Syrian Arab Republic

Higher Institute for Applied Sciences and Technology

Informatic Department

4th year, SEMINAR

Detecting breast cancer using AI

Written by: Ahmad Ali

Scientific supervisor: Dr. Yasser Rahal

Linguistic supervisor: Dr. Nada Mohanna

Seminar supervisor: Dr. Adel Alkefri

Contents

Abstraction

Breast Cancer is one of the diseases that causes a higher number of deaths in a year. Amongst woman, Breast Cancer is the second highest disease that causes death, and in Canada, it is a leading cause of death. Early detection of breast cancer makes it most curable cancer in among other types of cancer, early detection and accurate examination for breast cancer ensures an extends survival rate of the patients.

Breast cancer remains a leading cause of mortality among women, with early detection significantly improving survival rates. This research focuses on leveraging artificial intelligence (AI) to enhance the accuracy of breast cancer diagnosis. Various machine learning techniques, including Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs), are employed to classify benign and malignant breast cancer cells. The study utilizes multiple datasets, such as the Wisconsin breast cancer dataset, to train and evaluate different models. Key methodologies include the annotation of training and test sets, computerized quantification of tumor-infiltrating lymphocytes, and texture feature extraction using local binary patterns. The results demonstrate that AI models, particularly CNNs combined with SVMs, can effectively improve diagnostic accuracy, thereby supporting early detection and treatment of breast cancer. This research underscores the potential of AI in medical diagnostics, aiming to reduce human error and enhance diagnostic precision.

1.Introduction

Breast cancer is a significant health concern, accounting for one in six cancer-related deaths among women. Despite continuous advancements in medical research, a definitive method for preventing breast cancer remains elusive. Early detection is the most effective strategy to reduce the mortality rate associated with this disease. Identifying breast cancer at an early stage, before it metastasizes, is crucial for improving patient outcomes. Unfortunately, about 90% of breast cancer cases are diagnosed at advanced stages (III and IV), posing severe challenges for treatment.

Accurate and reliable diagnostic methods are essential to detect breast cancer early and reduce mortality rates. Traditional diagnostic approaches, even when conducted by experienced medical professionals, can be challenging and are often complicated by the necessity of interpreting multiple test results. Over the past few decades, advancements in diagnostic systems, particularly the integration of Artificial Intelligence (AI) techniques, have significantly improved breast cancer detection. AI technologies, including artificial neural networks and Computer-Aided Diagnosis (CAD) systems, assist radiologists in identifying breast cancer, thereby minimizing human error and enhancing diagnostic accuracy.

Various classification techniques are employed to predict and categorize breast cancer patterns. Implementing these classification systems reduces errors caused by fatigue or inexperience and allows for more thorough examination of medical data. This study explores the use of different artificial neural networks, such as Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and Deep Neural Networks (DNNs), to classify benign and malignant breast cancer cells. By comparing the diagnostic results of these models, we aim to identify the most effective network for early breast cancer detection. The primary objective is to employ these AI models for accurate classification of breast cancer cells, thereby facilitating early detection and improving patient prognosis.

2.Machine Learning Technique

In the field of machine learning, two primary types of learning exist: supervised learning and unsupervised learning. Supervised learning involves training the machine using a labeled set of data cases to obtain accurate results. On the other hand, unsupervised learning deals with unlabeled datasets where the expected outcomes are unknown, making the task more challenging. For this paper, we focused on supervised learning, specifically employing classification, which is a widely used method in this domain. Supervised learning utilizes labeled data to develop a model that can be used for future predictions. In the healthcare industry, machine learning models play a vital role, as hospitals maintain extensive databases. These datasets are often utilized in research to develop classification models that can provide insights based on the data. By implementing classification models on patient data, systems can be developed for future predictions. This paper discusses various classification models, including SVM (Support Vector Machine), which is a supervised classification model widely applied in the field of cancer diagnosis[1]. SVM (Support Vector Machine) is a powerful algorithm that selects crucial samples, known as support vectors, from all groups or classes. By constructing a linear function, SVM effectively separates the classes. In some cases, SVM can be applied to map the input vector to a higher-dimensional space, aiming to find the most suitable hyperplane that effectively separates the dataset into different classes. The goal of SVM is to identify the hyperplane that maximizes the margin between classes, thereby achieving optimal separation[2]. SVM (Support Vector Machine) is an influential algorithm used in machine learning. It selects important samples called support vectors from different groups or classes. By creating a linear function, SVM effectively separates these classes. In certain scenarios, SVM can also be employed to map the input vector to a higher-dimensional space. This approach helps in finding the most appropriate hyperplane that effectively separates the dataset into distinct classes. The primary objective of SVM is to identify the hyperplane that maximizes the margin between the classes, thereby achieving optimal separation. This capability makes SVM a powerful tool for classification tasks in machine learning. [3].

