EJMO

DOI: 10.14744/ejmo.2024.52521
EJMO 2024;8(2):152–164

Research Article

Correlation-Based Comparative Machine Learning Analysis for the Classification of Metastatic Breast Cancer Using Blood Profile

Mahendran Botlagunta,[1] Mdhavidevi Botlagunta,[2] Manjula Devarakonda Venkata,[3] Christina Kanakapudi,[4] Zeba Khan[5]

[1]School of Biosciences, Engineering and Technology, VIT Bhopal University, Bhopal, India
[2]Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad, India.
[3]Department of CSE, Pragati Engineering College(A), Andhra Pradesh
[4]Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad, India.
[5]School of Biosciences, Engineering and Technology, VIT Bhopal University, Bhopal, India

Abstract

Objectives: Histopathological and mammography image-guided diagnosis is a common practice for the detection of cancer grade, which is often associated with poor survival outcomes in breast cancer patients. A deep learning (DL) based clinical decision support system was developed for histologic grading of breast cancer, which often requires invasive procedures or expensive imaging equipment. Our study aimed to establish a machine learning model based on simple blood profile data.

Methods: The dataset consists of blood profiles of 1250 breast cancer patients and 259 normal subjects. Statistical methods were used to select the relevant feature for machine learning model development. Selected features were fitted into various Machine Learning classifiers to predict breast cancer with highest accuracy.

Results: Correlation-based feature selection revealed that blood profile ratio counterparts were statistically significant ($p<0.05$) and were used for the classification of metastatic breast cancer patients as compared to normal subjects.

Conclusion: The ensemble stacking classifier outperformed other algorithms with an accuracy, sensitivity, specificity, and F1 score with values of 96%, 98%, 98% and 98% respectively and it can be used for non-invasive laboratory-based diagnosis for early prediction of breast cancer.

Keywords: Ensemble stacking classifier, Breast Cancer, blood profile, Correlation, Machine Learning, TukeyHSD

Cite This Article: *Botlagunta M, Kanakapudi C, Polagangu M, Devarakonda Venkata M. Correlation-Based Comparative Machine Learning Analysis for the Classification of Metastatic Breast Cancer Using Blood Profile. EJMO 2024;8(2):152–164.*

Mammography is a well-known method for the diagnosis of breast cancer.[1] Following diagnosis, tissue is collected by fine-needle aspiration technique[2] and is subjected to histopathological analysis to identify the percentage of normal ducts (tubule formation), number of dividing cells (mitotic rate) and the appearance of the nucleus (nuclear pleomorphism) in the tumor specimen. According to Nottingham grading system greater than 75%, between 10% and 75% and less than 10% normal breast (milk) ducts in tumor tissue is used for tubule formation. Similarly, dividing cells, size and shape of the nucleus in tumor cells is used to score mitotic rate and nuclear pleomorphism.[3] Based on all three features pathologist tentatively grade the cancer (Table 1).

Address for correspondence: Mahendran Botlagunta, Ph.D. School of Biosciences Engineering and Technology, VIT Bhopal University, Bhopal, India
Phone: +917382145827 **E-mail:** bmnchowdary@gmail.com

Submitted Date: December 13, 2023 **Revision Date:** March 03, 2024 **Accepted Date:** March 13, 2023 **Available Online Date:** July 10, 2024

Table 1. Pathological Prognostic factors used for grading Breast Cancer

Score	Tubule formation	Mitotic rate	Nuclear pleomorphism	Total Feature Score	Grade	Appearance of Cells
1	>75%	<10	Small and uniform nuclei	3-5	1	Well differentiated (appear normal, growing slowly, not aggressive)
2	10 -75%	10-19	Intermediate variations in size and shape	6-7	2	Moderately differentiated (semi-normal, growing moderately quickly)
3	<10%	>19	Marked variations	8-9	3	Poorly differentiated (abnormal, growing quickly, aggressive)

Uniform cells with small nuclei similar to the size of normal breast epithelial cells (minimal nuclear pleomorphism) are considered grade 1, Cells larger than normal breast epithelial cells with hyperchromatic nuclei and prominent nucleoli, moderately variable shape, are considered grade 2 and glandular pattern of tumor cells with moderate amount of cytoplasm and more number of mitotic cells is classified as grade 3. On the other hand, invasion of tumor cells into Lympho vascular, Perineural, is absent in grade 1pateints, Whereas Invasive ductal carcinoma with secondary fibrotic and necrotic inflammatory and foreign body giant cells, lymphocytic infiltration atrophic ducts, epithelial hyperplasia, lymphoid aggregates, hemosiderin-laden pigments around the nucleus, proliferating blood vessels, and dense fibrosis around the tumor is seen in majority of the grade 2 patients.[4-8] Hemosiderin-laden pigments are the end product of hemorrhage, which is due to the released hemoglobin into the extracellular space and dead red blood cells.[9] In addition, infiltrated tumor cells surrounding tissues, Perinodal spread and Micro calcification are seen in grade 3 patients.[10,11] Breast cancerous cells commonly spread to the lungs, liver, bone, and brain. Brain parenchyma with a cellular lesion comprised of blood elements, foamy histiocytes, a few lymphocytes, and neutrophils is observed in the brain of breast cancer metastatic patients.[12] Whereas tumor cells in tubules and nests lie in pools of extracellular mucin and a moderate amount of cytoplasm is seen in lung breast cancer metastatic patients.[13,14] Overall, it suggests that reports generated by pathologists contain a substantial amount of useful information. Therefore, extracting useful information from text documents and applying those for cancer prediction using Machine-learning algorithms is a daunting task.[15,16]

Machine learning has emerged as a potential technique for managing high-dimensional data for the development of clinical decision support systems for breast cancer.[1,17,18] In our previous work, NLP-based techniques were employed to extract clinical data from breast cancer patients. Extracted features in conjunction with machine learning techniques were used for the Classification of breast cancer. Among the various ML models tested with our dataset, the decision tree showed the highest accuracy 83% with an AUC of 0.87.[19] Low accuracy may be due to less number of features or attributes, and it can be improved by adding more data, or better feature engineering. Feature engineering helps to select the most informative and relevant features in our dataset, which enhances model performance. Various techniques such as correlation coefficients from Pearson and Spearman, Euclidean distance Support vector machines, multi-layer perceptrons, k-nearest neighbors, and structure-adaptive self-organizing maps are some of the methods that can be used for feature selection and classification. In this present work, we used Pearson's correlation-based feature selection approach for the development of a Machine Learning model for the classification of breast cancer metastasis using simple non-invasive laboratory tests.

Breast Cancer is one of the major leading causes of cancer death in women, it is mainly due to ineffective cancer prediction systems. Mammography, PET, and MRI are generally used to diagnose breast cancer spread. On the other hand, tissue specimens are subjected to histopathological analysis to determine the cancer grade. However, these methods are unable to predict the circulatory tumor cells (CTCs) in blood. CTCs in the blood either negatively or positively impact complete blood count (CBC), which is frequently used to monitor therapy responders and non-responders. To determine the hostile cellular environment in the blood and to develop early cancer diagnosis, we developed a machine-learning model for classifying breast cancer using blood profile data. We anticipate that the deployment of our model in the hospital help physicians to constantly monitor patients' health between two visits using blood parameters. Moreover, our model can also be implemented by designing a suitable Patient-accessible Website to monitor their health on a mobile platform.

