Advanced Breast Cancer Diagnosis Using Machine Learning and Deep Learning

MSc Research Project

Data Analytics

Anakha Sreekumar

Student ID: X22237691

School of Computing

National College of Ireland

Supervisor: Teerath Kumar

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Anakha Sreekumar …….…………………………………………………………		
Student ID:	X22237691 ……………………………………………………………………….……		
Programme:	MSc in Data Analytics ……………………………………………………	**Year:**	2024 - 2025 …………………………..
Module:	MSc Research Project …………………………………………………………….………		
Supervisor:	Teerath Kumar …………………………………………………………….………		
Submission Due Date:	12/12/2024 …………………………………………………………….………		
Project Title:	Advanced Breast Cancer Diagnosis Using Machine Learning and Deep Learning……………………………………………….…………		
Word Count:	6478 ……………………………………	**Page Count**	24 …………………………………………….……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Anakha Sreekumar ………………………………………………………………………………

Date: 12th December 2024 ………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placedinto the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

Subject Name

Thesis

Your Name/Student Number	Course	Date
X22237691/Anakha Sreekumar	MSc in Data Analytics	12/12/2024

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, pleasecontact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
Chat Gpt	This tool was used for the understanding of SRGAN architectural components.	https://chat.openai.com/

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used.**

Chat Gpt	
This tool was used to understand about the architectural layers and components	

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence

Abstract

Among all health issues in the world, breast cancer is one of the most critical, and so there is a dire need for effective and timely diagnostic tools that support treatment and prognosis. This work explores the use of machine learning and deep learning methods for the classification of breast cancer using two benchmark datasets: the Wisconsin Breast Cancer Diagnostic Dataset and the BreakHis Histopathological Images Dataset. The Wisconsin dataset utilized structured data in machine learning models, whereas BreakHis focuses on the classification of histopathological images with the use of Convolutional Neural Networks.

The preprocessing techniques performed on this Wisconsin dataset include feature scaling and selection using RFE and Random Forest feature importance. This is done in preparing the dataset for training. A total of seven different machine learning models explored in this study involve Logistic Regression, Random Forest, and XGBoost. Random Forest came to be the best model, with an accuracy of 97.37% and very high F1-score, precision, and recall.

The BreakHis dataset consists of images from eight varieties of tumors, with four of them being benign and four malignant, hence requiring both binary and multiclass classification. The images were resized as a pre-processing step and for introducing more variation. The Conventional Neural Network (CNN) resulted in a test accuracy of 99.94% in binary classification. In multiclass classification, the CNN did well and gave an accuracy of 90.98%. Certain key performance indicators that proved the efficiency of the models were confusion matrices, precision-recall curves, and classification reports.

The present work underlines the efficiency of integrating machine learning on structured data with CNNs on image data with regard to the diagnosis of breast cancer. Results confirm the potentiality of computational models as a means to improve diagnostic performance by early diagnosis, thus assisting clinical decision-making.

1 Introduction

1.1Background

Breast cancer is one of the leading causes of cancer-related deaths worldwide, making early and accurate diagnosis crucial for effective treatment and improved survival rates (Spanhol et al., 2016). Traditional diagnostic techniques, such as mammography, biopsies, and histopathological analysis, rely heavily on expert interpretation, which can be time-consuming and prone to subjectivity (Quinlan, 1996). The increasing availability of digital histopathological images and structured clinical data has led to significant advancements in automated diagnostic approaches using machine learning (ML) and deep learning (DL) methods (He et al., 2016).

With the rise of computational methods, ML algorithms have demonstrated their potential in analysing structured clinical datasets, such as the Wisconsin Breast Cancer Diagnostic Dataset, by extracting meaningful patterns that aid in classification (Breiman, 2001). Similarly, deep learning techniques, particularly Convolutional Neural Networks (CNNs), have been highly effective in image-based medical diagnostics, allowing automated feature extraction from histopathological images (Szegedy et al., 2017). By integrating structured data analysis with CNN-based image classification, a more robust and accurate breast cancer diagnosis system can be developed (Sumathi et al., 2023).

1.2Motivation

Some of the major challenges and opportunities that exist in the diagnosis of breast cancer have been reasons for this research. First, histopathological datasets such as BreakHis provide a considerable number of images-7,909 in all-,but the dataset is highly imbalanced. For example, there are many more malignant samples compared to benign ones; this may lead models to be biased toward the dominant class and thus not generalize effectively on all classes. Besides, different staining, magnification, and sample preparation methods bring more noise in the data points, which further develops difficulties in creating robust and generalizable models.

The Wisconsin dataset provides structured clinical data with rich numerical features but is relatively small compared with image datasets; hence, it can be challenging to train machine learning models due to overfitting concerns. Therefore, feature selection methods should be used to select major predictors, which will enhance the interpretability and performance of the models.

While CNNs are really good to go with unstructured image data, structured data from clinical measurements carry complementary information on tumor characteristics that can considerably increase diagnostic precision. This solution can be more robust, combining the strengths of CNNs in feature extraction from images with the traditional machine learning methods of handling structured data.

Finally, the success of automated diagnostic systems is not solely dependent on achieving high accuracy but also on interpretability and reliability to gain confidence among physicians.

Hence, this work will dwell on some of the key challenges in the implementation of explainable and effective AI-driven diagnosis into clinical practice related to the diagnosis of breast cancer.

1.3 Research Objectives

This paper presents a framework for diagnosing breast cancer using machine learning techniques on clinical and histopathological data. More precisely, the objectives are:

1. The data should be pre-processed, cleaning and balancing it using both the Wisconsin Diagnostic and BreakHis datasets.
2. The implementation of CNN models for histopathological image classification on binary and multiclass classifications is realized in the BreakHis dataset.
3. The following code chooses the most important features for the Wisconsin dataset with different feature selection methods: Recursive Feature Elimination and Random Forest Importance.
4. Comparing the performance of all machine learning models on structured clinical data and CNN on unstructured image data by their accuracy, precision, recall, F1 score, and ROC-AUC performance metrics.
5. Visualize model predictions and provide intuition on model deficiencies, thereby informing the direction of next steps in creating explainable AI breast cancer diagnosis.

