Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

Normalization and Classification of Histopathology Electronic Health Records on Breast Cancer using NLP and ML Approaches

[1]Prathibha R J, [2]Ananya Tomar, [3]Anup Shandilya, [4]Chandana M , [5]Pruthvi Bhat

[1]-[5] JSS Science and Technology University, Department of Information Science and Engineering, Mysuru, India

Abstract. Breast cancer is a prevalent form of cancer that poses a significant threat to women globally, being one of the major causes of cancer-related fatalities among this population. Detecting the disease at an early stage plays a vital role in reducing both the number of cases and the mortality rate associated with it. To achieve this, valuable information for clinical and academic research is typically found within pathology reports, which provide essential insights into the nature and characteristics of the disease. However, these reports are often complex and detailed, making it difficult to extract relevant and qualitative data efficiently. Consequently, there is a need for keyword extraction techniques specifically designed for pathology reports, allowing for the effective summarization of educational content and minimizing the laborious and time-consuming process of report analysis. One such approach is the utilization of natural language processing, which involves the application of computational algorithms to extract keywords from histopathological reports, facilitating the identification and organization of critical information.

Keywords: Electrocardiogram, Arrhythmia, Convolutional neural network, K-Fold cross validation

1. Introduction

In India, every four minutes a woman is diagnosed with breast cancer. With around 1.78 lakh new cases being diagnosed every year, the incidence of breast cancer is the most common among various types of cancer. Histopathology is the study of changes in a tissue caused by a disease. It plays an important part in determining the treatment strategy for breast cancer, with the evaluation of breast specimens determining the surgical and the oncological therapeutic options to be used

Electronic Health Records (EHR) consists of raw data which is in an unstructured and text format. It is of utmost important to study the various biological terms present in the EHR to make any concrete decision. The pathology report is the primary source of information used to diagnose a patient. The pathologist examines and describes all types of specimens from all operations and biopsy procedures in the pathology report. The pathology report is a mandatory document in all clinical departments of the hospital since it is a source of comprehensive pathological information about the patient. However, because the pathology report is narrative in character, it is quite difficult to extract and generate research data from the original report. Because pathology reports are presented as narrative documents, data management for them often takes an inordinate amount of time, effort, and money.

Developing a model using machine learning and Natural Language Processing (NLP) to identify local recurrences in breast cancer patients can reduce the time-consuming work of a manual chart review. Natural language processing involves feeding an algorithm, large amounts of Electronic Health Records notes from which it “learns” a set of rules to identify what is meaningful. Machine learning can make patterns evident but only if the data used is clean, normalized and complete. NLP is a critical part of obtaining data from specialist documents and clinical notes.

2. Related works

The case of Breast Cancer Recurrence, by David S Carrell, Scott Halgrim, Diam Thi Tran[1], developed an NLP based system using open source software to process EHR from 1995 to 2012 for women with early stage breast cancer to identify whether and when recurrences were diagnosed. The study was limited to

Stage one or two cancer and machine learning techniques were not incorporated which could have enhanced the accuracy of status annotations.

Natural Language Processing approaches to detect the timeline of metastatic recurrence of breast cancer, by Imon Banerjee, Selen Bozkurt, Jennifer Lee Caswell-Jin[2], curated vocabulary by processing radiology and pathology reports. Developed and evaluated 2 NLP approaches to analyze free-text notes. Trained the NLP models and extracted results for future data. Limitations: Limited documentation of metastatic recurrence in clinical notes. Relatively short median follow-up time of 5 years. Understanding the basis for determinations made by neural networks are obscure.

Machine Learning methods to extract documentation of breast cancer symptoms from electronic health records, by Alexander W. Forsyth MEng, Regina Barzilay, Kevin S. Hughes[3], Manually annotated sentences and trained a conditional random field model to predict words indicating symptoms. Sentences labeled by human coders were divided into training, validation, and test data sets. Final model performance was determined on test data unused in model development or tuning. The major limitation is ambiguous descriptions in the free text which is in narrative manner. Accuracy could be increased if more stringent labelling was used. Used manual labeling which is time consuming. The small and restricted size.

Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records, by Yoojoong Kim,Jeong Hyeon Lee, Sunho Choi[4], Proposed a keyword extraction method for pathology reports based on the deep learning models for NLP, to extract pathological keywords, namely specimen, procedure, and pathology, from pathology reports. The algorithm was developed using the pathology reports of a single institution, which might limit the generalizability of its application to other institutions.

3. Methodology

This section will explain the approach we used in our project with step-by-step procedure and a detailed explanation on the optimized model created with its mathematical expressions and the reason behind their usage. Architecture for the proposed work is given in Figure 1.

3.1 Block diagram and Procedure

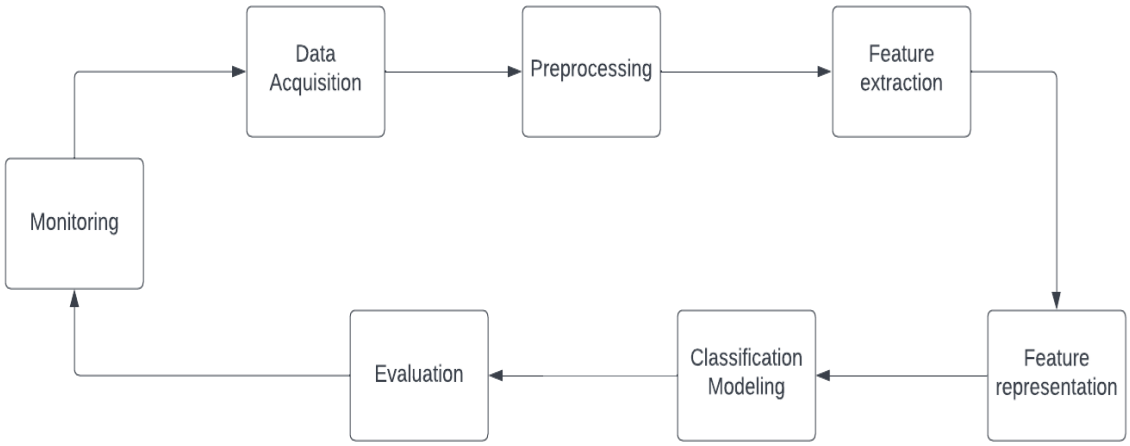

Figure 1: Architecture for the proposed work

A detailed description about the steps followed in the proposed model is given below.

- **Collection of raw data**

In this analytical study, electronic health records are collected from the hospital. EHRs are real-time, records that make information available securely and instantly to authorized users. The histological data from the HER is extracted from the collected dataset.

- **Pre-processing**

Natural language processing techniques are used to extract the useful information from the EHR. NLP involves feeding an algorithm large amounts of EHR data from which it "learns" a set of rules to identify what is meaningful. Machine learning can make patterns evident but only if the data used is clean, and normalized.

- **Feature Extraction**

Identifying the important keywords that help in determining whether the sample is malignant. For example, the number, size and color of cells.

- **Feature Representation**

Using the Bloom-Richardson's grading system, the features are tabulated column wise with the records being valued row wise.

- **Classification Modeling**

Various Machine Learning algorithms are used to classify whether the tissues are malignant or not, using the structured data.

- **Evaluation**

The accuracy of the findings produced by the various algorithms is compared and the optimal one is selected.

- **Monitoring**

To achieve the most optimal results, fresh samples are fed to the trained model.

3.2 Model Architecture

The workflow for the proposed work involves several key steps in the normalization and classification of histopathology Electronic Health Records (EHRs) on breast cancer using NLP and ML approaches. Firstly, digital EHRs related to breast cancer are acquired from the hospital, ensuring data compatibility and integrity. The acquired data is then pre-processed, where relevant attributes are extracted from the EHRs and handled for missing values and inconsistencies. Next, NLP techniques are applied to extract cancer-related features from textual data within the EHRs, identifying significant medical terms and contextual information. Feature engineering may be performed to create meaningful features. The ML model is trained using appropriate algorithms, splitting the dataset into training and testing sets. Model evaluation is conducted to assess its performance and generalization capability. Finally, the trained model is deployed and integrated into a user-friendly interface or application, ensuring secure handling and privacy of patient data. The detailed work flow of the proposed work is given in Figure 2. The explanation of each step is explained below.

