

# DEVELOPING MECHANIZED FRAMEWORK TO INVESTIGATE AND CLASSIFY BREAST CANCER TUMOURS BY EMPLOYING NATURAL LANGUAGE PROCESSING (NLP) TOOLS AND TECHNIQUES, 2018

Drishti Arora

## ABSTRACT

*Breast Cancer is the prime reason for death in Indian ladies. Medical clinics in India utilize electronic methods for assortment and detailing of information. One such report is the Pathology report which has regular language portrayals of the states of patients. This work plans to extricate the subtleties on Tumour (T) in the bosom utilizing design coordinating standards and determine the neurotic characterization of T by applying the PTNM arrangement convention by American Joint Committee on Cancer (AJCC). Data Retrieval (IR), Natural Language Processing (NLP) undertakings and Information Extraction (IE) strategies are applied to build up a mechanized framework to achieve the errand. The framework investigates the removed and the ordered estimations of T against the Gold Standard Values, which are inferred by manual examination of the reports. The assessment of the exhibition of the computerized framework performed utilizing three arrangements of Pathology reports, brought about a normal Precision of 86%, Recall of 82.7%, Specificity of 75.1% and Accuracy of 79.53%.*

## 1. INTRODUCTION

Application of computers in health care and medical diagnostics are increasing day by day. Numerous work using medical data and development of automated systems, especially on Breast Cancer have been done in the past all over the world. The criticality of breast cancer in India necessitates the study and analysis of regional data. The work presented in this paper focuses on processing of the 'Impression Section' of de-identified natural language Breast Cancer Pathology reports obtained from a hospital in South India. The size of Tumour (T) in the breast and the medical conditions associated with T narrated in natural language are derived from this section and used to pathologically classify T. The automated system developed applies the process on three datasets and evaluates the process by calculating the parameters precision, recall and accuracy. Several pre-processing steps are applied on the natural language narratives before the classification process. Simple queries on the results show the most prevalent tumour and the distribution of the various T classifications among the patient population. In the future, Parameters corresponding to Lymph Node N and Metastases M would also be extracted and classified from the Impression section, to automatically derive the stage of cancer. The automated system would help to identify the regional distribution of the patient population and the age group of patients with the most prevalent cancer. The ultimate aim of this work is to develop a Decision Support System for Breast Cancer Pathology to assist the Medical practitioners in quick diagnosis of cancer

stage of patients. The challenges in processing Electronic Health documents in the Indian context as compared to similar works in the Western countries are listed below:

- The rare and limited availability of medical reports for analysis and research.
- Hospitals in India do not adhere to uniform coding and structuring of medical data at the data collection phase. The format and language style used in reporting vary from one person to another.
- Pathology reports in India do not use any Medical codes (SNOMED/ICD-O) to represent various medical conditions and diagnostics.
- Absence of standard formats and coding systems in clinical records necessitate frequent and extensive support from Medical experts to apply domain knowledge for application development and analysis of the results.

In spite of the above challenges, the fact that India is ranked number one in Breast-cancer deaths in the world, is a huge motivating factor for this work. This paper is organized as follows: Section II describes the Materials and Methods used. Section III describes the Technique used and Section IV discusses the results obtained and the inferences made. Section V presents the Conclusions.

Various researches have been carried out in the past using the Medical data using NLP, and IE. A few related works are discussed below. K. Saravana Kumar and A.M. Arthanarieel evaluated the risk factors associated with breast cancer. K. Vaidehi and T.S. Subashini<sup>2</sup> classified the breast tissue using K-NN classifier. P. Yasodha and N.R. Ananthanarayanan<sup>3</sup> built a knowledge-based system for early detection of Ovarian Cancer. The work by Nelson et al.<sup>4</sup> focuses on developing a Pathology database. Erik Cambria and White<sup>5</sup> extracted information from pathology reports and also identify the various Schools of thought on NLP research namely Approaches that use Production rules by Chomsky (1956), Semantic Pattern matching approaches that use Semantic categories and Semantic case frames by Ceccato (1967), Approaches based on First-Order Logic (FOL) that use Axioms and Rules of Inferences for NLP tasks by Barwise (1977), Bayesian networks that use Variables represented by a probabilistic directed acyclic graph by Pearl (1985), Semantic networks which have Patterns of interconnected nodes and arcs used for NLP by Sowa (1987) and Ontology Web Language (OWL) which represents information as Hierarchical classes and relationships between them used by Mc Guinness and Van Harmelen (2004).

