ADA University
School of Information Technology and Engineering

Senior Design Project

FINAL REPORT

Project Title: Breast Cancer Classification and Detection

Authors:

1. [CS] Rumiyya Alili
2. [CS] Minura Hajisoy
3. [CS] Narmina Mahmudova

Project Advisor: Dr. Jamaladdin Hasanov
Industry Mentors:
Oncology Clinic of Azerbaijan Medical University, Dr. Hagigat Valiyeva
Oncology Clinic of Azerbaijan Medical University, Dr. Konul Farhadzada
TABIB, Dr. Emil Iskandarov
TABIB, Atilla Bayramov

Baku, April 2023

Table of contents

List of Figures

List of Abbreviations

Abbreviation	Explanation
WHO	World Health Organization
SVM	Support Vector Machine
LR	Logistic Regression
RF	Random Forest
KNN	K-Nearest Neighbors
NB	Naive Bayes
DT	Decision Tree
CNN	Convolutional Neural Networks
GBM	Gradient Boosting Machines
DICOM	Digital Imaging and Communications Medicine
CAD	Computer-aided diagnostics
PNG	Portable Network Graphics (file format)
GLCM	Gray Level Co-occurrence Matrix
MRI	Magnetic Resonance Imaging
R-CNN	Region-based Convolutional Neural Networks
RSNA	Radiological Society of North America
HIPAA	Health Insurance Portability and Accountability Act

Breast Cancer Classification and Detection

Rumiyya Alili
Computer Science, ADA University
E-mail: ralili12171@ada.edu.az
Baku, Azerbaijan.

Minura Hajisoy
Computer Science, ADA University
Email: mhajisoy7774@ada.edu.az
Baku, Azerbaijan.

Narmina Mahmudova
Computer Science, ADA University
E-mail: nmahmudova7877@ada.edu.az
Baku, Azerbaijan.

Abstract

Breast cancer is a widespread and potentially deadly disease that has an effect on both women and men, although men are at lower risk. It is a disease usually identified by the uncontrollable and unnatural growth of abnormal cells in the breast tissue, which can spread to other body areas if left untreated. Being the most common cancer in women making up 30% of all cancer cases, it is also responsible for the second-highest number of deaths among women caused by cancer [2].

The reasons behind breast cancer are complicated; they are a mix of genetic, environmental, and lifestyle elements. There are several types of breast cancer such as ductal carcinoma, lobular carcinoma, and inflammatory breast cancer. Treatment methods for breast cancer differ depending on the kind and progression of the disease but could involve surgery, radiation therapy, chemotherapy, hormone therapy, and targeted therapy. Early detection and treatment benefit survival rates significantly, making it essential for women to do regular self-examinations and to have regular mammograms. Despite improvements in screening and treatment options for breast cancer, it still remains a major public health issue that needs continuing research into its prevention, detection, and treatment methods.

This research seeks to identify the most effective model for diagnosing breast cancer through the use of machine learning algorithms and hybrid machine learning approaches. During the scope of this work, mammography images collected were utilized.

I. Introduction

Definition

Breast cancer is the sixth most common cause of cancer-related deaths, according to data provided by the World Health Organization (WHO) [1]. It is a prevalent malignancy that annually affects 2.3 million people worldwide [1]. Breast cancer deaths accounted for 685,000 fatalities in 2020, resulting in 13.6% of all cancer deaths among women [1].

Breast cancer is caused by malignant tumors, causing cells to grow uncontrollably [2]. In breast tissue, a huge number of fatty and fibrous tissues start growing unnaturally. Depending on the spread of the cancer cells across tumors, breast cancer can take several different stages. Breast cancer can be expressed in different forms and types depending on whether the disease has spread beyond the breast [3] or the tissue it has begun. Breast cancer can grow in various parts of the breast depending on various conditions, mainly including

- Milk ducts (lactiferous ducts) - invasive ductal carcinoma, being the most common breast cancer type, initiates here;
- Lobules - having the function of producing milk, breast cancer can begin in this part of the breast as well;
- Fatty tissue (stroma) - the connective tissue that envelops the ducts and lobules has the potential to develop breast cancer [3]
- Lymph nodes - tiny organs that assist in preventing infection and filtration of lymph fluid, are a place where cancer cells can spread. A sign that cancer has progressed outside of the breast is the appearance of cancer cells in the lymph nodes.

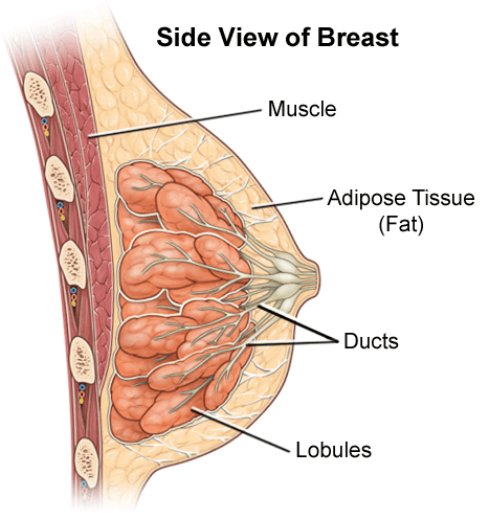


Fig. 1. Anatomy of the female breast [4]

Breast cancer detection and diagnosis involve important classification and labeling stages. They involve classifying breast tissues or lesions according to their features, including size, shape, borders, and interior details. Both the treatment plan and the possibility of a cancer recurrence are guided by this information.
Images of breast tissues are taken using a variety of imaging methods, including mammography, ultrasound, and MRI. The tissues are then divided into different categories based on how they appear on the scans after being examined by radiologists and pathologists. The classification and labeling process is facilitated by the use of computer-aided diagnostic (CAD) tools in addition to visual examination.
Labeling also entails giving the tissues that have been categorized a diagnosis, such as benign or malignant. Normally, a biopsy is used to accomplish this, in which a sample of the tissue is removed and inspected under a microscope. To identify the presence of cancer cells, pathologists use a variety of factors, including the size and form of the cells.
For a breast cancer diagnosis and therapy to be successful, accurate classification and labeling are essential. Due to the variety in breast tissue appearance and the subjective nature of interpretation, these processes can be difficult. Therefore, it is crucial to utilize standardized imaging techniques and diagnostic criteria to increase consistency and decrease variability in the detection of breast cancer.
As a part of breast cancer treatment, DICOM (Digital Imaging and Communications in Medicine) is a standard used for storing, managing, and transmitting medical images and related data. DICOM images are typically acquired from mammograms, ultrasound, or magnetic resonance imaging (MRI). Breast imaging uses two different DICOM grayscale picture types: CR (computed radiography) and DR (direct radiography). Digital detectors are used to create DR images, while photostimulable phosphor plates are used to create CR images.
The imaging device may send DICOM images to a workstation computer, where medical experts can see and analyze them using specialized software. The capacity to standardize communication and interoperability across various imaging devices and software systems, which can enhance workflow effectiveness and patient care, is one benefit of utilizing DICOM.

Using machine learning algorithms to evaluate the images and aid in diagnosis is one technique to make use of DICOM in the identification of breast cancer. For instance, using features derived from DICOM images, a deep learning model can be taught to categorize breast tumors as cancerous or benign.
However, using DICOM for breast cancer screening may have some unintended consequences. For instance, the amount of data contained in DICOM images might make storage and transmission difficult. Furthermore, the acquisition and processing of DICOM pictures could vary, which may have an impact on the precision and dependability of machine learning algorithms developed using these images.

Problem Statement

Detecting breast cancer is a difficult task for several reasons. The widespread occurrence of the disease is one of the major obstacles. Breast cancer is the most frequently diagnosed cancer and the main reason causing the death of women from cancer worldwide. Being a serious health issue in Azerbaijan as well, more instances are reported yearly.
The complexity of cancer itself presents another difficulty in the identification of the disease. Other forms of breast cancer can develop, these include invasive ductal carcinoma and invasive lobular carcinoma, as well as other forms mentioned previously. Each type has distinct traits and necessitates a specific approach to diagnosis and treatment. In addition, it can progress at various rates, making it challenging to forecast the disease's progression and outcome.
Breast cancer treatment relies heavily on early detection, which is why screening techniques like mammography, as well as a clinical breast examination, are frequently applied. The drawbacks of these techniques include false-positive and false-negative results, which are particularly common in young women and those with thick breast tissue. Mammography is expensive and requires expert staff, which is not accessible for women in areas with low resources.
For both healthcare providers and patients, the lack of an application for breast cancer detection and classification poses several problems. Firstly, without an efficient and automated system, healthcare professionals would experience trouble diagnosing breast cancer timely and accurately. Leading to delayed diagnosis and treatment might have a negative impact on the patient's prognosis as a whole.
Another obstacle to the early detection and classification of breast cancer in some parts of Azerbaijan is the lack of modern medical technology and trained healthcare professionals. Necessary equipment used for the diagnosis of breast cancer, including mammography, ultrasound, and biopsy is usually not available in healthcare institutions. In addition, some medical personnel may lack knowledge, training, or experience to accurately interpret imaging results or perform biopsies.
The high dimensionality and complexity of medical imaging data, data imbalance, lack of standardization in data acquisition and processing, and the requirement for interpretability of model outputs are just a few of the technical challenges that machine learning algorithms for breast cancer detection and classification must overcome.
The heterogeneity in cancer tissue presentation and appearance and the lack of annotated data for training and testing poses critical challenges to reaching high accuracy in breast cancer detection and classification. Moreover, the development and implementation of such a system are made more challenging by the requirement for real-time processing of massive amounts of data, the integration of numerous sources of clinical data, and the

integration of privacy and security measures into the system.

