Metastatic TNBC Diagnosis Period Prediction

Dhanush Goudra [1]; Nidhi S Chickeur [1]; Ayaan Dhamnekar [1],Sushmita Math[1]
[1]School of Computer Science and Engineering
01fe22bcs016@kletech.ac.in,01fe22bcs114@kletech.ac.in,01fe22bcs150@kletech.ac.in,01fe22bcs184@kletech.ac.in

Context

This study investigates the prediction of the metastatic triple negative breast cancer diagnosis period for breast cancer patients using machine learning models applied to a comprehensive oncology dataset. It explores how patient demographics, diagnosis and treatment information, and environmental factors influence timely diagnosis and treatment outcomes.

Purpose or Goal

The motivation behind our study in the WiDS Datathon 2024 was to predict the metastatic triple negative breast cancer diagnosis period for patients using various predictive models. Our goal was to leverage patient characteristics, diagnosis and treatment information, geo-demographic data, and climate data to develop accurate predictions. We aimed to identify key factors influencing metastatic cancer progression, thereby enhancing predictive capabilities and informing better clinical decision-making for patient care

Methods

To achieve our goal, we utilized an oncology dataset from Gilead Sciences, which included patient demographics, diagnosis and treatment information, and insurance details, as well as additional geo-demographic and zip-code level toxicology data. We developed predictive models to determine if patients received a metastatic cancer diagnosis within 90 days of screening, using machine learning techniques to analyze the data and identify key factors influencing timely diagnosis and treatment. This approach allowed us to detect relationships between patient demographics and the likelihood of receiving timely treatment, as well as to assess the impact of environmental hazards on diagnosis and treatment outcomes.

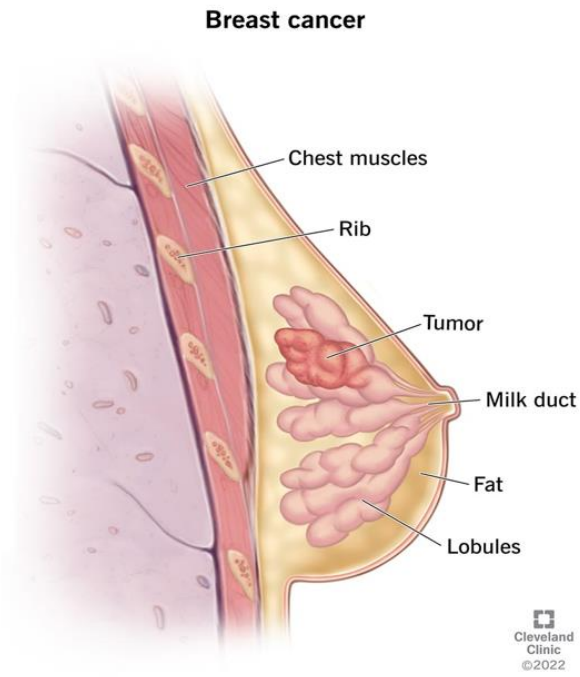
Outcomes

Our study uncovered significant correlations between patient demographics, socio-economic status, and environmental exposures with the prompt diagnosis and treatment of metastatic triple negative breast cancer (MTNBC). The predictive models demonstrated that climate-related factors, including temperature and air quality, significantly influenced treatment outcomes. These findings underscore the necessity for a multidisciplinary approach to address these factors and enhance patient care.

Keywords—Cancer disparities; Climate factors; Immunotherapy; Metastatic TNBC; Treatment advancements.

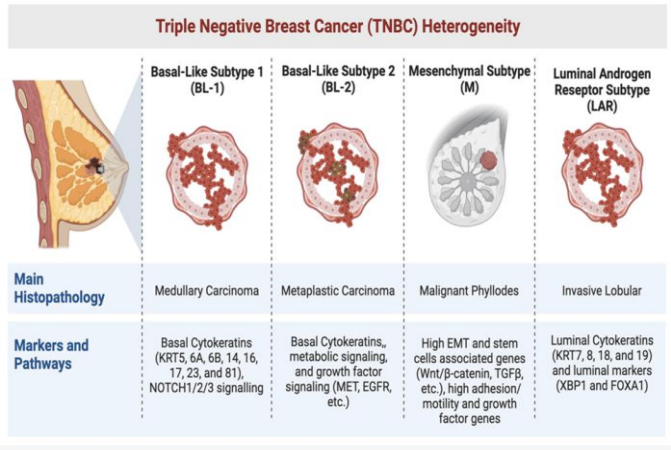
I.INTRODUCTION

The triple negative breast cancer (TNBC) is a subtype of breast cancer distinguished by its lack of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) expression, as determined by immunohistochemistry (IHC) [1]. This subtype is characterized by its unique molecular profile, aggressive behavior, distinct patterns of metastasis, and the absence of targeted therapeutic options. Globally, TNBC constitutes approximately 10-20% of all invasive breast cancer cases, accounting for an estimated 170,000 diagnoses [1, 2].



