scientific reports

Check for updates

OPEN

Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI

Taminul Islam[1], Md. Alif Sheakh[2], Mst. Sazia Tahosin[2], Most. Hasna Hena[2], Shopnil Akash[3], Yousef A. Bin Jardan[4], Gezahign FentahunWondmie[5]✉, Hiba-Allah Nafidi[6] & Mohammed Bourhia[7]✉

Breast cancer has rapidly increased in prevalence in recent years, making it one of the leading causes of mortality worldwide. Among all cancers, it is by far the most common. Diagnosing this illness manually requires significant time and expertise. Since detecting breast cancer is a time-consuming process, preventing its further spread can be aided by creating machine-based forecasts. Machine learning and Explainable AI are crucial in classification as they not only provide accurate predictions but also offer insights into how the model arrives at its decisions, aiding in the understanding and trustworthiness of the classification results. In this study, we evaluate and compare the classification accuracy, precision, recall, and F1 scores of five different machine learning methods using a primary dataset (500 patients from Dhaka Medical College Hospital). Five different supervised machine learning techniques, including decision tree, random forest, logistic regression, naive bayes, and XGBoost, have been used to achieve optimal results on our dataset. Additionally, this study applied SHAP analysis to the XGBoost model to interpret the model's predictions and understand the impact of each feature on the model's output. We compared the accuracy with which several algorithms classified the data, as well as contrasted with other literature in this field. After final evaluation, this study found that XGBoost achieved the best model accuracy, which is 97%.

Keywords Breast cancer prediction, Machine learning, Cancer prediction, Hyperparameter tuning, Explainable AI

Breast cancer begins when some cells in the breast start to grow uncontrollably, forming a mass called a tumor[1]. A breast cancer diagnosis typically falls into one of two main categories—benign (non-cancerous) or malignant (cancerous). Malignant tumors are dangerous as they can spread to distant sites in the body through the bloodstream or lymph system, a process known as metastasis[2,3]. Figure 1a illustrates the distinction between benign and malignant tumors in terms of the normal cells and tumor cells, and (b) shows the benign and malignant masses. Benign tumors generally stay localized in one area and do not metastasize. Breast cancer manifests through several symptoms—a noticeable lump or mass in the breast, changes in breast size or shape compared to the other breast, alterations in the skin overlying the breast like dimpling or puckering, newly inverted nipple, redness or scaliness of breast skin, breast pain, and nipple discharge other than breast milk[4].

Breast cancer is the second largest killer of women after cardiovascular disease. And more than 8% of women will experience it. Every year, more than 500,000 women are diagnosed with breast cancer, as stated in the

[1]School of Computing, Southern Illinois University Carbondale, Carbondale, IL, USA. [2]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. [3]Department of Pharmacy, Faculty of Allied Health Sciences, Daffodil International University, Dhaka, Bangladesh. [4]Department of Pharmaceutics, College of Pharmacy, King Saud University, P.O. Box 11451, Riyadh, Saudi Arabia. [5]Department of Biology, Bahir Dar University, P.O. Box 79, Bahir Dar, Ethiopia. [6]Department of Food Science, Faculty of Agricultural and Food Sciences, Laval University, 2325, Quebec City, QC G1V 0A6, Canada. [7]Laboratory of Biotechnology and Natural Resources Valorization, Ibn Zohr University, 80060 Agadir, Morocco. ✉email: resercherfent@gmail.com

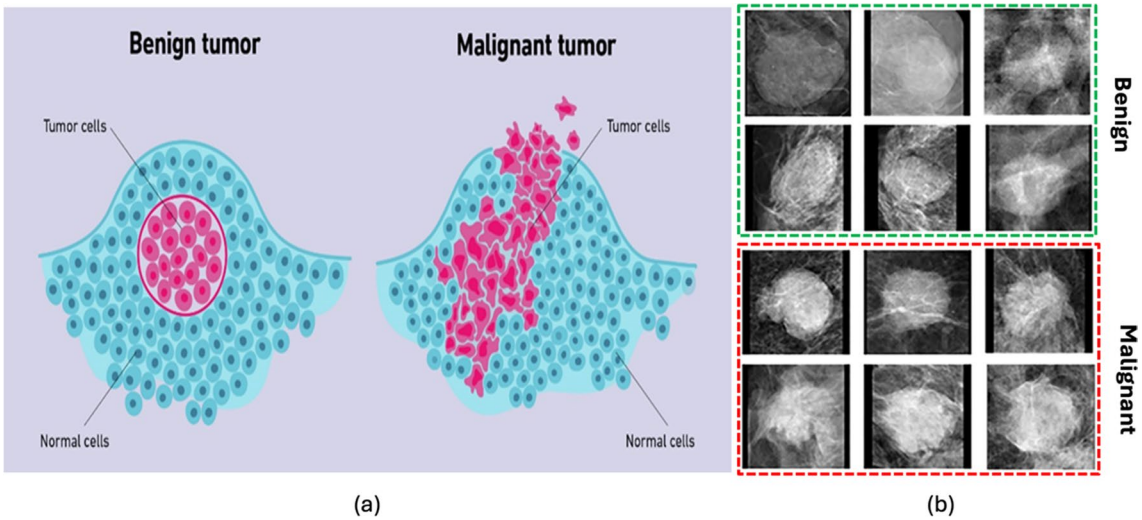

Figure 1. Visualization of breast cancer: (**a**) benign and malignant tumor cells, (**b**) benign and malignant masses.

World Health Organization's annual report[5]. In developing countries, due to a lack of screening programs and awareness, women often present at an advanced stage where treatment options are limited. Known risk factors for breast cancer include genetic mutations in BRCA genes, reproductive history (nulliparity, early menarche, late menopause), hormonal factors (use of hormone replacement therapy, oral contraceptives), obesity, alcohol consumption, smoking, radiation exposure at a young age and family history[6].

Many people were affected by cancer during this period. We can't pinpoint the origin of the sickness since it's tied to factors beyond our control. This is also a screening technique for identifying the cancer's aggressiveness. Several assessment items are connected to cancer detection, including clamp thickness, cell size consistency, and shape regularity. The outcome is difficult even for those tasked with inspiring others to take action, and yet the use of machine learning and other computer science techniques as general diagnostic tools has expanded in recent years. Countless numbers of people's lives have been saved by computer diagnostic programs that use diseases that have killed millions. In the realm of surgery, robotics is indispensable. Aside from other artificial intelligence's widespread usage in cancer detection, the system deployed in the intensive care unit is highly effective[7]. One in eight American females may get cancer between the ages of 15 and 19[8]. Breast cancer is the result of unchecked cell division, which can also cause breasts to sag (called tumors)[9]. In most cases, the tumor poses no health risk. The necessity for precise categorization in the clinic may be a severe challenge for doctors and health workers, especially when the correct identification of the determinants might contribute to survival, regardless of whether the condition is benign or malignant.

The diagnosis of breast cancer involves a step-wise approach starting with a thorough clinical examination and radiological tests like mammograms and breast ultrasounds. This may be followed by tissue sampling through fine needle aspiration cytology (FNAC) or biopsy from suspicious areas and microscopic assessment to confirm malignancy[10]. As symptoms of breast cancer can be non-specific with wide variation across patients, the combination of these investigations is needed for accurate diagnosis in the majority of cases. So, to monitor and diagnose diseases, a human observer must be able to pick out very particular signal features. Due to the large number of patients in the critical care unit and the need for round-the-clock monitoring, several CAD approaches[11] for computer-aided medical systems have emerged in the recent decade to meet this issue. With these strategies, the challenge of classifying quantitative features may be posed rather than relying on qualitative diagnostic criteria. Machine learning algorithms can predict breast cancer diagnosis and prognosis[12]. The purpose of this work is to evaluate the efficacy and performance of these algorithms in terms of their accuracy, sensitivity, range, and precision. In the past 25 years, the importance of artificial intelligence has increased[13]. As scientists realize the importance of making firm decisions about how to treat certain diseases, the use of computers and machine learning as diagnostic tools has become deadly, which is the most serious disease screening task in the medical field[14]. One of the most important functions of disease is the definition of cancer. Using machine learning technology, doctors can detect, identify, and classify tumors as benign or malignant. There are some challenges in analyzing patient data and choosing doctors and specialists, but cognitive systems and computational methods (such as ML for classification) will ultimately help doctors and professionals[15]. However, as machine learning models become more complex, there is a need for Explainable AI (XAI) techniques to interpret these models and understand how they arrive at their predictions. Explainable AI methods like SHAP (SHapley Additive exPlanations) can shed light on how different features contribute to a model's output, increasing trust and transparency in the model's decision-making process[16]. In this study, we employ SHAP analysis on our best performing XGBoost model to explain its predictions and understand which factors have the greatest impact on determining if a patient has early-stage breast cancer or not. This study makes several key contributions to the prediction of early-stage breast cancer using supervised machine learning approaches:

- Employing hyperparameter tuning to optimize each machine learning algorithm and enhance performance.