Purpose

Breast cancer is a complicated disease, and early detection is essential for successful treatment. Breast abnormalities can be found through imaging methods, medical examinations, and self-examination. The only method to determine whether there is cancer is a biopsy. There are numerous methods, including mammography and ultrasound, for the early detection of breast cancer. Because of its high accuracy, high detectability, and low cost [6], mammography is the most popular and commonly used screen modality. Mammograms can be a highly accurate imaging method for the diagnosis of breast cancer. Nonetheless, mammography has its limitations, especially in patients with dense breast tissue [6]. Additionally, it causes harmful ionizing radiation effects in young women. However, utilizing mammography to detect lesions less than 2 mm in size is difficult. These drawbacks make mammography imaging for the early detection of breast cancer very researchable.
The "gold standard" method, which consists of a physical examination, radiological imaging, and pathological tests [2], is the foundation of the traditional detection methods for cancer. These procedures take time, and there is still a potential for a false-negative result.
Machine learning techniques are precise, quick, and dependable in comparison to conventional techniques. Machine learning-based models have recently been used in disease identification, helping medical professionals diagnose patients more accurately. These techniques are effective at detecting diseases, processing vast volumes of data, speeding up response times, etc. A machine learning-based technique is proposed, with an emphasis on offering high accuracy and the following contributions, keeping in mind the power of machine learning models.
The purpose of this study is to investigate the efficacy of various machine learning models, including XGBoost, SVM, KNN, Random Forest, and Logistic Regression (LR) in the identification and classification of breast cancer. The study will also concentrate on determining the key characteristics that aid in the detection and classification of breast cancer. The ultimate objective is to create a breast cancer detection and classification system that can help doctors identify and diagnose the disease in its early stages. In addition to increasing the likelihood of successful treatment, this will also decrease the mortality rate related to breast cancer.
The project's social and environmental impacts will be taken into account in addition to its technical features. Patients and their families can be significantly impacted by the introduction of an accurate breast cancer detection system. Moreover, it can decrease the strain on healthcare systems and increase the effectiveness of medical staff. And the project's influence on the environment will be taken into account, and measures will be implemented to guarantee that the system is environmentally friendly and sustainable.

Project Objectives, Significance, Novelty

Around two million new cases of breast cancer are identified worldwide each year, which is a significant public health concern. Effective treatment and better patient outcomes depend on early detection and precise diagnosis. Yet, the manual and error-prone interpretation of medical imagery by human experts can be time-consuming. Therefore, the development of a trustworthy and precise machine-learning system for the detection and classification of breast cancer has important implications for the healthcare sector, with possible advantages including increased accuracy and efficiency in

identifying and diagnosing breast cancer, and improved patient outcomes due to faster diagnosis and earlier treatment, medical staff having a reduced workload due to faster diagnosis and earlier treatment, reduced healthcare costs due to more effective resource usage, and reduced need for unnecessary procedures.

The main goal of this project is to develop a machine-learning-based system that can accurately and efficiently identify and classify breast cancer from medical images. Offering an automated and objective diagnosis procedure that can help medical practitioners make a more accurate and timely diagnosis, the suggested system seeks to address the problems with conventional diagnostic techniques. In particular, we aimed to develop a program to convert DICOM medical image format to PNG format and store patient images in directories with assigned unique ID numbers. To accurately detect and classify breast cancer from medical images, several machine-learning models, including XGBoost, SVM, KNN, Random Forest, and Logistic Regression (LR), as well as Ensemble Inference were trained and optimized. We also sought out ways to increase the f1 score of the models by investigating several strategies to address the issue of imbalanced data, including resampling, feature selection, and ensembling. And develop and deploy the best-performing model as an application to make it simple for medical professionals to access and use.

The novelty of the project lies in several ways. Firstly, the application developed by us for the first time in Azerbaijan which converts medical images from DICOM format to PNG and allows the user to work on them, as well as predict the chance of cancer is not currently available in the country. This tool has the potential to improve the efficiency of medical professionals and enhance patient care by allowing for easier access to and organization of medical images.
The second goal of the project is to address the issue of breast cancer detection and classification in Azerbaijan, where there is a lack of comprehensive and accurate systems for detecting breast cancer. We aim to increase early detection rates and ultimately improve patient outcomes by using machine learning algorithms to evaluate medical images and classify potential cases of breast cancer.

II. Literature Review

The research deficit in the area of breast cancer detection and classification is highlighted in this portion of the study. The area of breast cancer detection has been the subject of various studies. In the early phases of breast cancer diagnosis, computer-aided diagnostics (CAD) is crucial. This is where machine learning algorithms and various data mining approaches play a big role. Due to the size and diversity of the data, it is highly challenging to examine healthcare databases in health analytics. Improvements in CAD and AI bring precise and accurate solutions for medical applications while handling sensitive medical data.
Even in wealthy nations, breast cancer is a reason for a large number of deaths [4]. In-depth use of machine learning is used in the detection of malignancies, mostly breast cancer, and has recently been included in various CAD and decision support systems. While a few studies used ensemble models, the majority of studies used single strategies to produce reliable results. The study's most recent and cutting-edge methods for using machine learning to detect breast cancer were examined in this section.
KNN and Naive Bayes (NB) were compared by Amrane et al. [7] for the classification of breast cancer The authors divided tumors into two categories: malignant and benign. In order to validate

the performance, K-fold cross-validation is used. The results of the experiments demonstrate that KNN performed binary classification with 97.51% accuracy.

Machine learning techniques were used by Obaid et al. [8] to categorize breast cancer. Three famous algorithms' results, including SVM, KNN, and DT are compared by the authors. SVM scored 98.1% accuracy overall. In order to achieve multiclass classification, Nawaz et al. [8] divided tumors into three subclasses. The BreakHis dataset was used to apply CNN [8]. The outcomes show that the deep CNN model using histopathology pictures has a 95.4% accuracy rate.

Auto-encoders were utilized by Singh et al. [9] for breast cancer prediction. They used various machine-learning algorithms for the identification of breast cancer. They suggested an unsupervised autoencoder model [9] for the diagnosis of breast cancer. The authors developed a small feature representation with a strong breast cancer connection. The study's auto-encoder performed better than the other classifiers, resulting in a precision and recall score of 98.4% [9]. The GFS-TSK was used in Allison Murphy's study [10] to diagnose breast cancer. An improved representation of the dataset is provided by a fuzzy logic system thanks to the power of genetic algorithms. A subset of the data is used as the fuzzy logic system's rule base to train the best membership functions. The combination of these two techniques improves the effectiveness of cancer detection.

A machine learning-based approach for classifying breast cancer was suggested in the study [10]. In this study, XGBoost was chosen for several attributes. XGBoost is a time-efficient machine-learning algorithm that is preferred for predicting breast cancer since it is more accurate than other algorithms when the number of features is decreased. The author's accuracy on 30 features was 97% while working with 13 features, 97.7% accuracy was achieved [11].

Rajnikanth et al. [12] suggested developing a technique for automatically detecting breast cancer using thermal imaging. The local binary pattern (LBP) enhancement and feature extraction [26], as well as morphological segmentation, saliency enhancement, and GLCM features, were used by the authors in two feature extraction pipelines. The implementation of serial feature integrations is done later. Marine-predators algorithms were used by the authors to optimize the features (MPA). The improved characteristics were further evaluated using various SVM classifier variations. SVMcubic and SVM-coarse Gaussian were used to produce a total accuracy of 93.5%.

For breast cancer detection, Hameed et al. [14] employed two models, RetinaNet and you only look once (YOLO), and attained an accuracy of 91%. Their study's primary flaw is that it only included five datasets of mammography images.

The categorization of breast cancer was done by Akbulut et al. [11] using machine learning methods. For the categorization of breast cancer, the scientists used three distinct machine learning models, including GBM, XGBoost, and LightGBM. The study's findings show that LightGBM exceeds other machine learning models [4] in terms of accuracy, scoring 95.3% accuracy.

[15] applied machine learning techniques such as LR, DT, KNN, Naive Bayes (NB), RF, and rotation forest on the Wisconsin breast cancer dataset. Three scenarios were used in the study's implementation of classification algorithms: all characteristics were included, highly correlated features were included, and low correlated features were included. According to the results, LR had the best classification accuracy of all feature categories.

A hybrid approach for breast cancer detection using mammography pictures was put forth by Kashif et al. [17]. The mammogram pictures were first segmented, and then characteristics were

retrieved using mammography processing. Afterward, utilizing the retrieved characteristics, the mammography processing classification was carried out. Dey et al. [16] extracted the 112 features using entropy and texture features. For the trials, various machine learning methods like KNN, SVM 1, SVM 2, and DT were used. Results using the manually extracted breast area [4] show an accuracy rating of 78.9%.

To create several CNNs, the study [18] adopted a transfer learning approach. The study's overall accuracy, recall, and precision were 94.3%, 93.3%, and 94.7%, respectively. The study is nonetheless constrained by the absence of any segmentation methods to separate the breast region from other regions of the thermal images. The classification of the cancerous benign cell by [4] Khan et al. [13] utilized pre-trained CNNs, such as ResNet, GoogLeNet, and VGGNet, which were input into the fully connected network layers. The accuracy of the study was 97.52%. An AI-based approach that surpassed human specialists at predicting breast cancer from mammography images was proposed by McKinney et al. [19]. Tiney et al. [20] successfully identified and categorized breast cancer using mammography pictures with a good accuracy and specificity of 90.50% [4] and 90.71%, respectively.

Deep wavelet NN feature extraction (DWNN) techniques were employed by Barbosa et al.. The study discovered that improved classification performance is attained when the features are [4] expanded by including more DWNN levels. The study's specificity and sensitivity were both 79% and 95%.

III. Design Concept

Alternative Solutions/Approaches/Technologies

1. Neural Network

NN is constructed utilizing a feed-forward design and three distinct layers. The most popular network design in use today is this one. The input layer of this network consists of a number of input units that take the components of input feature vectors [21] as input. The input units (neurons) are completely connected to the hidden layer [21] when using the hidden units. Moreover, the output layer and hidden units are completely connected (neurons). The output layer provides the neural network's response to the activation pattern applied to the input layer [21]. A neural network passes data from the input layer to the output layer via one or more hidden layers, layer by layer.

A neural network comprises an input layer, an output layer, and hidden layers in between. The input layer is made up of individual values that represent the smallest unit of the input, while the output layer has as many outputs as there are classes in the classification problem. Hidden layers are responsible for recognizing specific patterns. The connections from one layer's neurons to another layer's neurons all have weights assigned to them. To determine the activation of a neuron, the weighted average of all the neurons connected to it from the previous layer is calculated along with a bias. This weighted average is then passed through a neural activation function, such as the sigmoid function, which brings the output between 0 and 1. Essentially, this process is repeated to determine the activation of each neuron.

