IJCTA, 9(10), 2016, pp. 4321-4330

Summarization and Visualization of Textual Data in Breast Cancer Pathology Reports

***Johanna Johnsi Rani G **Dennis Gladis ***Joy John Mammen**

Abstract : The efficiency of patient care depends on how fast a Medical expert reads and interprets reports. Natural language text reports are elaborate and Summarization of data from such reports of individual patients or a group of patients is a necessity for quick decision-making. In this work, an automated system was developed to process Breast cancer Pathology reports applying Natural Language Processing and Information Extraction techniques and to generate summary reports of patients. The template for summarization of individual reports was designed based on The College of American Pathologists (CAP) protocol for breast cancer. The system also presents the visual summary of cancer stage on the entire dataset and Query-based summaries. The computer-generated summaries are an important component of the Domain-specific Decision support system that is being developed, using which Medical experts can read and interpret the condition of patients to provide efficient service.

Keywords : Breast Cancer; Summarization; Information Extraction; Natural Language Processing; Decision Support System

1. INTRODUCTION

Statistics in 2016 states that India ranks number two, with 17.84% of the total world population [13]. With 641 million of this population being female [14], a global study commissioned by GE Healthcare, estimated that by 2030, the incidence of new cases of breast cancer in India will increase from 115,000 to around 200,000 per year [15]. Processing the reports of every cancer patient who approaches for diagnosis and treatment and summarizing the processed data in every region is a necessary proactive measure to address this critical situation. In India, natural language text narrations are the most common form of reporting. According to Ellis DW and Srigley J, the timeliness or turnaround time of a cancer pathology report is as important as the accuracy of the diagnostic and prognostic observations [3]. With numerous patients to treat every day, Pathologists have limited time to summarize details and study the patient population for reporting. Hence providing a computer-assisted support to a Pathologist for quick decision-making is the best possible solution to shorten the turnaround time. Summaries provide a fast overview [2]. Processing natural language text and extracting essential details requires numerous pre-processing steps as the reports are in a highly heterogeneous form. Bringing the textual content to a summary report form also requires a precise extraction process that converts the unstructured report into a structured form and store in a database. The Natural language processing and information extraction techniques provide important functions and tools for generation of summaries. With technology enhancement, hospitals now use electronic forms of medical reporting that index and store them in databases. Processing the existing electronic medical documents such as breast cancer pathology reports and summarizing the data would support quick diagnosis and treatment for patients and also provide an overview of the patient population.

* Department of Computer Science Madras Christian College Chennai 600 059, South India gjojora@yahoo.com

** Department of Computer Science Presidency College Chennai 600 005, South India gladischristopher@gmail.com

** Department of Transfusion Medicine Christian Medical College Vellore - 632 004, South India joymammen@cmcvellore.ac.in

The proposed system summarizes text data in de-identified breast cancer pathology reports obtained from a renowned hospital in South India. The system focuses on generation of individual patient's pathology report summary, graphical summarization of the cancer stage of the patient population and query-based summaries. The summary of a patient's report is obtained using pattern-matching rules and presented in a format derived from the scientifically validated breast cancer checklist defined by 'The College of American Pathologists' (CAP) [11]. The automated system derives the Pathological classification pTNM, in which T represents Tumour, N represents the Lymph node and M represents the Distant Metastases. The stage of Cancer of patients is then determined based on the Pathological stage grouping proposed by the 'American Joint Committee on Cancer' (AJCC) [12]. The graphical summarization of cancer stage details of patients thus derived from the dataset highlights the severity of the disease in the patient population in the region. Query-based summarization can be used to project details as requirements arise, based on factors such as gender, age group, geographic region, or any vital Medical parameter. The automatic summary generation and visual presentation is an important component in a Decision Support System for Breast Cancer Pathology that is being developed to help the medical experts in quick decision-making.

The paper is organized as follows: Section II describes Related Works in Data Summarization and Visualization. Section III explains the Method of Summarization and Visualization. Section IV presents the Results obtained and their interpretation and Section V presents the Conclusion.

2. RELATED WORKS

Most of the medical documents generated in patient care are narrative reports. Joshua et al., highlighted the importance of Computer-assisted Summarization as it presents essential data in a format that assists in communication and decision-making and categorized Clinical summaries into source-oriented, time-oriented and concept-oriented summaries [1]. The individual patient's report summary presented here is a simple source-based structuring of the textual data for easy reading and interpretation. Donia Scott, Catalina Hallett, and, Rachel Fettiplace assessed the efficiency of AI-based computer-generated textual summaries of patient histories and found that computer-generated reports are more accurate and efficient than human-produced patient records [2]. The system used generic and medical domain-specific rules to generate summaries.