- Utilizing a primary dataset for algorithm evaluation.
- Demonstrating that XGBoost achieved the highest accuracy of 97% and F1 score of 0.96, surpassing other algorithms.
- Conducting SHAP analysis on the XGBoost model to interpret its predictions and comprehend the impact of each feature on the model's output.

Literature review

Every day, the medical sector discovers new machine learning applications. The development is beneficial to scientific research. There is an abundance of research being conducted on this topic. Several research articles pertinent to our study have been uncovered. This project aims to provide a mechanism for predicting breast cancer. The majority of the dataset was obtained from the Dhaka Medical College Hospital. During this study, we were exposed to a few novel methodologies. Not a straightforward undertaking on our end. This notion will be discussed in further detail in the next chapter. To completely apply this study and to learn this new term, we examined prior research about the prediction of heart attacks.

Using their model, V. Chaurasia and T. Pal determined which machine learning algorithms performed the best in predicting breast cancer. In their study, they used Support Vector Machine (SVM), Naive Bayes (NB), Radial basis function Neural Networks (RBF NN), Decision Tree (DT), and a simplified version of Classification and Regression Trees (CART)[17]. After adopting their successful model, they obtained the highest Area Under the Curve (AUC) (96.84%) using Support Vector Machine on the original Wisconsin Breast Cancer datasets.

Djebbari et al.[18] evaluated if a machine learning ensemble might predict breast cancer survival time. Their breast cancer dataset they achieve a greater rate of accuracy using their technique than was seen in previous studies. S. Aruna and L. Nandakishore examine the performance of Decision Tree, Support Vector Machine, Naive Bayes, and K-Nearest Neighbors (K-NN) for classifying White Blood Cell (WBC)[19]. Their top Support Vector Machine classifier AUC was 96.99%.

M. Angrap used six machine learning techniques to categorize tumor cells. Gated Recurrent Unit, a variant of the long short-term memory neural network, was created and implemented Gated recurrent unit (GRU). The SoftMax layer of the neural network was switched out for a layer of Support Vector Machine. With an accuracy of 99.04%, GRU Support Vector Machine performed best in that study[20]. Cross-validation was used by Karabatak et al.[21] to improve the accuracy of a model trained with association rules and a neural network to 95.6%. Naive Bayes classifiers were utilized, using a novel weight adjustment method.

Mohebian et al.[22] looked at the feasibility of using ensemble learning to foretell cancer recurrence. Three machine learning models that performed very well when fed a relevance vector were compared and contrasted by Gayathri et al.[23]. Payam et al.[24] employed a number of methods for preprocessing and data reduction, including a radial basis function network (RBFN), to achieve their goals.

Using information from breast cancer studies reported in[25], researchers created survival prediction models. In this study, they have used survival prediction methods to both benign and malignant tumor of breast cancers. Extensive historical research shows that machine learning algorithms for breast cancer diagnosis have been investigated at length, as illustrated in[26]. They suggested that data augmentation strategies might help address the problem of having insufficient data. In[27], the authors showed how to automatically detect and identify cell structure using features of computer-aided mammography images. Many different methods of categorization and clustering have been evaluated, as reported in[28].

Fatih Muhammed and Ak[29] compared detection and diagnosis of breast cancer using data visualization and machine learning. Using Dr. William H. Walberg's breast tumor data, they used a variety of techniques including Logistic Regression (LR), nearest neighbor (NN), Support Vector Machine, simple Bayes, Decision Tree, random forest (RF), and convolutional forest using R, Minitab, and Python. Logistic regression with all features achieved the highest accuracy (98.1%), indicating high performance. Their research showed the benefits of data visualization and machine learning in cancer diagnosis, opening up new opportunities for cancer diagnosis.

Md. Islam et al.[30] compared five supervised machine learning methods for breast cancer prediction using the Wisconsin Breast Cancer Database. These methods include Support Vector Machine, Nearest Neighbor, Random Forest, Artificial neural networks (ANN), and Logistic regression. ANNs outperformed others by achieving the highest accuracy (98.57%), precision (97.82%) and F1 score (0.9890). The researchers concluded that machine learning for disease detection could provide medical staff with reliable and rapid responses to reduce the risk of death.

Vikas Chaurasia and Saurabh Pal[31] used machine learning to predict breast cancer using the Wisconsin Diagnostic Breast Cancer Database. They compared six algorithms, reduced features to 12 using statistical methods, and used ensemble methods to combine models. The results show that all the algorithms performed well, with a test accuracy of over 90%, especially in the refined feature section. His contributions include the use of feature selection and ensemble methods to improve breast cancer prediction accuracy.

Kabiraj et al.[32] Creating a breast cancer risk prediction model using Extreme Gradient Boosting (XGBoost) and Random Forest algorithms. The dataset used is from the UCI Machine Learning Repository. This approach includes the use of Random Forest and XGBoost methods, and the model achieves a classification accuracy of 74.73%.

Meerja Jabbar et al.[33] proposed a new ensemble method using Bayesian Network and Radial Basis function to classify breast cancer data. This method achieves 97% accuracy, better than existing approaches. The trial was conducted on the Wisconsin Breast Cancer Dataset (WBCD) using a variety of metrics to measure performance. The proposed ensemble study can help cancer specialists make accurate tumor diagnoses and support patients in making treatment decisions.

Shalini and Radhika[34] are working to predict breast cancer using different machine learning techniques. They use the UCI machine learning database and use artificial neural networks, Decision Tree, Support Vector Machine and Naive Bayes algorithms. As a result, a classification accuracy of 86% was found.

Naji et al.[35] Machine learning algorithms are used to predict and diagnose breast cancer. They compared five different algorithms, including Support Vector Machine, Random Forest, Logistic regression, Decision Tree (C4.5), and KNN, using the Wisconsin breast cancer diagnostic database. The main goal is to determine the best algorithm for breast cancer diagnosis. The results revealed that the support vector machine outperformed the other classifiers and achieved the highest accuracy of 97.2%. Research conducted with the Scikit learning library in the Anaconda Python environment contributes important insights to update breast cancer therapy and improve patient safety standards.

Puja Gupta and Shruti Garg[36] investigated breast cancer prognosis using six supervised machine learning algorithms and deep learning. The study includes a parametric analysis of each algorithm to achieve greater accuracy. The data set used in the study is not mentioned. The article describes data pre-processing, machine learning algorithms and their key parameters. Their research results show that deep learning using Human Gradient Descent Learning is the most accurate with an accuracy rate of 98.24%. The paper concludes that proper hyperparameter machine learning tools can help identify tumors effectively.

Methodology

This research aims to anticipate breast cancer and achieve the highest level of precision possible. To run the model, we first build the dataset and choose which methods to employ. Supervised Learning[37] and Unsupervised Learning[38] are two techniques to develop a method in machine learning algorithms. We employed supervised machine learning methods in this study. In supervised learning, some classification methods are utilized to address classification issues. We employed the Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, and XGBoost method in this study. In the forthcoming part on the suggested technique, we will explore all of the algorithms, how they function, and which one is the best, and we will attempt to determine which algorithm produces the best results based on the data set we gathered. After the model demonstrated robust predictive accuracy and generalization capability, we performed explainable AI using SHAP to interpret the model's predictions and understand the impact of each feature on the model's output.

Dataset description

In this research, we have gathered a total of 500 patient's information from a government hospital in Bangladesh known as Dhaka Medical College Hospital. In this dataset, seven features were extracted from the image by Dhaka Medical College Hospital experts. The data has been used in this research with the consent of all patients. The dataset features are given below with a short description in Table 1. According to Fig. 2, the data shows 254 noncancerous cases, while the remaining 246 cases are considered to be cancerous. Table 1, shows a short description of our dataset:

In this research we considered geometric features and age for the classification because geometric features, such as shape irregularity and size, play a crucial role in identifying abnormal growth patterns associated with malignant tumors. On the other hand, Age is a significant factor as breast cancer incidence increases with age,

Label of the Dataset	Description
age	From 22 to 55 years women data is collected here
mean_radius	The average of the radius of the tumor cells
mean_texture	The average of the gray-scale values in the texture of the tumor cells
mean_perimeter	The average size of the tumor cell's boundary
mean_area	The average area of the tumor cells
mean_smoothness	The average smoothness of the tumor cell's surface
diagnosis	The classification of the breast tissue as benign is 0 or malignant is 1

Table 1. Short description of our dataset.

Figure 2. Ratio of malignant and benign data based on overall dataset.

and the disease can manifest differently in younger versus older patients. By incorporating these features into our classification model, we aim to capture the distinct characteristics of benign and malignant tumors, improving the accuracy of our predictions. The full dataset was considered throughout the analysis of the dataset. It is seen in Fig. 3 that the mean radius of the dataset is a counterpoint. Patients believed to have cancer have a radius bigger than 1, whereas those without symptoms have a radius nearer to 1. The full dataset was considered throughout the analysis of the dataset. It is seen in Fig. 3 that the mean radius of the dataset is a counterpoint. Patients believed to have cancer have a radius bigger than 1, whereas those without symptoms have a radius nearer to 1.

