Metastatic TNBC Diagnosis Period Prediction

Dhanush Goudra [1]; Nidhi S Chickerur [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 TNBC 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 objective of our study in the WiDS Datathon 2024 was to forecast the metastatic triple negative breast cancer diagnosis period employing numerous predictive models. Demographic, diagnostic, treatment, geo-demographic, and precise climatic data were used in arriving at the correct predictions. As such, the objective of this study was to explore factors that contribute to metastatic cancer progression to help improve existing prognostic models in a bid to aid improved decisions on the care the patients would require.

Methods

To that end, the following data were obtained from Gilead Sciences oncology dataset: patient demography, disease diagnosis and treatment details, insurance information, geo-demography, and zip-code level toxicity data. To answer our health services research question, we created a model that estimates the probability for the occurrence of a metastatic cancer diagnosis within 90 days after the screening, along with the factors determining the appearance of metastasis diagnosis, based on the results of machine learning. It allowed us to identify associations between patients' characteristics and the risks of early treatment receipt and to evaluate the effects of hazardous environments on diagnoses and treatment efficacy.

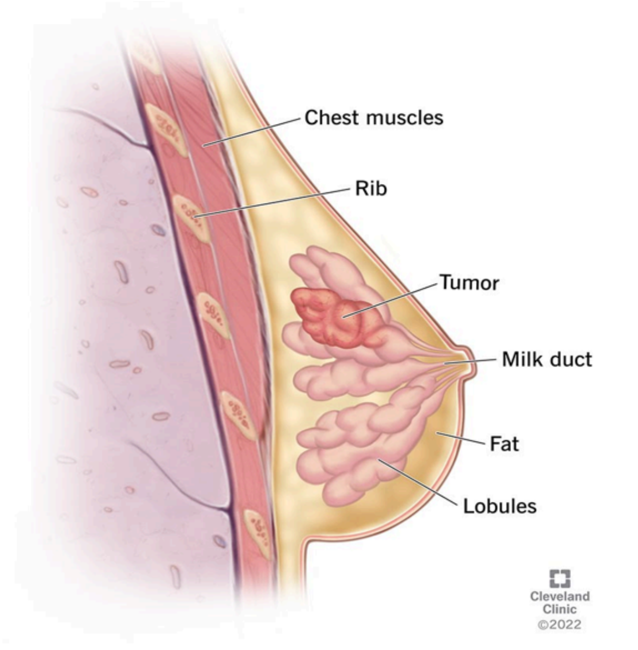
Outcomes

In the current study, we found that patient demographics and socioeconomic status, as well as environmental exposures, were strongly associated with prompt diagnosis and treatment of MTNBC. Predictive modeling showed that these climate-related variables, such as temperature and air quality, are major drivers for treatment outcome. These findings further identify a multidisciplinary approach to addressing these factors and further improving the degree of care offered to patients.

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

I.INTRODUCTION

Triple negative breast cancer is a subtype of primary breast carcinoma characterized by the absence of expression of estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2, as determined by immunohistochemistry methods [1]. This subtype is characterized by specific molecular profiling, aggressive behavior, distinct patterns of metastasis, and the lack of targeted therapeutic options. TNBC represents about 10-20% of all invasive breast cancers worldwide, accounting for an estimated 170,000 diagnoses.