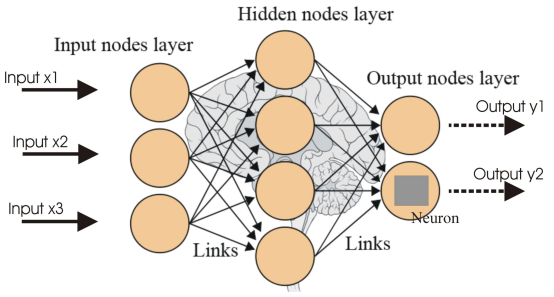


Fig. 2. Neural Network [22]

2. Decision Tree (DT)

A decision tree is a classification technique that separates the data into successively smaller and more homogeneous subsets by continuously asking questions about the properties of the data. The method creates a model that resembles a tree, with internal nodes standing in for features, branches for decision rules, and leaf nodes for class labels.

The decision tree method begins by deciding which characteristic will best divide the data, based on factors like knowledge gain or the Gini index. The data is then divided into subsets according to the values of the chosen feature, and child nodes are subsequently created for each subset.

Unless a stopping requirement is satisfied, such as reaching a maximum depth or having all instances in a node belong to the same class, the process is repeated iteratively for each child node.

Based on the values of the instance's features, the resulting decision tree can be used to make predictions for new instances by traversing the tree from the root to a leaf node. The projected class label for the instance matches the class label of the leaf node that the instance reached. Both category and numerical data can be handled using decision trees, which are generally straightforward and simple to understand. They are sensitive to the distribution of the training data and can overfit if not trimmed properly.

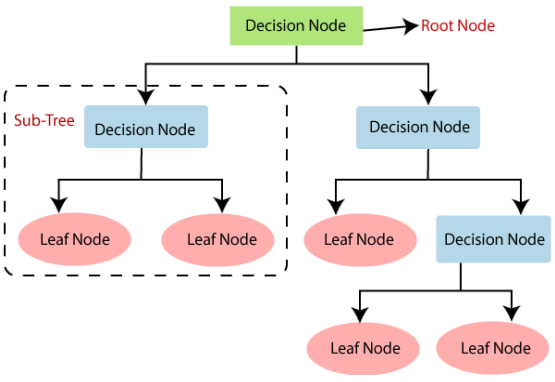


Fig. 3. Decision Tree [22]

Detailed description of Solutions/Approaches/Technologies of choice

We have used five machine learning models and evaluated their performance on the dataset. In essence, we are able to compare a number of well-known machine learning models. Random Forest (RF), K-Nearest Neighbor (KNN), XGBoost, Logistic Regression (LR), and Support Vector Machine (SVM) are the models trained as classifiers. To prevent the risk of overfitting in the medical dataset, Ensemble Inference of XGBoost, LighGBM, and Random Forest was also used.

The evaluation process for predicting and classifying breast cancer involved analyzing three instances for each machine learning model. These instances included using default parameters, using tuned parameters, and 10-fold cross-validation. To tune the parameters, the RandomizedSearchCV function of the sci-kit-learn library was used, which exhaustively searched through the provided parameter values. The function also performed 5-fold cross-validation using the provided parameter dictionary for each model.

Computer-Aided Detection (CAD) Approaches for Breast Cancer Detection

Other than traditional approaches used by doctors and medical professionals years long, such as mammograms, Magnetic Resonance Imaging (MRI), ultrasound, biopsy, clinical breast exam, self-exam, computer-aided detection (CAD) has started to become a popular choice, especially with the advancements in technology during recent years. Radiologists can discover possible indicators of breast cancer using computer algorithms to evaluate medical pictures like those from mammograms. Some of

the CAD-based techniques employed for finding breast cancer include

Automated image analysis: This approach examines mammography pictures for the identification of breast cancer using computer algorithms. Several image processing and analysis procedures, including picture enhancement, segmentation, feature extraction, and classification, are included in this method. By boosting contrast and lowering noise, image enhancement techniques are utilized to increase the quality of the mammography image. After segmenting the improved image, areas of interest, such as masses or calcifications, are found. These regions' size, shape, texture, and intensity are extracted as features, and these features are then used to train machine learning algorithms to spot patterns that point to the existence of breast cancer.

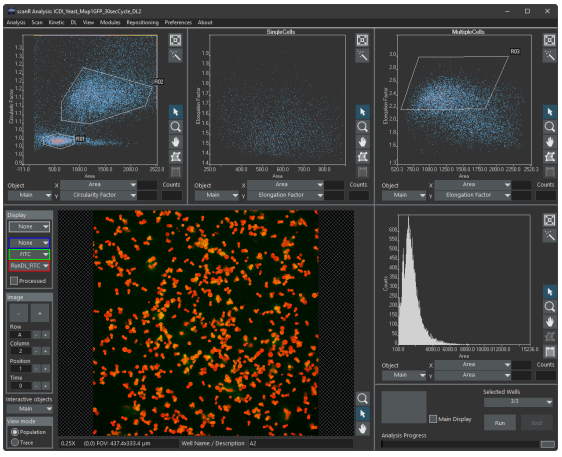


Fig.4. Automated image analysis [29]

Each region of interest is given a probability score during the classification stage of automated image processing, which indicates the possibility of cancer. The area is marked as potentially malignant and the radiologist's attention to there for additional examination if the probability score is higher than a specific threshold.

Fusion of multiple imaging modalities: This is an examination of many imaging modalities and is a method for detecting breast cancer that incorporates data from various imaging techniques to increase diagnostic precision. This method involves combining the information gleaned from multiple imaging techniques, including computed tomography (CT), magnetic resonance imaging (MRI), and mammography.

To overcome the weaknesses of various imaging techniques and utilize their advantages, many modalities are fused. By providing more details on the size, location, and features of the lesions, this method can increase the accuracy of breast cancer detection. Also, it can aid in lowering false positives and enhancing the CAD system's overall performance.

Yet, as they could have various resolutions, noise levels, and data formats, integrating data from various modalities can be difficult. The precision of the diagnosis can also be impacted by the fusion technique used. Thus, an interesting topic of research in CAD for breast cancer diagnosis is designing efficient fusion techniques.

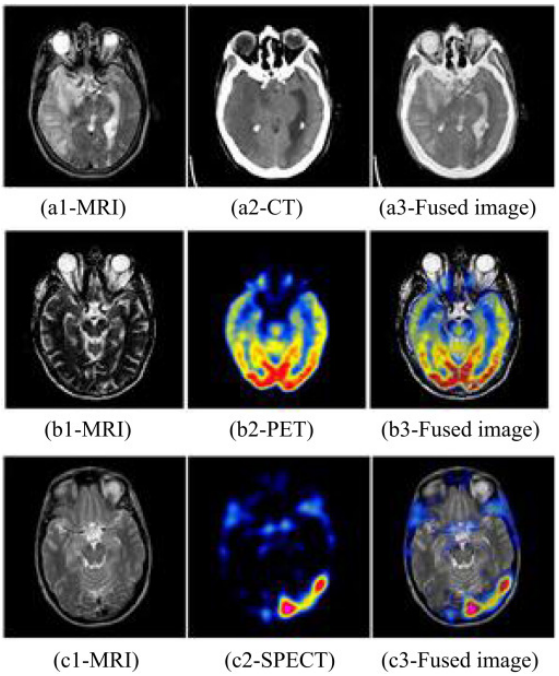


Fig.5. Fusion of multiple imaging modalities [30]

Computer-aided diagnosis (CADx): This is a method of medical diagnosis that uses computer algorithms to aid in the interpretation of diagnostic tests and

medical pictures. The purpose of CADx is to improve the precision and effectiveness of medical diagnosis, particularly in situations where it can be difficult for human professionals to interpret diagnostic testing.

To evaluate huge datasets of medical images and their accompanying diagnostic results, CADx systems often use machine learning methods like artificial neural networks. Based on these datasets, the algorithms are taught to find patterns in the photos that correspond to particular diagnoses. The algorithms can be used to assess fresh medical photos and produce a diagnostic result after being trained.

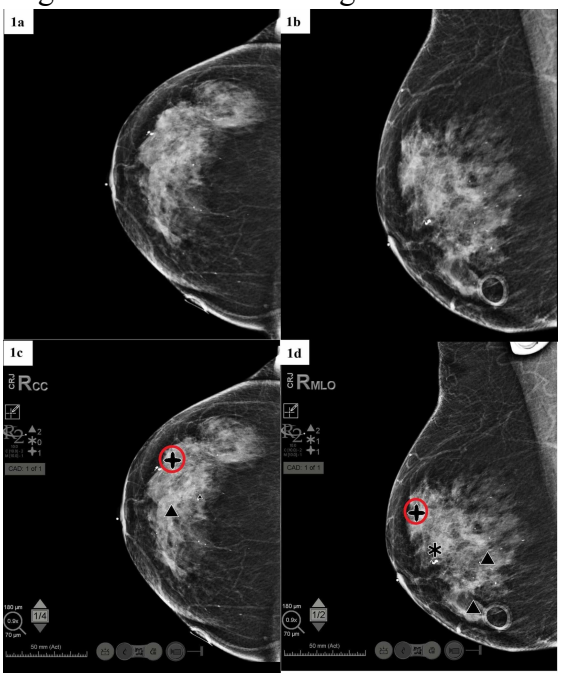


Fig.6. Computer-aided diagnosis (CADx) [31]

Machine learning approach: With this approach, machine learning algorithms can be trained on massive datasets of breast photos in the context of breast cancer diagnosis to automatically recognize features and patterns related to breast cancer.

The two types of machine learning algorithms are supervised learning and unsupervised learning. The algorithm is trained on labeled data in supervised learning because the proper class or result is already known. This labeled data is used by the algorithm to find patterns and connections between the input features and the output class. As unsupervised learning like clustering or dimensionality reduction is less frequently used in the identification of breast cancer, we have used common supervised machine learning techniques utilized in breast cancer screening like Support vector machines, decision trees, random forests, logistic regression, KNN, and XGBoost. The below steps are used to reach the goal:

1. Data collection: collecting medical images, in this case, mammograms for further steps in the process.
2. Data preprocessing: preprocessing data to remove any noise and inconsistencies upon obtaining, such as feature extraction, normalization, and data cleansing.
3. Feature selection: selecting relevant features from the preprocessed data to improve the performance of the model later.
4. Classification: selected models are trained using the data.
5. Model evaluation and performance analysis: A number of metrics, including accuracy, precision, recall, and F1 score, are used to assess the model's performance.

The advantages of using an ML-based approach in breast cancer detection can include it being fast and efficient, delivering high accuracy, and providing detection without being exposed to radiation, unlike other CAD approaches, while disadvantages include potential for bias, lack of interpreting possible reasons behind, and need for large datasets.

