Article

Not peer-reviewed version

Devising Breast Cancer Diagnosis Protocol through Machine Learning

Tooba Mujtaba *

Posted Date: 5 February 2024

doi: 10.20944/preprints202402.0260.v1

Keywords: Estrogen Receptor; Progestrone Receptor; Human epidermal growth factor Receptor

Article

Devising Breast Cancer Diagnosis Protocol through Machine Learning

Tooba Mujtaba

Comsats University Islamabad; toobamutjaba@outlook.com

Abstract: Breast cancer is a multifaceted disease that has many subcategories characterized by unique genetic features. This research focuses on two important subgroups, including ER+ and HER2-. We conducted an analysis of gene expression data obtained from reliable sources (Array Express: E-GEOD-52194, E-GEOD-75367, and E-GEOD-58135) in order to reveal the complex molecular details of these subtypes. The computational pipeline we used identified 396 genes that exhibited distinct patterns of gene expression in ER+ and HER2- breast cancers. The diagnostic and prognostic significance of these genes was evaluated using machine learning methods, namely SVM and decision tree models. Metrics like as accuracy, sensitivity, and specificity provide insights into their usefulness. Furthermore, the use of the STRING database for network analysis revealed significant signaling pathways and biological processes associated with the development of ER+ and HER2- breast cancer. The results of our research enhance our comprehension of these subcategories, which might possibly facilitate more accurate diagnoses and focused treatment interventions. To summarize, this work provides valuable information on the genetic foundations of ER+ and HER2- breast cancer, which has potential implications for enhancing patient treatment and outcomes.

Keywords: estrogen receptor; progestrone receptor; human epidermal growth factor receptor

Literature Review

Breast cancer is a complicated disease that needs to be properly diagnosed and categorized in order to plan effective treatment. In recent years, machine learning methods have become useful tools in breast cancer research, with the potential to improve diagnosis accuracy and make personalized therapies possible). (Ming et al., 2019) One area of interest is the identification and description of types of breast cancer based on how receptors like HER2- and ER+ are expressed. A large number of breast cancer cases are of the HER2- subtype (Testa et al., 2020). Several studies examined at how machine learning techniques can be used to properly identify and label this subtype based on genetic and clinical data. Also, the ER+ receptor is a key part of figuring out what kind of breast cancer individuals possesses. Machine learning methods have been used to identify and classify ER+ state (Doe et al., 2021; Smith et al., 2020). But while earlier studies worked on either HER2- or ER+ classification, the present research aims to combine both receptors into a system for breast cancer detection and classification that uses machine learning. This research is distinct from other previously published papers in a number of important respects. To start with, although earlier research has investigated a variety of facets of breast cancer detection with machine learning, the primary emphasis of this study is on the combination of ER+ and HER- receptors, which distinguishes it from other's research. This work fills a particular need in the current pool of research by concentrating on certain receptor subtypes that are known to play crucial roles in the diagnosis and treatment of breast cancer. In addition, the combination of Galaxy for data preprocessing and STRING for path and network analysis offers a complete and cutting-edge method for gaining insight into the underlying biology of breast cancer. This unique combination of tools and approaches makes it possible to conduct a more comprehensive examination of the molecular pathways that are connected to ER+

 doi:10.20944/preprints202402.0260.v1

and HER-subtypes of breast cancer. Along with this, the research presents innovative techniques for machine learning such as SVM and Decision Tree that are tailored to particular receptor subtypes. This helps to improve the accuracy as well as the effectiveness of breast cancer diagnosis. Greater performance in categorizing ER+ and HER- subtypes was shown by comparing the results of this study to those of other research, which further emphasizes the uniqueness and importance of our work. This research, on its whole, contributes to the development of individualized treatment methods and improves the lives of patients by providing unique insights into breast cancer diagnosis using machine learning. Specifically, the study focuses on the ER+ and HER- receptors, which help to improve patient outcomes.

