Medical & Biological Engineering & Computing (2021) 59:1751–1772
https://doi.org/10.1007/s11517-021-02399-7

ORIGINAL ARTICLE

Information extraction for prognostic stage prediction from breast cancer medical records using NLP and ML

Pratiksha R. Deshmukh[1,2] · Rashmi Phalnikar[1]

Received: 1 December 2020 / Accepted: 1 July 2021 / Published online: 23 July 2021

Abstract

For cancer prediction, the prognostic stage is the main factor that helps medical experts to decide the optimal treatment for a patient. Specialists study prognostic stage information from medical reports, often in an unstructured form, and take a larger review time. The main objective of this study is to suggest a generic clinical decision-unifying staging method to extract the most reliable prognostic stage information of breast cancer from medical records of various health institutions. Additional prognostic elements should be extracted from medical reports to identify the cancer stage for getting an exact measure of cancer and improving care quality. This study has collected 465 pathological and clinical reports of breast cancer sufferers from India's reputed medical institutions. The unstructured records were found distinct from each institute. Anatomic and biologic factors are extracted from medical records using the natural language processing, machine learning and rule-based method for prognostic stage detection. This study has extracted anatomic stage, grade, estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) from medical reports with high accuracy and predicted prognostic stage for both regions. The prognostic stage prediction's average accuracy is found 92% and 82% in rural and urban areas, respectively. It was essential to combine biological and anatomical elements under a single prognostic staging method. A generic clinical decision-unifying staging method for prognostic stage detection with great accuracy in various institutions of different regional areas suggests that the proposed research improves the prognosis of breast cancer.

Keywords Prognostic stage · Anatomic stage · Grade · Hormone receptor · HER-2 · Medical reports · Natural language processing · Information extraction

1 Introduction

Breast cancer is a harmful disease worldwide and makes it a prominent cause of mortality rate in India. A recent study stated that nearly 1.15 million new cancer victims get diagnosed in India per year, and deaths due to cancer are more than 0.78 million [1]. Breast cancer is the most common cancer in Indian women. Identifying stage encounters in breast cancer and its monitoring is essential at the starting stage to decrease patients' mortality rate [2]. In most Indian cancer centers, breast cancer detection uses an unstructured medical report of clinical and pathological records, including information regarding the stage of cancer. In the medical field, the extraction of information significantly identifies informative data from medical records [3, 4]. Different types of medical records like histopathology, immunohistochemistry, and clinical records contain the prognostic stage details. Extracting anatomic and biologic elements from unstructured medical documents is challenging in natural language processing [5, 6]. Extraction of prognostic stage data may help many health professionals to determine appropriate care for cancer victims.

In the prognosis of cancer, the anatomic stage has a significant role [7]. The anatomic stage describes information regarding anatomic factors [8, 9]. It also informs whether cancer is invasive or not [10]. As the cancer stage increases, cancer seriousness increases. Hence, medical specialists must carefully study the unstructured medical records to extract the

✉ Pratiksha R. Deshmukh
deshmukhpratikshar@gmail.com

[1] School of Computer Engineering and Technology, MIT World Peace University, Pune, India 411029

[2] Department of Computer Engineering and Information Technology, College of Engineering, Pune 411005, India

anatomic stage details [11]. The anatomic stage depends solely on TNM elements; TNM is tumor (T), affected lymph node (N) and distant metastasis (M) information. T represents the size and starting position of the tumor, N represents number of infected lymph nodes and M represents whether the lesion initiates expanding to other areas of the body. In contrast, the prognostic effect of tumor biology was not address in the anatomic stage. Hence, there is a need to combine the biological and anatomical elements under one staging method. It says that cancer victims with the higher anatomic stage but desired tumor biology have a good level of care than those with lower anatomic stage and adverse tumor biology. Hence, predicting accurate health status with the anatomic stage and biological elements may improve prognosis with optimized results.

The 8th edition of the American Joint Committee on Cancer (AJCC) presented the prognostic stage, which contained anatomic and biologic factors, and this prognostic stage is verified and preferable to the anatomic stage [12]. In breast cancer treatment, tumor biology is considered an essential element. It helps doctors to recognize the difference in cancer victims' results with different tumor biology but the same anatomic stage of cancer [13]. The additional features like grade, estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status should also use to identify cancer level precisely. A prognostic method can help to assist the specialist in selecting appropriate treatment decisions for cancer patients. This prognostic stage information is also helpful in cancer data analysis of reported patients for survival. Mostly in cancer registries, the data related to cancer information has been recorded manually, which causes an unintended mistake [14]. Very few Indian Population-Based Cancer Registries (PBCRs) aggregate cancer stage information details [11]. There is a need for the extensive study of medical records from pathology labs, clinical records, and immunohistochemistry reports to get collective prognostic details. There are massive differences in the structure of such documents and self-defined medical terms in various cancer centers and institutes. Hence, this is a challenging task in natural language processing to extract prognostic factors from medical documents.

This paper presents a generic clinical decision-unifying cancer staging method to develop a generalized system for prognostic stage details extraction from unstructured medical records of breast cancer. The authors explore the combination of natural language processing, machine learning and rule-based method for developing this research. This paper describes the novel method for information extraction and feature selection based on dictionary concept. To make an extensible or generalized system, this research work uses the data from two different hospitals belongs to different regions with variety in medical record format, the medical terms' description process, the context, and a technique to represent information regarding the prognostic stage.

2 Related work

The rapid progress in machine learning (ML) and Natural Language Processing (NLP) boosts the medical community's interest in using these techniques to improve cancer screening accuracy. Recent research works present importance of machine learning in medical and biological domain [15–22]. Many machine learning and NLP professionals developed a system to automatically identify a cancer TNM stage from pathology records. A coherent framework for extracting data from head and neck cancer pathology reports proposed by Muhammad et al. [23]. This framework implemented using CART software (Classification and Regression Tree algorithm implementation) which is one of the decision tree algorithms from ML methods to support and recommend health professionals for further decisions. The cancer staging framework using ML and NLP designed by Martinez et al. [24] used to analyze colorectal cancer reports of 2 separate organizations and found a 20% discrepancy between program findings and gold standards. They found 84%, 81%, and 91% F-score results of T, N, and M values. Nguyen et al. [25] implemented the TNM stage extraction method with the rule-based mechanism on 718 lung cancer pathology reports and achieved 72%, 78%, and 94% accuracy for T, N, and M values, respectively. Furthermore, they reported inefficacy in early-stage lesion identification results. Rani et al. [26] analyzed TNM detection from 150 pathology records of breast cancer with natural language and SNOMED annotation. SNOMED is a systematized Nomenclature of Medicine which includes medical codes for various medical terms. SNOMED was mostly used in developed countries. They noticed that documents with SNOMED annotation performed better for TNM detection, but they mainly focused on the impression section.

Warner et al. [27] presented a new method to identify TNM values from 2,327 lung cancer pathology records of single hospital and reached stage accuracy with 72%. Martinez and Yue Li [28] introduced a framework using ML for extracting important terms from 217 colorectal cancer pathology reports with minimal clinical expertise guidelines. McCowan et al. [29] presented TNM categorization from 700 lung cancer histology records using ML techniques and could get 74% and 87% accuracy for T and N, respectively. Johanna et al. [30] implemented an automated model for stage identification from 150 pathology records of breast cancer sufferers. They focused on only the impression section of pathology records; they did not concentrate on the remaining portions of the documents containing essential details about the stage of cancer. Their work could get 76%, 66%, and 72% accuracy for T, N, and stage, respectively. Nguyen et al. [31] introduced a multi-class categorization methodology, categorized TNM from 710 pathology reports of lung cancer, and achieved positive results. Rajaguru and Prabhakar [32] implemented a breast cancer categorization model with logistic regression to detect TNM from structured data. McCowan et al. [33] developed a model to identify stage details using a support vector machine (SVM) from ML methods. They presented a model based on the specific dataset of 710 lung cancer pathology records with similar structure, which belongs to a single institute.

Recent studies show a significance of machine learning in medical and biological domain for prognosis and research purpose [15–22]. Some studies present a key role of machine learning methods in data extraction from pathology reports and cancer stage detection [23, 24, 28, 29, 31, 33]. Machine learning methods like decision tree, support vector machine, and naïve bayes are relevant in breast cancer prediction and prognosis [8, 10, 34, 35].

According to the analysis, most previous studies used a single hospital's medical records, which has a unique formatting or collection style. Some of them considered only a subsection of records with structured or semi-structured inputs. Moreover, many studies focused only on TNM stage extraction from histopathology reports with and without sub-values of T, N, and M. Earlier research did not extract additional prognostic elements like grade, ER status, PR status, and HER2 status to predict prognostic stage from the number of free-text or unstructured medical records. Additional prognostic elements should be extracted from medical reports to identify cancer and get the exact measure of cancer to improve care quality [36]. During this extraction process, every prognostic factor's accuracy is vital because the wrong result in the single prognostic factor affects prognostic stage result.

This research aims to obtain collective information on the cancer stage from medical documents. Past work focused on the machine learning approach or the framework based on symbolic rules. This paper discusses a new method of prognostic stage detail extraction from unstructured medical records. This study relates a generic clinical decision-unifying staging method to the most reliable prognostic stage information extraction from different hospitals' breast cancer patient's medical records. Instead of analyzing only a subsection of the record, this research examines the entire document to gather valuable knowledge about the stage of cancer.

The proposed research uses a novel combination of NLP and machine learning techniques to extract prognostic stage information from unstructured medical records. NLP includes named entity recognition (NER) uses for feature extraction which is based on dictionary concepts. IE uses to extract information from records which includes confirmation word method for concrete filtering and range analysis during pattern matching and searching for precise results. Finally, machine learning uses for prediction of cancer stage from extracted information. Most of the previous studies were limited for anatomic factors' extraction. This study includes anatomic and biologic factors' extraction, and anatomic and prognostic stage prediction, thus ensuring progress and positive results in an accuracy of extracting prognostic stage details.

3 Workflow method

It is an essential requirement to combine biological and anatomical elements under a single prognostic staging method. TNM staging guidelines of the 7th edition of AJCC are referred for anatomic stage detail extraction, and the prognostic staging guidelines of AJCC 8th edition are referred for prognostic stage detail extraction [36]. This study extracts the most reliable prognostic stage information of breast cancer from medical records. Figure 1 shows the architecture of the proposed study of prognostic stage detection. Firstly, the collection of medical reports from two cancer hospitals and their labs is used as input. Section 4.2 gives details regarding data collection.

Initially, the medical reports get processed in the data preprocessing stage. After that, data standardization is performed on reports with three different stages, viz. measure standardization, numerical values, and numeric representation. Using information extraction, the relevant information gets extracted from these reports in terms of prognostic factors. Finally, the machine learning classifiers are applied to prognostic factors for prognostic stage prediction.

3.1 Input to system

The pathological and clinical reports are given as input to the system to extract anatomic and biologic factors' information; these reports are represented in natural language. Details regarding the dataset are provided in Sect. 4.2.

3.2 Report preprocessing

The preprocessing of the data is an important task to sort out in the document filtering process. Different types of reports have various subsections for differentiating the medical reports of cancer patients—many techniques and procedures are used for preprocessing.

- Section segmentation—The pathological and clinical reports have different sections for various remarks. Some previous studies [26, 30] have worked on a particular section like the impression section of pathology records. This study considers every section of pathological and clinical reports such as clinical history, specimen, microscopic, macroscopic, impression, and gross sections. Sentence segmentation distinguishes the contents of every sentence for processing on it.

Figure 2 shows a snap of the histopathology record which consists of different sections. Regex pattern search method is used to search section heading, and then, segmentation is performed accordingly. After that, the report gets to proceed to text processing.

