EARLY DETECTION OF BREAST CANCER USING NAN0-PARTICLE SENSOR URINE ANALYSIS DATABASE WITH TUMOR CLASSIFICATION USING MACHINE LEARNING MODELS

Charishma Karri, Naga Srihitha Vatsavayi, Hrashitha Lekkala

Computer Science Engineering Department, Gandhi Institute of Technology and Management, Visakhapatnam, India.

Assistant Prof. Computer Science Engineering., Dept. D NVR. Swaroop Kumar Mane Group of Institutes, Kolhapur, India,416112---

**ABSTRACT:** This research is mainly focused on the early detection of breast cancer in women by their urine samples with nanoparticle sensors. It detects certain enzymes and proteins that can be the main cause of cancer by machine learning algorithms, correlation analysis, and logical regression methods. We trained a model to identify the stages of cancer using a nanoparticle sensor database set which we created and gives predictions of which stages the person is, and we developed a website where by entering the values of the urine sample and sensor data and biomarker value it predicts This study mainly focuses on non-invasive cancer. We analyze non-invasive cancer by urine sample test data collected from the patients by utilizing advanced technology to access DNA signatures associated with breast cancer biomarkers. Our approach involves the barcode of nanoparticle sensors and matches with the urine samples. After the urine samples match, they are applied to the sensors, translated into digital data, and transmitted to a centralized system. Now, the centralized system collaborates with datasets derived from previous breast cancer cases. The algorithm now analyzes the urine data and identifies the patterns. Then, it correlates with different stages of breast cancer.

**KEYWORDS:** breast cancer, nanoparticle sensor, machine learning, SQL, urine test

**INTRODUCTION:** Detection of breast cancer in the early stages is still an issue worldwide. Out of many cancers, breast cancer is the 2nd top all over the world. Breast cancer is caused by unhealthy cells that change the shape or color of the breast. This change is called a tumor, and if this tumor is increased and not treated. Then it could called it has breast cancer. There are two types of breast cancer: invasive and non-invasive breast cancer. Most of the patients are affected by non-invasive breast cancer. There are four stages, each defining how much the breast is affected. Besides these types, there are sub-types of breast cancer. They are Hormone receptor-positive or Hormone receptor-negative; Ancient Egyptians first detected HER2-positive or HER2-negative triple-negative breast cancer. In 460 B.C., A Breast Lump is caused when the tumor (unhealthy cells/tissues) spreads and forms like a clot. Even the Breast becomes hardened and thicker around the nipple. It is caused by Paget's disease of the breast, known for changing the breast shape and the color of the nipple. It can also be altered after or before swelling. After the change of its shape and color, the outward nipple turns into an inward nipple. The skin around the breast changes color into red or orange. There's even a possibility of blood discharge from the outward/inward nipple. This is mainly caused by reproductive history, which means the person can get breast cancer more than once. There's even a high possibility of women getting breast cancer because of Genetics or Family History. Women who consume alcohol or smoke can develop breast cancer. A breast cancer patient should quit smoking and drinking. Some women who have dense breasts might end up getting breast cancer, too. To remove this rampant breast cancer, there's a therapy called breast removal, where the surgeon removes both breasts. There are multiple therapies and surgeries such as Lumpectomy, Mastectomy, Sentinel node biopsy, Axillary lymph node dissection, Radiation therapy, Chemotherapy Immunotherapy, and much more to remove breast cancer.

Table 1: Symptoms, Causes, Treatment, Durability

|  |  |  |  |
| --- | --- | --- | --- |
| Symptoms | Causes | Treatment | Durability |
| A Breast Lump | Fibroadenomas | Excisional biopsy | 2 Weeks |
| Uneven skin | Inflammatory breast cancer | Chemotherapy | 3 Months |
| Breast Swelling | Mastitis | Antibiotics/drugs | 5 Days to 3 Months |
| Flaking of pigmented skin | Eczema | Emollients | 3 to 4 Weeks |
| Pain in/around the nipple | Hormonal fluctuations | Antithyroid drugs | Depends upon surgery |

In the above table, you can witness some of the other symptoms and their treatments and durability of breast cancer symptoms leads to early detection of cancer, which can be identified easily.