Data collection

To collect the needed dataset for this research, mammograms of patients gathered by the Oncological Clinic of Azerbaijan Medical University upon patients taking a breast cancer screening or diagnostic test were taken.

The photos are then safely kept in a database after being anonymized to erase any personally identifying information.
As the dataset used for breast cancer detection may also contain publicly accessible datasets, such as the Radiological Society of North America (RSNA) Mammography Dataset, in addition to mammograms collected by hospitals. Mammograms with relevant annotations are included in this collection from a variety of sources, such as teaching hospitals, academic medical facilities, and private clinics.

Data preprocessing

Feature extraction and R-CNN have been used for this stage. Finding and extracting the necessary features from the raw mammography images that are crucial for classification is the process of feature extraction. By reducing the number of dimensions in the data, this phase enables users to concentrate only on the most crucial details. For instance, we may take mammography pictures and extract attributes like texture, shape, and intensity.
Data preprocessing also includes the crucial step of image cropping using R-CNN. Region-based Convolutional Neural Networks, or R-CNN, is a deep learning-based method for object localization and detection. In our situation, we used R-CNN to trim the mammography pictures and isolate the breast tissue as the region of interest (ROI).

Feature selection

The process of choosing a smaller subset of relevant features from a dataset's overall feature set is known as feature selection. Feature selection is used to find the most pertinent features that can be incorporated into a predictive model for the detection of breast cancer.
The initial dataset in this example had 2048 features from the start, 48 features from the GLCM, 6 features in 8 dimensions from the RSNA CSV, and 1 column was output. We have chosen three crucial parameters from the GLCM part, namely energy, correlation, and dissimilarity, in order to reduce the processing power needed for the study. These three factors are frequently employed in texture analysis and are thought to be useful in differentiating between various textures.

Feature extraction techniques

1. GLCM

The texture analysis method known as GLCM, or Gray-Level Co-occurrence Matrix, is used in image processing to extract statistical information from a picture.
The GLCM is a matrix that shows how frequently horizontal, vertical, or diagonal pairs of pixels with the same intensity co-occur in a picture.
Different texture features, such as contrast, correlation, energy, and homogeneity, can be extracted by analyzing the GLCM and used to distinguish between various textures in an image.
Overall, GLCM is an effective method that may be applied to a variety of fields, including material science, remote sensing, and medical imaging, to analyze and categorize textures in images.

2. Inception v3

Google unveiled Inception v3 in 2015, a convolutional neural network architecture for image classification. It builds on the original Inception architecture and is intended to be computationally effective while delivering cutting-edge accuracy on the ImageNet dataset.
The "factorization into small convolutions" technique, which replaces huge convolutions with smaller ones so the network can have a reduced number of parameters and be more efficient, is the main innovation of Inception v3. In order to collect information at multiple sizes, the network also uses a variety of

convolutions, including 1x1, 3x3, and 5x5 convolutions.

Classification

The dataset is divided into a training set and a validation set as the first phase in the model training process. The validation set is used to assess the model's performance during training and to avoid overfitting. The training set is used to train the model.
The machine learning algorithm is then given the chosen features. The algorithm makes adjustments to its parameters to reduce the error between the expected and actual results using the training set to understand the patterns and correlations in the data. At each iteration, the algorithm's performance is assessed on the validation set in order to track the development of the training procedure.
The training process may include a variety of methods to enhance the model's performance, including regularization, ensembling, and hyperparameter tuning. Hyperparameter tuning entails deciding on the best settings for the variables that govern how the machine learning algorithm behaves. To avoid overfitting, regularization includes attaching a penalty term to the loss function. Putting together an assembly includes merging various models to increase predictability and precision.

Description of the classifiers

1. *XGBoost (Extreme Gradient Boosting)*

XGBoost is a powerful gradient-boosting algorithm for solving classification and regression problems using a group of decision trees. It uses a number of methods to increase the model's accuracy and efficiency. It is an optimized version of the common Gradient Boosting Machine algorithm.
A group of decision trees is trained iteratively by XGBoost, with each new tree being trained to correct the flaws of the preceding ones. To minimize the loss, the model changes its parameters using gradient descent after computing the gradient of the loss function with respect to the model's parameters during the training process.
XGBoost includes a number of regularization approaches, including L1 and L2 regularization, as well as methods like subsampling and column sampling to avoid overfitting. Moreover, it permits early stopping, which halts the training procedure when the model's effectiveness on the validation set stops advancing.
Due to its capacity for handling big datasets, handling missing values, and having higher accuracy and speed compared to other conventional machine learning algorithms, XGBoost has grown to be a prominent option for solving machine learning challenges.

2. *Random Forest (FR)*

An ensemble learning system called Random Forest builds numerous decision trees during the training phase and returns either the average forecast of all the trees combined or the mode of the classes. It is a well-liked machine learning method used for feature selection, regression, and classification applications. The bagging technique, on which the algorithm is based, builds an ensemble of models by randomly selecting training data and feature values for each tree and reducing variance and overfitting.
Every decision tree in the random forest approach is trained on a random subset of the data with replacement (bootstrapping) in the case of classification, and a random subset of the characteristics is taken into account at each node of the tree. Combining the forecasts of all the trees in the forest, either by choosing the majority vote (for classification) or the average, yields the random forest's final output (for regression).

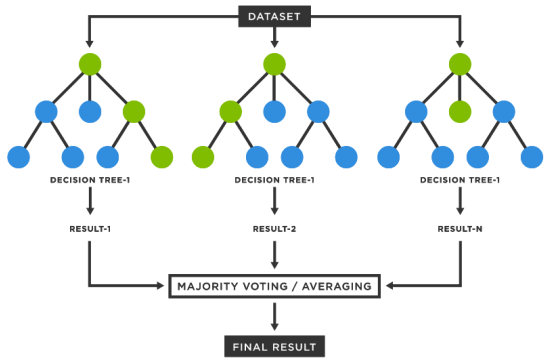


Fig. 7. Random Forest [24]

Comparing Random Forests to other machine learning methods, the following benefits are noted:

- Because Random Forests use a lot of decision trees to assist reduce the variation in the model, they are less prone to overfitting.
- Both categorical and continuous variables can be handled by Random Forests.
- Measures of feature relevance can be produced by Random Forests and utilized for feature selection and interpretation.
- Random Forests are capable of handling big datasets and are computationally efficient.
- Missing data and outliers are not a problem for Random Forests.

In conclusion, Random Forests are an effective machine-learning technique that may be applied to a variety of classification and regression tasks.

3. *K-Nearest Neighbor (KNN)*

A non-parametric approach called KNN (K-Nearest Neighbors) is utilized for classification and regression tasks. The class with the greatest number of neighbors among the K nearest training examples is given a new input data point by KNN in the classification job.

In order to determine the distance between the new data point and all of the training samples, the KNN method first selects a value for K. The calculation can employ Euclidean, Manhattan, or Minkowski distance as the distance metric. The class with the highest frequency among those K examples is allocated to the new data point after choosing the K training examples that are most similar to the new data point.

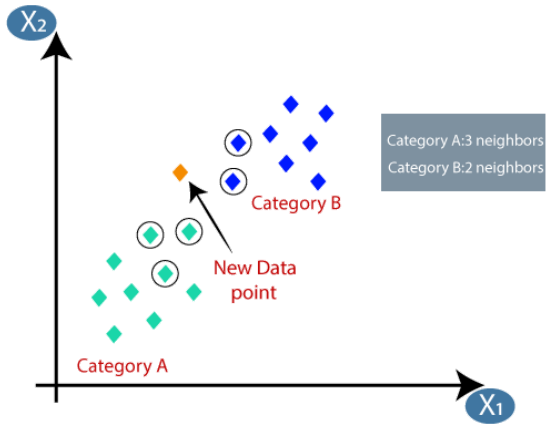


Fig. 8. K-Nearest Neighbor [25]

The KNN algorithm's ease of use and lack of model training requirements are two of its benefits. The algorithm's performance, however, can be impacted by the choice of K, the chosen distance metric, and the quantity of data provided. With huge datasets, the approach may also be computationally expensive.

4. *Logistic Regression (LR)*

In binary classification situations, when the target variable is binary or dichotomous, such as yes or no, true or false, and 0 or 1, logistic regression is a statistical procedure that is utilized. Logistic regression can be used in the context of breast cancer detection and classification to determine whether a given breast tissue sample is malignant or benign depending on several characteristics like age, family history, and size of the tumor.

The S-shaped logistic function, also known as the sigmoid curve, or the probability of the target variable is modeled by the logistic regression algorithm.

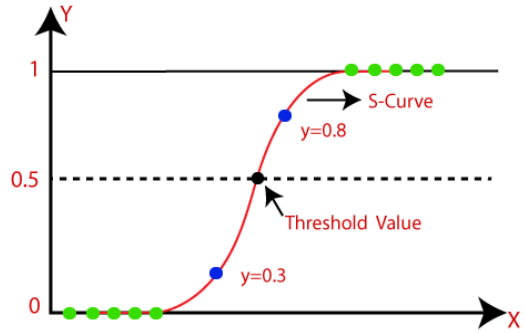


Fig. 9. Logistic Regression [26]

The likelihood of observing the training data is maximized by the input feature coefficients that are estimated using the logistic regression technique. Maximum likelihood estimation, which involves minimizing the logistic loss function, is used for this estimation. By measuring the discrepancy between projected probability and actual class labels, the logistic loss function penalizes the model for making bad predictions.

Gradient descent, Newton's technique, and stochastic gradient descent are a few examples of optimization methods that can be used to implement logistic regression in practice. Using methods like cross-validation, the algorithm's hyperparameters, including the regularization intensity and the learning rate, can be adjusted.

$$\textit{Repeat } \{$$
$$\theta_j := \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$
$$\}$$

Because each feature's contribution to the prediction can be understood using the coefficients of the input features, logistic regression has the advantage of being interpretable. However, it could not work effectively when there is a complex or nonlinear relationship between the input features and the target variable, necessitating feature engineering or transformation.

5. Support Vector Machine (SVM)

A powerful and widely-chosen approach for classification and regression issues is called Support Vector Machines (SVM). SVM looks for the ideal hyperplane in a high-dimensional space that can best divide the classes. This is done by transforming the input data into a high-dimensional feature space, where it attempts to locate the hyperplane with the highest margin—that is, the distance that the hyperplane has to each class's data points—in order to fulfill its goal. A higher margin indicates that the classifier is performing better at generalization, which is measured by the margin.