Logistic regression (LR) is an extension of regression analysis methods that is specifically suited for situations where the dependent variable is categorical. It allows for the examination of conditions where specific outcomes are common. For

instance, when evaluating the effectiveness of an educational program, LR can be used to predict dichotomous outcomes such as success/failure or advanced/not advanced. Similarly, in a medical context, LR can be utilized to determine the presence or absence of a disease.

On the other hand, an Artificial Neural Network (ANN) is a computational model inspired by the functioning of biological neural systems, such as the human brain. It is designed to process data and learn patterns or relationships within the data. ANN employs interconnected nodes, known as neurons, which simulate the behavior of biological neurons. Through a process of learning and adjustment of connection weights, ANNs can recognize complex patterns and make predictions based on input data. [4]. The distinguishing characteristic of Artificial Neural Networks (ANNs) lies in their unique data processing structure. They consist of a vast number of interconnected processing units, or neurons, which work together harmoniously to solve specific problems. Similar to humans, ANNs learn through examples. They are trained for specific purposes, such as pattern recognition or data analysis, using a learning algorithm. Learning in biological systems involves adjusting the synaptic connections between neurons. ANNs, with their exceptional ability to extract information from complex or ambiguous data, can uncover patterns and identify trends that may be too subtle to be observed by humans or other computer algorithms. A well-trained neural network can be viewed as an "expert" in the domain of knowledge it has been trained on.

Another classification model is Gaussian Naive Bayes. In addition to Gaussian Naive Bayes, there are two other variations: Multinomial Naive Bayes and Bernoulli Naive Bayes. Each of these models applies the Naive Bayes algorithm with different assumptions about the underlying distribution of the data. Gaussian Naive Bayes assumes that the features follow a Gaussian (normal) distribution, Multinomial Naive Bayes is suitable for discrete features following a multinomial distribution, and Bernoulli Naive Bayes is used when dealing with binary features following a Bernoulli distribution. These variations allow for the application of Naive Bayes in different types of data analysis tasks, depending on the nature of the features and their distributions. [5].

When comparing the performance of classification models on the Breast Cancer dataset, it is important to consider appropriate evaluation measures that can accurately reflect the predictive performance of the models. One commonly used measure is accuracy, which is calculated by dividing the number of correctly predicted samples by the total number of samples in the dataset.

Another widely used evaluation measure is the Confusion Matrix, as shown in Table 1. The Confusion Matrix aids in evaluating the classification predictions. In the prediction process, each sample can be classified into two categories: Cancer and No-Cancer. The Confusion Matrix consists of four elements: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

- True Positive (TP) represents the samples that are correctly predicted as Cancer and are consistent with their actual class.

- True Negative (TN) indicates the samples that are correctly predicted as No-Cancer and are consistent with their actual class.

- False Positive (FP) represents the samples that are predicted as Cancer while their actual class is No-Cancer.

- False Negative (FN) indicates the samples that are expected to be No-Cancer, but are predicted as Cancer.

Using these four basic factors, various evaluation measures can be calculated to assess the performance of the classification models. These measures include precision, recall, F1 score, and the area under the Receiver Operating Characteristic (ROC) curve, among others. Each measure provides valuable insights into different aspects of the model's performance and can help in selecting the most suitable model for breast cancer classification. [5].