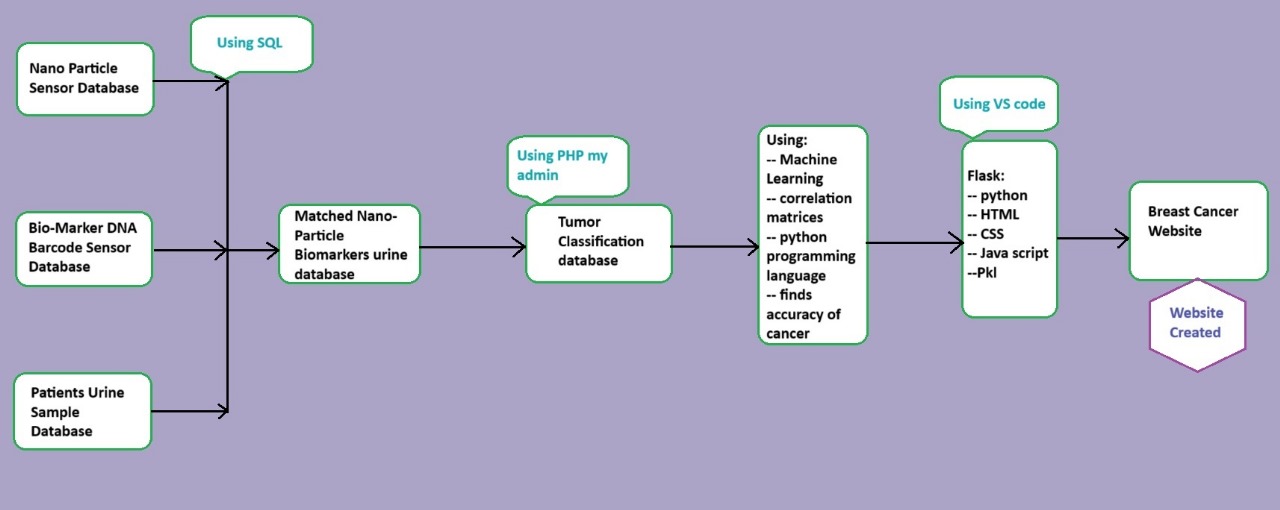
Graph 1: Breast cancer analysis in INDIA

The above bar graph represents the statistics of breast cancer patients from the past 25 years which are in blue color and red in present scenario where at the age of 50-60 years we can witness more cases where more hormonal changes occur in the women due to menopause which comes with a lot of hormonal changes so early detection can be a huge advantage

**LITERATURE REVIEW:** Our area of research mainly focuses on early detection, which should be simple, like a COVID-19 rapid test, which we studied apart from the quick test. ]Shoko Kure 1, Shinya Iida 2,3,4[1} researched by using volatile organic compounds(VOCs), we can train the sniffer dogs to detect breast cancer. The trained dogs acknowledge the potential of VOC analysis and conduct the research without using ML algorithms/devices. GCMS stands for gas chromatography-mass spectrometry, used when the dogs start to sniff, and VOC analyzes the potential. It is an entirely new approach to detecting breast cancer. Rapid tests even specify the cancer stage. In their researched way, it takes time and has potential only to certain limits. The conclusion is that dogs can only predict whether breast cancer is alive. whereas rapid tests even specify the cancer stage. Jee Yee Kim[2] did research and found out that the detection of breast cancer can be done by using a blood test. The research aims to identify RNA results where they compare the blood of breast cancer patients and non-breast cancer patients by using developed machine learning techniques. The study predicts the values like mammograms. Then, the values are entered into the machine. The aim of the research is created by using prediction and detection tests, but the results might turn into poor outcomes. In rapid tests, there needs to be more usage of developed tests and better outcomes. Only one outcome is breast cancer with the stage in the following research paper; Julia Beretov and Yong Li[3] used urine to detect breast cancer. The urine sample is monitored as potential for non-invasive breast cancer. The evaluation of urine protein specifies breast cancer tissues and tests the urine sample; first, it is kept under surveillance and prevention for early diagnosis. This research paper is similar to the rapid test. The only difference between those two is the nano biosensor. This nano biosensor value matches the value of urine protein.[1] did not use this nano biosensor concept; thus, it makes the detection of early breast cancer take time to detect cancer. Compared to all the above, early detection can be possible by nanoparticles nearby, which are used mainly in rapid tests by comparing these research using nanoparticles undergoing (MIT)[4]. Some students from MIT have researched and shown that they could use the sensors to detect cancer tumors by five different enzymes, but creating a nanoparticle sensor in real-time takes so instead of that, we worked on a database that contains the biomarker values and DNA barcode values when the urine sample it predicts cancer as that database and symptoms and structure of tumor database is merged and provided to machine learning model using logical regression it shows accuracy and using flask we created a website where you enter the values of urine\_sample, biomarker value, and sensor\_value in specified range it predicts the person has cancer or not these make everyone's work simple as technology in our hand we provided a model made by machine learning to the website it gives 95%accuracy in every case which you can witness in our results.