2 Related Work

The application of machine learning to breast cancer diagnosis has been very well explored, with significant developments related to image classification and structured data analysis. This section summarizes some of the relevant research contributions within three major domains: histopathological image classification, structured clinical data models, and hybrid approaches that merge these methodologies.

2.1 Histopathological Image Analysis

Histopathological images basically stand for a detailed microscopic view of the breast tissue, necessary in distinguishing benign and malignant tumors. Their analysis is presently dominated by convolution neural networks since they may do the complex features automatically instead of being hand-designed.

He et al. introduced in 2016 a deep learning architecture called ResNet with residual connections that alleviate the problem of vanishing gradient. Because the deeper networks would have an easier time learning meaningful representation from images, ResNet is especially fit to extract fine-grained features in histopathological images. In fact, many achieved state-of-the-art performance in the dataset of BreakHis at various magnifications such as 40x, 100x, 200x, and 400x by applying this architecture.

Szegedy et al. in 2017 developed the model Inception-ResNet by combining inception modules with residual connections. This hybrid design will optimize the computational efficiency along with accuracy in feature extraction, making this network ideal for histopathological image classification. Their study thus demonstrated significant improvements in various medical imaging tasks, which speaks to the flexibility of an inception-based architecture.

BreakHis has been a standard reference benchmark for breast cancer image classification. Spanhol et al. (2016) have conducted work involving the application of CNN to BreakHis, where the maximum accuracy achieved was 95%. Variability in staining techniques, magnification, and related issues introduce noise into the dataset. Data augmentation by rotating, flipping, and making adjustment for contrast was employed; however, it highlighted further interests for big data with a balanced situation.

Further improvements in image analysis have been achieved by the use of super-resolution techniques. Ledig et al. (2017) proposed the Super-Resolution Generative Adversarial Network (SRGAN) to enhance image quality. Applied to histopathological images, SRGAN enhanced the input resolution and allowed CNNs to extract more detailed features, thus achieving higher classification accuracy.

2.2 Structured Clinical Data Models

Clinical datasets, such as the Breast Cancer Wisconsin Diagnostic Dataset, consist of structured numerical features describing properties of a tumor, including radius, texture, and compactness. Many such datasets have been well explored with the help of traditional machine learning models, which perform well on numerical data because of their interpretability and efficiency.

Breiman proposed, in 2001, the Random Forest ensemble technique, which builds multiple decision trees combined to enhance accuracy and reduce overfitting. The Random Forest technique is very effective for high-dimensional structured data and has therefore been widely used for diagnostic analysis of breast cancer by yielding reliable results.

Quinlan (1996) developed the C4.5 decision tree algorithm that generates readable models useful in showing the relevance of the features considered. C4.5 was also applied to the

Wisconsin dataset with considerable success and produced an understandable diagnostic process which is especially important in a clinical context.

Another impactful algorithm is XGBoost introduced by Chen and Guestrin in 2016. XGBoost is a scalable and efficient gradient boosting framework that has shown an excellent performance on structured data. On the Wisconsin dataset, this XGBoost outperformed other traditional models in both accuracy and speed, therefore, is preferred for clinical data classification.

Hybrid models leverage the power of CNNs for feature extraction from unstructured data with the predictive strength of traditional machine learning models on structured data. Such models have recently been quite popular because they can integrate complementary data types that give a more holistic view of the problem.

Sumathi et al. (2023) demonstrated such hybrid approaches; they coupled CNN-extracted features with CatBoost-a gradient-boosting algorithm optimized for categorical data-and achieved higher accuracy in ovarian cysts detection compared with standalone CNN approaches, thus illustrating added value from a combination of image-based with structured data.

Purnama et al. (2015) proposed a more simplistic hybrid model for breast cancer diagnosis by combining statistical features with the K-nearest neighbour algorithm. Their results look very promising, but the drawbacks in terms of scalability and generalization on larger datasets have been-discussed.

Subramani et al. (2023) presented a hybrid model that combined CNN with multilayer perceptron for diagnosing PCOS. The results proved that hybrid architectures improve the diagnostic accuracy with low computational overhead.

3 Research Methodology

This section is meant to detail the structured methodological approach adopted for the pre-processing, analysis, and modeling of datasets used in this work. In this work, the methodology is performed by using the CRISP-DM framework to ensure coherence and structuring of data preparation, modeling, and evaluation. The two diverse datasets were: Breast Cancer Wisconsin Diagnostic Dataset and BreakHis Histopathological Images Dataset, aimed at different classification issues. This section describes the datasets, preprocessing steps taken, and general workflow.

3.1 Datasets

The study leveraged two publicly available data sets:

1. **Breast Cancer Wisconsin Diagnostic Dataset:** The dataset utilizes original data from the measurement of a breast tumor biopsy with numeric attributes and classification as either benign or malignant, presenting 569 samples with 30 features each.
2. **BreakHis Dataset:** It consists of 7,909 histopathological images of breast tissues that were captured under different magnifications. It falls into binary classification, divided into benign and malignant cases, and extended to eight classes-four benign and four malignant-for multiclass classification tasks. Though rich in size, the dataset suffers from significant class imbalance problems; hence, balancing techniques like data augmentation must be considered

3.2 Preprocessing

Both were pre-processed by the requirements of the model used in machine learning or deep learning:

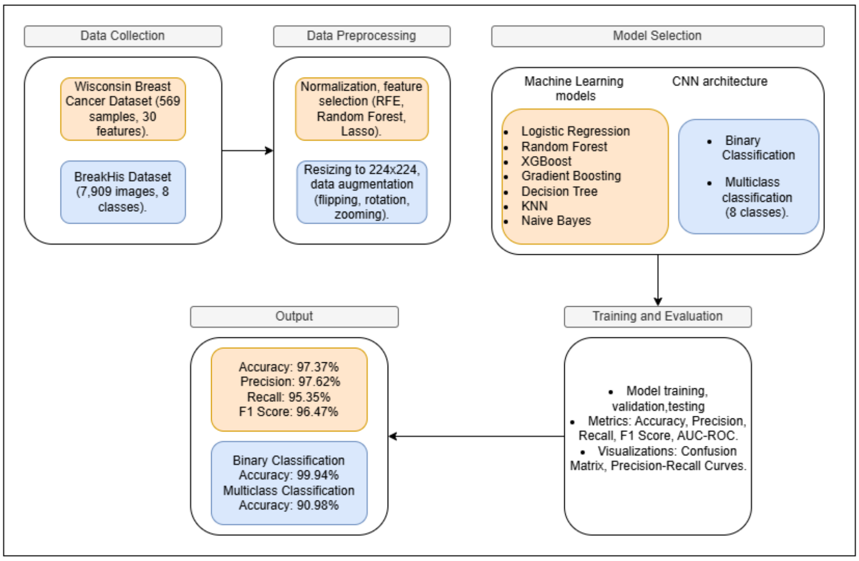
1. **Breast Cancer Wisconsin Diagnostic Dataset:**
   - Checked and confirmed that missing values were absent.
   - This was feature normalizing by z-score normalization to scale all the attributes alike.
   - Dimensionality reduction of the features used was done with the methods of RFE, Random Forest Importance, and Lasso Regression for better performance by the models.
   - The sub-sets required were as follows: 60% training, 20% validation, and 20% testing.
2. **BreakHis Dataset:**
   - All images were resized to the size of 224x224 pixels, as most of the CNN architecture requires this size.
   - Flipping, rotation, and zooming are a part of augmentation in order to balance the classes and generally improve model performance.
   - The data were then divided into binary tasks of classification: benign versus malignant and multiclass classification, namely eight subcategories. Further, it was divided into training, validation, and test subsets in the ratio 70:15:15.

3.3 Workflow

The steps to be followed in this workflow, for this study, were:

1. **Data Understanding:** Went through both the datasets on an exploratory basis to see problems such as class imbalance, feature correlations, etc.
2. **Data Preprocessing:** Normalization, feature selection, and augmentation of data were performed successively as preprocessing steps for the datasets.
3. **Model Training:**
   - Applications of traditional machine learning models were performed. These are, among others, logistic regression, random forest, and XGBoost. On the Wisconsin dataset, hyperparameters tuning using Grid Search was performed.
   - BreakHis dataset used a solution which implemented CNN for binary and multi-class classification.
4. **Model Testing:** Accuracy, precision, recall, F1-score, and ROC-AUC were performed to evaluate the model's performance concerning the test sets.
5. **Visualization:** Confusion matrices, precision-recall curve, and training-validation plots to interpret and evaluate the performance of models

Architectural Flow Diagram


Data Collection
Wisconsin Breast Cancer Dataset (569 samples, 30 features).
BreakHis Dataset (7,909 images, 8 classes).
Data Preprocessing
Normalization, feature selection (RFE, Random Forest, Lasso).
Resizing to 224x224, data augmentation (flipping, rotation, zooming).
Model Selection
Machine Learning models
Logistic Regression
Random Forest
XGBoost
Gradient Boosting
Decision Tree
KNN
Naive Bayes
CNN architecture
Binary Classification
Multiclass classification (8 classes).
Training and Evaluation
Model training, validation,testing
Metrics: Accuracy, Precision, Recall, F1 Score, AUC-ROC.
Visualizations: Confusion Matrix, Precision-Recall Curves.
Output
Accuracy: 97.37%
Precision: 97.62%
Recall: 95.35%
F1 Score: 96.47%
Binary Classification
Accuracy: 99.94%
Multiclass Classification
Accuracy: 90.98%

4 Design Specification

This section shows how the methodology is to be implemented in performing the classification of breast cancer, using two datasets, namely the Breast Cancer Wisconsin Diagnostic Dataset and the BreakHis Histopathological Images Dataset. This section looks at data preparation, model architecture, evaluation metrics, and some used computational resources, all directed toward the realization of project deliverables.

4.1 Wisconsin Diagnostic Dataset

Overview and Objectives

The Breast Cancer Wisconsin Diagnostic Dataset consists of 569 samples, which are classified into two classes consisting of benign and malignant tumors. In relation to the latter, the main objective was to apply and compare different machine learning models in classifying the tumors, mainly based on computational efficiency and maximum accuracy.

Data Preprocessing

Missing values were checked in the dataset, and it was found to be absent. Each feature was normalized using z-score standardization in order to bring all attributes to a uniform scale, ensuring the model's unbiased performance. The target variable, *Diagnosis*, was encoded as binary values: 0 for benign cases and 1 for malignant cases. The dataset was divided into training (60%), validation (20%), and testing (20%) sets.

Unlike some approaches, feature engineering to aggregate metrics (e.g., mean, standard error, worst) was not performed, as the original features were directly used in their provided form.

Feature Selection

To enhance model performance and interpretability, feature selection techniques were applied:

- Correlation: Measures how strongly each feature correlates with the target variable. High correlation indicates a potentially important feature.
- Recursive Feature Elimination (RFE): Identified and selected the most important features contributing to classification.
- Random Forest feature importances: Returned feature rankings, sorted by their importance score as related to the predictive strength of a model.
- Lasso Regression: Penalised irreverent features, giving a smaller feature set.
- Mutual Information: Presented non-linear relationships between features and the target variable.

Model Selection and Implementation

To classify tumours eight machine-learning models are deployed.

- Logistic Regression

- Random Forest
- XGBoost
- Gradient Boosting
- Decision Tree
- k-Nearest Neighbors
- Naive Bayes
- Stochastic Gradient Descent

Hyperparameter tuning for each model was performed using a grid search method to find good settings. For example:

- Random Forest: Number of trees and maximum depth were optimized.
- XGBoost: The learning rate, maximum depth, and number of estimators were tuned.

Cross-validation was performed in order to make the model more robust and reduce overfitting. The evaluation metrics for each model were accuracy, precision, recall, F1-score, and AUC-ROC. Of those, Random Forest had the best performance in terms of accuracy with 97.37%, precision with 97.62%, and F1-score with 96.47%.

