PROCEEDINGS OF
INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS AND NEW APPLICATIONS
https://proceedings.icisna.org/
2nd International Conference on Intelligent Systems and New Applications (ICISNA'24), Liverpool, April 26-28, 2024.

Advancing Breast Cancer Subtype Prediction and Mutation Analysis: Integrating Deep Learning and Machine Learning Techniques in Genomic Research

Samhita Gadamsetty[1], Dr. Pitchumani Angayarkanni[2]

[1]*ICER, VIT Bangalore, Bengaluru, Karnataka, India*
gsamhita99@gmail.com

[2]*Professor, ICER, VIT Bangalore, Bengaluru, Karnataka, India*
pitchumca@gmail.com, ORCID: https://orcid.org/0000-0001-9621-190X

Abstract*— Breast cancer, a heterogeneous disease, can be classified into several subtypes, each associated with distinct genetic mutations and clinical outcomes. As per the research article by National Institutes of Health it was stated that 25% of hereditary cases are due to the mutation of highly penetrant genes which leads to 80% lifetime risk of breast cancer [15]. This study aims to apply advanced deep learning and machine learning algorithms to predict breast cancer subtypes and to identify key genetic mutations contributing to the disease using a comprehensive gene expression dataset. We analyzed a dataset comprising 1904 samples, encompassing 331 genes and 175 gene mutations, sourced from a public platform and including PAM50 and Claudin low categorizations. Due to limited observations, SMOTE was employed for data augmentation, and Principal Component Analysis (PCA) was used to assess data variance. Several machine learning models, including Random Forest Classifier, Support Vector Machine, K-Nearest Neighbor, XGBoost, and Stacked models, were applied alongside deep learning techniques like Convolutional Neural Network and Multi-Layer Perceptron. The Stacked model demonstrated superior performance with an accuracy of 0.955, outperforming other models. The deep learning models achieved accuracies of 0.911 (CNN) and 0.936 (MLP). KNN analysis revealed potential clusters based on gene and mutation data, with the silhouette metric identifying "siah1_mut," "nras_mut," and "hras_mut" as significant mutations. The optimal clustering achieved a silhouette score of 0.997 for two clusters. These mutations may play pivotal roles in breast cancer pathogenesis and could serve as targets for therapeutic interventions. Our findings demonstrate the effectiveness of integrating stacked algorithms and deep learning models in predicting breast cancer subtypes. The identification of key mutations through clustering techniques provides valuable insights into the genetic underpinnings of breast cancer, which could guide future research and the development of targeted therapies. This study highlights the potential of advanced computational approaches in elucidating the complex landscape of breast cancer genomics and paves the way for personalized medicine in oncology.

Keywords— **Machine Learning, Deep Learning, Breast Cancer, Silhouette metric, Clustering**

I. Introduction

Breast cancer remains the most common malignancy among women worldwide, posing a significant public health challenge. Recent statistics reveal a staggering incidence, with over 2.3 million people globally diagnosed and approximately 685,000 succumbing to the disease. This prevalence underscores the critical need for advanced diagnostic and treatment strategies, particularly for women over the age of 50 who are at increased risk of developing abnormal breast tissue that can lead to cancer.

Genetic alterations are recognized as a primary risk factor in the development of breast cancer. Women inheriting specific cancer genes often exhibit mutations that lead to the formation of cancerous cells in breast tissue. Understanding these genetic underpinnings is crucial for early detection and effective treatment, which are key to improving survival rates and reducing healthcare costs.

The focus of our study is the analysis of a gene expression dataset that quantifies gene activity levels, providing a means to differentiate between normal and abnormal breast tissue. Using PAM50 and claudin low, breast cancer is categorized into six molecular subgroups: Luminal A, Luminal B, HER2, Basal, normal-like, and claudin low. These subtypes are instrumental in understanding the heterogeneity of breast cancer and are closely linked to genetic alterations. Accurate

classification of these subtypes is challenging yet essential for personalized treatment approaches.

