
			Weighted avg	0.89	0.89	0.88 134	Weighted avg	0.94	0.94	0.94 134
3	4	DT, SVM, LR & LDA	Class	Precision	Recall	F1-score Support	Class	Precision	Recall	F1-score Support
			0	0.88	1.00	0.94 38	0	0.93	1.00	0.96 38
			1	0.87	0.66	0.75 41	1	0.91	0.76	0.83 41

			2 55	0.85	0.93	0.89	2 55	0.88	0.95	0.91
			Accuracy 134			0.87	Accuracy 134			0.90
			Macro avg 134	0.87	0.86	0.86	Macro avg 134	0.91	0.90	0.90
			Weighted avg 134	0.87	0.87	0.86	Weighted avg 134	0.90	0.90	0.90
4	5	DT, SVM, LR, LDA & KNN	Class Support	Precision	Recall	F1-score	Class score Support	Precision	Recall	F1-
			0 38	0.88	1.00	0.94	0 38	0.93	1.00	0.96
			1 41	0.90	0.63	0.74	1 41	0.91	0.73	0.81
			2 55	0.84	0.95	0.89	2 55	0.87	0.95	0.90
			Accuracy 134			0.87	Accuracy 134			0.90
			Macro avg 134	0.87	0.86	0.86	Macro avg 134	0.90	0.89	0.89
			Weighted avg 134	0.87	0.87	0.86	Weighted avg 134	0.90	0.90	0.89
5	6	DT, SVM, LR, LDA, KNN & NB	Class Support	Precision	Recall	F1-score	Class score Support	Precision	Recall	F1-
			0 38	0.88	1.00	0.94	0 38	0.93	1.00	0.96
			1 41	0.86	0.73	0.79	1 41	0.79	0.83	0.81
			2 55	0.89	0.91	0.90	2 55	0.92	0.84	0.88
			Accuracy 134			0.88	Accuracy 134			0.88
			Macro avg 134	0.88	0.88	0.88	Macro avg 134	0.88	0.89	0.88
			Weighted avg 134	0.88	0.88	0.88	Weighted avg 134	0.88	0.88	0.88

Table 2 represent the classification comparison of various ensembled models for class 0, class 1 and class 2 as normal, BRCA1 and BRCA2 datasets with the help of classification performance metrics. It is inferred from the report that the soft voting classifier for model 1(DT & SVM) and model 2(DT, SVM & LR) has achieved maximum value for precision, recall and F1-score, each of 94% respectively, as compared to other ensembled classifiers. Hard voting classifier for model 3(DT, SVM, LR & LDA) and model 4(DT, SVM, LR, LDA & k-NN) performed comparatively less on the given dataset with an accuracy of 86.57% and 86.56% respectively.

The accuracy rate of the classification is identified by calculating the proportion rate of classes properly recognized to the total size of testing data. The classification accuracy in prediction for all 5 ensembled hard voting and soft voting classifiers are depicted in Table 3.

Table 3. Ensembled Classifiers Classification Accuracy

S.No	No of Classifiers ensembled	Classifiers ensembled	Classification Accuracy	
			Hard Voting	Soft Voting
1	2	DT & SVM	91.04	94.03
2	3	DT, SVM & LR	88.80	94.03
3	4	DT, SVM, LR & LDA	86.57	90.30
4	5	DT, SVM, LR, LDA & KNN	86.56	89.55
5	6	DT, SVM, LR, LDA, KNN & NB	88.06	88.06

Comparing both hard voting and soft voting, the classification accuracy in prediction is more for soft voting classifiers because every individual classifier offers a probability value that a specific data point which goes to a specific target class. It was inferred that the classification accuracy in prediction, 94.03 for soft voting ensemble was maximum, when two classifiers (DT & SVM) as well as three classifiers (DT, SVM & LR) were ensembled. Even with hard voting ensemble also, the classification accuracy was maximum for both models, 91.04 and 88.80 respectively. The accuracy declined gradually when four (DT, SVM, LR and LDA), five (DT, SVM, LR, LDA and k-NN) and six (DT, SVM, LR, LDA, k-NN and NB) classifiers were ensembled. Certainly, the classification accuracy of every ensembled model exceeded accuracy of corresponding individual base classifiers. The comparative classification accuracy graph of all 5 ensembled hard voting and soft voting classifiers are depicted in Fig 2.

