International Journal of Information Retrieval Research
Volume 12 • Issue 1

Intelligent Information Retrieval for Reducing Missed Cancer and Improving the Healthcare System

Madhu Kumari, The NorthCap University, India*
Prachi Ahlawat, The NorthCap University, India

ABSTRACT

This study presents an intelligent information retrieval system that will effectively extract useful information from breast cancer datasets and utilized that information to build a classification model. The proposed model will reduce the missed cancer rate by providing a comprehensive decision support to the radiologist. The model is built on two datasets, Wisconsin Breast Cancer Dataset (WBCD) and 365 free text mammography reports from a hospital. Effective pre-processing techniques including filling missing values with regression, an effective natural language processing (NLP) parser, is developed to handle free text mammography reports. Balancing the dataset with synthetic minority oversampling (SMOTE) was applied to prepare the dataset for learning. Most relevant features were selected with the help of filter method and tf-idf scores. K-NN and SGD classifiers are optimized with optimum value of k for K-NN and hyper tuning the SGD parameters with grid search technique.

KEYWORDS

Breast Cancer, Classification, Fine Needle Aspiration, FNA, K-NN, Mammography, Natural Language Processing, NLP, Processing, SGD, SMOTE, WBCD

INTRODUCTION

Cancer is the major chronic health risk worldwide, with 12.7 million cases reported in 2008 and is predicted to increase to 21 million by 2030 (Society, 2011). Breast cancer is the most invasive life-threatening disease among females. Late diagnosis and the high cost of treatment lead to high mortality rates. Cancer is the lump in which cells begin to grow recalcitrant and can be mortal. These lumps are known as tumors, which can be benign (non-cancerous) or malignant (cancerous). Most breast cancers are discovered in the milk-producing glands, called lobules, or in the ducts that connect to the nipple. Tumors are small in the initial stages and may not cause noticeable symptoms; therefore, it is difficult to diagnose in the early stages. However, advancement in diagnostic techniques allows the oncologist to detect breast cancer during the developing stages. Accurate and timely detection of cancer helps oncologists make effective treatment strategies that can increase patient survival (Jemal, 2005). Early diagnosis requires a reliable and robust diagnostic system that can accurately

DOI: 10.4018/IJIRR.2022010102 *Corresponding Author

distinguish between malignant and benign tumors. Machine learning practices are gradually being brought together to improve diagnostic capabilities (Osareh, 2010; Kumari, 2017). With the assistance of machine learning techniques, the possibility of human error can be minimized, and healthcare data can be analyzed rapidly with a higher degree of accuracy (Dutra, 201). Statistically, early tumor detection increases the chance of successful treatment by 30% and improves overall survival rates (Elmore, 2003; Veronesi U, 2005). Consequently, efficient diagnostic techniques are required to detect tumors at an early stage in order to prepare effective treatment plans and strategies for long-term survival. Medical experts and researchers are increasing efforts to improve detection rates of the disease in the initial stages.

Mammography is used as a preliminary diagnostic screening exam to visualize potentially malignant breast tumors using a low dose x-ray with a detection accuracy of 80% (Elmore J. G., 2005). Each breast screening results in a minimum of one x-ray image and one free text report narrated by a radiologist. Each mammography report is assessed and categorized according to the Breast Imaging-Reporting and Data System (BI-RADS) (Liberman, 2002), a standardized classification system given by the American College of Radiology for risk assessment and offers uniformity to radiologist reports. Mammograms have the potential to identify tumors several years before the development of physical symptoms; however, false positives and negatives are not uncommon. The double evaluation of mammograms by two different radiologists is recommended to reduce the proportion of misdiagnoses; however, this practice increases the workload, is costly and time-consuming (Brown, 1996).

To confirm detection, Fine Needle Aspiration (FNA) is used as an additional microscopic analysis and has a detection accuracy of 65-98% (Giard, 1992). A fine needle is used to extract breast tissue for pathological assessment. A comprehensive report is provided on the cell type, including comments on malignancy. Moreover, a surgical biopsy is a diagnostic technique with a detection accuracy of ~100%. The accuracy of the visual interpretation of mammograms and FNA fluctuates extensively and is not the most reliable breast cancer detection method. Surgical biopsy reveals most of the malignant cases; however, this technique is invasive and expensive. Regardless of the availability of contemporary diagnostic procedures and advances in healthcare systems, missed breast cancer continues (Singh, 2007). Missed cancer during the diagnosis is the most damaging and expensive kind of investigative error, also known as diagnostic errors.

Diagnostic errors are defined as missed, false, or delayed detection of a medical disease by a definitive test or screening (Maskell, 2019). The failure to detect a tumor during cancer screening increases the mortality rates of cancer patients. Common failures that lead to diagnostic errors include inadequate sampling or radiologic technique; human error (pathologist or radiologist); human error while documenting; inexperience of the physician in recognizing symptoms. Approximately 30% of errors occur during radiology investigations, with 75% of medical negligence claims against radiologists associated with diagnostic errors (Lee, 2013). Vigilant analysis and implementing standardized practices are required to overcome such unsolicited events.

Machine learning and data science techniques have the potential to discover hidden knowledge and statistical regularities from the available historical data to make accurate predictions using new data. A large number of studies are reported in the literature that utilizes different machine learning techniques to detect the malignant tumor (Cruz, 2006). In recent years, researchers proposed many data-driven rational classification models for breast cancer classification (Liu, 2019; Montelongo González, 2020; Esmaeili, 2020; Schaffter, 2020). The major limitations of present state-of-art methods is that they only aim to increase the classification accuracy and do not attempt to improve the precision and recall value. The precision value of the classification system is the measure of false-positives. The false-positive value is the number of samples/patients that are incorrectly identified as positives by the classifier. This situation could lead to unnecessary procedures or treatment. Moreover, recall value measures the false-negatives. The false-negative is the measure of the number of positives samples that the classification model fails to identify. An effort must be made to reduce the missed positive samples as it results in patients foregoing needed treatment and leads to severe consequences that can

be mortal. Additionally, the existing classification model cannot work on both types of datasets, i.e., structured medical data and semi-structured medical data. It is known that a considerable amount of structure, semi-structured and unstructured data are generated in the medical domain corresponding to the different medical activities.

Therefore, the objective of this study is to develop a novel decision support system based on a machine learning approach that provides complete decision support for breast cancer. The key contribution of the proposed work are as follows:

1. In a medical environment, data is collected and processed in a wide variety of formats for different purposes. Therefore, this work presents a novel approach that can simultaneously handle two types of datasets, i.e., structured as well as semi-structured data.
2. The performance of the machine learning model is greatly influenced by the quality of data utilized for the training and testing of the classifier. Therefore, to extract the most useful information for the medical data, two different intelligent information retrieval systems are proposed one for the structured dataset and the other for the semi-structured dataset. The information extraction systems utilize effective pre-processing practices to improve the quality of data. It involves handling missing values, standardizing the data, balancing the unbalanced dataset, NLP parser, reducing the dimensionality by selecting the most relevant and best features for the model learning in the feature selection step.
3. The extracted information became the input to the machine learning model for the decision support. The selection of the classifier for the machine learning model is very crucial and greatly affects the model's performance. Therefore, the two most widely and successfully used classifier, i.e., K-NN and SGD, are used to train and test the machine learning model.
4. The performance of the classifiers is greatly influenced by the choice classifier parameters. Therefore, both the classifiers, K-NN and SGD, are optimized to the best parameter value.
5. Finally, the proposed model's performance is evaluated on the model accuracy, precision, recall, and F1-score and compared with the present state-of-the-art methods.

BACKGROUND

The researchers used a range of machine learning techniques for predicting susceptibility (Jerez-Aragonés, 2003; Kate, 2017), diagnosis (Kim, 2012; Bilal, 2013; Kumari M. &., 2018), recurrence (Macías-García, 2020; Alzu'bi, 2021; Tseng, 2019) and survivability (Sasikala, 2018; Bouyer, 2017) of breast cancer in women.

By examining the available literature, it has been observed that most of the reported studies utilize WBCD structured data for the classification of benign and malignant tumors. An automatic diagnostic model was proposed for detecting breast cancer. The dimensionality of the WBCD is reduced by using association rules and the model is trained with a neural network classifier. A three-fold CV is used in the testing phase, which results in a classification accuracy of 95.6% (Karabatak, 2009). The classification accuracy was improved to 96.87% for the same dataset using a rough set algorithm to select the most predictive feature and SVM classifier for training and testing (Chen, 2011).

Setiono et al. (2000) proposed a method that improves WBCD classification performance with an accuracy of 98.10%. Before applying the classifier, the dataset is first pre-processed. Attributes with missing values are ignored and a Neural Network classifier is trained on most predictive features selected by one hidden layer of neural networks. This technique reduces the training time and enhances the accuracy of the model. A classification modal least square support vector machine (LS-SVM) was proposed by Polat et al. (2007) to detect breast cancer in women using cross-validation and attained an accuracy of 98.53%. Yeh et al. designed a breast cancer detection model with an accuracy of 98.71% by using machine learning techniques with swarm optimization (Yeh, 2009). For breast cancer detection, Akay, M. F. (2009) proposed a new method using SVM and attained the classification

accuracy of 99.02%. Another hybrid method for breast cancer diagnosis was developed where the size of the training dataset was reduced by using an artificial immune system (AIS) artificial intelligence algorithm. The fuzzy weighing technique was adopted to consolidate the effect of using a distance-based algorithm. The above-proposed model was then trained using a K- Nearest Neighbor (KNN) classifier. The fuzzy-AIS-KNN model with 10-fold Cross-Validation (CV) achieving an accuracy of 99.14% (Şahan, 2007). The classification accuracy for breast cancer is further improved to 99.26% by utilizing Artificial Neural Network (ANN) over biological metaplasticity (Marcano-Cedeño, 2011).