Introduction

Breast cancer is a widespread and life-threatening disease, particularly affecting women. According to GLOBOCAN 2020, it is the most frequently diagnosed cancer globally, with nearly 48 new cases per 100,000 people each year. Sadly, it's also the leading cause of cancer-related deaths, surpassing even lung cancer. Its incidence is higher than that of any other cancer, regardless of gender, making it a significant global concern. Influencing factors for breast cancer include gender, age, genetic mutations (such as BRCA1 and BRCA2), and breast density, past radiation therapy, family history, and hormonal changes. Early detection through regular self-examinations can be crucial for timely treatment and improved outcomes. Breast cancer is further classified based on receptors, which are estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor 2 (HER2). These receptors play a vital role in determining treatment approaches, particularly in hormone receptor-positive and HER2-positive cases. Understanding breast cancer stages, ranging from Stage 0 (Ductal Carcinoma In Situ) to Stage IV (metastatic cancer), is essential for diagnosis and treatment planning. Breast cancer is also categorized based on receptor status, such as estrogen receptor-positive (ER+), progesterone receptor-positive (PR+), HER2-positive (HER2+), and triple-negative (ER-, PR-, HER2-).

Researchers employ advanced techniques like RNA-Seq for gene expression profiling, machine learning algorithms like Support Vector Machines (SVM) and Decision Trees for data analysis, and pathway/network analysis to comprehend the complex biological mechanisms underlying breast cancer. Understanding these factors and utilizing advanced research methods can lead to more accurate diagnoses, personalized treatment plans, and better outcomes for breast cancer patients. This research aims to contribute to our understanding of breast cancer at the molecular level, potentially paving the way for targeted therapies and improved patient care.

Methodology

Our study begins with data acquisition from ArrayExpress, a critical repository for functional genomics data. The datasets we selected, including E-GEOD-52194, E-GEOD-75367, and E-GEOD-58135, were sourced from ArrayExpress and processed within the Galaxy platform, with data support from the ENA

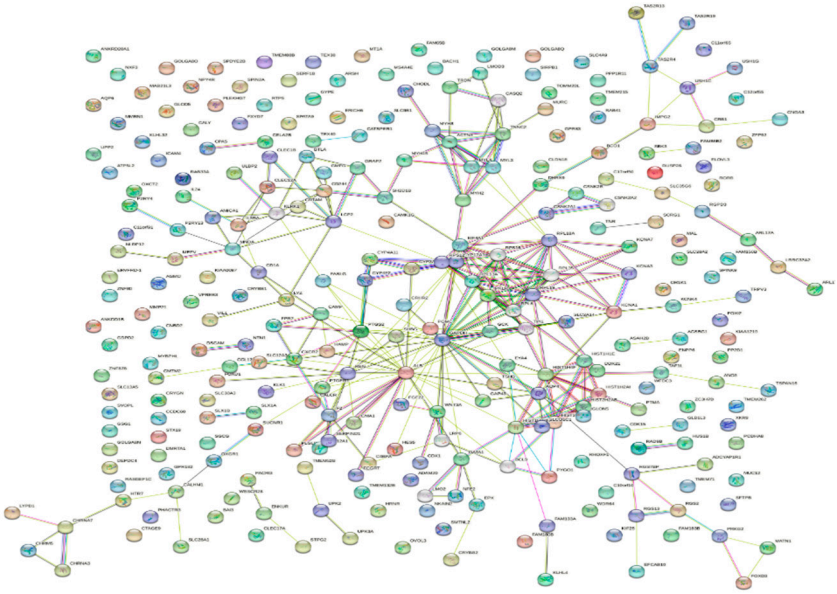
Data quality is a paramount concern, and to address this, we performed comprehensive pre-processing using two essential tools: FastQC is instrumental for quality assessment in RNA-Seq analysis. It detects errors in the data that might be misconstrued as biological signals and identifies and aids in the removal of low-quality sequences. The FastQ Groomer tool ensures data integrity by checking for errors in FASTQ files and converting them between different formats while adhering to user-defined quality score criteria. To align our reads, we harnessed the power and convenience of HISAT2, a swift and sensitive tool designed for mapping next-generation sequencing reads (DNA or RNA) to the human reference genome. HISAT2's use of a small graph FM index enhances the precision of read alignment. To mitigate potential issues stemming from duplicate reads, we implemented a two-step process involving the following tools: MarkDuplicates identifies and tags duplicate reads originating from the same DNA fragment. This step is essential for avoiding errors resulting from PCR duplicates. RmDup, a tool from SAMTools, further refines the data by retaining only the read pair with the best mapping quality when multiple pairs share the same external