Our research aims to categorize breast cancer according to claudin low and PAM50 subtypes, with a particular focus on the aggressive claudin low subtype commonly associated with triple-negative breast cancer (TNBC). To address the challenge of dataset imbalance, we employ SMOTE sampling strategies and analyze 331 genes and 175 gene mutations to identify PAM50 subtypes. The Z-score of mRNA serves as an indicator of tissue abnormality.

In this study, we employ a range of machine learning techniques, including SVM, KNN, NN, NB, DT, XGBoost, and LR, as well as deep learning algorithms, to enhance the prediction accuracy of breast cancer subtypes. This approach builds upon previous studies that have combined decision trees, symbolic classifiers, and neural networks for subtype classification. Additionally, we explore the feasibility of dimensional reduction using PCA to understand the dependencies of data points with respect to subtypes.

Structured into four sections—methodology, calculation results, discussion, and conclusion—this paper aims to not only improve the categorization of breast cancer subtypes but also to identify key gene mutations through clustering techniques. These efforts are directed towards facilitating early detection and enabling more personalized treatment strategies for breast cancer, ultimately contributing to better patient outcomes.

II. LITERATURE REVIEW

*Multiple AI Pipelines for Neoadjuvant Chemotherapy Response Prediction***:** A study by Shen et al. [12] developed multiple AI pipelines to predict the response of breast cancer to neoadjuvant chemotherapy using H&E-stained tissues. This approach used a combination of CNN, SVM, and random forest models. Our work differs by focusing on subtype prediction and mutation analysis, employing a broader range of machine learning and deep learning models, including stacked algorithms, which have demonstrated superior performance.

Optimization of Deep CNN Techniques for Classification and Relapse Prediction: Prasad et al. [10] optimized deep CNN techniques for classifying breast cancer and predicting relapse, achieving high accuracy with hypercomplex-valued CNNs. Our research, while also utilizing deep learning models like CNN and MLP, extends beyond classification to explore genetic mutations using clustering techniques, providing a more comprehensive understanding of the disease.

*Radiomics and Machine Learning for Recurrence-Free Survival Prediction***:** Yu et al. [15] used machine learning radiomics of MRI to predict recurrence-free survival after surgery in breast cancer patients. Our study, in contrast, employs a dataset focused on gene expressions and mutations, offering a different perspective on subtype prediction and the identification of key mutations.

Deep Learning for Breast Cancer Grading Using Synthetic Imaging: Tai et al. [14] introduced a deep learning approach for breast cancer grading using synthetic correlated diffusion imaging. Our approach is novel in its application of machine learning and deep learning for subtype prediction and mutation analysis, rather than grading, and does not rely on imaging data.

The proposed research stands out for its comprehensive use of various machine learning and deep learning models for subtype prediction and mutation analysis in breast cancer, a different focus compared to the recent studies that primarily concentrated on chemotherapy response prediction, classification and relapse prediction, recurrence-free survival prediction, and cancer grading.

III. METHODOLOGY

A. *General Idea*

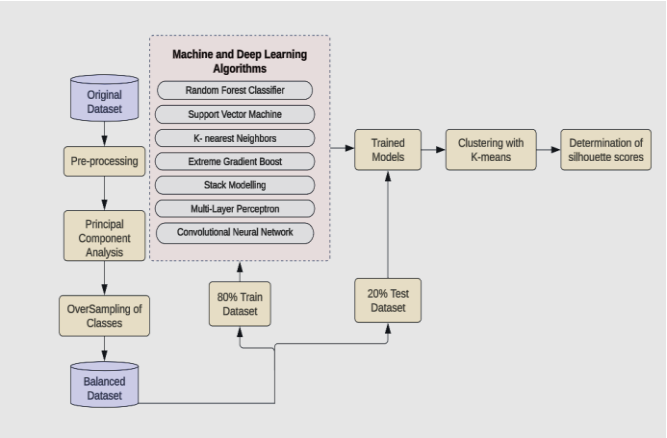

Fig. 1 Architecture Diagram

The dataset, integrating clinical and genetic expression data, was sourced from a public platform. In the preprocessing stage, data was managed using label encoding and visualized via principal component analysis (PCA). To address data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied [1]. The dataset was divided into 80% for training and 20% for testing. The final stage involved employing K-means clustering with silhouette scoring to understand gene expression clustering.