The aforementioned state-of-art methods work for only the structured dataset and fail to capture and utilize the information available with mammographic screening. It gained the attention of many researchers and became the area of interest.

A number of statistical models were developed where mammogram images are enhanced to extract the most predictive features and then became the input for the statistical learning model to classify benign and malignant tumors (Sharma, 2006; Tang, 2009; Anitha, 2016). Burnside et al. developed a probabilistic model to classify mammography reports (Burnside, 2009). Wieneke et al. gave a solution for abstracting free text structured data from the breast pathology report. The SVM classifier was used in combination with appropriate NPL techniques to analyze the data and recognize hidden patterns. A total of 6,965 manually abstracted pathology reports (2009-2011) were divided into 80% of the training sample and 20% of the test sample. The most relevant features were obtained through the chi-square statistical test, which fed to the classifier for learning produces promising results with an 80.0 F1 score and 77.0% precision value (Wieneke, 2015). Gao et al. developed a rule-based NLP system to extract mammographic findings from free-text mammography reports accurately. Mass, calcification, asymmetry, and architectural distortion were the findings extracted from 93,705 mammography reports using a dictionary lookup method on the SAS platform for ensuring portability and ease of implementation. The developed system coded 96-99 samples correctly out of 100 samples and performed better than earlier studies (Gao, 2015). Thiebaut et al. analyzed 14,029 textual clinical reports from breast cancer patients from 2000 to 2017. They developed an innovative NLP-based solution that allows for a multi-targeted analysis of free-text medical records. They used many techniques such as automatic synonyms detection and typographic error corrections to get several indicators such as tumor size, type, hormonal response to create various statistical studies on the corpus (Thiebaut, 2017).

Gupta et al. proposed an unsupervised information extraction system using the dependency-based parse tree with distributed semantics to generate controlled relation about observations from the mammography reports (Gupta, Automatic information extraction from unstructured mammography reports using distributed semantics, 2018). Rani et al. developed an automated system to extract the breast tumor details using pattern matching rules and applied a PTNM protocol to generate the pathological classification of breast tumors. The extracted and classified breast tumor value was analyzed against the gold standard values obtained from manually scrutinized reports. The developed system has showcased an average accuracy of 79.53% (Rani, 2015). Bozkurt et al. proposed a model to extract breast cancer diagnostic information from free text mammography narrative reports by combining NPL with the Bayesian network to provide decision support (Bozkurt, 2016). BI-RADS descriptors and diagnostic breast cancer information have been extracted from mammography reports and supplied as an input to the decision support system, which helps dictate the report as an output.

The state-of-art methods reported in the literature considered either structured or semi-structured data associated with breast cancer classification and analysis. Therefore, an advanced breast cancer prediction model is required to work with both structured and semi-structured datasets. Additionally, present state-of-art methods aim to improve the classification accuracy and little efforts have been made to reduce false-negative cases, i.e., the missed cancer rate. This study proposes an automated decision support system to provide a reliable and quick recommendation from screening sample analysis to address the above issues. The proposed system supports both FNA (Structured dataset) and mammography breast cancer screening (semi-structured) techniques. The approach relied on a

strong amalgamation among standard NLP and supervised machine learning methods that present an effective trade-off between required manual effort and generalizability of the system. The objective of the work is to provide real-time support to radiologists' to accurately predict malignancy, reducing the number of false negatives and the chance of human error. Consequently, it endows with a process that could integrate decision support into the radiologist and pathologist analysis process, devoid of the need for equivalent data entry processes.

MATERIALS AND RESEARCH METHODOLOGY

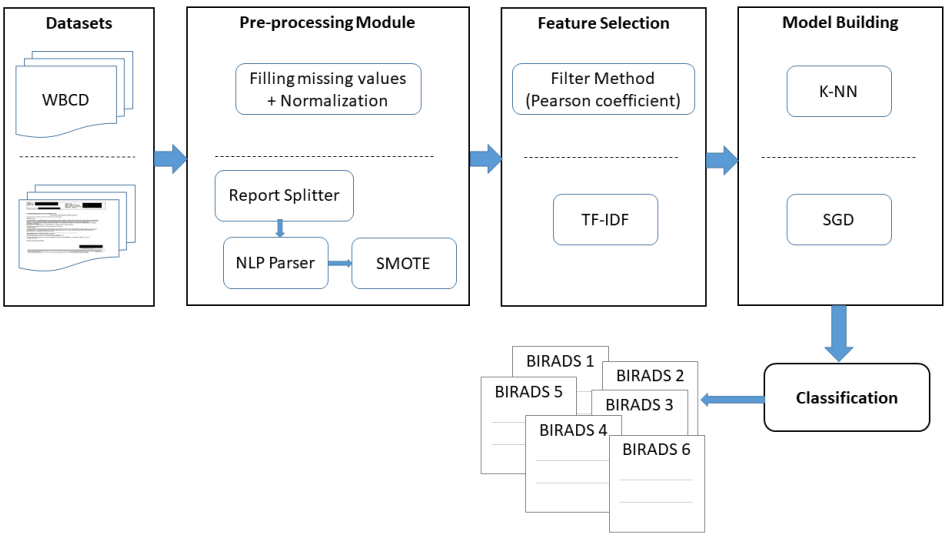
This section presents the different materials and methods applied in the proposed study. The architecture of the proposed work has been illustrated in Figure 1. The breast cancer screening practices generate two types of the dataset, one is structured and the second one is semi-structured. The proposed model can handle both types of datasets and provide a reliable recommendation to the radiologist during decision-making. The detailed functionality of the processing blocks is illustrated in the following subsections.

Datasets

In this work, two datasets, i.e., WBCD and free text mammography reports, have been used to work upon structured and semi-structured datasets.

Corpus-I (Structured breast cancer dataset): Data was collected from the precise domain to facilitate learning during analysis. The Original WBCD was obtained from the UCI [58] for this experiment. This dataset comprises nine numeric-valued continuous data type attributes. The statistical properties of all the attributes have been computed to gain better insight into the dataset, as shown in Table 1. The target attribute for this dataset is a dichotomous variable, i.e., a binary response variable. The 2 and 4 value of the target attribute represents benign and malignant cancer, respectively. The dataset has a total of 699 instances; among them, 35.0% are malignant and 65.0% are benign instances. The dataset suffers from a considerable number of missing values. Sixteen examples with missing values are identified, which are represented by "?" in the

Figure 1. Proposed system architecture



data set. These unknown or missing values could have a significant effect on the interpretations derived from the data. The detailed descriptions of all nine attributes are shown in Table 1.

A heat map chart is also created to understand the correlation between dataset features and given in figure 2. Heat map chart is the most effective statistical test to visualize the association between different dataset features. The correlation score between each feature with every other feature is calculated and represented in the heat map chart. The correlation score ranges between -1 and +1. The features with a correlation score close to 0 indicate no linear relationship between the two attributes. Moreover, the features that are strongly correlated are having correlation values close to 1.

Corpus-II (Unstructured data from mammography reports): For this study, a total of 357 mammography reports are collected from one of the reputed hospital in Northern India. An example of a mammography report is given in Figure 3 and documents the following information: patient's personal information (name, age, sex, address); referred by (physician's name, designation); specimen type; instrument specification (a brief description of the screening instrument); observations (dictations of the observed findings); impression (a conclusion about the lesion status based on radiologist observation during a screening).

Radiologist dictated impressions of lesions found in each mammogram from 0 to 6: incomplete screening, negative, benign, probably benign, suspicious, highly suggestive for malignancy and malignant, respectively, as per the ACR BI-RADS. The BI-RADS terms facilitated uniformity of terminology used in the text about the named entity through the ontologies and became the typical approach for mapping observations to canonical meaning. The mammography reports are further analyzed to understand the data distribution among different BI-RADS categories. According to the BI-RADS categories, the distribution of data samples in the mammography dataset is shown in Figure 4. It is observed that BI-RADS-1 and BI-RADS-2 account for 68 percent of dataset samples, while the remaining 32 percent are distributed among BI-RADS-3, BI-RADS-4, BI-RADS-5, and BI-RADS-6.

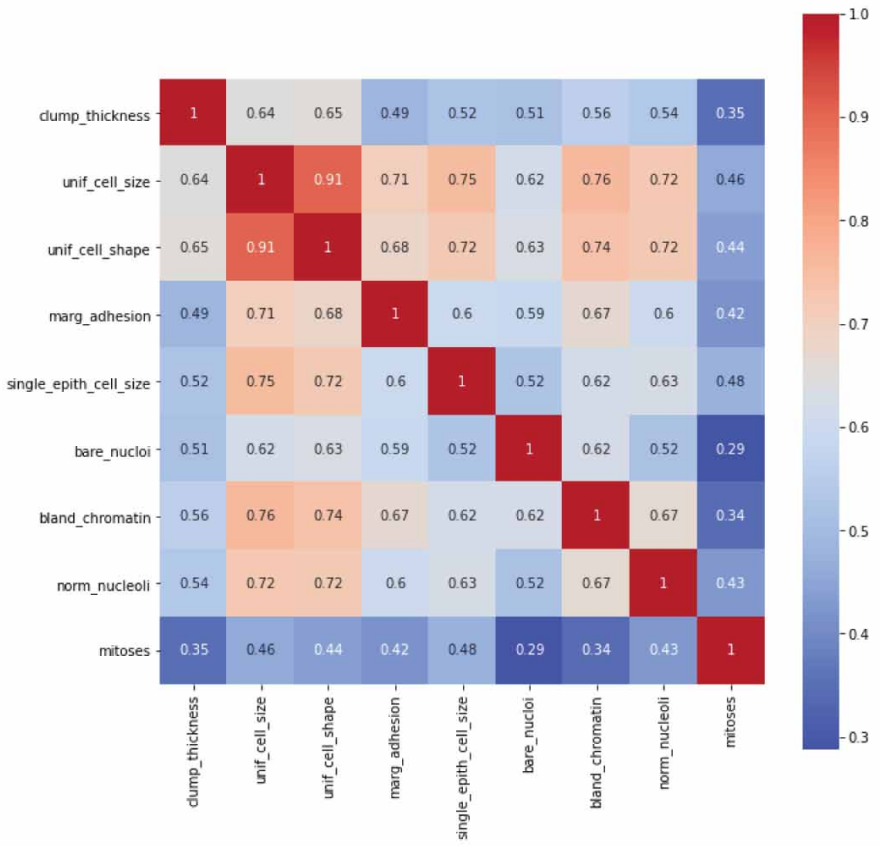
Pre-Processing Module

Data gathered for the analysis was raw and contained missing and superfluous values. This data was inappropriate for the analysis and significantly affected the performance of the classifier. Therefore, it was important to pre-process the dataset under consideration before training it on a classifier in order to enhance the ability to learn the unseen patterns in the dataset. Two different pre-processing