Confusion Matrix includes a table that usually applied to represent the fulfillment of a classification model, (or "classifier") on a collection of test data for

which the actual rates are known. Performance of such a method is regularly estimated utilizing the data in the model.

		Predicted	
		Positive	Negative
Real	Positive	True Positive (TP)	False Negative (FN)
	negative	False Positive (FP)	True Negative (TN)

Table 1. Confusion matrix

The classifier can accurately forecast states into their correct section. It is the number of accurate predictions divided by the total amount of cases in the dataset. Accuracy The classifier accuracy is a standard of how great

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

F-Measure. A compressive evaluation method based on precision and recall. It defines in equation (2). Precision and Recall are calculated by equation (3) and equation (4).

$$F - Measure = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (2)$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (3)$$

$$\mathrm{Recall} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \quad (4)$$

G-Mean. The geometric mean is the comprehensive evaluation method constructed with sensitivity and specificity which is presented in equation (5), and Sensitivity and Specificity are calculated in equation (4) and equation (6).[5]

$$G - Mean = \sqrt{\mathrm{Sensitivity} \times \mathrm{Specificity}} \quad (5)$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} \quad (6)$$

3. Datasets

In the context of breast cancer analysis, there are multiple types of datasets available, including imaging data and genetic data. It is common to have more imaging datasets compared to genetic datasets. One widely used imaging dataset is the Wisconsin breast cancer dataset, which is commonly obtained from the UCI Machine Learning Repository. This dataset consists of features extracted from digitized images of Fine Needle Aspirate (FNA) of breast masses.

Apart from the Wisconsin breast cancer dataset, there are several other publicly available breast cancer imaging datasets. These datasets can provide valuable resources for research and analysis purposes. In cases where the dataset is sufficiently large, such as the Digital Database for Screening Mammography (DDSM) dataset, it can be used independently to train and evaluate machine learning models.

However, if the available dataset is not large enough to produce high-performance results, it is possible to merge it with another dataset to enhance the data pool. Merging datasets can help increase the diversity and quantity of data, potentially leading to improved model performance and generalization. By combining multiple datasets, researchers can leverage the collective information contained within them for more comprehensive and accurate analyses.

Overall, the availability of various imaging datasets, along with the option of combining datasets, provides researchers with opportunities to explore and develop robust models for breast cancer analysis.

4. Models

Table 2 Models, classes, and performance for gene sequencing data in selected papers.

Models/ Algorithm	Binary or Multiclass	Classes	Accuracy	Other Performance Evaluation Parameters	Anomaly Application/ Task
IABC-EMBOT, IHM-FFNN, PSORM, ABCO-BCD and DNN-BCD	Binary	Negative, positive	0.975	--	BC detection
FNN, ANFIS, ANNFIS	Binary	Negative, positive	0.92	Precision: 0.944, recall: 0.944, F1:0.944	BC detection
CNN	Multiclass	--	0.956	--	BC subtype classification
DA	Multiclass	--	0.95	--	BC detection
DNN+SVM (separately)	Multiclass	Binary, Miotic/non	0.94	F-score: 0.98	Detection
CNN, SVM	Multiclass	4 tissue categories	0.9	F-score = 0.94 AUC = 0.99	BC detection
CNN	Binary	Tumor or not	0.967	--	--
CNN	Multiclass	7 Cancer types	0.846	--	BC subtypes identification

It is correct that Convolutional Neural Networks (CNNs) are commonly used for breast cancer detection and subtype classification, especially when dealing with MRI images. CNNs have demonstrated excellent performance in computer vision tasks, making them well-suited for analyzing medical images.

In the field of breast cancer research, different studies have employed a variety of algorithms. Some papers have utilized multiple models in a series, combining the strengths of different algorithms to improve performance. On the other hand, some studies have focused on a single model for their analyses.