According to Catalina Hallett, Richard Power and Donia Scott, it is desirable to generate summaries that provide a 30-second overview and fits entirely on the computer screen. [10]. The patient report summary generated by our system presents such a single, complete snapshot of the textual report. Boyd A.D et al. developed an application to summarize details in discharge summaries using SimpleNLG [6]. Sneh Garg and Sunil Chhillar proposed a document summarization method that involves corpus coverage, sentence coverage and term coverage weight [7]. Lankshear S, et al mentions the possibility of essential information being omitted by the pathologist, if he is not guided by a tumor-specific template or checklist during summarization [4]. Hence, the automated system developed uses a standard breast cancer checklist by CAP as a reference to avoid omissions while generating individual patient report summary.

A population-based overview of breast cancer would throw light on the geographic area and the age group to be focussed for treatment of the disease. Xue Qin Yu et al. used population-based data and innovative statistical methods to study breast cancer prevalence in a geographic region [8]. Though summarization is performed on a small dataset of 150 Pathology reports in the proposed system, future testing with a large dataset would provide an overview of the entire breast cancer patient population treated at the hospital. The Visual summarization of patient population is concept-oriented as it categorizes the patient population based on the stage of cancer. Simple graphs are used for this summarization for easy interpretation.

Ahmed A. Mohamed, Sanguthevar Rajasekaran proposed Query-based text summarization based on document graphs [9]. The system presented processes each de-identified unstructured report and extracts the data to a structured form, thus enabling numerous query-based summarizations on medical details. With the availability of demographic information of patients later, more query summarizations based on age, region etc. can be performed by the Pathologists for deeper and wider understanding of the patients and the affect of the disease.

3. METHOD OF SUMMARIZATION AND VISUALIZATION

Clinical Summarization has become a necessity in medical practice due to several reasons. Summarization of a single patient's reports and chronologically linking of them can help in the study of the patient's medical history. Genetic disorders can be diagnosed for treatment. Such summarizations can be transmitted to hospitals across the globe when a standard summarization template is used. Population based summarizations are essential to predict the spread of diseases and propose health-care measures. Three types of summarizations are generated by the system namely, Individual patient report summarization, Population-based summarization, and Query-based summarization.

A. Input

The corpus used in this work is a set of 150 de-identified breast cancer pathology reports. The reports were obtained as PDF files and were converted to text files for processing. A sample report from the dataset is presented in Fig. 1.

2) 18605/12

SPECIMEN : Right MRM.

CLINICAL : Carcinoma right breast, cT2N0M0, post NACT x 4 cycles.
Presently no lump palpable.

GROSS : Right modified radical mastectomy specimen measuring 20.5x20x2cm with a ellipse of nipple and areola bearing skin measuring 15x3.5cm.External surface of skin appears unremarkable.On sectioning there is a ?tumor measuring 2x1.5cm in the upper outer quadrant with a tan firm cut surface and it measures 4.5cm from the deep resection margin.Rest of the breast tissue appears predominantly fibrofatty.Attached axillary pad of fat measures 10x9.3x0.5cm,sectioning reveals 17 lymph nodes and the largest measures 1cm in dia with grey white firm cut surface.
A)Nipple and areola 8 all(A1-A8)
B)?Tumor 11 bits (B1-B11,B1 to B3 composite bits,B4&B5 composite bits,B6&B7 composite bits,B8&B9 composit bits,B10&B11 composite bits,the respective deep margins are marked with india ink) BFB1-3, ?tumour.
C)Upper outer quadrant 1 bit (1 block)
D)Lower outer quadrant 1 bit (1 block)
E)Upper inner quadrant 1 bit (1 block)
F)Lower inner quadrant 1 bit (1 block)
G)17 Lymph nodes 34 all (G1-G17) FIX.RS/lv

MICRO : A) Shows sections of nipple and areola with underlying breast tissue, no specific lesion.
B) Shows sections of breast tissue with mild stromal sclerosis and a focus of usual ductal hyperplasia(B10). The inked deep resection margin is free of tumour. There is no evidence of malignancy.
BFB) Shows breast tissue with microscopic foci of high grade ductal carcinoma in situ with a focus of invasion. The tumour cells have round to oval nuclei, vesicular chromatin, prominent nucleoli and moderate eosinophilic cytoplasm. The stroma shows desmoplastic reaction with mononuclear infiltrates. The largest tumour size is 0.5cm(BFB3).
C-F) Shows sections of breast tissue with no specific lesion, no tumour.
G) Shows seventeen lymph nodes with reactive hyperplasia, there is no evidence of viable tumour.