| **Researcher(s)** | **Method** | **Outcomes** | **Limitations** | **Comparison to Nanoparticle Sensor Approach** |
| --- | --- | --- | --- | --- |
| Shoko Kure et al. | Training sniffer dogs using VOCs | Dogs can detect the presence of cancer | Time-consuming, not machine learning-based, limited to detection without staging | Proposed method is quicker, stages cancer, and uses machine learning for analysis |
| Jee Yee Kim | Blood tests to identify RNA | Predictive values similar to mammograms | Possibility of poor outcomes, does not specify cancer stage | Proposed method specifies cancer stage and has higher accuracy |
| Julia Beretov et al. | Urine sample analysis for proteins | Monitors potential for non-invasive breast cancer | Requires surveillance and prevention for early diagnosis, no rapid test | Proposed method is similar but enhanced with nanoparticle biosensors for quicker detection |
| MIT students | Sensors to detect cancer enzymes | Show potential for tumor detection | Creating sensors in real-time is challenging | Proposed method uses a ready database for immediate prediction without the time-consuming sensor creation |

**Proposed model:** the model on which we worked is represented in a systematic diagram



In the following diagram, we developed a database containing proteins' biomarkers and DNA bar code values that match the cancer patients in the urine. These are created by MySQL language and PHP by admin. We got biomarker values from the Cancer Genome Atlas (TCGA)[5] European Bioinformatics Institute (EBI)[6] National Cancer Institute (NCI)[7], which helped us get real-time data and patient urine sample data collected from clinical doctors and ClinicalTrials.gov:[7]these two databases are merged into one and provided to symptoms database set which we got from Kaggle[8]that is an open source database when patients proteins values are matched with the biomarker value of the protein we can identify there is cancer in that patient for further classification we developed a machine learning model which provides and gives logical accuracy based on data provided and apply analytical regression model and shows a correlation matrics which can help to predict to know manually we developed a website which contains real-time data and at the bac ed our machine learning model is helping when the user is entering the data manually it provides the prediction based on specified values

**methodology**

***Algorithm*** developed a database of urine samples and a nanoparticle sensor database, which contains the attributes of

* ***Patient Table****:* Patient ID FirstName, LastName, DOB (Date of Birth), Other relevant patient information,
* ***DNA Barcode Table***: Barcode ID (Primary Key), Patient ID (Foreign Key referencing Patient Table), Barcode Sequence (The actual DNA barcode sequence), Collection Date, Other relevant DNA barcode information
* ***Nano Particle Sensor data Table:*** Sensor Data ID (Primary Key)Patient ID (Foreign Key referencing Patient Table)Sensor Reading (Data from nanoparticle sensors) Reading DateOther relevant sensor data information
* ***Urine Sample Table*:** Urine Sample ID (Primary Key)Patient ID (Foreign Key referencing Patient Table)Sample Analysis Result (Results from urine sample analysis)Sample Collection Date
* Merge the above data set with the locally available database; it gives the symptoms, tumor structure, and tumor classification. Using SQL statements, we got a unique data set which helped in the classification of stages using a database set with a machine learning model
* ***Machine learning model using Python: first, we will import the necessary*** libraries to simplify code. After that, we classify the databases as containing people with cancer and average persons. We remove unwanted rows and columns, get exact data, and remove null values
* We create a pair polt in which the same data row is matched with another variable value and print the new databases
* To get the correlation analysis, we use function df. corr(). After implementing the function, you can get the heat map of correlation where we can explore clear data for training the machine learning model
* We will split the data set into the variables X (target data)and Y (dependent data)
* Again, we will split the data into 75% training and 25% testing
* And Scale the data and bring all features into the same level of magnitude
* Create a function that holds many different values to get the accuracy from all the training tests and give precise accuracy and metrics values
* Using pkl, we imported all the accuracy values into the doc. pkl and given to Python, we created a script using (HTML, CSS, and JAVASCRIPT)
* And set certain threshold values for urine\_value ,biomarker\_value, sensor\_value to predict cancer using a easily accessible websitem