Methods

Data Processing

The dataset contains both continuous and categorical variables such as medical record number and continuous variables (clinical parameters). The dataset used in this study contains 1509 instances and two class categories, namely Normal and Cancer. The attributes used in this study for machine learning are Haemoglobin, Red Cell Count, Neutrophils, Lymphocyte, Monocyte, Haemoglobin/Red Cell Count, Red Cell Count/Haemoglobin, Neutrophils/Haemoglobin, Neutrophils/Monocyte, Monocyte/Red Cell Count, Monocyte/ Neutrophils and Monocyte/Lymphocyte. This data is further subjected to cleaning, which involves removing duplicates, removing data with a total cell count of less than 100. Cleaned data was further subjected to statistical techniques to identify patterns or relationships in the data. Python programming language is used for data analysis and visualization. Numpy is a general-purpose array processing package, which is used for match operations and handling multidimensional data. Matplotlib libraries are used for data visualization and SciPy library is used for scientific computing and data analysis. Scikit Learn is used to implement Random Forest (RF), Decision Tree (DT), Support vector Machine (SVM), K-Nearest Neighbor (K-NN), Naïve Bayes (NB), and Ensemble Stacking Machine Learning (ML) algorithms.

Feature Selection

The Pearson correlation feature selection method is employed to identify the most useful correlated features. The degree of correlation is expressed by a Pearson correlation coefficient (PCC), which ranges from -1 to 1. Using equation (1), the PCC is computed.

$$\rho = \frac{{}^{\sigma}ML}{{}^{\sigma}M^{*\sigma}L} \qquad (1)$$

Where,

ρ - denotes the Pearson correlation coefficient

${}^{\sigma}ML$ - denotes the covariance of M and L variables

${}^{\sigma}M$ - denotes the standard deviation of M

${}^{\sigma}L$- denotes the standard deviation of L

Implementation of Machine Learning Models

In addition to our previously validated models in this study, we implemented Naive Bayes (NB), and Stack ensemble model classifier to identify the proper ML algorithm with good accuracy, F1, precision, recall, and Roc for the target by the testing data. The naive Bayes classification method works based on Bayes' Theorem, which assumes that all features or attributes are conditionally independent. This method frequently produces decent results in a variety of applications despite its naïve assumptions. For the stacking classifier, we created multilayer stacking consisting of RF, SVM, and KNN classifiers as base models in the first layer of stacking. The identical training data is used to train all three models. Results from the base models are used to train the meta-classifier model. For the Meta-classifier, the Logistic Regression classifier is used. Hyper parameters used in this work are shown in Table 2.

Table 2. Type of hyper parameters used in the ML algorithms

S.No	Name of the Classifier	Hyper Parameter
1	Random Forest Classifier	max_depth=64, random state=43
2	Decision Tree Classifier	max_depth=10, random state=34
3	Support Vector Machine	probability = True, kemel = 'linear', C = 1
4	Kneighbors Classifier	neighbors=5, metric='minkowski'
5	GaussianNB	Default
6	Stacking Classifier	estimators=base_models, final estimator=lr

Performance Testing of Machine Learning Models

The performance and effectiveness of the proposed classification models were evaluated as described earlier [19]. In addition, the statistical parameters for regression, such as R, R-squared, and standard error, are used to evaluate the model's performance. The correlation coefficient R shows how strongly the variables are related to one another. Even if the additional predictor is not significant, by including it in the model, the R squared value always increases. The regression model's performance is measured by its standard error. The model becomes more accurate as the error decreases.

Statistical Analysis

R statistical package is used for Exploratory Data Analysis (EDA) to identify the optimal set of features. ANOVA and post hoc tests were performed to identify, the p-value associated with the F-statistic and the mean value difference between each group using the TukeyHSD () function.

Outline of the Proposed Model

The algorithm for predicting breast cancer metastasis

Input: Hematological Features

Procedure:

Step 1: Start

Step 2: Statistical analysis

- ANOVA analysis
- TukeyHSD

Step 3: Feature selection process

- Pearson correlation coefficient

Step 4: Divide the dataset into training and testing

Step 5: Build the regression model

- RFR
- DT
- SVM
- K-NN
- NB
- Ensemble Stacking

Step 6: Calculate the evaluation metrics

- Accuracy, Precision, Recall, F1 score

Step 7: Calculate the statistical measure

- R, R-squared, and Standard Error

Step 9: Finding the best ML model

Step 10: End

Output:

- Suggest the best ML model to predict breast cancer metastasis

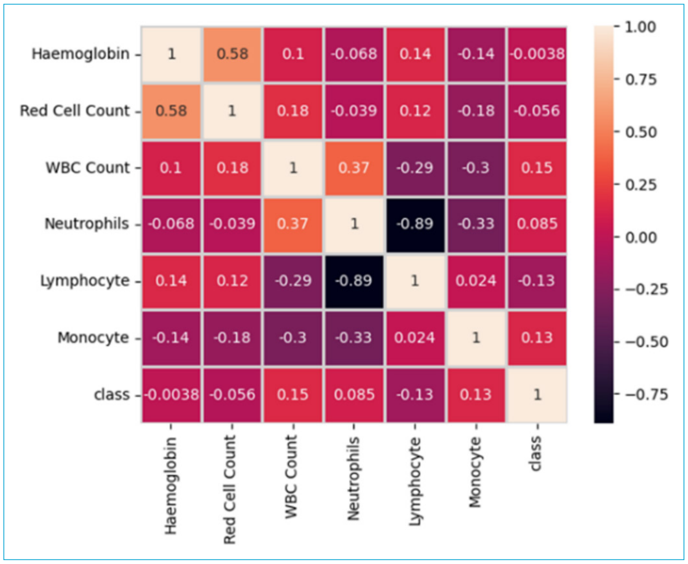


Figure 1. Correlation values between hematological features. Numbers displayed on the right bar denotes most and least correlated features.

Results

Multiple Comparison and Correlation Analysis of Hematological Features

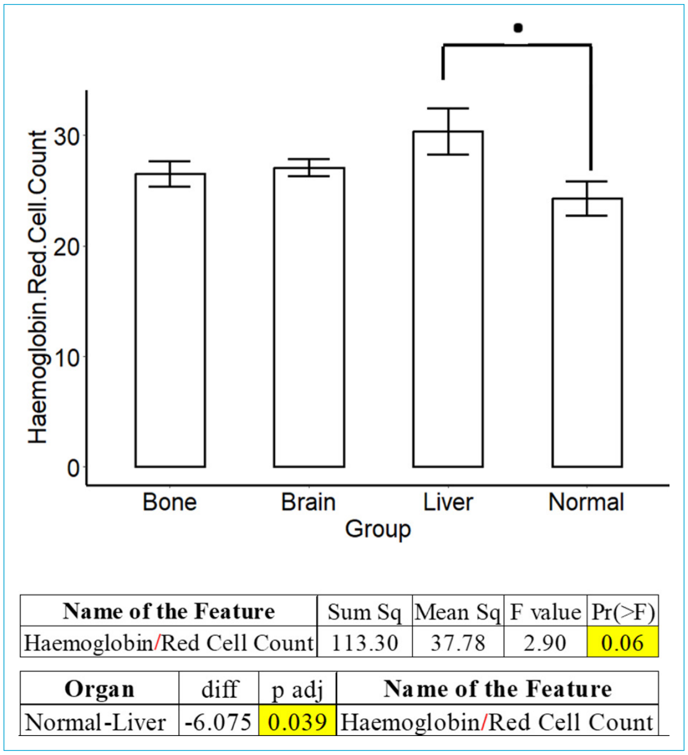
One of the simplest and best-used techniques for feature selection is the Pearson correlation. The correlation plot for the hematological features is shown in Figure 1. According to the correlation analysis, hemoglobin showed a moderate correlation with Red cell count. WBC count showed a fair correlation with Neutrophils. On the other hand, hemoglobin showed a slight correlation with WBC and Lymphocyte counts. Red cell count showed a slight correlation with WBC count and Lymphocyte. Similarly, Lymphocytes showed a slight correlation with hemoglobin and Red Cell Count. Whereas Monocytes showed no correlation with any feature. The correlation values for the moderate, fair, and slight values are shown in the heatmap. To identify the statistically significant feature, we performed a One-way analysis of variance (ANOVA) between normal and cancer subjects. Results showed the Pr(F) value associated with Red Cell Count, WBC, Neutrophils, Lymphocyte, and Monocyte is less than 0.001 (Table 3) and are statistically significant. Overall, it suggests the hematological features positively correlated with each other and they can be used for the classification of normal and cancer samples.