Proposed model workflow

The model workflow proposed in Fig. 4 includes several steps, such as collecting the dataset, performing data preprocessing, splitting the data into 80% training and 20% testing sets, selecting the most relevant features, selecting a suitable supervised machine learning algorithm, classifying the samples into benign tumor or malignant tumor classes, and evaluating the model's performance. The initial step is to obtain the dataset needed to train the model. Once we obtain the dataset, we preprocess it by cleaning and translating the raw data into a machine-learning-friendly format. Following that, we divided the preprocessed dataset into two subsets: one for training and another for testing the model. After splitting, we select the most appropriate supervised machine learning method, which is a Random Forest classifier, Decision Tree, XGBoost, Naive Bayes, and Logistic Regression, and train this on the features. After training the model, we use it to categorize new samples into benign or malignant groups. Finally, we test the model's performance using several measures like AUC, precision, recall, F_1 score, and accuracy. During the training process, the model acquired the ability to identify and differentiate patterns and features that are indicative of malignant and benign instances of breast cancer, thereby enhancing

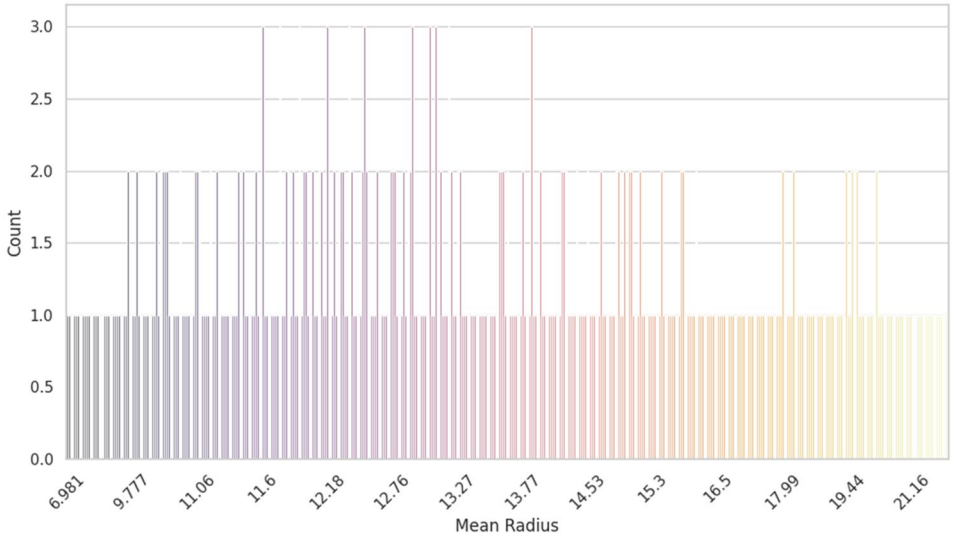

Figure 3. Mean radius of this work on our dataset.

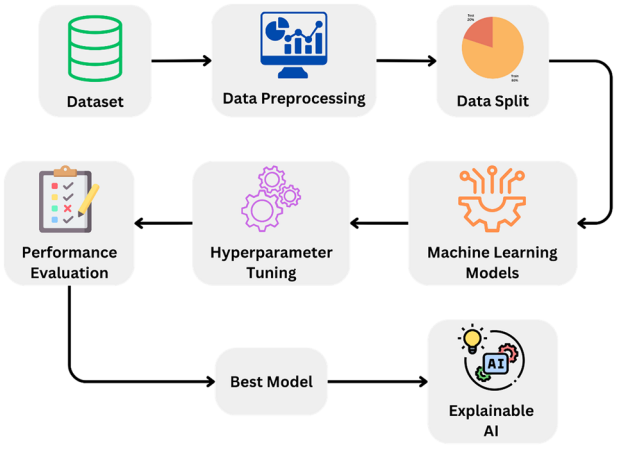

Figure 4. Visualizing the workflow of the proposed model.

its predictive capability. The performance of the model on the validation set was consistently monitored, and iterative fine-tuning was conducted until satisfactory results were achieved. After the model demonstrated robust predictive accuracy and generalization capability, we performed explainable AI using SHAP to interpret the model's predictions and understand the impact of each feature on the model's output. Additionally, we proceeded to implement the k-fold cross-validation technique. This proposed method ensures that the model is well-trained, correctly classifies samples, and performs optimally.

Data preprocessing
To make it suitable for machine learning we do cleaning and transforming raw data. This includes handling null values, scaling features using a standard scaler, encoding categorical variables, and normalizing data. The goal is to prepare data by improving the performance and efficiency of machine learning models.

Data splitting
After preprocessing we divided the dataset into a training set and a testing set. Here we use a common split ratio of 80:20, where 80% of the data is used to train the model and 20% is reserved for evaluating the model's performance. This ensures that the generalization ability of the model can be evaluated and prevents overfitting of the training data.

Machine learning model

Machine learning is the most practical way of predicting breast cancer sickness. Reading through the literature review, it becomes clear that the bulk of the work has been accomplished using machine learning and deep learning techniques. It is often understood that deep learning falls within the umbrella of machine learning. Five separate machine learning methods were used to this new dataset to find the most accurate method. Decision Tree, XGBoost, Logistic regression, Naive Bayes, and Random Forest are the categories used to organize these methods[39]. In this section, we'll get a brief overview of a few of these designs.

Decision tree (DT)

Decision trees are widely used in machine learning for data classification and prediction. It can have different structures depending on the application. Although they work well for some classifiers, they can struggle with a large number of classes and limited training data. The resulting decision tree is easy to understand for domain experts, making it valuable for problem solving. In addition, it can be combined with ensemble methods to further improve performance. In summary, Decision trees are versatile and effective in many industries, including finance[40]. Decision trees classifier can be written as –

$$f_{\Theta}(x) = \sum_{\ell \epsilon leaves(\mathcal{T})} \theta_{\ell} \mathcal{T}_{\ell}(x) \tag{1}$$

In Eq. 1, $f_{\Theta}(x)$ represents the model's final estimated for input vector x. $\mathcal{T}$ denotes the decision tree, θ_{ℓ} is the weight associated with leaf node l, and $\mathcal{T}_{\ell}(x)$ denotes an indicator function that returns 1 if x falls within the region defined by leaf node l and 0 otherwise. The total of all the leaves in the tree ensures that the final prediction includes contributions from all of the different trees in the ensemble[41].

Random forest (RF)

Random Forest is a popular machine-learning technique used for both classification and regression tasks. It creates multiple decision trees from different parts of the training data. Each tree classifies the data separately and the final prediction is the sum of all the individual forecasts. This approach reduces the risk of over fitting, leading to more accurate and reliable predictions. Additionally, as the number of trees increases, the method becomes more robust to noise and outliers in the data. However, there is a trade-off between accuracy and computational efficiency, as training more trees requires more time and resources. These algorithms divide the data recursively[42]. The Random Forest algorithm is provided below—

Node t, randomly selection v of the p independent variables.

∀of the k=1, …, vis the variable that was sampled; for each possible split of kth, determine the optimal split sk.

In s*, gain the optimal splitting k from k=1 to k=m; Determine the optimal split sk for splitting node t; jth variable is defined cut point cs* that is used for splitting node t.

The data is separated here, with the i=1, …, n; observation with xij<cs* going to the left descending node and all other observations going to the right descending node.

To grow a tree of maximum size Tb, simply repeat steps 1–4 for each node in the tree's descendent set.

Algorithm 1. Random Forest classifier.

XGBoost (XGB)

XGBoost (Extreme Gradient Boosting) is a widely used method for building robust predictive models, especially in machine learning[39]. It uses aggregated decision trees to make accurate predictions from complex datasets. Although decision trees are easy to interpret, XGBoost can be difficult to understand at first glance. However, data scientists and machine learning experts prefer XGBoost because it efficiently processes large datasets and quickly builds accurate models. This powerful and versatile tool strikes a balance between model complexity and prediction accuracy using gradient descent and regularization techniques. It is highly adaptable, making it valuable for extracting insights from complex datasets. However, careful optimization of the hyperparameters is necessary to achieve the best results. With the right approach, XGBoost enables data scientists to build accurate and reliable models that offer valuable insights into complex datasets[43]. XGBoost algorithm can be written as –

$$Obj = \sum_{i=1}^{m} l\left(y_i, \widehat{y}_i\right) + \sum_{k=1}^{m} \Omega\left(f_k\right) \tag{2}$$

where f_k is the leaf node's regular term of the kth classification tree, $l\left(y_i, \widehat{y}_i\right)$ is the training error of sample xi, and Obj is the objective function[44].

