BMC Medical Informatics and Decision Making

DATABASE

Open Access

TECRR: a benchmark dataset of radiological reports for BI-RADS classification with machine learning, deep learning, and large language model baselines

Sadam Hussain[1*], Usman Naseem[2], Mansoor Ali[1], Daly Betzabeth Avendaño Avalos[3], Servando Cardona-Huerta[3], Beatriz Alejandra Bosques Palomo[1] and Jose Gerardo Tamez-Peña[3]

Abstract

Background Recently, machine learning (ML), deep learning (DL), and natural language processing (NLP) have provided promising results in the free-form radiological reports' classification in the respective medical domain. In order to classify radiological reports properly, a high-quality annotated and curated dataset is required. Currently, no publicly available breast imaging-based radiological dataset exists for the classification of Breast Imaging Reporting and Data System (BI-RADS) categories and breast density scores, as characterized by the American College of Radiology (ACR). To tackle this problem, we construct and annotate a breast imaging-based radiological reports dataset and its benchmark results.

The dataset was originally in Spanish. Board-certified radiologists collected and annotated it according to the BI-RADS lexicon and categories at the Breast Radiology department, TecSalud Hospitals Monterrey, Mexico. Initially, it was translated into English language using Google Translate. Afterwards, it was preprocessed by removing duplicates and missing values. After preprocessing, the final dataset consists of 5046 unique reports from 5046 patients with an average age of 53 years and 100% women. Furthermore, we used word-level NLP-based embedding techniques, term frequency-inverse document frequency (TF-IDF) and word2vec to extract semantic and syntactic information. We also compared the performance of ML, DL and large language models (LLMs) classifiers for BI-RADS category classification.

Results The final breast imaging-based radiological reports dataset contains 5046 unique reports. We compared K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Adaptive Boosting (AdaBoost), Gradient-Boosting (GB), Extreme Gradient Boosting (XGB), Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT) and Biomedical Generative Pre-trained Transformer (BioGPT) classifiers. It is observed that the BioGPT classifier with preprocessed data performed 6% better with a mean sensitivity of 0.60 (95% confidence interval (CI), 0.391-0.812) compared to the second best performing classifier BERT, which achieved mean sensitivity of 0.54 (95% CI, 0.477-0.607).

*Correspondence:
Sadam Hussain
a01753094@tec.mx
Full list of author information is available at the end of the article

Conclusion In this work, we propose a curated and annotated benchmark dataset that can be used for BI-RADS and breast density category classification. We also provide baseline results of most ML, DL and LLMs models for BI-RADS classification that can be used as a starting point for future investigation. The main objective of this investigation is to provide a repository for the investigators who wish to enter the field to push the boundaries further.

Keywords BI-RADS classification, Breast radiological reports, TF-IDF, Word2vec, NLP, ML

Introduction

Breast cancer is the leading type of cancer diagnosed in women worldwide [1]. It poses a significant public health concern and carries a substantial economic burden. Early breast cancer diagnosis is widely recognized as a critical factor in reducing mortality rates [2]. Mammography screening is a recommended method for the early detection of breast cancer in average-risk women [3, 4]. To standardize terminology and categorize results for each breast imaging modality (i.e., mammography, ultrasound, MRI, and DBT), the ACR developed the BI-RADS. This system consists of seven categories, ranging from (0 through 6), where 0 is inconclusive, 1 is negative, 2 is benign, 3 is probably benign, 4 is suspicious, 5 is highly suggestive of malignancy, and 6 is known biopsy-proven malignancy. In addition to clinical settings, the BI-RADS system is a quality assurance tool in research [5].

The manual method is the state-of-the-art (SOTA) of extracting information from free text, which is costly, error-prone, and time-consuming, especially when dealing with large datasets. Free-form text is a way of writing radiology reports without a template, which can be more expressive and efficient but also more challenging to standardize and interpret [6–8]. To overcome these challenges, different NLP methods have been developed. These methods have revealed promising results in extracting crucial information from radiological reports, enabling easy access to appropriate information to be analyzed for various clinical applications [9–11].