IMPRESSION : Right modified mastectomy specimen (Post NACT 4 cycle):
Multifocal microscopic residual high grade ductal carcinoma in situ with a focus of invasion.
Largest tumour size 0.5cm.
Deep surgical resection margin free of tumour.
Nipple and areola free of tumour.
No lymphovascular or perineural invasion.
Seventeen out of seventeen right axillary lymph nodes with reactive changes, no viable tumour.
ypT1N0Mx

Fig. 1. A sample Breast Cancer Pathology Report.

The report has the following sections: Demographic information with Serial No. and Patient-ID, Specimen section, Clinical history section, Gross description, the Microscopic description section and the Impression section which has short descriptions of the important findings, along with the Pathological Tumour-Lymph node-Metastasis Classification (pTNM Classification). The entire report is considered for Individual Patient report summarization, while Population-based summarization derives the cancer stage from the Impression section alone. The structured data derived from the textual report and stored in a database is used for Query-based summarization.

B. The Workflow

Information obtained though pre-processing, cancer staging, and gold standard editing is used for summarization and visualization of results. The workflow of the summarization and visualization of summary is presented in Fig. 2.

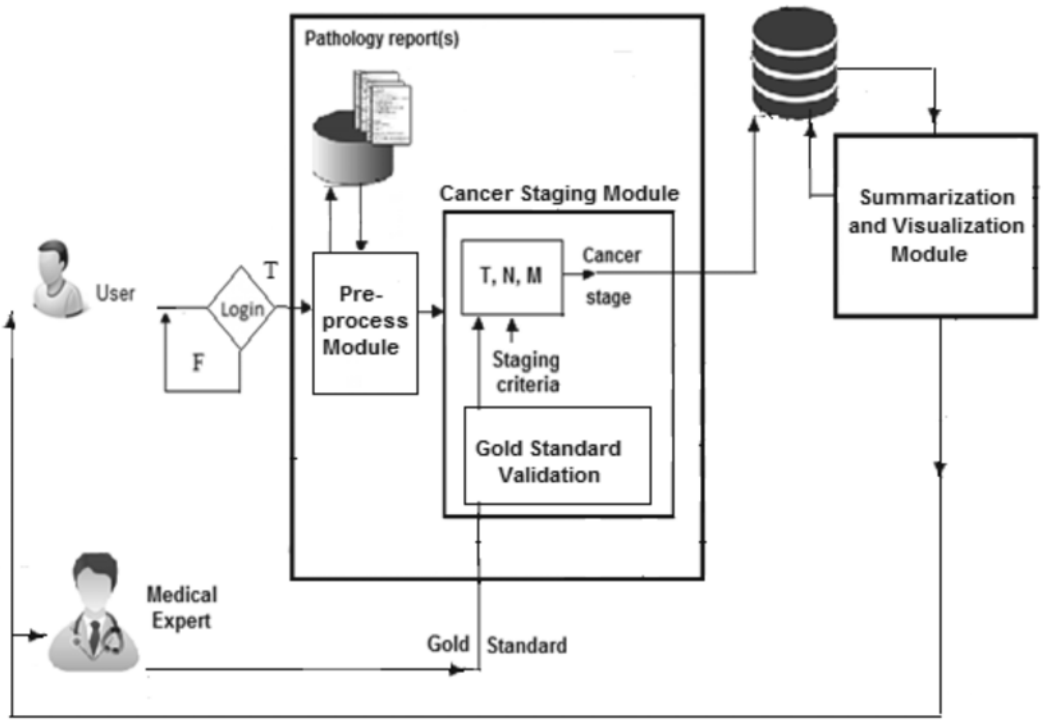

Fig. 2. Work Flow of Summarization and Visualization Process

The summary report of individual patients is obtained through application of natural language pre-processing tasks both on the entire dataset containing multiple reports and the contents of every pathology report. The pre-processing tasks homogenizes the contents and extracts essential data using pattern-matching rules thus converting the narrative report content to a structured form. The report segregation task separates multiple reports in the dataset to individual reports. On the textual content of each report, section segmentation, standardization of measures, homogenization of date information, sentence segmentation and, standardization of numerical values are performed on the textual content of each report in the dataset. In addition to the above, expansion of abbreviations, standardization of spelling variations, whitespace removal, handling of parenthesized terms, handling the case-sensitivity of medical terms and insertion of missing headers are performed. The main parameter that is summarized is the stage of cancer of patients that is derived by grouping the components of the pTNM classification.