Table 1. Detailed description of WBCD dataset

S.No.	Features name	Values	Mean Value	Standard deviation
1	clump_thickness	1 to 10	4.44	2.83
2	unif_cell _size	1 to 10	3.15	3.07
3	unif_cell_shape	1 to 10	3.22	2.99
4	marg_adhesion	1 to 10	2.83	2.86
5	single _epith_cell_ size	1 to 10	2.23	2.22
6	bare _nuclei	1 to 10	3.54	3.64
7	bland _chromatin	1 to 10	3.45	2.45
8	norm_nucleoli	1 to 10	2.87	3.05
9	mitoses	1 to 10	1.60	1.73

Figure 2. Statistical correlation between all the WBCD features



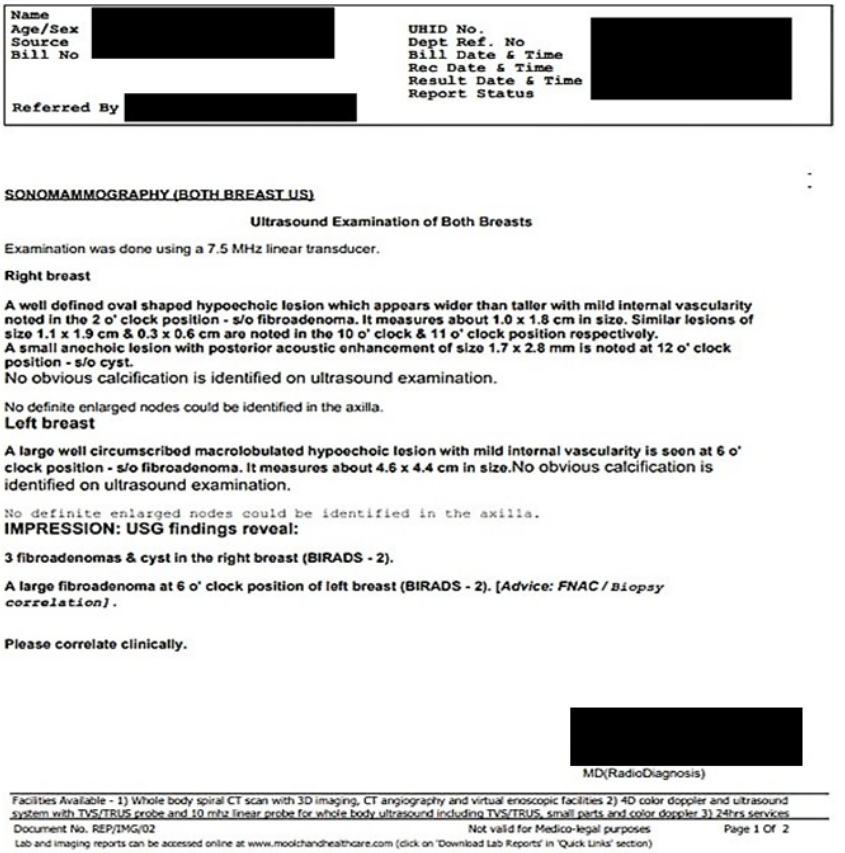
architecture was proposed for the corpus-I and corpus-II. A detailed explanation of the proposed pre-processing architecture is given in the below sections.

Corpus-I: Structured Breast Cancer Dataset

The dataset suffers from missing values and these values can adversely degrade the classifier's performance. To address this concern, this paper proposes a pipelined pre-processing module that can adequately handle the aforementioned challenge. The proposed pre-processing pipeline is illustrated in Figure 5. It comprises of filling missing values, standardizing data and selection of features.

- **Filling Missing values:** The dataset is processed for the imputation of missing values. The dataset is scanned for missing values that were represented by"?". Such values can significantly influence the results derived from the data. In this study, the regression algorithm is used for finding and filling the missing values. The correlation between features contains valuable information which is required during the feature selection process and must be preserved. Therefore, the missing values are filled with the regression algorithm as it maintains the original correlation between features of the dataset. Pseudo-code for filling up the missing values is given in algorithm 1. It replaces all those null values initially with "nan". Each missing value is calculated and imputed by using all other values of the specific feature in the dataset.
- **Standardization:** Data in a medical setting is collected for a wide range of purposes and may be stored in different formats. Data standardization is used to remove all such internal inconsistencies

Figure 3. Sample mammography report



Name [redacted]
Age/Sex
Source
Bill No

UHID No.
Dept Ref. No
Bill Date & Time
Rec Date & Time
Result Date & Time
Report Status

Referred By [redacted]

SONOMAMMOGRAPHY (BOTH BREAST US)

Ultrasound Examination of Both Breasts

Examination was done using a 7.5 MHz linear transducer.

Right breast

A well defined oval shaped hypoechoic lesion which appears wider than taller with mild internal vascularity noted in the 2 o' clock position - s/o fibroadenoma. It measures about 1.0 x 1.8 cm in size. Similar lesions of size 1.1 x 1.9 cm & 0.3 x 0.6 cm are noted in the 10 o' clock & 11 o' clock position respectively.
A small anechoic lesion with posterior acoustic enhancement of size 1.7 x 2.8 mm is noted at 12 o' clock position - s/o cyst.
No obvious calcification is identified on ultrasound examination.

No definite enlarged nodes could be identified in the axilla.

Left breast

A large well circumscribed macrolobulated hypoechoic lesion with mild internal vascularity is seen at 6 o' clock position - s/o fibroadenoma. It measures about 4.6 x 4.4 cm in size. No obvious calcification is identified on ultrasound examination.

No definite enlarged nodes could be identified in the axilla.

IMPRESSION: USG findings reveal:

3 fibroadenomas & cyst in the right breast (BIRADS - 2).

A large fibroadenoma at 6 o' clock position of left breast (BIRADS - 2). [*Advice: FNAC / Biopsy correlation*].

Please correlate clinically.

[redacted]
MD(RadioDiagnosis)

Facilities Available - 1) Whole body spiral CT scan with 3D imaging, CT angiography and virtual enoscopic facilities 2) 4D color doppler and ultrasound system with TVS/TRUS probe and 10 mhz linear probe for whole body ultrasound including TVS/TRUS, small parts and color doppler 3) 24hrs services

Document No. REP/IMG/02 Not valid for Medico-legal purposes Page 1 Of 2

Lab and imaging reports can be accessed online at www.moolchandhealthcare.com (click on 'Download Lab Reports' in 'Quick Links' section)

Figure 4. Data distribution in the mammography dataset according to the BI-RADS categories

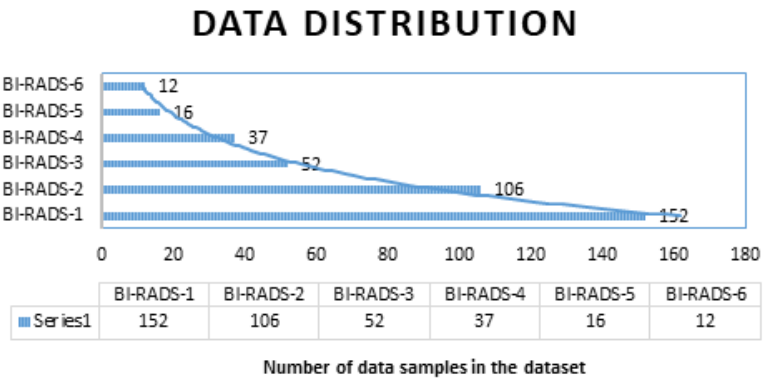
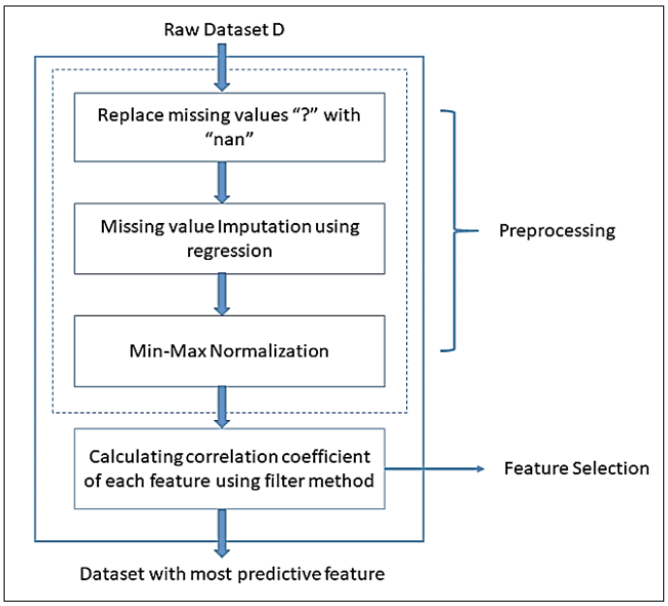


Figure 5. Proposed pipelined pre-processing module for WBCD



Algorithm 1. Pseudocode for filling missing values

```
//Pseudocode for filling missing values
//Dataset: d, Features: f, index: x,
//number of instances in feature: n
begin
set x=0
repeat while(j<=n)
  if f[j] is null
    predict value "v" by fitting to
    regression
    impute f[j] with value "v"
  end of if
  x=x+1
end of while
end
```

and bring all the data in a consistent format. The proposed work uses the min-max normalization technique to standardize the entire dataset into one standard range. Algorithm 2 contains the pseudo-code of the min-max standardization algorithm.