According to Figure 1, which is referenced in your statement, Artificial Neural Networks (ANNs) and CNNs were the most frequently used algorithms for breast cancer analysis, both for gene sequence data and image data. ANN is a broad term that encompasses different neural network architectures, while CNN is a specific type of neural network designed for image processing tasks.

While other algorithms such as Deep Neural Networks (DNNs) and Support Vector Machines (SVMs) have been utilized in breast cancer research, the majority of papers have employed CNNs and ANNs with various parameter configurations and properties. This reflects the effectiveness and popularity of these algorithms in the context of breast cancer detection and subtype classification.

In summary, CNNs are widely recognized as one of the most suitable algorithms for analyzing MRI images in breast cancer research. ANNs, including CNNs, have been extensively used, while other algorithms have also been explored to a lesser extent. The choice of algorithm depends on the specific research objectives, available data, and desired performance outcomes. [6].

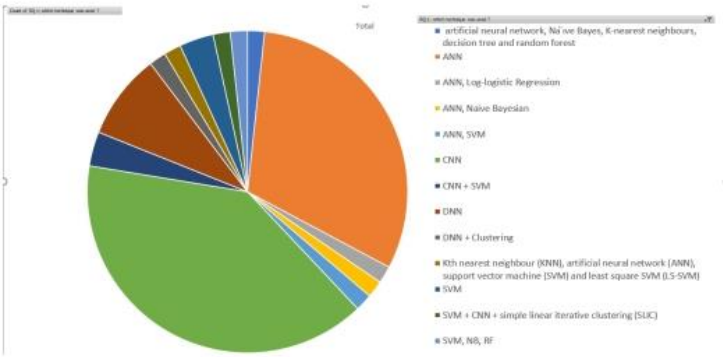

Figure 1 Most commonly used algorithms in the papers

5. Model use CNN and SVM

5.1 Annotation of the Training Set

Based on the CD45-expression, we annotated a training set of image regions (n = 1,116) in the twenty H&E-stained whole-slide images [Figure 1]. While viewing the

consecutively cut H&E and CD45 sections side-by-siFde, we labeled the regions with a raster graphic editor (Adobe Photoshop, Adobe Systems, Mountain View, CA, USA) in downscaled H&E-stained image (1:10, 2.2 µm/pixel). Five entities, four representing different tissue categories and one representing background (BG), were labeled: (1) leukocyte-rich (LR) regions – tissue regions in epithelium and stroma densely populated with TILs. (2) Epithelial (EP) tissue – regions of normal and malignant epithelium with none or few TILs. (3) Stroma predominant regions (SR) – regions of stromal tissue including tissue folds and other tissue types not separately defined with none or few TILs. (4) Adipose tissue (AD) and (5) BG. The TIL-rich and TIL-poor regions were confirmed and selected based on the CD45 expression in the consecutive section.

5.2 Annotation of the Test set

To compare our approach to pathologists' visual assessment at the patient level, we randomly selected 10 images (1,000 × 1,000 pixels, 440 × 440 µm2) excluding areas containing BG from each of the 20 whole-slide image. Two pathologists (P.E.K. and M.M.) visually estimated relative proportions of the different tissue categories of interest (LR, EP, SR, and AD) in this test set of 200 images. The experts were blinded from the results of the automated quantification, and they were asked to estimate the proportions of tissue categories on a continuous scale and independently of each other

5.3 Computerized Quantification of Tumor-infiltrating Lymphocytes

The proposed approach for analysis of whole-slide images in this study consists of three main components: regional segmentation into superpixels, classification of superpixels using CNN activations, and postprocessing. Here is a breakdown of the steps involved:

1. Image Preprocessing: The whole-slide images were downscaled to a resolution of 0.44 µm/pixel and divided into non-overlapping tiles of size 3,000 × 3,000 pixels. Prior to superpixel segmentation, the image tiles were further downscaled to 0.88 µm/pixel and median filtered with a radius of ten pixels. The images were then filtered with an average filter using a radius of 15 pixels and converted to the Lab color-space for subsequent superpixel segmentation.