Recently, there has been a great interest in using AI algorithms to increase the accuracy of BI-RADS prediction from breast imaging-based radiological reports [12–17]. Nonetheless, the development and evaluation of such AI algorithms require quality and large-scale datasets, including radiological reports, that are very rare. Therefore, we are annotating and releasing a novel dataset. This work presents a new dataset on breast imaging-based radiological reports and its baselines. The dataset contains over 5,000 radiological reports from 2D mammography, 3D mammography, and breast ultrasound (US). The data was collected from TecSalud Hospitals (Monterrey, Mexico). The dataset was translated into English from Spanish and preprocessing techniques were applied to remove the duplicates and missing values with the consultation of radiologists. We tried different word embedding techniques to vectorize the radiological reports in order to extract the syntactic and semantic meaning that is subsequently helpful for BIRADS classification. We also used different ML, DL and LLMs architectures to provide baselines for classifying BI-RADS. This dataset and baselines can serve as a valuable resource for researchers working on AI algorithms for breast imaging-based radiological reports and can contribute to enhancing the accuracy and efficacy of breast cancer diagnosis.

Methods

Breast imaging is the angular stone for early detection and diagnosis of breast cancer and other breast-related conditions [4, 18]. Currently, the reporting system for breast imaging studies differs significantly among radiologists and institutions and is predominantly based on traditional free-text reporting [19]. However, structured reporting is being promoted to improve reporting in radiology, which would benefit radiological, clinical practice, and data mining in an ongoing project. Meanwhile, we have substantial pre-existing data that requires retrospective analysis. To address this issue, the ACR developed the BI-RADS, a standardized system for describing and communicating breast imaging results [20]. The BI-RADS provides a common language and a classification scheme for mammography, ultrasound, MRI, and DBT of the breast, as well as guidance for follow-up and outcome monitoring. The BI-RADS also enables radiologists to perform quality assurance and improvement through medical audits and data analysis.

This paper presents a new breast imaging-based radiological reports dataset that follows the BI-RADS framework and covers multiple imaging modalities. Our dataset contains over 5,000 radiology reports from different patients, annotated by board-certified radiologists according to the BI-RADS lexicon and categories [21, 22]. The BI-RADS category 0 was assigned when additional evaluation was required. The BI-RADS category 1, commonly known as negative, was assigned when there was no evidence of malignancy in either breast, and they are symmetrical with no masses, architectural distortion, or suspicious calcifications. The BI-RADS category 2, i,e benign, was assigned when a benign finding did not

require further evaluation or follow-up, such as a simple cyst, fibroadenoma, intramammary lymph node, or benign calcifications. The BI-RADS category 3 (probably benign) was assigned when there was less than a 2% chance of being malignant and can be safely monitored with short-interval follow-up imaging, such as a probably benign mass, focal asymmetry, or clustered microcysts. The BI-RADS category 4 (suspicious) was assigned when there was a 2-94% chance of being malignant, and biopsy should be considered, such as a spiculated mass, architectural distortion, or suspicious calcifications. The BI-RADS category 5 (highly suggestive of malignancy) was assigned when there was a chance of greater than 95% of being malignant and appropriate action should be taken. The BI-RADS category 6 (known biopsy-proven malignancy) was assigned when there was histological confirmation of malignancy in the breast before any treatment had been initiated [23, 24]. We use our dataset to benchmark several ML, DL and LLMs methods on the BI-RADS correct category classification. We also provide clinical applications and insights based on our dataset and results. This dataset can be requested to use.

We introduce a novel and comprehensive breast imaging-based radiological reports dataset that adheres to the BI-RADS standard and covers multiple modalities reports. Using our dataset, we propose SOTA NLP pipeline-based ML, DL and LLMs benchmarks on the BI-RADS correct category classification, demonstrating its potential for advancing research in both domains.