B. *Model Implementation*

In the preprocessing phase, intrinsic subtypes were encoded as shown in Table 1.

TABLE 1
REPRESENTATION OF CLASS NAMES AFTER ORDINAL ENCODER APPLICATION

S.no	Class Name	Representation
1.	Basal	0
2.	HER2	1
3.	LumA	2
4.	LumB	3
5.	Normal	4
6.	Claudin Low	5

PCA was utilized to understand data point variances [1][11], employing three eigenvectors to represent variance in gene data points. This analysis highlighted the significance of all genes and gene mutations in subtype identification.

To address potential outliers in the dataset, a robust scaler was used, employing the formula:

$$\text{Scaled Value} = \frac{X - Q1}{Q3 - Q1} \quad (1)$$

X is original feature
Q1 is 25th percentile of the feature
Q3 is 75th percentile of the feature

A stratified split followed, using 20% for testing and 80% for training. Classification employed machine learning methods including stacked models, K-nearest neighbor, Random Forest Classifier, Support Vector Machine, and Extreme Gradient Boost, alongside deep learning algorithms like convolutional neural networks and multi-layer perceptrons.

C. *Machine Learning Models*

Random Forest Classifier: This model aggregates multiple trees, with the highest voting leading to the best result. Default settings were used, focusing on criteria like gini impurity and splitter strategy.

Support Vector Machine (SVM): SVM was employed with a focus on the cost parameter "C" and the kernel type, set to "linear" for handling multiclass classification.

K-nearest Neighbors (KNN): This model was used for class clustering, with the primary hyperparameter, n_neighbors, left unspecified to allow the algorithm to learn from the training set.

Extreme Gradient Boost (XGBoost): XGBoost settings included the "gbtree" booster for a tree-based model, the "multi:softmax" objective for multi-class classification, and the default n_estimators value of 100.

Stacked Modeling: This approach combined Random Forest Classifier, SVM, Multi-layer Perceptron, and XGBoost, with SVM as the meta-model.

D. *Deep Learning Models*

Multi-Layer Perceptron: A feedforward neural network with a single hidden layer of 100 neurons and "ReLu" activation function. The max_iter was set to 100.

Convolutional Neural Network (CNN): Employed 64 filters in a one-dimensional network with "ReLu" activation and max pooling.

IV. CALCULATIONS RESULTS

Performance metrics of the models were evaluated, focusing on accuracy, precision, recall, and F1-score.

The accuracy is defined as, [1] the ratio of correctly predicted instances to the total number of instances.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (2)$$

Precision is a metric that determines the ratio of true positive predictions to the total number of positive predictions[1].

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positives + False Positives}}. \quad (3)$$

Recall is known as sensitivity, it is evaluated by the ratio of true positive predictions to total number of actual positive observations[1].

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positives + False Negatives}}. \quad (4)$$

F1-score is combination of both Recall and Precision[1], it provides balance between both and calculated as,

$$\text{F1-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

The Table 2 below shows all the metrics with respect to the models implemented.

TABLE 2
PERFORMANCE METERICS OF IMPLEMENTED MODELS

Model Implemented	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest Classifier	0.92	0.92	0.92	0.92
Support Vector Machine	0.93	0.93	0.93	0.93
K-nearest neighor	0.76	0.82	0.76	0.69
Extreme gradient boost	0.92	0.92	0.92	0.92
Stack Modelling	0.96	0.96	0.96	0.96
Multi-layer Perceptron	0.92	0.91	0.92	0.91
Convolutional Neural Network	0.94	0.94	0.94	0.94

V. DISSCUSSIONS

The figure 2 represents a three-dimensional scatter plot derived from a Principal Component Analysis (PCA) of the dataset, encompassing a variety of sample categories including 'Basal', 'HER2', 'LumA', 'LumB', 'Normal', and 'Claudin-low'. The axes correspond to the first three principal components which have been extracted to capture the maximum variance within the dataset, with the x-axis representing Principal Component 1 (PC1), the y-axis representing Principal Component 2 (PC2), and the z-axis representing Principal Component 3 (PC3).