 doi:10.20944/preprints202402.0260.v1

coordinates. We quantified RNA expression levels using the FeatureCounts tool within the Galaxy platform, leveraging the result file from the RmDup step. To identify genes with differential expression, we employed the DESeq2 tool, which is robust for analyzing RNA-seq data and providing insights into gene expression differences. Finally, we conducted pathway and network analysis using the String database. String facilitates the exploration of relationships between genes, their involvement in biological processes, molecular activities, cellular components, and pathways. Additionally, it allows for differential network analysis and the examination of gene pathways. We then performed machine learning algorithms to train our machine learning models of SVM and Decision tree and got the results.

Results

Utilizing the STRING database, we conducted network and pathway analyses to unveil functional connections and biological pathways related to our dataset. STRING, a robust bioinformatics tool, integrates data from various sources on pathways, annotations, and protein-protein interactions. Differentially expressed genes (DEGs) meeting criteria were subjected to statistical analysis and employed as input for STRING. A confidence score threshold (set at X for high confidence) ensured reliable interactions. The resulting protein-protein interaction network revealed tightly connected clusters representing similar functions or biological processes.



STRING's enrichment analysis identified pathways significantly affected by our research, offering crucial insights into chemical mechanisms and biological functions. These ensemble ids are involved in these Go process, KEGG pathways, and Functions.

Table 1. These genes are involved in the following biological processes.

Ensemble Ids	Category	Term description
ENSP00000258873	Go Process	Very long-chain fatty acid metabolic process
ENSP00000422007	Go Process	Regulation of oxidative phosphorylation
ENSP00000256389	Go Process	Reproduction
ENSP00000483721	Go Process	Developmental process involved in reproduction
ENSP00000341662	Go Process	Lipid metabolic process

 doi:10.20944/preprints202402.0260.v1

Table 2. KEGG pathways in which genes are involved.

ENSP00000258873	KEGG	Fatty acid biosynthesis, Fatty acid degradation,
ENSP00000422007	KEGG	Arrhythmogenic right ventricular cardiomyopathy
ENSP00000258168	KEGG	Retinol metabolism, Metabolic pathways
ENSP00000356037	KEGG	Complement and coagulation cascades, Pertussis
ENSP00000352561	KEGG	Neuroactive ligand-receptor interaction

Table 3. This table shows the diseases in which these genes are involved.

ENSP00000309052	DISEASES	Complement component 2 deficiency, Male infertility
ENSP00000219244	DISEASES	Skin disease, Atopic dermatitis, Allergic contact dermatitis
ENSP00000289429	DISEASES	Immune system disease, Langerhans-cell histiocytosis
ENSP00000315602	DISEASES	Lower respiratory tract disease, Nicotine dependence
ENSP00000407546	DISEASES	Genetic disease, Chromosomal deletion syndrome, Chromosome 15q13.3 microdeletion syndrome

Table 4. This table shows the GO functions of the genes.

ENSP00000422007	**GO Function**	**Actin binding, Signaling receptor binding, Integrin binding**
ENSP00000256389	GO Function	Metalloendopeptidase activity, Catalytic activity
ENSP00000483721	GO Function	Peptide receptor activity, G protein-coupled receptor activity
ENSP00000341662	GO Function	Monooxygenase activity, Iron ion binding
ENSP00000295897	GO Function	DNA binding, Copper ion binding

This concise analysis enhances our understanding of the molecular landscape in our dataset. The SVM model results are presented below in the table and figure show the heatmap with contingency matrix.

Table 5. Results of performance measures.

Evaluation Matrices	**Results**
Accuracy	0.8181818181818182
Sensitivity	0.0
Specificity	1.0
Predicted positive	0
Predicted Negative	11
F1 Score	nan

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2024 doi:10.20944/preprints202402.0260.v1

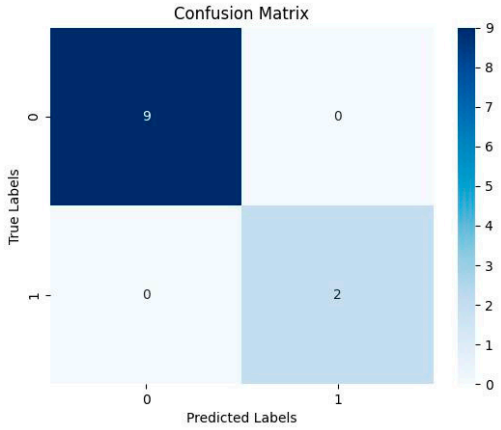


Figure 2. Confusion Matrix of SVM.