- **Feature Selection:** The most predictive features improve accuracy and reduce the overall complexity of the prediction system. The most predictive features in this dataset are retrieved using the filter method. The filter method uses a mathematical function given in equation 1 to find the correlation among independent and dependent (target) variables. The features are selected based on their correlation coefficient values. This technique is known as Pearson's

Algorithm 2. Pseudo-code for min-max normalization

```
//Pseudocode for min-max normalization
// D: Dataset
begin
define dataset_minmax(D)  // Find the min and max values for each column
  min_max=list()
  k=1
  repeat for i=1 to len(D[0])
    column_value=[row[i] foreach row in D]
    smallest_value=min(column_value)
    largest_value=max(column_value)
    append to min_max([smallest_value],[largest_value])
  end of for
  return min_max
end of dataset_minmax
define normalize(D,min_max) // Normalize dataset columns to the range 0-1
  row=1
  repeat for each row in D
    repeat while j<=len(row)
      row[j]=(row[j]-min_max[j][0])/ (min_max[j][1]-min_max[j][0])
    end of while
  end of for
end of normalize
```

linear correlation. Highly correlated features with the target variable are included in the final set. Pearson's linear correlation: Consider the n-dimensional dataset D^n, where f ϵ D^n and target outcome, t ϵ D. Pearson's linear correlation coefficient is defined as:

$$r = \frac{\sum_{i=1}^{n}\left(f_i - \overline{f}_i\right)\left(t_i - \overline{t}_i\right)}{\sqrt{\sum_{i=1}^{n}\left(f_i - \overline{f}_i\right)^2} \cdot \sqrt{\sum_{i=1}^{n}\left(t_i - \overline{t}_i\right)^2}} \quad (1)$$

where f_i and t_i are the ith value of f and t, respectively. The value of r = +1 signifies a positive correlation among the independent variables and the target variable and r= -1 represents a negative correlation among the independent variables and the target variable. Table 2 shows the correlation score between independent features and the target variable. Based on the correlation coefficient score, the top five (bare_nuclei, unif_cell_size, unif_cell_shape, bland_chromatin, norm_nucleoli) features are selected for model building.

Corpus-II: Unstructured Data From Mammography Reports

Mammography pdf reports are narrated by the radiologist in free-text natural language. It contains all the necessary diagnostic information about the sample, such as the interpretation of mammography screening, its evaluation and its findings. The mammography report data are highly unstructured and complex and thus not suitable for the machine learning model. Therefore, to effectively extract and utilize the valuable information from this rich source of available data, an efficient information extraction module named "NLP parser" is proposed using NLP techniques.

NLP is the application area that offers compelling exploration, understanding, and extraction of valuable information from natural language datasets for building machine learning models. The detailed architecture of the NLP parser is given in figure 7 and explained in detail below:

Figure 6. Correlation between independent and dependent target variable of WBCD dataset features

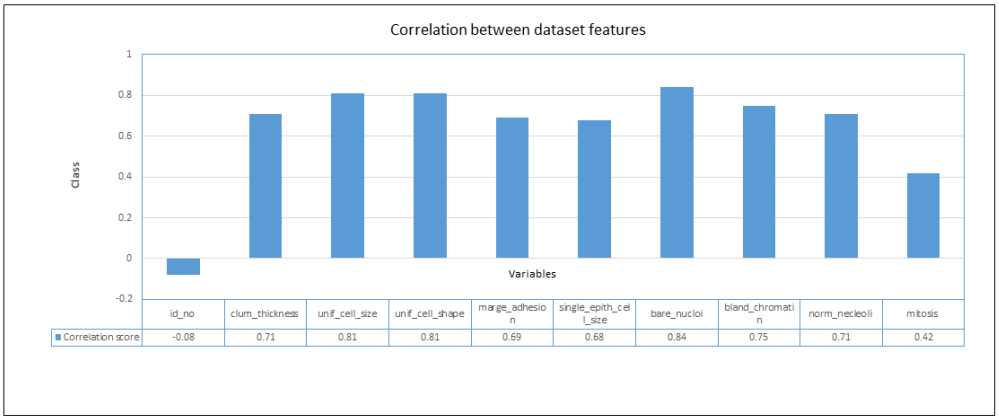
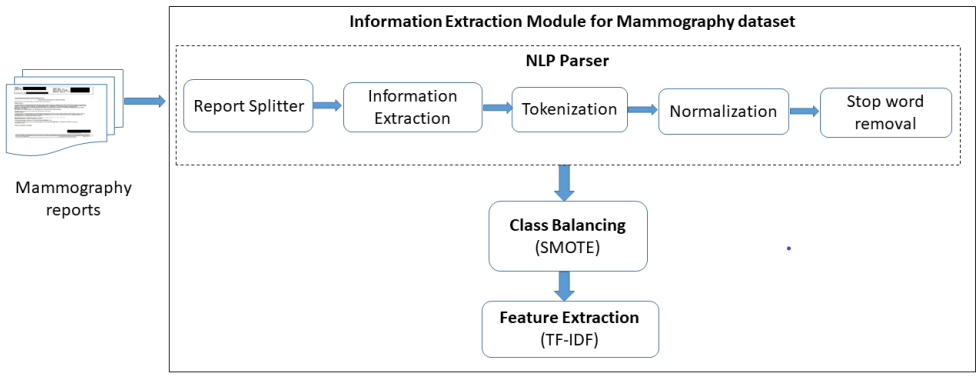


Figure 7. The architecture of the proposed NLP parser for information extraction from mammography reports



- **NLP Parser:** The mammography reports are divided into different sections; for example, section 1 contains personal information of the patient, section 2 contains the specification of an instrument used, observations and impression and section three contains radiologist information (name, signature etc.) and some other information. Only radiologist observations regarding the lesion and impressions, including the textual findings and corresponding BI-RAIDS ratings, are required for the learning. Therefore, a python-based section splitter is developed to separate the different sections of mammography reports. To extract the information from the mammography reports, a Python-based module is developed.

The PyPDF2 library was used to extract the specific information from the mammogram pdf and stored in a structured format to be consumed for learning. An object of pdf file reader was created and initialized by giving the path of the mammogram pdf files. Each page from the pdf was retrieved by providing the page number. The observations and impressions are identified and extracted into a separate text file(mamodata.txt).

The structured file extracted from the pdf mammography report was still not suitable for the statistical model as the observations and impressions were in the free text format narrated by the radiologist and flooded with many irrelevant and noisy terms. This document firstly required cleaning and pre-processing before applying any learning. The various phases involved in the NLP parser are tokenization, normalization and stop word removal and feature extraction.

Firstly, a free text dictated mammography reports, consisting of a sequence of characters, words, codes etc., required segmentation into linguistic units such as words, punctuation, numbers, alpha-numeric codes etc. and this was completed by tokenization. Here, the sequence of string from the medical records was segmented into smaller pieces, termed tokens. These tokens are methodologically helpful in recognizing patterns displaying significant collocation. Here, we have used a standard white space approach for tokenization. Normalization was a process of standardizing the document in one uniform format. Mammography narrative reports were not in the uniform format; therefore prior to any learning application, tokens derived from the mammography reports were converted into lower case. The final step was to remove stop words. Stop words are frequently occurring insignificant words used as connectors in sentences and do not contain any meaningful information. These words were removed to boost the performance of the feature selection practice.

- **Class Balancing:** During the statistical analysis of the mammography dataset, it was observed that the dataset is unbalanced. 68% of the dataset samples belong to the BI-RADS-1 and BI-RADS-2, and the rest 32% data samples are distributed among BI-RADS-3, BI-RADS-4, BI-RADS-5 and BI-RADS-6. This imbalanced data distribution can cause biasness against some BI-RADS categories and adversely affect the classifier's performance. To address this concern, the SMOTE technique is applied to balance the dataset. SMOTE counteracts the dataset by generating synthetic samples of the minority class.
- **Feature Selection:** Selecting appropriate features to construct vector space was a crucial step in pre-processing. High dimensional feature space consists of a high proportion of redundant, irrelevant, and noisy features that affect classifier performance. Thus, identifying and extracting only the relevant features by ignoring the others was essential for the learning algorithm. Hence, feature selection is widely used to reduce the dimensionality of the feature space and enhance the classifier's accuracy and efficiency. Selecting the most promising features from the original document should consider the domain and algorithm characteristics. All the features in the document were rated according to the predetermined measure of relevance of word importance. In feature selection processes, only the highest scored features were kept. The selected features preserve the original meaning and provide a clear and improved understanding of the data and the statistical algorithm. The most relevant features enhance the accuracy of the classifier, efficiency of the model and improve scalability. Interesting features can be found by simply including words with the highest word count in each document, but the problem with this approach was that it gave higher scores to longer documents when compared to shorter ones. To overcome this barrier, term frequency (TF), defined as #count (word)/number of total words, was used for each document. We could also reduce the score of more common words like (the, is, an, etc.), which occur commonly in all documents. This was termed term frequency time's inverse document frequency (TF-IDF). Here, we have used a TF-IDF approach to extract the most relevant features.

The frequency of bi-grams was represented by numerical values in the document for reckoning on the attribute variable. All the equivalent frequencies formed a document vector, which later refers as a dictionary in the document. For the classification task, all terms were considered as features and participated in the learning process.

