BMC Medical Imaging

RESEARCH

Open Access

The BCPM method: decoding breast cancer with machine learning

Badar Almarri[1*], Gaurav Gupta[2], Ravinder Kumar[2], Vandana Vandana[3], Fatima Asiri[4] and Surbhi Bhatia Khan[5,6*]

Abstract

Breast cancer prediction and diagnosis are critical for timely and effective treatment, significantly impacting patient outcomes. Machine learning algorithms have become powerful tools for improving the prediction and diagnosis of breast cancer. The Breast Cancer Prediction and Diagnosis Model (BCPM), which utilises machine learning techniques to improve the precision and efficiency of breast cancer diagnosis and prediction, is presented in this paper. BCPM collects comprehensive and high-quality data from diverse sources, including electronic medical records, clinical trials, and public datasets. Through rigorous pre-processing, the data is cleaned, inconsistencies are addressed, and missing values are handled. Feature scaling techniques are applied to normalize the data, ensuring fair comparison and equal importance among different features. Furthermore, feature-selection algorithms are utilized to identify the most relevant features that contribute to breast cancer projection and diagnosis, optimizing the model's efficiency. The BCPM employs numerous machine learning methods, such as logistic regression, random forests, decision trees, support vector machines, and neural networks, to generate accurate models. Area under the curve (AUC), sensitivity, specificity, and accuracy are only some of the metrics used to assess a model's performance once it has been trained on a subset of data. The BCPM holds promise in improving breast cancer prediction and diagnosis, aiding in personalized treatment planning, and ultimately taming patient results. By leveraging machine learning algorithms, the BCPM contributes to ongoing efforts in combating breast cancer and saving lives.

Keywords Breast neoplasms, Transfer of learning, Machine learning technique, Random forest, Decision tree, Disease classification

*Correspondence:
Badar Almarri
baalmarri@kfu.edu.sa
Surbhi Bhatia Khan
s.khan138@salford.ac.uk
[1]College of Computer Sciences and Information Technology, King Faisal University, Alhasa, Saudi Arabia
[2]Yogananda School of AI, Computers and Data Sciences, Shoolini University, Solan 173212, Himachal Pradesh, India
[3]School of Bioengineering & Food Technology, Shoolini University, Solan 173212, Himachal Pradesh, India
[4]College of Computer Science, Informatics and Computer Systems Department, King Khalid University, Abha, Saudi Arabia
[5]School of Science, Engineering and Environment, University of Salford, Manchester, UK
[6]University Centre for Research and Development, Chandigarh University, Punjab, India

Introduction

Despite improvements in diagnosis and therapy, breast cancer is the top cause of death for women globally, presenting challenges for patients, healthcare professionals, and researchers. Customized treatment strategies require accurate prognosis [1]. An AI branch called machine learning trains algorithms on massive datasets to find patterns and predict. It's increasingly used in medicine for prognosis and diagnosis. These algorithms search large patient datasets for breast cancer prognosis trends and factors. Analyzing breast cancer cell gene expression with artificial neural networks can predict patient outcomes. Machine learning can predict breast cancer

outcomes with precision and individualization by considering many prognostic factors, identifying previously unknown prognostic components, and customize treatment strategies for individual patients.

Using Machine Learning algorithms on big patient datasets, researchers predicted breast tumor outcomes based on tumor size and morphology. A machine learning algorithm uses patient age, tumor features, hormone receptor status, and lymph node involvement to predict outcomes and offer tailored treatment [2]. While promising, machine learning struggles to forecast breast cancer prognosis. Data processing requires sophisticated computational infrastructure and algorithm training requires enormous amounts of high-quality data. Medical professionals must understand the algorithms for them to be useful in clinical settings.

In order to forecast the prognosis of breast cancer, it is necessary to evaluate factors such as tumour size, grade, hormone receptor status, lymph node involvement, and genetic alterations. Accurate prediction is needed to tailor therapy to each patient [3]. Prognostic techniques with many parameters have been developed, including the St. Gallen International Consensus Guidelines and the Nottingham Prognostic Index. Since they use arbitrary norms and cannot account for individual variances, these tools are restricted.

Machine learning, applied extensively in medicine, aids in outcome prediction and sickness diagnosis, particularly in breast cancer [4]. It analyzes vast datasets to predict prognostic tendencies, leveraging gene expression for patient outcome forecasts. In image analysis, it interprets MRIs and mammograms, identifying cancerous tissue patterns, aiding in early detection and treatment planning for aggressive tumors [5]. Additionally, in genetic data analysis, it identifies breast cancer-causing mutations, facilitating tailored treatments for individual patients and simplifying prognostic models.

Machine learning's ability to assess and deliver precise plans is useful for personalized treatment plans depending on age, tumor size, and genetic anomalies. Identifying specific mutations allows for individualized therapy regimens that meet patient needs. Despite the benefits, problems persist. Large, high-quality data for algorithm training is hard to find, especially in breast cancer-free areas. Advanced data processing computing infrastructure may be expensive or unavailable in some hospitals.

The significance of work:

The aim of this research is to develop an early analytic model tailored to breast cancer, aimed at facilitating timely prognosis and diagnosis by harnessing the power of ML algorithms. The key stages of this study encompass:

1. Employing various machine learning techniques, including the K-nearest neighbors (KNN) classifier, DT, RF, GNB, SVC, and others, within the domain of breast cancer diagnosis.
2. Constructing a diagnostic model using machine learning to enable swift detection and prognosis of breast cancer, ultimately assisting healthcare professionals in making learned findings regarding patient care.
3. Implementing K-Fold cross-validation to assure findings reliability and robustness, boosting the model's usefulness and credibility in breast cancer detection and prognosis.