Data

The dataset was collected from the Breast Radiology department, TecSalud hospitals (Monterrey, Mexico). It was approved by the ethical committee to give access to the data upon reasonable request. This dataset comprises digital mammography (DM), digital breast tomosynthesis (DBT), and breast ultrasound (US) based radiological reports. All the reports were anonymized. These reports were originally in Spanish and collected from January to December 2014. The reports were then translated online using Google Translate and were verified by the radiologist, who is better aware of English and Spanish. The reports were created and evaluated by (n=2) trained radiologists. There were (n=7904) actual entries in our dataset. However, after preprocessing, i.e., removing reports with missing values, duplicate records, and BI-RADS categories (0-inconclusive) and (6-biopsy proven), we were able to utilize 5,046 unique reports with an average age of 53 years and 5046 (100.0%) women from 7,904 reports for the final model building and evaluation. The overview of the dataset is given in Table 1.

Preprocessing

Our original dataset was in the Spanish language. The dataset was translated online using Google Translate. The single entry of the patient consists of a brief clinical indication of the study, and personal risk history, both imaging description, findings, and diagnostic impression. As shown in Fig. 1, each original DM, DBT, and breast US report consists of multiple paragraphs. We developed an NLP-based pipeline to extract associated imaging features for the final BI-RADS category classification. The pipeline consists of various analysis filters for different clinical and linguistics tasks, like sentence segmentation, sentence detection, tokenization, detection of concepts, and data normalization. The pipeline is described in Fig. 2.

Data curation

Data curation was conducted on the TECRR dataset to ensure its quality and suitability for analysis. The process involved several key steps: first, all references to doctors were removed to maintain privacy and confidentiality. Duplicated structured radiological reports were then identified and eliminated to prevent redundant information. Dates within the reports were removed to minimize

Digital mammography with bilateral volumes and ultrasound:\n Indication: Routine study in 58-year -old patient. It has an excisional biopsy history in the left breast due to calcifications with benign histopathological result.\n Technique: conventional projections are carried out, volumetric reconstruction (tomosynthesis) and bilateral ultrasound are included. The previous studies of this institution of #### are reviewed.\n Findings:\n The breast tissue is #### dense (pattern #### of the American College of Radiology, ACR ####).\n There are no relevant changes in the glandular pattern compared to the previous study.\n There is an asymmetry in the #### breast seen only in the CC projection in the internal quadrants, a medium third that changes in the different projections and that by volumes it is confirmed that it is #### by tissue summation.\n There are no masses, distortion areas, or suspicious calcifications of malignancy.\n by ultrasound there are some simple #### and cysts. There are no #### nodules. The axillary regions with lymph nodes of #### size and morphology.\n

Fig. 1 Breast imaging based radiological report (This is a fabricated report for demonstration purpose)