Identification of Statistically Significant Ratio Counterparts

Metastasis is the spread of cancer in different organs from the site of origin. Previously, we have developed a text mining method for the extraction of important attributes associated with the progression of breast cancer metastasis. Using a similar approach, we retrieved the blood profile data for the brain (n=6), bone (n=6), and lung (n=6) metastatic breast cancer patients, and they were matched with the same number of normal subjects for data analysis. ANOVA and Tukey HSD analysis revealed none of the hematological features are statistically significant to distinguish the organ-specific spread of cancer. Based on the correlation studies and to explore further we created ratios

Table 3. ANOVA analysis between normal and cancer subjects for hematological features

S.No	Name of Feature	Mean Sq	F	Pr (>F)
1	Haemoglobin	6	0.02	0.882
2	Red Cell Count	1.4	4.71	0.030
3	WBC Count	243	33.17	0.000
4	Neutrophils	1050	11.03	0.001
5	Lymphocyte	2026	27.46	0.000
6	Monocyte	236	26.12	0.000



Name of the Feature	Sum Sq	Mean Sq	F value	Pr(>F)
Haemoglobin/Red Cell Count	113.30	37.78	2.90	0.06

Organ	diff	p adj	Name of the Feature
Normal-Liver	-6.075	0.039	Haemoglobin/Red Cell Count

Figure 2. Identification of statistically significant Ratio counterparts-Upper histogram represent a mean difference of Haemoglobin to Red cell count ratio between the groups. Bottom panel represents statistical ANOVA and Tukey HSD values. Dot (.) indicates significant difference between the groups.

for each attribute, the results data file consists of 31 attributes. Statistical analysis showed, no mean difference between the independent features however, the hemoglobin / Red cell count ratio showed close to the significance and Tukey posthoc analysis further confirmed the same with an average difference of -6.075 ($p<0.05$) between the normal to the liver (Fig. 2). Next, we analyzed a similar study to identify the statistically significant attributes, which can be used to differentiate normal, grade 2, and grade 3 cancer subjects using a separate data file. Statistical analysis showed that Haemoglobin and Monocyte counts, Haemoglobin/Red Cell Count, Red Cell Count/Haemoglobin, Neutrophils/Haemoglobin, Neutrophils/Monocyte, Monocyte/Red Cell Count, Monocyte/Neutrophils, Monocyte/Lymphocyte ratios were statistically-significant ($p<0.05$) between normal and grade 2 and 3 breast cancer patients (Table 4a). A Tukey posthoc test revealed that Haemoglobin content and Monocyte count showed significant pairwise differences between normal and grade 2 and normal and grade 3 with an average difference of -9.83 and -1.57 ($p<0.05$) respectively. On the other hand, Haemoglobin/Red Cell Count, Red Cell Count/Haemoglobin, Neutrophils/Monocyte, Monocyte/Red Cell Count, Monocyte/Neutrophils, and Monocyte/Lymphocyte ratios showed significant pairwise differences between normal and grade 3 with a p value less than 0.05 (Table 4b). Overall it suggests that independent hematological Features in conjunction with their respective ratios counterparts can be used to distinguish in-depth classification of breast cancer metastasis using blood profile data.

Distribution of Hematological Ratio Attributes Between Normal and Grade 3 Breast Cancer Patients

Next, we analysed the distribution of the ratio-associated hematological features between normal and grade 3 breast cancer subjects, because many of the ratio-associated features showed a significant difference between normal and grade 3 breast cancer patients. For this analysis, we used the blood profiles of 62 grade 3 breast cancer patients and 62 normal subjects. Results showed that the mean value difference for the Haemoglobin, monocyte, and Haemoglobin/Red Cell count is higher in grade 3 breast cancer patients as compared to normal individuals. On the other hand, Neutrophil/Monocyte ratio and Neutrophil/WBC count ratio values are significantly lower in cancer subjects as compared to normal subjects (Fig. 3). The overall mean difference in hemoglobin content in grade 3 breast cancer patients is 117.89 g/L as compared to 110.95 g/L in normal subjects with a p-value less than 0.005. Similarly, for monocytes, 7.56% as compared to 5.97%, the Haemoglobin to Red cell count ratio is 26.67 as compared to 24.99 and the neutrophils to monocyte ratio is 9.75 as compared to 11.56 in normal subjects. Except Neutrophil to WBC count ratio p-value for all the ratios of hematological features is significant. Overall, it suggests that features associated with hematological ratio attributes differentiate normal and grade 3 patients and these ratio parameters can be used as a prognostic indicator for grading breast cancer patients.

Correlation-Based Feature Selection for Machine Learning

Next, we performed a correlation-based feature selection analysis to identify the redundant and unnecessary features. According to the kappa value interpretation by[20] monocytes revealed a substantial correlation range (0.61-0.81) with monocyte/neutrophils and monocyte/Lymphocyte (0.87 and 0.66) respectively. Similarly, neutrophils also showed a substantial correlation with neutrophils/hemoglobin (0.69). Neutrophils and hemoglobin showed a moderate correlation range (0.41-0.60) with neutrophils/monocyte (0.53) and hemoglobin/red cell count (0.49) and red cell count alone (0.58). Similarly, a fair correlation range (0.21-0.40) is observed among many features such as Lymphocyte with monocyte/neutrophils (0.31), neutrophils with monocyte/lymphocyte (0.27), WBC counts with neutrophils/monocyte (0.32), neutrophils/hemoglobin (0.22)

Table 4a. ANOVA analysis for hematological features and their ratio counterparts. Yellow shade indicates statistically significant (p<0.05)