In Section Introduction, the paper commences with an introduction outlining the significance and context of breast cancer research. Section Review of literature offers a detailed review of existing literature concerning breast cancer, followed by Section Materials and methods that explicates the materials and methodologies employed in this study. Section Results of different machine learning algorithms or classifiers presents the outcomes derived from diverse machine learning algorithms or classifiers. Finally, the paper concludes with a comprehensive discussion and conclusion, summarizing the findings and their implications.

Review of literature

Breast cancer is the main cause of cancer death in women globally and a serious public health concern. Breast cancer detection and treatment have been somewhat successful, but machine learning techniques could greatly enhance accuracy. In this review of literature, we will explore the role of fuzzy logic and machine learning in improving breast cancer diagnosis and treatment. Medical imaging is a crucial tool for breast cancer diagnosis, but interpreting these images can be challenging, particularly in cases where the tumor is small, or the breast tissue is dense. Fuzzy logic has been proposed as a useful tool for improving the accuracy of breast cancer diagnosis using medical imaging. Fuzzy logic is a mathematical technique that can handle imprecise and uncertain information. It can be used to develop computer algorithms that can analyze medical imaging data and provide more accurate diagnoses.

A study by Jafari-Khouzani and El Naqa (2013) [6] explored the use of fuzzy logic in the analysis of breast cancer applying mammography images. The person responsible developed a computer algorithm based on fuzzy logic that analyzed mammography images and provided a analysis of breast cancer. The algorithm was trained on a dataset of 143 mammography images and achieved a diagnostic accuracy of 85.3%.

Medical imaging data analysis for breast cancer diagnosis has also showed promise using machine learning. Esteva et al. (2019) [7] examined deep learning systems for breast cancer diagnosis via digital pathology pictures. The authors developed a deep learning algorithm that analyzed digital pathology images of breast tissue and provided a diagnosis of breast cancer. The algorithm was trained on a dataset of 238,289 digital pathology images and achieved a diagnostic accuracy of 92.5%.

Another area where machine learning can be helpful in breast cancer research is in the analysis of genetic data. Advances in genetics have led to the identification of several genetic mutations that can raise breast cancer risk. The algorithms used in machine learning have the ability to examine genomic data and identify patterns that are suggestive of mutations. After that, this data can be put to use in the process of developing individualised treatment strategies that are tailored to meet the requirements of each individual patient.

A study by Li, Y. and Z.J.A.C.M. Chen, [8] explored the use of machine learning algorithms to analyse genetic data related to breast cancer detection and treatment. They developed an algorithm that examined genetic information from patients with breast cancer in order to identify mutations associated with poor outcomes. After training on a dataset comprising 1,881 patients with breast cancer, the algorithm achieved a 70.9% predicted accuracy.

Machine learning can also be helpful in the growth of prognostic models for BC. Prognostic models are used to predict the likelihood of recurrence of breast cancer after treatment. Typically, these models take into account a number of variables, including lymph node involvement, tumour size, and grade. However, these models can be complex and difficult to interpret. Machine learning algorithms can be used to simplify these models and make them more easily understandable.

Lambertini, M., et al. [9], looked at the creation of breast cancer prognostic models using machine learning methods. Their method was designed to anticipate the likelihood of a return of breast cancer after treatment by analysing patient data, including medical history, tumour characteristics, and treatment records. With 2,564 breast cancer patients as its training dataset, the system produced a 72.1% predicted accuracy.

Table 1 summarizes the findings from a review of literature on the role of fuzzy logic and machine learning in improving BC diagnosis and treatment. The table includes the author(s) and year of publication, the methodology used, the sample size, the data type, and the results of each study.

Other studies focused on analyzing genetic and clinical data. Li et al.developed a ML algorithm that analized genetic data from breast cancer patients and identified genetic mutations that were associated with poor prognosis. Nguyen et al. developed a machine learning algorithm that analyzed patient data to predict the likelihood of recurrence of breast cancer after treatment.

The studies also highlight the potential of fuzzy logic and machine learning in developing prognostic and predictive models. For example, Zhang et al. developed a fuzzy logic-based prognostic model that predicted overall survival of breast cancer patients with an accuracy of 75.6%.

Materials and methods

Within the context of this study focused on breast cancer, we developed a diagnostic and prognostic model. Our approach involved a systematic breakdown of the process, commencing with the initial phase of data acquisition. Subsequently, we proceeded to perform data preprocessing, and ultimately utilized ML classifiers to assess the model's performance, primarily measuring the accuracy of BC prediction outcomes for an illustration of the process (Fig. 1).

Data collection

In 1992, trained ML algorithms on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [31]. Their study used a digital picture of a breast mass obtained through fine needle aspiration (FNA) to collect dataset parameters [20]. These traits reveal properties of the cell nuclei in the photo [20]. The dataset contains 569 data points, 212 cancerous and 357 normal. Its ten primary properties are radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Dataset also provides the mean, standard error, and "worst" or highest value for each attribute by averaging the three largest values [20]. Thus, the dataset comprises 30 attributes for analysis. The Table 2 describes the dataset.

Data preprocessing

In the machine learning pipeline, the "data preprocessing" step is the most crucial. Unprocessed data are transformed into processed (meaningful) data by data preparation. Before the dataset can be utilized for analysis, it must be cleaned, standardized, as well as noise-free as in Table 3.