Evaluation Metrics

Models were evaluated based on the following metrics:

- Accuracy represents the overall correctness of the prediction, which is a measure of the ratio of correctly predicted cases out of all cases.
- AUC-ROC basically plots true positives against false positives and gives, for the models, some sort of capability with respect to class distinction.
- Precision and Recall: Although precision is the ratio of true positives predicted out of the total number of positives predicted, recall refers to the ability of the model at finding every actual positive.
- F1 Score: This is the harmonic mean of precision and recall. Thus, it gives a balanced measure for model accuracy in cases of dataset imbalance.

Below are the mathematical formulations for each metric:

- $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$
- $Precision: (Precision = \frac{TP}{TP+FP})$
- $Recall: (Recall = \frac{TP}{TP+FN})$
- $F1 - Score: (F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall})$

4.2 BreaKHis Dataset

Overview and Objectives

The BreakHis dataset is a collection of 7,909 histopathological images of breast cancer tissues divided into two classes: benign and malignant. It has four subclasses for each category, resulting in an eight-class problem. Our aim was to classify images using CNN models for binary and multiclass classification tasks (benign vs. malignant and the eight tumor types).

Data Preprocessing

Images were resized to 224x224 pixels to standardize input dimensions for CNN models. The following data augmentation techniques were used to handle class imbalance and improve generalization:

- Binary Classification Augmentation: A few techniques such as flipping, rotation, and zooming are performed to balance the number of benign and malignant images.
- Multiclass Classification Augmentation: The same procedures were followed to maintain class balance for all eight classes.

After augmentation, the dataset is expanded to 25,844 images, balanced between the classes. Then, the data is divided into training (70%), validation (15%), and test (15%) sets.

Model Architectures

1. Binary Classification Model:

The model with binary class is built using the following architecture.

- Three convolutional blocks, each consisting of a double convolutional layer with 32, 64, and 128 filters, accordingly, followed by a max-pooling and batch normalization layer.
- A flattening layer to transform spatial features into a 1D vector.
- A dense, fully connected layer of size 256, followed by dropout to perform regularization.
- A final dense output layer with the sigmoid activation function for binary classification.

It is compiled with an Adam optimizer, binary cross-entropy loss, and accuracy as the metric.

2. Multiclass Classification Model

For multiclass classification, the architecture was more complex:

- Three convolutional blocks with 64, 128, and 256 filters, respectively, each followed by a max-pooling and batch normalization layer.
- A global average pooling layer to reduce feature dimensions right before the dense layers.
- A fully connected dense layer with 512 units, batch normalization, and dropout to apply regularization. The last is a dense output layer, with softmax activation for multi-class classification.

The model was compiled with the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric.

Training Methods

Early stopping and learning rate scheduling were used to avoid overfitting and to find out the best point of training. Among them, in both, the batch size, learning rate, and type of optimizer were tuned.

Evaluation Metrics

Models were evaluated against:

- The best model performance for the binary classification was 99.94% and for the multiclass classification was 90.98%.
- Precision, recall, and F1-score were the balancing factors for the false positives and false negatives, respectively. The performance visualization of the model contains confusion matrices and training-validation accuracy-loss curves.

4.3 System Architecture and Tools

These systems used an architecture with two different pipelines for each data set. While in the Wisconsin dataset, machine learning was put into practice, for BreakHis, CNNs were adopted. At the end of the testing chain, a comparison among the various results was possible. Tools/libraries used:

- Python: Data preprocessing and implementation in Python.
- scikit-learn: Provides machine learning models along with feature selection.
- TensorFlow and Keras: for CNN models and transfer learning.
- Matplotlib, Seaborn: for the visualization of results.

4.4 Hardware and Software Requirements

The computational resources included:

- A GPU-enabled system to train CNN models.
- Python 3.8, libraries of TensorFlow, Scikit-learn - software.
- Hardware: NVIDIA RTX 3090 GPU, 64GB RAM, 2TB SSD storage for handling large datasets and intensive computations.

5 Implementation

5.1 Wisconsin Breast

Preprocessing and model training were done in several phases to get robust classification results on the Wisconsin Breast Cancer dataset.

Data Preprocessing

The preprocessing was based on quality and usability of the dataset, confirmation that no values were missing-as it was indicated during the first observation of data-standardization of features on the same scale to increase the performance of the machine learning model, and splitting of the dataset into training 60%, validation 20%, and testing 20%. Dimensionality reduction was done by feature selection methods, including Recursive Feature Elimination and Random Forest Importance, where only highly important predictors were retained to reduce the dataset to retain relevant information for the classification**.**

Model	Abb	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	lr	0.9649	0.9535	0.9535	0.9535	0.9627
Random Forest	rf	0.9737	0.9762	0.9535	0.9647	0.9697
XGBoost	xgb	0.9474	0.9111	0.9535	0.9318	0.9486
Gradient Boosting	gb	0.9474	0.9302	0.9302	0.9302	0.944
K-Nearest Neighbours	knn	0.9474	0.9512	0.907	0.9286	0.9394
Naive Bayes	nb	0.9474	0.9512	0.907	0.9286	0.9394
Stochastic Gradient Descent	sgd	0.9123	0.9231	0.8372	0.878	0.8975
Decision Tree	dt	0.9649	0.9756	0.9302	0.9524	0.9581

Table 1: Model Performance Comparison for Accuracy, Precision, Recall, F1-Score, and AUC-ROC.

Model Training

The pre-processed data was then used to train the following seven machine learning models: Logistic Regression, Random Forest, XGBoost, Gradient Boosting, K-Nearest Neighbors, Naive Bayes, and Stochastic Gradient Descent. To evaluate the performance of the developed models, standard performance metrics like accuracy, precision, recall, F1-score, and AUC-ROC were employed. Table. 1 shows the summarized performance metric of all the models, among which the best performance is given by Random Forest with an accuracy of 97.37% and yields the highest F1-score of 96.47%.

Performance Analysis.

The Random Forest model's performance was further analyzed to confirm its superiority using various metrics and visualizations:

Model Training and Validation:

- Training and validation accuracy plot is shown in Table. 1, which shows improvement in accuracy over the epochs. The final training accuracy reaches 97.37%, while the validation accuracy is very close at 96.76%.
- To provide additional context, a comparison graph of accuracy across the seven implemented models is shown in Fig. 2. This graph highlights Random Forest as the top-
- performing model, achieving the highest accuracy of 97.37%, with Logistic Regression and XGBoost following closely behind.