Natural Language Processing of Pathology reports have been attempted in the past. Moh'd Rasoul et al.<sup>6</sup> detected breast cancer by processing images. Buckley et al.<sup>7</sup> have identified the widespread variation of how pathologists report common pathologic diagnosis. The dataset they utilized had 124 different ways of saying intrusive ductal carcinoma and 95 different ways of saying obtrusive lobular carcinoma. There were > 4000 different ways of saying that obtrusive ductal carcinoma was absent. They utilized International ICD-9 and Current Procedural Terminology (CPT) codes to distinguish those reports relating to the breast. Nguyen et al.<sup>8</sup> separated malignant growth notice things dependent on Symbolic guideline-based grouping strategy, by distinguishing SNOMED CT ideas in the free content. Napolitano et al.<sup>9</sup> has removed information utilizing design coordinating. Annie Coden et al.<sup>10</sup> consequently extricated malignant growth ailment attributes from pathology

reports. Meystre et al.<sup>11</sup> endeavored extraction of data utilizing Pattern-Matching Techniques and Full/halfway parsing. Chen et al.<sup>12</sup> have handled clinical accounts to separate practice design patterns utilizing NLP. McCowan et al.<sup>13</sup> have dealt with assortment of Cancer Stage Data by Classifying Freetext Medical Reports. They separated malignant growth warning things dependent on Symbolic standard-based characterization technique, by distinguishing SNOMED CT ideas in the free content. The trouble in handling Breast Pathology reports legitimately utilizing NLP, because of the absence of accentuation in the content and data on different examples in the report was accounted for by Xu and Friedman<sup>14</sup>. The Clinical reports in the Western world utilize medicinal coding frameworks (SNOMED/ICD) and UMLS Knowledge assets. Schadow G. furthermore, McDonald C.J.<sup>15</sup> built up a strategy that concentrates organized data about examples and their related discoveries in free-content careful pathology reports. NLP systems have been applied in various examinations in Medicine. Hripcsak et al.<sup>16</sup> built up a Natural Language Processor Med LEE and utilized it to code more than 800,000 Chest Radiograph reports. Numerous Clinical NLP frameworks have been created in the Western world, for example, LSP-MLP (Fortran and C++), Med LEE (Prolog), SPRUS/SynText/MPLUS (LISP, C++), Meta Map (Perl C, Java, Prolog), HITex (Java) and cTAKES (Java). The content parts were coded against the UMLS.

David Martinz, Yue Li<sup>17</sup>, have extracted information from pathology reports. The Breast Cancer related details were obtained from the online resource<sup>18</sup>. Most of the Breast Cancer research carried out in India use downloaded Breast Cancer datasets from the Western world, the most common one is the Wisconsin Breast Cancer Dataset. In contrast, this work uses the dataset consisting of Pathology reports of patients in the region. Hence the results obtained have relevance and practical applicability to the local patient population.