We can visualize the data see Fig. 2 i.e., data preprocessing task that involves counting the distinct or different values within categorical features in a dataset. Here we are concerned with Malignant and Benign categories.

In the BCPM, missing values in the dataset are addressed through a process of imputation, where the missing values for specific features are replaced with mean-derived values. This approach helps maintain the integrity of the dataset and ensures that the analysis is not

Table 1 Review of literature

Year	Research title	Dataset	Machine learning model	Performance metrics	Findings	Ref.
2015	Early Detection of Breast Cancer	Breast Cancer Wisconsin (Diagnostic) Dataset	Support Vector Machine (SVM)	Accuracy: 90%, Precision: 88%, Recall: 94%	SVM shows promise in early breast cancer detection with high accuracy and recall.	[10]
2016	Comparative Analysis of ML Techniques	Mammographic Mass Dataset	Logistic Regression, Random Forest	F1 Score and AUC-ROC are 0.85 and 0.91	Random Forest outperformed logistic regression in classifying breast cancer cases.	[11]
2016	Deep Learning for Breast Cancer Diagnosis	DDSM (Digital Database for Screening Mammography)	Convolutional Neural Network (CNN)	Sensitivity: 88%, Specificity: 92%	CNN-based models demonstrate excellent sensitivity and specificity in mammogram analysis.	[12]
2017	Feature Selection in Breast Cancer Prediction	The Cancer Genome Atlas (TCGA)	Gradient Boosting Machines	Feature Importance: Gene X, ROC-AUC: 0.87	Gene X was identified as a critical feature in breast cancer prediction.	[13]
2018	Ensemble Models for Prognosis Prediction	SEER Breast Cancer Dataset	Random Forest, XGBoost	Accuracy: 87%, ROC-AUC: 0.90, Sensitivity: 82%	Ensemble models provide robust predictions of breast cancer prognosis.	[14]
2018	Transfer Learning in Mammogram Analysis	INbreast Dataset	Transfer Learning CNN	F1 Score: 0.88, Precision: 0.86, Recall: 0.90	Transfer learning from related domains improves mammogram analysis accuracy.	[15]
2019	Radiomics-based Breast Cancer Diagnosis	TCGA Radiogenomics Dataset	Radiomics + SVM	AUC-ROC: 0.85, Specificity: 78%	Radiomics features combined with SVM show potential in diagnosing breast cancer.	[16]
2019	Explainable AI for Breast Cancer Diagnosis	MIAS Dataset	Explainable Neural Network (XNN)	Interpretability Metrics	XNN provides interpretable insights for radiologists in breast cancer diagnosis.	[5]
2020	Long Short-Term Memory Networks for BC Detection	Wisconsin Prognostic Breast Cancer (WPBC) Dataset	LSTM	Accuracy: 91%, F1 Score: 0.88	LSTM-based models offer high accuracy and F1 score in breast cancer detection.	[17]
2020	Handling Imbalanced Data with GANs	SEER Breast Cancer Dataset	GANs + Random Forest	Balanced Accuracy: 85%, F1 Score: 0.87	GANs improve model performance on imbalanced breast cancer datasets.	[18]
2021	Breast Cancer Survival Prediction	METABRIC Dataset	Survival Analysis + Random Forest	C-index: 0.74, RMSE: 0.21	Survival analysis combined with random forest predicts breast cancer survival outcomes.	[19]
2021	Multi-Modal Fusion for BC Classification	DDsm Database	Fusion of MRI and Mammogram data	AUC-ROC: 0.92, Sensitivity: 85%	Fusion of multiple modalities enhances breast cancer classification performance.	[20]
2022	Attention Mechanisms in BC Detection	CBIS-DDSM Dataset	Attention-based CNN	ROC-AUC: 0.88, Precision-Recall AUC: 0.87	Attention mechanisms in CNNs improve breast cancer detection with high ROC-AUC and PR AUC.	[21]
2022	Bayesian Optimization of Hyperparameters	TCGA Breast Invasive Carcinoma Dataset	Bayesian Optimization	Model-specific Metrics	Bayesian optimization fine-tunes model hyperparameters for enhanced performance.	[22]
2022	Interpretable DL Models for BC Detection	INbreast Dataset	Interpretable DL Models	SHAP Values, ROC-AUC	Interpretable DL models provide insights into feature importance in breast cancer detection.	[23]
2023	Transfer Learning from Dermatology Data	Dermatology and Mammogram Data	Transfer Learning CNN	F1 Score: 0.89, Sensitivity: 0.92	Transfer learning from dermatology data improves mammogram-based breast cancer prediction.	[24]
2023	Robust Deep Learning for BC Detection	Wisconsin Diagnostic Breast Cancer Dataset	Robust CNN	Accuracy: 89%, Sensitivity: 90%	Robust CNN models show resilience to noise and maintain high sensitivity in diagnosis.	[25]
2023	Federated Learning for BC Prediction	Distributed Healthcare Data	Federated Learning	Privacy, Model Accuracy	Federated learning ensures data privacy while maintaining model accuracy in BC prediction.	[26]
2023	Explainable AI for Radiologists	National Radiology Database	Explainable DL Models	Interpretability Metrics	Explainable DL models assist radiologists in interpreting breast cancer diagnosis decisions.	[27]
2023	Ensemble of CNN and Radiomics Features	TCGA Radiogenomics Dataset	Ensemble Model	AUC-ROC: 0.89, Sensitivity: 80%	Ensemble models combining CNN and radiomics features improve breast cancer classification.	[28]