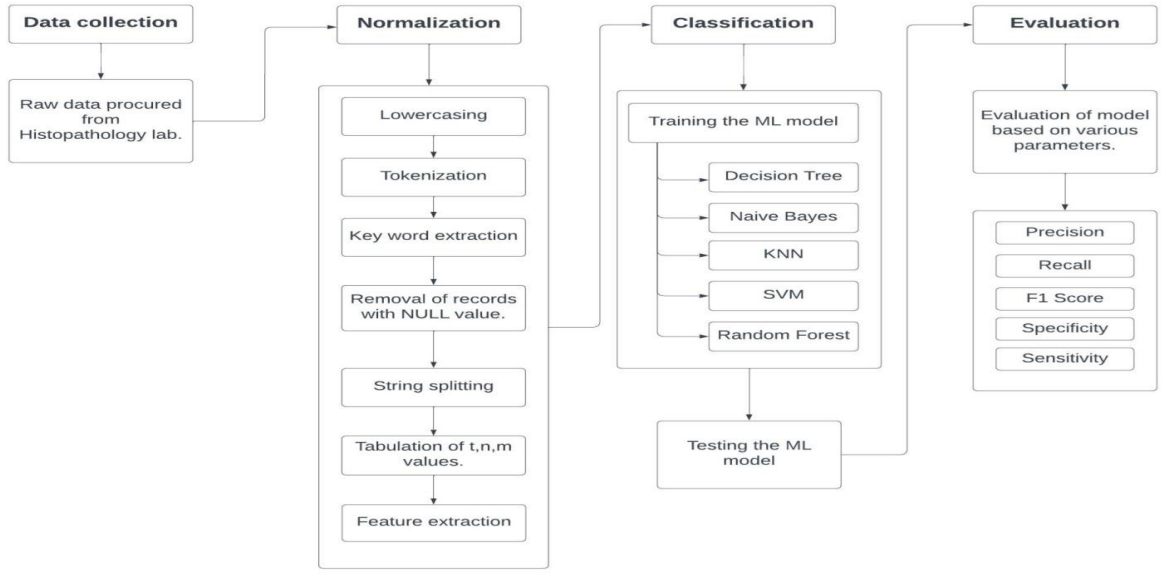

Figure 2: Detailed work flow of the proposed work

3.2.1. Data collection and Cleaning

This analytical study makes advantage of the hospital's computerised health records. In an EHR, a patient's paper chart is transformed to digital form. EHRs are patient-centered, real-time records that make information available to authorised users safely and quickly. We remove the histological information from the EHR. To extract relevant information from the EHR, the data is pre-processed using natural language processing algorithms. NLP entails training an algorithm to discover significant patterns in a huge set of EHR notes. Data must be clean, normalised, and complete in order for pattern recognition to be effective. NLP is critical in extracting information from specialised texts and clinical notes.

3.2.2. Normalization

It is a crucial step in data pre-processing that involves transforming text data into a consistent and standardized format for analysis. It ensures that the data is in a suitable form for further processing. Within the normalization process, several sub-steps are typically performed:

- Lowercasing: This involves converting all text to lowercase letters. It helps by avoiding duplication and treating words with different cases as the same.
- Tokenization: Tokenization involves breaking down the text into individual tokens or words. It separates the text into meaningful units, such as words or punctuation marks.
- Keyword Extraction: Keyword extraction is the process of identifying and extracting important words or phrases from the text. These keywords carry significant meaning and help in understanding the content of the data.
- Removal of Records with Null Values: Null values refer to missing or empty data points. The records with null values are removed to ensure the integrity of the data.
- String Splitting: String splitting involves dividing a text string into separate substrings based on a specific delimiter or separator. It helps in extracting specific information or splitting composite data into individual components.
- Tabulating the T, N, M Sections: In medical data analysis, the T, N, M sections refer to specific sections or attributes related to tumor staging. T denotes tumor size and extent, N denotes the presence or absence of involvement of lymph node, and M denotes the presence or absence of distant metastasis. Tabulating these sections involves organizing the extracted information from EHRs into a structured format, typically in a table or spreadsheet, where each section is represented as columns, and each record or instance is represented as rows.