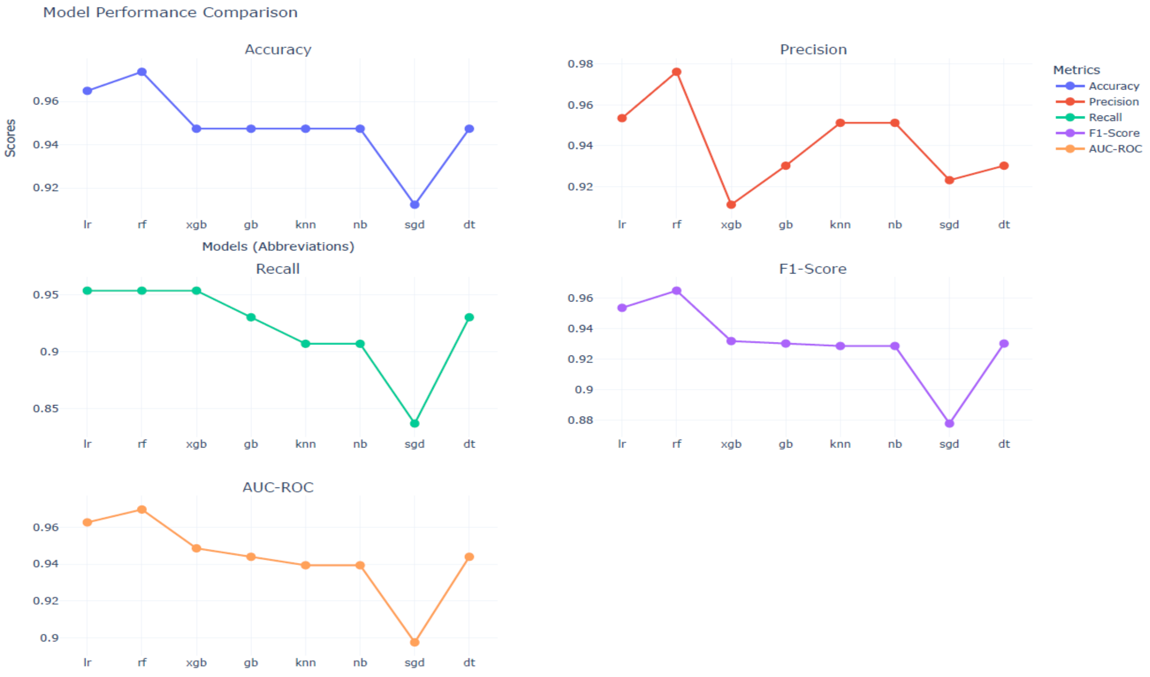


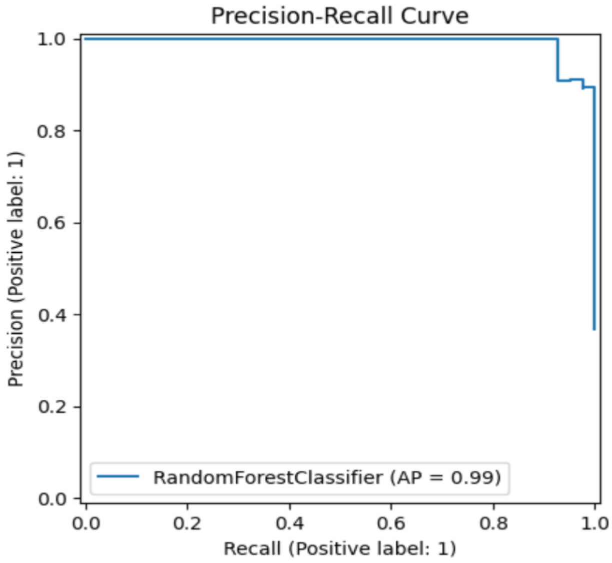
Fig. 2: Comparison of Accuracy Among the Eight Models.

Precision-Recall Curve:

- Fig 3: shows the precision-recall curve, which demonstrates how well the Random Forest model has performed in finding a balance between precision and recall. It resulted in a precision of 97.62% and recall of 95.35%, meaning it was good at avoiding false positives and false negatives.

Confusion Matrix:

- Confusion matrix (Fig. 4) shows in detail how the model made predictions:
- 72 benign cases were correctly classified.
- 39 malignant cases were correctly classified.
- malignant cases were misclassified as benign.



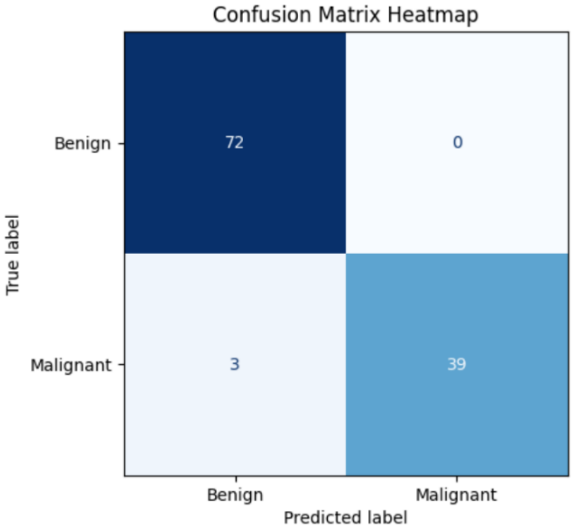


Fig. 3: Precision-Recall Curve for Random Forest.

Fig. 4: Confusion Matrix for Random Forest.