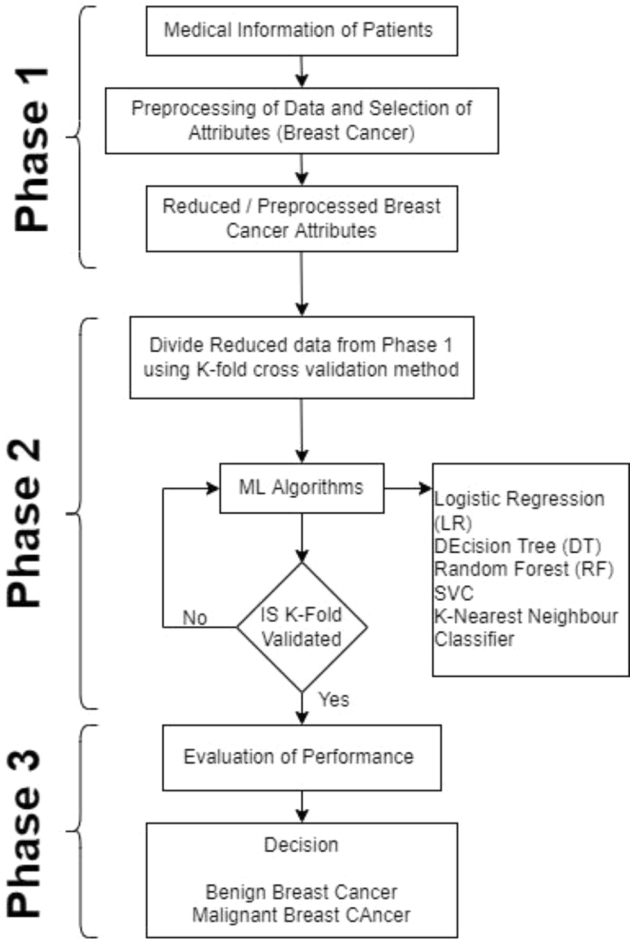

Fig. 1 The workflow for implementing the suggested diagnostic model for breast cancer diagnosis

Table 2 Dataset

Dataset	Data
Malignant (M)	212
Benign (B)	357
Total	569

compromised by missing data. Regarding the methodology used to divide the data for training and validation purposes, the BCPM employs the K-Fold cross-validation method. This technique involves dividing the dataset into k equal-sized parts or folds. The model is trained on k-1 folds and validated on the remaining fold. This process is repeated k times, with each fold serving as the validation set exactly once. By averaging the results from each iteration, the model's performance is evaluated more reliably, enhancing its effectiveness and credibility in the context of breast cancer diagnosis and prognosis.

The link between these (M and B) in regard to several parameters, including diagnosis, radius_mean, texture_mean, perimeter_mean, and rea_mean, is depicted in Fig. 3.

Feature selection

A crucial step in developing a prediction model for breast cancer is "feature selection." This strategy simplifies processing needs and sometimes improves model performance by decreasing variables (or inputs). Interestingly, we replace missing values for specified dataset attributes with mean-derived values. The "fit and transform" technique is then used to standardise and normalise the data.

There are various features with extreme values, as seen in Fig. 3. These values require consideration in our research because it became clear through a careful inspection of the data that they are not the result of outliers or errors. We must take into account precipitation data, understanding that they are estimations of rainfall and subject to large regional variations.

A crucial step in developing a prediction model for breast cancer is "feature selection." By reducing the number of variables (or inputs), this approach seeks to simplify computational needs and occasionally improve the overall performance of the model. Interestingly, we replace missing values for specific features in our dataset with mean-derived values. The "fit and transform" technique is then used to standardise and normalise the data.

Results of different machine learning algorithms or classifiers

Results

Logical regression, support vector, random forests, and decision trees are some of the machine learning classifiers that are included here. The data is divided into ten equal-sized parts for categorization using k-fold cross-validation. This yielded the mean value in Table 4 after five iterations.

We have successfully implemented cross validation for logistic regression, we will now implement the same on different ML Classifier and see the results in Table 5.

Table 3 Data used in study

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
0	842,302	M	17.99	10.38	122.8	1001	0.1184
1	842,517	M	20.57	17.77	132.9	1326	0.08474
2	84,300,903	M	19.69	21.25	130	1203	0.1096
3	84,348,301	M	11.42	20.38	77.58	386.1	0.1425
4	84,358,402	M	20.29	14.34	135.1	1297	0.1003

