UMassAmherst | Data Analytics and Computational Social Science Program

Devising a solution to the problems of Cancer awareness in Telangana, India

Authors: Priyanka Avhad, Vedanti Kshirsagar, Mahek Nakhua, and Urvi Ranjan
VJTI, Mumbai, India

TSIC ఇంటింటా ఇన్నోవేటర్

ABSTRACT

According to the findings, the percent of women who underwent screening for cervical cancer, breast, and oral cancer in Telangana in the year 2019-2020 was 3.3 percent, 0.3 percent, and 2.3 percent respectively. Although early detection is the only way to reduce morbidity and mortality, people have very low awareness about cervical and breast cancer signs and symptoms and screening practices. We developed an ML classification model to predict if a person is susceptible to breast or cervical cancer based on demographic factors. We devised a system to provide suggestions for the nearest hospital or Cancer treatment centres based on the user's location or address. In addition to this, we can integrate the health card to maintain medical records of all individuals and conduct awareness drives and campaigns. For ML classification models, we used decision tree classification and support vector classification algorithms for cervical cancer susceptibility and breast cancer susceptibility respectively. Thus, by devising this solution we come one step closer to our goal which is spreading cancer awareness, thereby, decreasing cancer mortality and increasing cancer literacy among the people of Telangana.

METHODOLOGY

Data Collection

The NFHS dataset was studied to understand the current statistics for: Screening for Cancer among Women (30-49 years) for cervical cancer, breast cancer and oral cancer. The report Profile of Cancer and Related Factors-Telangana, 2021 by IMCR and NCDIR was studied as well.
The cervical cancer training dataset was obtained from Kaggle, gathered at the 'Hospital Universitario de Caracas' in Caracas, Venezuela and contains 858 examples. The breast cancer training dataset was obtained from BCSC Research, originally collected by the Breast Cancer Surveillance Consortium in 1996-2002, and consists of 4,62,563 examples. Both datasets consist of demographic data, habits and relevant medical records of patients.

Surveying Previous Projects

A free cancer screening and training camp was organized at the Appolo Hospital. 60+ women were trained in procedures like self-breast examination. According to Dr. Nalini, the program aimed at spreading the message across the state that cancer is a curable disease if detected early.
Dr. Charanjith Veeramalla said breast and cervical cancer were most common among Indian women and the number is on the rise. The DLSA of Warangal and Omega Bannu Hospitals organized a free cancer screening test camp in Aug 2021. Tests like Mammography for Breast Cancer, Pap smear for Cervical Cancer, Anaemia Screening by Blood Test and ECG, were conducted at the camp for a total of 80 women.

Studying Current Statistics

According to a report "Profile of Cancer and Related Factors- Telangana, 2021", the five leading sites of cancers in females are Breast, Cervix Uteri, Ovary, Corpus Uteri, Lung. In the age group of 0 to 74 years, the cumulative risk of developing cancer in females is 1 in every 7. The projected incidence of cancer cases in Telangana for the years 2020 and 2025 was calculated and found to be 25434 cases and 28708 cases for females respectively.
The NFHS dataset was examined for the indicators: Screening for Cancer among Women (30-49 years) for cervical cancer, breast cancer, oral cancer. This was then classified into two categories- Residence in rural and urban areas, District-wise and visualization was performed for both.

Understanding Importance of Analysis

Analyzing the dataset obtained by the American Cancer Society, helped visualize the trend of the number of diagnosed cancer patients of 15 years (2007 - 2021). It also enlisted the number of deaths which was visualized for better understanding. The death rate calculated was 133.5 per 100,000 from 2014 to 2018. The entire dataset combined was sliced to focus on Breast, Cervical, and Oral cancer. It portrayed the rapid increase in patients and also highlighted the importance of cancer literacy.

Defining ML Model Design

Developed an ML classification model to predict if a person is susceptible to breast or cervical cancer using training datasets that consisted of demographic data, habits and relevant medical records of patients. Multiple classification algorithms were trained and the best model was finalized for both cancers.