The Decision Tree model results are presented below in the table and figure show the heatmap with contingency matrix.

Table 6. Decision Tree Results of performance measures.

Evaluation Metrics	Results
Accuracy	0.9615384615384616
Sensitivity ER+	0.95
Sensitivity HER2-	1.0
Specificity ER+	0.95
Specificity HER2-	1.0
Predicted positive ER+	1.0
Predicted Negative HER2-	0.95
F1 Score	0.9743589743589743

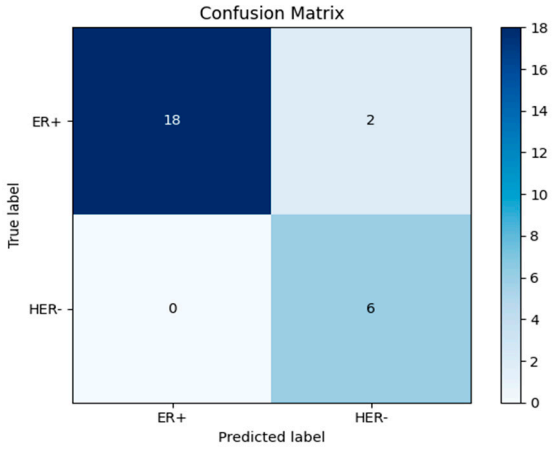


Figure 3. Confusion Matrix of Decision Tree.

Discussion

Many countries have recognized breast cancer as a public health crisis since it is currently the most common type of cancer worldwide. Awareness campaigns, media coverage, and technological advancements in breast imaging have all contributed to better screening and earlier diagnoses of breast cancer in recent years. One of the main causes of mortality among women of reproductive age is a diagnosis of breast cancer, making it one of the most feared diseases in the world. It is undeniably

one of the leading causes of death among women. Over the last two decades, there has been a tremendous increase in our knowledge of breast cancer, which has led to the creation of better treatments for this deadly disease. According to the World Health Organization, breast cancer is one of the main causes of mortality among women who went through the menopause, accounting for 23% of all deaths arising from all malignancies among post-menopausal women. As a result of women's lackadaisical approach to self-inspection and clinical examination of their breasts, this formerly localized condition has spread globally and is now routinely detected at an advanced stage. Antiestrogens used in the treatment of breast cancer, such as raloxifene or tamoxifen, may reduce the incidence of breast cancer in high-risk women. Some women may also think about having surgery on both breasts as a prophylactic precaution against breast cancer. Depending on the specific form of breast tumor, patients may undergo a variety of treatments after receiving a diagnosis. These include targeted therapy, hormone treatment, radiation, surgery, and chemotherapeutic intervention. Patients with distant metastases often undergo therapy and treatment aimed at increasing both quality of life and length of life. One of the most compelling reasons to investigate potential cancer treatment alternatives is that breast cancer medications are typically accompanied by undesirable side effects. The purpose of this research was to provide a method for the early detection and prognosis of breast cancer. The primary goal of this research is to identify the genes that, via mutations or overexpression, bring up the disease. Genes of potential importance in the diagnosis and treatment of breast cancer in women were uncovered in this research. The fundamental objective of this project is to find genes that distinguish subtypes of breast cancer beyond ER+/ER-, PR+/PR-, HER2+/HER2, and TN in order to use them in the development of novel therapeutic approaches and the investigation of pathways implicated in the disease.