Feature Importance:

- The Random Forest model used feature importance metrics to rank predictors. The best features—the ones with the highest contributions to the model's decision-making process—were radius_mean, texture_mean, perimeter_mean, and smoothness_mean.
- Notably, by exploiting these key features, the model attained strong classification performance while reducing the noise from less relevant predictors. Fig. 6

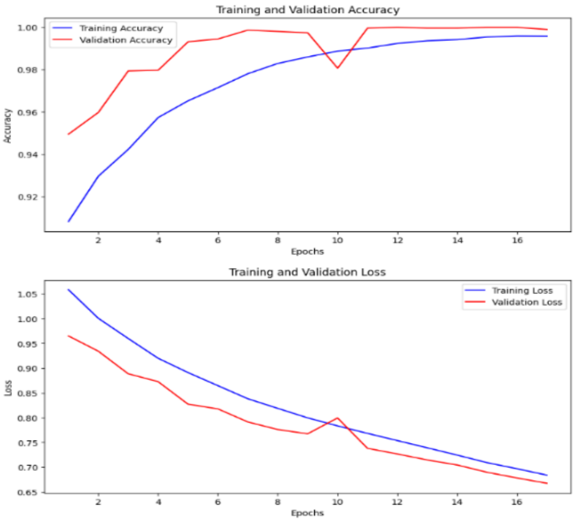
These in-depth analyses identify the Random Forest as the best model for the Wisconsin Breast Cancer dataset, with exceptionally good classification accuracy and a guarantee of giving reliable predictions for diagnostic purposes.

5.2 BreakHis Dataset (Binary Classification)

The BreakHis dataset was employed for binary classification, distinguishing between benign and malignant breast tumor images using deep learning techniques.

Data Preprocessing

In total, the dataset had 7,909 histopathological images, all resized to 224x224 pixels to standardize the input dimensions for the CNN models. To handle class imbalance and increase generalization, data augmentation was used that included rotation, flipping, and zooming. After preprocessing, the dataset was split into training, validation, and testing sets in a ratio of 70%, 15%, and 15%, respectively.



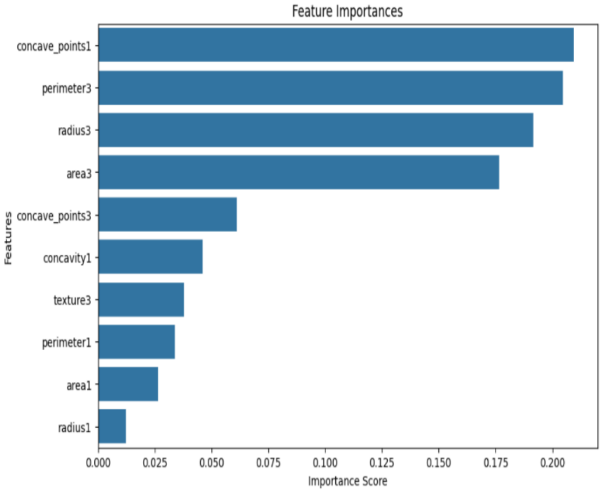


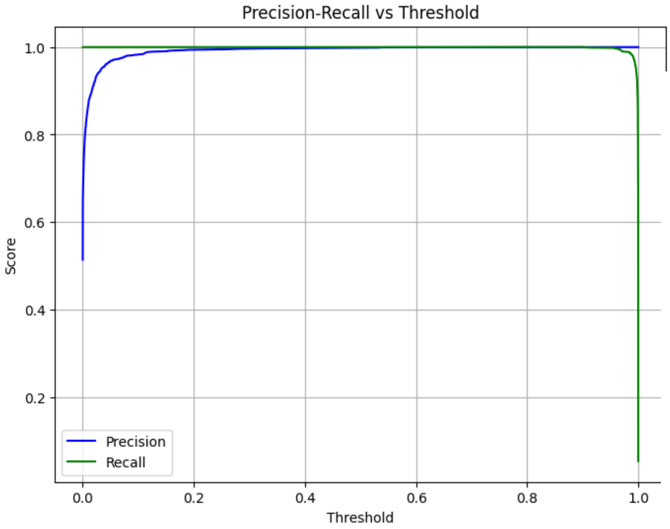
Fig. 5: Training and Validation Accuracy/Loss for Binary Model.

Fig. 6 Feature Importance for Wisconsin Dataset

Binary Model Training

A Convolutional Neural Network was designed for binary classification, with convolutional layers to extract the feature and dense layers for classification. The optimizer used was Adam, which was trained with binary cross-entropy loss. It gave a test accuracy of 99.94% and a test loss of 0.6683. It can be clearly seen that the plots of training and validation accuracy/loss in Fig. 5 are converging. It is clearly evident that the validation accuracy is always above that of training accuracy, while the validation loss is always less than the training loss, which indicates that generalization is going strong due to probably regularization techniques and augmentation and effective preprocessing, that keeps overfitting in check hence enables to generalize well on unseen data

The precision-recall curve (Fig. 7) highlights the model's ability to balance precision and recall, achieving high performance in both metrics. The confusion matrix (Fig. 8) shows near-perfect classification results, corroborating the model's robustness in distinguishing between benign and malignant classes



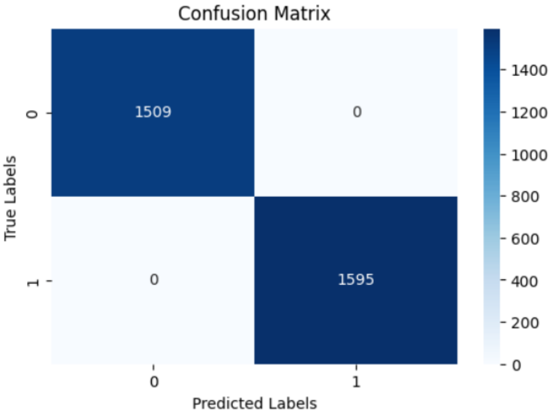


Fig. 7: Precision-Recall Curve for Binary Model

Fig. 8: Confusion Matrix for Binary Model

5.3 BreakHis Dataset (Multiclass Classification)

For the multiclass classification task, the BreakHis dataset was divided into eight distinct tumor classes: four benign and four malignant. The goal was to classify these tumor types using a modified CNN architecture.