S.No	Feature	Sum Sq	Mean Sq	F value	Pr(>F)
1	Haemoglobin	3207.0	1603.70	4.592	0.011
2	Red.Cell.Count	1.440	0.719	2.063	0.130
3	WBC.Count	13.800	6.888	0.659	0.519
4	Neutrophils	222.000	110.800	0.790	0.456
5	Lymphocyte	27.000	13.430	0.162	0.851
6	Monocyte	81.500	40.740	5.169	0.007
7	Haemoglobin/Red Cell Count	91.500	45.770	3.529	0.031
8	Haemoglobin/WBC Count	171.000	85.250	1.309	0.273
9	Haemoglobin/Neutrophils	5.900	2.950	2.226	0.111
10	Haemoglobin/Lymphocyte	19.200	9.580	0.948	0.390
11	Haemoglobin/Monocyte	25.000	12.730	0.189	0.828
12	Red Cell Count/Haemoglobin	0.000	0.000	3.588	0.030
13	Red Cell Count/WBC Count	0.241	0.121	1.769	0.173
14	Red Cell Count/Neutrophils	0.005	0.002	1.570	0.211
15	Red Cell Count/Lymphocyte	0.015	0.007	0.478	0.621
16	Red Cell Count/Monocyte	0.190	0.095	1.000	0.370
17	WBC/Haemoglobin	0.002	0.001	0.845	0.431
18	WBC/Red Cell Count	0.110	0.054	0.092	0.912
19	WBC/Neutrophils	0.009	0.004	1.328	0.267
20	WBC/Lymphocyte	0.240	0.122	0.229	0.795
21	WBC/Monocyte	0.900	0.450	0.537	0.585
22	Neutrophils/Haemoglobin	0.192	0.096	3.251	0.041
23	Neutrophils/Red Cell Count	21.200	10.610	0.870	0.421
24	Neutrophils/WBC Count	66.500	33.260	2.639	0.074
25	Neutrophils/Lymphocyte	0.600	0.277	0.041	0.960
26	Neutrophils/Monocyte	117.000	58.600	3.318	0.038
27	Monocyte/Haemoglobin	0.003	0.002	1.822	0.165
28	Monocyte/Red Cell Count	4.280	2.141	4.143	0.017
29	Monocyte/WBC Count	2.810	1.405	2.049	0.132
30	Monocyte/Neutrophils	0.065	0.032	4.645	0.011
31	Monocyte/Lymphocyte	0.292	0.146	3.248	0.041

Table 4b. Post-hoc analysis between the grades of breast cancer subjects as compared to normal subjects. Yellow shade indicates statistically significant (p<0.05)

Group	Grade 3-Grade 2		Normal-Grade 2		Normal-Grade 3	
Feature	diff	p adj	diff	p adj	diff	p adj
Haemoglobin	-2.695	0.700	-9.839	0.011	-7.144	0.085
Monocyte	0.459	0.632	-1.113	0.073	-1.572	0.006
Haemoglobin/Red Cell Count	0.48	0.74	-1.18	0.16	-1.66	0.03
Red Cell Count/Haemoglobin	-0.001	0.58	0.002	0.23	0.003	0.02
Neutrophils/Haemoglobin	0.00	1.00	0.07	0.07	0.07	0.07
Neutrophils/Monocyte	-0.19	0.96	1.58	0.10	1.77	0.05
Monocyte/Red Cell Count	0.17	0.40	-0.20	0.26	-0.37	0.01
Monocyte/Neutrophils	0.02	0.25	-0.02	0.32	-0.05	0.01
Monocyte/Lymphocyte	0.03	0.67	-0.06	0.23	-0.10	0.03

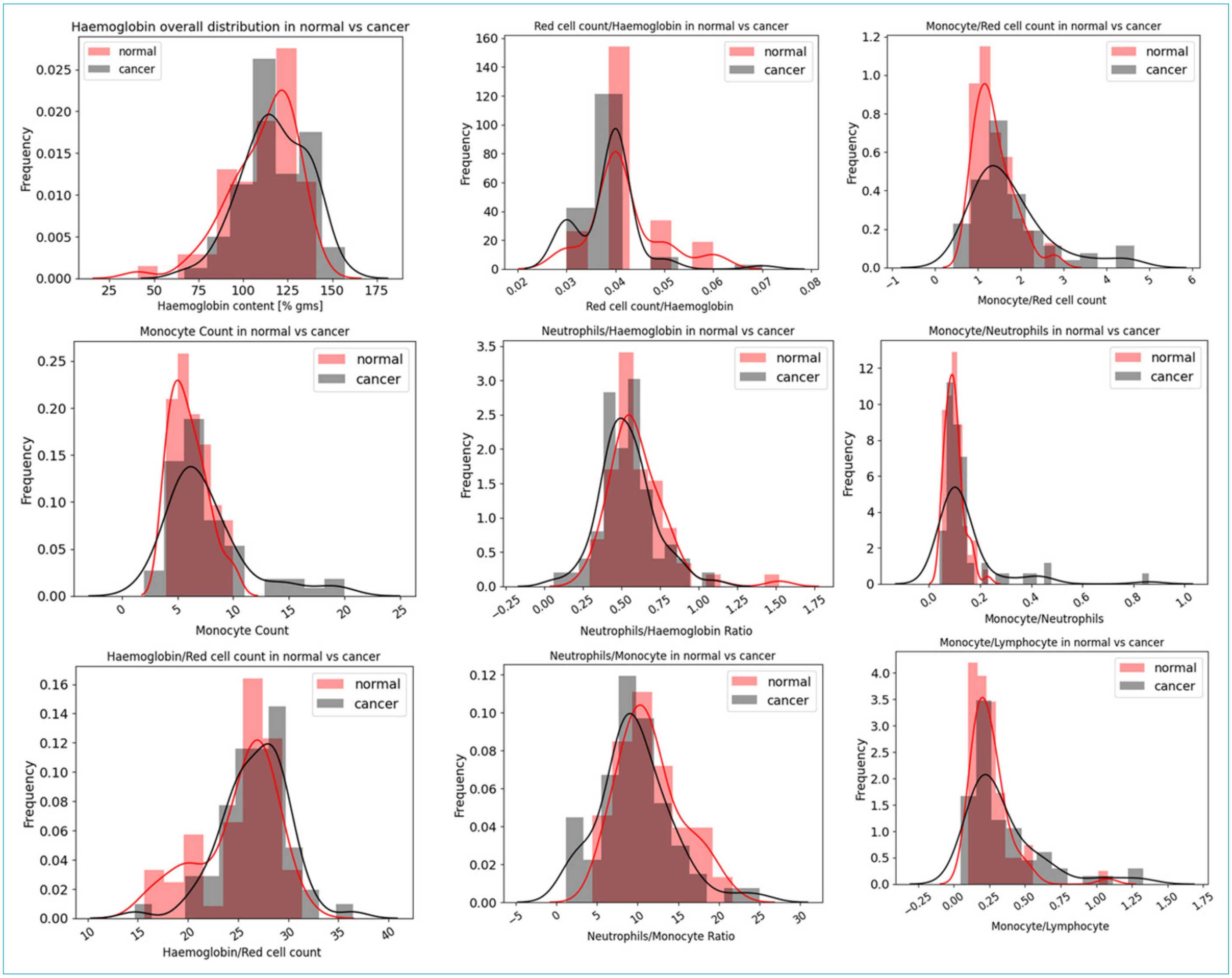


Figure 3. Overall distribution of hematology and the selected ratio attributes between normal and grade 3 breast cancer subjects. Here, frequency on the Y-axis shows how frequently a specific value appeared in both normal and cancer participants (number of repetitions/total observations).

and with neutrophils (0.37). Red cell count showed a correlation with red cell count/hemoglobin (0.31) (Fig. 4). In addition, many features show a slight correlation with each other. Overall, it suggests that independent hematological features are well correlating with their ratio counterparts and they can be used to increase the overall accuracy and performance of the machine learning.

Machine Learning Model Development

The initial data set is subjected to K-Means clustering to cluster the dataset based on similarities in feature space.[19] As a result, the initial dataset was reduced to a total of 893 instances (782:111 cancer: normal). Based on the selected features, the dataset is divided into four categories. Category 1) hematological features alone, 2) Selected ratio attributes of hematological features such as Haemoglobin/Red Cell Count, Red Cell Count/Haemoglobin, Neutrophils/Haemoglobin, Neutrophils/Monocyte, Monocyte/Red Cell Count, Monocyte/ Neutrophils and Monocyte/Lymphocyte with a same number of instances and 3) Category 1+2 and 4) All features (Category 1 + all ratio attributes). Every category consists of two classes labeled "1" (cancer) and "0" (normal). To choose the best ML model, our datasets are processed using 6 ML models. For all the models data is split into 70:30 ratios, with 70% of the data used for training and 30% for testing. Performance indicators of all ML models such as Accuracy, Recall, Precision, and F1 score were represented in Table 5.