The scatter plot elucidates the distribution and separation of the samples across the principal components, highlighting the inherent clustering by sample type. The 'Basal' and 'Claudin-low' samples display a distinctive pattern along PC1, suggesting that the variables contributing to PC1 are particularly divergent for these categories compared to the 'LumA', 'LumB', 'HER2', and 'Normal' samples. The distribution of samples along PC2 and PC3 further illustrates the multidimensional variance within the dataset, allowing us to discern patterns that are not observable in lower-dimensional spaces.

This PCA plot is instrumental in understanding the underlying structure of the data, providing insight into the characteristics that differentiate the sample types. The discernible separation along the principal components supports the hypothesis that significant molecular variations underpin the categorization of the samples. It should be noted, however, that while this visualization captures a significant portion of the dataset's variability, it does not encapsulate all the multidimensional relationships. PCA analysis revealed significant data loss, emphasizing the importance of every gene and mutation in subtype identification. The datasets were thus fully utilized for classification.

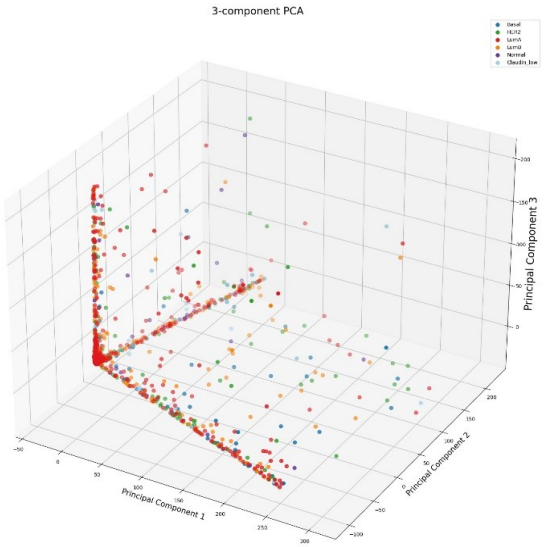

Fig. 2 The visual provides us with information about the distribution of gene expression, they are highly concentrated

Afterwards, running models on the relevant dataset, research was done to identify the best biomarkers that would be crucial in identifying breast cancer. Elbow graph analysis identified optimal clusters, with K-means used for further exploration.

The figure presented illustrates the results of a Kmeans clustering analysis, specifically the distortion scores for different numbers of clusters (k). The distortion score, which measures the sum of squared distances from each point to its assigned center, is a common metric used to evaluate the quality of a clustering model. An elbow method has been employed to determine the optimal number of clusters by identifying the point where the distortion score begins to diminish at a slower rate, which is indicative of a natural division within the data.

As demonstrated in the figure, the elbow point is identified at k=2, with a silhouette score of approximately 0.568. This inflection point suggests that increasing the number of clusters beyond three yields diminishing returns in terms of reducing the distortion score. The elbow at k=2 indicates that the within-cluster sum of squares (WCSS) does not significantly decrease with the addition of more clusters, hence the dataset is optimally partitioned into two clusters.

This finding has significant implications for the structure of the dataset being analyzed. It implies that the data can be naturally divided into two distinct groups, each representing a potentially different category, behavior, or type within the overall population. This partitioning can be particularly useful for subsequent analyses, where the characteristics of each cluster may be examined to uncover patterns or trends that were not previously apparent.

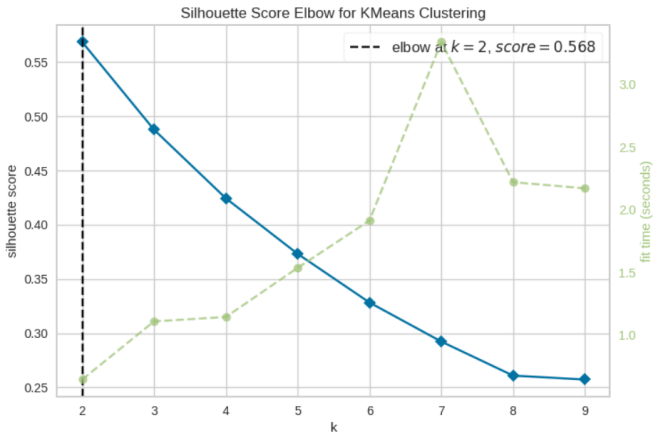

Fig. 3 The Elbow determines the K value of the dataset with respect to the measure of WCSS (Within clusters sum of squares)