Naive Bayes (NB)

Naive Bayes is a classification algorithm that assumes that features are conditionally independent, given class labels. Although this assumption is often violated in real databases, it is still useful in practical applications. Although not independent, classification can derive information from features. For high-dimensional data, is fast and efficient with less training data than complicated and complex models because it estimates the probability of each feature separately. It can handle both discrete and continuous data, making it versatile for different databases. In natural language processing, it is essential for text categorization and spam filtering. Overall, Naive Bayes is a useful machine learning tool for solving classification problems[45]. Naive Bayes classifier can be written as –

$$f_c^{NB}(x) = \prod_{j=1}^{n} P\left(X_j = x_j | C = c\right) P\left(C = c\right) \tag{3}$$

In Eq. 3, Naive Bayes(x) is the probability that observation x is in class c, f_c is the class-c observations, and n is the number f observations. $P\left(X_j = x_j | C = c\right) P\left(C = c\right)$ is the conditional probability of seeing feature j in class c. Divide class c's observations by feature j's x_j usage. Training data can be used to calculate $P(C = c)$. Multiplying feature conditional probabilities and prior probabilities classifies new data into the most likely class[46].

Logistic regression (LR)

Using labeled data, we train our model in supervised learning. Logistic regression is used for categorization problems in supervised learning. Logistic regression's discrete output variable (y) is usually 0 or 1. A sigmoid function simulates X's effect on the output variable. This function provides a probability between 0 and 1 indicating the input's likelihood of being positive (1). Finance, marketing, and healthcare use logistic regression. Based on medical history and demographic data, logistic regression can estimate patients' cancer risk. Logistic regression handles nonlinear input–output relationships. For massive datasets, it requires fewer computing resources. Logistic regression's simplicity and effectiveness make it a common classification problem solution. Logistic regression can also provide relative relevance information for feature selection and model interpretation[47]. Logistic regression classifier can be written as –

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_0 x_1 + \cdots + \beta_m x_m \tag{4}$$

where $\pi(x)$ denotes the probability of a binary outcome (such as success or failure) given the values in the predictor vector x. The log odds are predicted as a linear combination of the predictor variables, and the coefficients $\beta_0, \beta_1, \ldots, \beta_m$ show the impacts of each predictor on the log chances. Exponentiating the equation allows one to determine the odds of a successful outcome given specific values of the predictors[48].

Ethical approval

The current research is approved by the Dhaka Medical College Hospital Cancer Sample & Research Center Informed consent was obtained, and ethical guidelines were followed. The approval letter confirms that all necessary precautions were taken to ensure the protection of human subjects and adherence to ethical standards.

Experimental result

A lot of people make blunders in training or when extrapolating their results. Because the training error rate decreases with increasing model complexity, increasing the model's complexity can assist reduce training mistakes. The Bias-Variance Decomposition (Bias + Variance) method can be used to reduce the number of incorrect generalizations. Over fitting occurs when a reduction in training error results in an in-crease in test error rates. Each classification method may be judged by its accuracy, precision, recall, and F1 score.

When gauging the success of their models, writers used a wide range of techniques. While most studies looked at a combination of markers to determine how well they did, some just used one. The work is evaluated here using the criteria of accuracy, precision, recall, and the F1 score. For analyzing prediction data, this four-factor system is ideal. The capacity to appropriately recognize and categorize incidents is related to accuracy. Equation 5[49] shows the formula of accuracy.

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{TruePositive} + \text{FalsePositive} + \text{TrueNegative} + \text{FalseNegative}} \tag{5}$$

Specifically, accuracy in statistics is defined as the ratio of actual positive occurrences to the total predicted positive events. The mathematical expression of accuracy is given by Eq. 6[50].

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \tag{6}$$

The term "harmonic mean" describes this method since it balances accuracy and memory. A version of the mathematical equation for the F_1 score is given by Eq. 7[51].

$$F_1\text{score} = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{7}$$

Result analysis

Table 2 provides information on hyperparameter tuning and each metric for different machine learning algorithms: Decision Tree, Random Forest, XGBoost, Naive Bayes, and Logistic regression. Hyperparameter tuning is an important step to improve the performance of machine learning models and involves finding the best combination of hyperparameters to achieve the highest precision, accuracy, recall, and F_1 score. Three hyperparameters are set for the decision tree algorithm: max_depth, min_samples_leaf, and min_samples_split. The best

Algorithms	Hyperparameter tuning	Range	Best	Accuracy	Precision	Recall	F_1 Score
Decision tree	max_depth	None, 5, 10	5	0.91	0.94	0.89	0.9
	min_samples_leaf	2, 5, 10	4				
	min_samples_split	1, 2, 2004	5				
Random forest	max_depth	None, 5, 10, 20	None	0.96	0.93	0.95	0.94
	min_samples_leaf	1, 2, 2004	1				
	min_samples_split	2, 5, 10	5				
	n_estimators	100, 300, 500	300				
XGBoost	learning_rate	0.01, 0.1, 0.3	**0.01**	**0.97**	**0.94**	**0.95**	**0.96**
	max_depth	3, 5, 2007	**3**				
	n_estimators	100, 300, 500	**500**				
	subsample	0.8, 1.0	**1**				
Naive Bayes	No			0.94	0.99	0.9	0.94
Logistic Regression	Regularization strength	0.001, 0.01, 0.1, 1, 10	10	0.93	0.93	0.93	0.93

Table 2. Hyperparameter tuning with performance metrics for all algorithms. E.g. Significant values are in [bold]

combination of hyperparameters that resulted in the highest F_1 score of 0.90 was a max_depth of 5, min_samples_leaf of 4, and min_samples_split of 5. Next, the Random Forest algorithm was tuned with four hyperparameters: max_depth, min_samples_leaf, min_samples_split, and n_estimators. The best combination of hyperparameters, which resulted in an impressive F1score of 0.94, included a min_samples_leaf of 1, min_samples_split of 5, and n_estimators of 300. The max_depth hyperparameter was not specified, indicating that the default value or automatic selection method might have been used. For the XGBoost algorithm, four hyperparameters were tuned: learning_rate, max_depth, n_estimators, and subsample. The best combination of hyperparameters achieved the highest F_1 score of 0.96, with a learning_rate of 0.01, max_depth of 3, n_estimators of 500, and subsample of 1.0. The Naive Bayes algorithm did not require hyperparameter tuning, and its default settings were used. It still achieved a respectable F_1 score of 0.94. Finally, the Logistic Regression algorithm was tuned for the hyperparameters regularization strength, max_iter, and penalty. The best combination of hyperparameters resulted in an F_1 score of 0.93, with a regularization strength of 10, max_iter of 100, and penalty using L2 regularization.

In terms of performance metrics, the XGBoost algorithm outperformed others with an effective precision of 0.97 and high precision, recall, and F_1 scores. The random forest algorithm also performed well with an accuracy of 0.96 and a balanced accuracy trade-off. The decision tree algorithm achieved good results with an accuracy of 0.91 and a balanced F_1 score of 0.90. Naive Bayes and Logistic regression show competitive performance with F_1 scores of 0.94 and 0.93, respectively. Overall, hyper-parameter tuning plays an important role in improving the performance of the model, and the choice of algorithm significantly influenced the final result, with XGBoost and Random Forest standing out as high-performance models.

As can be seen in Table 2, it is evident that the accuracy of Random Forest and XGB is significantly greater than that of the other five machine-learning methods. When compared to the other algorithms that were used, the results of the Decision Tree method are significantly inferior to those of the Naive Bayes and Logistic Regression algorithms. In terms of the AUC comparison, the best results were achieved by Random Forest and XGB.

Confusion matrices (CM) can be used to rapidly and easily summarize a classification system's efficacy. When the quantity of observations across categories differs significantly, even when there are only two categories in the dataset, the categorization may be incorrect. For more insight into the precision of the classification approach, we can compute a CM (Fig. 5, 6, 7, 8 and 9).

Additionally, receiver operating characteristic (ROC) curves illustrate the diagnostic ability of a binary classifier as its discrimination threshold is varied. The area under the ROC curve (AUC) provides an aggregate measure across all possible classification thresholds. For our XGBoost model, the ROC AUC was 0.98, indicating excellent overall performance in distinguishing between the two classes shown in Fig. 10. This high ROC AUC means the model is reliably assigning higher scores to positive instances than negative instances. We can have increased confidence in its ability to generalize well to new data. Further analysis into the confusion matrix for specific probability thresholds would provide deeper insight into preferred operating points along the ROC curve.

Performance analysis using explainable AI

Here we perform SHAP analysis on our best performing XGBoost model.

In Fig. 11, the SHAP summary plot shows the impact of each feature on the model's output. The x-axis represents the SHAP value, where higher positive values indicate a higher probability of predicting early-stage breast cancer, and lower negative values indicate a lower probability. The features are ordered by their importance, with the most important features at the top. From the plot, we can observe that the mean_perimeter feature has the highest positive SHAP values, indicating that higher values of mean_perimeter contribute significantly to predicting early-stage breast cancer. On the other hand, the mean_radius feature has predominantly negative SHAP values, suggesting that lower values of mean_radius are associated with a higher likelihood of early-stage breast cancer.