TF(t, d) represents the term frequency. It is measured as the number of times term (t) appeared in the document (d). A high TF value only indicates that the particular term frequently appears in the document and does not provide any information about the term's significance. However, many times

the term with high term frequency was found least significant for the learning process. To understand the significance of the term in the document, the inverse document frequency was calculated. Free text mammography reports were narrated by a radiologist depending on the observation derived during the screening process. Therefore, identifying the most significant and important terms for learning was essential. Inverse document frequency for each term in the document was calculated in order to distinguish between significant terms for the learning process and commonly used insignificant terms. The calculation considered the frequency of terms across the whole document:

$$IDF\left(t\right) = \log\left(\frac{\left|D\right|}{\left(DF\left(t\right)\right)}\right) \tag{2}$$

D/DF (t) was the total frequency of all documents under consideration having term (t). The calculated value was normalized by using a logarithm. The significance of a particular term was evaluated by calculating the TF-IDF score by taking the product of TF(t) and IDF(t):

$$TF - IDF = TF\left(t,d\right) * IDF\left(t\right) \tag{3}$$

The significance of any term in the document is evaluated on the basis of their corresponding TF-IDF score. The highest TF-IDF score achieved here was 4.7. Figure 8 demonstrates the TF-IDF score per token and the corresponding pseudo-code is given in algorithm 3.

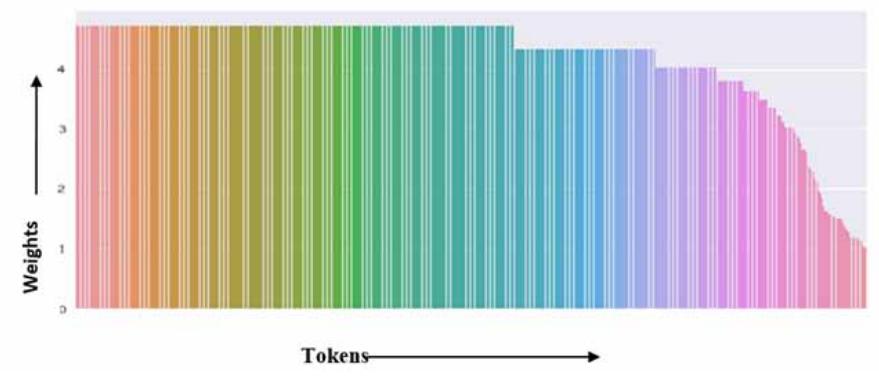
Model Building

The proposed model is built on optimized K-NN(corpus-I) and SGD(corpus II) . The methodology is further explained in the subsequent sections.

Corpus-I: K-Nearest Neighbors (KNN)

K-NN is a versatile algorithm that performs extremely well for binary classification problems (Kumari M. &., 2018; Şahan, 2007). It is a non-parametric classifier and does not make any assumptions on the data under consideration. This non-parametric behavior of KNN makes it a low bias classifier. Therefore, the K-NN classifier is utilized to build the proposed machine learning model for corpus-I. The K-NN classifier is trained on the most relevant features identified in the previous sub-section to

Figure 8. Term frequency-inverse document frequency per token of mammography dataset



Algorithm 3. Pseudo-code for TF-IDF feature extraction technique

```
//Pseudo-code 3: Feature Extraction
// Dataset: df

define feature_extraction_txt
  df_t=TD ("features", "Final class")
  df_t.dropna (inplace=True)

  //initializing Tfidf vectorizer with bigarms
  count_vect_x=Tfidf(ngram_range(1,2))

  //Crearting Tfidf document term matrix
  X_train_count_x=count_vect_x.fit_transform(df_t["features"])
  features=count_vect_x.get_feature_name()
  tfidf=dict(zip(count_vect_x.get_feature_names(),count_vect_x.idf_)

  //dictionary mapping the tokens to their tfidf score
  tfidf=df(columns=['tfidf']).from_dict(dict(tfidf), orient='index')
  tfidf.column=['tfidf']

  //saving the vocabulary generated by the document term matrix and saving it
  in the variable model_vocab model_vocab=count_vect_x.vocabulary

  print(model_vocab)
end of feature_extraction_txt
```

predict the presence of diabetes for unknown samples automatically. The pseudo-code of the same is given in algorithm 4. K-NN classifier calculates the similarity distance among each labeled training sample and the unknown test samples. It can use different distance metrics for identifying the nearest neighbors such as Euclidean distance, Manhattan distance, Minkowski distance for a continuous variable and hamming distance for categorical variables. In this study, the Euclidean distance metric is applied to calculate the relationship among the data samples which is termed as similarity distance. The formula for the same is given below:

$$SimilarityDistance = \sqrt{\sum_{i=1}^{k}\left(x_i - y_i\right)^2} \tag{4}$$

where x_i and y_i are two data points in the feature space.

The computed similarity distance is used to identify the K nearest neighboring data samples. The K-NN classifier returns the class of the majority of K nearest data samples as the class of the test sample.

Algorithm 4. Pseudo-code for optimized K-NN model for structured FNA dataset

```
//Pseudo-code 4: Knn model
//T: training dataset, C : Class labels of T, y: unknown samples,
  d: Distance Matrix

KNeighborsClassifier(T, C, y)
 for (i=0; i<=m; i++)
      Calculate distance d (Ti  y)
calculate set Z having indices for the k smallest distances d (Ti  y).
return class with {Ci where i ∈ Z}
```

- **Optimizing K-NN classifier:** The performance of KNN is significantly dependent on the value of K. Higher value of K brings more bias, whereas a too low value makes it more sensitive to noise.

To find the optimal value of K, a validation set is created from the existing training data set. Therefore, the original training data set is divided randomly into a new 80% of the training set and 20% of a validation set. This validation set is used to evaluate the K-NN classifier's performance for all K values from 1 to 100. This experiment showed that the K-NN attained the highest accuracy with the value of K as 15. Therefore, the final model is trained with K=15.

- **k- Fold Cross-Validation (CV):** k fold CV is a statistical method that minimizes the biasness associated with random training data samples. It involves splitting data samples into mutually exclusive k identical size subsamples. This approach takes k iterations. Each iteration takes one subsample to validate the model and the remaining k-1 subsamples for training the model. Therefore, every kth subsample is used just once for training or validating the model. The value of k is set to 10 for this study. The final accuracy of the model is the average accuracy achieved in all the k iterations. Mathematically, it can be given as:

$$CrossValidationAccuracy = \sum_{i=1}^{k} a_i \tag{5}$$

Corpus 2: Stochastic Gradient Descent

A supervised statistical classifier, stochastic gradient descent (SGD) has been used successfully for learning the insights and relationships among the high-dimensional dataset and to make predictions based on acquired learning (Bottou, 2010; Kabir, 2015; Kumar, 2015; Nguyen, 2016). Therefore, the SGD classifier is utilized to carry out the text classification under the proposed setting. Moreover, SGD is an improvement over simple gradient descent, which is quite slow for large training examples. The gradient descent algorithm scans all training samples to calculate the gradient and find the update required for the optimization parameter. This is a slow process for large dataset problems. SGD is an improvement over standard gradient descent as the optimization parameter is updated by calculating the gradient of only one randomly selected training instance each time, rather than considering the complete training sample each time. The SGD process by choosing w to minimize loss function Q(w). To make an initial guess for w, a search algorithm is used, and the value of Q is modified repeatedly to generate output from Q. Mathematically, formulation of the update process is given as:

$$w_{t+1} = w_t - \eta \frac{\delta}{\delta w} Q(w;x^i;y^i) \tag{6}$$

where w is the parameter that determines its behavior. Therefore, with some given w_t and considering one example at a time at constant step size, the next move (w_{t+1}) is calculated to reach the solution. $x^i;y^i$ is the training example and η is the step size for the algorithm. SGD was initiated by randomizing the data sample to avoid any biasness. It randomly selected one training sample each time and updated the weights on the basis of the calculated gradient for that sample only, instead of using a complete gradient as in simple gradient descent. As a result, the algorithm converges faster by taking large numbers of tiny steps. The pseudo-code of the SGD is given in algorithm 5.

- **Optimizing SGD by hyper-tuning the parameters:** The SGD algorithm performs well with sparse and high dimensional data. Moreover, the performance of the SGD is greatly influenced by choice of hyper-parameter of the algorithm. For example, a large learning rate would converge faster but may overshoot the minima, whereas smaller learning rate results are more stable with a

Algorithm 5. Pseudo-code for SGD classifier for semi-structured mammography model

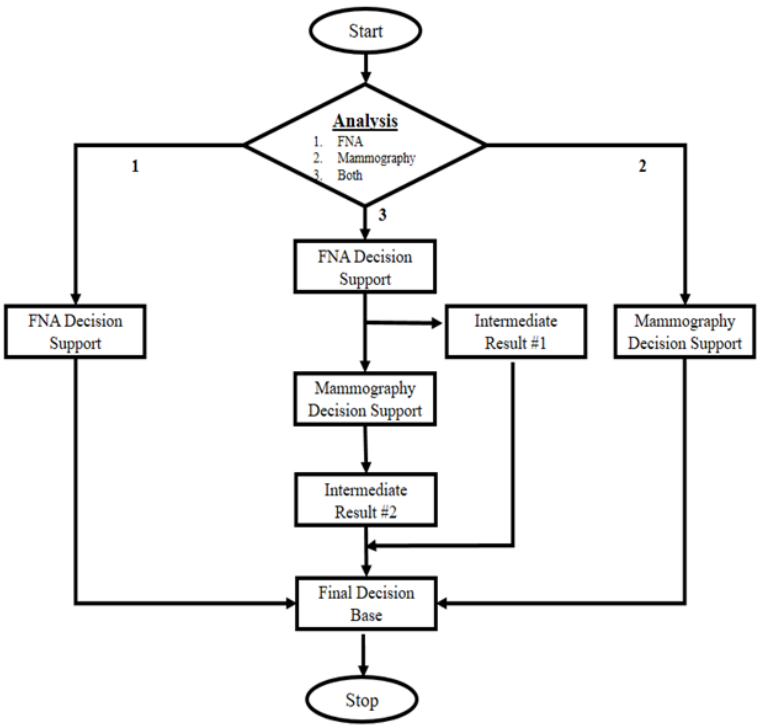
```
Pseudo-code 5: Stochastic Gradient descent
 Randomly initialize parameter weight and learning rate
      while Not Converged do
      Shuffle examples in training set
            for i = 1, · · · , n do
                  weight + = weight -   η Δweight L(f weight (x i , y i ))
            end
   end
```

high chance of converging into minima, but the performance was slow. Therefore, to select the best set of hyper-parameters for the algorithm grid-search technique was applied. Grid-search method search and return the specific subset of the hyper-parameter where the algorithm performs best. The final algorithm (optimized SGD) was then implemented on the same set of hyper-parameters.