2. Superpixel Segmentation: Superpixel segmentation is a low-level segmentation method that divides an image into locally similar segments called superpixels. Here, the Simple Linear Iterative Clustering (SLIC) algorithm was used for superpixel segmentation. The SLIC parameters were set to a region size of 50 and a shape parameter of 150. This step over-segments the tissue structures into homogeneous regions.

3. Feature Extraction: A pretrained CNN model, specifically the VGG-F network trained on the ImageNet dataset, was used as a feature extractor. From each superpixel, activations from the penultimate layer of the VGG-F network were extracted. The superpixels were scaled to fit the input size of the pretrained network (224 × 224 pixels).

4. Superpixel Classification: A linear multiclass Support Vector Machine (SVM) classifier with L2 regularization and L2-loss was used for the classification of superpixels. The SVM classifier was trained using a one-vs-rest approach, and the cost parameter (C) was optimized using a 3-fold cross-validation grid search.

5. Postprocessing: To fine-tune the final segmentation, spatial filtering was applied to smooth the classification results. The decision value channel of the background (BG) class was dilated with a circular structuring element (radius of 50 pixels) to minimize classification errors on tissue borders. Each decision value channel was then filtered with a disk-shaped average filter bank using specific radii for different channels. The final segmentation result was obtained through pixel-wise majority voting across the decision channels. Additionally, any adipose tissue (AD) regions smaller than 40,000 pixels were labeled as stroma (SR).

6. Implementation Details: The analysis methods were implemented using MATLAB, utilizing libraries for computer vision and classification, namely LIBLINEAR, VLFeat, and MatConvNet. The computations were performed on a system with a 3.1 GHz quad-core processor and 16GB memory. On average, the analysis of an image tile

(3,000 × 3,000 pixels) took approximately 1 minute, including tasks such as downloading from a remote server, superpixel segmentation, and feature extraction.

This approach allowed for efficient analysis of the whole-slide images, providing segmentation and classification results for different tissue categories.

5.4 Texture Feature

A joint distribution of local binary patterns (LBPs)[27] and local variance (VAR) was used to capture the texture content of the superpixels. We extracted rotation uniform 2 LBP/VAR features without (LBP/VAR) and with (LBP/VAR-KCHI2) an explicit kernel mapping.[28] The LBP/VAR features were computed in a neighborhood of 16 sampling points (n) on radius (r) of 4 pixels. The quantification limits for VAR were set based on 10 random images of the different tissue categories by dividing the VAR distribution into quartiles.

5.5 Results

First, we evaluated how features extracted with the deep learning (VGG-F) network compare to texture features (LBP/VAR) in discrimination of TILs and different tissue categories. Details of the features are listed in Table 1. The training samples (n = 1,116) were segmented into superpixels with SLIC, resulting in a set of 123,442 superpixels representing different categories (LR n = 9,995, EP n = 25,749, SR n = 28,784, AD n = 31,269, and BG n = 27,645). For statistical evaluation, we ran a 3-fold cross-validation 10 times. Overall F-scores for VGG-F, LBP/VAR, and LBP/VAR-KCHI2 were 0.96, 0.89, and 0.92, respectively, while corresponding AUCs were 0.996, 0.983, and 0.984, respectively. The mean sensitivity in the discrimination of TIL-rich regions with the VGG-F based model was 91% (range: 88%–92%), specificity 100% (range: 100%–100%), and precision 96% (range: 96%–97%), respectively.

In all pairwise/inter-category comparisons, features derived with the VGG-F CNN outperformed the LBP/VAR texture features [Table 2]. Kernel mapping improved performance of the texture features in all categories although the difference in classification of TIL-rich regions was marginal (F-score: 0.88 vs. 0.87). AD tissue and BG are more homogeneous in comparison to other tissue categories, and all features discriminated AD and BG superpixels better than others. By definition, the TIL-rich category is composed of the immune cells mixed with different tissue morphologies, and overall it obtained the lowest F-scores[7].