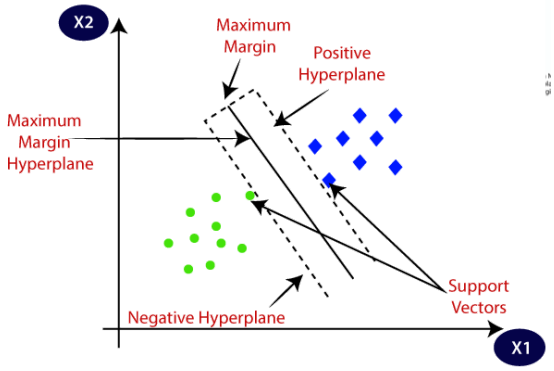


Fig. 10. Support Vector Machine [27]

SVMs perform effectively with data that can be separated into linear and non-linear categories. When dealing with non-linearly separable data, SVM employs a method known as the kernel trick to transfer the data into a higher-dimensional space where it is linearly separable, and then it locates the best hyperplane in that space.

Compared to alternative classifiers like decision trees and logistic regression, SVMs have a number of advantages. For instance, SVMs work well on high-dimensional data with few samples and are less prone to overfitting. SVMs are capable of addressing both binary and multi-class classification issues and are effective at handling noisy data.

6. Ensemble inference

A machine learning technique called ensemble learning uses different models to increase the system's overall accuracy and durability. The core concept is to train various independent models on various subsets of data, then aggregate their predictions to produce a final result. The

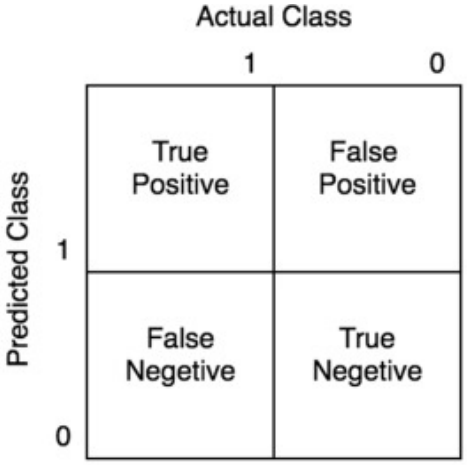
methods of bagging, boosting, and stacking are among those used for ensemble learning.
Multiple models are trained on arbitrary subsets of data in bagging, and the predictions from these models are then averaged or voted upon to provide the final result. Each model is trained using a random subset of the features in this method, which is frequently used for decision trees.

Model Evaluation and performance analysis

A critical stage in the creation of machine learning models, particularly those employed in the detection and classification of breast cancer, is performance evaluation. The below metrics are used to assess how well these models perform:

- Precision
- Recall
- Accuracy
- F1 score

Above metrics are found using the confusion matrix.



A confusion matrix is a machine learning performance evaluation parameter used to assess the precision of a classifier. It is a table that lists the classifier's predictions made in relation to the dataset's actual target values. The number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) predicted by the model is shown in the matrix.

- True Positive (TP): The number of instances when the model accurately identified them as positive.
- False Positive (FP): The proportion of genuine negative cases that the model misclassified as positive.
- False Negative (FN): The proportion of genuine positive cases that the model misclassified as negative.
- True Negative (TN): The number of genuine negative cases that the model properly identified as being negative.

Measurement of precision is the percentage of true positives (positive instances that were correctly identified) among all positive classifications. The ratio of true positives to the total of true positives and false positives is used to compute it. A model with high precision produces fewer false positive mistakes.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

Recall is a performance evaluation metric that assesses a classification model's capacity to locate each successful classification in a dataset. True positives (TP) to the total of true positives and false negatives are measured as a ratio (FN).
Even at the cost of misclassifying some negative examples as positive, recall is a useful metric in situations where it is crucial to identify every occurrence of a specific class (i.e., false positives). Recall might be a more significant number than accuracy, for example, in a medical diagnosis scenario where false negatives can have catastrophic repercussions.

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

$$Recall = \frac{TP}{TP + FN}$$

A classifier's accuracy is a measurement of how effectively it anticipates the solution to a binary classification problem. It is obtained by dividing the total number of samples in the dataset by the number of samples that were correctly categorized. One of the most used criteria for assessing a classification model's performance is accuracy. However, when the data is unbalanced, with one class having much more samples than the other, accuracy might be deceptive. This may result in the dominant class being overfitted and the minority class being underfitted.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Lastly, the F1 score is a precision and recall-balanced index of a classifier's accuracy. Its values range from 0 to 1, and it represents the harmonic mean of the precision and recall. An F1 score of 1 indicates a powerful classifier, whereas a score of nearly 0 indicates a poor classifier.

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Engineering Standards

To assure the project's quality and dependability, specific engineering criteria were adhered to during this research. The pursuing criteria were put into practice:

1. DICOM Standards

The initiative adhered to the Digital Imaging and Communications in Medicine (DICOM) standards, which are the global norm for exchanging medical pictures and related data. This guaranteed that the algorithms employed in the project correctly interpreted and processed the medical images.

2. Data Privacy Standards [28]

The project adhered to the Health Insurance Portability and Accountability Act (HIPAA) rules to guarantee the privacy and confidentiality of patient data. This required putting in place a number of security procedures to guard the information against unwanted access or disclosure.

3. Performance standards

The project was created to adhere to a number of performance standards, including accuracy and processing speed. The algorithms were tuned and put to the test on numerous hardware setups in order to accomplish this.

4. Usability standards

The application's user interface was created using user-centered design concepts to be simple to use and intuitive. Enhanced the user experience, required by conducting usability testing and implementing user feedback.

Research methodology and techniques

In this research, we developed a breast cancer detection and classification model through an experimental investigation employing multiple machine learning methods. We created an application to convert the DICOM-formatted medical images in our dataset to PNG, as well as labeling the mammograms. We trained and tested our models using these mammograms.

Data collection

Our data came from the Azerbaijan Medical University's Cancer Clinic as well as the RSNA dataset, which contained breast tissue images from mammography. The dataset underwent pre-processing to guarantee that the photos were of excellent quality and to get rid of any unnecessary information. In order to use the DICOM

images in our machine learning models, we developed an application to convert the DICOM images to PNG format.
Due to class imbalance, the models overfitted the data, resulting in f1 score of 0. Later, positive images were increased to make sure that our models weren't overfitting the training data. Later, dataset was into training and testing sets and employed cross-validation techniques.

Tools and materials

In doing the research, several tools and resources were used, including

- Python programming language for ML model development as well as data analysis
- C# programming language for developing the application for converting mammograms from DICOM file format to PNG, as well as allowing the user to format those pictures, add comments and show prediction for each image, working integrated with the ML model built
- Scikit-learn library, Pandas, numpy libraries for implementing machine learning models

Reasons for methodology choices

We determined our methodology based on the nature of the issue we were attempting to resolve. It takes sophisticated machine learning techniques to accurately identify and categorize breast cancer. To ensure that we looked at a range of possible ways to solve the problem, we selected a variety of algorithms. In order to guarantee the reliability and robustness of our models, we also applied cross-validation methods. Lastly, to make sure we could work with our dataset in our machine learning models without difficulty, we created our own application for converting DICOM images to PNG format, and integrating the model to predict the chance of cancer.

Architecture and Model

The project's architecture includes an application that lets doctors upload DICOM images and use machine-learning algorithms to diagnose breast cancer. The program is made up of various components that work together to give doctors an effective and convenient experience. Figure 7 depicts the proposed architecture of the system.

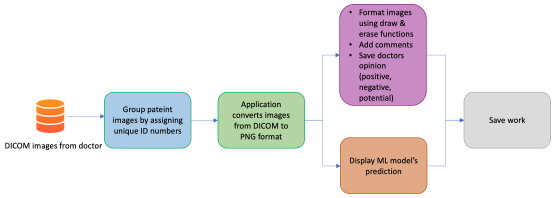


Fig. 11. Architecture of the system

The application gives each patient a unique ID when the doctor uploads the DICOM images, and it organizes the images into folders based on this ID. After that, a DICOM to PNG converter module is used to process the images. The PNG format, which is more frequently used and is simpler for machine learning algorithms to comprehend, is used by this module to convert DICOM images.
The doctor can inspect and format the converted images in the program after they have been converted. The doctor can also decide which forecast to use and enter it into the program depending on their analysis. The doctor's prediction is displayed alongside the machine learning model's prediction, which is also made based on the input photos. This enables the doctor to compare their analysis with the prediction made by the model and come to a well-informed conclusion.
Classifiers including XGBoost, Random Forest, KNN, Logistic Regression, and SVM, are the machine learning models employed in this research. Each classifier learns from a set of breast cancer images with associated labels and then uses a voting-based method to combine predictions to provide the final prediction.
Ultimately, the architecture and model utilized in this project are intended to give

medical professionals a trustworthy, precise, user-friendly tool for diagnosing breast cancer.

Social and environmental impact

There are numerous social and environmental effects of breast cancer detection studies. First off, early detection of breast cancer can raise a woman's quality of life by increasing her odds of survival and decreasing the need for more invasive and expensive therapies. Also, early detection lessens the psychological and physical toll on women and their families.
Also, by fewer cases of advanced breast cancer that need more resources to treat, the research may lessen the strain on healthcare systems. Cost reductions and more effective use of healthcare resources could result from this.
Early detection can help lessen the quantity of medical waste created by therapies for advanced cases of breast cancer, which is beneficial for the environment. This is so that unnecessary invasive and expensive therapies can be avoided with the help of early detection.
Also, the creation of a breast cancer screening app could make healthcare services more readily available to women in rural or underserved locations, hence lowering access disparities. Decreasing the burden of sickness and enhancing population health can positively affect the social and economic growth of these regions.
The social and environmental effects of this research could ultimately result in more sustainable healthcare systems, lower healthcare expenditures, and better health outcomes.