- Preprocessing of text with normalization, tokenization, and stemming functions used for preprocessing of medical reports. Natural language toolkit (nltk) is used to perform

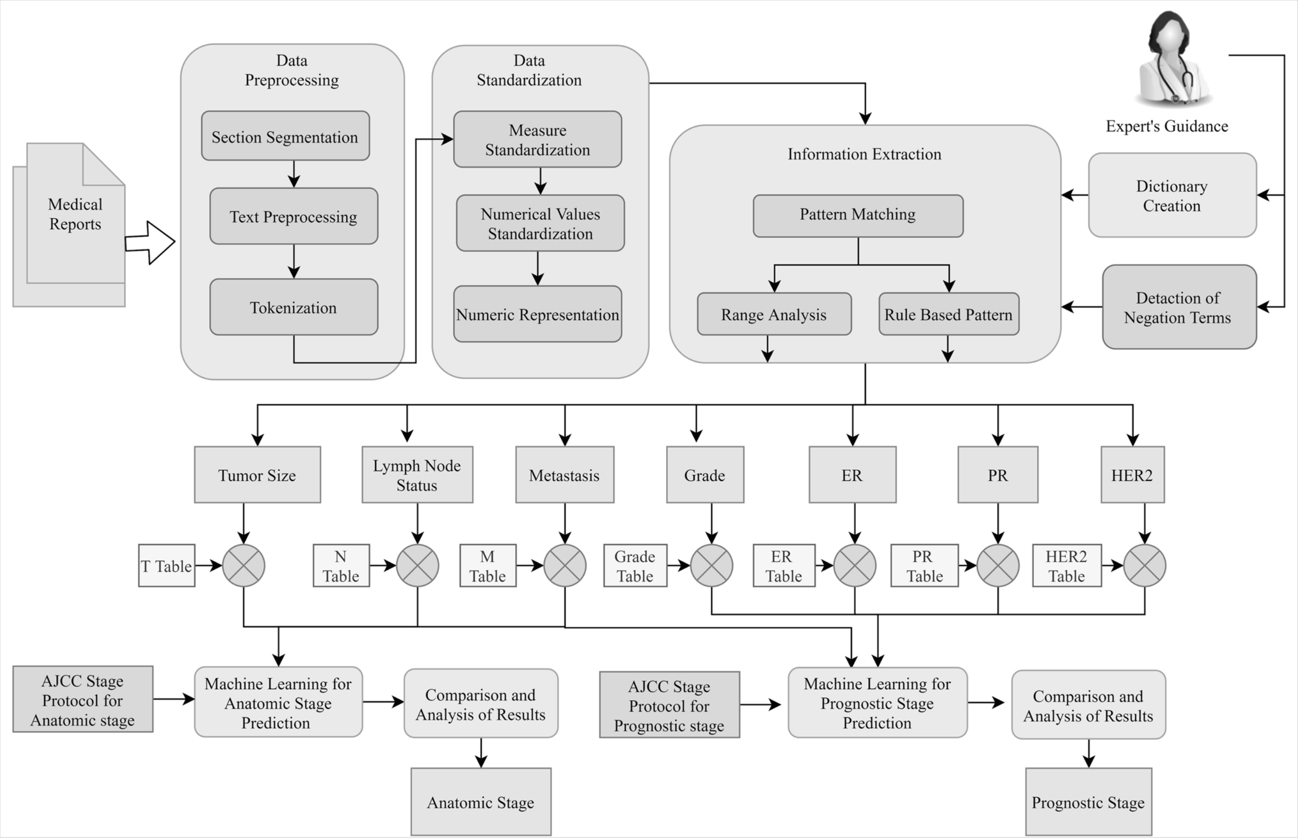


Fig. 1 Architecture of proposed study of prognostic stage detection

all text preprocessing functions because NLTK is a platform for building a Python program for NLP, including text processing libraries. Packages like *nltk.tokenize*, *nltk.stem*, *nltk.corpus* used for text data filtering functions.

3.3 Standardization of data

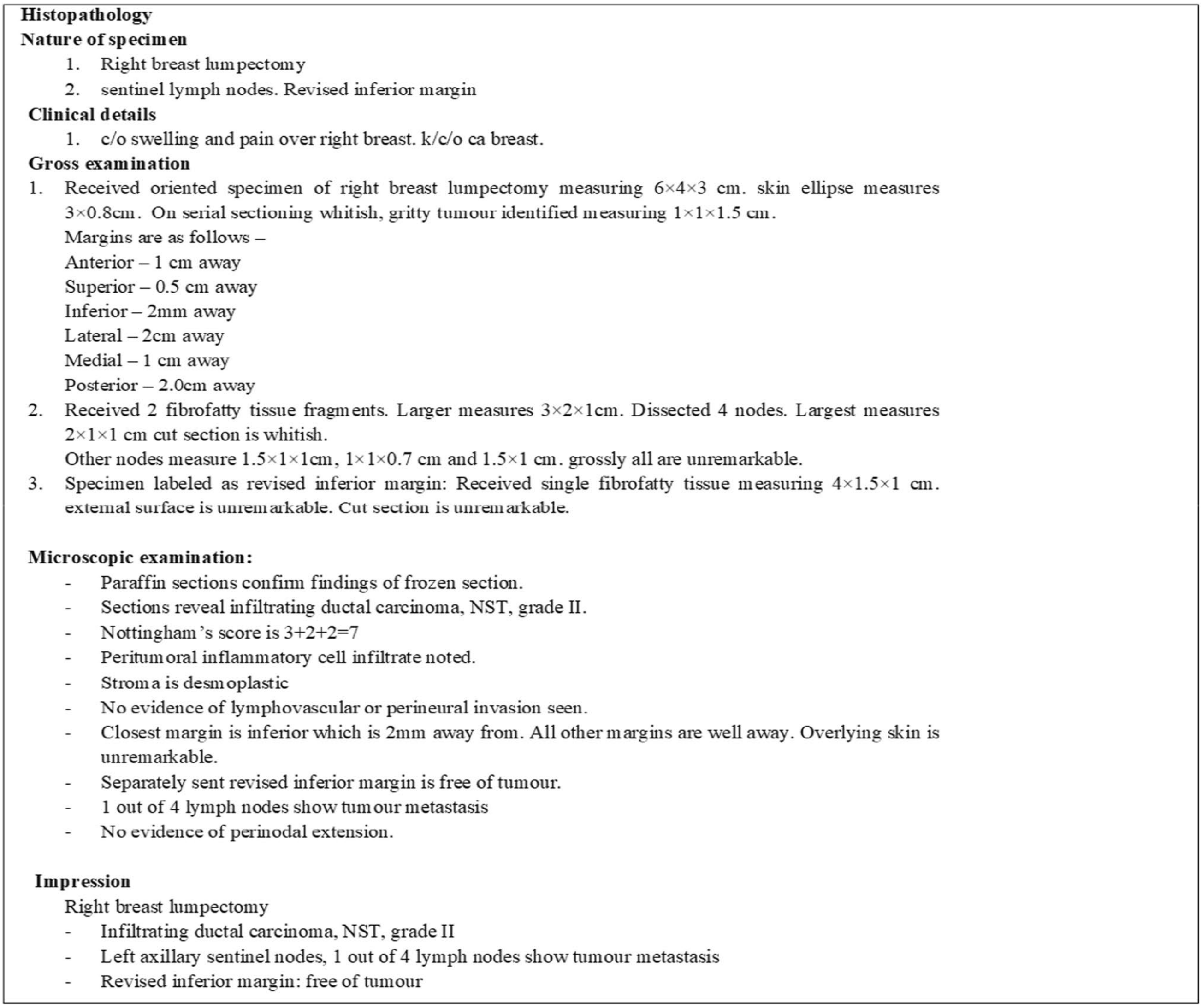
Various medical professionals from different organizations have different writing styles and representations of medical records; this module conducts the standardization of such distinct formats received from different report collections. More details about these processes are explained below:

- Measure standardization—This feature is used to standardize measures. For example, lesion size can express in several ways of centimeters and millimeters like ‘25 × 10 mm in size’, ‘7 × 6 × 5 cm noted’. This function converts all values in one unit into specific units (e.g., centimeter to millimeter). This feature standardizes all these measures in a single form. It uses to find out exact size of tumor.
- Value standardization—This feature standardizes values that are expressed in different forms. For example, to represent grade-1, different forms can be used as ‘grade 1’, ‘Grade-I’, ‘GR 1st’, and ‘low grade’, and all these will be converted into a value 1 to represent the grade output.
- Numeric representation—This feature shows numeric. For example, the number of affected lymph nodes is described in several ways like ‘2/10’, ‘2 of 10’, ‘two out of ten’, ‘2 out of 10’, and ‘2 out of ten’; hence, numeric will be extracted to get the affected lymph node value.

In this function, the textual mentions of numbers are converted into their numeric representation using normalization and regular expressions such as <numeric1> out of <numeric2>. After that, numeric 1 and 2 get extracted, and smaller numeric get selected. Hence, all numeric will be extracted to get the affected lymph node value.

3.4 Creation of dictionary

- Term dictionary—A huge diversity was found in medical record format, medical terms’ context, and description [37]. Representation of prognostic stage information can vary at various cancer centers and institutions. Hence, there is a need to standardize these terms using a predeveloped dictionary in the system. A dictionary is created, and while creation, a Unified Medical Language System [38] is referred to knowing the semantic details regarding medical terms. It also identifies contextual medical terms from medical documents. Some of the medical terms such as ‘tumor’ described in several ways like ‘suspicious-mass’, ‘irregular-lump’, ‘malignant-lesion’, and ‘ill-defined-tumour’. Hence, ‘medical term dictionary’ is developed for such medical concepts related to cancer with different descriptions.



Histopathology

Nature of specimen

1. Right breast lumpectomy
2. sentinel lymph nodes. Revised inferior margin

Clinical details

1. c/o swelling and pain over right breast. k/c/o ca breast.

Gross examination

1. Received oriented specimen of right breast lumpectomy measuring 6×4×3 cm. skin ellipse measures 3×0.8cm. On serial sectioning whitish, gritty tumour identified measuring 1×1×1.5 cm.
   Margins are as follows –
   Anterior – 1 cm away
   Superior – 0.5 cm away
   Inferior – 2mm away
   Lateral – 2cm away
   Medial – 1 cm away
   Posterior – 2.0cm away
2. Received 2 fibrofatty tissue fragments. Larger measures 3×2×1cm. Dissected 4 nodes. Largest measures 2×1×1 cm cut section is whitish.
   Other nodes measure 1.5×1×1cm, 1×1×0.7 cm and 1.5×1 cm. grossly all are unremarkable.
3. Specimen labeled as revised inferior margin: Received single fibrofatty tissue measuring 4×1.5×1 cm. external surface is unremarkable. Cut section is unremarkable.

Microscopic examination:

- Paraffin sections confirm findings of frozen section.
- Sections reveal infiltrating ductal carcinoma, NST, grade II.
- Nottingham's score is 3+2+2=7
- Peritumoral inflammatory cell infiltrate noted.
- Stroma is desmoplastic
- No evidence of lymphovascular or perineural invasion seen.
- Closest margin is inferior which is 2mm away from. All other margins are well away. Overlying skin is unremarkable.
- Separately sent revised inferior margin is free of tumour.
- 1 out of 4 lymph nodes show tumour metastasis
- No evidence of perinodal extension.

Impression

Right breast lumpectomy

- Infiltrating ductal carcinoma, NST, grade II
- Left axillary sentinel nodes, 1 out of 4 lymph nodes show tumour metastasis
- Revised inferior margin: free of tumour

Fig. 2 Sample histopathology record with different sections

- Negation detection—This module identifies negation concepts in medical records like 'no evidence', 'not suspicious', 'absences', 'unaffected', and 'free from malignancy'. Negex method was used by Chapman et al. [39], which determines medical terms from text reports associated with negation. In this study, the dictionary implemented all terms associated with negation terms and possibility terms and this helps in minimizing the false positive count.
- Variations in spelling—It is observed that there are variations in medical records and the number of medical terms described with spelling variations. For example, the medical term 'tumor' is also described as 'tumour'; similarly, the term 'estrogen' is also described as 'oestrogen' in some records; hence, the proposed work considers all these variations in the spelling of medical terms.