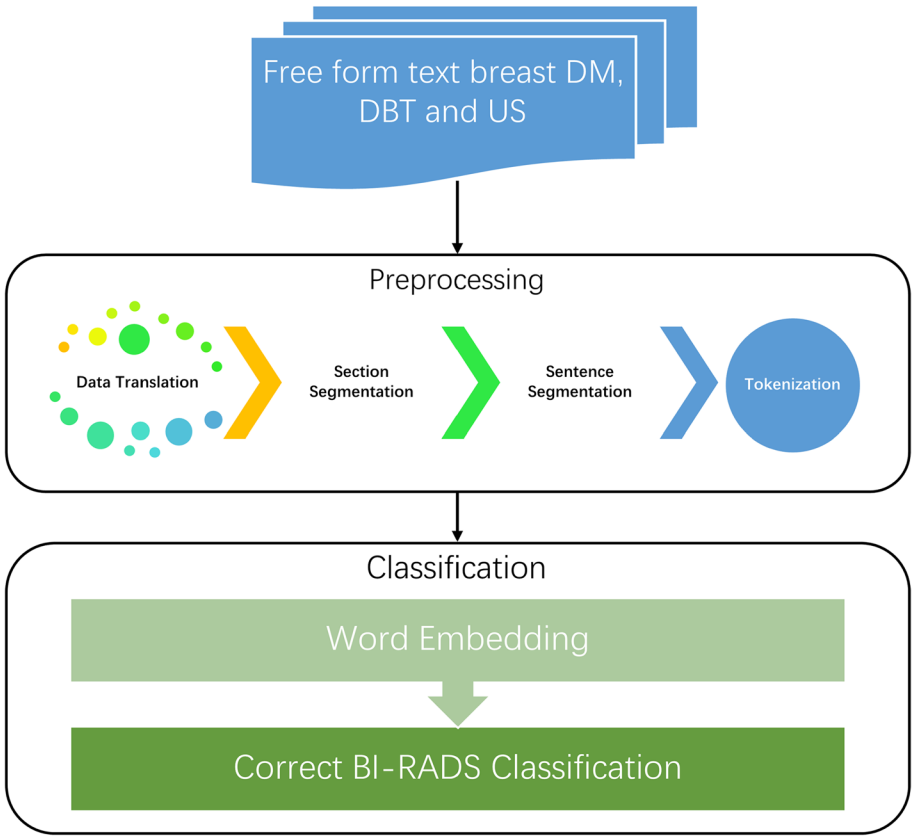

Fig. 2 NLP workflow for preprocessing reports and BI-RADS classification

potential bias related to temporal factors, and any non-radiological information was excluded to focus solely on the relevant radiological data. Additionally, density scores and BI-RADS scores were standardized across the dataset to enhance comparability and consistency. Finally, the structured reports were segmented into sections, including the main body, patient description, and report conclusion.

Word embedding techniques

We employed word-level NLP-based embedding techniques such as TF-IDF [25] and word2vec [26] to extract semantic and syntactic information from the free-form text of radiological reports. TF-IDF is a statistical measure that evaluates the importance of words based on their frequency in a document and their rarity across the entire corpus. It assigns higher weights to words that frequently occur in a document but are rare in the corpus. The formula for TF-IDF is:

$$\text{tf-idf}(w,d,D) = \text{tf}(w,d) \times \text{idf}(w,D) \tag{1}$$

where w is a word, d is a document, and D is the collection of documents. The term frequency, $tf\,(w,d)$, represents how often word w appears in document d, while the inverse document frequency, $idf\,(w,D)$, is calculated as:

$$\text{idf}(w,D) = \log\frac{|\text{D}|}{|\text{d} \in \text{D} : \text{w} \in \text{d}|} \tag{2}$$

Here, $|D|$ is the total number of documents, and $|d \in D : w \in d|$ represents the number of documents that contain the word$w$. TF-IDF increases for words that appear frequently in a specific document but are rare across the collection, thus emphasizing distinctive terms. Word2vec [26], on the other hand, is a neural network-based model that generates word embeddings-vector representations of words that capture both semantic and syntactic relationships. Word2vec uses either a skip-gram model, which predicts context words given a target word, or a continuous bag-of-words (CBOW) model, which predicts the target word from its surrounding context. The word2vec skip-gram objective is to maximize the likelihood of context words given a target word:

$$\text{word2vec}(w,c,D) = \log\frac{\exp\left(\text{v}_\text{w}^\text{T}\text{v}_\text{c}\right)}{\sum_{\text{w}' \in \text{D}} \exp\left(\text{v}_{\text{w}'}^\text{T}\text{v}_\text{c}\right)} \tag{3}$$

where w is the target word, c is a context word, D is the vocabulary, and $v_w^T v_c$ represents the dot product of the word embeddings of w and c. This formula can be

interpreted as the probability of observing a context word given a target word, normalized by a softmax function over the vocabulary. The word embeddings are learned by maximizing this probability over all word-context pairs in the corpus.