Overall, these steps in normalization play a crucial role in preparing the text data for further analysis and modeling tasks, ensuring consistency, and extracting relevant information from the raw EHR data.

3.2.3. Classification

It is a machine learning task that involves assigning predefined labels or categories to input data based on its features. To train a classification model, various algorithms can be used, such as decision trees, naive bayes, k-nearest neighbours (KNN), support vector machines (SVM), and random forests. In the training phase, the dataset is generally divided into two sets namely training set and testing set. The training set, typically consisting of 70%, 80%, or 90% of the data, is used to train the model. The remaining portion, the testing set, is used to evaluate the model's performance. Each ML model has its own underlying principles and learning mechanisms.

- Decision trees use a hierarchical structure of decisions to classify data.
- Naive Bayes applies probabilistic principles based on feature independence.
- KNN classifies data based on the similarity of neighboring instances.
- SVM separates data into different classes using hyperplanes.
- Random forests combine multiple decision trees for classification.

The 'tnm' dataset is labeled based on the reference table provided by the American Joint Committee on Cancer (AJCC) given in Table 1. The AJCC reference table assigns specific categories to tumor size (t), lymph

node involvement (n), and metastasis (m) to facilitate standardized classification and staging of cancer cases. There are four primary stages and several sub-stages.

Table 1: The reference table for labeling the 'tnm' dataset by American Joint Committee Cancer (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	IIIB	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

3.2.4. Evaluation

The evaluation metrics provide a comprehensive understanding of the model's performance by considering different aspects like accuracy, identification of positive and negative cases, and trade-offs between precision and recall. Depending on the specific task and requirements, one or more of these metrics can be used to evaluate and compare ML models. In this proposed work, the evaluation metrics like Precision, Recall, F1-Score, Specificity and Sensitivity are used.

4. Result and Discussion

This section will show the results of the project which includes Accuracy, precision, F1- score, recall, sensitivity and specificity comparison graphs for 90:10, 80:20 and 70:30 splits. Also includes a confusion matrix with real time testing of the model.

4.1 Results obtained with graphs/plots.

Total number of records dataset consists of **1692** records. The data set is Normalized through various NLP techniques like lower casing, tokenization, keyword extraction, removal of null values and string splitting to extract the substrings of the t, n, m values and the values are tabulated. Each column has the subdivision values of the t, n, m parameters.

Labels present in the dataset:

The dataset includes labels representing tumor size (t), lymph node involvement (n), and metastasis (m). Tumor size is categorized into various stages (0, 1a, 1b, 1c, 2, 3, 4a, 4b, 4c), lymph node involvement has different stages (0, 1a, 1b, 1c, 2a, 2b, 3a, 3b, 3c), and metastasis is either present (1) or absent (x). The performance obtained by the proposed work using various ML algorithms is given in Table 2.

Table 2: Performance obtained by the proposed work using various ML algorithms

	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity
Decision tree	71.992	0.660	0.720	0.686	0.451	0.943
Naive Bayes	6.706	0.008	0.067	0.012	0.281	0.857
KNN	95.66	0.959	0.957	0.957	0.881	0.991
SVM	92.11	0.959	0.957	0.957	0.821	0.987
Random forest	96.449	0.967	0.964	0.965	0.954	0.991

The table displays the performance metrics of different machine learning models on a dataset divided into a 70:30 training-testing split. Models like KNN and Random Forest demonstrate higher accuracy and overall performance, while Gaussian Naive Bayes performs relatively poorly. These metrics provide insights into the models' predictive capabilities and aid in model comparison. Random Forest and KNN perform better due to their ability to capture complex relationships and adapt to different data types. Random Forest leverages ensemble learning, while KNN utilizes nearest neighbors for accurate predictions, making them robust and effective in handling various datasets. The graph plotted for the obtained results is given below.