IV. Implementation

Hardware Design

- Processor: Intel core i5 or higher
- RAM: 16 GB or more
- Hard Disk: 500 GB or more, preferring SSD
- Graphics card (optional): Nvidia GTX 1060 or higher

Software Design

- Operating system: Windows 10, Ubuntu 18.04 or later
- C#, Python 3.7 or later
- Anaconda distribution of Python (recommended)
- ML libraries: sci-kit-learn, XGBoost, Numpy, Pandas
- Jupyter notebook or similar software

Essential Components of Project

Many different strategies have been suggested in the literature for breast cancer diagnosis, which is a significant field of research in medical image analysis. In this area, supervised learning techniques in particular have been demonstrated to be quite promising. Unfortunately, despite the high incidence of breast cancer in Azerbaijan, there has never been a version of this program.
Our study concentrated on the use of mammography for the early diagnosis of breast cancer. Mammograms are frequently used for diagnosis and screening, and they are a valuable source of data for radiologists. The mammograms were preprocessed by transferring them from DICOM to PNG format, and we took the positive, negative, and prospective results for each mammography into account.
To train and assess our models, we used supervised learning techniques. To train our models, we chose a variety of variables that have previously been demonstrated to be important for diagnosing breast cancer. Following training, we assessed our models' performance using a range of metrics, including accuracy, precision, recall, and

F1 score. Figure 12 depicts the process of the ML part in more detail.
In medical applications, it can be difficult to achieve high accuracy, and this was true for our study as well. However, our method produced encouraging results, and we think that our application has the potential to be an effective tool for the diagnosis of breast cancer in Azerbaijan.

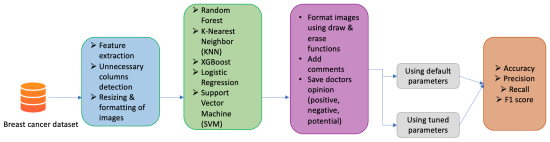


Fig. 12. ML architecture

Labeling application

Being a significant step towards improving the quality of breast cancer detection in Azerbaijan, the application built during this project is a novel tool for labeling medical images for breast cancer detection.
This solution fills the gap left by the absence of similar programs in the country and offers a simple user interface to help medical practitioners with labeling.
The application consists of 4 Forms. A form is a window or dialog box that provides a user interface for your application. A form contains controls, such as buttons, labels, and text boxes, that enable the user to interact with your application. You can use forms to display data, receive input from the user, and perform various other tasks.
The second form contains important functions of the application.
read_Directories() function reads the path to the folder where the dicoms are located in. Later, this folder will be used for converting pictures from DICOM to PNG format.

```
1 reference
private void read_Directories()
{
    int countL = 0;
    StreamReader sr = new StreamReader("path_to_dcm.txt");
    String line = sr.ReadLine();
    while (line != null)
    {
        Console.WriteLine(line);
        if (countL == 0)
            dicomPath = line;
        else
            pythonPath = line;
        line = sr.ReadLine();
        countL++;
    }
    sr.Close();
}
```

dcm_to_png() function converts medical images which are in DICOM format to png format. As one patient has more than one image, this function also creates separate folders for each patient and assigns unique ID numbers to each patient. Later, colored and original images of patients are stored in their respective folders.
To create different colored interpretations of images, the dcmread function from the PyDICOM package was used to read the DICOM image. The image's pixel information is then extracted and scaled with NumPy to a range of 0-255. The scaled pixel data is then converted to an unsigned 8-bit integer using NumPy's uint8 function.
As it is not possible to open DICOM images on regular computers, they have been converted to PNG. The pixel array is first extracted from the DICOM file once it has been read using the pydicom package. The grayscale image is then represented by scaling the pixel array to a numpy array of 8-bit unsigned integers. After that, a PNG file is created and the numpy array is transformed into a PIL image.

```
1 reference
private void dcm_to_png()
{
    string directory = Directory.GetCurrentDirectory();

    var psi = new ProcessStartInfo();
    psi.FileName = pythonPath;
    var script = directory + "\\Dicom_png_original_colored.py";
    var fname = dicomPath;
    psi.Arguments = $"\"{script}\"  \"{fname}\"";

    psi.UseShellExecute = false;
    psi.CreateNoWindow = true;
    psi.RedirectStandardOutput = true;
    psi.RedirectStandardError = true;

    var errors = "";
    var results = "";

    using (var process = Process.Start(psi))
    {
        errors = process.StandardError.ReadToEnd();
        results = process.StandardOutput.ReadToEnd();
    }


}
```

LoadData() function contains a function named loadjson() checks whether the JSON file is empty or not and changes it to the list. The program, a JSON file is used for storing the information marked by the doctor for each patient. frontend view of the button, panel, picture box, and textbox are assigned in this function. It loads the image of the picture box from where it was left from.

```
1 reference
private void LoadData()
{
    LoadJson();
    c = dataList.Count();
    int icount = dataList.Count;
    ii = dataList.Count;

    PictureBoxClicked[0] = false;
    PictureBoxClicked[1] = false;
    PictureBoxClicked[2] = false;
    PictureBoxClicked[3] = false;

    this.WindowState = FormWindowState.Maximized;
    this.KeyPreview = true;
    this.BackColor = Color.FromArgb(34, 34, 34);

    panel1.BackColor = Color.FromArgb(45, 45, 45);
    panel2.BackColor = Color.FromArgb(55, 55, 55);
    panel4.BackColor = Color.FromArgb(45, 45, 45);

    radioButtonNegative.BackColor = Color.FromArgb(55, 55, 55);
    radioButtonNegative.ForeColor = Color.White;
    radioButtonPositive.BackColor = Color.FromArgb(55, 55, 55);
    radioButtonPositive.ForeColor = Color.White;
    radioButtonPotential.BackColor = Color.FromArgb(55, 55, 55);
    radioButtonPotential.ForeColor = Color.White;

    textBoxComment.Text = textCom;
    textBoxComment.ForeColor = Color.White;
    textBoxComment.BackColor = Color.FromArgb(55, 55, 55);
```

CheckIfWeHave() function checks if there already exists a rectangle drawing inside the JSON file, if yes, it takes the position of the image and shows which out of all images the current one is.

```
private void CheckIfWeHave()
{
    ch = dataList.Count() - 1;
    int jd = dataList.Count;
    if (jd > ii)
    {
        Update(ii);
        var item = dataList[ch];
        string[] dirs = Directory.GetDirectories(pathOri);
        string[] dirsC = Directory.GetDirectories(pathCol);
        string dirName = new DirectoryInfo(dirs[ii]).Name;
        string testdir = dirs[ii] + "\\" + item.ImageName;
        string testdirC = dirsC[ii] + "\\" + item.ImageName;
        if (File.Exists(testdir) || File.Exists(testdir))//dirs[ii] + "\\" + item.ImageName.ToString()))
        {
            if (textBoxComment.Equals(textCom))
                this.textBoxComment.Text = item.Comment;

            if (item.Diagnosis == 1)
            {
                this.radioButtonPositive.Checked = true;
                d = 1;
            }
            else if (item.Diagnosis == 2)
            {
                this.radioButtonPotential.Checked = true;
                d = 2;
            }
            else if (item.Diagnosis == 3)
            {
                this.radioButtonNegative.Checked = true;
                d = 3;
            }
```

Although the buttonPencilClick() function is built in, there need to be some changes like adding penClick which checks if the pen is clicked or not.

```
private void buttonPencil_Click(object sender, EventArgs e)
{
    picBoxDraw = true;

    if (!penClick)
    {
        var bitmap = (Bitmap)Properties.Resources.minipencil;
        this.Cursor = CreateCursor(bitmap, new Size(bitmap.Width / 15, bitmap.Height / 15));
        penClick = true;
    }
    else
    {
        this.Cursor = Cursors.Default;
        penClick = false;
    }
}
```

When the pictureBox3_MouseDown() function is called, it takes the initial coordinates of the x and y. pictureBox3_MouseMove() function keeps refreshing, according to the temporary final location, until you stop pressing the mouse. pictureBox3_MouseUp() shows the location whenever the mouse is released.

```
private void pictureBox3_MouseDown(object sender, MouseEventArgs e)
{
    PictureBoxClicked[2] = true;
    if (penClick)
    {
        IsMouseDown = true;
        Location3XY = e.Location;
        tempData.Rect3X1 = e.Location.X;
        tempData.Rect3Y1 = e.Location.Y;
    }
}
1 reference
private void pictureBox3_MouseMove(object sender, MouseEventArgs e)
{
    if (IsMouseDown == true)
    {
        Location3X1Y1 = e.Location;
        tempData.Rect3X2 = e.Location.X;
        tempData.Rect3Y2 = e.Location.Y;
        Refresh();
    }
}
1 reference
private void pictureBox3_MouseUp(object sender, MouseEventArgs e)
{
    if (IsMouseDown == true)
    {
        Location3X1Y1 = e.Location;
        tempData.Rect3X2 = e.Location.X;
        tempData.Rect3Y2 = e.Location.Y;
        IsMouseDown = false;
    }
}
```

pictureBox1_DoubleClick() function takes the picture number and adds the picture to the data list and it first checks if the data is new or not, when the rectangle is drawn, it is added to the data list.

```
private void pictureBox1_DoubleClick(object sender, EventArgs e)
{
    picNum = 0;
    AddToDataList();
    CreateStaticData(ii);
    Form4 form4 = new Form4();
    form4.Show();
}
```

CreateStaticData() function sends data from one form to another one.

```
private void CreateStaticData(int i)
{
    j = i;
    var item1 = dataList[i];
    StaticData.DataList1.Image1Id = item1.Image1Id;
    StaticData.DataList1.DoctorId = item1.DoctorId;
    StaticData.DataList1.Diagnosis = item1.Diagnosis;
    StaticData.DataList1.Comment = item1.Comment;
    StaticData.DataList1.Rect1X1 = item1.Rect1X1;
    StaticData.DataList1.Rect1Y1 = item1.Rect1Y1;
    StaticData.DataList1.Rect1X2 = item1.Rect1X2;
    StaticData.DataList1.Rect1Y2 = item1.Rect1Y2;

    StaticData.DataList1.Image2Id = item1.Image2Id;
    StaticData.DataList1.Rect2X1 = item1.Rect2X1;
    StaticData.DataList1.Rect2Y1 = item1.Rect2Y1;
    StaticData.DataList1.Rect2X2 = item1.Rect2X2;
    StaticData.DataList1.Rect2Y2 = item1.Rect2Y2;

    StaticData.DataList1.Image3Id = item1.Image3Id;
    StaticData.DataList1.Rect3X1 = item1.Rect3X1;
    StaticData.DataList1.Rect3Y1 = item1.Rect3Y1;
    StaticData.DataList1.Rect3X2 = item1.Rect3X2;
    StaticData.DataList1.Rect3Y2 = item1.Rect3Y2;

    StaticData.DataList1.Image4Id = item1.Image4Id;
    StaticData.DataList1.Rect4X1 = item1.Rect4X1;
    StaticData.DataList1.Rect4Y1 = item1.Rect4Y1;
    StaticData.DataList1.Rect4X2 = item1.Rect4X2;
    StaticData.DataList1.Rect4Y2 = item1.Rect4Y2;
}
```

buttonClean_Click() function makes all the locations zero and refreshes the picture boxes.