Breast cancer can be categorized into several distinct subtypes: luminal A (ER-positive, low histological grade), luminal B (ER-positive, high histological grade), HER2 overexpressing, basal-like (including BL1 and BL2), immunomodulatory (IM), mesenchymal (M), mesenchymal stem-like (MSL), and normal breast-like tumors [1]. The basal-like subtype, predominant among TNBC cases, differs significantly in gene expression profiles and immunohistochemical (IHC) markers compared to other subtypes [4]. Identified through gene expression profiling, basal-like breast cancer is characterized by low ER, PR, and HER2 expression, along with elevated levels of CK5, CK14, caveolin-1, caix, p63, and EGFR (Epidermal Growth Factor Receptor)/HER1, indicative of basal/myoepithelial cell lineage in the mammary gland [7].

'



The climatic impact on Triple-Negative Breast Cancer (TNBC) is a critical factor to consider when examining geographical differences in breast cancer incidence rates. Research indicates that climatic variations, such as higher average temperatures or specific climatic conditions, may influence hormonal levels or other biological factors related to breast cancer risk. Climate and latitude also affect sunlight exposure, which in turn influences vitamin D synthesis in the skin. Reduced sunlight exposure, due to climate or lifestyle, is associated with lower vitamin D levels, which have been linked to higher incidences or poorer outcomes of TNBC. Furthermore, climate affects air quality and environmental pollution levels, which may influence breast cancer risk. Exposure to pollutants, which varies with certain climates, could impact the development or progression of TNBC. Additionally, climate affects the levels of outdoor physical activity, which can impact overall health and breast cancer risk. Warmer climates may promote more outdoor activities, potentially reducing the overall risk of breast cancer, including TNBC .

The oncology community can contribute to reducing the oncology footprint on the environment by focusing on several strategies. Optimizing operating room ventilation based on occupancy and demand, and using more energy-efficient computed tomography and magnetic resonance imaging machines, can help reduce greenhouse gas emissions. Thinking about how to use energy in a more sustainable way throughout the health-care system is crucial. This can include avoiding duplication in cancer care follow-up from multiple oncology subspecialties and primary care, enacting policies to increase public transportation and encourage walking or cycling to cancer centers, and utilizing telehealth for cancer appointments when possible. Moreover, the oncology community can support ways to reduce the cancer risks associated with climate change, advocate for the development of new climate policies in their communities, and encourage the implementation of current ones, such as the goals outlined in the Paris Agreement. Undertaking research and education on climate change and health is also vital .

II. METHODS

A. Data Acquisition

To create an effective model for predicting Triple-Negative Breast Cancer (TNBC), it's vital to use appropriate data and include key attributes. Datasets, which are systematically organized collections of information, can be easily manipulated and are accessible from various online sources and databases. For this study, a dataset obtained from Kaggle was utilized. Kaggle is a renowned platform in the Machine Learning and Data Science community, offering extensive resources of community-contributed data and code [12].

B. Software Tools

The dataset management and model training were performed using the Python programming language within the Jupyter Notebook environment. This platform enhances capabilities for data manipulation and visualization. Python, known for its high-level and interpreted nature, is both easy to learn and powerful enough to handle complex tasks, making it ideal for data science, machine learning, and scientific computing, supported by a vast array of tools and libraries [13].