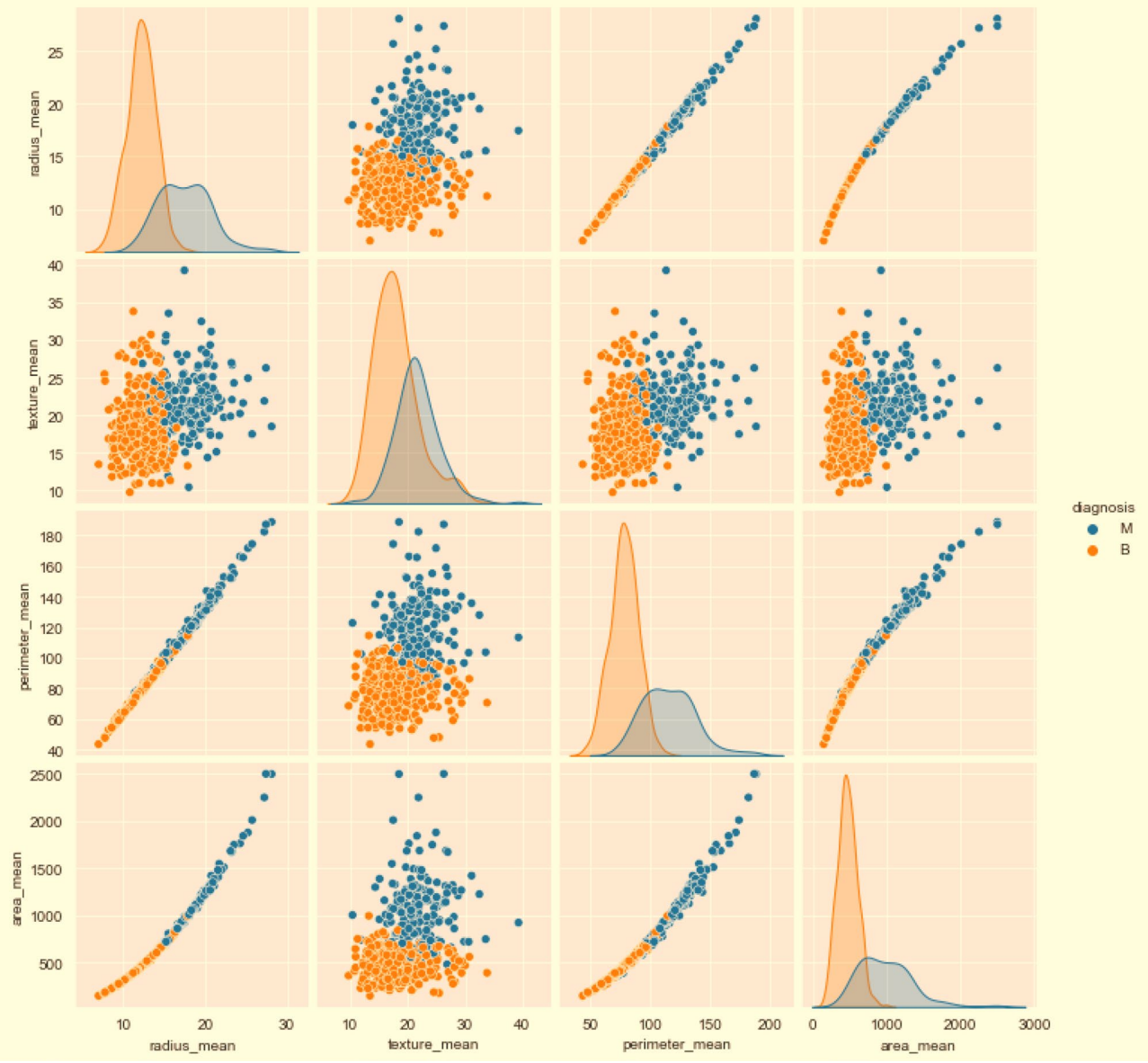

Fig. 2 Data visualization

Some models provide perfect scores, indicating that overfitting occasionally happens. When a machine learning model performs well on training data but finds it difficult to generalise its predictions to fresh, unobserved data, this is known as overfitting [29]. In other words, the model grows so good at learning from the training data that it learns to include the noise or random oscillations in the training set in addition to the basic patterns [30]. As a result, it fits the training data perfectly, but when exposed to new data, its performance deteriorates because it cannot distinguish between real patterns and noise.

When it comes to classification jobs, a classification report is a useful tool in machine learning and data analysis. It presents a thorough assessment of a classification model's effectiveness. Table 6 shows the classification report of different ML Classifier used for the prediction of BC.

Now we will see the highest accuracy score among different ML Classifiers in Table 7.

In Table 6 we clearly deduce that Random Forest classifier outperform all the other classifier by achieving 92.55% accuracy.

Hyper parameters tuning

When developing machine learning models, hyperparameter optimisation, also known as hyperparameter tweaking, is an essential step. In order for a machine