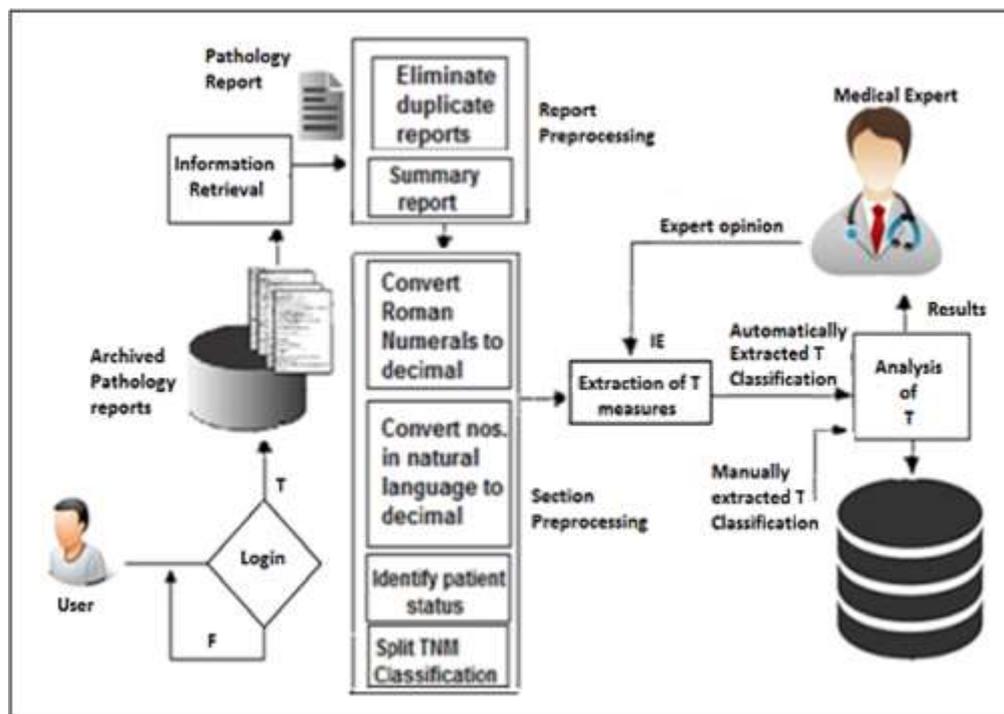
## 2. MATERIALS AND METHODS

### 2.1 Dataset

The Dataset for this project was acquired from one of the reputed hospitals in South India that has treated numerous cancer patients across India. The 150 de-identified Breast Cancer Pathology reports are in pdf format with the structure described in the following section.

### 2.2 Structure of the Pathology Reports

A pathology report is a document that contains the diagnosis determined by Examining cells and Tissues under a Microscope.



**Figure 1.** System Architecture for Extraction and Analysis of T.

Pathology reports play an important role in Cancer diagnosis and Staging. A general pathology report constitutes of the following sections as indicated in online resource17:

- Personal information: Name, age, sex, and address of the patient.
- Specimen: This section describes where the tissue samples are taken from. Breast Cancer Tissue samples could be taken from the Breast, Lymph nodes under arm (axilla), or both.
- Clinical history: This is a short description of the patient and how the breast abnormality was found. It also describes the kind of surgery that was done.
- Clinical diagnosis: This is the diagnosis expected by the doctors before the tissue sample was tested.
- Gross description: This section describes each piece of the tissue removed - its Size, Weight, and Colour.
- Microscopic description: Depicts the manner in which the disease cells looked under the magnifying instrument, their relationship to the ordinary encompassing tissue, and the size of the malignant growth.
- Exceptional markers: results for Proteins, Genes, and Cell growth rate.
- Final diagnosis Summary: this section has a short description of tissues examined

### 2.3 TNM Classification

A staging system for cancer indicates how far the cancer has spread. TNM classification define Breast Cancer by AJCC has designated staging by. There are two methods of staging, namely Clinical Staging and Pathological Staging. Between the two, Pathological staging is more accurate. The T classification of the primary tumour is the same regardless of whether it is based on clinical or pathologic criteria or both. This work derives the T classification of pathology reports, and hence uses pTNM staging.

## 3. TECHNIQUE USED

The System Architecture shown in Figure 1 contains three important tasks namely Pre-processing of

reports, Extraction of T and Analysis of T. According to Napolitano<sup>9</sup>, the goal of information extraction is to extract structured and semantically well-defined concepts from unstructured data sources in order to facilitate access and retrieval of information. This work applies Pattern-based extraction to extract the measures and factors required to classify T.

### 3.1 Pre-processing

Pre-processing is a necessity when we process natural language text, in order to bring homogeneity to the documents for computer processing. The Pathology reports in pdf format are first converted to plain text. Pre-processing at the report level eliminates duplicate reports using the combination of the Patient Id and the pTNM classification and generates a summary of the reports in the dataset. At the Section level pre-processing, numerical values represented as Roman numerals or English words are converted to their corresponding Arabic numeral form. The measure of tumour is also homogenized across the dataset by converting all measures to centimeter.

### 3.2 Segmentation

The pre-processed Pathology reports are segmented into sections and stored into various section tables in the database. The Impression section of the reports is then processed by segmenting the sentences corresponding to the Tumour, for extraction of T and its classification.