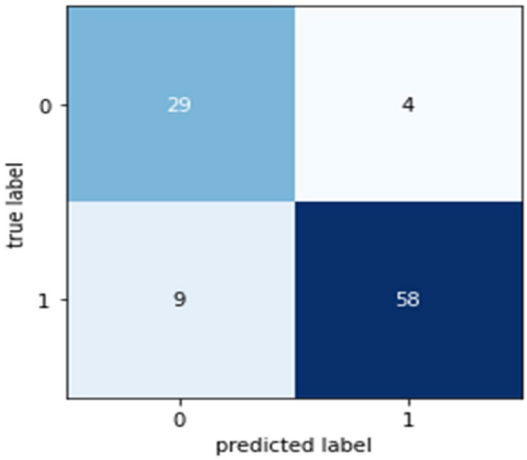

Figure 5. CM of DT.

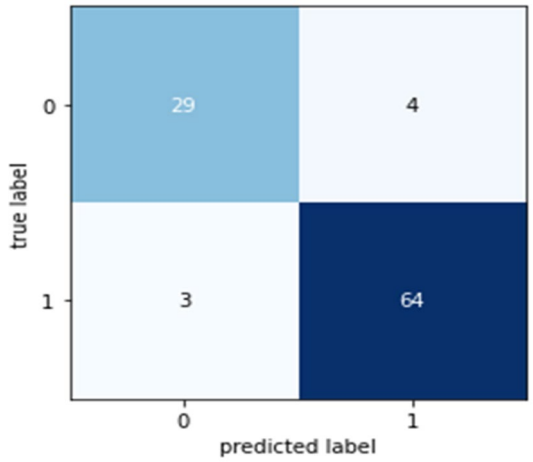

Figure 6. CM of RF.

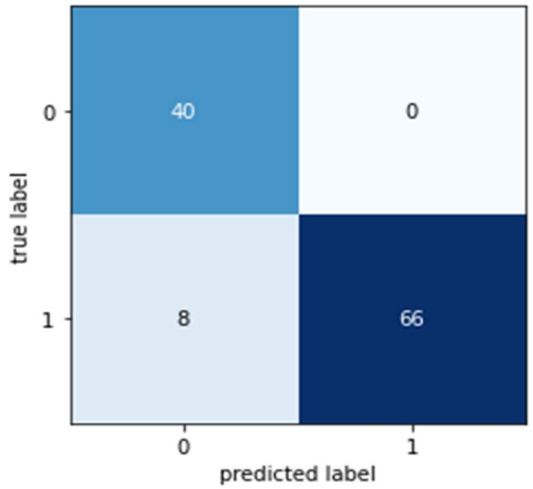

Figure 7. CM of NB.

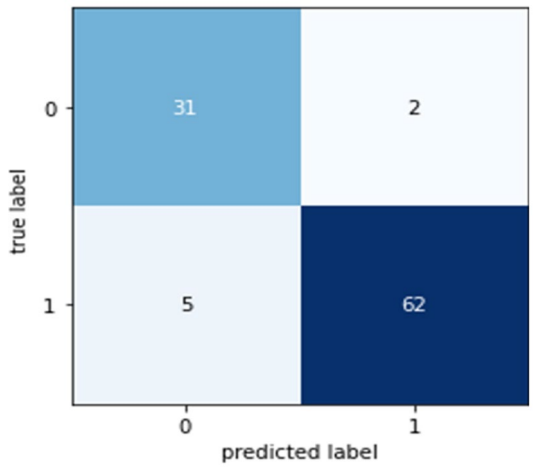

Figure 8. CM of XGB.

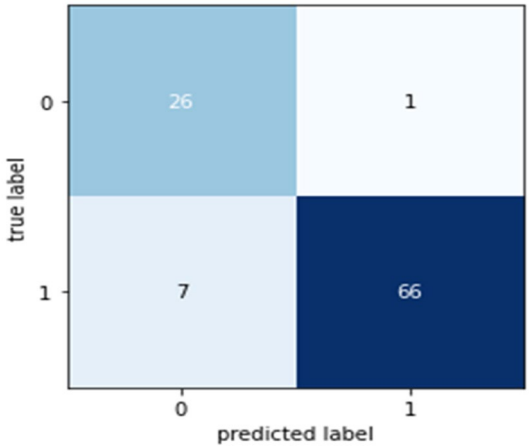

Figure 9. CM of LR.

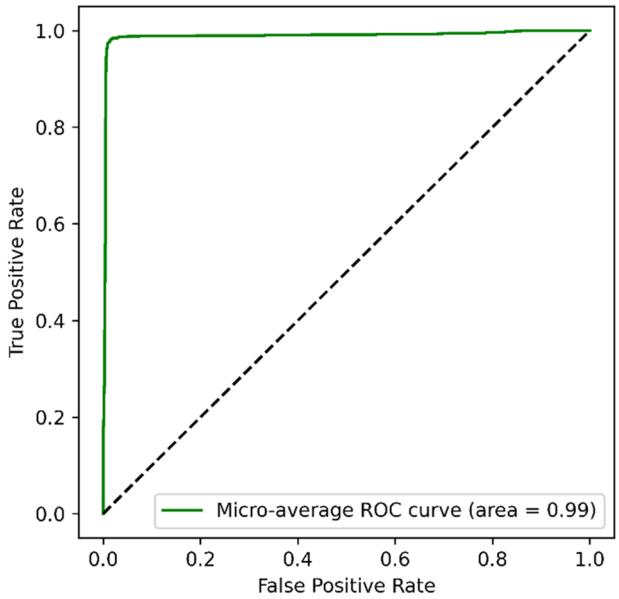

Figure 10. ROC Curve of XGBoost Model.

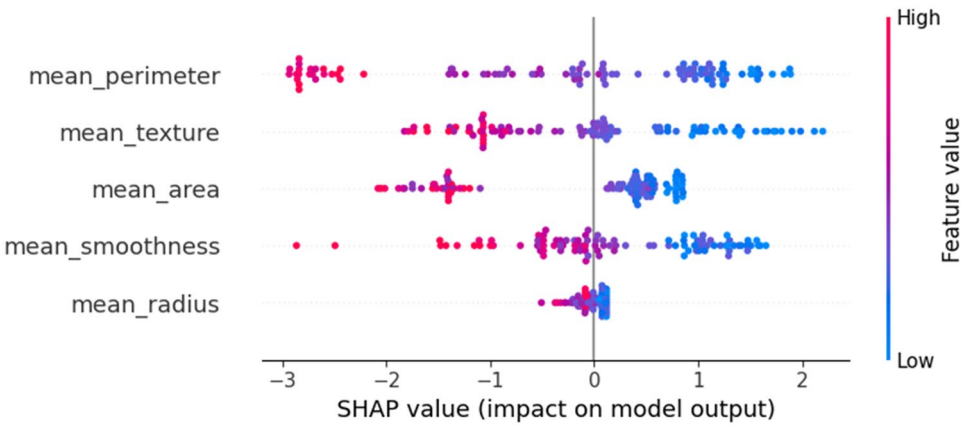

Figure 11. SHAP Summary Plot for XGBoost Model.

Figure 12 presents the SHAP dependence plots for various feature pairs, illustrating the relationship between the SHAP values and the feature values. In Fig. 12a, the dependence plot for mean_perimeter shows a weak positive correlation with the SHAP value, indicating that higher mean_perimeter values contribute to a higher probability of predicting early-stage breast cancer. The scatter plot for mean_smoothness and mean_texture in Fig. 12b suggests a weak positive correlation, implying that as mean_smoothness increases, mean_texture also tends to increase. Figure 12c shows the plot for mean_area and mean_smoothness, which exhibits no clear linear relationship, with data points scattered throughout the area, suggesting no strong correlation or causation. Interestingly, the data points for mean_smoothness and mean_area in Fig. 12d appear to exhibit a weak positive correlation, indicating that smoother surfaces tend to have larger areas, but with a fair amount of variation. Finally, Fig. 12e