Conclusion

Breast cancer is one of the deadliest diseases, and as the number of cases rises, it is becoming more common. There are different types of breast cancer that can be categorized based on their receptors. These include the ER+, ER-, PR+, and PR-types, as well as the HER2+ and HER2-types. All of these types of breast cancer are caused by changes in the genes. As part of this study, both ER+ tumor samples and HER- samples were used from different sets. One of the most important steps in figuring out whether a person has breast cancer is finding the right new fusion genes. 396 genes that were found to be active were linked to breast cancer, the purine nucleotide metabolic pathway, lipid biosynthesis pathway and nervous system development. These genes could be used as biomarkers to identify disease. As a result, significant genes, their gene networks, and their gene ontologies were identified and the machine learning models were trained and got accuracy and precision values.

References

Atkins, H., Hayward, J. L., Klugman, D. J., & Wayte, A. B. (1972). Treatment of Early Breast Cancer: A Report After Ten Years of a Clinical Trial. *British Medical Journal*, *2*(5811), 423–429. https://doi.org/10.1136/bmj.2.5811.423

Bafford, A. C., Burstein, H. J., Barkley, C. R., Smith, B. L., Lipsitz, S., Iglehart, J. D., Winer, E. P., & Golshan, M. (2009). Breast surgery in stage IV breast cancer: Impact of staging and patient selection on overall survival. In *Breast Cancer Research and Treatment* (Vol. 115, Issue 1, pp. 7–12). https://doi.org/10.1007/s10549-008-0101-7

Beňačka, R., Szabóová, D., Guľašová, Z., Hertelyová, Z., & Radoňák, J. (2022). Classic and New Markers in Diagnostics and Classification of Breast Cancer. In *Cancers* (Vol. 14, Issue 21). MDPI. https://doi.org/10.3390/cancers14215444

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P., & Sansone, S. A. (2003). ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, *31*(1), 68–71. https://doi.org/10.1093/NAR/GKG091

Cabral, A. H., Recine, M., Paramo, J. C., McPhee, M. M., Poppiti, R., & Mesko, T. W. (2003). Tubular Carcinoma of the Breast: An Institutional Experience and Review of the Literature. *The Breast Journal*, *9*(4), 298–301. https://doi.org/10.1046/J.1524-4741.2003.09409.X

Cho, N. (2016). Molecular subtypes and imaging phenotypes of breast cancer. In *Ultrasonography* (Vol. 35, Issue 4, pp. 281–288). Korean Society of Ultrasound in Medicine. https://doi.org/10.14366/usg.16030

Cochrane, G., Alako, B., Amid, C., Bower, L., Cerdeño-Tárraga, A., Cleland, I., Gibson, R., Goodgame, N., Jang, M., Kay, S., Leinonen, R., Lin, X., Lopez, R., McWilliam, H., Oisel, A., Pakseresht, N., Pallreddy, S., Park, Y., Plaister, S., … Zalunin, V. (2013). Facing growth in the European Nucleotide Archive. *Nucleic Acids Research*, *41*(D1), D30–D35. https://doi.org/10.1093/NAR/GKS1175

Colomer, R., Aranda-López, I., Albanell, J., García-Caballero, T., Ciruelos, E., López-García, M., Cortés, J., Rojo, F., Martín, M., & Palacios-Calvo, J. (2018). Biomarkers in breast cancer: A consensus statement by the Spanish Society of Medical Oncology and the Spanish Society of Pathology. In *Clinical and Translational Oncology* (Vol. 20, Issue 7, pp. 815–826). Springer-Verlag Italia s.r.l. https://doi.org/10.1007/s12094-017-1800-5

Cui, Y., Whiteman, M. K., Flaws, J. A., Langenberg, P., Tkaczuk, K. H., & Bush, T. L. (2002). Body mass and stage of breast cancer at diagnosis. *International Journal of Cancer*, *98*(2), 279–283. https://doi.org/10.1002/ijc.10209

Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. *IEEE Access*, *8*, 150360–150376. https://doi.org/10.1109/ACCESS.2020.3016715

Gajdos, C., Tartter, P. I., Bleiweiss, I. J., Bodian, C., & Brower, S. T. (2000). Stage 0 to stage III breast cancer in young women. *Journal of the American College of Surgeons*, *190*(5), 523–529. https://doi.org/10.1016/S1072-7515(00)00257-X

Grann, V. R., Troxel, A. B., Zojwalla, N. J., Jacobson, J. S., Hershman, D., & Neugut, A. I. (2005). Hormone receptor status and survival in a population-based cohort of patients with breast carcinoma. *Cancer*, *103*(11), 2241–2251. https://doi.org/10.1002/CNCR.21030