The results indicated equal contribution of all genes and mutations in forming breast cancer subtypes. Silhouette scores were calculated for cluster validation, with the highest score of 0.997 for two clusters, identifying key mutations.

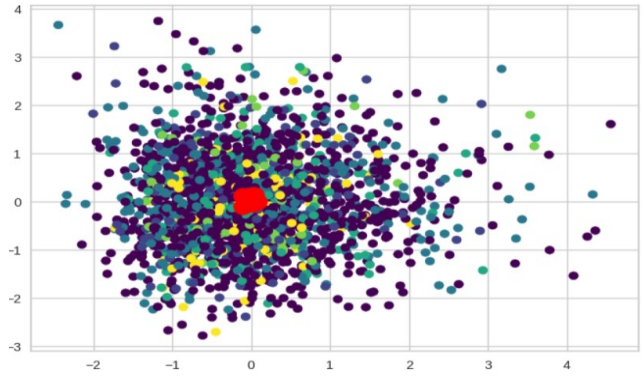

Fig 4. The scatter plot of K-means with the centroids overlapping for the given 6 clusters

The outcome in *figure 4* suggests a complex and intertwined relationship among genes and their mutations in relation to the development of breast cancer subtypes. The depicted scatter plot, characterized by merging clusters with overlapping centroids, illustrates the challenge in segregating breast cancer subtypes based solely on the classification of gene expression data. This overlap implies a high degree of similarity in the genetic profiles across different subtypes, indicating that there may not be distinct sets of genes or mutations uniquely defining each subtype.

The lack of clear demarcation between clusters can be interpreted as a sign that the genetic underpinnings of breast cancer are multifaceted, with a multitude of genes and mutations contributing collectively to the phenotype of the cancer subtypes. As a result, it appears that no single gene or mutation is solely responsible for the differentiation of subtypes; rather, it is the combined effect of multiple genetic factors that contributes to the disease's heterogeneity.

This outcome challenges the expectation of identifying discrete genetic signatures for each breast cancer subtype. It underscores the necessity for more sophisticated analytical methods or the inclusion of additional data types, such as epigenetic or proteomic data, to enhance the resolution of subtype classification. The findings call for a holistic approach to understanding the genomic landscape of breast cancer, moving beyond the identification of individual genes or mutations to a more integrative perspective that considers the complex interplay of the entire genomic milieu.

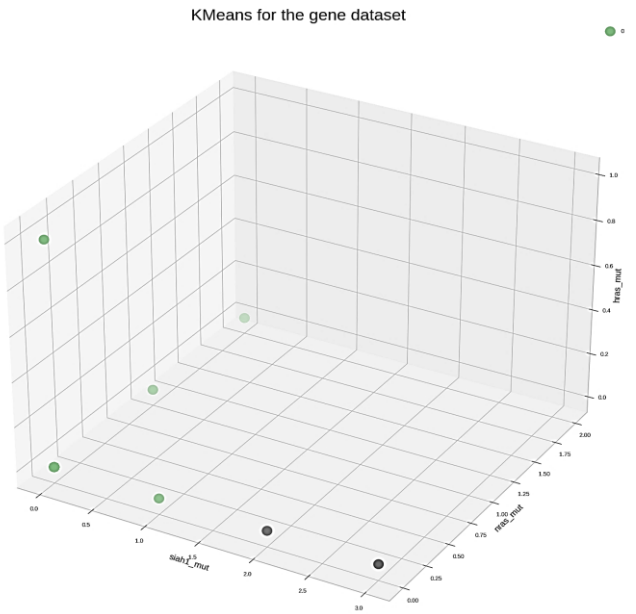

Fig 5. The silhouette scores for 2 clusters in gene expression

The figure appears to be a 3D scatter plot, but the axes labels are not fully readable in the provided image, which constrains the ability to give a precise interpretation. However, the provided context suggests that this plot is related to clustering analysis outcomes for breast cancer tissue samples based on gene mutations.