### 3.3 Extraction of T

In the Breast Cancer staging using pTNM, T represents the Primary Tumour, N represents Lymph node and M represents Metastases. The size of the tumour is the most important parameter that is used to classify T. In addition to the tumour size, certain medical conditions in natural language such as 'No evidence of primary tumour', 'Chest wall', 'Skin oedema/ulceration/satellite skin nodules', 'Inflammatory carcinoma', are used in classifying T. The various conditions for classifications of T are listed in Table 1. The pattern-based extraction method uses the pTNM classification protocol to extract details corresponding to T and classifies the tumour. The classification is further refined by applying exception conditions provided by the Medical expert

as tabulated in Table 2. The conditions listed in the table refer to the benign conditions that may occur in women that are not indicative of breast cancer.

**Table 1.** Conditions for Classification of T Classification

S. No.	Conditions	T-Classification
1	Primary Tumour cannot be assessed	TX
2	No evidence of primary tumour	T0
3	Ductal Carcinoma in situ/DCIS	Tis(DCIS)
4	Lobular Carcinoma in situ / LCIS	Tis(LCIS)
5	Paget's disease	Tis(Paget)
6	Tumour size $\leq 0.1$ cm	T1mic
7	Tumour size $> 0.1$ cm to 0.5 cm	T1a
8	Tumour size $> 0.5$ cm to 1 cm	T1b
9	Tumour size $> 1$ to 2 cm	T1c
10	Tumour size $> 2$ to 5 cm	T2
11	Tumour size $> 5$ cm	T3
12	Chest wall	T4a
13	Skin oedema/ulceration/satellite skin nodules	T4b
14	Both 4a and 4b	T4c
15	Inflammatory carcinoma	T4d

**Table 2.** Exception Conditions for T Classification

S. No.	Conditions/T-Classifications	T Classification
1	Hyperplasia, Phyllodes tumour, Fibrosis, Cysts, Apocrine metaplasia	No classification
2	Tis(DCIS)	T1
3	T1mic	T1
4	T1a	T1
5	T1b	T1

The discrepancy report generated after extraction of T presents the details of Pathology reports in which the Gold standard and the automatically extracted T Classifications with the medical experts, as medical applications need more accuracy.

## 4. RESULT

### 4.1 Evaluation Parameters

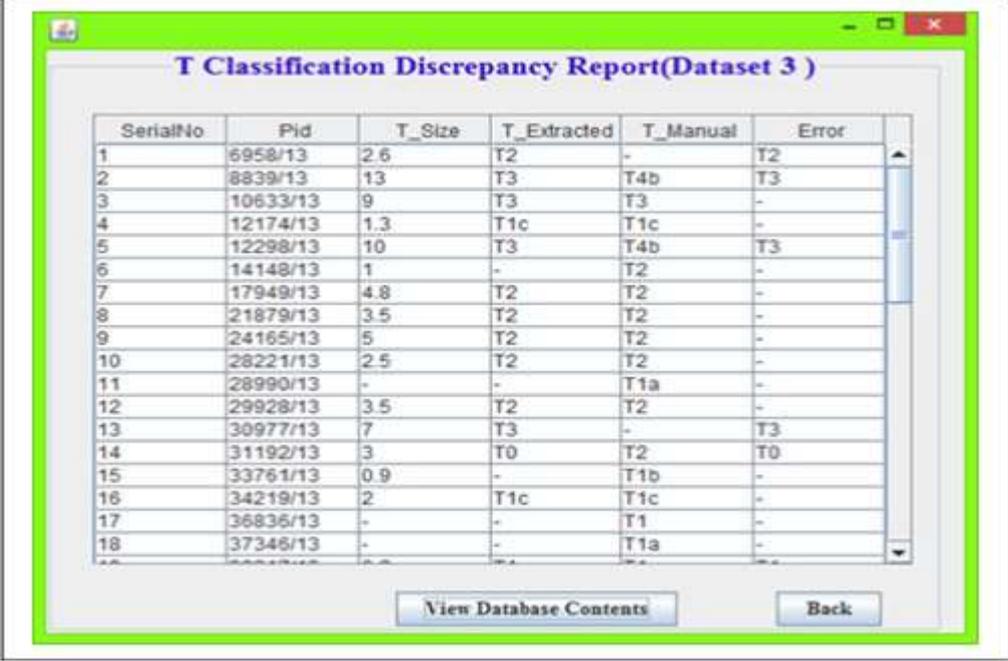
The Evaluation Parameters derived for the three datasets and listed in Table 3. Dataset I has 48 reports since it has two duplicate reports which are eliminated during pre-processing.