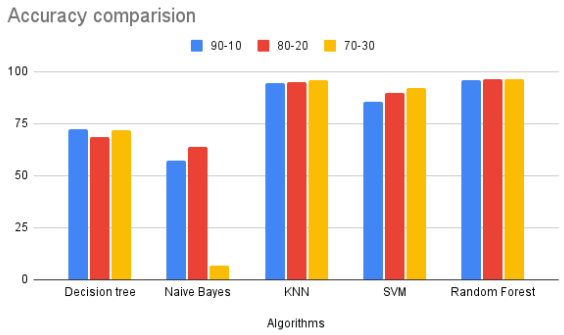

Figure 3: Accuracy comparison graph

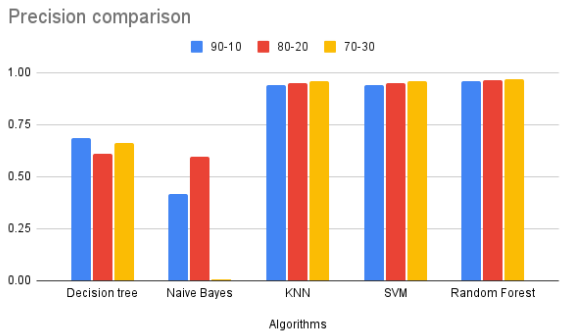

Figure 4: Precision comparison graph

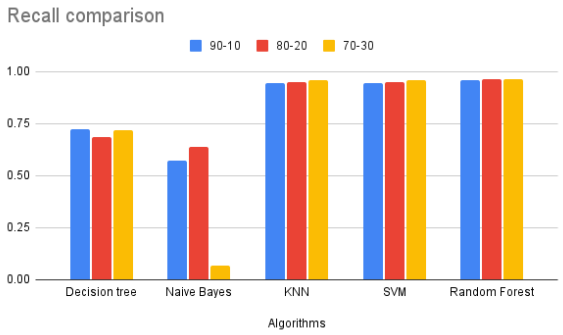

Figure 5: Recall comparison graph

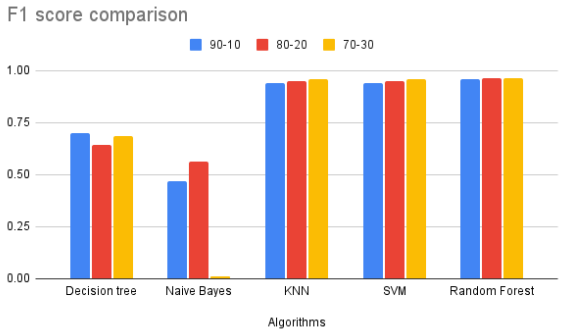

Figure 6: F1 score comparison graph

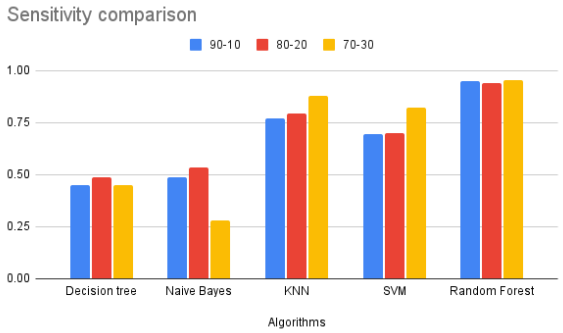

Figure 7: Sensitivity comparison graph

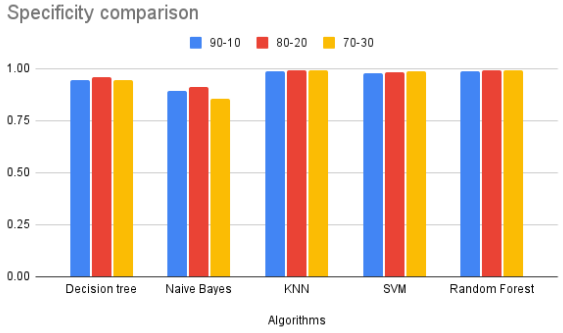

Figure 8: Specificity comparison graph

The x-axis would represent the models (Decision Tree, Gaussian Naive Bayes, KNN, SVM, Random Forest), and the y-axis would represent the metric values. The height of each bar would correspond to the value

of the metric, allowing for a quick and easy comparison of accuracy, precision, recall, F1 score, sensitivity and specificity across the models. The confusion matrix obtained on the collected dataset is given in Figure 9.

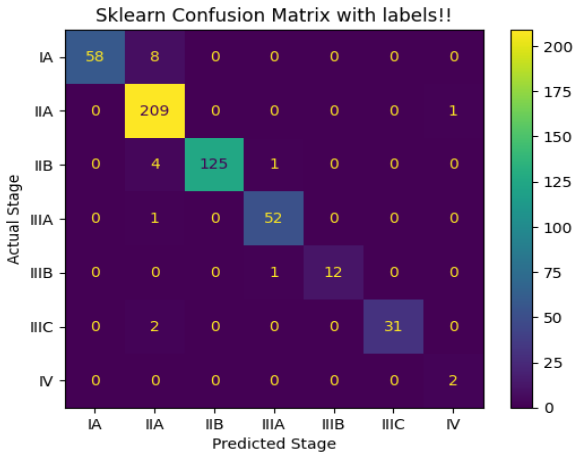

Figure 9: Confusion Matrix

After examining the dataset with five different algorithms, the random forest algorithm using the **70:30** split approaches produced the best results with 96.449**%** accuracy.

5. Conclusion

In conclusion, the proposed work on the normalization and classification of histopathology EHRs on breast cancer using NLP and ML approaches has successfully demonstrated the potential for leveraging advanced technologies to improve the analysis and utilization of healthcare data. By acquiring digital EHRs from the hospital and extracting relevant attributes, the project has enabled the identification of key information for determining the presence or absence of breast cancer. The findings in this work highlight the importance of data normalization and feature extraction in enabling effective analysis and