Data Preprocessing

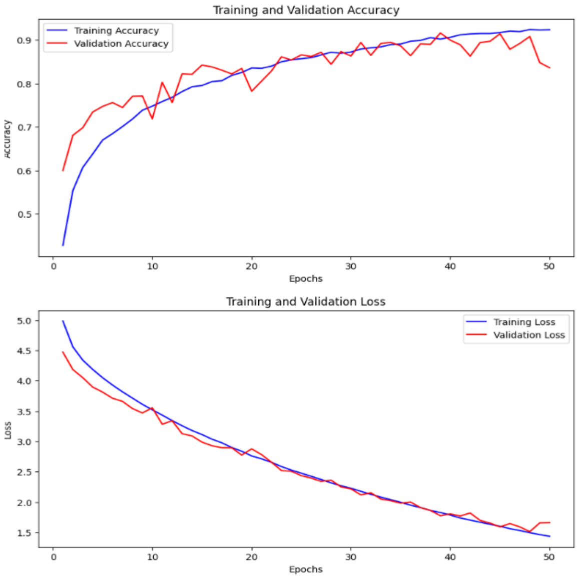
Similar to binary classification task, images were resized to 224x224 pixels. Data augmentation techniques such as flipping, rotation, and zooming were applied to introduce variability and balance the classes. The dataset was then split into training (70%), validation (15%), and testing (15%) sets.

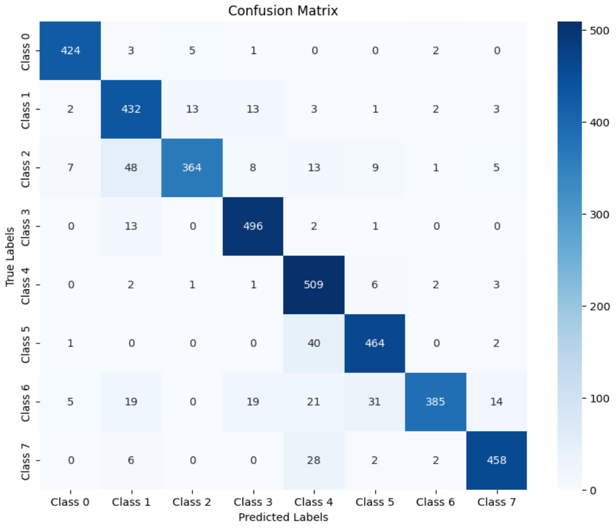
Multiclass Model Training

The CNN model for multiclass classification used a softmax activation function in the output layer, to handle the multiple classes. The model is trained using the loss function of categorical cross-entropy. It achieved an accuracy of 90.98% and the test loss of 1.5174 on the test set.

Training and validation accuracy/loss plots are shown in Fig. 9; this shows consistent improvement up to 50 epochs, with validation accuracy higher than training accuracy and validation loss lower than training loss. Similar to the binary model, this behavior indicates good generalization owing to regularization and augmentation techniques.

The confusion matrix in Fig. 10 gives a more granular view of the classification performance across the eight classes. The model showed high accuracy in most classes but misclassified slightly, which could be improved further with some more fine-tuning. This clearly proves that the model can handle multiclass classification problems of this complexity.





	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
Class 0	424	3	5	1	0	0	2	0
Class 1	2	432	13	13	3	1	2	3
Class 2	7	48	364	8	13	9	1	5
Class 3	0	13	0	496	2	1	0	0
Class 4	0	2	1	1	509	6	2	3
Class 5	1	0	0	0	40	464	0	2
Class 6	5	19	0	19	21	31	385	14
Class 7	0	6	0	0	28	2	2	458

Fig. 9: Training and Validation Accuracy/Loss for Multiclass Model.

Fig 10: Confusion Matrix for Multiclass Model.

6 Evaluation

The results of the experiments conducted to study the Wisconsin Diagnostic Dataset, and the BreakHis Dataset are analyzed in detail in this section. The findings are supported by visual aids, and statistical comparisons are performed; hence, strengths and limitations of each model are highlighted.

6.1 Experiment 1: Evaluation of Machine Learning Models on Wisconsin Dataset

Wisconsin dataset was run on seven machine learning models and evaluated by the criteria of accuracy, precision, recall, F1-score, and AUC-ROC.

Model Comparison

The comparative performance of models is visualized in Table. 1, in which Random Forest emerges as the best-performing model considering its accuracy of 97.37%, precision of 97.62%, and F1-score of 96.47%. Other good performers are Logistic Regression and

XGBoost; however, Random Forest consistently led the performance across all criteria compared to other models.

The confusion matrix for Random Forest (Fig 4) shows that:

- 72 benign cases were correctly classified.
- 39 malignant cases were correctly classified.
- malignant cases were misclassified as benign.

Validation Metrics

- Accuracy: The overall accuracy of the model is 97.37%, which proves its high reliability in distinguishing between benign and malignant cases.
- Precision and Recall: Random Forest yielded the best performance with 97.62% precision and 95.35% recall, hence very few false positive or false negative.
- Confusion Matrix Analysis: Fig. 4 reveals the confusion matrix, wherein the model can rightly identify most of the benign and malignant cases with a few misclassifications. This confirms that Random Forest is robust against imbalanced data.
- Feature Importance: The feature importance from the Random Forest showed that concave_points1, peremeter3, radius3 and area3 were the strongest predictors that contributed most to the classifier's performance.

6.2 Experiment 2: Binary Classification on BreakHis Dataset

Model Performance

The binary CNN model was able to achieve 99.94% test accuracy with a test loss of 0.6683. The performance shows the capability of the model in classifying the images with minimum errors.

Validation Metrics

The training vs. validation accuracy/loss plots - Fig. 5 depict the following:

- Validation accuracy was always higher than training accuracy and reached a higher final value.
- Validation loss was lower compared to the training loss, which depicts good generalization.

This behaviour shows the consequence of regularization and data augmentation not to overfit.

Precision-Recall Curve

The precision-recall curve of Fig. 7 is a balance between precision and recall, with a minimum number of false positives and false negatives. The model has kept a high precision with a good recall rate to prove its reliability in distinguishing the images as either benign or malignant.

Confusion Matrix Analysis

The confusion matrix in Fig. 8 confirms the model's performance. No major misclassifications happen, thus showing the model's accuracy across the dataset.

6.3 Experiment 3: Multiclass Classification on BreakHis Dataset

The BreakHis dataset was also used for multiclass classification, categorizing images into eight tumor types

Model Performance

The multiclass CNN model yielded a test accuracy of 90.98% and a test loss of 1.5174. This suggests excellent performance, though a little worse than the binary model due to increased difficulty in distinguishing eight classes.