Integrating FNA and Mammography Models

The amalgamation of two proposed methodology presents an intelligent information extraction and decision support system that can provide decision support for both two most widely used breast cancer screening practices, i.e., FNA and mammography. Figure 9 shows the organizational workflow design of the proposed model. To integrate both the proposed model, an interface was developed, allowing the radiologist to choose the screening practices for which the decision support is required.

Figure 9. The proposed architecture of the complete Breast Cancer Classification Model



The radiologist can choose any of the two, FNA or mammography or can choose both simultaneously. If the radiologist wants the decision support for only one screening practice at a time, by selecting that screening practice when prompted. The corresponding module will be called and executed, which provides the results in the form of benign or malignant or its associated BI-RADS category. In some cases, in order to ensure the diagnosis, both screening tests are recommended. The proposed system can handle such events also. To address the above-mentioned event, the first module for FNA screening will be called and executed. The result of the first FNA module is stored in a separate file (Intermediate Result#1) to contribute to the final decision in the future. Afterward, the second Mammography module will be called and executed. The result of the second Mammography module is again stored in a separate file (Intermediate Result#2) to participate in the final result. The intermediate file, Intermediate Result#1 and Intermediate result#2 are then called and processed to give the final result. The result provides quick recommendations and decision support to the radiologist during decision-making about the diagnosis.

Hence, in the presented study, an intelligent information retrieval system was developed that automatically extracts the most useful and relevant information from the two different types of datasets and further utilizes it to build the decision support system based on the machine learning approach. As the system was completely automated, diagnostic errors during the screening process due to human error can be avoided. The proposed model could be used to assist during the decision-making process for FNA and mammography screening, individually or in parallel.

EXPERIMENTAL ANALYSIS

Experiments are executed on a PC with Intel(R) Core (TM) i7-8565U at 1.80 GHz CPU with 8 GB RAM. Scientific Python Development Environment (SPYDER) is used with Python 3.7.3 for implementing the machine learning algorithm. It is an open-source Integrated Development Environment for writing and executing python codes.

Evaluation Metrics

The performance of the proposed model is assessed on the basis of the following assessment metrics: classification accuracy, recall, precision and F1-score. Details of the performance assessment measures are given in Table 2.

Table 2. Evaluation metrics

Measures	Explanation	Mathematical Formulation
Accuracy	The percentage of correctly categorized data samples from the total number of data samples.	$Accuracy = \frac{TP + TN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100$
Recall	The percentage of correctly classified positive samples from the total positive samples	$Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \times 100$
Precision	The percentage of correctly labelled positive samples that the classifier has found from the total number of positive samples that have been identified.	$Precision = \frac{TP}{\mathrm{TP} + \mathrm{FP}} \times 100$
F1-score	The subcontrary mean of accuracy and precision	$F1\ score = \frac{2TP}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \times 100$

Note: TP, TN, FP, FN are True Positives, True Negatives, False Positives and False Negatives, respectively.

Experimental Results and Discussion

The proposed experiments were coded in Python 3.7.3 and implemented on the SPYDER environment. In order to evaluate the efficiency of the proposed model, a range of comparative tests were carried out. The experimental results are further presented and analyzed in the section below.

Corpus-I Analysis: WBCD Structured Dataset From FNA Screening

The contribution of the proposed pipelined pre-processing module is evaluated by analyzing the performance of the classification model. Classification models have been trained and tested to understand the impact of pre-processing techniques on the fundamental (with random K value) and optimized K-NN classifier (with tuned K value). The details of the performance of each model variant are given in table 3. It is evident from the results that effective pre-processing techniques and optimized K-NN classifier improve classification accuracy and other performance metrics to a great extent.

The classification accuracy achieved by the fundamental K-NN classifier on the raw dataset is 87.27%. Optimizing the K-NN classifier for the best K value shows an improvement of 4.15% in the classification accuracy. The raw dataset is not suitable for learning as it suffers from missing values, unstandardized data and irrelevant features. Thus, the classifier was not able to attain good performance results with the raw dataset. To address this concern, different pre-processing practices were applied to refine the raw dataset. It is clear from table 3 that the optimized K-NN classifier's performance improved significantly after applying pre-processing techniques, i.e., missing value imputation, data standardization, and feature selection. Accuracy, precision, recall and F1-score of optimized K-NN improved by 9.46%, 7.57%, 5.53% and 3.86% with fundamental K-NN classifier. Further, accuracy, precision, recall and F1-score improved by 7.86%, 4.4%, 4.38% and 5.01%, respectively, with the tuned K-NN classifier for the pre-processed dataset in comparison to the raw dataset. In addition to the accuracy precision, recall and F1-score are also the necessary measures to calculate the classifier's performance on an imbalanced medical dataset. The proposed model under this study can attain significant improvement in recall and precision value. The high precision value of 99.38% indicates a low false-positive rate. The false-positive rate is the number of samples/patients that are falsely identified as malignant by the classifier. This situation could lead to unnecessary procedures or treatment. Furthermore, the classifier's missed positive cancer samples/patients is a critical and severe issue measured by the recall value. It induces patients to forego required care, resulting in severe effects that may be fatal. The recall value for the proposed model is 99.36% which is evidently good to reduce the possibility of missed positive samples. 99.36% is the F1-score, the harmonic mean of recall and precision. Therefore, it is clearly apparent that the proposed classification and prediction model performed really well. Table 4 present the performance of the proposed approach for WBCD classification in comparison to other state-of-art methods. Compared with the present state-of-art methods, the proposed model shows a significant improvement in the classification accuracy ranging from 10.28% to 0.02%. The proposed work is able to achieve these results due to efficient pre-processing techniques and effective classifier learning. The different classification techniques

Table 3. Evaluating the effect of classifier performance under different experiment setup

Input Dataset	K-NN classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Raw Dataset	Random K value	87.27	86.59	83.27	84.21
	Tuned k to 15	91.42	94.98	94.98	94.35
Dataset + missing value imputation + feature selection	Random K value	89.82	91.81	93.83	95.50
	Tuned k to 15	99.28	99.38	99.36	99.36

Table 4. Performance evaluation of this study with present state-of-art methods

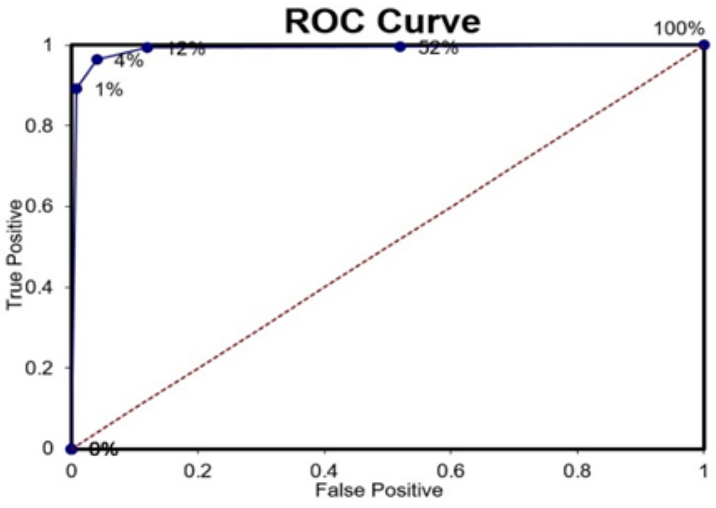
Dataset	Authored By	Approach	Accuracy achieved	This study Accuracy
WBCD (Original)	(Chen, 2011)	Rough set theory and SVM	89.20%	99.28%
	(Karabatak, 2009)	Association rule and Neural Network	95.6%	
	(Setiono, 2000)	Feature selection and neural network	98.10%	
	(Şahan, 2007)	Fuzzy, AIS and KNN	99.14%	
	(Marcano-Cedeño, 2011)	Metaplasticity Neural Network	99.26%	

and validation approaches adopted in the compared studies may affect the classifier performance, but the dataset used for the experiment and the objective of improving the classification accuracy is the same. Therefore, it is apparent that the proposed model under this study achieves the best classification accuracy in comparison to the other work reported in the present literature. The receiver operating characteristic (ROC) curve illustrates the diagnostic ability of a binary classifier system of the proposed model given in figure 10.

Corpus-II Analysis: Semi-Structured Dataset From Mammography Report

The number of features extracted from text datasets was highly uncertain. Thus, the proposed model has been evaluated to study the effect of unigrams and bigrams on accuracy. Bigram-based models performed better than unigram-based models, with the average accuracy improved from 86.47% to 97.40%, a statistically significant accuracy increase of 9.15%. Unigrams fail to capture the structure of a particular language and do not contain much information about a term's context. However, when the same term is combined with the next successive term, it becomes more powerful for learning as it contains defined, concise information about the context and structure of the term. For example, "fatty" was the highest TF-IDF scoring term in the dictionary, but it did not provide much

Figure 10. ROC graph for FNA model using K-NN



information about its context. However, "fatty hilum" is bigram, having the same TF-IDF score but is more informative than the unigram "fatty". Bigrams provide more concise information about the importance of consecutive terms, giving the frequency of occurrence of a particular term X followed by term Y. The terms were combined according to their proximity in the document, therefore preserving the semantic relationship of the terms and decreasing the bias. They incorporate knowledge into the purely statistical task of text categorization. Tables 5 and 6 contain some of the highest generated unigram scores and bigrams. Accuracy fluctuations when using unigrams or bigrams for training the model are shown in figure 11.