The medical experts were provided with a Gold Standard Editor (GSE) to manually scrutinize the reports and approve the final classifications for T and N alone as M is a clinically determined parameter that is given a default value of M0. The GSE permits the experts to alter the values in the printed report after providing the reason for the change. This step removes the discrepancies in the reports such as wrong or missing classifications. The extraction process is evaluated by comparing the Gold standard values and the extracted values and calculating the parameters Precision, Recall, Accuracy and Specificity. This ensures that the visual summary of population-based reports reflect reality. A simple visualization tool such as bar graph is used to present the graphical summarization of data.

C. The Individual Patient Report Summarization

The pre-processing steps are vital to the accuracy of the information extraction process. The extraction process converts the unstructured data to a structured form and stores the extracted values in a database. Presenting the summary in a standard, globally-acceptable format is essential for decision-support. Several standards and checklists are available for cancer pathology reporting. The College of American Pathologists (CAP) is an organization with certified pathologists who advocate excellence in the practice of pathology and laboratory medicine worldwide. They provide Cancer protocol templates for reporting essential data on malignant tumours. Two protocols for Breast cancer pathology reporting namely, DCIS Breast and Invasive Breast were used for the summarization process. The templates were customized to the hospital's needs and it can be used as a GUI for the Pathologists in the future.

D. Population-based Summarization of Cancer stage

The visual reporting of cancer stage of a population is a complex process which has the following components – extraction of cancer stage of each patient, obtaining the gold standard data from the reports and resolving discrepancies through the GSE and graphical representation of the cancer stage of patients. The accuracy of the summarization depends on the accuracy of the extraction process. To be compatible with globally accepted medical practices, the cancer stage was derived through extraction of the Pathological classification pTNM and the Pathological stage grouping of AJCC which is summarized in Table I and applied to the proposed system.

Table 1. Breast Cancer Stage Grouping by AJCC.

Stage	*T*	*N*	*M*	*Stage*	*T*	*N*	*M*
0	Tis	N0	M0	IIIA	T0	N2	M0
IA	T1	N0	M0		T1	N2	M0
IB	T0	N1mi	M0		T2	N2	M0
	T1	N1mi	M0		T3	N1	M0
IIA	T0	N1	M0		T3	N2	M0
	T1	N1	M0	III B	T4	N0	M0
	T2	N0	M0		T4	N1	M0
IIB	T2	N1	M0		T4	N2	M0
	T3	N0	M0	IIIC	Any T	N3	M0
				IV	Any T	Any N	M1

E. The Output

The automated system generates several outputs under the three categories of summarizations namely, individual patient report summary, population-based summary and query-based summary. The individual patient summary is a single-screen, online snapshot of the textual report. Population-based summary is presented in graphical form, while query summarizations are generated as listings according to user needs. The objective of the summarization component in the system is to provide the medical expert with a quick and efficient decision-support tool. While individual patient summary helps the medical expert to focus on quick diagnosis of a patient's condition, population based summary is valuable for understanding the spread of the disease in the region and the vital statistics on the cancer stage in which patients report for diagnosis of the disease. Query summaries provide lists as per requirements that arise in the day to day treatment of patients.

F. Evaluation of the results

The evaluation of the individual patient summarization process is through automatic comparison of the textual report data and the generated patient data summary. The graphical summary of cancer stage is performed through a series of manual comparisons of the parameters associated with cancer staging namely T-Classification, N-Classification, and the Cancer stage. M-Classification representing Distant Metastasis is given a default value of M0, since it is not pathologically determined. The extraction process is evaluated using the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values and calculating the following parameters.

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Specificity} = TN / (TN + FP)$$

The cancer staging process yielded 88.33% Precision, 100% Recall, 90.54% Accuracy and 66.66% Specificity. The deficiency in the process can be attributed to the heterogeneity of the reports. Applying natural language processing tasks can improve the results in the medical decision support system. This would ensure that the graphical summaries are true representations of the textual reports and the query reports are highly reliable to apply for decision-making.

4. RESULTS

The results of the processed textual data in pathology reports is presented using structured data view and graphical view. It provides an easy human and machine-readable form of the contents for machine analysis for decision-making. The results obtained in the Summarization and Visualization tasks, applying Information Extraction and Natural Language Processing methods are presented in this section.

A. Individual Patient Report Summary

The automated system generates a single-screen snapshot summary of the pathological diagnostic of a patient from the textual reports. The report has five sections – Specimen, Clinical, Gross, Micro and Impression. Pattern-matching rules are applied for the extraction process on each of the sections. The template to present the extracted summary is a customized one based on the CAP checklist [11]. The accuracy of the generated summary is fully dependent on the accuracy of its input. Fig. 3 presents the summary report for the patient report shown in Fig. 1. The result shows that the summary is easily readable and enables a faster interpretation of data than the textual report.