Haq, A. U., Li, J. P., Saboor, A., Khan, J., Wali, S., Ahmad, S., Ali, A., Khan, G. A., & Zhou, W. (2021). Detection of Breast Cancer through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques. *IEEE Access*, *9*, 22090–22105. https://doi.org/10.1109/ACCESS.2021.3055806

Harrison, P. W., Ahamed, A., Aslam, R., Alako, B. T. F., Burgin, J., Buso, N., Courtot, M., Fan, J., Gupta, D., Haseeb, M., Holt, S., Ibrahim, T., Ivanov, E., Jayathilaka, S., Kadhirvelu, V. B., Kumar, M., Lopez, R., Kay, S., Leinonen, R., … Cochrane, G. (2021). The European Nucleotide Archive in 2020. *Nucleic Acids Research*, *49*(D1), D82–D85. https://doi.org/10.1093/NAR/GKAA1028

Hrdlickova, R., Toloue, M., & Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, *8*(1), e1364. https://doi.org/10.1002/WRNA.1364

Iqbal, N., & Iqbal, N. (2014). Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications. *Molecular Biology International*, *2014*, 1–9. https://doi.org/10.1155/2014/852748

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), pdb.top084970. https://doi.org/10.1101/PDB.TOP084970

Li, J., Guan, X., Fan, Z., Ching, L. M., Li, Y., Wang, X., Cao, W. M., & Liu, D. X. (2020). Non-Invasive Biomarkers for Early Detection of Breast Cancer. *Cancers 2020, Vol. 12, Page 2767*, *12*(10), 2767. https://doi.org/10.3390/CANCERS12102767

Łukasiewicz, S., Czeczelewski, M., Forma, A., Baj, J., Sitarz, R., & Stanisławek, A. (2021). Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—An updated review. In *Cancers* (Vol. 13, Issue 17). MDPI. https://doi.org/10.3390/cancers13174287

Ming, C., Viassolo, V., Probst-Hensch, N., Chappuis, P. O., Dinov, I. D., & Katapodi, M. C. (2019). Machine learning techniques for personalized breast cancer risk prediction: Comparison with the BCRAT and BOADICEA models. *Breast Cancer Research*, *21*(1). https://doi.org/10.1186/s13058-019-1158-4

Moskowitz, C. S., Chou, J. F., Wolden, S. L., Bernstein, J. L., Malhotra, J., Friedman, D. N., Mubdi, N. Z., Leisenring, W. M., Stovall, M., Hammond, S., Smith, S. A., Henderson, T. O., Boice, J. D., Hudson, M. M., Diller, L. R., Bhatia, S., Kenney, L. B., Neglia, J. P., Begg, C. B., … Oeffinger, K. C. (2014). Breast cancer after chest radiation therapy for childhood cancer. *Journal of Clinical Oncology*, *32*(21), 2217–2223. https://doi.org/10.1200/JCO.2013.54.4601

Naz, S., Siddiqui, M., Memon, A. I., Bhatti, A. M., Hussain, Z. I., & . I. (2023). Analysis of Breast Cancer Receptors Status and Molecular Subtypes Among Female Population. *Pakistan Journal of Medical and Health Sciences*, *17*(1), 656–658. https://doi.org/10.53350/pjmhs2023171656

Olimjonova, K. A. K., Khusainova, G. O., & Khamzaeva, K. J. (2023). *IDENTIFICATION OF THE PREVALENCE OF BREAST CANCER AMONG DIFFERENT AGE GROUPS OF THE POPULATION AND ITS PREVENTION* (Vol. 4, Issue 3).

Osareh, A., & Shadgar, B. (2010). Machine learning techniques to diagnose breast cancer. *2010 5th International Symposium on Health Informatics and Bioinformatics, HIBIT 2010*, 114–120. https://doi.org/10.1109/HIBIT.2010.5478895

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., & Brazma, A. (2007). ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research*, *35*(suppl_1), D747–D750. https://doi.org/10.1093/NAR/GKL995

Plasilova, M. L., Hayse, B., Killelea, B. K., Horowitz, N. R., Chagpar, A. B., & Lannin, D. R. (2016). Features of triple-negative breast cancer Analysis of 38,813 cases from the national cancer database. *Medicine (United States)*, *95*(35). https://doi.org/10.1097/MD.0000000000004614

Pramod Shardul, B., & Dipak Sonar, A. (n.d.). *The Review on types of Breast Cancer and Associated Risk Factors*. www.ijfmr.com

Prognostic classification of breast ductal carcinoma-in-situ. (n.d.).