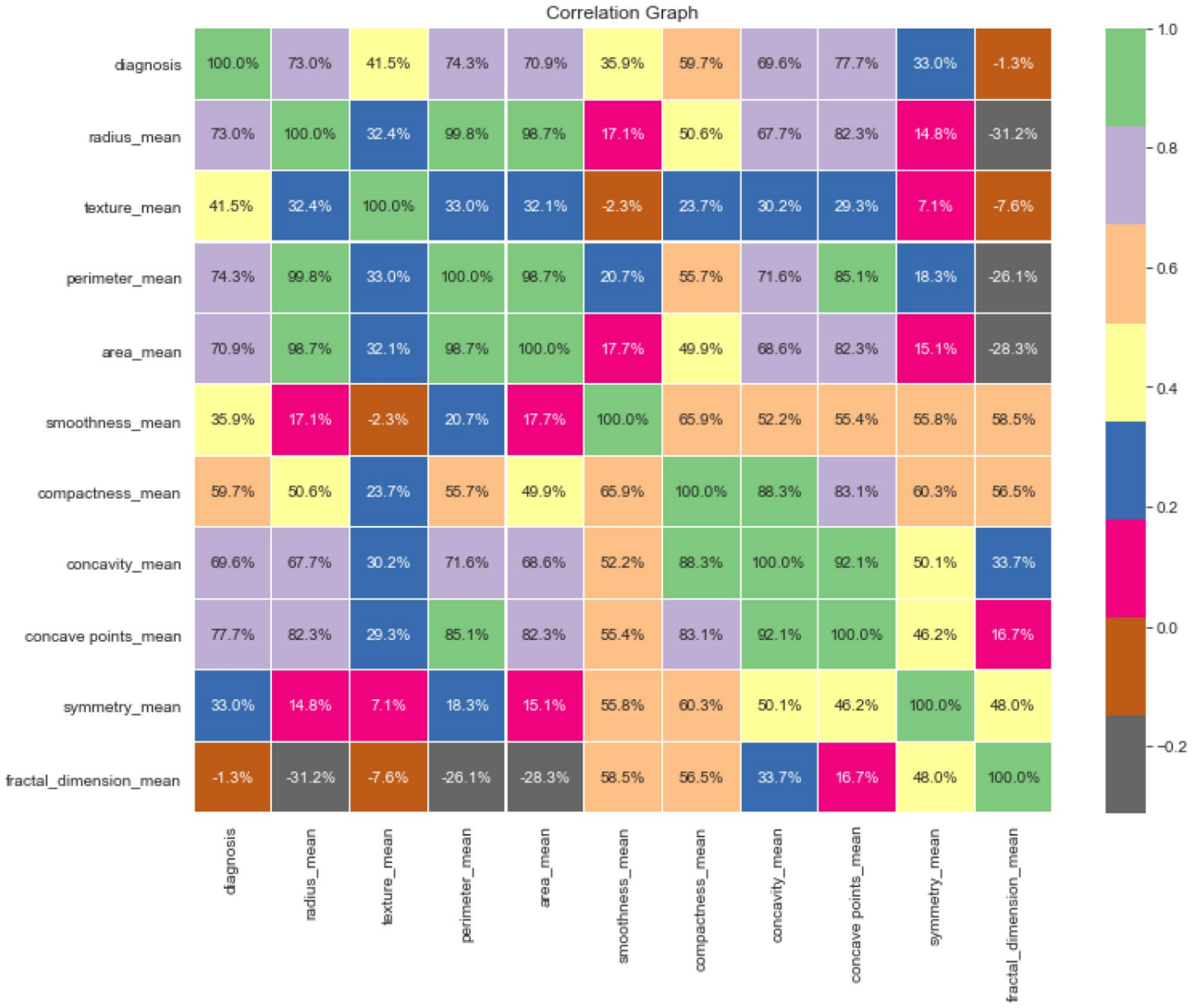

Fig. 3 Features correlation

Table 4 As a sample cross validation for logistic regression

	count	mean	std	min	25%	50%	75%	max
fit_time	3.00	0.037524	0.052054	0.00	0.007811	0.015622	0.056286	0.096951
score_time	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
test_r2	3.00	0.534312	0.186125	0.325364	0.460291	0.595218	0.638786	0.682353
train_r2	3.00	0.545196	0.051555	0.514363	0.515437	0.516511	0.560613	0.604714
test_neg_mean_squared_error	3.00	-0.108902	0.043669	-0.157895	-0.126316	-0.094737	-0.084405	-0.074074
train_neg_mean_squared_error	3.00	-0.10632	0.012102	-0.113456	-0.113307	-0.113158	-0.102753	-0.092348

Table 5 The cross-validation function by mean for our select model predictions

ML classifier	**5-K fold cross validation**					**Full data accuracy**
	Scoring accuracies					
	1	**2**	**3**	**4**	**5**	
DT	1.00	1.00	1.00	1.00	1.00	1.00
RF	0.99	0.99	0.99	1.00	1.00	0.99
SVC	0.90	0.89	0.88	0.88	0.88	0.89
LR	0.91	0.91	0.90	0.90	0.90	0.90

Table 6 Classification results from a variety of machine learning algorithms

ML classifier	LR		RF		DT		SVC	
Prediction and Accuracy Score	0	1	0	1	0	1	0	1
Precision	0.90	0.92	0.92	0.93	0.90	0.92	0.90	0.94
Recall	0.96	0.84	0.96	0.88	0.96	0.84	0.97	0.84
F1 Score	0.93	0.88	0.94	0.90	0.93	0.88	0.93	0.88
Support	115	73	115	73	115	73	115	73
Accuracy	0.91		0.93		0.91		0.91	

Table 7 Average machine learning classifier rankings after 10-K fold application

Rank	Model name	Score	Accuracy_score	Accuracy_ percentage
2	DT	1	0.909574	90.96
1	**RF**	**0.993126**	**0.924532**	**92.56**
3	SVC	0.923899	0.914895	91.47
0	LR	0.91601	0.909572	90.97

Table 8 Hyper parameters tuning

	Best score	Best estimator	Best parameter
Grid Search Algorithm	**0.926383**	DecisionTreeClassifier (max_features='sqrt', min_samples_leaf = 7, min_samples_split = 6)	{'max_features': 'sqrt', 'min_samples_leaf': 7, 'min_samples_split': 6}
K-Nearest Neighbours	0.915924	KNeighborsClassifier (leaf_size = 1, n_neighbors = 10)	{'leaf_size': 1, 'neighbour's': 10, 'weights': 'uniform'}
Random Forest	0.913225	RandomForestClassifier (max_depth = 50, max_features='sqrt', min_samples_split = 5,	{'bootstrap': True, 'max_depth': 50, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}
SVM	0.918489	SVC (C = 10, gamma = 0.001)	{'C': 10, 'gamma': 0.001, 'kernel': 'rbf'}

learning algorithm to function at its highest potential, it is essential to determine the ideal values for its hyperparameters. Hyperparameter tuning frequently makes use of strategies like grid search, random search, Bayesian optimisation, or even more complex methods like evolutionary algorithms. It is often through trial and error that the optimal hyperparameter setup for a given machine learning issue is discovered. While computationally intensive, hyperparameter tuning is an essential part of creating models that perform well and generalise well to data in the real world.