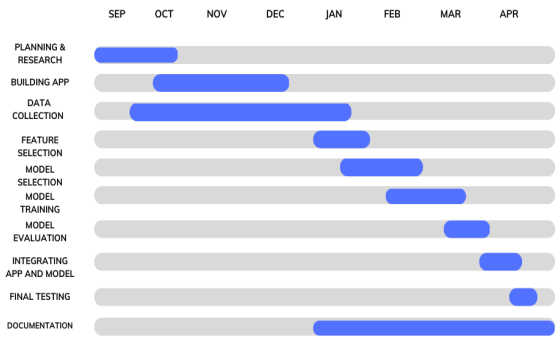
```
private void buttonClean_Click(object sender, EventArgs e)
{
    Location1XY.X = 0;
    Location1XY.Y = 0;
    Location1X1Y1.X = 0;
    Location1X1Y1.Y = 0;

    Location2XY.X = 0;
    Location2XY.Y = 0;
    Location2X1Y1.X = 0;
    Location2X1Y1.Y = 0;

    Location3XY.X = 0;
    Location3XY.Y = 0;
    Location3X1Y1.X = 0;
    Location3X1Y1.Y = 0;

    Location4XY.X = 0;
    Location4XY.Y = 0;
    Location4X1Y1.X = 0;
    Location4X1Y1.Y = 0;
    Refresh();
    if (dataList.Count > ii)
    {
        var item = dataList[ii];
        item.Rect1X1 = Location1XY.X * 11;
        item.Rect1Y1 = Location1XY.Y * 11;
        item.Rect1X2 = Location1X1Y1.X * 11;
        item.Rect1Y2 = Location1X1Y1.Y * 11;

        item.Rect2X1 = Location2XY.X * 11;
        item.Rect2Y1 = Location2XY.Y * 11;
        item.Rect2X2 = Location2X1Y1.X * 11;
        item.Rect2Y2 = Location2X1Y1.Y * 11;

        item.Rect3X1 = Location3XY.X * 11;
        item.Rect3Y1 = Location3XY.Y * 11;
        item.Rect3X2 = Location3X1Y1.X * 11;
        item.Rect3Y2 = Location3X1Y1.Y * 11;

        item.Rect4X1 = Location4XY.X * 11;
        item.Rect4Y1 = Location4XY.Y * 11;
        item.Rect4X2 = Location4X1Y1.X * 11;
        item.Rect4Y2 = Location4X1Y1.Y * 11;
    }
}
```

To keep them organized, the application converts DICOM images to PNG format and gives each patient's images a different ID. Additionally, it has built-in features that let the user label photographs with squares and choose between three options: negative, positive, or potential. These features support the categorization and labeling of medical pictures, which is essential for the creation of precise and trustworthy machine-learning models for the detection of breast cancer.

The directory of the original DICOM images on the doctor's PC is specified in the application's configuration file. The program loads those DICOM images from that directory at startup, changes them to PNG format, saves them, and then deletes the DICOMs. As the user hits the "back" or "next" button, the application determines whether the JSON file is empty or not, and shows the labeled or original photos as appropriate. The up and down arrows are used to navigate between different image formats, while the pen and

eraser buttons are used to draw and remove squares from the screen. Based on the ML models created, the "Predict" button offers a prediction for each image. Moreover, the doctor can add comments or notes in the textbox. Any changes made are automatically saved when the "back" or "next" buttons are clicked.

Gantt Chart

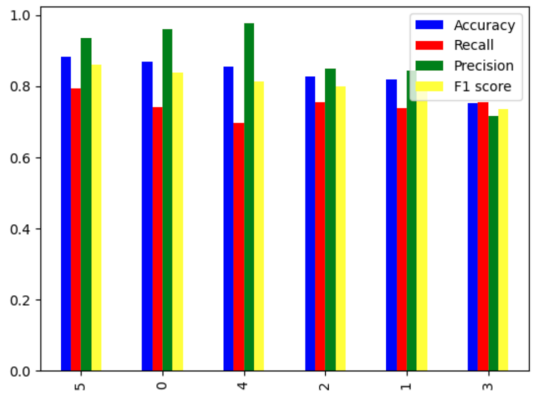


Testing/Verification/Validation of Results

The software we've created is intended to make it easier to detect breast cancer from mammograms. The program accepts input in the form of DICOM images, which are then collected and saved in a folder with a specific ID assigned to the patient. This makes it easier to access and arrange all of the images for a specific patient.
As the user uploads the photographs, the application gives them the option to be formatted for better visibility and clarity. The integrated machine learning model is then given structured photos to forecast the likelihood of breast cancer.
The application is easy-to-navigate and user-friendly.

For the machine learning part of the project, the XGBoost model's accuracy score of 88.22% was the highest. This indicates that 88.22% of the model's total predictions were accurate. The accuracy of the Random Forest Classifier, which came in second, was 86.98%. The SVM model had the highest accuracy (85.41%), followed by KNeighbour Classifier (82.83%) and Logistic Regression with relatively low accuracy (81.93%).
The XGBoost model once again achieved the top score when it came to the recall statistic, which measures the percentage of true positives that were properly detected by the model. KNeighbour Classifier came in second with a score of 75.50%, followed by Random Forest Classifier with a score of 74.26%, and Logistic Regression with a score of 73.76%.
The SVM model outperformed all other models in terms of precision, which is the percentage of true positives among all positive predictions generated by the model (97.57%). With a precision score of 96.15%, the Random Forest Classifier placed second, ahead of XGBoost (93.59%), Logistic Regression (84.42%), and KNeighbour Classifier (84.96%).
Lastly, XGBoost once more outperformed the competition with a score of 85.94% when considering the F1 score, which represents the harmonic mean of precision and recall. With an F1 score of 83.80%, the Random Forest Classifier placed second, ahead of the K-Neighbour Classifier (79.95%), and the Logistic Regression (78.73%).
Overall, the Random Forest Classifier came in second place to the XGBoost model in all criteria. As opposed to other sorts of data, medical cases might sometimes have relatively low accuracy, thus the findings of this study are nonetheless highly encouraging.

	Model	Accuracy	Recall	Precision	F1 score
5	XGBoost	0.882155	0.794554	0.935860	0.859438
0	Random Forest Classifier	0.869809	0.742574	0.961538	0.837989
4	SVC	0.854097	0.695545	0.975694	0.812139
2	KNeighbour Clasifier	0.828283	0.754950	0.849582	0.799476
1	Logistic Regression	0.819304	0.737624	0.844193	0.787318



Machine learning (ML) applications for medical image analysis frequently use ensemble learning. To make better forecasts than a single model, it aggregates the results of several models. Ensemble learning can help increase the generalization of the model and lower the risk of overfitting, which are two of its key advantages. The accuracy of tasks involving the classification, segmentation, and detection of medical images has been significantly increased by the application of ensemble approaches.

By mixing the outputs of various models, each of which may be able to capture a distinct element of the data, ensemble learning can produce more reliable findings in the specific situation of breast cancer detection. This is crucial since there can be variances in medical images as a result of various imaging modalities, acquisition settings, and patient-related factors. The effects of data imbalance, which are frequent in datasets of medical picture data, can also be lessened with the aid of ensemble learning.

A well-liked ensemble technique that combines the predictions of various models by majority vote is the voting classifier. This means that in the case of breast cancer detection, the final prediction is based on the XGBoost, RandomForest, and Lightgbm models' combined vote. Through the confidence of each model's vote, this method can serve to increase the precision of the final prediction while simultaneously supplying a degree of uncertainty.

Overall, ensemble learning is a useful method for enhancing the reliability and accuracy of medical imaging machine-learning models, which can ultimately improve patient outcomes by enabling earlier and more precise breast cancer identification.

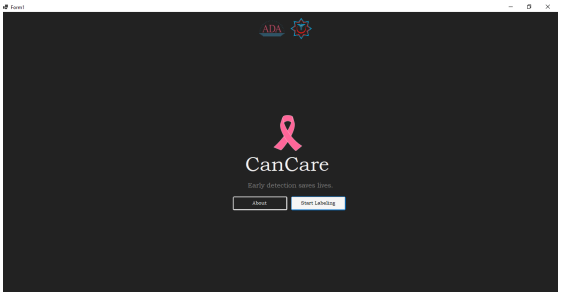
The ensemble learning model's accuracy score of 89.67% was the highest of all models. The F1 score of the model was the highest again being 87.53%. The ensemble learning model had the second highest precision which was 96.71%. The model had a score of 79.95% was again the highest.

V. Conclusion

Discussion of Results

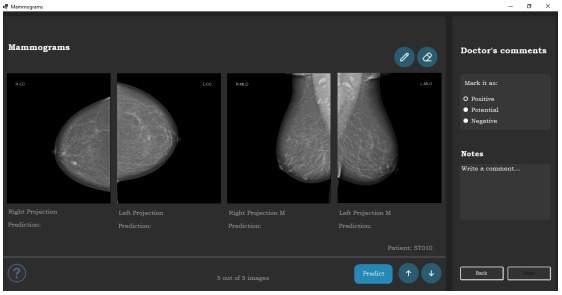
The study investigated the application of supervised learning algorithms for mammogram-based breast cancer screening. Six well-known algorithms' performances were assessed and contrasted. Of them, XGBoost displayed the best F1 score and accuracy. The created application enables the upload of DICOM images, the grouping of patient images into a folder with a distinct ID, formatting, and cancer status prediction based on the integrated model. The application may help with early breast cancer detection and diagnosis, improving healthcare outcomes.

Below is the home page of the application:



Upon navigating to label the images, the user can find the below interface. Three radio buttons help mark the doctor’s prediction. The built-in model predicts the

accuracy when the “Predict” button is clicked.



Future work

To sum up, the creation of a reliable and effective tool for the detection and prognosis of breast cancer has the potential to significantly enhance patient outcomes and lower medical expenses. For the development of healthcare technology, further research and development in this area are essential.