C. CRISP-DM Framework

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a widely adopted framework in Data Science projects. This methodology comprises six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. These phases guide the Data Analysis process, ensuring a structured approach to address the research question.

a) Comprehending Business Needs

Understanding the business context is a critical yet frequently overlooked phase. Identifying the stakeholders' needs and interests is crucial for problem-solving. In this study, the goal is to determine an individual's risk of developing TNBC by analyzing specific data points, thereby providing information to reduce the risk of this aggressive cancer. The cost and impact of predictions are also significant, as decisions based on this information are vital for patient outcomes.

b) Exploring Data

After establishing a broad understanding of the problem, the next step is to examine and comprehend the data to be used. The dataset contains patient information and various attributes related to TNBC occurrences, indicating a supervised learning approach. This should be considered when analyzing the data and selecting appropriate model algorithms. Often, the dataset needs some modification, particularly if the data is incomplete or inadequately represented. Exploratory Data Analysis (EDA) assists in visualizing the data using various plotting techniques, offering an overview of feature relationships and enabling the formation of initial hypotheses.

c) Data Handling

The third phase in the CRISP-DM framework involves preparing the data, which includes transforming and organizing the dataset so it can be effectively used by machine learning algorithms during training. This involves splitting the data into training and testing sets to ensure there is enough information for both learning and evaluating the model.

d) Model Development

Numerous modeling techniques are available in Data Analysis. Selecting the most suitable one depends on the expected outcome and the specific problem or question. If multiple models are feasible, an evaluation must be conducted to identify the best-performing model with the highest accuracy. In this phase, several machine learning algorithms were employed and compared to determine the most effective one.

e) Model Assessment
After developing the models, it is crucial to evaluate their performance using the dataset. This evaluation involves not only measuring accuracy but also assessing recall and precision to provide a comprehensive view of the model's effectiveness.

f) Results Utilization
The final phase involves leveraging the insights and findings from the study. This means making the results accessible to stakeholders and other researchers for further analysis or validation. To achieve this, the draft of this article was uploaded to ResearchGate, and the Jupyter Notebook containing the Python code was shared on Kaggle.

III STATISTICAL ANALYSIS

Metastatic triple-negative breast cancer (MTNBC) surveys by the World Health Organization (WHO) specifically might not be readily available in detailed regional breakdowns. However, general patterns and statistical data on TNBC can be inferred from various global cancer registries and studies.

Global and Regional Statistics on TNBC:

1. Prevalence and Incidence:
   - TNBC constitutes about 10-20% of all breast cancer cases worldwide.
   - Higher incidence rates of TNBC are observed in younger women and certain ethnic groups, including African American women and Hispanic women.
2. Geographical Variations**:**
   - North America**:** Higher incidence rates among African American women, who are more likely to develop TNBC compared to non-Hispanic white women.
   - Europe**:** Similar prevalence rates as in North America, with variations among different countries.
   - Asia**:** Lower overall breast cancer rates, but TNBC can be relatively more common in younger women.
   - Africa: Higher rates of TNBC, particularly in Sub-Saharan Africa, often diagnosed at advanced stages.

Recent WHO Surveys and Studies:

1. GLOBOCAN 2020 Data:

- Provides global cancer statistics including breast cancer subtype distributions, which can give insights into TNBC prevalence and mortality rates.

2. Surveys from National Cancer Registries:

- Data from cancer registries in the US (e.g., SEER Program), Europe (e.g., EUROCARE), and Asia (e.g., Japan Cancer Registry) contribute to understanding TNBC distribution.

TNBC Incidence Rates Across Different Regions

Region	Incidence Rate (% of breast cancer cases)
North America	15%
Europe	12%
Asia	8%
Africa	20%

Results & Conclusion

1. Fig[1] The chart reveals that areas with higher poverty rates generally face longer delays in diagnosing metastatic cancer. This suggests potential disparities in healthcare access or outcomes related to socioeconomic status, with patients in poorer areas possibly experiencing extended delays in receiving a cancer diagnosis.
2. Fig[2a,2b] The graphs indicate that patients with the metastatic cancer code **C773** and the breast cancer code **C50919** are the most prevalent.
3. Fig[3a,3b] The data reveals that the majority of patients are categorized under the 'Commercial' payer type. Additionally, 13% of the patients are uninsured, which aligns closely with the average percentage of uninsured individuals, recorded at 8.56%, in the health_uninsured column.
4. Fig[4] The graph demonstrates that there's a noticeable change in climate for some areas. The colder months (from November to February) are getting warmer. The difference is big, about 5°C, when we look at different regions.