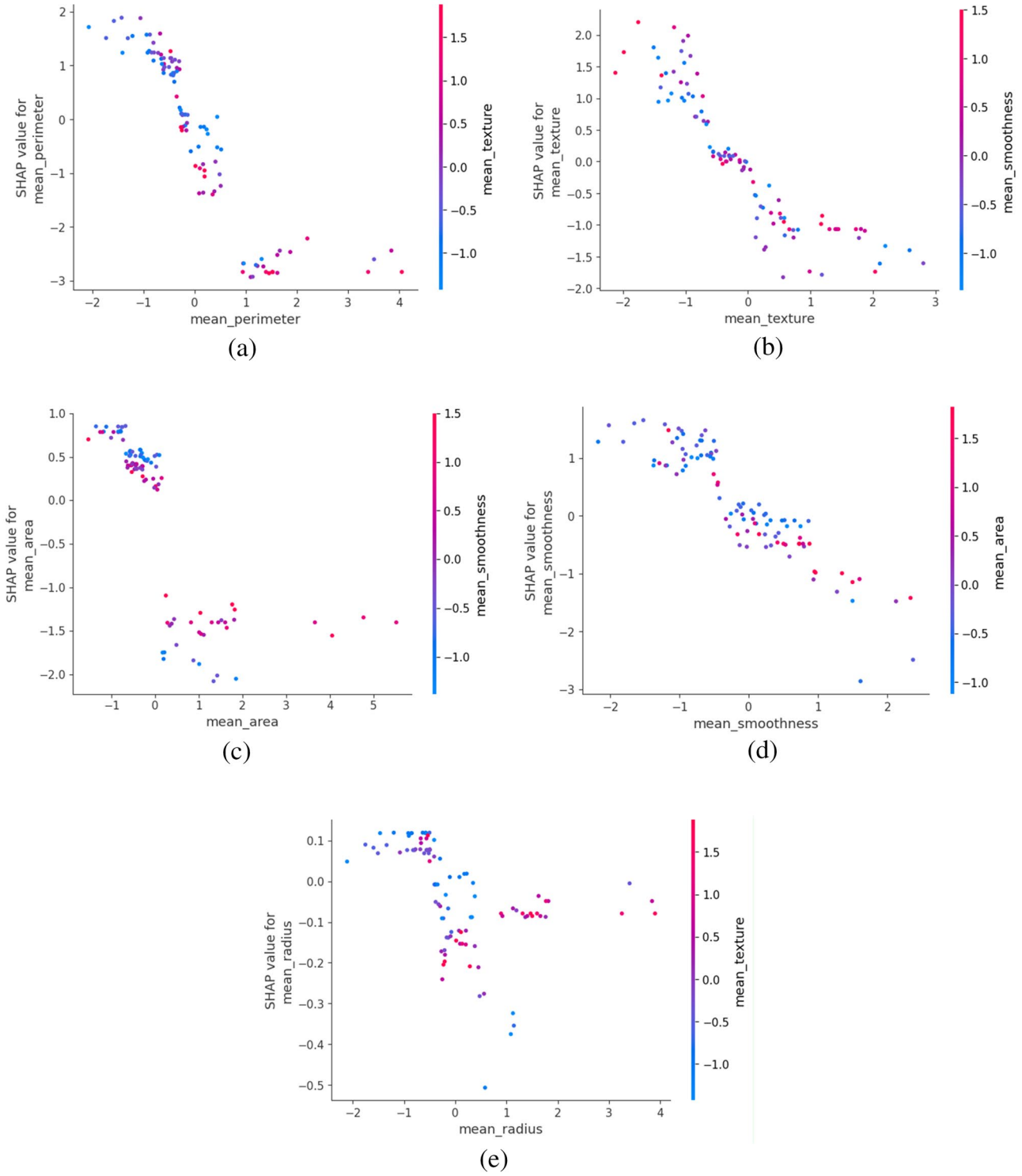

Figure 12. SHAP Dependence Plot for XGBoost Model.

displays the scatter plot for the SHAP value and mean_radius, which shows no discernible relationship, with data points scattered throughout the area, suggesting no strong correlation or causation between these two variables.

Performance analysis using cross-validation

When evaluating the transferability of statistical findings to a new dataset, researchers often employ a model validation technique known as cross-validation[52]. The result was calculated using K-fold cross-validation in this study. Using k-fold cross-validation, the dataset is split up into k smaller subsets. The remaining k1 subsets are combined for use as training samples, while the remaining subset is utilized to validate the others. The optimal value of k is dependent on the number of variables and the nature of the predictor, according to statistical theory. The sole adjustable aspect of the method is the number of subsamples (K) into which each data sample is divided. This method is typically referred to as k-fold cross-validation. For instance, k = 10 would be referred to be tenfold cross-validation if used in the model reference. Accuracy for all of the models in this study is shown in Table 3 using k-fold cross-validation.

In this study, five machine learning algorithms were implemented to determine the optimal model performance. Based on the findings presented in the results section, the XGBoost method generated the highest model accuracy of 97%. To assess the performance, the k-fold cross-validation technique was employed in this study. The detailed outcome can be observed in Table 3. The accuracy cross-validation revealed that five algorithms, namely Decision Tree, Random Forest, Logistic regression, Naive Bayes, and XGBoost, exhibited strong performance. A 90% mean accuracy score was obtained through tenfold cross-validation for the Decision Tree model, while the Random Forest model achieved a 95% mean accuracy score. In contrast, XGBoost exhibits superior performance compared to alternative algorithms. A tenfold cross-validation mean score of 97% was obtained from the algorithm, while Naive Bayes achieved a score of 92%. In the tenfold cross-validation, Logistic regression achieved a mean score of 92%. The results of the accuracy tenfold cross-validation score indicate that the quality of the model employed in this research is satisfactory.

Discussion

Accurately predicting breast cancer development is a critical objective in current research efforts. However, the current utilization of data in this particular field is still limited. ML techniques have been employed to improve the accuracy of cancer prediction. The efficiency of the categorization method employed in this study justifies its comparison with other research attempts to evaluate its public importance. The primary goal of this study was to determine the most effective machine learning methodologies for accurately evaluating breast cancer risk. It acknowledges the potentially surprising effects of the predictive capabilities within this domain.

While most research in this field focuses on a limited amount of publicly accessible datasets, resulting in comparable evaluations of algorithms, the current study aims to explore new data sources. Out of the five machine learning methods utilized in this dataset, XGBoost and RF demonstrated exceptional performance, achieving 97% and 96% accuracy. Significantly, XGBoost demonstrated superior effectiveness in handling a recently produced dataset. It is reasonable to consider that a more extensive and equally distributed dataset may result in even higher levels of precision in the long term. The Random Forest algorithm showed significant capability by achieving the second-highest accuracy in performance. The Random Forest algorithm demonstrates its superiority in managing intricate interactions and mitigating the issue of overfitting. On the other hand, XGBoost exhibits remarkable abilities in enhancing model performance by utilizing gradient-boosting approaches. Upon evaluating their respective performances within the context of our study, it is evident that both algorithms contribute substantially to the accuracy of predictions. The ability of Random Forest to effectively handle different data features is a valuable complement to XGBoost's capability to capture complicated patterns. These unique traits improve our comprehension of breast cancer prediction. However, it is essential to acknowledge that these methods have certain disadvantages, including sensitivity to noisy data and computational complexity. When considering the broader effects of a study, it is essential to acknowledge that the generalizability of the results may be influenced in ways such as the size and diversity of the dataset. Reliability could be enhanced by ensuring a more balanced and broader dataset. This discourse highlights the considerable prospects of these algorithms while also acknowledging their limitations and recognizing the need for further refinement of prediction models in Table 4.

Table 4 summarizes several research papers related to the prediction and diagnosis of breast cancer using various machine learning and deep learning techniques. Each paper addresses different aspects of breast cancer detection and prediction, and they vary in terms of accuracy, applied algorithms, and limitations.

Yu et al. introduce a nine-layer Convolutional Neural Network (CNN) for identifying abnormal breast tissue in mammography images. Their method achieves a sensitivity of 93.4% and specificity of 94.6%, showing good performance for image-based breast cancer detection. The limitation mentioned is a potential for further enhancing the method's performance. Parampreet et al. propose a Bayesian hyperparameter optimization technique for stacked ensemble models, achieving good Area under the Curve (AUC) scores for various datasets. They use machine learning models like Deep Neural Network (DNN), Gradient Boosting Machine (GBM), and Distributed Random Forest (DRF). The limitation here is the availability of a limited dataset. Alessandro et al. focus on automated image analysis of breast ultrasound images, achieving an accuracy of 91% using ensemble methods with CNN architectures. The limitation is the potential for improving accuracy. Hiba et al. primarily employ machine learning algorithms for breast cancer risk prediction and diagnosis, with SVM yielding the highest accuracy of 97.13%. They also use DT, NB, and KNN. A limitation mentioned is data pre-processing. Puja et al. use deep learning with Adam Gradient Descent Learning, achieving a high accuracy of 98.24%. They also employ various traditional machine learning models. The sources do not mention any specific limitations. Comparing this work to the others, the novelty lies in the application of multiple supervised machine learning algorithms for early-stage breast cancer prediction, providing a comprehensive approach. While the accuracy is