Algorithm BreastCancerDetectionSystem

Input: None

Output: Web-based Prediction System

// Initialize databases

1: Initialize four tables: Patient, DNA\_Barcode, Nano\_Sensor, and Urine\_Sample with relevant attributes.

// Merge datasets

2: Merge the four tables on Patient\_ID to form a comprehensive dataset D.

3: Integrate D with a local database containing additional attributes such as symptoms, tumor structure, and classification.

4: Clean D by removing unwanted attributes and null values to form a cleaned dataset D'.

// Exploratory Data Analysis

5: Generate pair plots and a correlation heatmap from D' to identify relevant features.

// Data preparation

6: Split D' into features X and target Y based on the defined outcome variable.

7: Split (X, Y) into training and testing sets using a 75:25 ratio.

8: Normalize X to have a uniform scale across all features.

// Model training

9: Train a machine learning model using the training set (X\_train, Y\_train).

10: Evaluate the model's performance on the testing set (X\_test, Y\_test) to obtain accuracy and other metrics.

// Model deployment

11: Serialize the trained model into a .pkl file for deployment.

12: Create a web interface using HTML, CSS, and JavaScript that takes user input for urine\_value, biomarker\_value, and sensor\_value.

13: Set thresholds for these values to enable the prediction of cancer presence.

14: Deploy the web interface linked to the machine learning model, allowing users to receive predictions based on their input.

return Web-based Prediction System with the trained model

End Algorithm

**Results & Discussion:**

ADABOOST CLASSIFIER REPORT:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| accuracy |  |  | 0.95 | 144 |
| B | 0.99 | 0.95 | 0.96 | 81 |
| M | 0.86 | 0.94 | 0.91 | 33 |
| Macro avg | 0.93 | 0.95 | 0.94 | 114 |
| Weighted avg | 0.95 | 0.95 | 0.95 | 114 |