Comparative ML models revealed that RF, SVM, and Ensemble models showed 93% accuracy for the category-1 data-

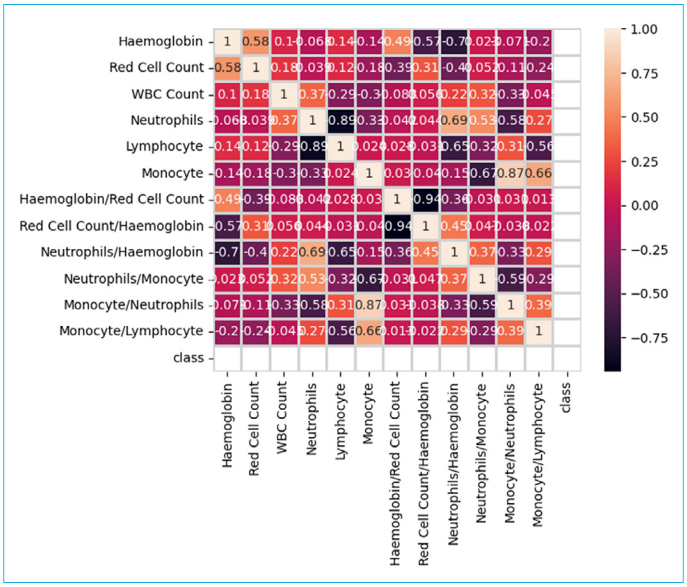


Figure 4. Correlation values between hematological and the selected ratio counterparts. Numbers displayed on the right bar denotes Most and Least Correlated features.

set, and RF and Ensemble models showed greater than 93 % accuracy for category-3 and category-4 datasets respectively. Whereas the category-2 dataset is unable to classify the model with an accuracy greater than 93%. Among all models and all the datasets, ensemble classier is a common model and it showed the highest accuracy (97%) for the category-3 dataset using default hyperparameters. Hyperparameter tuning (HPT) is performed to increase the model accuracy of the models. Following tuning (RF, SVM, and Ensemble), (DT, SVM, Ensemble), (RF, DT, SVM, and Ensemble) and (RF, KNN, and Ensemble) models classified the breast cancer with an accuracy greater than 93 % using category-1, 2, 3 and 4 datasets respectively. The accuracy of the DT before and after hyperparameter tuning is increased from 92 to 97% using the category-3 dataset. Overall, it suggests that Ensemble (stacking) and DT with a maximum depth of 10 and random state 34 showed better accuracy using the category-3 dataset. To confirm the proposed models with specific parameters can be used to predict breast cancer with higher accuracy 10-fold cross-validation (CV) is per-

Table 5. Comparative analysis of machine learning models with and without hyper parameter tuning.

	ML models (Hyper Parameter Tuning)						**ML models (Default Parameters)**					
	RF	**DT**	**SVM**	**KNN**	**NB**	**Ensemble**	**RF**	**DT**	**SVM**	**KNN**	**NB**	**Ensemble**
Category-1												
Accuracy	0.96	0.94	0.96	0.94	0.94	0.96	0.93	0.91	0.93	0.91	0.91	0.93
Recall	0.98	0.97	0.98	0.97	0.97	0.98	0.99	0.98	0.99	0.98	0.97	0.99
Precision	0.96	0.96	0.96	0.95	0.95	0.97	0.94	0.93	0.93	0.92	0.93	0.93
F1 Score	0.99	0.97	0.99	0.99	0.99	0.99	0.96	0.95	0.96	0.95	0.95	0.96
AUC	0.92	0.86	0.89	0.86	0.88	0.88	0.94	0.71	0.94	0.83	0.88	0.88
CV (10)	93.98	92.02	92.44	91.74	91.46	92.58	94.23	91.20	92.31	91.36	92.00	92.30
Category-2												
Accuracy	0.93	0.94	0.94	0.93	0.87	0.94	0.89	0.87	0.91	0.91	0.89	0.92
Recall	0.96	0.97	0.97	0.96	0.93	0.97	0.97	0.95	0.99	0.99	0.95	0.99
Precision	0.94	0.94	0.94	0.95	0.94	0.94	0.90	0.91	0.91	0.92	0.93	0.92
F1 Score	0.98	1.00	0.99	0.98	0.91	0.99	0.94	0.93	0.95	0.95	0.94	0.95
AUC	0.84	0.78	0.85	0.80	0.84	0.84	0.97	0.84	0.93	0.92	0.92	0.92
CV (10)	90.20	89.50	90.61	90.89	90.07	91.46	91.04	86.41	91.84	91.84	89.60	92.16
Category-3												
Accuracy	0.96	0.97	0.95	0.93	0.91	0.97	0.96	0.92	0.94	0.94	0.91	0.97
Recall	0.98	0.98	0.97	0.96	0.95	0.98	0.98	0.94	0.99	0.99	0.94	0.99
Precision	0.98	0.98	0.96	0.94	0.96	0.97	0.98	0.97	0.95	0.95	0.96	0.97
F1 Score	0.98	0.99	0.99	0.99	0.94	0.99	0.98	0.96	0.97	0.97	0.95	0.98
AUC	0.95	0.89	0.93	0.91	0.82	0.82	0.95	0.81	0.96	0.91	0.82	0.82
CV (10)	93.99	93.02	93.43	91.75	90.63	92.59	93.85	90.77	92.31	91.61	90.63	92.59
Category-4												
Accuracy	0.94	0.90	0.93	0.94	0.89	0.94	0.95	0.93	0.93	0.92	0.85	0.94
Recall	0.98	0.94	0.98	1.00	0.93	0.99	0.99	0.96	1.00	0.99	0.87	0.99
Precision	0.96	0.95	0.94	0.93	0.95	0.94	0.96	0.96	0.93	0.93	0.96	0.95
F1 Score	0.97	0.95	0.96	0.97	0.94	0.97	0.97	0.96	0.96	0.96	0.91	0.97
AUC	0.93	0.77	0.86	0.84	0.87	0.87	0.86	0.79	0.91	0.78	0.85	0.85
CV (10)	94.57	92.02	92.97	91.04	88.33	93.29	93.92	89.76	92.00	89.76	86.88	93.28

formed. Results showed that the accuracy of the CV score for Ensemble and DT classifiers for the category-3 dataset was found to be greater than 90%. On the other hand, the Random forest model with a maximum depth of 64 and random state of 43 showed 94% accuracy for the dataset with all features (category-4) and its cross-validation is 94% with an ROC/AUC of 0.93 (Table 5).

In addition, the effectiveness of the model is assessed using the statistical parameters for regression, such as R, R-squared, and standard error. The correlation coefficient R shows how strongly the variables are related to one another. All 6 ML algorithms are applied to training and testing data to calculate all of these measures (Table 6). The category-1 the Ensemble (stacking) algorithm has a very high correlation coefficient (R) value of 0.93 in training and 0.79 in testing data, a coefficient of determination (R-squared) value of 0.86 in training and 0.61 in testing data, and when compared to the RF, DT, SVM, K-NN, NB. In comparison to SVM, K-NN, and NB, the standard error value for stacking is 0.12 in training and 0.20 in testing data for category 1, which is low. The performance of the Ensemble (stacking) algorithm is better than other algorithms for the most correlated training and testing for our datasets. For category 3 the Ensemble (stacking) algorithm possesses consistent training and testing similar to category 1. Unlike Ensemble (stacking), RF model unable to satisfy the statistical parameters, in spite of the similar accuracy and CV score (94%) for category 4. Based on model accuracy, CV score, AUC and statistical measures, we propose that Ensemble (stacking) algorithm is the best fit model to classify the breast cancer using blood profile and the selected ratio attributes.