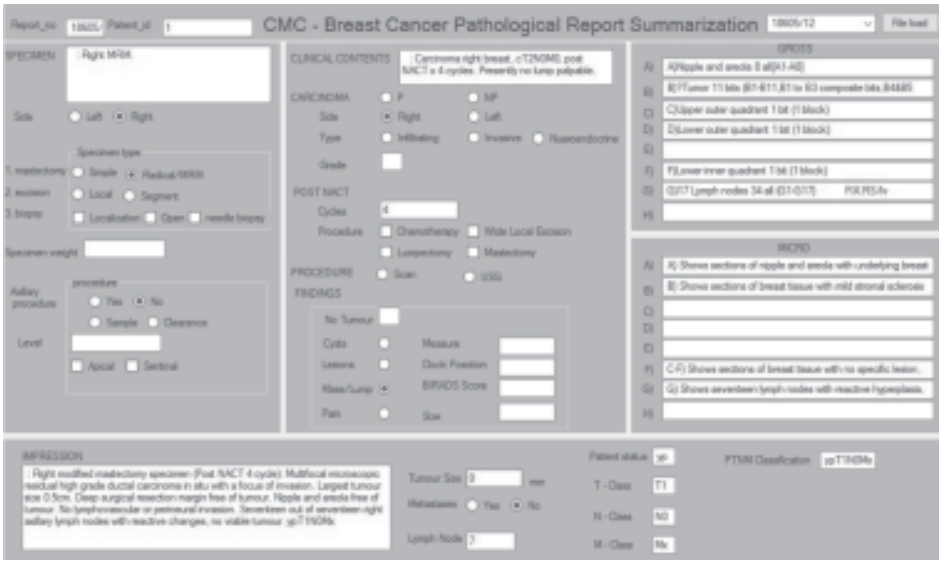

Fig. 3. Sample Individual Patient Report Summary

B. Population based Summary

The visual reporting of cancer staging is based on the distribution of various T-Classification values and N-Classification values. The graphical summarization of these parameters are color-coded and displayed against the number of patients. The advantage of a visual display is easy readability and interpretation. Fig. 4 presents the graphical summarization of the Tumour (T) classification. Among the patient population of 150, T-Classification could not be derived for 6 patients. 20 patients were identified with T0, 4 with T1b, 28 with T1c. 62 patients were classified with T2 and 6 patients had advanced cancer stage of T4b. This summary helps in understanding at what stage of the disease patients reported for pathological analysis. Fig. 5 summarizes the classification of Lymph node N. In the dataset, the lymph node classification is not available for 14 patients, 55 are classified N0, 35 are classified N2a and 13 are classified with N3a.