6. Model use DNN and SVM separately

6.1 Methods

We have applied a deep learning approach that extracts the important gene expression relationships using SDAE. After training the SDAE, we selected a layer that has both low-dimension and low validation error compared to other encoder stacks using a validation data set independent of both our training and test set. As a result, we selected an SDAE with four layers of dimensions of 15,000, 10,000, 2,000, and 500. Consequently we used the selected layer as input features to the classification algorithms. The goal of our model is extracting a mapping that possibly decodes the original data as closely as possible without losing significant gene patterns. We evaluated our approach for feature selection by feeding the SDAE-encoded features to a shallow artificial neural network (ANN) and an SVM model.30 Furthermore, we applied a similar approach with PCA and KPCA as a comparison. Lastly, we used the SDAE weights from each layer to extract genes with strongly propagated influence on the reduced-dimension SDAE-encoding. These selected “deeply connected genes” (DCGs) are further tested and analyzed for pathway and Gene Ontology (GO) enrichment. The results from our analysis showed that in fact our approach can reveal a set of biomarkers for the purpose of cancer diagnosis. The details of our method are discussed in the following subsections

6.2 Data

For our analysis, we analyzed RNA-seq expression data from The Cancer Genome Atlas (TCGA) database for both tumor and healthy breast samples. These data consist of 1097 breast cancer samples, and 113 healthy samples. To overcome the class imbalance of the data, we used synthetic minority over-sampling technique (SMOTE) to transform data into a more balanced representation for pre-training. We used the imbalanced-learn package for this transformation of the training data. Furthermore, we removed all genes that had zero expression across all samples

6.3 Dimensionality Reduction Using Stacked Denoising Autoencoder

- Autoencoder (AE): An AE is a feedforward neural network that aims to reconstruct its input layer as closely as possible using a lower-dimensional hidden layer. It consists of an encoder, which maps the input layer to a hidden layer, and a decoder, which reconstructs the input from the hidden layer.

- SDAE: An SDAE is constructed by stacking multiple AEs, with each AE learning a representation at a specific layer. This layer-wise training helps capture complex relationships and reduces information loss compared to reducing dimensionality in one step.

- Training: The parameters (weights and biases) of the SDAE are learned using stochastic gradient descent (SGD) algorithm. The weights are initialized through unsupervised pre-training, where each layer is trained to reconstruct its input. This initialization helps in achieving better optimization.

- Fine-tuning: After pre-training, the SDAE is fine-tuned on the full training set using supervised learning. This step updates the parameters to optimize the SDAE for the specific classification task.

- Regularization: Dropout regularization is applied to prevent overfitting during training. Dropout randomly excludes fractions of hidden units by setting them to zero, which prevents co-adaptation of nodes. Additionally, denoising is used by providing partially corrupted input values to the SDAE, helping it find better representations from noisy data.

- Implementation: The SDAE model was implemented using the Keras library with Theano backend and trained on an Nvidia Tesla K80 GPU. The batch training and GPU utilization enable faster training and testing.

- Performance: The SDAE achieves good performance in a few minutes of training and a few seconds for testing on a sample. Pre-training and the use of denoising and dropout contribute to smoother convergence and better generalization. Fig 2 shows the SDAE encoded, decoded, and denoised representations on the subset of genes.