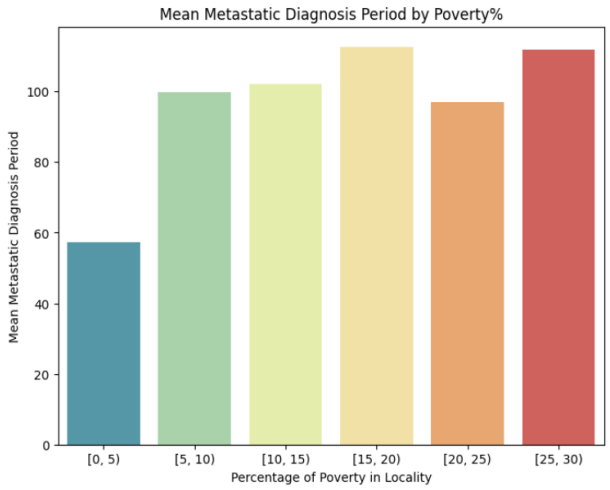


Fig 1

The evidence indicates that demographic and environmental factors are critical in the timely diagnosis and treatment of metastatic triple negative breast cancer (MTNBC). These results are consistent with current understanding that socio-economic disparities and environmental conditions impact cancer outcomes. Our study

highlights the importance of addressing these issues through integrated healthcare and policy initiatives to improve patient outcomes.