SerialNo	PatientId	T_Size	T_Extracted	T_Manual	Error
1	18462/12	4	T2	T2	-
2	18605/12	0.5	T1	T1	-
3	19408/12	-	-	-	-
4	21598/12	2	T1c	T2	T1c
5	18148/12	4.5	T2	T2	-
6	18464/12	0.5	T1	Tis(DCIS)	T1
7	19286/12	2.5	T2	T2	-
8	20609/12	2.5	T2	T2	-
11	21026/12	4	T2	T2	-
12	19737/12	3.5	T2	T2	-
13	20183/12	0.9	-	T1	-
14	20276/12	3.5	T2	T2	-
15	20614/12	-	-	-	-
16	19288/12	5.3	-	-	-
17	20186/12	4	T2	T2	-
18	20707/12	-	-	-	-
19	21327/12	-	-	-	-
20	21493/12	2	T1c	T2	T1c
21	19291/12	-	-	-	-
22	21178/12	-	-	-	-

Figure 2. T-classification Discrepancy Report for Dataset 1.

SerialNo	Pid	T_Size	T_Extracted	T_Manual	Error
1	1017/13	-	-	-	-
2	3270/13	3.5	T2	-	T2
3	3310/13	-	-	T0	-
4	332/13	-	-	-	-
5	1060/13	-	-	-	-
6	214/13	2.2	T2	T2	-
7	1867/13	4	T2	T2	-
8	1986/13	4.5	T2	T4b	T2
9	2876/13	13	T3	-	T3
10	220/13	3	T2	T2	-
11	1588/13	-	-	-	-
12	1884/13	2.5	T2	-	T2
13	1920/13	6.5	T3	T3	-
14	1982/13	4	T2	T2	-
15	2374/13	2	T1c	-	T1c
16	3272/13	1.5	T1c	T1c	-
17	1377/13	-	-	-	-
18	2006/13	6.5	-	T3	-
19	2276/13	2	T2	T2	-

Figure 3. T-classification Discrepancy Report for Dataset 2.



SerialNo	Pid	T_Size	T_Extracted	T_Manual	Error
1	6958/13	2.6	T2	-	T2
2	8839/13	13	T3	T4b	T3
3	10633/13	9	T3	T3	-
4	12174/13	1.3	T1c	T1c	-
5	12298/13	10	T3	T4b	T3
6	14148/13	1	-	T2	-
7	17949/13	4.8	T2	T2	-
8	21879/13	3.5	T2	T2	-
9	24165/13	5	T2	T2	-
10	26221/13	2.5	T2	T2	-
11	28990/13	-	-	T1a	-
12	29928/13	3.5	T2	T2	-
13	30977/13	7	T3	-	T3
14	31192/13	3	T0	T2	T0
15	33761/13	0.9	-	T1b	-
16	34219/13	2	T1c	T1c	-
17	36836/13	-	-	T1	-
18	37346/13	-	-	T1a	-