Dictionary creation uses Named Entity Recognition (NER) based on dictionary concepts. Dictionary is created using Python where dictionary elements are stored in the form of "dict = {'key': 'value'} pairs" and can be accessed using the key's name and index position. Dictionary elements are changeable and ordered without duplicates. Different methods such as 'get (), keys (), values (), items (), update ()' are used to perform operation on the dictionary. While creating a dictionary, authors also focused on issues like abbreviations, grammatical mistakes, context learning, phrase incorrectness, case sensitivity or insensitivity, and typing error.

3.5 Extraction of prognostic factors

This function extracts all prognostic factors from medical records of breast cancer using a novel technique. Machine learning algorithms can be used to extract prognostic factors; but these techniques need huge data, and prognostic factors that have a small occurrence (for example, anatomic factors of stage-0 or stage-IV) may not be adequate for these techniques. Furthermore, the extensible or generic nature of the system may be limited due to diversity in reports. In contrast, this study performs effectively on small data also for more generic system implementation. This section presents an information extraction process to extract prognostic factor details from medical records. The following algorithm explains prognostic factors' extraction.

```
Input: set D {D: $d_1$, $d_2$, ……...$d_n \in$ D} (D is set of medical records)
Output: TNM attributes
for all d ϵ D do
function record_preprocessing D
        R ← tokenize, stop_word, normalize, stemming,
        R ← segmentation, data standardization,
return (R)
function get_T
        apply pattern _search ($t_i \in$ T_dictionary)
        get_position → index($t_i$)
        perform range_analysis (T_conf_word ϵ $C_T$_ dictionary)
                  »where range < len (T_statement) with L_bound and U_bound.
        extract T_info ← (location, dimension immediate to T_conf_word)
        T_size ← (T_info, measure_std, L_R_preference)
        T_h ←set_priorities (h_priority ← high_dimen)
        T_extract ← (T_h, T_size)
        if (n (T_extract)> 1)
                T_extract ← select T_h
        end if
        if (T_extract)
                search (T_extract ϵ T4_dictionary || T_extract ϵ Tis_dictionary|| T_extract ϵ T0_T3_info || T_extract ϵ
                T_negation_dict)
                T_extract ← T_info
        end if
        T_final ←match (T_extract to T_AJCC)
return (T_final)
function get_N
        apply pattern _search ($n_i \in$ N_dictionary)
        get_position → index($n_i$)
        perform range_analysis (N_conf_word ϵ $C_N$_ dictionary)
                              »where range < len (N_statement) with L_bound and U_bound.
        extract N_info ← (N_conf_word, measure, N_neg ϵ N_neg_dict, N_pos ϵ N_pos_dict)
        N_value ← (minimum(value) immediate to measure)
        set_priorities (N_extract)
        if (N_extract ϵ N_negation_dict)
            1 ← h_priority
        else
            2 ← l_priority
            N_extract ← minimum(N_value)
        end if
        N_final ←match (N_extract to N_AJCC)
return (N_final)
function get_M
        apply pattern _search ($m_i \in$ M_dictionary)
        get_position → index($m_i$)
        perform range_analysis (M_conf_word ϵ $C_M$_ dictionary)
                            »where range < len (M_statement) with L_bound and U_bound.
        extract M_info ← (M_conf_word, M_neg ϵ M_neg_dict, M_pos ϵ M_pos_dict)
        set _priorities (M_extract)
        if (M_neg)
            1← h_priority
        else if (M_pos)
            2←l_ priority
        end if
        M_final ←match (M_extract to M_AJCC)
return (M_final)
```

Algorithm description—the algorithm represents the information extraction process from unstructured medical reports to get prognostic factors. Initially, medical reports are preprocessed. After that, pattern matching and rule-based technique with range analysis functions are performed to search information related to each prognostic factor. Section 3.5.1 explains working of this function. While applying pattern matching and rule-based technique, medical terms for prognostic factors are matched with their respective contents in dictionary. Section 3.4 explains dictionary concept. During feature selection process, confirmation keywords are searched for concrete filtering and priorities are set as per importance and weight of all medical terms related to prognostic factors with positive and negative findings. Finally, extracted prognostic factors' information matches with AJCC information to get value of each prognostic factor as an output.

3.5.1 Pattern matching and rule-based technique

In this study, pattern matching and rule-based techniques are used to search a particular match according to their importance and weight from a developed medical term dictionary. Positive and negative terms in the information extraction process are set with priorities, e.g., during T extraction, priorities are set towards terms associated with the tumor, followed by measures like the dimension of a tumor and its occurrence. Most of earlier research [24, 28] proposed regex patterns (regular expression) to extract TNM information like tumor size and location, affected lymph node, and distant metastasis information. However, concrete filtering with such a pattern for prognostic factors extraction is quite tricky, e.g., any numeric with dimension can be lesion size, lesion position, different breast sizes, item size, or margin distance. The proposed study performs concrete filtering and separates all these aspects to extract the lesion size.

The following example shows how the feature selection process takes place. "Received oriented specimen of left breast lumpectomy measuring 12 × 7 × 5 cm. Overlying skin with nipple-areola measures 11 × 7 cm, grossly is unremarkable. On serial sectioning, a well-circumscribed whitish, gritty tumor measuring 3.5 × 3 × 3 cm was identified." In this example, numeric with dimensions are given three times, and out of them, one is the size of tumor; hence, feature selection process to extract lesion size is:

1. Apply dictionary search for term t_i ϵ tumor. (where $T\_dictionary = \{‘t_1’, ‘t_2’, ‘t_3’ \ldots\ldots ‘t_n’\}$)
2. Perform range analysis with upper bound and lower bound (where $range < len(T\_statement)$).
3. Apply confirmation keyword search immediate to medical term tumor in given range.
4. Extract features as term related to tumor, confirmation keyword, measure, dimension, and location.

Thus, this function includes range analysis, dictionary concept, priorities of terms, concrete filtering and feature selection.

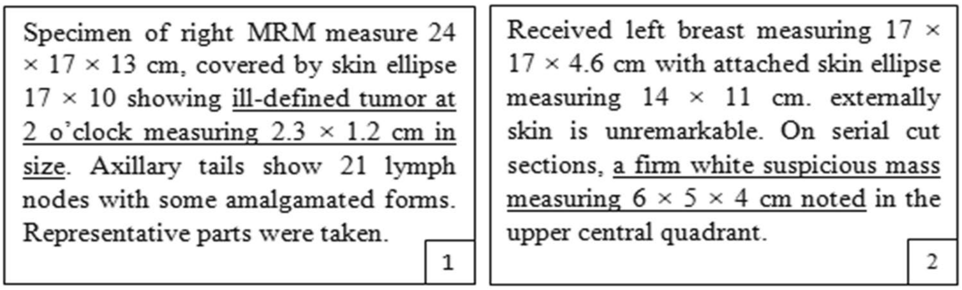
As mentioned in Section 3.5, the proposed algorithm extracts each prognostic factor, including tumor details, affected lymph node, metastasis status, grade, HER-2, and hormonal status information from medical reports. For better understanding, Sections 3.5.2 to 3.5.7 describe extraction of each prognostic factor with sample examples.

3.5.2 Tumor extraction

This function is used to detect the lesion size and location of the tumor from pathological records. As an example, snaps of records containing tumor information are given in Fig. 3. As explained above, diversity can be found in pathological records in the representation of tumor details.

The following are the few examples found in unstructured pathological reports to demonstrate the extracted tumor information.

(a) 'Suspicious mass at 2 o′clock measuring 25 × 10 mm in size.'
(b) 'A firm white irregular lump was measured 7 × 6 × 5 cm noted.'
(c) 'An ill-defined tumour measuring 3.8 × 3 × 2.8 cm noted.'
(d) 'A large irregular hypoechoic lesion of size 33 × 23 × 21 mm is seen.'



Specimen of right MRM measure 24 × 17 × 13 cm, covered by skin ellipse 17 × 10 showing ill-defined tumor at 2 o'clock measuring 2.3 × 1.2 cm in size. Axillary tails show 21 lymph nodes with some amalgamated forms. Representative parts were taken.

1

Received left breast measuring 17 × 17 × 4.6 cm with attached skin ellipse measuring 14 × 11 cm. externally skin is unremarkable. On serial cut sections, a firm white suspicious mass measuring 6 × 5 × 4 cm noted in the upper central quadrant.

2

Fig. 3 Snap of records containing tumor information

There is diversity in the representation of medical terms, variation in spelling, and variation in measure from the above examples. Extracted tumor information matches with AJCC protocol details and shows the output of tumor values in terms of ‘Tx, Tis, T1, T2, T3, and T4’. ‘Tx’ is considered if the value of T cannot be checked. ‘Tis’ if carcinoma in situ, ‘T1’, ‘T2’ are used if the values of $T \leq 2$ cm and in between $T > 2$ cm and $T \leq 5$ cm, respectively, while T3 is used when $T \geq 5$ cm. T4 is used for an extension to skin or chest wall with any size of T.

3.5.3 Lymph node extraction

This function extracts the affected lymph node (LN) value. Pathological records are used to extract lymph node information. Pathological records have diversity in the representation of lymph node details. As an example, snaps of documents containing lymph node information are given in Fig. 4.

The following are some examples to illustrate the extracted lymph node information from unstructured pathological reports.

(a) ‘Metastasis to 8 of 22 lymph nodes (8/22)’
(b) ‘Lymph node is free of tumor’
(c) ‘Negative for lymph node metastasis (0 out of 5 lymph nodes)’
(d) ‘All 18 axillary lymph nodes are negative for tumor’

By referring AJCC manual for N’s details, the output of LN is detected as ‘Nx, N0, N1, N2, and N3’. Similarly, ‘Nx’ stands when regional lymph node (LN) cannot be checked. If there is no regional LN metastasis, then ‘N0’ is considered and extends to ‘N1’ to ‘N3’ as number of affected lymph node increases.

3.5.4 Metastasis status extraction

This function is used to perceive the information about the extension of cancer in other body parts of the victims. Clinical (mammography and ultrasonography) records are used to extract metastasis (M) information. Extracted metastasis information gets match with AJCC protocol for M details and shows M values’ output in the form of ‘Mx, M0, and M1’. ‘Mx’ stands when distant metastasis cannot be checked. If there is no distant metastasis, then ‘M0’ is considered and ‘M1’ if there is distant metastasis.

3.5.5 Grade extraction

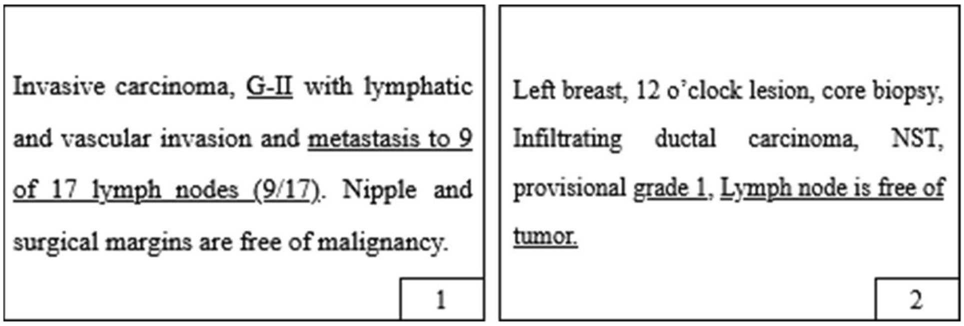
Grade values differentiate cancer cells and normal cells [40]. Grade values are extracted from pathological reports. As an example, Fig. 3 represents snaps of records containing grade information.

Grade values divide into three parts, viz. grades 1, 2, and 3. These entire grades can represent different forms in various medical reports, e.g., ‘grade 3’ can represent in different ways like ‘G 3,’ ‘grade III,’ ‘high grade,’ ‘poorly differentiated,’ and ‘fast growing’. Hence, three standard grade terms are considered to extract different grade information from all reports. This function shows the output of grade values in ‘1, 2, and 3’.

3.5.6 Hormone receptor status extraction

Recognizing the status of hormone receptors (HR) is essential [41] to decide the alternative treatment. If cancers have hormone receptors, they are ‘HR positive’; otherwise, they are ‘HR negative’. Estrogen receptor (ER) and progesterone receptor (PR) are the two types of hormone receptors and can extract from immunohistochemistry reports. As an example, a snap of records containing ER, PR, and HER-2 information is given in Fig. 5.