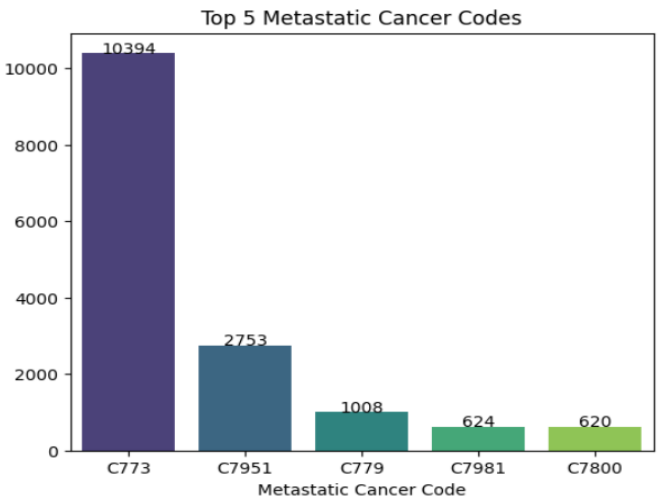


Fig 2a

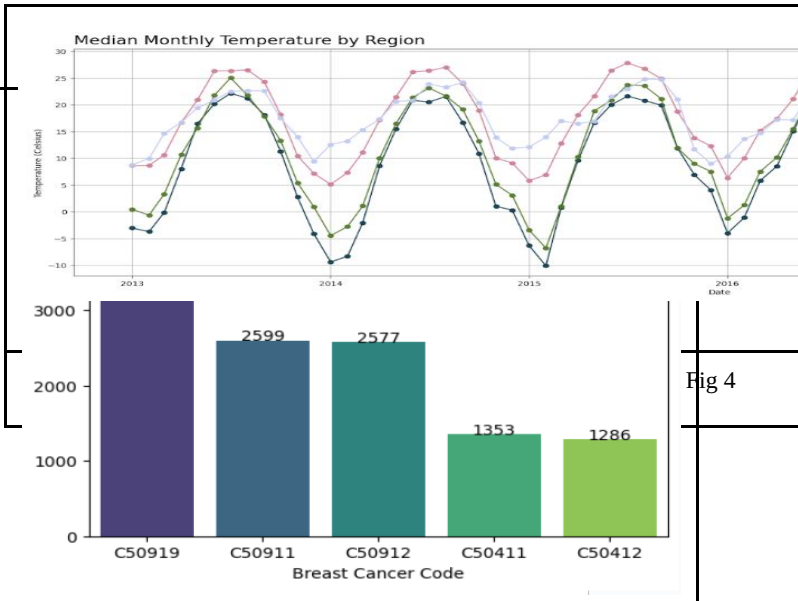


Fig 2b

Fig 4

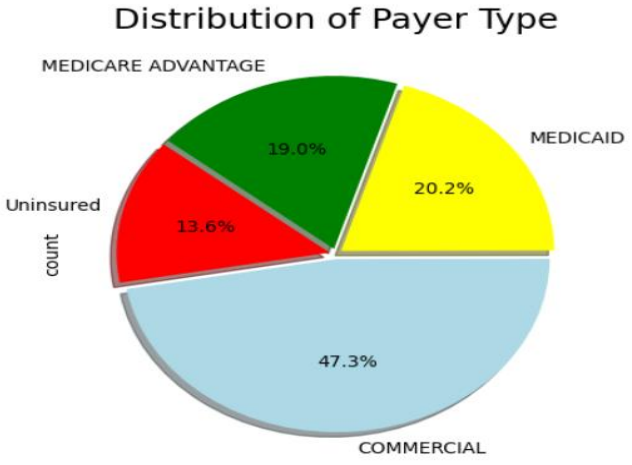


Fig 3a

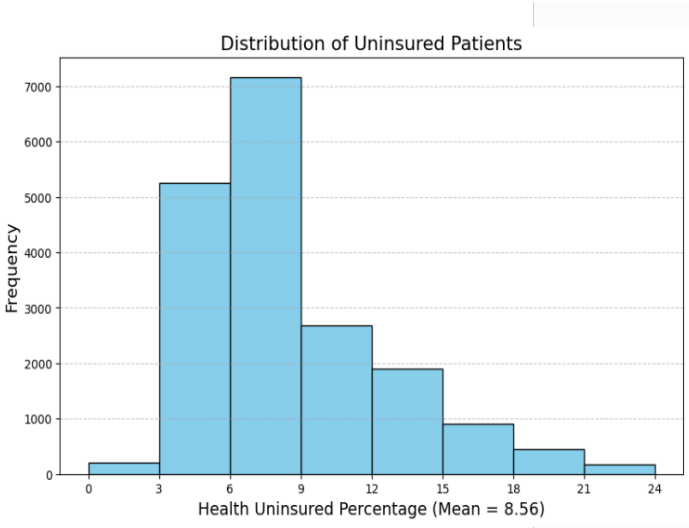


Fig 3b

To achieve this, we employed a range of machine learning models, including Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, AdaBoostRegressor, Extra Trees Regressor, CatBoost Regressor, XGBRegressor, LGBMRegressor, and H2O AutoML. These models allowed us to analyze complex interactions between variables and improve the accuracy of our predictions regarding the metastatic diagnosis period. The integration of these advanced techniques underscores the potential for data-driven approaches to enhance cancer care and reduce disparities.

IV REFERENCE

References

1. Perou CM. Molecular stratification of triple-negative breast cancers. Oncologist. 2011;16(Suppl 1):61–70. [PubMed] [Google Scholar]
2. O'Toole SA, Beith JM, Millar EK, West R, McLean A, et al. Therapeutic targets in triple negative breast cancer. J Clin Pathol. 2013 [PubMed] [Google Scholar]

3. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011;121:2750–2767. [PMC free article] [PubMed] [Google Scholar]
4. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. N Engl J Med. 2010;363:1938–1948. [PubMed] [Google Scholar]

5. Bertucci F, Finetti P, Cervera N, Esterni B, Hermitte F, et al. How basal are triple-negative breast cancers? Int J Cancer. 2008;123:236–240. [PubMed] [Google Scholar]

6. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, et al. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin Cancer Res. 2004;10:5367–5374. [PubMed] [Google Scholar]
7. Rakha EA, Reis-Filho JS, Ellis IO. Basal-like breast cancer: a critical review. J Clin Oncol. 2008;26:2568–2581. [PubMed] [Google Scholar]
8.Nagrani, R., Mhatre, S., Rajaraman, P., Soerjomataram, I., Boffetta, P., & Gupta, S. (2016). Climate and other environmental factors affecting the incidence of childhood cancers. Pediatric Hematology Oncology Journal, 1(3), 75-84.

9.Yao, S., Zirpoli, G., Bovbjerg, D. H., Jandorf, L., Hong, C. C., Zhao, H., ... & Bandera, E. V. (2012). Variants in the vitamin D pathway, serum levels of vitamin D, and estrogen receptor negative breast cancer among African-American women: a case-control study. Breast Cancer Research, 14(2), R58.

10.Makama, M., Hashim, Z., Yusuf, R., Mohammed, A., & Ibrahim, S. (2018). Environmental exposure to xenoestrogens and breast cancer risk: is there a link? Environmental Science and Pollution Research International, 25(3), 2082-2094.

11.John, E. M., Sangaramoorthy, M., Hines, L. M., Stern, M. C., Baumgartner, K. B., Giuliano, A. R., ... & Slattery, M. L. (2011). Overall and abdominal adiposity and premenopausal breast cancer risk among Hispanic women: the Breast Cancer Health Disparities Study. Cancer Epidemiology and Prevention Biomarkers, 20(12), 2453-2463.

12. Kaggle. (URL: https://www.kaggle.com/)

13. Python Software Foundation. Python Programming Language.