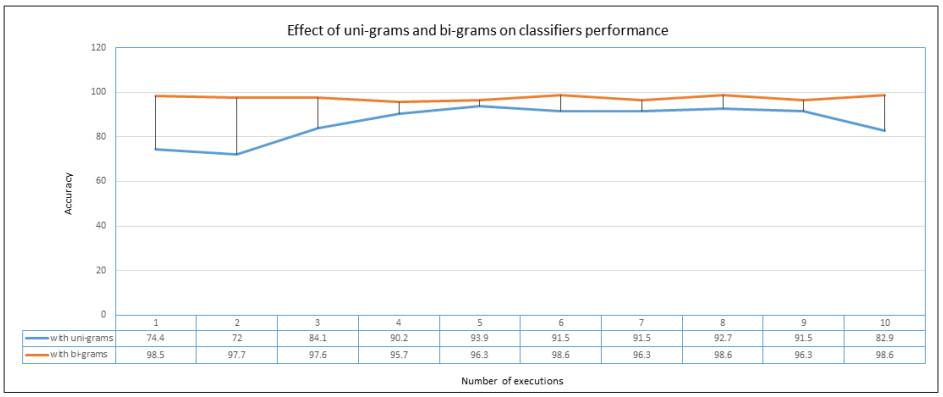
Table 5. Some of the unigrams with the highest TF-IDF score

axillary	heterogeneity	macrocalcific	significant
circumscribed	hilum	lactating	microlobulated
fatty	lymph	inflammatory	distal
evidence	large	insinuating	lesions
extending	prominent	periareolar	hyperechogenecity

Table 6. Some of the Bigrams with the highest TF-IDF score

ovoid lobulated	thickening measuring	area insinuating	contour lesion
anechoic lesion	surrounding perilesional	axilla heterogenously	fatty hilum
breast anechoic	tissue hyperechogenecity	echoes seen	fibroglandular tissues
surrounding fibroglandular	suggestive mastitis	dilated mammary	perilesional hyperechogenecity
tall masses	seen subareolar	calcific foci	fibroadenoma axillae
transducer heterogenous	reveals hypoechoic	axillary lymph	cyst noted
vascularity noted	abscess subareolar	breast macrocalcific	cystic lesions

Figure 11. Classifier performance w.r.t uni-grams and bi-grams



The proposed method produced an F1 score of 0.981 to classify the mammography reports into benign and malignant classes. The precision and recall were both higher than those from the existing rule-based system. The proposed model's performance was compared with those from other research groups and summarized in table 8. Others have previously reported on the classification of mammography reports and various statistical techniques had been used to attain high classification accuracies. The proposed model outperformed state-of-the-art rule-based systems. The major limitation of a rule-based system is that the model is domain-specific and lacks generalizability with the problem of reporting local variation. Moreover, in these models, rules need to be defined explicitly by the domain expert for the detection of cancer. Therefore, the rule-based approach is very time-consuming and needs constant revisions to meet the changing symptoms and detection rules. Thus, in the proposed study, an intelligent information system was proposed to develop an automated complete cancer classification model that does not require expensive text annotation and learns readily from the large pool of unannotated corpora. To reduce the rate of false negatives, the recall value must be high. The proposed system successfully extracted more comprehensive information from the reports than the rule-based system described previously by Bozkurt et al. (Bozkurt, 2016). We also noted that the generic state-of-the-art regular expression-based NLP tools failed to detect negation of findings from report sample sentences. Furthermore, our model reduces the rate of false positives and increases the precision value.

It is clearly evident that the proposed system outperforms existing state-of-art methods for the classification of breast cancer samples for the structured data from FNA(WBCD) and free text mammography reports. Therefore, integrating both the proposed system presents a complete breast cancer decision-support for the classification and prediction.

Table 7. Performance evaluation of SGD classifier for semi-structured mammography reports

	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
Proposed study	97.53	98.82	98.88	98.82

Table 8. Evaluation of proposed system architecture with the present state-of-art methods

Year of study	Objective	Approach	Performance	Limitation
Gupta, 2018	To derive relationships from radiology records in an unsupervised manner	Rule-based + unsupervised clustering	Precision = 95% Recall = 94%	Not able to classify in BI-RADS categories.
Hussam, 2013	Developed a features extraction algorithm BI-RADS categorization	Rule-based	Precision =97.7%, Recall =95.5%	Not able to classify in BI-RADS categories.
Bozkurt, 2016	Providing decision-support by automatically extracting information from mammography reports	Rule-based + supervised ML	Accuracy = 97.58%	Lack generalizability.
Castro, 2017	Extracting BI-RADS categories from the reports	Rule-based + supervised ML	Precision = 98% Recall = 93%	The category BI-RADS can only be extracted if the radiologist records the category BI-RADS
Sippo, 2013	Extracting final evaluation categories for BI-RADS from the reports	Rule-based	Precision = 99.6% Recall = 100%	The model could classify the BI-RADS category only if it is reported in the report.

LIMITATIONS

The proposed system provides complete breast cancer decision support to the radiologist during crucial decision-making. The model performed statistically well compared with the present state-of-art methods by showing good improvement in accuracy and recall value, but the proposed approach has certain limitations. Firstly, both the structured and semi-structured datasets utilized for the study are from different but single institutional datasets, which might bias the classification performance of the proposed model. Secondly, the mammography report dataset's size is limited to only 356 reports; thus, the model tends to overfit and can give misleading results.

CONCLUSION

An automated complete decision support system for clinical practice is a useful artificial intelligence tool to assist medical experts in making informed diagnoses in the case of ambiguity or inadequate information. It may reduce the overall cost of treatment. FNA and mammography screening are two of the most widely used techniques for the detection and diagnosis of breast cancer. The proposed support system is a hybrid model capable of analyzing structured FNA observations and unstructured mammography observations within one system. An effective information extraction system was presented to extract valuable information from the dataset. Information extraction systems use efficient pre-processing techniques that increase data quality. It involves handling missing values, standardizing data, balancing unbalanced datasets, reducing dimensionality by selecting the most appropriate and best features for model learning in the feature selection process. The machine learning model is then built on the extracted information. The FNA structured model is trained and tested with a K-NN classifier with an optimum value of K. The semi-structured mammography model is trained and tested with an SGD classifier by hyper-tuning the parameter through the grid-search technique. It was evident that a KNN classifier for FNA analysis and the SGD classifier for mammography analysis performed best when used with the most predictive features. Our model is intended to assist medical experts by providing quick, precise and reliable recommendations that can be applied during crucial decision-making stages and has the potential to improve patient survival rates and overall quality of life.

REFERENCES

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *36*(2), 3240–3247. doi:10.1016/j.eswa.2008.01.009

Alzu'bi, A. N.-S. (2021). Predicting the recurrence of breast cancer using machine learning algorithms. *Multimedia Tools and Applications*, 1–14.

Anitha, J. (2016). A Multiresolution Ripplet Transform for Breast Cancer Diagnosis in Digital Mammograms. *Recent Patents on Computer Science*, *9*(3), 195–202. doi:10.2174/2213275908666150324223944

Bilal, E. D., Dutkowski, J., Guinney, J., Jang, I. S., Logsdon, B. A., Pandey, G., Sauerwine, B. A., Shimoni, Y., Moen Vollan, H. K., Mecham, B. H., Rueda, O. M., Tost, J., Curtis, C., Alvarez, M. J., Kristensen, V. N., Aparicio, S., Børresen-Dale, A.-L., Caldas, C., Califano, A., & Margolin, A. A. et al. (2013). Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS Computational Biology*, *9*(5), e1003047. doi:10.1371/journal.pcbi.1003047 PMID:23671412

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, (pp. 177-186). Physica-Verlag HD. doi:10.1007/978-3-7908-2604-3_16

Bouyer, A. (2017). Breast Cancer Diagnosis Using Data Mining Methods, Cumulative Histogram Features, and Gary Level Co-occurrence Matrix. *Current Medical Imaging Reviews*, *13*(4), 460–470. doi:10.2174/1573405613666161227162918

Bozkurt, S. G., Gimenez, F., Burnside, E. S., Gulkesen, K. H., & Rubin, D. L. (2016). Using automatically extracted information from mammography reports for decision-support. *Journal of Biomedical Informatics*, *62*, 224–231. doi:10.1016/j.jbi.2016.07.001 PMID:27388877

Brown, J. B., Bryan, S., & Warren, R. (1996). Mammography screening: An incremental cost effectiveness analysis of double versus single reading of mammograms. *BMJ (Clinical Research Ed.)*, *312*(7034), 809–812. doi:10.1136/bmj.312.7034.809 PMID:8608287

Burnside, E. S., Davis, J., Chhatwal, J., Alagoz, O., Lindstrom, M. J., Geller, B. M., Littenberg, B., Shaffer, K. A., Kahn, C. E. Jr, & Page, C. D. (2009). Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology*, *251*(3), 663–672. doi:10.1148/radiol.2513081346 PMID:19366902

Castro, E. T. (2017). Automated annotation and classification of BI-RADS assessment from radiology reports. *J. Biomed. Inform*177-187.