If cancer has an estrogen hormone receptor, it classifies as ‘ER positive’; else, it is ‘ER negative’. Similarly, if cancer has a progesterone hormone receptor, it classifies as ‘PR positive’; else, it is ‘PR negative’. ER and PR classification’s borderline has been counted positive since it has now been described as higher than 1% positive. Hence, it was taken as positive with 1% or higher than lesion cells. ER and PR scores can also be represented in Allred score ranges between 0 and 8, representing a negative value for 0 and strongly positive for 8. After extracting ER and PR information, it shows the output of ER and PR status as either ‘positive’ or ‘negative’.



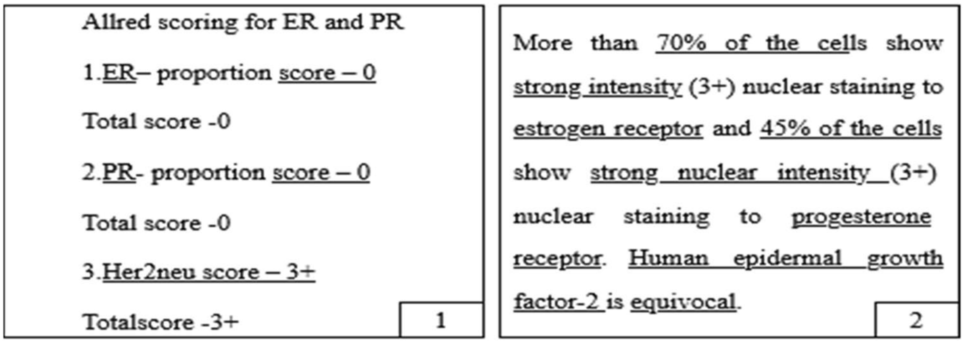
Invasive carcinoma, G-II with lymphatic and vascular invasion and metastasis to 9 of 17 lymph nodes (9/17). Nipple and surgical margins are free of malignancy.

1

Left breast, 12 o’clock lesion, core biopsy, Infiltrating ductal carcinoma, NST, provisional grade 1, Lymph node is free of tumor.

2

Fig. 4 Snap of records containing lymph node and grade information



Allred scoring for ER and PR

1.ER– proportion score – 0

Total score -0

2.PR- proportion score – 0

Total score -0

3.Her2neu score – 3+

Totalscore -3+

1

More than 70% of the cells show strong intensity (3+) nuclear staining to estrogen receptor and 45% of the cells show strong nuclear intensity (3+) nuclear staining to progesterone receptor. Human epidermal growth factor-2 is equivocal.

2

Fig. 5 Snap of records containing ER, PR, and HER-2 information

3.5.7 HER-2 extraction

A gene known as human epidermal growth factor receptor 2 is also an essential factor in breast cancer progression. HER2 receptors usually help to regulate how a cell of the breast develops, splits, and recovers. Because of the excess in the HER2 gene, breast cells generate many HER2 receptors, which results in uncontrolled growth and splitting of breast cells [41].

HER2 status information is extracted from immunohistochemistry reports. HER2 value has been taken as negative if its score is between 0 and 1 + . Value is considered borderline if the score is 2 + and the value has taken as positive if it is 3 + or higher than 10% of lesion cells. This function shows the output of HER-2 status in terms of positive and negative.

Immunohistochemistry records are used to extract hormone receptor status and HER-2 status information. Immunohistochemistry records have diversity in representation of hormone receptor status and HER-2 status details. Following are some examples to illustrate the extracted hormone receptor status and HER-2 status information from unstructured immunohistochemistry reports.

(a) ‘5/8 for oestrogen receptor (ER), and 7/8 for progesterone receptor (PR)’
(b) ‘80% of the cells show strong intensity (3 +) nuclear staining to the estrogen receptor, and 55% of the cells show strong nuclear intensity (3 +) nuclear staining to the progesterone receptor’
(c) ‘hormonal status (Er + + + , PR + + +) and HER-2 equivocal’
(d) ‘ER – proportion score – 0, PR—proportion score – 0, Her2neu score – 3 + ’

For ER information, few reports have used the term ‘estrogen’ or ‘oestrogen’, whereas some have used ‘ER’. In some reports, they have used the Allred score, whereas the percentage of the cell has been used in some reports; similarly, PR and HER-2 information is represented in different ways. These show the complexity of NLP tasks.

Information extraction also includes variations in statement descriptions because statement descriptions in medical records may vary from institute to institute. Buckley et al. [42] described a number of ways to represent medical terms in the statement, and their work has shown that this is a complex and challenging task in NLP. For example, according to observations, DCIS can be represented in several ways like ‘DCIS’, ‘Ductal_Carcinoma_in_Situ’, and ‘Carcinoma_in_Situ_with_Ductal_features’. The proposed work uses a technique of pattern matching [43, 44] to fix such issues. This technique searches for the pattern ‘ductal’ and then extracts information; it applies to scrutinize and match in the forward and the backward direction of pattern ‘ductal’. During this information extraction process, priorities are set to associate terms, positive terms, and negative terms such that it reduces false positive counts, although there are variations in the statement description.

3.6 Machine learning approach

Machine learning model—decision tree (DT), gaussian naïve bayes (GNB), and linear support vector machine (SVM) are used for classification, because these are mostly used classifiers in medical document classification [8, 10, 23, 34, 35].

Parameter setting is tuned for decision tree, gaussian naïve bayes and support vector machine to select optimal model architecture. Hyperparameters’ tuning is done by Grid Search method. This study has used GridSearchCV function of scikit-learn to tune hyperparameters.

```
1. paramet_tune_dt = GridSearchCV(DecisionTreeRegressor(), param_grid = {'max_depth' : [2,4,6,8,10,12], 'criterion' : ['mse', 'mae'], 'splitter' : ['best', 'random'], 'random state' : [0,1]}, cv = 5)
paramet_tune_dt.best_params
2. paramet_tune_svm = GridSearchCV(SVC(), param_grid = {'C': [1, 10, 100, 1000], 'kernel': ['linear']}, cv = 5)
paramet_tune_svm.best_params
3. paramet_tune_gnb = GridSearchCV(GaussianNB(), param_grid = { 'var_smoothing' : [1e-2,1e-3,1e-4,1e-5, 1e-6, 1e-7, 1e-8, 1e-9]}, cv = 5)
paramet_tune_gnb.best_params
4. Used optimal values of parameters of each ML method to get best model creation.
```

Machine learning (ML) model has been applied to the extracted prognostic factors for predicting anatomic and prognostic stage. The feature selection process is performed during prognostic factor extraction to reduce feature space. Feature selection includes dictionary creation, range analysis, negation finding, contextual finding, and regular expression with concrete filtering.

F1-score computes to compare the results obtained by each classifier (mentioned in Sect. 4). Classifier choice also depends on feature-space dimensionality. SVM classifiers perform precisely in high-dimensional feature space, whereas DT model performs better in low-dimensional feature space. SVM (black-box model) does not describe the classification reason in the form of a decision rule, whereas DT (white-box model) has an explanatory system. The decision tree has the better performance among three classifier models. After extracting all prognostic factors using the NLP method, this section describes anatomic and prognostic stage prediction using a decision tree.

3.6.1 Anatomic stage prediction

Anatomic stage prediction requires extracted T, N, and M factors. The 7th edition of AJCC is studied for anatomic stage detail prediction. Table 1 shows the details of anatomic stage classification according to the 7th AJCC manual. The anatomic stage varies from stage-0 to stage-IV as per changes in T, N, and M values. Our previous work [9] shows details regarding tumor (T) extraction, lymph node (N) extraction, and metastasis (M) status extraction from unstructured medical records. The T, N, and M factors are extracted from both regional areas' medical records. Then, machine learning algorithms are applied to the extracted factors to predict the anatomic stage.

3.6.2 Prognostic stage prediction

The prognostic stage table is created from prognostic factors according to the directives given in the 8th AJCC manual [45]. With the machine learning approach, the prognostic stage is identified as per the findings of all prognostic factors. A supervised machine learning technique is used to predict the prognostic stage from all extracted prognostic factors. Table 2 shows brief information on prognostic stage classification using prognostic factors according to the directives given in the 8th AJCC manual. The prognostic stage varies from stage-0 to stage-IV as per changes in each prognostic factor (T, N, M, grade, ER, PR, HER2 values). The subfields of each prognostic stage, like IA and IB, also have been extracted to achieve precise results.

Table 1 Anatomic stage classification according to the 7th AJCC manual

Sr. No	TNM value	Anatomic stage	Sr. No	TNM value	Anatomic stage
1	TisN0M0	0	10	T1N2M0	3A
2	T1N0M0	1A	11	T2N2M0	
3	T0N1miM0	1B	12	T3N1M0	
4	T0N1M0	2A	13	T3N2M0	
5	T1N1M0		14	T4N0M0	3B
6	T2N0M0		15	T4N1M0	
7	T2N1M0	2B	16	T4N2M0	
8	T3N0M0		17	AnyTN3M0	3C
9	T0N2M0	3A	18	AnyTAnyNM1	4

3.6.3 Decision tree model

A decision tree provides remarkable outcomes for assessing breast cancer survival and prediction of prognosis [34, 35, 46]. In this research, the ML algorithm was used to predict the prognostic stage from extracted prognostic factors. A decision tree is one of the supervised techniques, easy to understand outcomes of the tree model using the visualization part for analyzing data. A decision path gives adequate knowledge to recognize key parameters in decision-making.

This study used jupyter notebook of anaconda framework with scikit-learn to implement decision tree (DT) for the prediction of the anatomic stage and the prognostic stage. Sklearn is used for decision tree implementation in Python. Table 3 gives details regarding the specifications of the ML model. After extracting prognostic factors from medical reports, a structured database of prognostic factors is given as input to the DT model. 'train_test_split' method from the 'model_selection' library is applied to data. A 'fit' method is used to train the algorithm while a 'predict' method is used for prediction. Mean squared error metrics is used for performance evaluation. DT is divided, based on anatomic factors' values T, N, and M stages at the initial level. The next level division is based on biologic factors' value.

Figure 6 shows the DT model's snaps for anatomic stage prediction, and Fig. 7 shows snaps of DT model for prognostic stage prediction. The prognostic stage assessment for the extracted prognostic factors can be done by observing the DT model's decision paths. The decision rule for each prognostic factor is learned for the prediction of the prognostic stage.

For generic system development, experimentation is divided into three parts, viz. first—data collected from a single regional area for training and testing; second—data collected from one regional place considered for training and tested on the other regional areas; third—data collected from both regional areas are considered for training as well as for testing.