1. Expanding the analysis to include other imaging modalities, such as ultrasounds or MRIs, would be beneficial given that this study concentrated on mammograms.
2. Clinical data integration: Integrating imaging information with clinical information, such as patient demographics and medical history, may increase the precision of cancer diagnosis.
3. Application deployment in a clinical environment: Further testing and validation of the application in a clinical environment would be required to guarantee its safety and efficacy.
4. The creation of a mobile application version: A mobile application version would improve accessibility and user-friendliness for healthcare practitioners.
5. Examining deep learning methods: Using deep learning methods, such as convolutional neural networks, may help to increase the precision of cancer detection and prediction.

Acknowledgment

We want to sincerely thank Professor Jamaladdin Hasanov for his guidance and assistance during Senior Design Project (SDP) I and II.
We also appreciate the provision of the chance to conduct our research at the Center for Data Analytics Research (CeDAR) at ADA University.

References

[1] World Health Organization. (2021). Breast Cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Umer, M., Naveed, M., Alrowais, F., Ishaq, A., Al Hejaili, A., Alsubai, S., Eshmawi, A. A., Mohamed, A., & Ashraf, I. (2021). Breast Cancer Detection Using Convoluted Features and Ensemble Machine Learning Algorithm. Applied Sciences, 11(10), 4719. https://doi.org/10.3390/app11104719

[3] Cancer Treatment Centers of America. (n.d.). Types of Breast Cancer. Retrieved March 23, 2023, from https://www.cancercenter.com/cancer-types/breast-cancer/types

[4] Stanford Children's Health. (n.d.). Anatomy of the breasts. Retrieved March 23, 2023, from https://www.stanfordchildrens.org/en/topic/default?id=anatomy-of-the-breasts-85-P00132

[5] American Cancer Society. (2021). Inflammatory Breast Cancer. Retrieved September 28, 2021, from https://www.cancer.org/cancer/breast-cancer/about/types-of-breast-cancer/inflammatory-breast-cancer.html

[6] Cheng, H. D., Shan, J., Ju, W., Guo, Y., & Zhang, L. (2010). Automated breast cancer detection and classification using ultrasound images: A survey. Department of Computer Science, Utah State University, Logan, UT 84322, USA School of Mathematics and System Science, Shandong University, China.

[7] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast cancer classification using machine learning. In Proceedings of the 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, 18–19 April 2018; pp. 1–4.

[8] Nawaz, M., Sewissy, A. A., & Soliman, T. H. A. (2018). Multi-class breast cancer classification using deep learning convolutional neural network. International Journal of Advanced Computer Science and Applications, 9(11), 316-332.

[9] Murphy, A. (2021). Breast Cancer Wisconsin (Diagnostic) Data Analysis Using GFS-TSK. In North American Fuzzy Information Processing Society Annual Conference (pp. 302-308). Springer.

[10] Ghosh, P. (n.d.). Breast Cancer Wisconsin (Diagnostic) Prediction. Retrieved October 1, 2022, from https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

[11] Akbulut, S., Cicek, I. B., & Colak, C. (2022). Classification of Breast Cancer on the Strength of Potential Risk Factors with Boosting Models: A Public Health Informatics Application. Med Bull. Haseki/Haseki Tip Bul., 60, 196-203. doi: 10.4274/haseki.galenos.2021.0307.

[12] Rajinikanth, V., Kadry, S., Taniar, D., Damaševiˇcius, R., & Rauf, H. T. (2021). Breast-cancer detection using thermal images with marine-predators-algorithm selected features. In Proceedings of the 2021 Seventh International Conference on Bio Signals, Images, and Instrumentation (ICBSII), Chennai, India, 25–27 March 2021; pp. 1–6.

[13] Khan, S., Islam, N., Jan, Z., Din, I. U., & Rodrigues, J. J. C. (2019). A novel deep learning-based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters, 125, 1-6. doi: 10.1016/j.patrec.2019.06.004

[14] Hamed, G., Marey, M. A. E. R., Amin, S. E. S., & Tolba, M. F. (2020). Deep learning in breast cancer detection and classification. In The International Conference on Artificial Intelligence and Computer Vision (pp. 322-333). Springer.

[15] Ak, M.F. (2020). A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. Healthcare, 8(3), 111. https://doi.org/10.3390/healthcare8030111

[16] Dey, N., Rajinikanth, V., & Hassanien, A. E. (2021). An examination system to classify the breast thermal images into early/acute DCIS class. In International Conference on Data Science and Applications (pp. 209-220). Springer.

[17] Kashif, M., Malik, K. R., Jabbar, S., & Chaudhry, J. (2020). Application of machine learning and image processing for detection of breast cancer. In Innovation in Health Informatics (pp. 145-162). Elsevier.

[18] Cabıoğlu, Ç. & Oğul, H. (2020). Computer-aided breast cancer diagnosis from thermal images using transfer learning. In Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO), Granada, Spain, 6–8 May 2020 (pp. 716–726). Springer.

[19] McKinney, S.M., Sieniek, M., Godbole, V. et al. International evaluation of an AI system for breast cancer screening. Nature 577, 89–94 (2020). https://doi.org/10.1038/s41586-019-1799-6

[20] Ting, F.F.; Tan, Y.J.; Sim, K.S. (2019). Convolutional neural network improvement for breast cancer classification. Expert Syst. Appl. 120, 103–115.

[21] Sakib, S., Yasmin, N., Tanzeem, A. K., Shorna, F., Hasib, K. M., & Alam, S. B. (2022). Breast Cancer Detection and Classification: A Comparative Analysis Using Machine Learning Algorithms. In Computational Intelligence in Data Mining (pp. 659-670). Springer, Singapore.

[22] Neural Network: https://www.analyticsvidhya.com/blog/2016/08/evolution-core-concepts-deep-learning-neural-networks/

[23] Decision Tree: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[24] Random Forest: https://www.tibco.com/reference-center/what-is-a-random-forest

[25] K-Nearest Neighbor: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[26] Logistic Regression https://www.javatpoint.com/logistic-regression-in-machine-learning

[27] Support Vector Machine https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[28] U.S. Department of Health & Human Services. (n.d.). HIPAA Privacy Rule. Retrieved March 22, 2023, from https://www.hhs.gov/hipaa/for-professionals/privacy/index.html#:~:text=The%20HIPAA%20Privacy%20Rule&text=The%20Rule%20requires%20appropriate%20safeguards,information%20without%20an%20individual's%20authorization.

[29] Schnabel, J. A., Chakraborty, D. P., & Castellano-Smith, A. D. (2020). Automated breast ultrasound segmentation in whole‐breast images using deep convolutional neural networks. Journal of Microscopy, 277(3), 189-201. https://doi.org/10.1111/jmi.12869

[30] Kumar, S., & Sharma, A. (2017). A review on breast cancer diagnosis and treatment using intelligent techniques. Neural Computing and Applications, 28(7), 1651-1664. https://doi.org/10.1007/s00521-016-2338-2

[31] Sultan, A. A., Jerjes, W., & Upile, T. (2019). Current Available Computer-Aided Detection Catches Cancer but Requires a Human Operator!. Cureus, 11(3), e4438. doi: 10.7759/cureus.4438

[32] Koren G, King S, Knowles S, Phillips E. Adverse drug reactions to nonsteroidal anti-inflammatory drugs: Clinical and methodological issues. J Clin Pharmacol. 1993;33(4):296-315. doi: 10.1002/j.1552-4604.1993.tb03930.x. PMID: 8492349; PMCID: PMC61235.

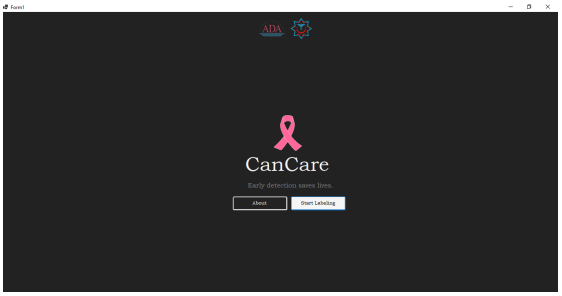
[33] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

Appendix 1

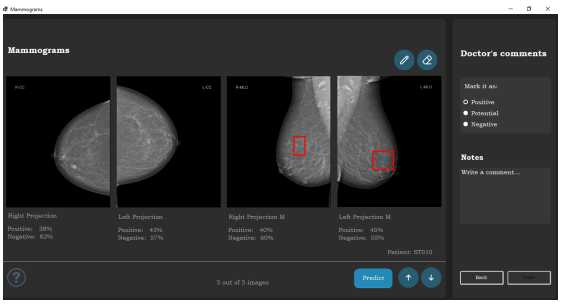
Program codes

Link to GitHub repository

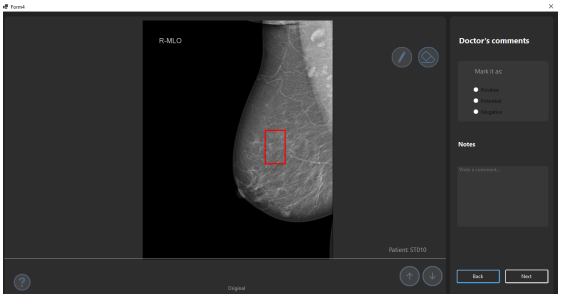
Screenshots of the software interface



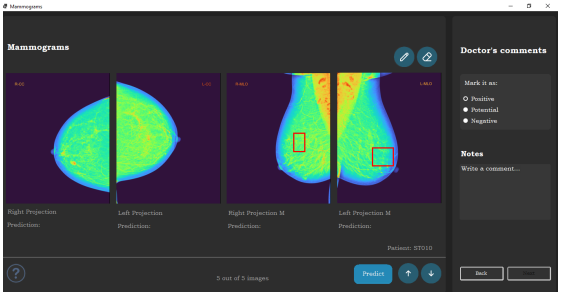
Home page



Labeling interface



Zoomed interface



Colored images

In-depth description of technologies used in the project

Scikit learn

A free, open-source machine learning package for the Python programming language called Scikit-learn, is also known as sklearn. It offers a wide range of machine learning algorithms for applications like classification, regression, clustering, and dimensionality reduction in addition to easy and effective tools for data mining and analysis.
In order to provide a comprehensive data analysis and machine learning pipeline, Scikit-learn is built on top of other scientific computing tools like NumPy, SciPy, and Matplotlib. It provides a full range of tools for selecting and analyzing models, feature extraction from data, and data preprocessing.