Data Pre-processing

The dataset for cervical and breast cancer was obtained from Kaggle and BCSC Research respectively. The data was preprocessed by discarding the null and duplicate values from both datasets. Certain columns with exceeding null values and no correlation with the dependent feature (i.e. Cancer) were dropped from both datasets. After this, a dataset of 688 patients and 15,203 patients remained for cervical and breast cancer analysis respectively.

Finalizing the Classsification Model

For cervical cancer, the Decision Tree Classifier produced the best results with a test accuracy of 99.39% and a train accuracy of 100%.
For breast cancer, the Support Vector Classifier produced the best results with a test accuracy of 99.89% and a train accuracy of 99.04%

Devising Solutions

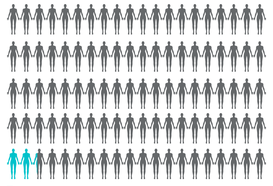
One of the solutions would be to deploy an open-source application that would allow the population to check their susceptibility towards Breast or Cervical Cancer based on their demographic data, habits, and relevant medical records. The application can be accompanied by our API-based searcher or a database supporting searcher which can locate the nearest hospital or cancer treatment center based on their location. In addition to this, we can integrate the health card to maintain medical records of all individuals and conduct awareness drives and campaigns in specific areas which can be decided using a k-means clustering algorithm.

Some awareness projects that took place in Telangana

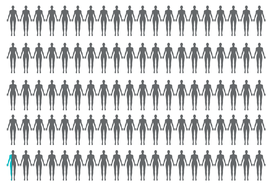
A free cancer screening and training camp was organized at the Appolo Hospital. Over 60 women were trained in procedures like self-breast examination. According to Dr. Nalini, the program aimed at spreading the message across the state that cancer is a curable disease if detected early. Dr. Veeramalla said breast and cervical cancer among Indian women is on the rise. The DLSA of Warangal and Omega Bannu Hospitals organized a free cancer screening test camp in Aug 2021. Tests like Mammography for Breast Cancer, Pap smear for Cervical Cancer, Anaemia Screening by Blood Test and ECG, were conducted at the camp for a total of 80 women.

Rural and Urban Areas

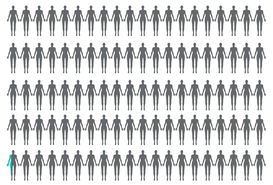
3.9% Percentage of women from Rural Telangana who have ever undergone a screening test for cervical cancer



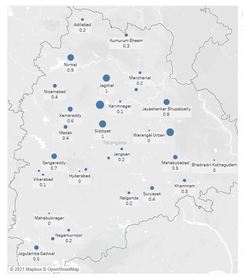
2.3% Percentage of women from Urban Telangana who have ever undergone a screening test for cervical cancer



0.4% Percentage of women from Rural Telangana who have ever undergone a breast examination for breast cancer

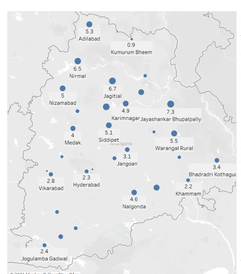


0.3% Percentage of women from Urban Telangana who have ever undergone a breast examination for breast cancer



Breast Cancer table

District	
Rajanna Sircilla	1.2
Warangal Rural	1.2
Jagitial	1
Siddipet	1
Nirmal	0.8
Jayashankar Bhupalpally	0.8
Sangareddy	0.7
Kamareddy	0.6
Jogulamba Gadwal	0.5
Mahabubabad	0.5
Medak	0.4
Nizamabad	0.4
Suryapet	0.4
Khammam	0.3
Kumurm Bheem	0.3
Peddapalli	0.3
Wanaparthy	0.3
Adilabad	0.2
Jangoan	0.2
Mancherial	0.2
Medchal-Malkajgiri	0.2
Nagarkurnool	0.2
Nalgonda	0.2
Yadadri	0.2
Karimnagar	0.1
Ranga Reddy	0.1
Vikarabad	0.1
Bhadradri Kothagudem	0
Hyderabad	0
Mahabubnagar	0
Warangal Urban	0