**GRADIENT CLASSIFIER REPORT:**

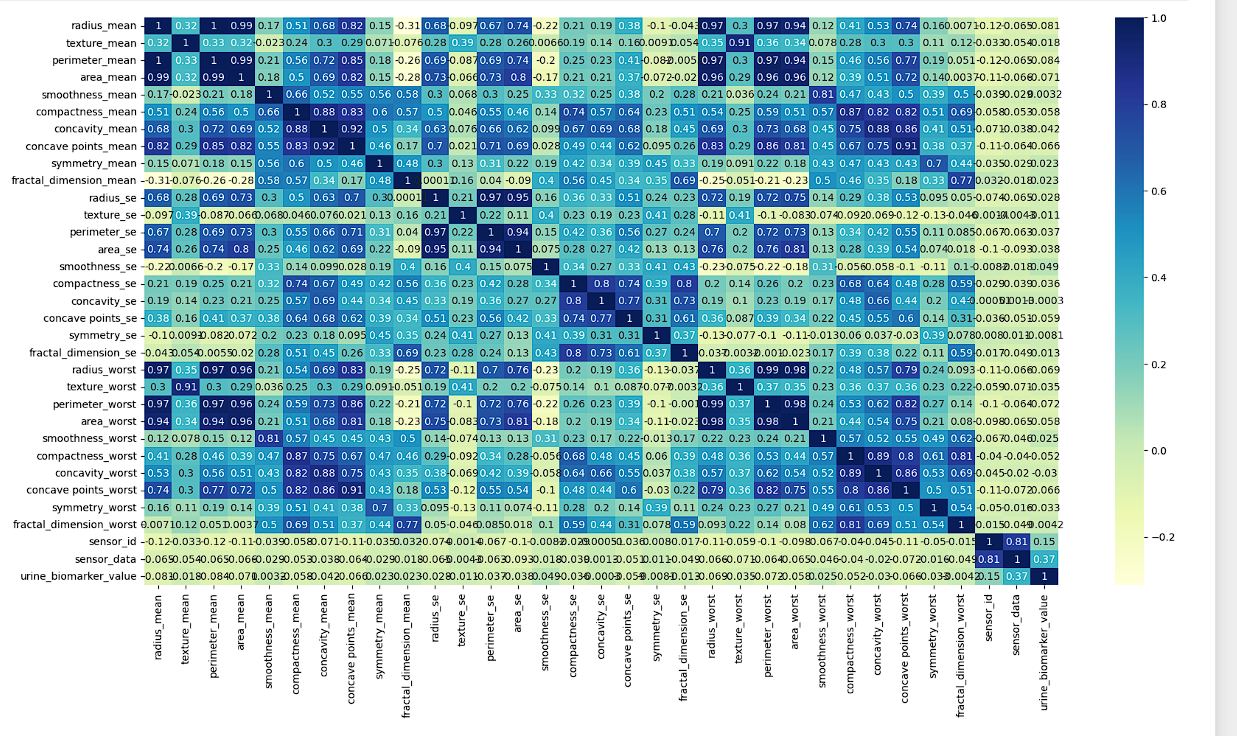
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| accuracy |  |  | 0.95 | 114 |
| B | 0.97 | 0.97 | 0.96 | 81 |
| M | 0.89 | 0.89 | 0.91 | 33 |
| Macro avg | 0.93 | 0.93 | 0.94 | 114 |
| Weighted avg | 0.95 | 0.95 | 0.95 | 114 |

**MLP FINAL REPORT:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| accuracy |  |  | 0.95 | 144 |
| B | 0.99 | 0.95 | 0.96 | 81 |
| M | 0.86 | 0.94 | 0.91 | 33 |
| Macro avg | 0.93 | 0.95 | 0.94 | 114 |
| Weighted avg | 0.95 | 0.95 | 0.95 | 114 |

Where these classifications report precise cut-off analysis of the detection of breast cancer in the early stages

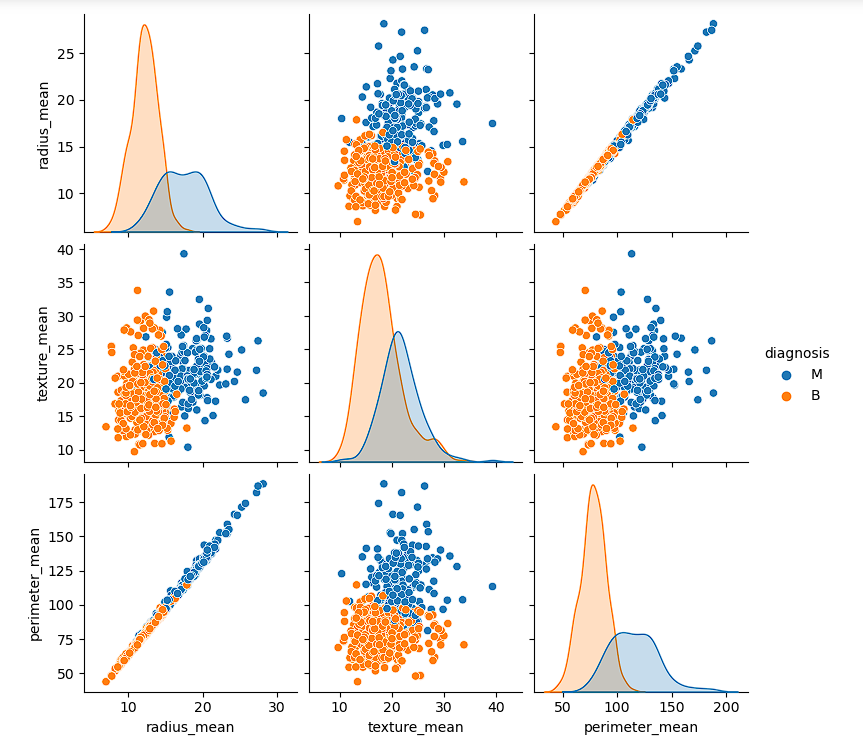
* The accuracy report and the final classification report value **(0.956140350877193)**