Incorporating the context provided and assuming the axes represent different mutation types or perhaps mutation frequencies, and one of the axes measures the silhouette score (a measure of how similar an object is to its own cluster compared to other clusters), the following interpretation can be modified accordingly:

Figure 5 provides an insightful visualization into the clustering analysis of breast cancer tissue samples based on the presence and mutations of certain genes. The 3D scatter plot emphasizes that when the dataset is partitioned into two clusters, mutations play a more significant role than the mere presence of certain genes in defining the subtypes of breast cancer tissue. This is evidenced by the more pronounced impact of gene mutations on the clustering outcome.

The silhouette scores, which gauge the appropriateness of the cluster formations, support this observation. With inputs ranging from 2 to 6 clusters, the silhouette score for the two-cluster solution is remarkably high at 0.997, suggesting excellent intra-cluster similarity and inter-cluster separation for mutations 'siah1_mut', 'nras_mut', and 'hras_mut'. This high score indicates a very strong cluster structure, where each cluster is well differentiated from the others.

For the three-cluster configuration, mutations 'prps2_mut', 'ndfip1_mut', and 'mbl2_mut' are identified as significant, with a silhouette score just a notch lower at 0.996, still indicative of robust cluster delineation. As the number of clusters increases to four, mutations 'smarcb1_mut', 'stmn2_mut', and 'klrg1_mut' emerge as significant, with the silhouette score slightly decreasing to 0.995. This trend continues with five clusters, where 'hist1h2bc_mut', 'smarcd1_mut', and 'nr2f1_mut' are highlighted, maintaining a silhouette score of 0.995.

The six-cluster solution presents a silhouette score of 0.993 for mutations 'agtr2_mut', 'ppp2cb_mut', and 'sgcd_mut', which, while still high, suggests that the distinctness of clusters may begin to diminish as the number of clusters increases. The consistent high silhouette scores across cluster configurations underscore the analysis's robustness, yet the optimal clustering, as depicted in the plot, is achieved with two clusters. This suggests a potential stratification of breast cancer tissue based on these genetic mutations, which may have implications for personalized treatment approaches and understanding the etiology of cancer subtypes.

VI. Conclusions

In summary, this investigation has successfully leveraged the strengths of machine learning and deep learning techniques to enhance the classification of breast cancer subtypes. The application of SMOTE to balance the dataset and the strategic decision to avoid dimensionality reduction have both been instrumental in preserving the integrity of genetic data. The K-means clustering outcomes, augmented by silhouette score assessments, have facilitated the identification of significant clusters that are paramount in understanding the genetic drivers of breast cancer. These findings not only contribute to the current body of knowledge but also hold promise for the development of more personalized and precise treatment protocols. The high accuracy of the predictive models developed in this study reaffirms the transformative impact that computational methodologies can have in the domain of genomic medicine.

References

[1] Anđelić, N., & Baressi Šegota, S. (2023, June 29). Development of Symbolic Expressions Ensemble for Breast Cancer Type Classification Using Genetic Programming Symbolic Classifier and Decision Tree Classifier. *Cancers*, *15*(13), 3411. https://doi.org/10.3390/cancers15133411

[2]Anđelić, N., Baressi Šegota, S., Glučina, M., & Lorencin, I. (2023, February 2). Classification of Faults Operation of a Robotic Manipulator Using Symbolic Classifier. *Applied Sciences*, *13*(3), 1962. https://doi.org/10.3390/app13031962

[3] Alromema, N., Syed, A. H., & Khan, T. (2023, February 13). A Hybrid Machine Learning Approach to Screen Optimal Predictors for the Classification of Primary Breast Tumors from Gene Expression Microarray Data. *Diagnostics*, *13*(4), 708. https://doi.org/10.3390/diagnostics13040708