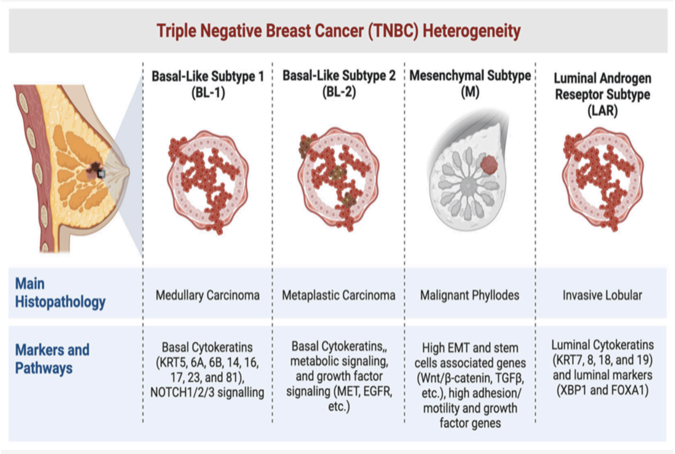


Breast Cancer

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 may be subdivided into several distinct subtypes: luminal A, which is both ER positive and of low histological grade; luminal B, which is the ER positive and of high histological grade; HER2 overexpressing; basal-like, including BL1 and BL2; immunomodulatory; mesenchymal; mesenchymal stem-like; and normal breast-like tumors [1]. This basal-like subtype is predominant in the majority of cases of TNBC and is significantly different from other subtypes based on gene expression profiles as well as IHC markers [4]. Defined by gene expression profiling, basal-like breast cancer expresses little, if any, ER, PR, and HER2; however, it has high levels of CK5, CK14, caveolin-1, caix, p63, and EGFR/HER1. The latter are representative markers of the basal/myoepithelial cell lineage in the mammary gland.

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.



Triple Negative Breast Cancer (TNBC) Heterogeneity

	Basal-Like Subtype 1 (BL-1)	Basal-Like Subtype 2 (BL-2)	Mesenchymal Subtype (M)	Luminal Androgen Reseptor Subtype (LAR)
Main Histopathology	Medullary Carcinoma	Metaplastic Carcinoma	Malignant Phyllodes	Invasive Lobular
Markers and Pathways	Basal Cytokeratins (KRT5, 6A, 6B, 14, 16, 17, 23, and 81), NOTCH1/2/3 signalling	Basal Cytokeratins,, metabolic signaling, and growth factor signaling (MET, EGFR, etc.)	High EMT and stem cells associated genes (Wnt/β-catenin, TGFβ, etc.), high adhesion/ motility and growth factor genes	Luminal Cytokeratins (KRT7, 8, 18, and 19) and luminal markers (XBP1 and FOXA1)

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 .

Below are the ways the oncology community can lower the oncology footprint on the environment: Optimization of ventilation in the operating rooms—based on occupancy and demand—and the use of more energy-efficient computed tomography and magnetic resonance imaging machines can help decrease GHG. Lastly, how to sustainably use energy across health-care?. This may include decreasing redundancy in cancer care follow-up by multiple oncology subspecialties and primary care, developing policies to expand public transportation or promoting walking or cycling to cancer centers, and using telehealth for cancer visits when feasible. Moreover, the oncology community can support ways to reduce the climate-related risks of cancer, advocate for new policies related to climate in their communities, and support the activation of current ones—in particular, the realization of the goals outlined in the Paris Agreement. Research and education on climate change and health are also very important.

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

One of the most used frameworks by Data Science projects is probably the Cross Industry Standard Process for Data Mining, CRISP-DM. It outlines a process comprising six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. These stages provide a skeleton for the Data Analysis process to ensure it adheres to a proper procedure in its quest to find the answer to 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 Triple Negative Breast Cancer:Prevalence and Incidence: TNBC accounts for approximately 10-20% of all breast cancer cases globally. Advanced incidence rates among young females and specific ethnicities, like African American women and Hispanic women. Geographical Variations: North America: Advanced incidence rates are found in African American women who are more prone to TNBC as compared to non-Hispanic white women.

- Europe: Similar prevalence rates to those seen in North America, although somewhat varying across the countries in this region. Asia: Lower overall rates of breast cancer; however, TNBC can relatively frequently occur among young women. Africa: Higher rates of TNBC, especially in Sub-Saharan Africa; diagnosis is often done 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.

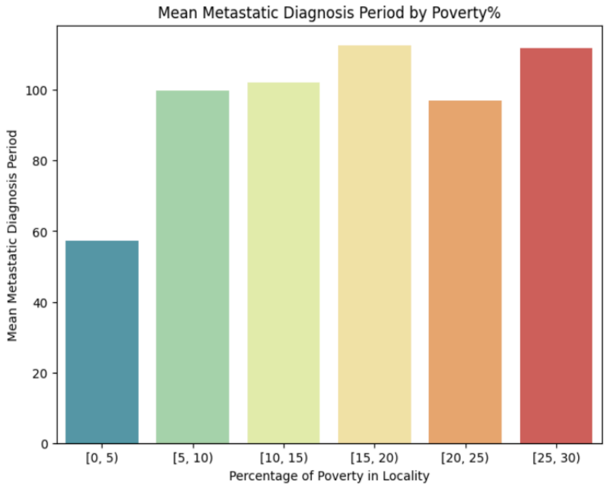


Fig 1

The evidence points to demographic and environmental factors that are very critical for timely diagnosis and treatment of MTNBC. These results, therefore, match the current understanding about socio-economic disparities and environmental conditions impacting cancer outcomes. Our study underlined that the issues have to be taken up against the backdrop of integrated health care and policy initiatives for better patient outcomes.