Graph 2: Breast cancer patients and normal patients diagnosis report

The above bar graph is about diagnosis report and are related to the correlation matrics and the tumor classification of the Breast cancer patients and the Normal patients in the specific ( radius\_mean, texture\_mean, perimeter\_mean) in the X-axis you can witness all these aspects and in the Y-axis we can witness the range where blue represents a normal person and orange we witness the breast cancer patient.

Graph 2.1:



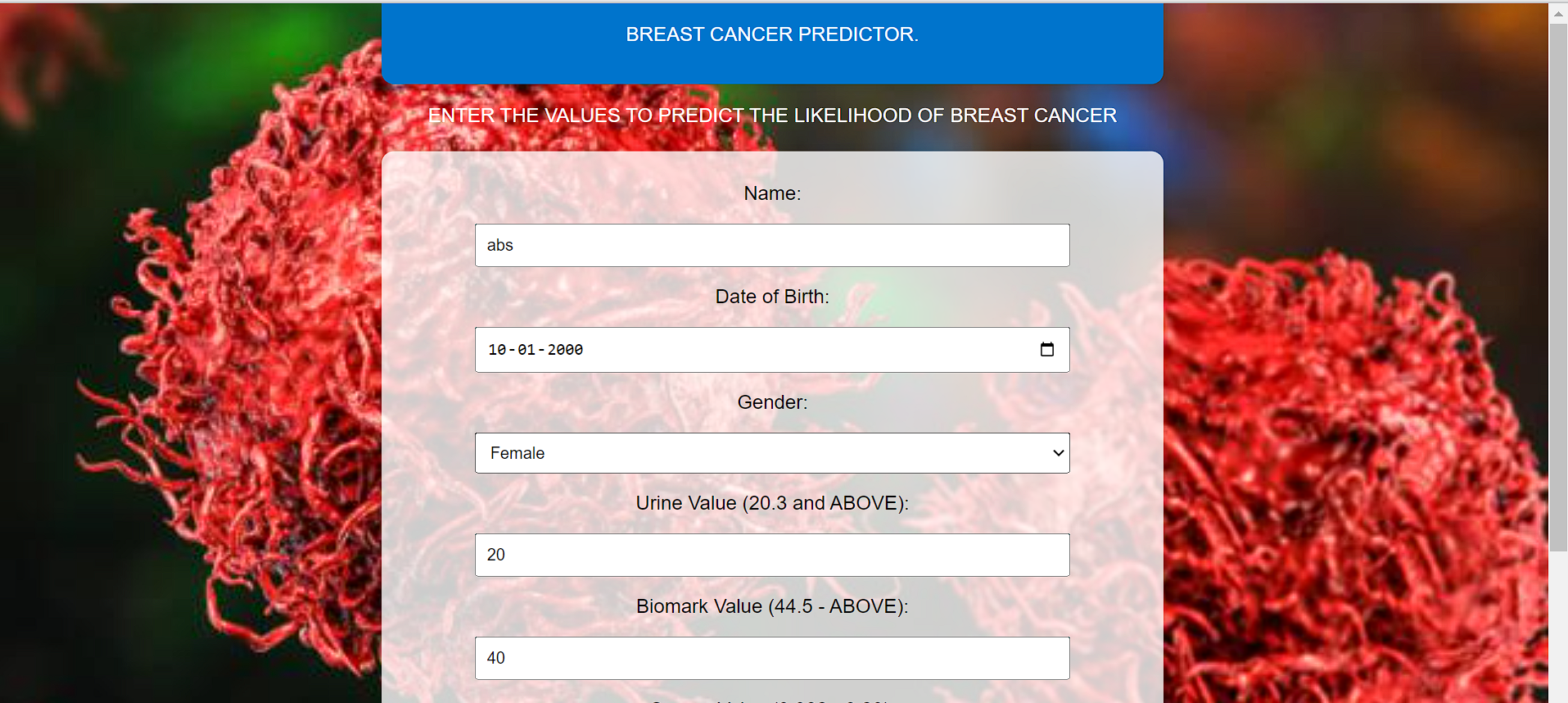
**USER ACCESSIBLE WEBSITE**

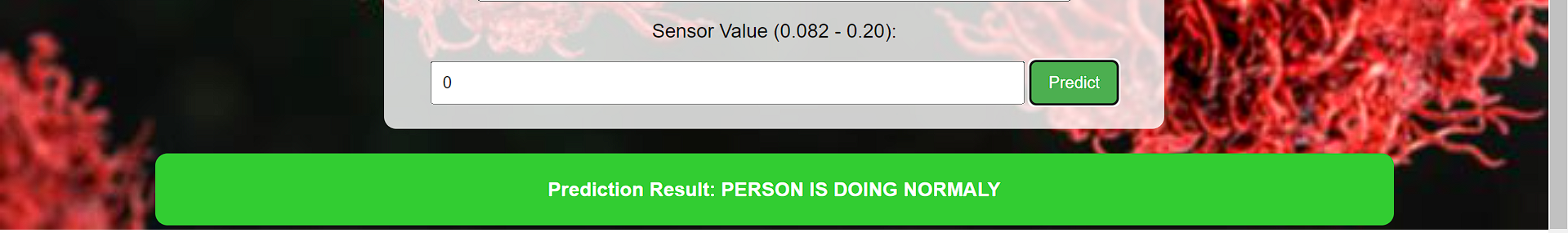
The website below, which is made using machine learning model prediction, is user-friendly. Where you need to add the urine\_value that you got during the test and analysis sensor\_value and biomarker\_value for some reasons we specified basic details of the patients (name, date of birth, gender)