Discussion

Histologic grading is a simple and inexpensive method, often used to measure the clinical behavior of tumor cells, which helps clinicians to choose appropriate therapeutic strategies. The cancer spread and its aggressiveness are often evaluated using the Bloom-Richardson grading system.[3] Accordingly, breast cancer is graded on a 1-3 scale, here 1, 2, and 3 represent well differentiated, moderately differentiated, and poorly differentiated cancer respectively. Poorly differentiated cancer is a rapidly growing cellular phenotype and spreads to visceral organs like the brain, bone, lung, and liver. The grade of the cancer is assessed by histopathological observation of the tissue sections by a pathologist under the microscope.[21] Despite the well-established grading system, significant inter and intra-laboratory variability is often noticed in grading cancer.[22,23] It may be due to variations in the thickness of the sections and a lack of trained pathologists.[24] Histological and mammography image data sets were used for the classification of breast cancer using machine learning and deep learning approaches.[25–28] Recently, deep learning-based breast cancer grading method was also developed using histopathological images.[29] Data acquisition based on histological methods requires invasive methods such as surgery.

Data used in this study is the follow-up on our ongoing research.[19] In our previous work, we found that a Deci-

Table 6. Statistical measure for the trained and testing data for hematological features alone (category 1), Selected ratio attributes alone (category 2, and combined hematological features alone, ratio attributes (category 3) and all features (category 4).

Measures	Training Data						Testing Data					
	RF	DT	SVM	KNN	NB	Stacking	RF	DT	SVM	KNN	NB	Stacking
(Category-1)												
R	1.00	0.99	0.65	0.72	0.60	0.93	0.76	0.55	0.76	0.65	0.69	0.79
R- Squared	1.00	0.97	0.37	0.50	0.29	0.86	0.55	0.04	0.55	0.38	0.44	0.61
SE	0.00	0.05	0.26	0.23	0.28	0.12	0.21	0.31	0.21	0.24	0.23	0.20
(Category-2)												
R	1	0.96	0.51	0.65	0.51	0.627	0.53	0.52	0.65	0.54	0.42	0.65
R- Squared	1	0.91	0.18	0.38	0.12	0.358	0.16	-0.01	0.38	0.21	-0.3	0.38
SE	0	0.10	0.29	0.26	0.31	0.259	0.29	0.32	0.24	0.28	0.36	0.24
(Category-3)												
R	1	0.97	0.69	0.66	0.60	0.88	0.71	0.72	0.71	0.67	0.53	0.81
R- Squared	1	0.95	0.44	0.39	0.21	0.75	0.41	0.41	0.47	0.41	0.00	0.65
SE	0	0.07	0.25	0.26	0.30	0.17	0.24	0.24	0.22	0.23	0.31	0.18
(All Features-Category-4)												
R	1	1	0.74	0.65	0.54	0.91	0.64	0.53	0.57	0.61	0.47	0.61
R- Squared	1	1	0.51	0.37	0.07	0.81	0.24	-0.04	0.28	0.32	-0.16	0.32
SE	0	0	0.23	0.26	0.33	0.14	0.24	0.34	0.25	0.24	0.33	0.25

sion Tree (DT) algorithm predicted breast cancer with an accuracy of 83% with an AUC of 0.87. It may be due small dataset with the least number of features and gathering more attributes such as hormonal status (ER, PR, and Her2), Liver and Kidney function tests, molecular markers (AFP, CA125, CA 15-3, CA 19-9 and CEA) from same cancer patients is expensive and is difficult. On the other hand, neutrophils lymphocyte ratio and absolute lymphocyte count/absolute monocyte count were shown to be prognostic indicators to monitor the progress of urothelial and lymphoma cancer, respectively.[30,31] Therefore, in this paper, we expanded our initial dataset with their ratio counterparts to increase the number of features. The Pearson correlation method is employed to identify the important features. Results showed that individual hematological features are moderately correlated with each other and their correlation is statistically significant. To identify whether the correlation attributes can help us to distinguish the cancer spread we generated a separate data file. This file consists of combined blood profile data of six individual cancer subjects with brain, bone, and liver metastasis (6+6+6=18) instances with 31 features compared with the same number (6) of normal subjects. Similarly, another data file is created for the grade-wise distribution of samples, consisting of grade 2, grade 3, and normal subjects with the same number of instances (62+62+62) with 31 features. ANOVA and Tukey multiple comparison analysis of means among the groups (brain, bone, and liver), as well as grades (normal, grade 2, and grade 3), are performed to identify the strongly correlated features. As shown in Figure 2 the hemoglobin / Red cell count ratio attribute was close to significant and this ratio can be used to distinguish the liver metastatic breast cancer patients as compared to normal subjects. Along the lines, we also found that Haemoglobin content alone can be used to identify moderately differentiated (grade-2) breast cancer patients as compared to normal subjects and monocyte count in combination with Haemoglobin/Red Cell Count, Red Cell Count/Haemoglobin, Neutrophils/Haemoglobin, Neutrophils/Monocyte, Monocyte/Red Cell Count, Monocyte/Neutrophils and Monocyte/Lymphocyte counts can be used to identify poorly differentiated (grade -3) breast cancer patients as compared to normal subjects. None of the features showed a statistically different mean to distinguish grade 2 to grade 3 breast cancer patients. Modified Red blood cells Protein profiles were noticed in metastatic breast cancer patients.[32] Our studies also by monocyte to red blood cell ratio (MRR)[33] and Monocyte/Lymphocyte ratio (0.39) in association with circulating tumor cells[34] for predicting locally advanced breast cancer. It suggests that the selected features can be used for the development of machine learning-based breast cancer classification.

Next, we fitted four different categories of datasets into six different machine learning algorithms like RF, DT, SVM, K-NN, NB, and Ensemble (Stacking). Results were compared before and after hyper parameter tuning. Hyper parameters are parameters that are not learned from the data but are set prior to training a ML model to improve the accuracy and specificity.[35] Various algorithms were developed to identify the relevant features to optimize the machine learning models.[36,37] As shown in Table 5, hyper parameter tuning improved the accuracy of the DT classifier with an accuracy of 97% for category 3 as compared to other models and other categories. To validate the effectiveness of the model we performed an AUC (Area under curve) analysis and 10-fold cross validation (CV) of all the selected models. Cross validation (CV) is commonly used to method to select features with proper diagnosis.[38,39] The FPR (False Positive Rate) and TPR are used to analyze the AUC performance. (True positive rate). As shown in Figure 5 the AUC performance of the ensemble stacking classifier is 97% as compared to the decision tree (86%). Variation in the cross validation scores is noticed for all the datasets before and after hyper parameter tuning. Moderate to less variation is detected for DT and RF models for category 3 and 4 datasets respectively. RF model with maximum depth of 64 and random state 43 showed 0.94 accuracy for dataset with all features and its cross validation score also found to be 94% with an ROC/AUC of 0.93. Based on model accuracy and its correlation with CV score, we believe that RF model can be used to classify the breast cancer. In addition, the effectiveness of the model is further assessed using statistical parameters such as the R, R-squared, and standard error values for all the categories. When considering the comparison of R and R-squared values between training and testing data, it can be observed that the ensemble stacking model has superior performance, as it demonstrates the lowest absolute error. Based on the statistical inference, we believe that ensemble stacking model for the classification of breast cancer spread using hematological features and their selected ratio counterparts (category- 3). Hematological features are routinely used for early diagnosis of hematologic Malignancies using Artificial intelligence. AI-based Artificial neural network (ANN) diagnosed the hematologic Malignancies with an accuracy of 82.5%.[40] Along the lines ratios attributes of hematologic features are often used to predict not only cancer diagnosis but also therapy response.[41,42] Overall, it suggests that features influencing the model performance for breast cancer classification, the Random Forest and Ensemble (stacking)