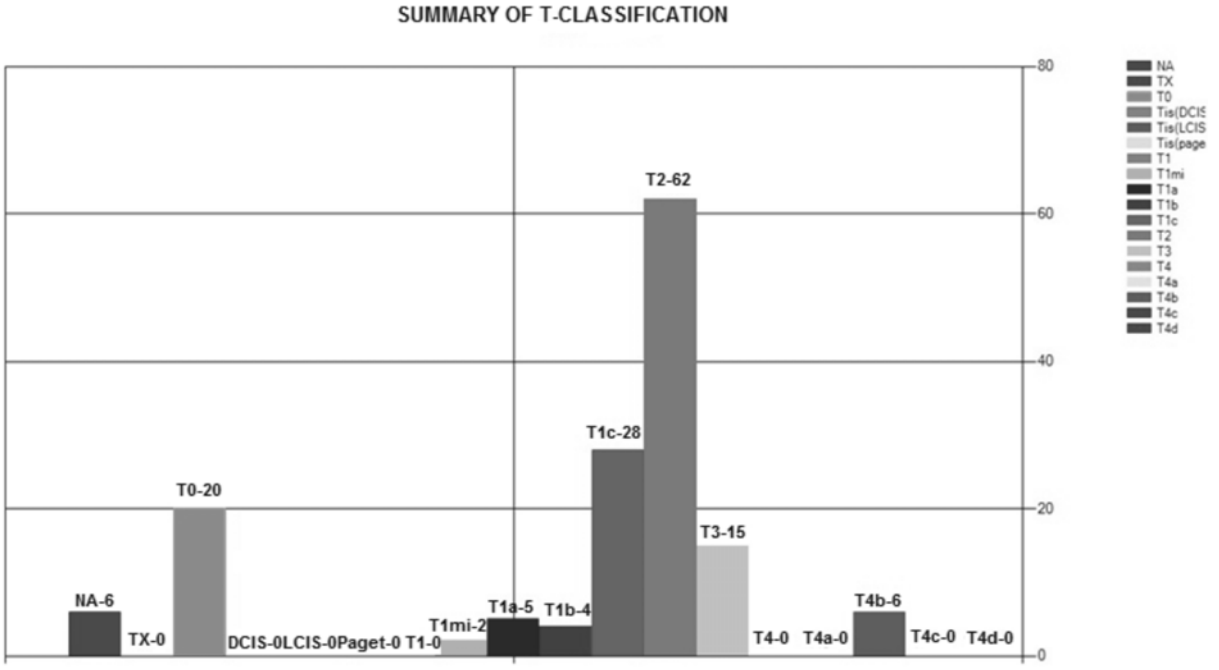

Fig. 4. Summarization of Tumour T-Classification.

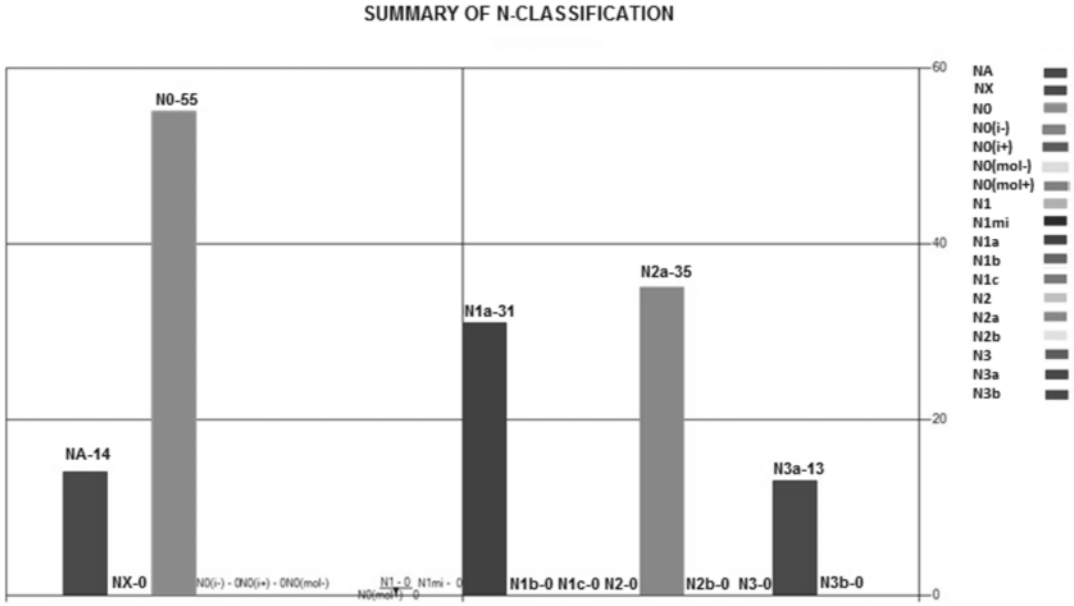

Fig. 5. Summarization of Lymph Node N-Classification

PT_patient_id	PT_T_class	PT_N_Class	PT_M_Class	PT_Stage	PT_report_no
1017/13	T0	N3	M0	IIIC	49
10633/13	T3	N2	M0	IIIA	101
12298/13	T4	N1	M0	IIIB	103
12481/13	T2	N2	M0	IIIA	137
1377/13	T0	N2	M0	IIIA	65
1588/13	T0	N2	M0	IIIA	59
17949/13	T2	N2	M0	IIIA	105
1879/13	T2	N2	M0	IIIA	68
1884/13	T2	N3	M0	IIIC	60
1920/13	T3	N2	M0	IIIA	61
19286/12	T2	N3	M0	IIIC	7
19330/13	T1	N3	M0	IIIC	76
19798/13	T4	N3	M0	IIIC	139

Fig. 6. Patients with Stage III of Breast cancer in 2013.

Fig. 6 presents the summary of cancer stage of the patient population. For 23 patients in the population, stage could not be derived. 16 patients reported in stage II, 40 were identified with stage IIA, 14 with stage IIB, 38 with stage IIIA, 4 with stage IIIB, 13 with stage IIIC. This indicates that majority of the patients in the region reported for diagnosis when they were at stage II or III of cancer. Information such as this provides an authentic base using which Pathologists can propose awareness campaigns and health care measures.