Table 2 Prognostic stage classification according to the 8th AJCC manual

TNM	Grade	HER-2	ER	PR	P-Stage	TNM	Grade	HER-2	ER	PR	P-Stage
TisN0M0	Any	Any	Any	Any	0	T0N2M0 T1N2M0 T2N2M0 T3N1M0 T3N2M0	G1	Positive	Positive	Positive	IB
T1N0M0 T0N1miM0 T1N1miM0	G1	Positive	Positive	Positive	IA					Negative	IIIA
				Negative					Negative	Positive	IIIA
			Negative	Positive						Negative	IIIA
				Negative				Negative	Positive	Positive	IB
		Negative	Positive	Positive						Negative	IIIA
				Negative					Negative	Positive	IIIA
			Negative	Positive						Negative	IIIA
				Negative			G2	Positive	Positive	Positive	IB
	G2	Positive	Positive	Positive	IA					Negative	IIIA
				Negative					Negative	Positive	IIIA
			Negative	Positive						Negative	IIIA
				Negative				Negative	Positive	Positive	IB
		Negative	Positive	Positive						Negative	IIIA
				Negative					Negative	Positive	IIIA
			Negative	Positive						Negative	IIIB
				Negative	IB		G3	Positive	Positive	Positive	IIA
	G3	Positive	Positive	Positive	IA					Negative	IIIA
				Negative					Negative	Positive	IIIA
			Negative	Positive						Negative	IIIA
				Negative				Negative	Positive	Positive	IIB
		Negative	Positive	Positive						Negative	IIIA
				Negative					Negative	Positive	IIIA
			Negative	Positive						Negative	IIIC
				Negative	IB	T4N0M0 T4N1M0 T4N2M0 AnyTN3M0	G1	Positive	Positive	Positive	IIIA
T0N1M0 T1N1M0 T2N0M0	G1	Positive	Positive	Positive	IA					Negative	IIIB
				Negative	IB				Negative	Positive	IIIB
			Negative	Positive	IB					Negative	IIIB
				Negative	IIA			Negative	Positive	Positive	IIIA
		Negative	Positive	Positive	IA					Negative	IIIB
				Negative	IB				Negative	Positive	IIIB
			Negative	Positive	IB					Negative	IIIB
				Negative	IIA		G2	Positive	Positive	Positive	IIIA
	G2	Positive	Positive	Positive	IA					Negative	IIIB
				Negative	IB				Negative	Positive	IIIB
			Negative	Positive	IB					Negative	IIIB
				Negative	IIA			Negative	Positive	Positive	IIIA
		Negative	Positive	Positive	IA					Negative	IIIB
				Negative	IIA				Negative	Positive	IIIB
			Negative	Positive	IIA					Negative	IIIC
				Negative	IIA		G3	Positive	Positive	Positive	IIIB
	G3	Positive	Positive	Positive	IA					Negative	IIIB
				Negative	IIA				Negative	Positive	IIIB
			Negative	Positive	IIA					Negative	IIIB
				Negative	IIA			Negative	Positive	Positive	IIIB
		Negative	Positive	Positive	IB					Negative	IIIC
				Negative	IIA				Negative	Positive	IIIC
			Negative	Positive	IIA					Negative	IIIC
				Negative	IIA	AnyTAnyNM1	Any	Any	Any	Any	IV

Table 2 (continued)

TNM	Grade	HER-2	ER	PR	P-Stage	TNM	Grade	HER-2	ER	PR	P-Stage
T2N1M0 T3N0M0	G1	Positive	Positive	Positive	IA	T2N1M0 T3N0M0	G3	Positive	Positive	Positive	IB
				Negative	IIB					Negative	IIB
			Negative	Positive	IIB				Negative	Positive	IIB
				Negative	IIB					Negative	IIB
		Negative	Positive	Positive	IA			Negative	Positive	Positive	IIA
				Negative	IIB					Negative	IIB
			Negative	Positive	IIB				Negative	Positive	IIB
				Negative	IIB					Negative	IIIA
	G2	Positive	Positive	Positive	IB	T2N1M0 T3N0M0	G2	Negative	Positive	Positive	IB
				Negative	IIB					Negative	IIB
			Negative	Positive	IIB				Negative	Positive	IIB
				Negative	IIB					Negative	IIB

3.6.4 Validation method

The K-fold cross-validation technique, also known as rotation estimation, is used to evaluate unbiased measure of the classification system by training and testing k times. Fivefold cross-validation is performed for training and testing purpose of a single organization records. This validation method is commonly performed for optimizing the use of the training set to validate the framework rigorously over the unknown test set. Finally, the effect of all folds is merged for evaluation of the anatomic stage and prognostic stage.

4 Results

4.1 Experimental setup

Python 3.7 is used for report preprocessing, knowledge extraction from reports, and outcome analysis. Anaconda framework is used with scikit-learn, pandas, numpy, and natural language toolkit packages for system development. Precision, recall, F-measure, and accuracy are used to evaluate the output of prognostic factors and prognostic stage because these are the basic measures to evaluate output related to text classification or information extraction. Table 4 describes the prognostic stage confusion matrix. TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

$$\text{Recall of P - Stage } (R) = \frac{TP}{FP + TP} \tag{1}$$

$$\text{Precision of P - Stage } (P) = \frac{TP}{FN + TP} \tag{2}$$

$$\text{Accuracy of P - Stage } (A) = \frac{TP + TN}{FN + FP + TN + TP} \tag{3}$$

4.2 Dataset

A total of 465 de-identified original and realistic medical records are collected from reputed cancer treatment institutions and laboratories in Maharashtra, India. Tables 5 and 6 provide data collection details from two hospitals in the urban and rural regions. Medical records are gathered from Nurgis Dutta Memorial Cancer Hospital (NDMCH) in the rural region and Jehangir Hospital (JH) and laboratories in the urban region. For this proposed research, an ethics committee approval is taken from each hospital. The first approval is taken from the ethics committee of JH (Ref. No. JCDC/18–19/00455), and the second approval is taken from the ethics committee of NDMCH (Ref. No. IEC NDMCH/2019/37). Records of breast cancer sufferers registered from January 2017 to December 2019 are collected for this study. A total of 342 medical records are collected from Nurgis Dutt Cancer Memorial Hospital, while 123 medical records are received from Jehangir Hospital and laboratories. Four hundred sixty-five histopathology reports, 465 immunohistochemistry reports, and 465 mammography and ultrasonography reports are collected. These collected medical text records are unstructured and described in natural language. The anatomic stage details and prognostic stage details of both hospitals are also collected and tabulated in Tables 7 and 8.

Table 3 Details of specification for ML model

Attribute description		T (Tis, T0, T1, T2, T3, T4)
		N (N0, N1, N2, N3)
		M (M0, M1)
		G (1, 2, 3)
		ER (ER + ve, ER-ve)
		PR (PR + ve, PR-ve)
		HER (HER2 + ve, HER2-ve)
Implementation platform		Anaconda framework (jupyter notebook), sklearn, language—Python
Libraries to import		DecisionTreeRegressor, SVC, GaussianNB, GridSearchCV, train_test_split, mean squared error, NumPy, Pandas, nltk
Parameters	DT	random_state = 0, splitter = 'best', prune = 'true', Criterion = 'mse', ccp_alpha = 0.015, max_depth = 4 (for anatomic), 8 (for prognostic)
	SVM	C = 10, kernel = 'linear'
	GNB	var_smoothing = 1e-7
Node		Attribute, MSE, numSample, values
Validation		Cross-validation—5

It is observed that the earlier studies worked on only pathology reports for anatomic factor's information extraction [24, 28, 31, 33]. Hence, for more precise outcome prediction, this study considered different medical records of breast cancer patients. This study has used clinical, pathological, and immunohistochemistry reports for prognostic factors' information extraction to reduce false negative output and get high accuracy.

4.2.1 Analysis and verification of data

The analysis and verification of the collected data and information are further distinguished in the following manner:

- Inadequate data—authors did not collect inadequate data. In some cases, histopathology records were there, but immunohistochemistry records were not available. There were some reports where T (tumor) information was given, but N stage information was not provided. This could happen if resection of the tumor was done, but lymph node surgery was not done. Such information is considered inadequate information.
- Missing data—data collection excluded records having missing data of breast cancer sufferers such as HER-2 data was missing in some immunohistochemistry records and grade data was also missing in some histopathology records; these records also excluded in data collection.

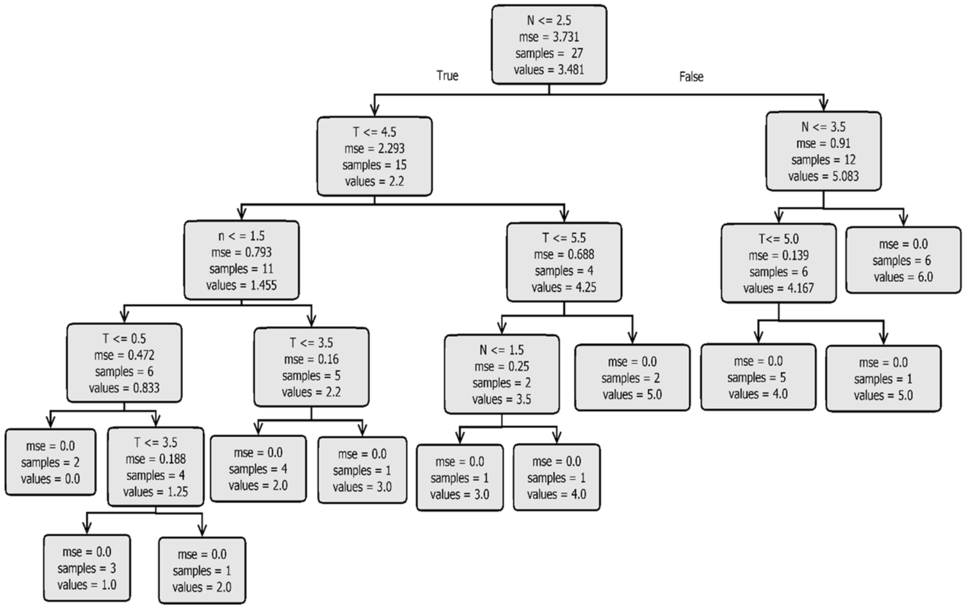


Fig. 6 Snap of decision tree model for anatomic stage prediction (for 27 samples)

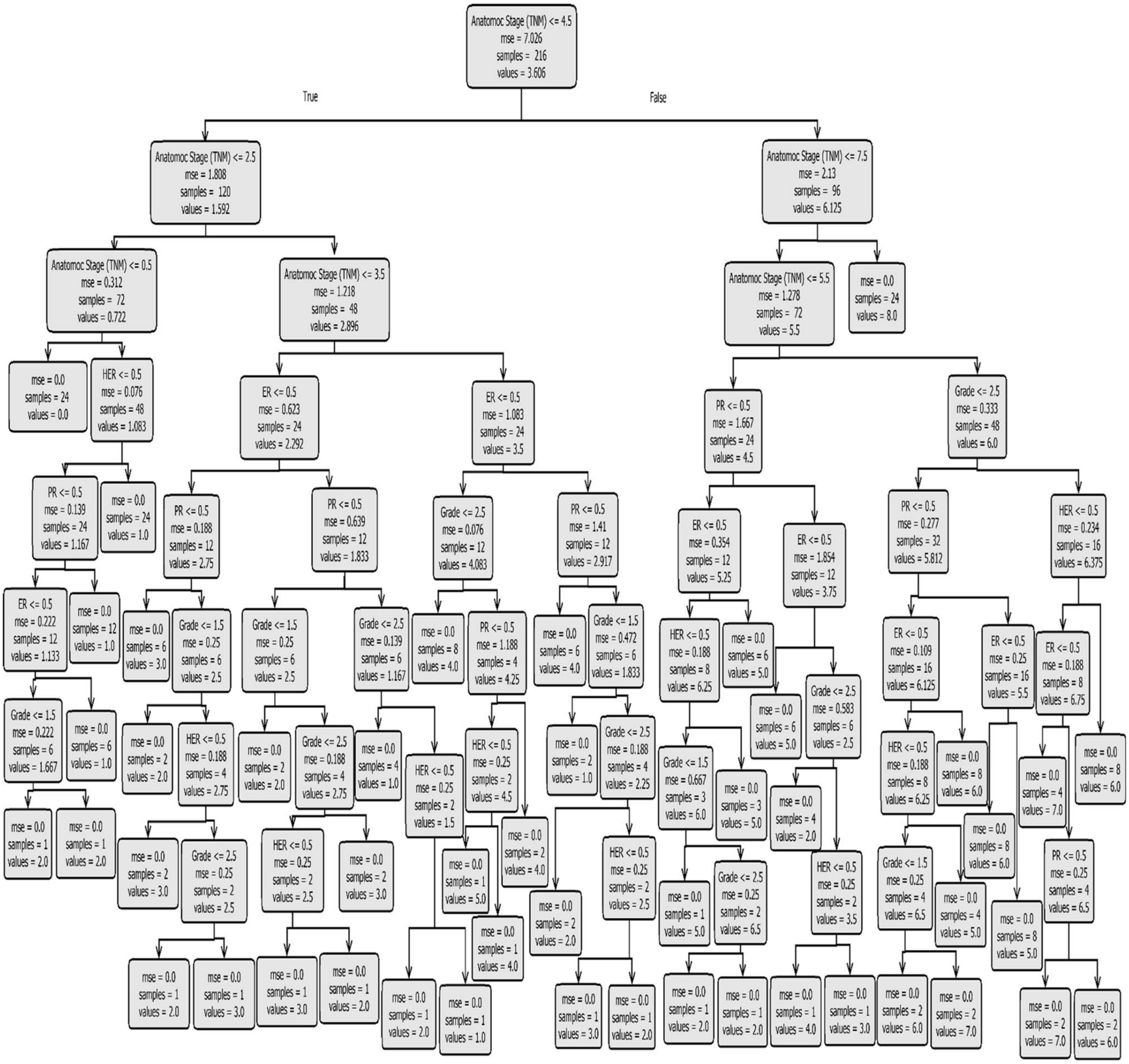


Fig. 7 Snap of decision tree model for prognostic stage prediction (for 216 samples)

- Quality data—analysis and verification of data are necessary to improve data collection because quality data is an important factor in accuracy prediction.
- Data security and confidentiality—De-identified records are obtained from hospitals to protect the patient's confidentiality and privacy. The authors performed plenty of surveys for report collection.
- Final data collection has 465 histopathology, immunohistochemistry, mammography, and ultrasonography reports of breast cancer sufferers after completion of verification and analysis of reports.