Validation Metrics

Training and validation accuracy/loss plots (Fig. 9) are presented below:

- Validation accuracy outperformed training accuracy throughout the training.
- Validation loss was always lower than the training loss, which shows good generalization.

Confusion Matrix Analysis

Fig. 10 shows the confusion matrix that represents the classification performance for each of the eight classes in detail. Overall, the model performed well, though slight misclassifications can be observed between closely related classes, such as some subtypes of benign and malignant classes. These results indicate possible further improvements using advanced architectures or extra data augmentation.

6.4 Comparative Analysis Across Experiments

Overall Performance

Comparative analysis shows that:

The Random Forest algorithm behaved the best on structured data, classifying 97.37% of Wisconsin data. The binary CNN model presented perfect results, achieving an accuracy of 99.94%. Multiclass CNN performance was at 90.98, yet it is quite good as far as the task was complex enough.

Challenges and Insights

These are the main challenges: class imbalance in the BreakHis dataset and the complexity of multiclass classification. These issues have been mitigated by using regularization and augmentation, thus ensuring strong generalization

7 Conclusion and Future Work

Conclusion

The given research was focused on the development of a machine learning framework for the classification of breast cancer using two of the most studied datasets: the Wisconsin Breast Cancer Dataset and the BreakHis dataset. Among other objectives, this included the pre-processing of these datasets, implementing machine learning models for structured data, designing and training deep learning models for histopathological images, and evaluating their performances using standard metrics. Throughout the study, we effectively addressed these objectives and achieved major results.

Key results of this analysis are:

- Random Forest was identified as the top model for Wisconsin with 97.37% accuracy, 97.62% precision, and 95.35% recall. The nature of the classes can be seen through the confusion matrix, which revealed very few misclassifications in this study between benign and malignant, adding weight to its robustness on structured data classification.
- In the case of the BreakHis dataset, binary classification with a CNN achieved a test accuracy of 99.94%, which showed that the model could tell the difference between benign and malignant cases.
- Multiclass classification of the BreakHis dataset into eight tumor subtypes using a CNN architecture was done, with an accuracy of 90.98% on test data. In fact, the model was able to perform such a complex task with an accurate detailed classification of the various tumor classes, as shown by confusion matrix and precision-recall analysis.

These findings emphasize how effective machine learning and deep learning techniques are in improving the accuracy of diagnosis concerning breast cancer. This could significantly improve early detection and treatment planning if integrated into clinical workflows.

Implications

The findings of the present study have significant implications for theoretical and practical perspectives:

- Academic Contribution: The findings add to the literature on machine learning and medical image analysis, especially in the field of diagnosis of breast cancer. The insights gained from this research will help guide future studies on how structured and image data are combined to obtain robust diagnostic models.
- Practical Implications: The models developed, especially the CNN regarding the classification of histopathological images, will help pathologists identify malignancies more accurately and with more speed. Feature importance analysis from Random Forest may also be used to glean further insight into the important factors contributing to the diagnosis of breast cancer.

Limitations

Despite these promising results, certain limitations were faced during the study:

1. Class Imbalance Problem: Even though data augmentation was done to reduce class imbalance, originally such imbalance present in the BreakHis dataset did create some problem during model training.
2. Computational Resources: Most deep learning-based models, if big datasets are considered, require high computational power; hence, scaling this research may be restricted.
3. Generalizability: All the models are trained and tested on particular datasets; whether they generalize well on other datasets or even in the real clinical world remains to be seen.

Future Work

The present research opens several avenues for further investigation:

1. **In corporealization of Advanced Methods**: Future studies can be done by using state-of-the-art architectures based on either a ViT-based or hybrid CNNs model combined with an attention mechanism to improve the classification performances.

2. **Explainability and Interpretability**: Developing explainable AI methods with embedded visual explanations for the AI model predictions assists in gaining trust and further enhances usability among the healthcare professional fraternity.
3. **Real-World Validation:** Generalisability and strength of the models need to be tested in various and larger clinical trials using real-world datasets.
4. **Clinical Workflow Integration:** Needless to say, the development of complete diagnostic solutions that integrate the models into clinical workflows will enhance value by way of real-time analytics and decision-making.
5. **Multimodal Approach:** Integrating different data from other modalities, including genetic and radiological imaging, will offer a comprehensive diagnosis.

The study has identified how machine learning and deep learning can be used to improve diagnostics of breast cancer. Further research into the identified limitations, following the proposed future directions, will go a long way in the development of more useful, interpretable, and impactful diagnostic tools in the struggle against breast cancer.

References

Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5-32.

Chen, T. and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A., 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence*.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. and Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4681–4690).

Quinlan, J.R., 1996. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4, pp.77–90.

Spanhol, F.A., Oliveira, L.S., Petitjean, C. and Heutte, L., 2016. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7), pp.1455–1462.

Ahmad, W., Ghafoor, A., Shafique, M.A., Jameel, T. and Yaqoob, I., 2022. Super-resolution techniques for medical imaging: A survey and taxonomy. *IEEE Access*, 10, pp.12115–12142.

Sumathi, M., Rani, S. and Revathi, M., 2023. Hybrid deep learning model for ovarian cyst classification. *International Journal of Advanced Research in Computer Science*, 14(3), pp.56–67.

Subramani, K., Venkatesh, R., Manoharan, R. and Kumari, S., 2023. CNN-MLP hybrid approach for PCOS detection using ultrasound images. *ICSES Proceedings* (pp. 1–7).

Purnama, B., Andriyanto, R. and Prabowo, P.S., 2015. Breast cancer classification using KNN and statistical features. *Proceedings of ICoICT* (pp. 396–401).

Liang, J., Zheng, Y., Cai, W. and Yang, F., 2022. Progressive GANs for realistic breast ultrasound image synthesis. *Medical Image Analysis*, 79, p.102461.

Shashank, H., Rajeshwari, R. and Vijaya, S., 2023. Explainable AI for CNN-based medical imaging systems. *IEEE Transactions on Medical Imaging*, 40(5), pp.1234–1247.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J. and Greenspan, H., 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321, pp.321–331.

Suha, S.A. and Islam, M.N., 2022. Machine learning for breast cancer classification using structured data. *Scientific Reports*, 12, p.17123.