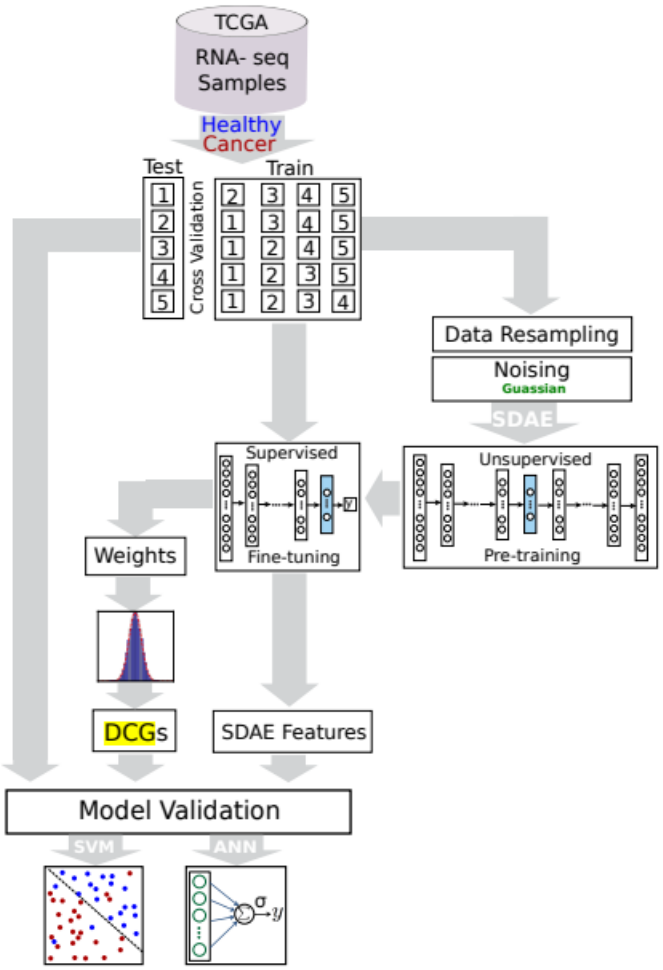

Figure. 2. The pipeline representing the stacked denoising autoencoder (SDAE) model for breast cancer classification and the process of biomarkers extraction

6.4 Differentially Expressed Genes

We used significantly differentially expressed genes as a comparison to our SDAE features for cancer classification. First, we computed the log fold change comparing the median expression n cancer tissue samples to that of healthy tissue samples. We then computed a two-tailed pvalue using a Gaussian fit, followed by a Benjamini-Hochberg (BH) correction.38 We identified two sets of differentially

expressed genes. The first, DIFFEXP0.05 was the 206 genes, 98 upregulated and 118 down-regulated, that were significant at an FDR of 0.05. The second set, DIFFEXP500, contains the top 500 most significant differentially expressed genes (the same dimension as the SDAE features) using the same 2-tailed p-values, containing 244 up-regulated and 256 down-regulated genes.

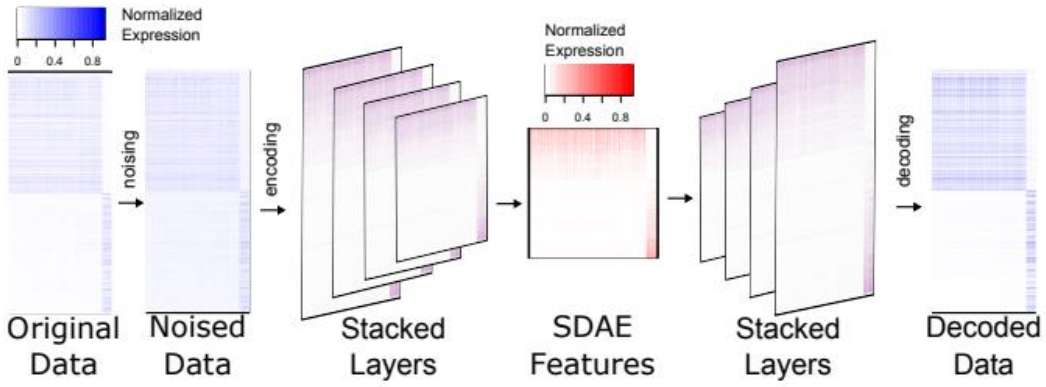

Figure. 3. SDAE representation using the enriched genes in the TCGA breast cancer. In this depiction for illustrative purposes, the top 500 genes with median expression across cancer samples enriched above health samples, and the top 500 genes with reduced median expression across cancer samples is shown

6.5 Results

In order to evaluate the effectiveness of our autoencoder-extracted features, we used two different supervised learning models to classify cancer samples from healthy control samples. First, we considered a single-layer ANN with input nodes directly connected to output layers without any hidden units. If we consider the input units as X = (x1, x2, ..., xn), the output values are calculated as y = σ(P i wixi + b). Second, we considered both an SVM with a linear kernel and with a radial basis function kernel (SVM-RBF). We applied 5-fold cross-validation for to exhaustively split the data into train and test sets to estimate the accuracy of each model without overfitting. In each split, the model was trained on 4 partitions and tested on the 5th, ensuring that training and testing are performed on non-overlapping subsets[8].