Cervical Cancer table

District	
Jayashankar Bhupalpally	7.3
Rajanna Sircilla	6.9
Jagitial	6.7
Nirmal	6.5
Warangal Rural	5.5
Adilabad	5.3
Suryapet	5.3
Siddipet	5.1
Nizamabad	5
Karimnagar	4.9
Peddapalli	4.9
Nalgonda	4.6
Medak	4
Bhadradri Kothagudem	3.4
Wanaparthy	3.3
Jangoan	3.1
Vikarabad	2.8
Jogulamba Gadwal	2.4
Yadadri	2.4
Hyderabad	2.3
Khammam	2.2
Mahabubnagar	2.2
Kamareddy	2
Nagarkurnool	2
Mahabubabad	1.8
Mancherial	1.8
Warangal Urban	1.4
Sangareddy	1.2
Kumurm Bheem	0.9
Medchal-Malkajgiri	0.5
Ranga Reddy	0.5

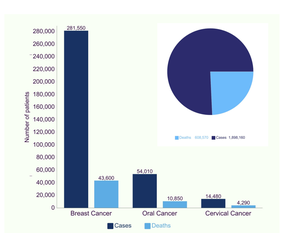
District-wise in Telangana

To make a comparison based on districts in Telangana, a new combined dataset was designed appending all the latitudes and longitudes of the respective districts.

Data visualization on the indicators (ever undergone a screening test for cervical cancer (%), ever undergone a breast examination for breast cancer (%)) was done such that the results were mapped onto the district along with a tabular representation of the percentages in all the districts in decreasing order.

Importance of Cancer Awareness

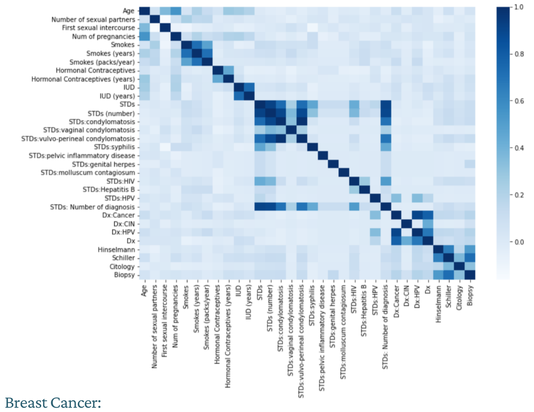
A dataset containing the estimated number of new cancer cases and deaths in 2021 by the American Society of Cancer was considered. The dataset consisted of multiple types of cancers and after data cleaning, it is observed that the incidence of cancer is increasing rapidly; highlighting the importance of cancer literacy amongst the population.



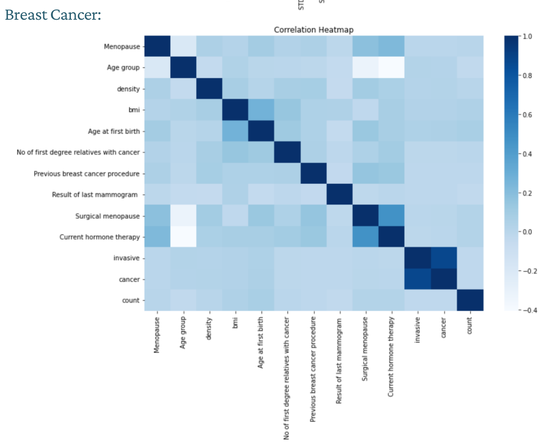
Exploratory Data Analysis

After loading the dataset we first look at the dimensions. Looking at the information of the dataset to get insights into the data like its features, data types of the feature, etc. Thereafter, we preprocess and clean the data. Statistical summary of the features can be useful in inspecting the feature distribution and anomalies if any. And before we can standardize our data, we need to know if we have columns that provide the same (or very similar) information, which could cause our model to perform poorly. This information can be obtained by creating a correlation matrix as shown in the 2 heatmaps below