State-of-the-art models

We compared the performance of ML, DL and LLMs models. The ML models are KNN, SVM, NB, RF, AdaBoost, GB and XGBoost; the DL model is LSTM [27]; and the LLMs models are BERT [28] and BioGPT [29]. LSTM networks capture long-term dependencies in sequential data, making them suitable for modelling temporal patterns in radiological reports. BERT, particularly the "bert-base-uncased" model, leverages a bidirectional transformer architecture to capture context from both directions in a sentence. This version of BERT consists of 12 layers, 12 attention heads, and approximately 110 million parameters, focusing on uncased English text. Furthermore, BioGPT, a specialized transformer model for biomedical text generation, is employed for tasks such as medical report generation and knowledge extraction. The tokenization for BioGPT is handled using the BioGPT Tokenizer, which processes biomedical vocabulary and ensures precise tokenization of medical terminology. These advanced models capture more complex contextual and semantic relationships than traditional methods, enhancing tasks such as information retrieval, classification, and summarization of radiological reports. These methods enable the extraction of key information from the often unstructured text of radiological reports, which may contain specialized medical terminology [30]. The semantic relationships between words, captured by these embeddings, facilitate tasks such as information retrieval, classification, and summarization of radiological reports [31].

Evaluation metrics

We report sensitivity and accuracy with 95% CI. A 95% CI indicates that the model is confident that, in 95 out of 100 cases, the true prediction will fall within the upper and lower bounds of the interval. Sensitivity or recall is the proportion of true positives among all positive cases. Sensitivity is a measure of how well a test can correctly identify true positives, i.e., cases of breast cancer. Sensitivity is good for BI-RADS prediction because it can help reduce the number of false negatives, i.e., cases of breast cancer that are missed by the test. A high sensitivity means that the test can capture most of the breast cancer cases and avoid unnecessary delays in diagnosis and treatment [32]. The formula for the sensitivity is given below,

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

Accuracy measures the total number of correct classifications divided by the total number of cases. The formula is given as follows,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Training and test sets

Our dataset contained 5046 studies from 5046 patients. Data were randomly split into training and testing to avoid overlap between subjects. The 80% of the data was allocated to the training set, while 20% was allocated to the testing set. The training set includes 4036 subjects, while the remaining data of 1010 subjects was used for testing.

Exploratory data analysis

In this work, we applied Exploratory Data Analysis (EDA) on dataset of radiological reports from the Breast Radiology department, TecSalud hospitals (Monterrey, Mexico) database. We used different graphical and statistical tools, like bar graphs, pie charts and word clouding methods, to examine the distribution of data set in different ways. The results of EDA provided insights into the structure and quality of the data set and the potential factors influencing radiological outcomes. The EDA also helped us selecting appropriate statistical models and hypotheses for further analysis.

To understand the distribution of each BIRADS category better, we have shown the BI-RADS distribution and percentage in Fig. 3. Moreover, an average number of letters and words per BI-RADS distribution is shown in Figure. S1 and Figure. S2 (Supplementary material), respectively. Furthermore, we used the word clouding technique to represent the 100 most common words in Fig. 4 in the dataset to understand the data further. Ultimately, in order to get further insights from each BI-RADS category (1,2,3,4,5), we have shown the frequency of the 25 most common words in Figures S3, S4, S5, S6 and S7 (Supplementary material), respectively.

Results

Evaluation results are presented in Table 2, where we report the mean sensitivity and accuracy for each model with 95% CI using TF-IDF and word2vec as word embedding techniques on unpreprocessed and preprocessed data.