	Model	Accuracy	Sensitivity	Specificity	Precision	F-measure
SDAE	ANN	96.95	98.73	95.29	95.42	0.97
	SVM	98.04	97.21	99.11	99.17	0.981
	SVM-RBF	98.26	97.61	99.11	99.17	0.983

Table 3.results

7. conclusion

This research highlights the significant potential of artificial intelligence (AI) in improving the early detection and accurate diagnosis of breast cancer. By employing various machine learning models, including Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and Deep Neural Networks (DNNs), we have demonstrated that AI can effectively classify benign and malignant breast cancer cells, thus supporting early diagnosis and improving patient outcomes.

The study utilized several datasets, such as the Wisconsin breast cancer dataset, to train and evaluate the performance of different AI models. We implemented key methodologies including the annotation of training and test sets, computerized quantification of tumor-infiltrating lymphocytes, and extraction of texture features. Our findings indicate that the combination of DNNs with SVMs yields superior results in terms of classification accuracy compared to CNNs with SVMs.

The superior performance of DNNs combined with SVMs can be attributed to the ability of DNNs to capture more complex patterns and relationships in the data, thanks to their deeper and more flexible architecture. DNNs can learn intricate features that might be overlooked by CNNs, especially when handling high-dimensional datasets. When these rich features are input into an SVM, which excels at handling high-dimensional spaces and avoiding overfitting, the combined model achieves higher accuracy. While CNNs are excellent for spatial feature extraction in images, DNNs provide a more comprehensive feature representation, enhancing the SVM's classification performance.

Moreover, integrating AI into breast cancer detection processes not only enhances diagnostic accuracy but also offers a more efficient and thorough examination of medical data. This can lead to earlier detection of breast cancer, which is crucial for improving patient survival rates. Future research should focus on expanding the datasets used and further refining the AI models to enhance their robustness and generalizability.

In conclusion, the integration of AI techniques into breast cancer diagnostics represents a significant advancement in medical technology. By improving the early detection and accurate classification of breast cancer cells, AI has the

potential to significantly impact patient outcomes and reduce the mortality rate associated with this disease.

References

[1] ZAKIA SULTANA, MD. ASHIKUR RAHMAN KHAN, NUSRAT JAHAN," Early Breast Cancer Detection Utilizing Artificial Neural Network", WSEAS TRANSACTIONS on BIOLOGY and BIOMEDICINE, DOI: 10.37394/23208.2021.18.4

[2] G. Williams, "Descriptive and Predictive Analytics," in Data Mining with Rattle and R, New York, NY: Springer New York, 2011, pp. 171–177

[3] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Comput. Struct. Biotechnol. J., vol. 13, pp. 8–17, Jan. 2015.

[4] Christos Stergiou and Dimitrios Siganos, "Neural Networks." [Online]. Available: https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html. [Accessed: 22-Oct-2018].

[5] S. Dhankhad, E. Mohammed, and B. Far, "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study," in 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018, pp. 122–125.

[6] Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, Yaman Afadar, Omar Elgendy, Breast cancer detection using artificial intelligence techniques: A systematic literature review, Artificial Intelligence in Medicine, Elsevier, Vol 127, May 2022

[7] Riku Turkki , Nina Linder , Panu E. Kovanen , Teijo Pellinen , Johan Lundin. Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples, Journal of Pathology Informatics | Published by Wolters Kluwer – Medknow 2016

[8] Danaee P, Ghaeini R, Hendrix DA. A Deep Learning Approach for Cancer Detection and Relevant Gene Identification. Biocomput. 2017, n.d., p. 219–29. https://doi.org/10.1142/9789813207813_0022.