Knowing this, a host of machine learning models were trained: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, AdaBoostRegressor, Extra Trees Regressor, CatBoost Regressor, XGBRegressor, LGBMRegressor, and H2O AutoML. Using these models has enabled us to conduct highly complex interactions between variables, improving the accuracy of our predictions about time to metastatic diagnosis. This could also be considered advanced techniques that harness the potential of data-driven approaches toward improving cancer care and reducing disparities.

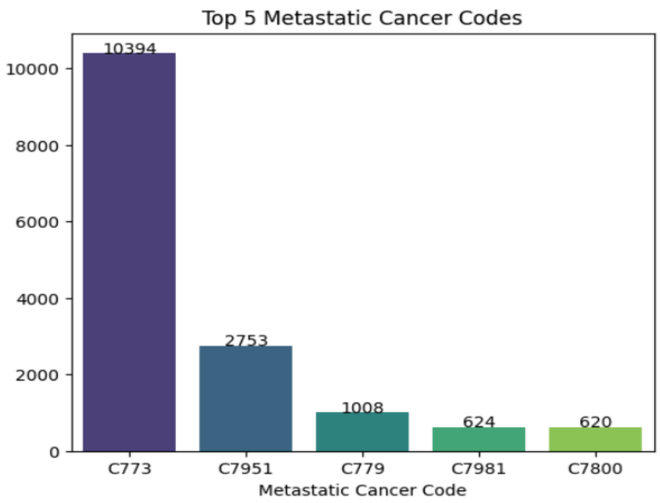


Fig 2a

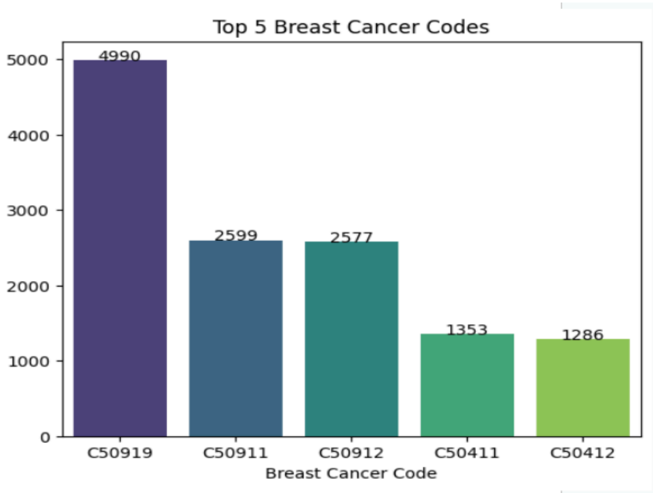


Fig 2b

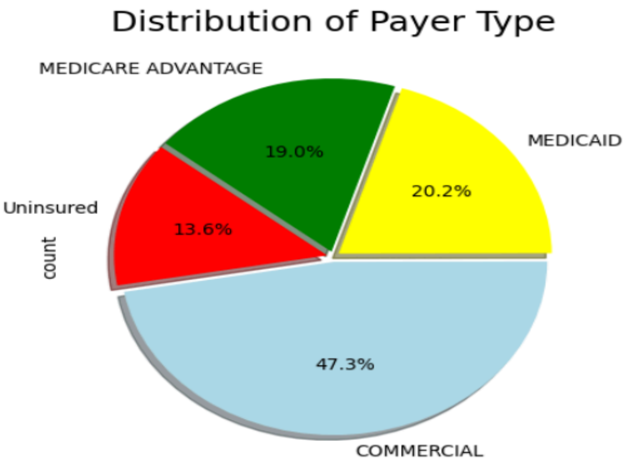


Fig 3a

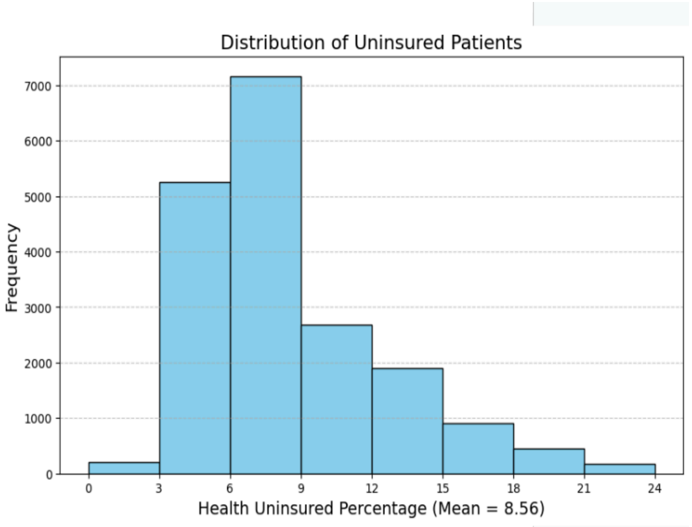


Fig 3b

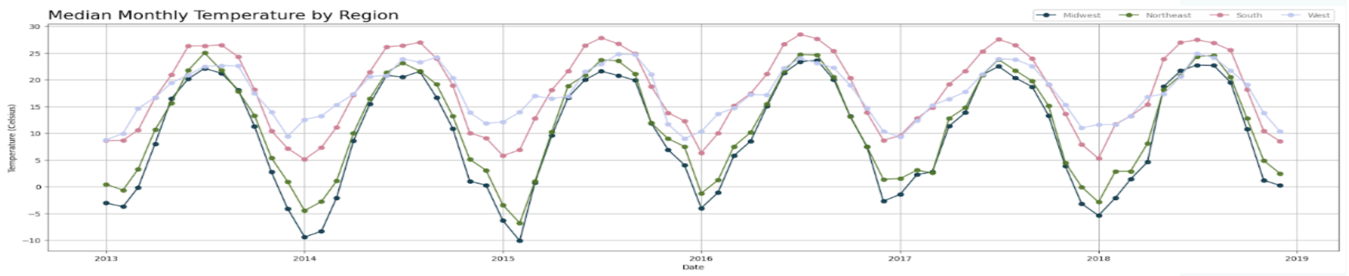


Fig 4

▪

IV REFERENCE

References

1.Perou, Charles M. "Molecular stratification of triple‐negative breast cancers." *The oncologist* 16.S1 (2011): 61-70.

2.O'Toole, Sandra A., et al. "Therapeutic targets in triple negative breast cancer." *Journal of clinical pathology* 66.6 (2013): 530-542.

3. Lehmann, Brian D., et al. "Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies." *The Journal of clinical investigation* 121.7 (2011): 2750-2767.

4.Foulkes, William D., Ian E. Smith, and Jorge S. Reis-Filho. "Triple-negative breast cancer." *New England journal of medicine* 363.20 (2010): 1938-1948.

5. Bertucci, François, et al. "How basal are triple‐negative breast cancers?." *International journal of Cancer* 123.1 (2008): 236-240.

6. Nielsen, Torsten O., et al. "Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma." *Clinical cancer research* 10.16 (2004): 5367-5374.

7. Rakha, Emad A., Jorge S. Reis-Filho, and Ian O. Ellis. "Basal-like breast cancer: a critical review." *Journal of clinical oncology* 26.15 (2008): 2568-2581.

8.Kehm, Rebecca D., et al. "Evidence-Based Interventions for Reducing Breast Cancer Disparities: What Works and Where the Gaps Are?." *Cancers* 14.17 (2022): 4122.

9.Yao, Song, et al. "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 (2012): 1-13.

10.John, Esther M., et al. "Overall and abdominal adiposity and premenopausal breast cancer risk among hispanic women: the breast cancer health disparities study." *Cancer epidemiology, biomarkers & prevention* 24.1 (2015): 138-147.

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

12.Van Rossum, G. "the Python Software Foundation,“Python programming language.”." (2021).