We used TF-IDF and word2vec as a word vectorizer to convert text into associated vectors to extract desired keywords and embeddings. We applied TF-IDF and

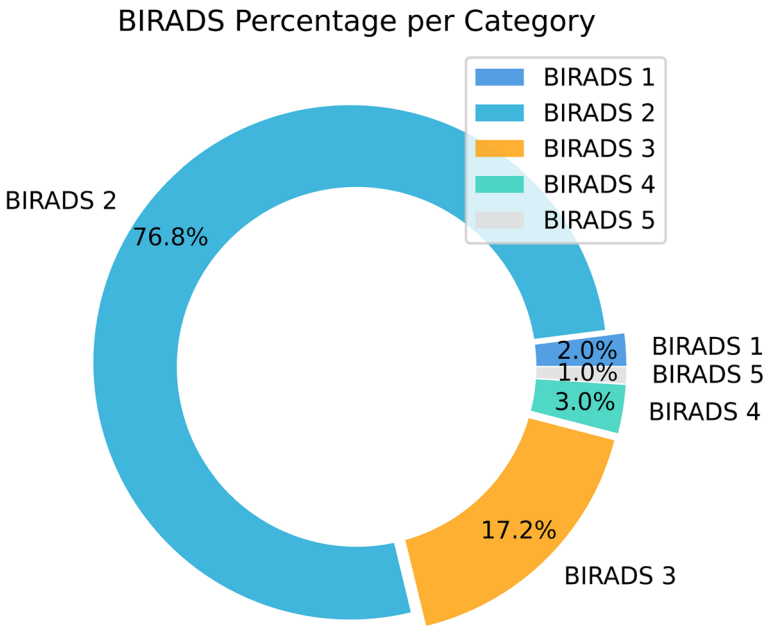

Fig. 3 Overview of the BIRADS percentage in the dataset

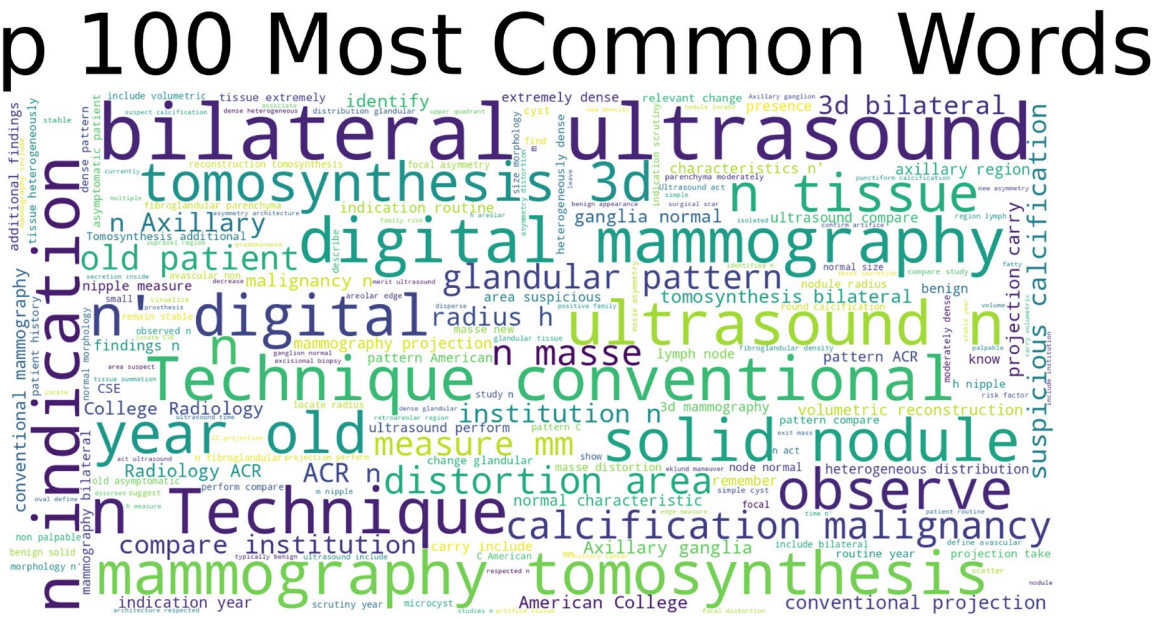

Fig. 4 Representation of 100 most common words in our dataset using word clouding

word2vec on the data that was preprocessed and the one that was not preprocessed. We used mean sensitivity and accuracy as evaluation metrics to measure the performance across all the models.