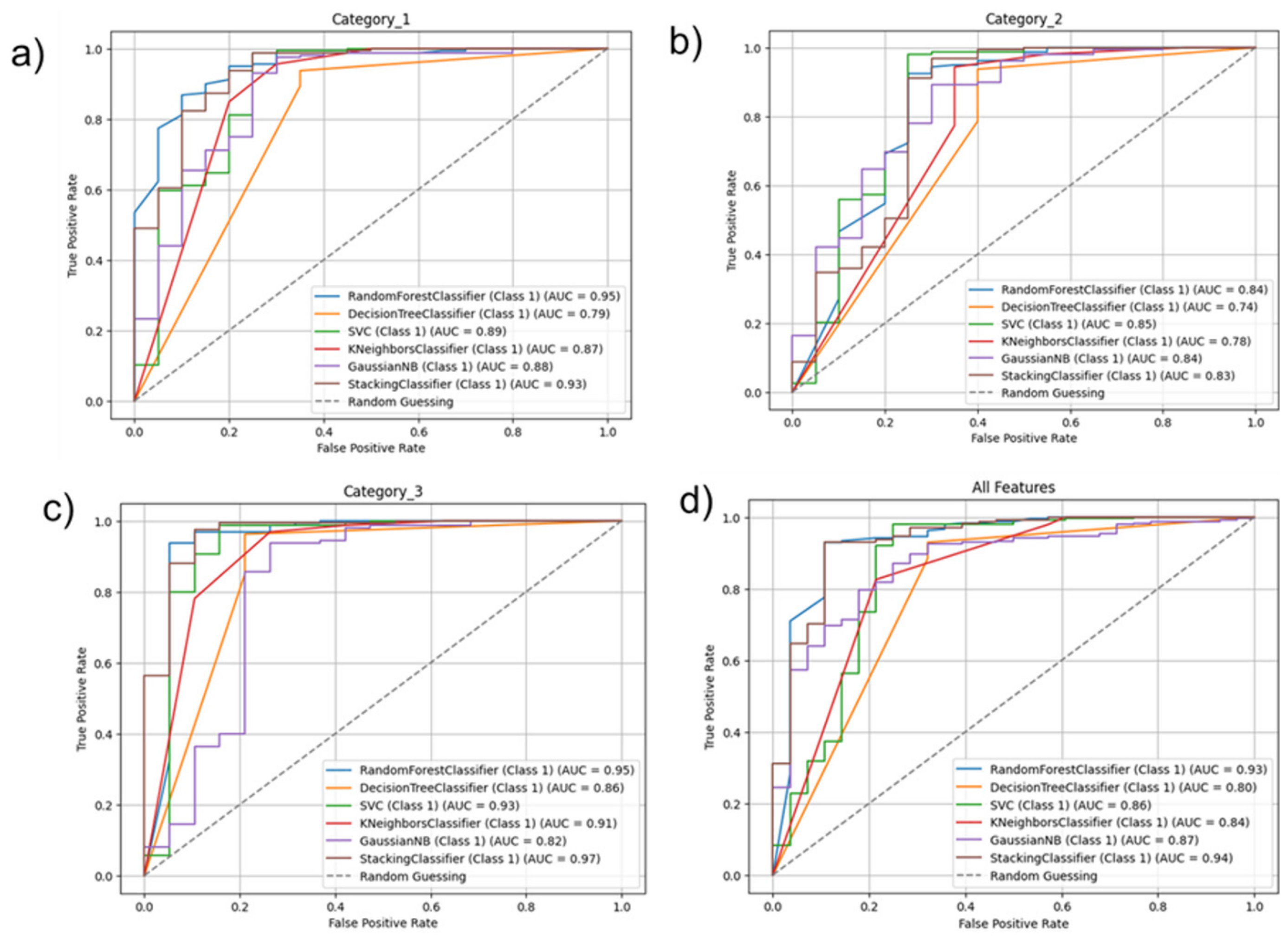


Figure 5. Comparative analysis of AUC between Category 1, 2, 3 and 4 (All features).

emerged as the most suitable and effective regression models within the scope of our research. A combination of graph convolutional network (GCN) and convolutional neural network (CNN), a nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling models were developed using breast mini-MIAS dataset for the effective detection of malignant breast masses and abnormal breast with an accuracy of 96.10% and 94% respectively.[43] Similarly, we hypothesize that the combination of Random Forest with a maximum depth of 64 and random state 43 and Ensemble (stacking) model can be used for the detection of breast cancer using a simple blood profile dataset.

Conclusion

Changes in the blood profile and their interactions within and other immune cells are equally important for diagnosing diseases. Early detection of breast cancer spread using traditional quantitative interpretations based on reference ranges for blood parameters with the greatest accuracy is a challenging task. Our research findings in conjunction with statistical data analysis, correlation-based feature selection, and machine learning models recognized the cancer-associated hematological features, resulting in higher diagnostic accuracy compared to hematological features alone. The ensemble stacking classifier predicted breast cancer with 97% accuracy and it outperformed our previously studied models. Our model can be used for the development of the self-diagnosis mobile app for self-diagnosis of breast cancer using simple blood profile data.

Disclosures

Acknowledgments: We acknowledge Basavatarakam Indo-American Cancer Hospital and Research Institute (BIACH & RI), Hyderabad, Telangana, India for providing data.

Ethics Committee Approval: The study was approved by the Local Ethics Committee.

Peer-review: Externally peer-reviewed.

Conflict of Interest: None declared.

Authorship Contributions: Concept – M.B.; Design – M.B.; Supervision – M.B.; Materials – C.K., Z.K.; Data collection &/or processing – M.B., C.K., Z.K.; Analysis and/or interpretation – M.B., M.B., M.D.V., C.K., Z.K.; Literature search – M.B., M.B., M.D.V., C.K., Z.K.; Writing – M.B.; Critical review – M.B., M.B., M.D.V., C.K., Z.K.