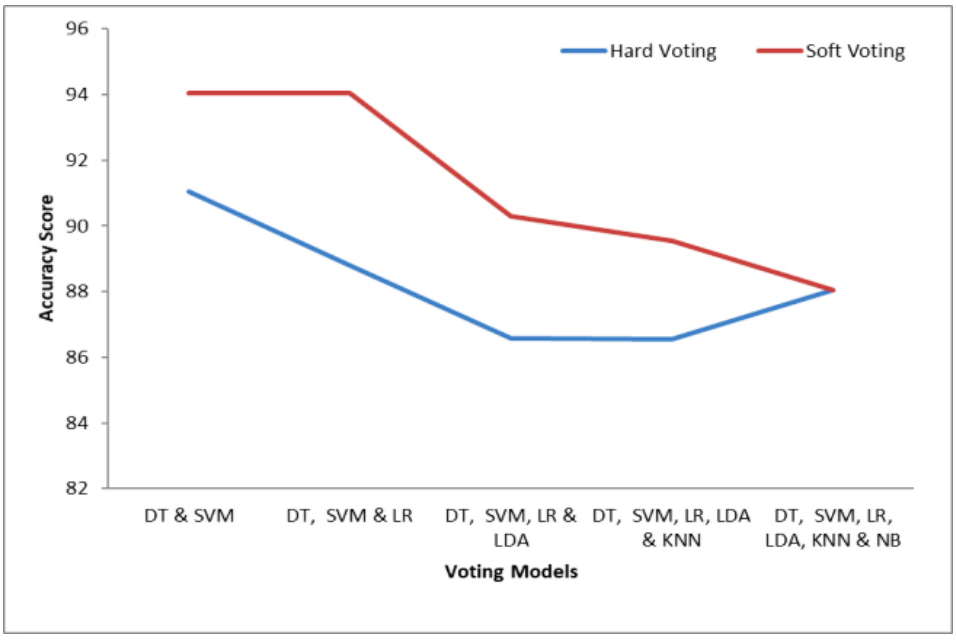

Fig 2. Comparative Classification Accuracy of Ensemble Hard Voting and Soft Voting Classifier

5. Conclusion

Breast cancer can have an enduring worst impression on lives and well-being of people. Prognosis of breast cancer is vital in spite of its complexity at an early stage. Machine learning methods such as ensemble voting classifiers enhance the precision of cancer prediction vulnerability, reappearance and impermanence. Sequences of normal human genome, BRCA1 and BRCA2 were extracted from NCBI as data instances. Breast cancer prediction was done using ensemble hard voting and soft voting classifiers which combined machine learning classifiers like DT technique, SVM algorithm, LR statistical model, Linear Discriminant analysis model, Naive Bayes classifier and k-nearest neighbours' algorithm. 5 ensembled models were formed with six machine learning models and performance was evaluated for both hard voting and soft voting classifiers. By calculating precision, recall, F1-score and support values, classification accuracy of ensemble hard voting and soft voting classifiers were evaluated. Maximum prediction accuracy of 94.03 was derived by soft voting classifier when two classifiers (DT & SVM) and three classifiers (DT, SVM & LR) were ensembled. The research can be expanded for earlier detection of contagious infections such as COVID, for which more SARS-CoV-2 virus characteristics may be extracted for classification and ensemble voting classifiers may be used for disease prediction.

Conflict of interest The author declares no conflict of interest

References

[1]. Abdar, M., &Makarenkov, V. (2019). CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. *Measurement*, *146*, 557-570.

[2]. Abdar, M., Zomorodi-Moghadam, M., Zhou, X., Gururajan, R., Tao, X., Barua, P. D., &Gururajan, R. (2020). A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recognition Letters*, *132*, 123-131.