4.3 Gold standard

Two specialists in the medical sector reviewed all medical records, and their evaluation is known as the gold standard. The authors evaluated and measured the performance of prognostic factors produced by the proposed research with the assessment of prognostic factors by specialists.

4.4 Performance analysis

Performance analysis of the proposed method for the rural and urban areas is measured using the precision, recall,

Table 4 Prognostic stage confusion matrix

		Predicted prognostic stage	
		P-stage	Not P-stage
Original prognostic stage	*P-stage*	TP	FN
	Not P-stage	FP	TN

F-measure and accuracy of prognostic factors and prognostic stage. T, N, and M values extracted from unstructured medical records. In previous work [9], the extraction of T, N and M factors with a 98%, 94% and 99% accuracy is achieved in rural areas, while 95%, 88% and 98% respective accuracy is attained in urban areas' hospital.

4.4.1 Anatomic stage prediction result

Using the machine learning technique, the anatomic stage is predicted from the extracted anatomic factors. Figure 8 describes the output assessment obtained from the gold standard and the proposed research for the rural and urban areas' anatomic stage. The accuracy of anatomic stage prediction using ML depends on the combination of T, N, and M factors as given in Table 1; that is, the wrong extraction of the values from TNM factors' combination gives the wrong prediction accuracy of the anatomic stage. Figure 8 shows that the anatomic stage's average accuracy is 93% and 86% in the rural and urban areas. The most common stage observed in both regional areas is stage IIA, almost 33% and 28% in the rural and urban areas. When the value of TNM is T0N1miM0, then the anatomic stage will be stage IB. In this TNM combination, this study did not receive reports on lymph node value N1mi, hence could not check the anatomic stage prediction accuracy for stage IB.

A model for NDMCH is developed first and then applied to JH. While comparing the accuracy's differences between both regions, error analysis is done manually to determine low recall causes. In some records of the urban region, lymph node information is represented as "multiple AXLN," but the exact number of LN was not mentioned. Also, "No. of LN" was given in some urban region records for representing the lymph node information, where they considered its meaning as 'number of' but in this work, "No" is considered a negative term. This study could receive a small set of reports for stage-0 and stage-IV cases from both regional areas, but for a practical system, a sufficient amount of reports are necessary.

4.4.2 ER extraction result

This section represents the ER extraction results. Figure 9 presents the details of performance measurement of the proposed model and gold standard for ER. Figure 9 shows that the average accuracy of ER values is 99% and 98% in the rural and urban areas, respectively. It is noticed that ER positive cases are more than ER negative in the urban area, while ER negative cases are more than ER positive cases in the rural area for this dataset.

In the feature selection process, the use of dictionary creation and negation finding method for each prognostic factor, having a

Table 5 Data collection details of urban region

Dataset	T value		N value		M value		Grade value		ER		PR		HER	
Reports 123 cases	Tx	20	Nx	20	Mx	20	1	10	+ ve	60	+ ve	42	+ ve	49
	Tis	3	N0	42										
	T1	34	N1	37	M0	100	2	40						
	T2	54	N2	12					-ve	43	-ve	61	-ve	54
	T3	9	N3	12	M1	3	3	53						
	T4	3												

Table 6 Data collection details of rural region

Dataset	T value		N value		M value		Grade value		ER		PR		HER	
Reports 342 cases	Tx	15	Nx	15	Mx	15	1	52	+ ve	166	+ ve	143	+ ve	102
	Tis	3	N0	163										
	T1	99	N1	84	M0	340	2	125						
	T2	221	N2	58					-ve	176	-ve	199	-ve	240
	T3	16	N3	37	M1	2	3	165						
	T4	3												

Table 7 Anatomic stage details of urban and rural regions

Sr. No	Anatomic Stage	Urban (123)	Rural (342)
1	0	3	3
2	IA	18	66
3	IIA	29	111
4	IIB	21	63
5	IIIA	17	60
6	IIIB	2	2
7	IIIC	10	35
8	IV	3	2

strong correlation with the respective prognostic factors, helped to minimize the false positive count. In addition, the use of rule-based techniques as explained in Section 3.5.1, which differentiate positive and negative occurrence of terms, leads to achieve high accuracy of ER + ve, ER-ve, PR + ve, PR-ve, HER + ve and HER-ve extraction from reports of both hospitals.

4.4.3 PR extraction result

A performance analysis has been performed for both regional areas in terms of PR status extraction. The average accuracy of PR values is represented in Fig. 10, which shows 99% and 98% accuracy in the rural and urban areas. It is found that PR negative cases are higher than the PR positive cases in both regional areas.

At the time of creating dictionary, the medical terms ‘strongly positive’, ‘moderately positive,’ and ‘weakly positive’ all are considered ‘PR positive’. According to Table 2, PR status could be assigned as either positive or negative for prognostic stage detection. This could help to improve the accuracy of the model.

4.4.4 HER-2 extraction result

HER status extraction results are presented in the form of performance assessment of the proposed model and the gold standard for HER status as given in Fig. 11. It shows that the average accuracy of HER status is 99% and 96% in the rural and urban areas respectively. It is noticed that HER negative cases are more than HER positive cases in both regional areas for this dataset.

Table 8 Prognostic stage details of urban and rural regions

Sr. No	Prognostic Stage	Urban (123)	Rural (342)
1	0	3	3
2	IA	27	134
3	IIA	20	76
4	IIB	23	43
5	IIIA	15	35
6	IIIB	5	25
7	IIIC	7	24
8	IV	3	2

There are 60, 42, and 49 positive cases and 43, 61, and 54 negative cases of ER, PR, and HER2 values, respectively, in urban region. Similarly, there are 166, 143, and 102 positive cases and 176, 199, and 240 negative cases of ER, PR, and HER2 values, respectively, in rural region. Out of these, 16 cases were triple positive, and 15 cases were triple negative in the extracted results of urban area. Twenty-four cases were triple positive, and 106 cases were triple negative in rural region’s extracted results.

4.4.5 Grade extraction result

The grade values are evaluated to study the performance of both regional areas and it is shown in Fig. 12. The average accuracy of grade value is 98% and 95% in the rural and urban areas. Grade 3 is the most common grade observed in both regional areas (48% in rural area and 51% in urban area). Initially, it is noticed that the system is biased towards the negative target values because multiple grade values were given in some records of the urban region. Hence, the priorities are set as high priority to high value which enhances the accuracy of the results.

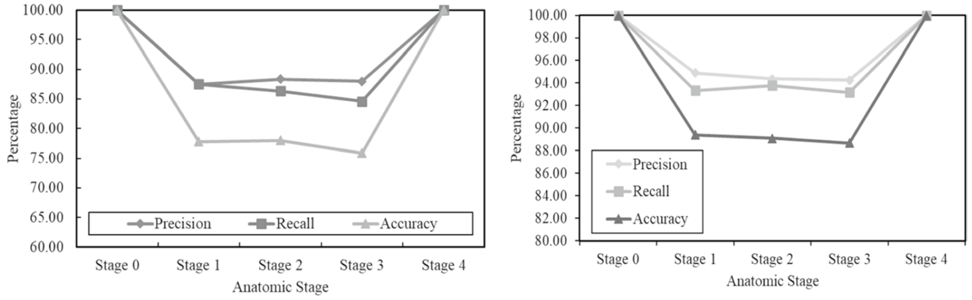
4.4.6 Prognostic stage prediction result

Predicting the prognostic stage is a challenging task as it includes several permutations of all prognostic factors. Figure 13 describes the output assessment obtained from the gold standard and the present study for the prognostic stage of the rural and urban areas. Figure 13 shows that the prognostic stage’s average accuracy is 92% and 82% in the rural and urban areas, respectively. Prognostic stage prediction using ML depends on the combination of prognostic factors as given in Table 2. Any single inaccurate factor out of T, N, M, grade, ER, PR, and HER-2 factors will affect the prognostic stage accuracy; hence, compared to the rural region, stage-2 and stage-3 show lower performance than remaining stages in the urban area.

The most common stage observed in both regional areas is stage-I, almost 39% and 40% in the rural and urban areas, respectively. This study could receive a small set of reports for stage-0 and stage-IV cases from both regional areas. For a more practical system, a sufficient amount of reports are necessary; hence, the authors aim to gather adequate reports of stage-0 and stage-IV cases in the future.

Table 9 gives a summary of the F1 score results of prognostic factors and cancer stage of both hospitals. After predicting the anatomic stage and prognostic stage from the medical reports of people with breast cancer, a comparison of the obtained results for both stages is performed. Figure 14 represents the comparison between the anatomic stage and the prognostic stage results of the rural and urban areas. The patients are described as up-staged, down-staged, or unchanged if the prognostic stage is greater than, smaller than, or similar to the anatomic stage.

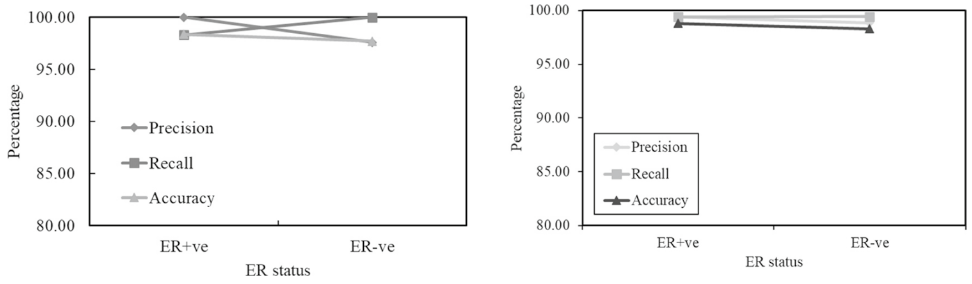
Fig. 8 Output evaluation from gold standard and proposed research for anatomic stage of (a) JH and (b) NDMCH



(a) JH

(b) NDMCH

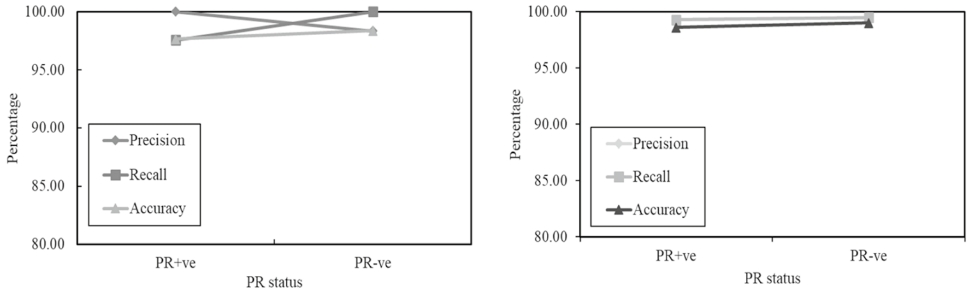
Fig. 9 Output evaluation from gold standard and proposed research for ER values of (**a**) JH and (**b**) NDMCH