References

1. Aruleba K, Obaido G, Ogbuokiri B, Fadaka AO, Klein A, Adekiya TA, et al. Applications of computational methods in biomedical breast cancer imaging diagnostics: A review. J Imaging 2020;6:105.
2. Tse GM, Tan PH. Diagnosing breast lesions by fine needle aspiration cytology or core biopsy: Which is better? Breast Cancer Res Treat 2010;123:1–8.
3. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up. Histopathology 1991;19:403–10.
4. Kuhn E, Gambini D, Despini L, Asnaghi D, Runza L, Ferrero S. Updates on lymphovascular invasion in breast cancer. Biomedicines 2023;11:968.
5. Granavel H, Joseph LD, Priya M, Chandrasekharan A, Dev B. Inflammatory Lessions. In: Dev B, Joseph LD, eds. Holistic approach to breast disease. Singapore: Springer; 2023. p. 187–210.
6. Sun H, Ding Q, Sahin AA. Immunohistochemistry in the diagnosis and classification of breast tumors. Arch Pathol Lab Med 2023;147:1119–32.
7. Tardivon AA, Guinebretière JM, Dromain C, Deghaye M, Caillet H, Georgin V. Histological findings in surgical specimens after core biopsy of the breast. Eur J Radiol 2002;42:40–51.
8. Harada TL, Nakashima K, Uematsu T, Sugino T, Nishimura S, Takahashi K, et al. Imaging features of breast cancer with marked hemosiderin deposition: A case report. Eur J Radiol Open 2019;6:302–6.
9. Leftin A, Ben-Chetrit N, Klemm F, Joyce JA, Koutcher JA. Iron imaging reveals tumor and metastasis macrophage hemosiderin deposits in breast cancer. PLoS One 2017;12:e0184765.
10. Gosling SB, Arnold EL, Davies SK, Cross H, Bouybayoune I, Calabrese D, et al. Microcalcification crystallography as a potential marker of DCIS recurrence. Sci Rep 2023;13:9331.
11. Tian Y, Zhao L, Gui Z, Liu S, Liu C, Yu T, et al. Clinical and pathological features analysis of invasive breast cancer with microcalcification. Cancer Med 2023;12:11351–62.
12. Sakibuzzaman M, Mahmud S, Afroze T, Fathma S, Zakia UB, Afroz S, et al. Pathology of breast cancer metastasis and a view of metastasis to the brain. Int J Neurosci 2023;133:544–54.
13. Joneja U, Palazzo J. The spectrum of mucinous lesions of the breast. Arch Pathol Lab Med 2023;147:19–29.
14. Tarin D. Understanding cancer. 1st ed. Cham, Switzerland: Springer; 2023. p. 157–73.
15. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: A practical introduction. BMC Med Res Methodol 2019;19:64.
16. Deo RC. Machine learning in medicine. Circulation. 2015;132:1920–30.
17. Jadoon EK, Khan FG, Shah S, Khan A, ElAffendi M. Deep learning-based multi-modal ensemble classification approach for human breast cancer prognosis. IEEE Access 2023;11:85760–9.
18. Chan RC, To CKC, Cheng KCT, Yoshikazu T, Yan LLA, Tse GM. Artificial intelligence in breast cancer histopathology. Histopathology 2023;82:198–210.
19. Botlagunta M, Botlagunta MD, Myneni MB, Lakshmi D, Nayyar A, Gullapalli JS, et al. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Sci Rep 2023;13:485.
20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
21. Rakha EA, Tse GM, Quinn CM. An update on the pathological classification of breast cancer. Histopathology 2023;82:5–16.
22. van Dooijeweert C, van Diest PJ, Willems SM, Kuijpers CCHJ, van der Wall E, Overbeek LIH, et al. Significant inter- and intralaboratory variation in grading of invasive breast cancer: A nationwide study of 33,043 patients in the Netherlands. Int J Cancer 2020;146:769–80.
23. Ginter PS, Idress R, D'Alfonso TM, Fineberg S, Jaffer S, Sattar AK, et al. Histologic grading of breast carcinoma: A multi-institution study of interobserver variation using virtual microscopy. Mod Pathol 2021;34:701–9.
24. Quinn C, Maguire A, Rakha E. Pitfalls in breast pathology. Histopathology 2023;82:140–61.
25. Zebari DA, Ibrahim DA, Zeebaree DQ, Haron H, Salih MS, Damasevicius R, et al. Systematic review of computing approaches for breast cancer detection based computer aided diagnosis using mammogram images. Appl Artif Intell 2021;35:2157–203.
26. Öztürk Ş, Akdemir B. Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA. Procedia Comput Sci 2018;132:40–6.
27. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol 2017;52:434–40.
28. Darweesh MS, Adel M, Anwar A, Farag O, Kotb A, Adel M. Early breast cancer diagnostics based on hierarchical machine learning classification for mammography images. Cogent Eng 2021;8:1968324.
29. Wetstein SC, de Jong VMT, Stathonikos N, Opdam M, Dackus GMHE, Pluim JPW, et al. Deep learning-based breast cancer

grading and survival analysis on whole-slide histopathology images. Sci Rep 2022;12:15102.

30. Shin SJ, Roh J, Kim M, Jung MJ, Koh YW, Park CS, et al. Prognostic significance of absolute lymphocyte count/absolute monocyte count ratio at diagnosis in patients with multiple myeloma. Korean J Pathol 2013;47:526–33.
31. Johnson PJ, Dhanaraj S, Berhane S, Bonnett L, Ma YT. The prognostic and diagnostic significance of the neutrophil-to-lymphocyte ratio in hepatocellular carcinoma: A prospective controlled study. Br J Cancer 2021;125:714–6.
32. Pereira-Veiga T, Bravo S, Gómez-Tato A, Yáñez-Gómez C, Abuín C, Varela V, et al. Red blood cells protein profile is modified in breast cancer patients. Mol Cell Proteomics 2022;21:100435.
33. Wang Y, Wang H, Yin W, Lin Y, Zhou L, Sheng X, et al. Novel lymphocyte to red blood cell ratio (LRR), neutrophil to red blood cell ratio (NRR), monocyte to red blood cell ratio (MRR) as predictive and prognostic biomarkers for locally advanced breast cancer. Gland Surg 2019;8:627–35.
34. Kasimir-Bauer S, Karaaslan E, Hars O, Hoffmann O, Kimmig R. In early breast cancer, the ratios of neutrophils, platelets and monocytes to lymphocytes significantly correlate with the presence of subsets of circulating tumor cells but not with disseminated tumor cells. Cancers (Basel) 2022;14:3299.
35. Burçak KC, Baykan ÖK, Uğuz H. A new deep convolutional neural network model for classifying breast cancer histopathological images and the hyperparameter optimisation of the proposed model. J Supercomput 2021;77:973–89.
36. Vakharia V, Shah M, Suthar V, Patel VK, Solanki A. Hybrid perovskites thin films morphology identification by adapting multiscale-SinGAN architecture, heat transfer search optimized feature selection and machine learning algorithms. Phys Scr 2023;98:025203.
37. Vakharia V, Shah M, Nair P, Borade H, Sahlot P, Wankhede V. Estimation of lithium-ion battery discharge capacity by integrating optimized explainable-AI and stacked LSTM model. Batter 2023;9:125.
38. Vakharia V, Gupta VK, Kankar PK. A comparison of feature ranking techniques for fault diagnosis of ball bearing, Soft Comput 2016;20:1601–19.
39. Falessi D, Huang J, Narayana L, Thai JF, Turhan B. On the need of preserving order of data when validating within-project defect classifiers. Empir Softw Eng 2020;25:4805–30.
40. Syed-Abdul S, Firdani RP, Chung HJ, Uddin M, Hur M, Hyeon J, et al. Artificial intelligence based models for screening of hematologic malignancies using cell population data. Sci Rep 2020;10:4583.
41. Jalali A, Miresse D, Fahey MR, Ni Mhaonaigh N, McGuire A, Bourke E, et al. Peripheral blood cell ratios as prognostic indicators in a neoadjuvant chemotherapy-treated breast cancer cohort. Curr Oncol 2022;29:7512–23.
42. Luo L, Tan Y, Zhao S, Yang M, Che Y, Li K, et al. The potential of high-order features of routine blood test in predicting the prognosis of non-small cell lung cancer. BMC Cancer 2023;23:496.
43. Zhang YD, Satapathy SC, Guttery DS, Górriz JM, Wang SH. Improved breast cancer classification through combining graph convolutional network and convolutional neural network. Inf Process Manag 2021;58:102439.