Algorithm	cv = 10	cv_score	cv_score (mean)
Decision tree	1	0.917718	0.907497
	2	0.923146	
	3	0.894783	
	4	0.899032	
	5	0.924925	
	6	0.90834	
	7	0.895156	
	8	0.913099	
	9	0.895104	
	10	0.903675	
Random forest	1	0.919113	0.955505
	2	0.972991	
	3	0.951697	
	4	0.940283	
	5	0.957985	
	6	0.95892	
	7	0.97818	
	8	0.957673	
	9	0.932415	
	10	0.981293	
XGBoost	1	0.972403	0.973807
	2	0.976671	
	3	0.98727	
	4	0.969338	
	5	0.967068	
	6	0.968395	
	7	0.963595	
	8	0.981956	
	9	0.988937	
	10	0.962441	
Naive Bayes	1	0.915354	0.925651
	2	0.917053	
	3	0.950712	
	4	0.914418	
	5	0.923784	
	6	0.908685	
	7	0.913652	
	8	0.921779	
	9	0.949879	
	10	0.941254	
Logistic regression	1	0.917948	0.927323
	2	0.917869	
	3	0.93195	
	4	0.929658	
	5	0.939511	
	6	0.924365	
	7	0.918985	
	8	0.935196	
	9	0.936036	
	10	0.921721	

Table 3. Evaluation of machine learning algorithms (accuracy) using k-fold cross-validation.

competitive, the acknowledgment of potential for improvement shows a commitment to enhancing the model's performance. Additionally, by leveraging various algorithms, this research showcases versatility and adaptability,

Ref.	Main idea of the paper	Accuracy	Applied algorithms	Limitations
This Work	Proposed supervised machine learning algorithms to predict early-stage breast cancer	Decision Tree: 91%, Random Forest: 96%, XGBoost:97%, Naïve Bayes: 94%, Logistic Regression: 93%	Decision Tree Random Forest XGBoost Naïve Bayes Logistic Regression	Need more analysis with models
[53]	The paper proposes an improved nine-layer convolutional neural network (CNN) for identifying abnormal breast in mammography using the open-access mini MIAS dataset	The proposed method achieved a sensitivity of 93.4%, specificity of 94.6%, precision of 94.5%, and accuracy of 94.0% on the test set	Convolutional neural network (CNN)	Could enhance the performance of the proposed method
[54]	Proposed a parallel Bayesian hyperparameter to optimize stacked ensemble models for breast cancer survival prediction	BSense model 83.9%, 87.3%, 91.1%, and 80.1% Area Under Curve (AUC) for TCGA, METABRIC, Metabolomics, and RNA-seq dataset, respectively	Stacking of machine learning models, i.e., Deep Neural Network (DNN), Gradient Boosting Machine (GBM), and Distributed Random Forest (DRF)	Limited dataset
[55]	Outline a complete automated process using advanced computer techniques to identify and categorize structures within breast ultrasound images	Achieved 91% accuracy in the classification	Ensembles methods to combine the performance of the individual CNNs architectures	Accuracy could be better
[56]	The paper focuses on using machine learning algorithms for breast cancer risk prediction and diagnosis	SVM has the highest accuracy of 97.13% with the lowest error rate	Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB), and k Nearest Neighbors (k-NN) algorithms	Data Pre-Processing
[36]	The paper focuses on predicting breast cancer using machine learning models and varying parameters	Deep learning using Adam Gradient Descent Learning is 98.24%	k-Nearest Neighborhood, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine	The provided sources do not mention any specific limitations of the paper

Table 4. Comparative analysis of different published studies.

which can be advantageous in real-world clinical settings where different types of data may be available. This approach makes it a robust and promising breast cancer prediction and diagnosis solution. However, it's essential to continue refining the model to achieve even higher accuracy rates and overcome its limitations.

Limitations and future work

The study although presents a promising result, but it has several limitations. As we have already discussed, the dataset has used here was limited in size and diversity, potentially limiting the generalizability of the findings. Due to the sensitive nature of patient data and the challenges associated with data collection, we were constrained in the amount of data we could gather for this study. Additionally, the features provided in the dataset were extracted image properties rather than raw scans, restricting direct image-based analysis. In future we are aiming to collect a large number of patient dataset from Bangladesh and United States and applying our method to get the comparative result. We plan to collaborate with multiple healthcare institutions to gather a more extensive and diverse dataset, incorporating data from different demographics and regions to improve model generalizability. Additionally, we will explore newer machine learning algorithms, deep learning techniques, or ensemble methods to enhance the accuracy and efficiency of our models. Conducting in-depth analyses on feature importance and selection methods will provide valuable insights into the key factors influencing breast cancer prediction. Furthermore, we intend to validate our models in real clinical settings through prospective studies, working closely with healthcare professionals for validation.

Conclusion

This study demonstrates the potential of using supervised machine learning algorithms for early prediction of breast cancer. This research has collected 500 patients primary data from Dhaka Medical College Hospital and applied five supervised machine learning algorithms. In the evaluation this research found that XGBoost achieved the highest accuracy of 97%. XGBoost also achieved the highest precision (0.94), recall (0.95), and F1 score (0.96) rather than other algorithms. However, some limitations need to be acknowledged. More extensive real-world data is required to confirm the model's generalization capability across larger populations. Furthermore, the dataset only included derived picture attributes rather than raw scans, which limited the possibility of conducting direct image-based analysis. Despite these constraints, this work illustrates an important proof-of-concept for leveraging artificial intelligence to improve breast cancer diagnosis. The model can enable clinicians to rapidly screen patients and identify high-risk cases needing further examination. This would significantly impact early intervention and tailored treatment planning to improve survival outcomes. As next steps, integrating medical imagery into the pipeline and validating performance over larger multi-center datasets could help strengthen model robustness. It would also be valuable to experiment with combining machine predictions with expertise of oncologists to develop an augmented diagnostics system. Such human-AI collaboration can lead to more accurate and transparent cancer care.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 12 September 2023; Accepted: 21 March 2024
Published online: 11 April 2024