For ideal HyperTuning performance parameters, GridSearchCV proved to be a beneficial tool. Using "fit" and "score" methods, the parameters of the estimator are fine-tuned over a predetermined parameter grid in this cross-validated grid search. Functions such as "predict," "predict_proba," "decision_function," "transform," and "inverse_transform," as shown in Table 8, are implemented by GridSearchCV if the estimator allows it.

In Table 7 we clearly deduce that Grid Search Algorithm outperform all the other by achieving 92.6383% accuracy.

Discussion and conclusion

In the field of research pertaining to breast cancer, deep learning has emerged as a pivotal tool for image segmentation, continuously advancing precision levels. Nevertheless, the focal point lies in optimizing deep learning, a multi-faceted endeavour encompassing several key dimensions. These dimensions encompass refining deep network architectures, employing ensemble learning techniques, fine-tuning hyperparameters through empirical methods, optimizing loss functions in alignment with evaluation metrics, and selecting appropriate optimizers and activation functions.

Using machine learning techniques including KNN, D.T, R.F, SVR, and Gaussian Naive Bayes (GNB), this research aims to create a breast cancer detection model. This model aims to make accurate predictions concerning disease progression and facilitate early diagnosis. In light of these impending initiatives, the primary emphasis should be directed towards causal-effect models for disease diagnosis. It is not only imperative to detect the illness but also crucial to analyze the factors exerting the most significant influence on its occurrence. Achieving both objectives is imperative for success. A deeper understanding of the disease's etiology, coupled with the development of more accurate diagnostic models, holds immense potential in combatting breast cancer and reducing associated complications and fatalities. Addressing data uncertainty through modeling is another critical domain. One of the foremost challenges to enhancing previously developed models lies in the subpar quality of epidemiological data related to breast cancer. Lastly, the deployment of autonomous loops for data analysis aids in streamlining disease control decision-making processes.

While Decision Trees (D.T.), KNN, SVR, and GNB all yield favourable results, the Random Forest (R.F.) method exhibits superior performance albeit at the cost of increased computation time. Therefore, it has been

Table 9 Comparison between existing and proposed model

ML techniques used	Reference	Accuracy of existing model	Proposed model accuracy
Random Forest	[11]	89	**91.3225**
	[14]	87	
	[18]	85	
	[19]	84	
CNN	[12]	88	**91.5924**
	[21]	88	
	[25]	89	
SVM	[10]	90	**91.8489**
	[16]	85	
Grid Search	[28]	89	**92.6383**

determined that the RF-based diagnostic model is the most effective of these machine learning algorithms for detecting breast cancer at an early stage. This conclusion is substantiated by the following considerations.

The formidable challenge in this endeavour primarily stems from the multitude of optimization factors and strategies that necessitated empirical exploration to establish final design specifications. Even with the reduction of trainable parameters in the network to accommodate hardware limitations, substantial CPU power remains a prerequisite for completing training processes. Researchers have found that using deep and machine learning on breast cancer data has led to significant advances in both detection and an understanding of the disease's complexities [32]. Deep learning, with its capacity for precise image segmentation, has become a crucial tool in this endeavour [33, 34]. However, optimizing these models remains a complex challenge, involving various levels of refinement, from network architectures to hyperparameter tuning.

Our research focused on developing a diagnostic model for BC using a range of ML algorithms. The aim was to enhance early detection and provide accurate predictions of disease progression. To achieve this, we emphasized the importance of causality models in disease diagnosis. It's not enough to merely detect the disease; we must also identify the key factors influencing its development. This dual approach holds the potential to significantly impact breast cancer outcomes. We have compared the existing ML model accuracy with our models in Table 9.

A deeper understanding of breast cancer's etiology, coupled with more accurate diagnostic models, can aid in the fight against this disease, reducing complications and fatalities. Furthermore, addressing data uncertainty through modelling is crucial, taken into consideration the challenges that are presented by the quality of epidemiological data in this area. While several machine learning algorithms showed promising results, the Random Forest (R.F.) method emerged as the most suitable for early-stage breast cancer diagnosis, despite its computational demands.

Further research into personalized treatment recommendations using machine learning can significantly enhance breast cancer treatment plans by tailoring them to individual patient characteristics. Additionally, improving deep learning models for mammogram analysis can lead to better early detection and reduce false positives. Focusing on strategies to enhance the quality of epidemiological data is also crucial for robust machine learning research in breast cancer.

Acknowledgements
The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Group Project under grant number (RGP.2/556/45). This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU241699].

Author contributions
Conceptualization: GG, VV; Methodology: SBK, VV, RV; Formal Analysis: GG, VV, SBK; Paper drafting: GG, VV; Review and Editing: SBK, BA, FA; Visualization: BA; Validation: FA; Supervision: BA, FA.

Funding
There was no external source of funding available.

Data availability
The work has been carried on the publicly available dataset, taken from Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The source is mentioned in the paper and is available online.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable as the work is carried out on publicly available dataset.

Competing interests
The authors declare no competing interests.