(a) JH

(b) NDMCH

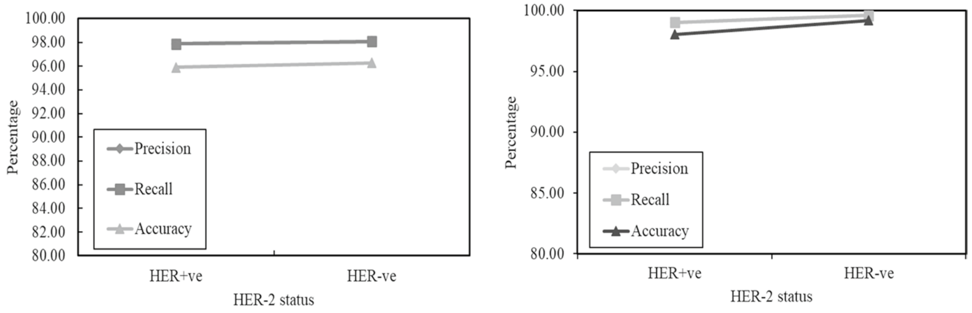
Fig. 10 Output evaluation from gold standard and proposed research for PR values of (**a**) JH and (**b**) NDMCH



(a) JH

(b) NDMCH

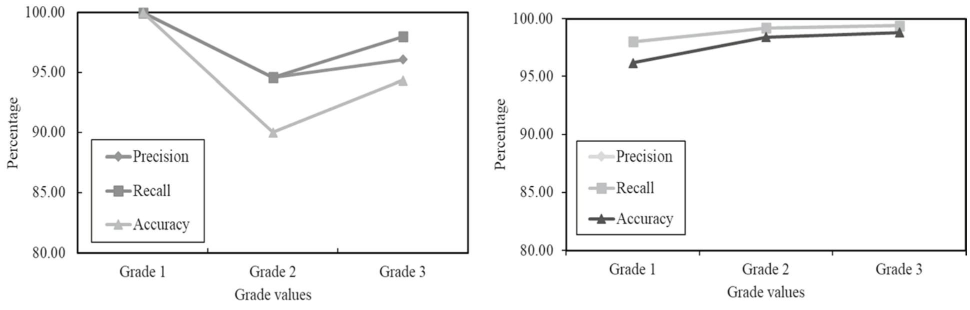
Fig. 11 Output evaluation from gold standard and proposed research for HER values of (**a**) JH and (**b**) NDMCH



(a) JH

(b) NDMCH

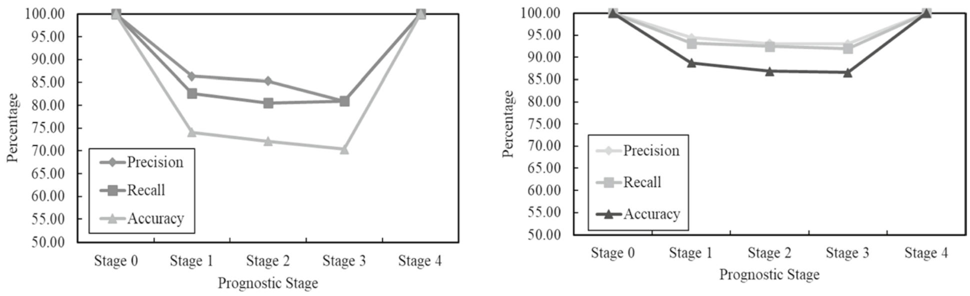
There was a significant difference found in the extracted anatomic stage and extracted prognostic stage for many cases. A total of 43% of cancer sufferers with extracted anatomic stage were either down-staged or up-staged in prognostic stage extraction. Tables 10 and 11 show the changes in stages for both regions. In both areas, the



(a) JH

(b) NDMCH

Fig. 12 Output evaluation from gold standard and proposed research for grade values of (**a**) JH and (**b**) NDMCH



(a) JH

(b) NDMCH

Fig. 13 Output evaluation for prognostic stage determination of (**a**) JH and (**b**) NDMCH

common change of stage is IIA (anatomic stage) to IA (prognostic stage).

Figure 15 shows 87% average accuracy of the prognostic stage achieved in both regions. It shows that, although there is diversity in medical reports of different hospitals from different regions, the proposed study achieved precise accuracy. The above results proved that the proposed work's primary goal is to make a generalized system for breast cancer stage detection achieved with strong performance accuracy.

Table 9 F1-score measurement of prognostic factors and cancer stage of JH and NDMCH

Prognostic factors		JH	NDMCH
T		97.43	98.70
N		93.12	96.64
M		98.81	99.70
ER		98.95	99.26
PR		98.96	99.39
HER		97.97	99.29
Grade		97.21	98.85
Cancer stage	Classifier	JH	NDMCH
Anatomic stage	GNB	87.15	89.30
	SVM	88.10	91.22
	DT	91.23	96.38
Prognostic stage	GNB	82.40	88.25
	SVM	81.80	89.54
	DT	87.65	95.21

4.5 Discussion

4.5.1 Observations

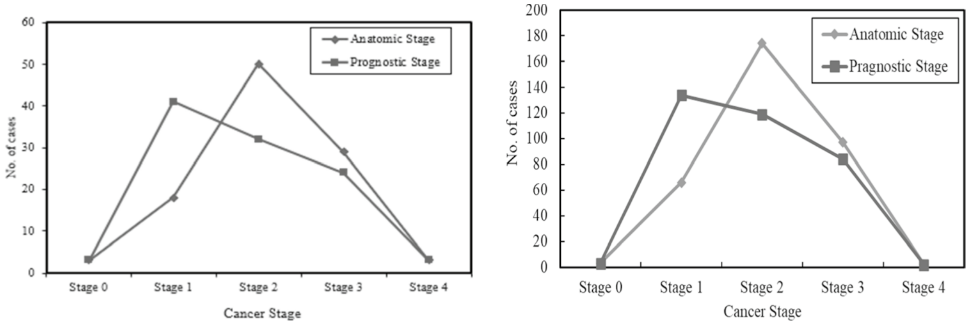
A manual analysis of the reports is performed to seek the potential causes of errors. It is noticed that there is a contradiction between the extracted prognostic factors' results obtained by our system and prognostic factors' results given by the gold standard. In some of the cases, the extracted T result is 'T1c,' and the result given by the gold standard is 'T2'. Similarly, during the extraction of ER, PR, and HER-2, in some cases, extracted HER-2 value result is 'positive,' and the result given by the gold standard is 'equivocal.' Our methodology has referred to the guidelines of AJCC while implementing prognostic factors' extraction algorithm. This contradiction could happen as experts may have reviewed some additional information or documentation for identifying prognostic factor details.

While at the time of implementation, extracting affected lymph node information from reports was quite tricky than extracting other prognostic factors from the reports. In this study, a few records are also collected which are not malignant and used for testing an output of Tx, Nx, and Mx.

4.5.2 Prognostic factor dependency

The only anatomic stage is not sufficient in the age of effective treatment. Hence, a tumor biology is also equally

Fig. 14 Comparison between anatomic stage and prognostic stage of (**a**) JH and (**b**) NDMCH



(a) JH

(b) NDMCH

considered for study. Histological grade, hormone receptor status, and HER-2 are not depending on the anatomic stage. The prognostic stage depending on the number of prognostic elements is considered more precise in predicting outcome than the anatomic stage alone.

4.5.3 Generic system

A generic and universally suitable model for predicting prognostic staging is the primary aim of this study. The prognostic stage integrates almost all significant biological elements like hormone receptor status, HER-2, and histological grade, along with TNM values under one system. Earlier studies reported some constraints for generic implementation, they developed an anatomic stage extraction system only for a particular dataset which may lead to over-fitting risk [31, 33]. Hence, this study used medical records of two different hospitals belonging to different regional areas having diversity in the dataset, improving the generic and portable system implementation.

Table 10 Changes in stage for JH

Anatomic stage	Prognostic stage		No. of cases changed
	Up	Down	
0	0	NA	0
IA	1	0	1
IB	0	0	0
IIA	0	14	14
IIB	4	6	10
IIIA	1	8	9
IIIB	0	1	1
IIIC	0	7	7
IV	NA	0	0

NA, not applicable

Implementing such a generic prognostic system, including several permutations of all the prognostic factors to predict the prognostic stage, is difficult. For a more precise practical approach, a sufficient amount of reports to test each permutation is necessary. Hence, in the future, this research aims to collect a sufficient amount of reports to test each permutation of all prognostic factors such that the prognostic system will be more precise and practical.

4.6 Constraints

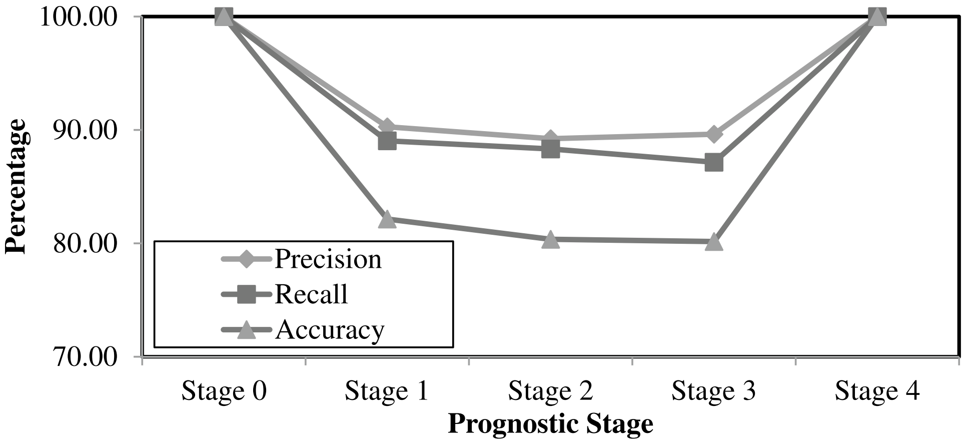
- This work primarily designed for people with breast cancer, but it could apply to other subcategories of cancer with some slight modifications.
- In some instances, the authors found a difference between the results for prognostic factors given in the gold standards and the present model's results. However, the system developed based on the 7th and 8th editions of the AJCC manual.
- During this extraction, the accuracy of every prognostic factor is important because the wrong result of any single prognostic factor affects the prognostic stage results.
- Point noted: this study has a small set of reports for stage-0 and stage-IV cases from both regional areas. But the authors specially mention that there is a need of a

Table 11 Changes in stage for NDMCH

Anatomic stage	Prognostic stage		No. of Cases Changed
	Up	Down	
0	0	NA	0
IA	12	0	12
IB	0	0	0
IIA	0	45	45
IIB	12	20	32
IIIA	16	25	41
IIIB	0	0	0
IIIC	0	21	21
IV	NA	0	0

NA, not applicable

Fig. 15 Output evaluation for prognostic stage determination of both regions



sufficient number of reports for better results. Hence, the authors aim to gather adequate reports for stage-0 and stage-IV cases in the future.

5 Conclusion and future work

This research study presented the method of information extraction from various unstructured medical records produced in prognostic stage detection of breast cancer patients using natural language processing. The authors proposed a rule-based method and a machine learning technique to predict the prognostic stage with higher accuracy. The data from clinical, pathological, and immunohistochemistry reports from two health centers located in urban and rural regions is used for the study. This study suggests a generic clinical decision-unifying staging method for the extraction of the most reliable prognostic stage information from the unstructured medical records of various health institutions. Higher average accuracy achieved for the anatomic stage, grade, ER, PR, and HER2 was 93%, 98%, 99%, 99%, and 99%, respectively, in rural region hospital, whereas in urban area, it found 86%, 95%, 98%, 98%, and 96%, respectively. The prognostic stage average accuracy is 92% and 82% in rural and urban regions. This study achieved 87% average accuracy of the prognostic stage in both regions. From the analysis presented, this study can make the following conclusion.