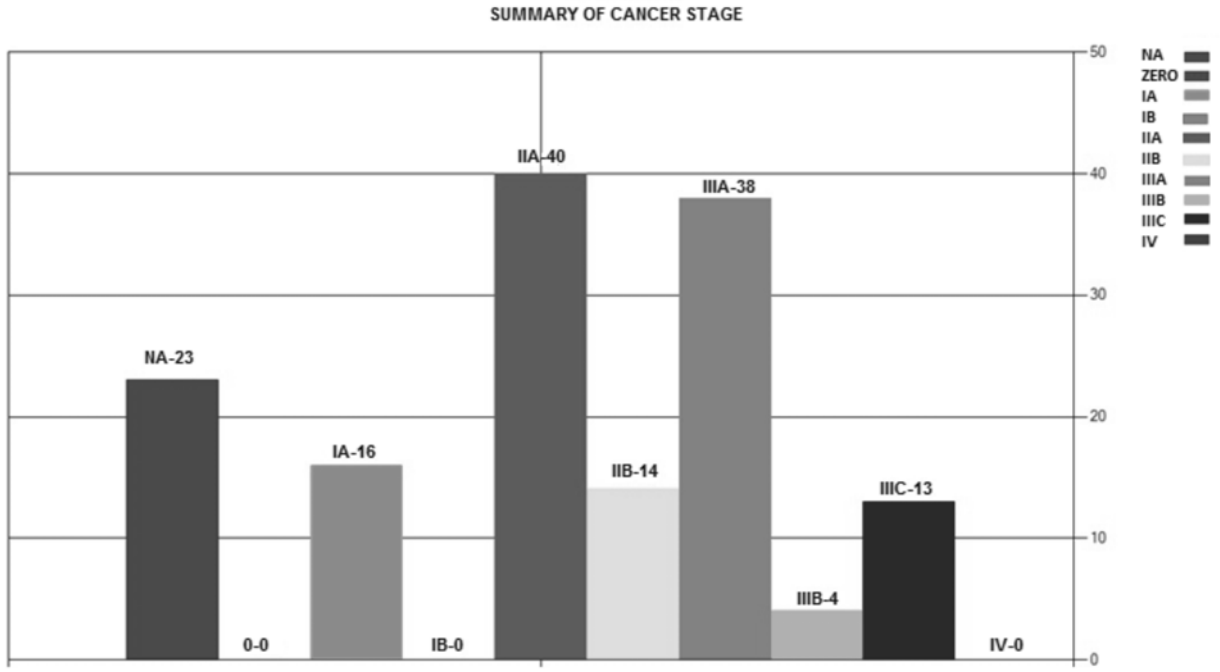

Fig. 7. Summarization of Cancer Stage of all the patients

C. Query-based Summary

The simplest form of summarization done over the years on structured data is query-based summarization. The Medical community would make quick decisions and propose health-care measures in a geographic location or for a particular age group among the patients, if the system provides structured information for querying. In the absence of demographic information of patients in the present dataset, the query summarization focuses on chronological listing and listing based on the cancer stage alone. Summarizing the list of patients who are critically ill with advanced stage of cancer is essential to know the severity of the disease and focus on the specific group of patients. Fig. 7 lists the patients with III stage of cancer, who reported in 2013 respectively. A total number of 42 patients reported at the pre-critical stage.

Listing of patients who reported for diagnosis in a particular year shows whether the disease is progressive among the population or at the decline over the years. Fig. 8 lists the patients with breast cancer, who reported in 2012. Results show that there were a total of 68 patients who reported in 2012, among the patient population.

PT_patient_id	PT_T_class	PT_N_Class	PT_M_Class	PT_Stage	PT_report_no
19286/12	T2	N3	M0	IIIC	7
21178/12	T0	N2	M0	IIIA	20
22495/12	T3	N2	M0	IIIA	35
22502/12	T4	N1	M0	IIIB	36
22520/12	T2	N3	M0	IIIC	42
22540/12	T3	N2	M0	IIIA	37
24037/12	T1	N2	M0	IIIA	30
24208/12	T1	N2	M0	IIIA	39
24323/12	T2	N2	M0	IIIA	47
24726/12	T1	N2	M0	IIIA	40

Fig. 8. Patients affected with breast cancer in 2012

5. CONCLUSION

The essential data to present the summary of individual patients were extracted and the required outputs were successfully generated by the automated system. The deficiency in the performance of the system can be attributed to the following limitations.