[4] Chen, Y., Gu, Y., Hu, Z., & Sun, X. (2020, October 30). Sample-specific perturbation of gene interactions identifies breast cancer subtypes. *Briefings in Bioinformatics*, *22*(4). https://doi.org/10.1093/bib/bbaa268

[5] Ghozy Ghulamul Afif, Adiwijaya, & Widi Astuti. (2021, August 26). Cancer Detection based on Microarray Data Classification Using FLNN and Hybrid Feature Selection. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(4), 794–801. https://doi.org/10.29207/resti.v5i4.3352

[6]Ibrahim, N. M., Ali, B., Jawad, F. A., Qanbar, M. A., Aleisa, R. I., Alhmmad, S. A., Alhindi, K. R., Altassan, M., Al-Muhanna, A. F., Algofari, H. M., & Jan, F. (2023, June 15). Breast Cancer Detection in the Equivocal Mammograms by AMAN Method. *Applied Sciences*, *13*(12), 7183. https://doi.org/10.3390/app13127183

[7]Jiang, Q., & Jin, M. (2021, February 26). Feature Selection for Breast Cancer Classification by Integrating Somatic Mutation and Gene Expression. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.629946

[8] Nassif, A. B., Talib, M. A., Nasir, Q., Afadar, Y., & Elgendy, O. (2022, May). Breast cancer detection using artificial intelligence techniques: A systematic literature review. *Artificial Intelligence in Medicine*, *127*, 102276. https://doi.org/10.1016/j.artmed.2022.102276

[9] Piccolo, S. R., Mecham, A., Golightly, N. P., Johnson, J. L., & Miller, D. B. (2022, March 11). The ability to classify patients based on gene-expression data varies by algorithm and performance metric. *PLOS Computational Biology*, *18*(3), e1009926. https://doi.org/10.1371/journal.pcbi.1009926

[10] Prasad, V. V., Venkataramana, L. Y., S Keerthana, & Subha R. (2023, November 27). Optimization of Deep CNN Techniques to Classify Breast Cancer and Predict Relapse. *Journal of Advanced Zoology*, *44*(4), 774–787. https://doi.org/10.17762/jaz.v44i4.2182

[11] Shaban, W. M. (2022, December 1). Insight into breast cancer detection: new hybrid feature selection method. *Neural Computing and Applications*, *35*(9), 6831–6853. https://doi.org/10.1007/s00521-022-08062-y

[12] Shen, B., Saito, A., Ueda, A., Fujita, K., Nagamatsu, Y., Hashimoto, M., Kobayashi, M., Mirza, A. H., Graf, H. P., Cosatto, E., Hazama, S., Nagano, H., Sato, E., Matsubayashi, J., Nagao, T., Cheng, E., Hoda, S. A., Ishikawa, T., & Kuroda, M. (2023). Development of multiple AI pipelines that predict neoadjuvant chemotherapy response of breast cancer using H&E-stained tissues. *The journal of pathology. Clinical research*, *9*(3), 182–194. https://doi.org/10.1002/cjp2.314

[13] Shiovitz S, Korde LA. Genetics of breast cancer: a topic in evolution. Ann Oncol. 2015 Jul;26(7):1291-9. doi: 10.1093/annonc/mdv022. Epub 2015 Jan 20. PMID: 25605744; PMCID: PMC4478970.

[14]Tai, C., et al. (2023). Cancer-Net BCa-S: Breast Cancer Grade Prediction using Volumetric Deep Radiomic Features from Synthetic Correlated Diffusion Imaging.

[15] Yu, Y., Ren, W., He, Z., Chen, Y., Tan, Y., Mao, L., Ouyang, W., Lu, N., Ouyang, J., Chen, K., Li, C., Zhang, R., Wu, Z., Su, F., Wang, Z., Hu, Q., Xie, C., & Yao, H. (2023, November 1). Machine learning radiomics of magnetic resonance imaging predicts recurrence-free survival after surgery and correlation of LncRNAs in patients with breast cancer: a multicenter cohort study. *Breast Cancer Research*, *25*(1). https://doi.org/10.1186/s13058-023-01688-3

.