Cervical Cancer:



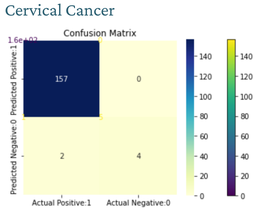
Breast Cancer:



Classification Model Results

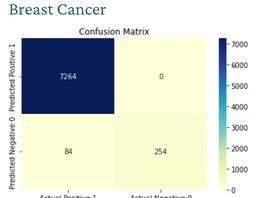
We are very pleased with the final results of our classification models. The models used to train the cervical cancer dataset were stochastic gradient descent, support vector classification, and decision tree classification. In the same way, the models used to train breast cancer data set were support vector classification, decision tree classification, and random forest classification. In the end, we chose the best classifier for each of the datasets.

Cervical Cancer



The decision tree classification algorithm produced the best results. For test data (25%) **accuracy = 99.39 %** while the train data (75%) **accuracy = 100 %**

Breast Cancer



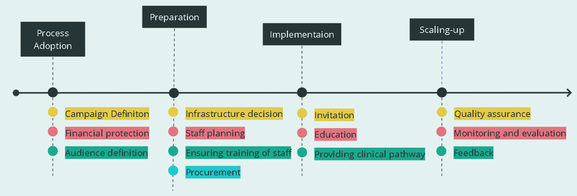
The support vector classification algorithm produced the best results. For test data (50%) **accuracy = 99.89 %** while the train data (50%) **accuracy = 99.04%**

SOLUTIONS

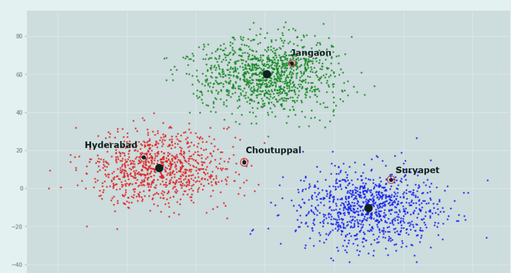
The work presented in this poster devises a classification model which can be used to check one's susceptibility for Breast and Cervical cancer. The **model can be an effective tool as an open-source application** available to everyone. Additionally, the application would consist of other approaches supporting our attempt towards spreading awareness and highlighting its importance.

1. Susceptibility tester along with nearest hospital/centres suggestions
The application will be accompanied by our system which provides suggestions for the nearest hospital or Cancer treatment centers based on the user's location or address. The system incorporates 2 APIs which help find the best suggestions. The first API from Position Stack helps find the latitude and longitude of the user and the second API from MapMyIndia helps find the nearest hospital or Cancer treatment centers. These suggestions will ensure that everyone also knows the correct source to contact in case of concern.

2. Awareness drives and Campaigns
Cancer awareness campaigns are crucial in cancer prevention programs. The aim of these campaigns is to create cancer awareness amongst the population of Telangana. It is important to dispel the myths that people wrongly believe, inform them about the signs and symptoms, and the importance of screening for early detection. Moreover, knowledge of cancer risk factors is a determinant element in this process. It can be implemented using the process timeline shown below.



3. Integration with Health Card
On September 27, 2021, Prime Minister Narendra Modi introduced the digital health id card which will be provided to all people. It will create a seamless online platform that will make all the health-related information portable and easily accessible to doctors. This can be used to integrate with our system for easy access. Everyone's health records will be maintained and used to identify the need for campaigns and drives based on locations and demographics using a k-means clustering algorithm as shown in figure.



CONCLUSION

Considering the district-wise cancer cases and mortality rates in Telangana with an amalgamation of the abovementioned suggestions; it could aid in creation new localized schemes, awareness drives, training camps, and financial support policies can be organized which can be tailored to the level of cancer literacy and degree of urbanization in that district in Telangana. This will take us a step closer to our goal of increasing cancer literacy among women.

metafour.developers@gmail.com