1. Prognostic stage detection using the proposed method can be possible with good accuracy in different health institutions of different regions. It suggests that the proposed research would probably help to improve care choices and accurate prognosis for breast cancer. The prognostic stage showed precise prognostic details compared to the anatomic stage by integrating biological and anatomical elements.
2. This study contributes to the predictability of prognosis and appropriate care, which will be a step forward in improving breast cancer's health outcome.

Acknowledgements The authors would like to convey thanks to Nurgis Dutta Memorial Cancer Hospital (NDMCH) in the rural region and Jehangir Hospital (JH) and different laboratories in the urban area for allowing access to their medical records. The authors also want to acknowledge Ethics Committees for approving proposed research. The authors would like to express their gratitude to Dr. Nene, Dr. Chauhan, Dr. Joshi, Dr. Shilpa, Dr. Vibhute, and Dr. Mane for their guidance in the medical field. Thanks to all patients with breast cancer for letting the authors use their medical records for this research. The authors would like to thank Harshal Ingale for his cooperation.

References

1. Cancer Statistics in India. http://cancerindia.org.in/cancer-statistics/. Accessed 25 Nov 2020
2. Mathur P, Sathishkumar K, Chaturvedi M, Das P, Sudarshan K, Santhappan S, Nallasamy V, John A, Narasimhan S, Roselind F (2020) Cancer Statistics, 2020: Report from National Cancer Registry Programme, India. JCO Global Oncol 6:1063–1075. https://doi.org/10.1200/GO.20.00122
3. Martinez D, Cavedon L, Pitson G (2013) Stability of text mining techniques for identifying cancer staging. In: Louhi, The 4th International Workshop on Health Document Text Mining and Information Analysis, NICTA, Canberra, Australia
4. Kim BJ, Merchant M, Zheng C, Thomas AA, Contreras R, Jacobsen SJ, Chien GW (2014) Second prize: "A natural language processing program effectively extracts key pathologic findings from radical prostatectomy reports." J Endourol 28(12):1474–1478. https://doi.org/10.1089/end.2014.0221
5. Wen-wai Y, Meliha Y (2016) Natural Language Processing in Oncology a Review. J Am Med Inform Assoc 2(6):797–804. https://doi.org/10.1001/jamaoncol.2016.0213
6. Cheng LTE, Zheng J, Savova GK, Erickson BJ (2010) Discerning tumor status from unstructured MRI reports: completeness of information in existing reports and utility of automated natural language processing. J Digit Imaging 23(2):119–132. https://doi.org/10.1007/s10278-009-9215-7

7. Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A (2011) AJCC cancer staging manual, 7th edn. Springer-Verlag, Berlin. ISBN 978-0-387-88440-0
8. Spasic I, Livsey J, Keane JA, Nenadic G (2014) Text mining of cancer-related information: Review of current status and future directions. Int J Med Informatics 83:605–623. https://doi.org/10.1016/j.ijmedinf.2014.06.009
9. Deshmukh PR, Phalnikar R (2020) TNM cancer stage detection from unstructured pathology reports of breast cancer patients. In: Bhalla S et al (eds) Proceeding of International conference on computational science and applications, algorithms for intelligent systems. Springer Nature Singapore Pte Ltd., CH 40:411–418. https://doi.org/10.1007/978-981-15-0790-8_40
10. Ravi K, Ramachandra GA, Nagamani K (2013) An Efficient Prediction of Breast Cancer Data using Data Mining Techniques. Int J Innov Eng Technol 2(4):139–144. SSN: 2319-1058
11. Chatterjee S, Chattopadhayay A (2016) Cancer Registration in India– Current Scenario and Future Perspectives. Asian Pac J Cancer Prev 17(8):3687–3696. https://doi.org/10.14456/apjcp.2016.154/APJCP.2016.17.8.3687
12. Wong RX, Wong FY, Lim J, Lian WX, Yap YS (2018) Validation of the AJCC 8th prognostic system for breast cancer in an Asian healthcare setting. Breast 40:38–44. https://doi.org/10.1016/j.breast.2018.04.013. Elsevier
13. Wang M, Chen H, Kejin W, Ang D, Mingdi Z, Peng Z (2018) Evaluation of the prognostic stage in the 8th edition of the American Joint Committee on Cancer in locally advanced breast cancer: An analysis based on SEER 18 database. Breast 37:56–63. https://doi.org/10.1016/j.breast.2017.10.011
14. National centre for Disease Informatics and Research, National Cancer Registry Program, http://www.ncrpindia.org/. Accessed 25 Nov 2020
15. Yokoyama S, Hamada T, Higashi M, Matsuo K, Maemura K, Kurahara H, Horinouchi M, Hiraki T, Sugimoto T, Akahane T, Yonezawa S, Kornmann M, Batra SK, Hollingsworth MA, Tanimoto A (2020) Predicted Prognosis of Patients with Pancreatic Cancer by Machine Learning. Clin Cancer Res 26:2411–2421. https://doi.org/10.1158/1078-0432,January28
16. Li J, Li Z, Luo J, Yao Y (2020) ACNNT3: Attention-CNN Framework for Prediction of Sequence- Based Bacterial Type III Secreted Effectors. Comput Math Methods Med Article ID 3974598:7. https://doi.org/10.1155/2020/3974598
17. Li Z, Zhu J, Xu X, Yao Y (2020) RDense: a protein-RNA binding prediction model based on bidirectional recurrent neural network and densely connected convolutional networks. IEEE Access 8. https://doi.org/10.1109/ACCESS.2019.2961260.
18. Jiang X, Zhao J, Qian W, Song W, Ning LG (2020) A generative adversarial network model for disease gene prediction with RNA-seq data. IEEE Access 8. https://doi.org/10.1109/ACCESS.2020.2975585.
19. Mignone P, Pio G, D'Elia D, Ceci M (2020) Exploiting transfer learning for the reconstruction of the human gene regulatory network. Bioinformatics 36(5):1553–1561. https://doi.org/10.1093/bioinformatics/btz781
20. Pio G, Ceci M, Prisciandaro F, Malerba D (2020) Exploiting causality in gene network reconstruction based on graph embedding. Mach Learn 109:1231–1279. https://doi.org/10.1007/s10994-019-05861-8
21. Barracchia EP, Pio G, Delia D, Ceci M (2020) Prediction of new associations between ncRNAs and diseases exploiting multi-type hierarchical clustering. BMC Bioinformatics 21:70. https://doi.org/10.1186/s12859-020-3392-2
22. Jiang D, Liao J, Duan H, Wu Q, Owen G, Shu C, Chen L, He Y, Wu Z, He D, Zhang W, Wang Z (2020) A machine learning-based prognostic predictor for stage III colon cancer. Sci Rep 10:10333. https://doi.org/10.1038/s41598-020-67178-0
23. Muhammad A, Maqbool H, Wajahat Ali K, Ali T, Lee S, Huh E-N, Hafiz Farooq A, Arif J, Hassan I, Muhammad I, Manzar Abbas H (2017) Comprehensible knowledge model creation for cancer treatment decision making. Comput Biol Med 82:119–129. https://doi.org/10.1016/j.compbiomed.2017.01.010. Science Direct, Elsevier
24. Martinez D, Pitson G, MacKinlay A, Cavedon L (2014) Cross-hospital portability of information extraction of cancer staging information. Artif Intell Med 62:11–21. https://doi.org/10.1016/j.artmed.2014.06.002. Elsevier
25. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, Colquist S (2010) Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc 17:440–445. https://doi.org/10.1136/jamia.2010.003707
26. Rani GJJ, Gladis D, Mammen JJ (2017) Comparison of breast cancer staging in natural language text and SNOMED annotated text. Int J Pure Appl Math 116(21):243–249
27. Warner JL, Mia AL, Michael NN (2016) Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. Am Soc Clin Oncol 12(2). https://doi.org/10.1200/JOP.2015.004622.
28. Martinez D, Li Y (2011) Information extraction from pathology reports in a Hospital setting. CIKM'11, 1877–1882, ACM 978-1-4503-0717-8/11/10, October 24–28
29. McCowan I, Moore D, Fry M-J (2006) Classification of cancer stage from free-text histology reports. International Conference of the IEEE Engineering in Medicine and Biology Society. https://doi.org/10.1109/IEMBS.2006.259563
30. Rani GJJ, Gladis D, Mammen JJ (2019) SNOMED CT annotation for improved pathological decisions in breast cancer domain. Int J Recent Technol Eng 8(3). https://doi.org/10.35940/ijrte.C6519.098319
31. Nguyen A, Moore D, McCowan I, Courage M Multi-class classification of cancer stages from free-text histology reports using support vector machines. 29th Annual International Conference of the IEEE EMBS, France IEEE 2007, pp 5140–5143, https://doi.org/10.1109/IEMBS.2007.4353497
32. Rajaguru H, Vasanthi NS, Balasubramani M (2012) Performance analysis of artificial neural networks and statistical methods in classification of oral and breast cancer stages. Int J Soft Comput Eng 2(3)
33. McCowan IA, Moore DC, Nguyen AN, Bowman RV, Clarke BE, Duhig EE, Fry M-J (2007) Collection of cancer stage data by classifying free-text medical reports. J Am Med Inform Assoc 14(6):736–745. https://doi.org/10.1197/jamia.M2130
34. Dursun D, Glenn W, Amit K (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 34:113–127. https://doi.org/10.1016/j.artmed.2004.07.002. Elsevier
35. Joseph AC, David SW (2006) Applications of machine learning in cancer prediction and prognosis. Cancer Informat 2:59–77. PMID: 19458758, PMCID: PMC2675494
36. Dechang C, Huan W, Li S, Matthew TH, Donald EH, Arnold MS, Jigar AP (2016) An algorithm for creating prognostic systems for cancer. J Med Syst 40:160. https://doi.org/10.1007/s10916-016-0518-1. Springer
37. Deshmukh PR, Phalnikar R Identifying contextual information in medical document classification using term weighting. IEEE 8th International Advanced Computing Conference at Bennett University, Greater Noida, India, 17th -18th Dec 2018
38. U.S. National Library of Medicine (2008) Unified medical language system (UMLS). https://www.nlm.nih.gov/research/umls/. Accessed 25 Nov 2020
39. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG (2001) A simple algorithm for identifying negated

findings and diseases in discharge sum-maries. J Biomed Inform 34(5):301–310. https://doi.org/10.1006/jbin.2001.1029

40. www.Breastcancer.org. Accessed 25 Nov 2020
41. Sanjay PB, Partha SR, Myung-Shin S, Xing Y, Jaime MS, Xiaojiang C, Armando EG (2014) Personalizing breast cancer staging by the inclusion of ER, PR, and HER2. JAMA 149(2):125–129. https://doi.org/10.1001/jamasurg.2013.3181
42. Buckley JM, Coopey SB, Sharko J (2012) The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 3:23. https://doi.org/10.4103/2153-3539.97788
43. Dixit A, Singh R (2017) Multiple sliding window based pattern matching algorithms: survey. International Journal of Creative Research Thoughts (IJCRT) 5(4):3453–3458
44. Amjad H, Rola A, Dima S (2015) Four sliding windows pattern matching algorithms. J Softw Eng Appl. https://doi.org/10.4236/jsea.2015.83016
45. Hortobagyi GN, Connolly JL, D'Orsi CJ, Edge SB, Mittendorf EA, Rugo HS, Solin LJ, Weaver DL, Winchester DJ, Giuliano A AJCC Cancer staging manual eighth edition. https://doi.org/10.1007/978-3-319-40618-3_48
46. Mogana DG, Nur AT, Yip CH, Pietro L, Sarinder KD (2019) Predicting factors for survival of breast cancer patients using machine learning Techniques. BMC Med Inform Decis Mak 19:48. https://doi.org/10.1186/s12911-019-0801-4

Pratiksha R. Deshmukh received the MTech Computer degree from College of Engineering Pune (COEP). She is an Assistant Professor in College of Engineering Pune, and she is a research scholar at School of Computer Engineering and Technology, MIT World Peace University, India.

Dr. Rashmi Phalnikar received a Ph.D. degree in Computer Engineering. She is an Associate Professor and Ph.D. guide in School of Computer Engineering and Technology, MIT World Peace University, India.