- **Totality :** The summarization is not exhaustive due to the fact that demographic details of patients are not available in the dataset. Due to this, graphical summarization and query summarizations could not be performed based on age and regional details.
- **Accuracy :** The visual summarization of cancer stage in the patient population is not a true indication of patient conditions because of the assumption of M-Classification to be M0 in the cancer staging process. However this is not an indication of failure in the automated system in processing the data but a medical limitation of not having the needed information for M-Classification.

- **Dataset Adequacy :** The work was performed with a limited dataset of 150 Pathology reports. Testing the system with numerous reports would provide a deeper and wider spectrum of results required for understanding the patient population.
- **Limited pTNM classes in the dataset :** The dataset does not have all of the possible classifications of T and N. Hence the process is yet to be validated for all classifications of T, N and all possible cancer stages.

  Results indicate that the summarization of individual patient report considerably reduces the Pathologist's time in reading and interpreting the report. Developing a tumour-specific standardized synoptic template or checklist for cancer staging based on CAP protocol in the future, would reduce the reporting time further for faster decision-making.

6. ACKNOWLEDGMENT

The authors would like to thank the Department of Pathology, Christian Medical College and Hospital, Vellore for providing them with the sample data for their study. The authors would also like to acknowledge S. Pradeep Vignesh, student of MCA in the Department of Computer Science, Madras Christian College for his contributions towards the development of the automated system.

7. REFERENCES

1. Joshua C. Feblowitz, Adam Wright, Hardeep Singh, Lipika Samal, Dean F. Sittig, *Summarization of clinical information: A conceptual model*, Journal of Biomedical Informatics, Volume 44, Issue 4, August 2011, Pages 688-699, ISSN 1532-0464. http://dx.doi.org/10.1016/j.jbi.2011.03.008.
2. Donia Scott, Catalina Hallett, and Rachel Fettiplace. *"Data-to-Text Summarisation of Patient Records: Using Computer-Generated Summaries to Access Patient Histories*." Patient Education and Counseling 92.2 (2013): 153–159. *PMC*. Web. 13 May 2016.
3. Ellis DW, Srigley J. *Does standardised structured reporting contribute to quality in diagnostic pathology? The importance of evidence-based datasets*. Virchows Arch. 2015.
4. Lankshear S, Srigley J, McGowan T, Yurcan M, Sawka C, *Standardised synoptic cancer pathology reports - so what and who cares? A population-based satisfaction survey of 970 pathologists, surgeons and oncologists*. Arch Pathol Lab Med., 2013, 137(11):1599-1602.
5. Brierley J, Srigley J, Yurcan M, Li B, Rahal R, Ross J, King ML, Sherar M, Skinner R, Sawka C, *The Value of Collecting Population- Based Cancer Stage Data to Support Decision-Making at Organizational, Regional and Population Levels*. Healthcare Quarterly 16(3):27-33, 2013.
6. Boyd, A.D., Balasubramanian, A., Burton, M., Di, B., Eugenio, Friedman, C., Keenan, G.M., Lugaresi, C., Lopez, K.D., Li, J., Lussier, Y.A., & Macieira, T.G.. *PatientNarr: Towards generating patient-centric summaries of hospital stays*, 2014.
7. Sneh Garg, Sunil Chhillar, *Document Summarization and Evaluation using Knowledge based Super Set Features*, International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 2, March 2015
8. Xue Qin Yu, Roberta De Angelis, Qingwei Luo, Clare Kahn, Nehmat Houssami and Dianne L O'Connell, *A population-based study of breast cancer prevalence in Australia: predicting the future health care needs of women living with breast cancer, BMC Cancer* 2014 14:936, DOI: 10.1186/1471-2407-14-936.
9. Ahmed A. Mohamed, Sanguthevar Rajasekaran, Department of Computer Science & Engineering University of Connecticut Storrs, CT 06268, *Query-Based Summarization Based on Document Graphs*, 2006
10. Hallett, Catalina; Power, Richard and Scott, Donia, *Summarisation and visualisation of e-Health data repositories.* In: UK E-Science All-Hands Meeting, 18-21 Sept 2006, Nottingham, UK.
11. http://www.cap.org/web/oracle/webcenter/portalapp/pagehierarchy/cancer_protocol_templates.jspx
12. Edge SB, Byrd DR, Compton CC, et al., *AJCC Cancer Staging Manual*, 7th ed. New York, NY: Springer, 2010, 00 347-76.
13. http://www.worldometers.info/world-population/india-population/
14. http://www.indiaonlinepages.com/population/india-current-population.html
15. http://www.medicaldaily.com/breast-cancer-rates-rising-india-especially-among-younger-women-261466