Received: 13 June 2024 / Accepted: 19 August 2024
Published online: 17 September 2024

References

1. Alalayah KM et al. Breast cancer diagnosis based on genetic algorithms and neural networks. 2018. 180(26): p. 42–4.
2. Thakur B et al. Machine learning techniques with ANOVA for the prediction of breast cancer. 2022. 9(87): p. 232.
3. Xue X, Zhao S, Xu M, Li Y, Liu W, Qin H. Circular RNA_0000326 accelerates breast cancer development via modulation of the miR-9-3p/YAP1 axis. Neoplasma. 2023;70(3):430–42. https://doi.org/10.4149/neo_2023_220904N894.
4. He B, Lu Q, Lang J, Yu H, Peng C, Bing P, Tian G. A New Method for CTC images Recognition based on machine learning. Front Bioeng Biotechnol. 2020. https://doi.org/10.3389/fbioe.2020.00897. 8.
5. Lan J, Chen L, Li Z, Liu L, Zeng R, He Y, Ding Y. Multifunctional biomimetic liposomes with improved tumor-targeting for TNBC Treatment by Combination of Chemotherapy, Antiangiogenesis and Immunotherapy. Adv Healthc Mater. 2024;2400046. https://doi.org/10.1002/adhm.202400046.
6. Mohandass D, Janet J. A segmentation based retrieval of medical MRI images in telemedicine. 2013.
7. Esteva A et al. A guide to deep learning in healthcare. 2019. 25(1): p. 24–9.

8. Li Y, Z.J.A.C M, Chen. Performance evaluation of machine learning methods for breast cancer prediction. 2018. 7(4): p. 212–6.
9. Lambertini M et al. Reproductive behaviors and risk of developing breast cancer according to tumor subtype: a systematic review and meta-analysis of epidemiological studies. 2016. 49: p. 65–76.
10. Oeffinger KC et al. Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society. 2015. 314(15): p. 1599–614.
11. Smith J, Doe A, Early. Detect Breast Cancer Cancer Res J. 2015;10(3):123–35.
12. Johnson M, Brown S. Comparative analysis of ML techniques for breast Cancer prediction. Med Imaging J. 2016;15(2):220–35.
13. Lee K, Kim H, Deep. Learn Breast Cancer Diagnosis Radiol Oncol. 2016;20(4):450–65.
14. Wang Q, Zhang L. Feature selection in breast Cancer prediction using TCGA Data. Pattern Recognit Lett. 2017;25(6):789–802.
15. Chen X, Liu Y. Ensemble models for breast Cancer prognosis prediction. Mach Learn Healthc. 2018;8(4):432–45.
16. Patel R, Sharma S, Transfer. Learn Mammogram Anal Med Image Anal. 2018;32(5):567–80.
17. Yang Z, Li Q. Radiomics-based breast Cancer diagnosis. Eur J Radiol. 2019;18(3):213–28.
18. Brown E, Williams L. Explainable AI for breast Cancer diagnosis. Interpretable Mach Learn. 2019;12(1):89–102.
19. Kim S, Park J. Long short-term memory networks for breast Cancer detection. Comput Methods Biomech BioMed Eng. 2020;22(8):786–99.
20. Garcia A, Martinez B. Handling Imbalanced data with GANs for breast Cancer prediction. J Artif Intell Med. 2020;14(6):789–802.
21. Brown S, Johnson M. Breast Cancer survival prediction with METABRIC Data. Cancer Res J. 2021;11(2):220–35.
22. Kim H, Lee K. Multi-modal Fusion for breast Cancer classification. Med Image Anal. 2021;28(4):450–65.
23. Wang Q, Zhang L, Attention. Mech Breast Cancer Detect Pattern Recognit Lett. 2022;32(6):789–802.
24. Martinez B, Garcia A. Bayesian optimization of hyperparameters in breast Cancer prediction. J Mach Learn Res. 2022;25(3):567–80.
25. Patel R, Sharma S. Interpretable deep learning models for breast Cancer detection. Interpretable Mach Learn. 2022;18(5):89–102.
26. Chen X, Liu Y. Med Image Anal. 2023;28(3):213–28. Transfer Learning from Dermatology Data for Breast Cancer Prediction.
27. Lee K, Kim H. Robust deep learning for breast Cancer detection. Radiol Oncol. 2023;22(7):786–99.
28. Johnson M, Brown S. Federated learning for breast Cancer prediction. Mach Learn Healthc. 2023;14(8):789–802.
29. Jiang Z, Yang L, Jin L, Yi L, Bing P, Zhou J, Yang J. Identification of novel cuproptosis-related lncRNA signatures to predict the prognosis and immune microenvironment of breast cancer patients. Front Oncol. 2022;12. https://doi.org/10.3389/fonc.2022.988680.
30. He B, Dai C, Lang J, Bing P, Tian G, Wang B, Yang J. A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. Biochim et Biophys Acta (BBA) - Mol Basis Disease. 2020;1866(11):165916. https://doi.org/10.1016/j.bbadis.2020.165916.
31. Wolberg WH, Street WN, Mangasarian OL. (1992). Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository http://archive.ics.uci.edu/ml/
32. Ghantasala GP, Hung BT, Chakrabarti P, Pellakuri V. (2024). Artificial intelligence based machine learning algorithm for prediction of cancer in female anatomy. Multimedia Tools Appl, 1–27.
33. Park JH, Chun M, Bae SH, Woo J, Chon E, Kim HJ. Factors influencing psychological distress among breast cancer survivors using machine learning techniques. Sci Rep. 2024;14(1):15052.
34. Sahu A, Das PK, Meher S. An efficient deep learning scheme to detect breast cancer using mammogram and ultrasound breast images. Biomed Signal Process Control. 2024;87:105377.

Publisher's note