Chen, H. L., Yang, B., Liu, J., & Liu, D.-Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *38*(7), 9014–9022. doi:10.1016/j.eswa.2011.01.120

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, *2*. doi:10.1177/117693510600200030 PMID:19458758

Dutra, I. N. (2011). Integrating machine learning and physician knowledge to improve the accuracy of breast biopsy. *AMIA ... Annual Symposium Proceedings - AMIA Symposium. AMIA Symposium*, *2011*, 349. PMID:22195087

Elmore, J. G. (2005). Screening for breast cancer. *Journal of the American Medical Association*, *293*(10), 1245–1256. doi:10.1001/jama.293.10.1245 PMID:15755947

Elmore, J. G., Nakano, C. Y., Koepsell, T. D., Desnick, L. M., D'Orsi, C. J., & Ransohoff, D. F. (2003). International variation in screening mammography interpretations in community-based programs. *Journal of the National Cancer Institute*, *95*(18), 13. doi:10.1093/jnci/djg048 PMID:13130114

Esmaeili, M. A., Ayyoubzadeh, S. M., Ahmadinejad, N., Ghazisaeedi, M., Nahvijou, A., & Maghooli, K. (2020). A decision support system for mammography reports interpretation. *Health Information Science and Systems*, *8*(1), 1–8. doi:10.1007/s13755-020-00109-5 PMID:32257128

Gao, H. B., Aiello Bowles, E. J., Carrell, D., & Buist, D. S. M. (2015). Using natural language processing to extract mammographic findings. *Journal of Biomedical Informatics*, *54*, 77–84. doi:10.1016/j.jbi.2015.01.010 PMID:25661260

Giard, R. W., & Hermans, J. (1992). The value of aspiration cytologic examination of the breast a statistical review of the medical literatur. *Cancer*, *69*(8), 2104–2110. doi:10.1002/1097-0142(19920415)69:8<2104::AID-CNCR2820690816>3.0.CO;2-O PMID:1544116

Gupta, A. B., Banerjee, I., & Rubin, D. L. (2018). Automatic information extraction from unstructured mammography reports using distributed semantics. *Journal of Biomedical Informatics*, *78*, 78–86. doi:10.1016/j.jbi.2017.12.016 PMID:29329701

Jemal, A. M., Murray, T., Ward, E., Samuels, A., Tiwari, R. C., Ghafoor, A., Feuer, E. J., & Thun, M. J. (2005). Cancer statistics. *CA: a Cancer Journal for Clinicians*, *55*(1), 10–30. doi:10.3322/canjclin.55.1.10 PMID:15661684

Jerez-Aragonés, J. M.-R.-J.-P.-C., Gómez-Ruiz, J. A., Ramos-Jiménez, G., Muñoz-Pérez, J., & Alba-Conejo, E. (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial Intelligence in Medicine*, *27*(1), 45–63. doi:10.1016/S0933-3657(02)00086-6 PMID:12473391

Kabir, F. S. (2015). Bangla text document categorization using stochastic gradient descent (sgd) classifier. *2015 International Conference on Cognitive Computing and Information Processing (CCIP)*, 1-4. doi:10.1109/CCIP.2015.7100687

Kate, R. J., & Nadig, R. (2017). Stage-specific predictive models for breast cancer survivability. *International Journal of Medical Informatics*, *97*, 304–311. doi:10.1016/j.ijmedinf.2016.11.001 PMID:27919388

Kim, W. K., Kim, K. S., Lee, J. E., Noh, D.-Y., Kim, S.-W., Jung, Y. S., Park, M. Y., & Park, R. W. (2012). Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of Breast Cancer*, *15*(2), 230–238. doi:10.4048/jbc.2012.15.2.230 PMID:22807942

Kumar, S. (2015). Enhancing text classification by stochastic optimization method and support vector machine. *International Journal of Computer Science and Information Technologies*, *6*(4), 3742–3745.

Kumari, M., & Singh, V. (2018). Breast Cancer Prediction system. *Procedia Computer Science*, *132*, 371–376. doi:10.1016/j.procs.2018.05.197

Kumari, M., & Vijendra, S. (2017). Big data analytics in healthcare: Opportunities, challenges and techniques. *International Journal of Social Computing and Cyber-Physical Systems*, *2*(1), 35–58. doi:10.1504/IJSCCPS.2017.088748

Lee, C. S.-T., Nagy, P. G., Weaver, S. J., & Newman-Toker, D. E. (2013). Cognitive and system factors contributing to diagnostic errors in radiology. *AJR. American Journal of Roentgenology*, *201*(3), 611–617. doi:10.2214/AJR.12.10375 PMID:23971454

Liberman, L., & Menell, J. H. (2002). Breast imaging reporting and data system (BI-RADS). *Radiologia Clinica*, *40*(3), 409–430. doi:10.1016/S0033-8389(01)00017-3 PMID:12117184

Liu, N. Q., Qi, E.-S., Xu, M., Gao, B., & Liu, G.-Q. (2019). A novel intelligent classification model for breast cancer diagnosis. *Information Processing & Management*, *56*(3), 609–623. doi:10.1016/j.ipm.2018.10.014

Macías-García, L. M.-B.-R.-H.-G.-S., Martínez-Ballesteros, M., Luna-Romera, J. M., García-Heredia, J. M., García-Gutiérrez, J., & Riquelme-Santos, J. C. (2020). Autoencoded DNA methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance. *Artificial Intelligence in Medicine*, *110*, 101976. doi:10.1016/j.artmed.2020.101976 PMID:33250148

Marcano-Cedeño, A. Q.-D., Quintanilla-Domínguez, J., & Andina, D. (2011). WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Systems with Applications*, *38*(8), 9573–9579. doi:10.1016/j.eswa.2011.01.167

Maskell, G. (2019). Error in radiology—Where are we now? *The British Journal of Radiology*, *92*(1096), 20180845. doi:10.1259/bjr.20180845 PMID:30457880

Montelongo González, E. E., Reyes Ortiz, J. A., & González Beltrán, B. A. (2020). Machine Learning Models for Cancer Type Classification with Unstructured Data. *Computación y Sistemas*, *24*(2). Advance online publication. doi:10.13053/cys-24-2-3367

Nassif, H., & F., C.-C. (2012). Extracting BI-RADS features from portuguese clinical texts Proceedings. *IEEE International Conference on Bioinformatics and Biomedicine, NIH Public Access*, 1. doi:10.1109/BIBM.2012.6392613

Nguyen, T. D. (2016). Fuzzy Kernel Stochastic Gradient Descent machines. *International Joint Conference on Neural Networks (IJCNN)*, 3226-3232. doi:10.1109/IJCNN.2016.7727611

Osareh, A. &. (2010). Machine learning techniques to diagnose breast cancer. *5th International Symposium on Health Informatics and Bioinformatics*, 114-120. doi:10.1109/HIBIT.2010.5478895

Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing*, *17*(4), 694–701. doi:10.1016/j.dsp.2006.10.008

Rani, G. G. (2015). Tumour Classification and Analysis from Breast Cancer Pathology Reports using Natural Language Processing. *Indian Journal of Science and Technology*, *8*(29).

Şahan, S. P., Polat, K., Kodaz, H., & Güneş, S. (2007). A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. *Computers in Biology and Medicine*, *37*(3), 415–423. doi:10.1016/j.compbiomed.2006.05.003 PMID:16904096

Sasikala, S. E., Ezhilarasi, M., & Senthil, S. (2018). Breast Cancer Diagnosis System Based on the Fusion of Local Binary and Ternary Patterns from Ultrasound B Mode and Elastography Images. *Current Medical Imaging Reviews*, *14*(6), 947–956. doi:10.2174/1573405613666170511125859

Schaffter, T. B., Buist, D. S. M., Lee, C. I., Nikulin, Y., Ribli, D., Guan, Y., Lotter, W., Jie, Z., Du, H., Wang, S., Feng, J., Feng, M., Kim, H.-E., Albiol, F., Albiol, A., Morrell, S., Wojna, Z., Ahsen, M. E., Asif, U., & Jung, H. et al. (2020). Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Network Open*, *3*(3), e200265–e200265. doi:10.1001/jamanetworkopen.2020.0265 PMID:32119094

Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in Medicine*, *18*(3), 205–219. doi:10.1016/S0933-3657(99)00041-X PMID:10675715

Singh, H. S., Sethi, S., Raber, M., & Petersen, L. A. (2007). Errors in cancer diagnosis: Current understanding and future directions. *Journal of Clinical Oncology*, *25*(31), 5009–5018. doi:10.1200/JCO.2007.13.2142 PMID:17971601

Sippo, D. A., Warden, G. I., Andriole, K. P., Lacson, R., Ikuta, I., Birdwell, R. L., & Khorasani, R. (2013). Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing. *Journal of Digital Imaging*, *26*(5), 989–994. doi:10.1007/s10278-013-9616-5 PMID:23868515

Society, A. C. (2011). *Global Cancer Facts and Figures*. American Cancer Society.

Tang, J. R. (2009). Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances. *IEEE Transactions on Information Technology in Biomedicine*, *13*(2), 236–251. doi:10.1109/TITB.2008.2009441 PMID:19171527

Thiebaut, N. S. (2017). *An innovative solution for breast cancer textual big data analysis.* arXiv preprint arXiv, 1712.02259.

Tseng, Y. H., Huang, C.-E., Wen, C.-N., Lai, P.-Y., Wu, M.-H., Sun, Y.-C., Wang, H.-Y., & Lu, J.-J. (2019). Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *International Journal of Medical Informatics*, *128*, 79–86. doi:10.1016/j.ijmedinf.2019.05.003 PMID:31103449

Veronesi, U. B. P. (2005). Breast cancer. *Lancet*. PMID:15894099

Wieneke, A. E. (2015). Validation of natural language processing to extract breast cancer pathology procedures and results. *Journal of Pathology Informatics*, 6. PMID:26167382

Yeh, W. C., Chang, W.-W., & Chung, Y. Y. (2009). A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Systems: International Journal of Knowledge Engineering and Neural Networks*, *36*(4), 8204–8211. doi:10.1016/j.eswa.2008.10.004