References

1. Park, M. Y. *et al.* Function and application of flavonoids in the breast cancer. *Int. J. Mol. Sci.* **23**, 7732 (2022).
2. (1) (PDF) Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. https://www.researchgate.net/publication/361228083_Breast_cancer_detection_based_on_thermographic_images_using_machine_learning_and_deep_learning_algorithms.
3. Uddin, K. M. M., Biswas, N., Rikta, S. T. & Dey, S. K. Machine learning-based diagnosis of breast cancer utilizing feature optimization technique. *Comput. Methods Progr. Biomed. Update* **3**, 100098 (2023).
4. Adekeye, A., Lung, K. C. & Brill, K. L. Pediatric and adolescent breast conditions: A review. *J. Pediatr. Adolesc. Gynecol.* **36**, 5–13 (2023).
5. Siegel Mph, R. L. *et al.* Cancer statistics, 2023. *pathologyinnovationcc.orgRL Siegel, KD Miller, NS Wagle, A JemalCa Cancer J Clin, 2023•pathologyinnovationcc.org* **73**, 17–48 (2023).
6. Akter, S. *et al.* Recent advances in ovarian cancer: Therapeutic Strategies, potential biomarkers, and technological improvements. *Cells* **11**, 650 (2022).
7. Tsochatzidis, L., Costaridou, L. & Pratikakis, I. Deep learning for breast cancer diagnosis from mammograms—A comparative study. *J. Imaging* **5**, 37 (2019).
8. Mahesh, T. R. *et al.* An efficient ensemble method using K-fold cross validation for the early detection of benign and malignant breast cancer. *Int. J. Integr. Eng.* **14**, 204–216 (2022).
9. Sheakh, M. A. *et al.* Child and maternal mortality risk factor analysis using machine learning approaches. In *ISDFS 2023—11th International Symposium on Digital Forensics and Security*. https://doi.org/10.1109/ISDFS58141.2023.10131826 (2023).
10. Ermakov, M. S., Nushtaeva, A. A., Richter, V. A. & Koval, O. A. Опухоль-ассоциированные фибробласты и их роль в опухолевой прогрессии. *Вавиловский журнал генетики и селекции* **26**, 14–21 (2022).
11. Lei, L., Ma, B., Xu, C. & Liu, H. Emerging tumor-on-chips with electrochemical biosensors. *TrAC Trends Anal. Chem.* **153**, 116640 (2022).
12. Boutry, J. *et al.* The evolution and ecology of benign tumors. *Biochim. Biophys. Acta (BBA) Rev. Cancer* **1877**, 188643 (2022).
13. Tadesse, A., Tafa Segni, M. & Demissie, H. F. Knowledge, attitude, and practice (KAP) toward cervical cancer screening among adama science and technology university female students, Ethiopia. *Int. J. Breast Cancer* https://doi.org/10.1155/2022/2490327 (2022).
14. Szczepski, K. *et al.* Metabolic biomarkers in cancer. *Metabol. Path Towards Pers. Med.* https://doi.org/10.1016/B978-0-323-99924-3.00005-4 (2023).
15. Srivani, M., Murugappan, A. & Mala, T. Cognitive computing technological trends and future research directions in healthcare—A systematic literature review. *Artif. Intell. Med.* **138**, 102513 (2023).
16. Rathore, A. S. *et al.* Erythemato-squamous diseases prediction and interpretation using explainable AI. *IETE J. Res.* https://doi.org/10.1080/03772063.2022.2114953 (2022).
17. Noninvasive hemoglobin sensing and imaging: optical tools for disease diagnosis. https://www.spiedigitallibrary.org/journals/journal-of-biomedical-optics/volume-27/issue-08/080901/Noninvasive-hemoglobin-sensing-and-imaging--optical-tools-for-disease/https://doi.org/10.1117/1.JBO.27.8.080901.full?&webSyncID=7620c89a-0ce4-6e9e-6ec7-a49dab6a0cba&sessionGUID=d059329a-d883-c9d9-02bc-9993ced268be#_=_.
18. Giaquinto, A. N. *et al.* Breast cancer statistics, 2022. *CA Cancer J. Clin.* **72**, 524–541 (2022).
19. Gophika, T., Sudha, S. & Ranjana, M. R. Introduction to Translating healthcare through intelligent computational methods. In *EAI/Springer Innovations in Communication and Computing* Part F282 3–17 (2023).
20. Bevilacqua, G. The viral origin of human breast cancer: From the mouse mammary tumor virus (MMTV) to the human betaretrovirus (HBRV). *Viruses* **14**, 1704 (2022).
21. Richards, G., Rayward-Smith, V. J., Sönksen, P. H., Carey, S. & Weng, C. Data mining for indicators of early mortality in a database of clinical records. *Artif. Intell. Med.* **22**, 215–231 (2001).
22. Djebbari, A., Liu, Z., Phan, S. & Famili, F. An ensemble machine learning approach to predict survival in breast cancer. *Int. J. Comput. Bio.l Drug Des.* **1**, 275–294 (2008).
23. Aruna, S., Rajagopalan, S. P., Nandakishore, L. V. & In, S. C. Knowledge based analysis of various statistical tools in detecting breast cancer. *Comput. Sci. Inf. Technol.* https://doi.org/10.5121/csit.2011.1205 (2011).
24. Agarap, A. F. M. On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset. In *ACM International Conference Proceeding Series* 5–9. https://doi.org/10.1145/3184066.3184080 (2018).
25. Toprak, A. Extreme learning machine (ELM)-based classification of benign and malignant cells in breast cancer. *Med. Sci. Monit.* **24**, 6537 (2018).
26. Thomas, T., Pradhan, N. & Dhaka, V. S. Comparative analysis to predict breast cancer using machine learning algorithms: A survey. In *Proceedings of the 5th International Conference on Inventive Computation Technologies, ICICT 2020* 192–196. https://doi.org/10.1109/ICICT48043.2020.9112464 (2020).
27. Livingston, F. Implementation of Breiman's random forest machine learning algorithm. *ECE591Q Mach. Learn. J. Pap. Fall* (2005).
28. Mitchell, R. & Frank, E. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Comput. Sci.* **3**, e127 (2017).
29. Ak, M. F. A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. *Healthcare* **8**, 111 (2020).
30. Islam, M. M. *et al.* Breast cancer prediction: A comparative study using machine learning techniques. *SN Comput. Sc.i* **1**, 1–14 (2020).
31. Chaurasia, V. & Pal, S. Applications of machine learning techniques to predict diagnostic breast cancer. *SN Comput. Sci.* **1**, 1–11 (2020).
32. Kabiraj, S. *et al.* Breast cancer risk prediction using XGBoost and random forest algorithm. In *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*. https://doi.org/10.1109/ICCCNT49239.2020.9225451 (2020).
33. Jabbar, M. A. Breast cancer data classification using ensemble machine learning. *Eng. Appl. Sci. Res.* **48**, 65–72 (2021).
34. Shalini, M. & Radhika, S. Machine learning techniques for prediction from various breast cancer datasets. In *2020 6th International Conference on Bio Signals, Images, and Instrumentation, ICBSII 2020*. https://doi.org/10.1109/ICBSII49132.2020.9167657 (2020).
35. Naji, M. A. *et al.* Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Comput. Sci.* **191**, 487–492 (2021).
36. Gupta, P. & Garg, S. Breast cancer prediction using varying parameters of machine learning models. *Procedia Comput. Sci.* **171**, 593–601 (2020).
37. Mustapha, M. T., Ozsahin, D. U., Ozsahin, I. & Uzun, B. Breast cancer screening based on supervised learning and multi-criteria decision-making. *Diagnostics* **12**, 1326 (2022).
38. Sun, R., Hou, X., Li, X., Xie, Y. & Nie, S. Transfer learning strategy based on unsupervised learning and ensemble learning for breast cancer molecular subtype prediction using dynamic contrast-enhanced MRI. *J. Magn. Reson. Imaging* **55**, 1518–1534 (2022).

39. Hasan, M., Tahosin, M. S., Farjana, A., Sheakh, M. A. & Hasan, M. M. A harmful disorder: Predictive and comparative analysis for fetal Anemia disease by using different machine learning approaches. *ISDFS 2023—11th International Symposium on Digital Forensics and Security*. https://doi.org/10.1109/ISDFS58141.2023.10131838 (2023).
40. Safavian, S. R. & Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* **21**, 660–674 (1991).
41. Khosravi, P., Vergari, A., Choi, Y., Liang, Y. & Broeck, G. Van den. *Handling Missing Data in Decision Trees: A Probabilistic Approach* (2020).
42. Rigatti, S. J. Random forest. *J. Insur. Med.* **47**, 31–39 (2017).
43. Zhu, H., Liu, H., Zhou, Q. & Cui, A. Towards an accurate and reliable downscaling scheme for high-spatial-resolution precipitation data. *Remote Sens.* **15**, 2640 (2023).
44. Liu, P. *et al.* Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer. *IEEE Trans. Biomed. Eng.* **68**, 148–160 (2021).
45. Webb, G. I. Naïve Bayes. *Encycl. Mach. Learn. Data Min.* https://doi.org/10.1007/978-1-4899-7502-7_581-1 (2016).
46. Nurhasanah, N., Sumarly, D. E., Pratama, J., Heng, I. T. & Irwansyah, E. Comparing SVM and Naïve Bayes classifier for fake news detection. *Eng. Math. Comput. Sci. J. (EMACS)* **4**, 103–107 (2022).
47. Romli, I. *et al.* Classification of breast cancer using Wrapper and Naïve Bayes algorithms. *J. Phys. Conf. Ser.* **1040**, 012017 (2018).
48. Hilbe, J. M. *Logistic Regression Models* 658.
49. Islam, T. *et al.* Review analysis of ride-sharing applications using machine learning approaches Bangladesh perspective. *Comput. Stat. Methodol. Model. Artif. Intell.* https://doi.org/10.1201/9781003253051-7 (2023).
50. Islam, M. T. *et al.* Convolutional neural network based partial face detection. In *2022 IEEE 7th International Conference for Convergence in Technology, I2CT 2022*. https://doi.org/10.1109/I2CT54291.2022.9825259 (2022).
51. Islam, T. *et al.* A proposed Bi-LSTM method to fake news detection. In *2022 International Conference for Advancement in Technology, ICONAT 2022*. https://doi.org/10.1109/ICONAT53423.2022.9725937 (2022).
52. Tahosin, M. S., Sheakh, M. A., Islam, T., Lima, R. J. & Begum, M. Optimizing brain tumor classification through feature selection and hyperparameter tuning in machine learning models. *Inform. Med. Unlocked* **43**, 101414 (2023).
53. Zhang, Y. D., Pan, C., Chen, X. & Wang, F. Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling. *J. Comput. Sci.* **27**, 57–68 (2018).
54. Kaur, P., Singh, A. & Chana, I. BSense: A parallel Bayesian hyperparameter optimized Stacked ensemble model for breast cancer survival prediction. *J. Comput. Sci.* **60**, 101570 (2022).
55. Podda, A. S. *et al.* Fully-automated deep learning pipeline for segmentation and classification of breast ultrasound images. *J. Comput. Sci.* **63**, 101816 (2022).
56. Asri, H., Mousannif, H., Al Moatassime, H. & Noel, T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Comput. Sci.* **83**, 1064–1069 (2016).

Acknowledgements

The authors would like to extend their sincere appreciation to the Researchers Supporting Project, King Saud University, Riyadh, Saudi Arabia for funding this work through the project number (RSP-2024R457)

Author contributions

Conceptualization: T.I., M.A.S., M.S.T., and S.A. Methodology: T.I., M.A.S., M.S.T. Validation: T.I., M.A.S., and Y.A.B.J. Formal analysis: T.I., Y.A.B.J., G.F.W., and H.A.N. Writing—original draft preparation: T.I., M.A.S., and M.S.T. Writing—review and editing: T.I., M.A.S, M.S.T., Y.A.B.J., M.B., and G.F.W. Supervision: M.H.H., and M.B. Resources: T.I., M.A.S., M.H.H. Project administration: M.B., and T.I.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to G.F. or M.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.