Then, it predicts the cancer based on the threshold value

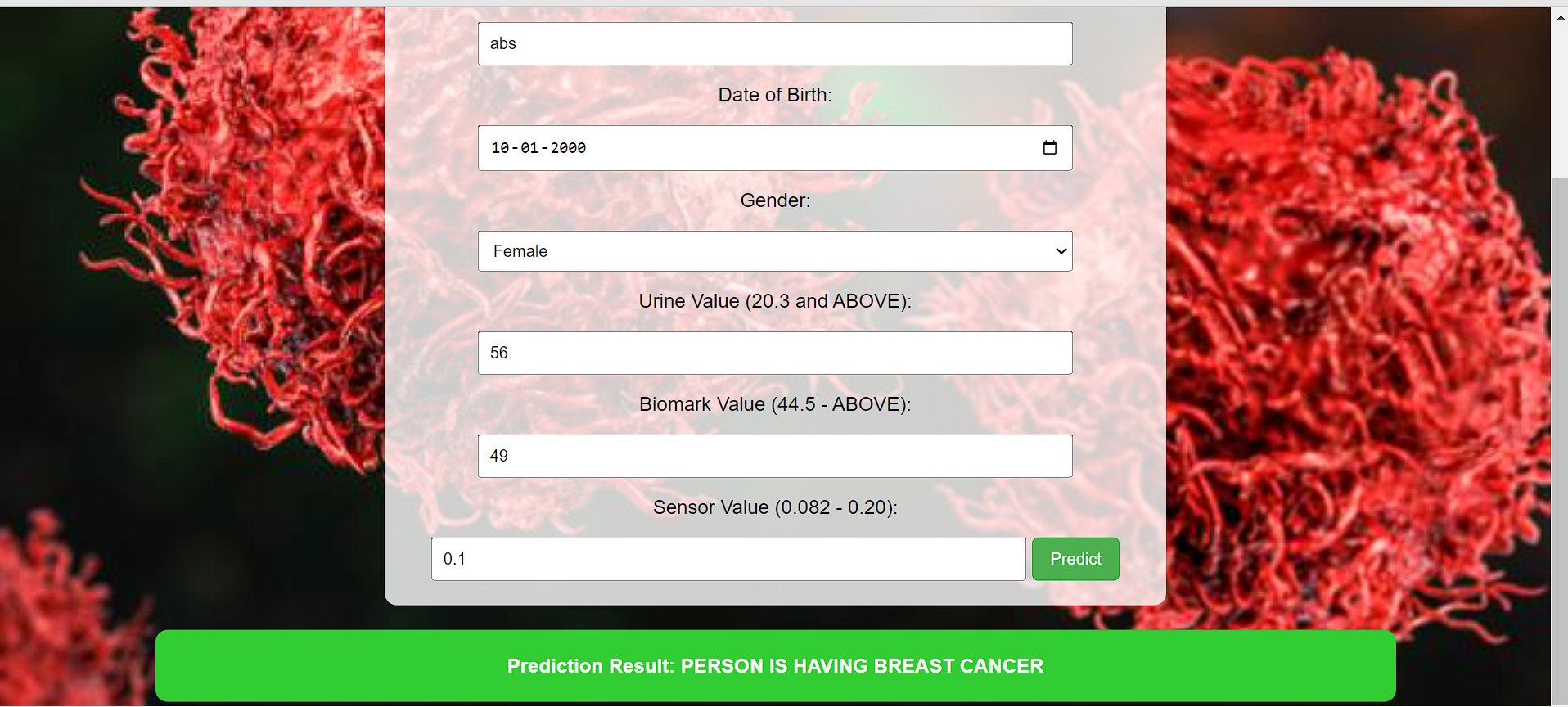
**Visualization of the website**

NORMAL REPORT





**BREAST CANCER REPORT**

****

**CONCLUSION:**

Based on the above research and results, breast cancer can be predicted using a nanoparticle sensor that has DNA bar code values and bio \_markers that indicate cancer. That can also be done by the database that we produced in the methodology. Future research can also be developed using nano.-particle sensor, which can make an instant kit that needs only urine samples to detect cancer, and with machine learning, we can say accuracy precisely, and we created a user-friendly platform that can be accessible by everyone, which makes their work simpler and we can mainly focus on artificial intelligence model which can analysis the image correctly after the tests it can identify minor lumps and tumors which creates which can help in early detection

**REFERENCES:**

[1] Giró Benet, J., Seo, M., Khine, M., Gumá Padró, J., Pardo Martnez, A., & Kurdahi, F.J. (2022). Breast cancer detection by analyzing the volatile organic compound (VOC) signature in human urine. Scientific Reports, 12.

[2] Park, S., Ahn, S., Kim, J.Y., Kim, J., Han, H., Hwang, D., Park, J., Park, H.S., Park, S., Kim, G.M., Sohn, J.H., Jeong, J., Song, Y.U., Lee, H., & Kim, S. (2022). Blood Test for Breast Cancer Screening through the Detection of Tumor-Associated Circulating Transcripts. International Journal of Molecular Sciences, 23.