[3]. Asri H, Mousannif H, Al Moatassime H, et al. (2016) Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science 83: 1064–1069.

[4]. Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, *2*, 117693510600200030.

[5]. Duijf, P. H., Nanayakkara, D., Nones, K., Srihari, S., Kalimutho, M., & Khanna, K. K. (2019). Mechanisms of genomic instability in breast cancer. *Trends in molecular medicine*, *25*(7), 595-611.

[6]. Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., ... & Ren, G. (2018). Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & diseases*, *5*(2), 77-106.

[7]. Huang Q, Chen Y and Liu L (2019) On combining biclustering mining and AdaBoost for breast tumour classification. IEEE Transactions on Knowledge and Data Engineering 32(4): 728–738.

[8]. Balusamy R., Kumaravel P., Renganathan N.G "Dielectric and electrical properties of lead zirconate titanate" *Der Pharma Chemica* (2015).

[9]. Jayapandian N., Rahman A.M.J.M.Z., Poornima U., Padmavathy P." Efficient online solar energy monitoring and electricity sharing in home using cloud system" *IC-GET 2015 - Proceedings of 2015 Online International Conference on Green Engineering and Technologies* (2016).

[10]. Li Y and Chen Z (2018) Performance evaluation of machine learning methods for breast cancer prediction. Applied and Computational Mathematics 7(4): 212–216.

[11]. Lu H, Wang H and Yoon SW (2019) A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis. Expert Systems with Applications 116: 340–350.

[12]. Mittal, M., Arora, M., Pandey, T., & Goyal, L. M. (2020). Image segmentation using deep learning techniques in medical images. In *Advancement of machine intelligence in interactive medical image analysis* (pp. 41-63). Springer, Singapore.

[13]. Mittal, M., Kaur, I., Pandey, S. C., Verma, A., & Goyal, L. M. (2019). Opinion mining

for the tweets in healthcare sector using fuzzy association rule. *MH*, *50*, S2.

[14]. Pe´rez-Ortiz M, Gutierrez PA and Herva´s-Martı´nez C (2014) Graph-based approaches for over-sampling in the context of ordinal regression. IEEE Transactions on Knowledge and Data Engineering 27(5): 1233–1245

[15]. Polley MYC, Freidlin B, Korn EL, et al. (2013) Statistical and practical considerations for clinical evaluation of predictive biomarkers. Journal of the National Cancer Institute 105(22): 1677–1683

[16]. Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1584-1589). IEEE.

[17]. Sun YS, Zhao Z, Yang ZN, et al. (2017) Risk factors and preventions of breast cancer. International Journal of Biological Sciences 13(11): 1387–1397

[18]. Ture M, Kurt I, Kurum AT, et al. (2005) Comparing classification techniques for predicting essential hypertension. Expert Systems with Applications 29(3): 583–588.

[19]. Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research*, *267*(2), 687-699.

[20]. Mathew O.C., Rahman A.M.J.Z." A novel energy optimization mechanism for medical data transmission using honeycomb routing" *Journal of Medical Imaging and Health Informatics* (2016).

[21]. Warner B and Misra M (1996) Understanding neural networks as statistical tools. The American Statistician 50(4): 284–293.

[22]. Zhao M, Tang Y, Kim H, et al. (2018) Machine learning with k-means dimensional reduction for predicting survival outcomes in patients with breast cancer. Cancer Informatics 17: 1176935118810215.

[23]. Li, Y., & Luo, Y. (2020). Performance-weighted-voting model: an ensemble machine learning method for cancer type classification using whole-exome sequencing mutation. *Quantitative Biology*, *8*(4), 347-358.

[24]. Mirsadeghi, L., Hosseini, R. H., Banaei-Moghaddam, A. M., &Kavousi, K. (2021). EARN: an ensemble machine learning algorithm to predict driver genes in metastatic breast cancer. *BMC Medical Genomics*, *14*(1), 1-19.