decision-making in healthcare. By standardizing and structuring the EHRs, the project has facilitated a more efficient and comprehensive understanding of breast cancer-related data. The integration of machine learning techniques in this work has enhanced the predictive power and efficiency of cancer stage classification, providing healthcare professionals with valuable tools for improved patient care and treatment planning. The project's outcomes underscore the potential of NLP and ML approaches to revolutionize the field of oncology by streamlining processes, reducing errors, and enhancing patient care coordination.

Overall, this work serves as a stepping stone towards harnessing the power of advanced technologies to unlock the vast potential of EHRs in the context of breast cancer diagnosis and treatment. It opens up new avenues for future research and development, paving the way for more accurate and personalized approaches to cancer care, ultimately leading to improved patient outcomes and a positive impact on healthcare systems as a whole.

References

[1] Banerjee, I., Bozkurt, S., & Caswell-Jin, J. L. (2019). Natural Language Processing approaches to detect the timeline of metastatic recurrence of breast cancer.

[2] Paweł Filipczuk, Thomas Fevens, Adam Krzyżak, and Roman Monczak “Computer-Aided Breast Cancer Diagnosis Based on the Analysis of Cytological Images of Fine Needle Biopsies”, IEEE Transactions on Medical Imaging, Vol. 32, No. 12, December 2017.

[3] Alexander W. Forsyth MEng, Regina Barzilay, Kevin S. Hughes “Machine Learning methods to extract documentation of breast cancer symptoms from electronic health records, 2018

[4] Yoojoong Kim,Jeong Hyeon Lee, Sunho Choi , “Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health record, 2022.