[3] Kure, S., Iida, S., Yamada, M., Takei, H., Yamashita, N., Sato, Y., & Miyashita, M. (2021). Breast Cancer Detection from a Urine Sample by Dog Sniffing: A Preliminary Study for the Development of a New Screening Device, and a Literature Review. Biology, 10.

[4] Beretov, J., Wasinger, V.C., Millar, E.K., Schwartz, P., Graham, P.H., & Li, Y. (2015). Proteomic Analysis of Urine to Identify Breast Cancer Biomarker Candidates Using a Label-Free LC-MS/MS Approach. PLoS ONE, 10.

[5] Rovshenov, A., & Peker, S. (2022). Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using Wisconsin Breast Cancer Dataset. 2022 3rd International Informatics and Software Engineering Conference (IISEC), 1-6.

[6] Elouerghi, A., Bellarbi, L., Afyf, A., & Talbi, T.F. (2020). A Novel Approach for Early Breast Cancer Detection based on Embedded Micro-BioHeat Ultrasensitive Sensors: IoT Technology. 2020 International Conference on Electrical and Information Technologies (ICEIT), 1-4.

[7] Paci, E. (2002). Mammography and beyond: developing technologies for the early detection of breast cancer. Breast Cancer Research, 4, 123 - 125.

[8] Amethiya, Y., Pipariya, P., Patel, S., & Shah, M. (2021). Comparative Analysis of Breast Cancer detection using Machine Learning and Biosensors. Intelligent Medicine.