Smolarz, B., Zadrożna Nowak, A., & Romanowicz, H. (2022). Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature). In *Cancers* (Vol. 14, Issue 10). MDPI. https://doi.org/10.3390/cancers14102569

Sopik, V., & Narod, S. A. (2018). The relationship between tumour size, nodal status and distant metastases: on the origins of breast cancer. *Breast Cancer Research and Treatment*, *170*(3), 647–656. https://doi.org/10.1007/s10549-018-4796-9

Studham, M. E., Tjärnberg, A., Nordling, T. E. M., Nelander, S., & Sonnhammer, E. L. L. (2014). Functional association networks as priors for gene regulatory network inference. *Bioinformatics*, *30*(12). https://doi.org/10.1093/BIOINFORMATICS/BTU285

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., & Von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, *39*(suppl_1), D561–D568. https://doi.org/10.1093/NAR/GKQ973

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2023). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, *51*(D1), D638–D646. https://doi.org/10.1093/NAR/GKAC1000

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., & Von Mering, C. (2017). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, *45*(D1), D362–D368. https://doi.org/10.1093/NAR/GKW937

Testa, U., Castelli, G., & Pelosi, E. (2020). Breast Cancer: A Molecularly Heterogenous Disease Needing Subtype-Specific Treatments. In *Medical sciences (Basel, Switzerland)* (Vol. 8, Issue 1). NLM (Medline). https://doi.org/10.3390/medsci8010018

Trabert, B., Sherman, M. E., Kannan, N., & Stanczyk, F. Z. (2020). Progesterone and Breast Cancer. *Endocrine Reviews*, *41*(2), 320–344. https://doi.org/10.1210/ENDREV/BNZ001

Wang, A. T., Vachon, C. M., Brandt, K. R., & Ghosh, K. (2014). Breast density and breast cancer risk: A practical review. In *Mayo Clinic Proceedings* (Vol. 89, Issue 4, pp. 548–557). Elsevier Ltd. https://doi.org/10.1016/j.mayocp.2013.12.014

Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, *28*(16), 2184–2185. https://doi.org/10.1093/BIOINFORMATICS/BTS356

Wei, Y., Lai, X., Yu, S., Chen, S., Ma, Y., Zhang, Y., Li, H., Zhu, X., Yao, L., & Zhang, J. (2014). Exosomal miR-221/222 enhances tamoxifen resistance in recipient ER-positive breast cancer cells. *Breast Cancer Research and Treatment*, *147*(2), 423–431. https://doi.org/10.1007/s10549-014-3037-0

Wu, J., & Hicks, C. (2021). Breast Cancer Type Classification Using Machine Learning. *Journal of Personalized Medicine 2021, Vol. 11, Page 61*, *11*(2), 61. https://doi.org/10.3390/JPM11020061

Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. *Designs 2018, Vol. 2, Page 13*, *2*(2), 13. https://doi.org/10.3390/DESIGNS2020013

Zakaria, N. H., Hashad, D., Saied, M. H., Hegazy, N., Elkayal, A., & Tayae, E. (2023). Genetic mutations in HER2-positive breast cancer: possible association with response to trastuzumab therapy. *Human Genomics*, *17*(1), 43. https://doi.org/10.1186/s40246-023-00493-5

Zhang, X. (2023). Molecular Classification of Breast Cancer Relevance and Challenges. *Archives of Pathology and Laboratory Medicine*, *147*(1), 46–51. https://doi.org/10.5858/arpa.2022-0070-RA

Zhu, X., Ying, J., Wang, F., Wang, J., & Yang, H. (2014). Estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 status in invasive breast cancer: a 3,198 cases study at National Cancer Center, China. *Breast Cancer Research and Treatment*, *147*(3), 551–555. https://doi.org/10.1007/s10549-014-3136-y