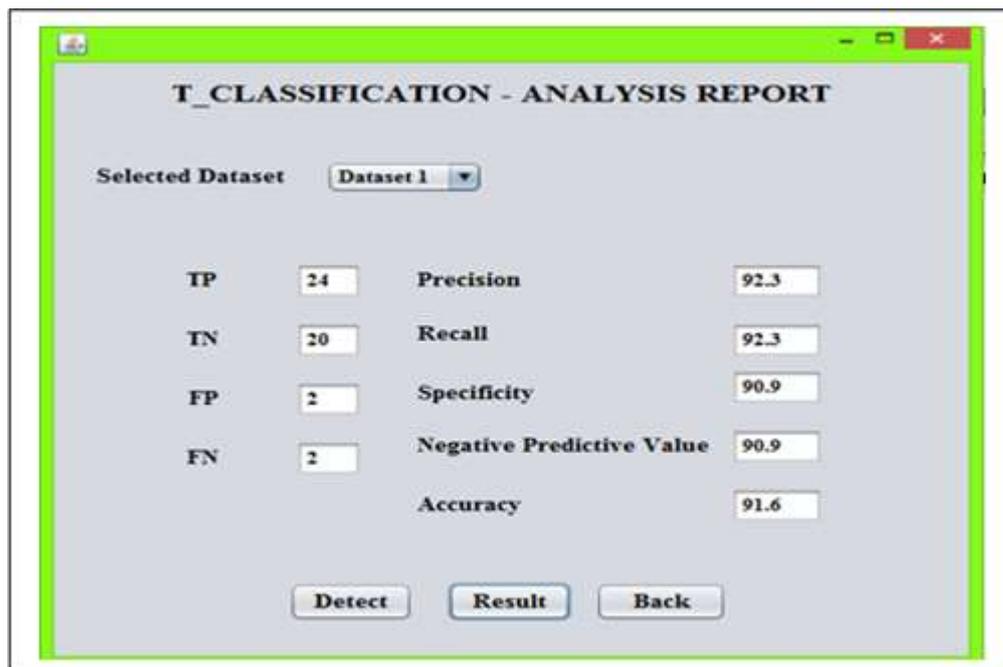
**Figure 4.** T-Classification Discrepancy Report for Dataset 3.

#### 4.2 T-Classification Discrepancy Reports

The report listing the manually extracted T classification, automatically extracted T classification and the reports with discrepancy are generated for the three datasets as shown in Figure 2-4.

#### 4.3 T-Classification Analysis Reports

The extraction and classification of T performed on three datasets consisting of 150 de-identified Pathology reports yielded considerably good results. The Analysis reports for the three datasets are shown in Figure 5. The results of analysing the three datasets, with respect to the extraction and classification of T are summarized in Table 4.



**Figure 5.** T-Classification Analysis Report – Dataset 1.

**Table 4.** T Classification Analysis Summary

Parameters	Dataset		
	I	II	III
<i>Precision (%)</i>	92.3	80.0	85.7
<i>Recall (%)</i>	92.3	87.5	68.5
<i>Specificity (%)</i>	90.9	61.1	73.3
<i>Accuracy (%)</i>	91.6	78.0	70.0
<i>Negative Predictive value</i>	90.9	73.3	50.0

4.4 Evaluation Metrics Query The evaluation parameters TP, TN, FP and FN are determined for the extracted T values. Using these values Precision, Recall, Specificity, Negative Predictive Value and Accuracy are derived.

#### 4.5 Reports

Queries executed on the T-classification indicates that 38.66% of the patient population were treated in the category T2 of tumour.

**Table 3.** Evaluation Parameter for Datasets 1-3

Parameters	Dataset		
	<i>I</i>	<i>II</i>	<i>III</i>
<i>TP</i>	24	28	24
<i>TN</i>	20	11	11
<i>FP</i>	2	7	4
<i>FN</i>	2	4	11
<b>Total</b>	<b>48</b>	<b>50</b>	<b>50</b>

\*Dataset I has 2 duplicate reports

## 5. CONCLUSION

The processing of Natural Language Pathology reports to extract and classify T has offered reasonably good results, in spite of many constraints mentioned in the introduction section. The shortfall in the precision level can be attributed to the following reasons.

- Omissions: The Pathologist might have not reported the pTNM classification due to oversight/the T classification may be not be entered during report writing, while the application would automatically derive and classify T.
- Natural Language Expression: The Pathology reports are documented by different Pathologists whose language style both in English and mention of medical terms may vary. For example, the tumour could be mentioned as Lump, Tumour or Lesion. An exhaustive consideration of language variations in medical terms would improve the precision, however, this is a human impossibility and requires the use of a coding reporting system. The future work would incorporate the extraction of information regarding two more important parameters relating to Cancer staging namely, Lymph Nodes (N) and Distant Metastasis (M). This would help in a comprehensive understanding of cancer stage of patients, success rate of treatment and development of a Decision Support System for the Pathology Department of the hospital.