The KNN classifier using TF-IDF achieved mean sensitivity and accuracy scores of 0.34 and 0.81, respectively, on unprocessed data, and 0.36 and 0.80 on preprocessed data. In contrast, using Word2Vec, it achieved 0.33 and 0.78 on unprocessed data, and 0.30 and 0.78 on preprocessed data. The SVM classifier using TF-IDF achieved mean sensitivity and accuracy scores of 0.44 and 0.85, respectively, on unprocessed data, and 0.42 and 0.85 on preprocessed data. When using Word2Vec, it obtained 0.20 and 0.77 on both unprocessed and preprocessed

Table 1 The dataset comprises data from 5046 female participants of Mexican descent. It includes demographic details such as age and medical history, including breast implant status, prior cancer history, and previous surgeries or biopsies. Furthermore, it contains information on family cancer history, breast density classifications (A, B, C, D), and BI-RADS assessments across five categories. The dataset also records biopsy recommendations and confirmed cancer cases, as well as the mean time to a cancer event

Characteristics of dataset	
No. of Examples	5046
Sex	Women(5046), 100%
Race	Mexican Population
Age (Mean, SD, Range)	53, 9.99, 25-90
Patients Implants	No: 4185, Yes: 861
History of Previous Cancer	No: 4525, Yes: 521
Previous Surgeries/Biopsies	No: 4160, Yes: 886
Family History of Cancer	No: 4296, Yes: 750
Breast Density Distribution	A: 641, B: 1271, C: 2053, D: 1064
Distribution of BIRADS	1: 117, 2: 3921, 3: 802, 4: 129, 5: 77
Biopsy Recommendation	No: 4796, Yes: 250
Patients with Confirmed Cancer	30
Cancer Development in Five Years	61
Mean Time to Cancer Event	Mean: 1000.008 Days, SD: 898.91

data. For the NB classifier, using TF-IDF resulted in sensitivity and accuracy scores of 0.20 and 0.78, respectively, on unprocessed data, and 0.22 and 0.79 on preprocessed data. Using Word2Vec, the scores were 0.20 and 0.76 on unprocessed data, and 0.24 and 0.70 on preprocessed data.

The RF classifier using TF-IDF achieved mean sensitivity and accuracy scores of 0.34 and 0.81, respectively, on both unprocessed and preprocessed data. With Word2Vec, the scores were 0.20 and 0.77, consistent across both unprocessed and preprocessed data. Using TF-IDF, the AdaBoost classifier achieved sensitivity and accuracy scores of 0.33 and 0.41, respectively, on unprocessed data, and 0.33 and 0.74 on preprocessed data. When using Word2Vec, the scores were 0.33 and 0.65 on unprocessed data, and 0.35 and 0.61 on preprocessed data. The GB classifier using TF-IDF achieved mean sensitivity and accuracy scores of 0.43 and 0.84, respectively, on unprocessed data, and 0.45 and 0.85 on preprocessed data. Using Word2Vec, it achieved 0.30 and 0.79 on unprocessed data, and 0.31 and 0.79 on preprocessed data. The XGB classifier using TF-IDF achieved mean sensitivity and accuracy scores of 0.49 and 0.85, respectively, on unprocessed data, and 0.52 and 0.86 on preprocessed data. Using Word2Vec, it achieved 0.33 and 0.80 on both unprocessed and preprocessed data.

The LSTM classifier achieved mean sensitivity and accuracy scores of 0.42 and 0.70, respectively, on unprocessed data, and 0.53 and 0.78 on preprocessed data. The BERT classifier achieved scores of 0.40 and 0.72 on unprocessed data, and 0.54 and 0.79 on preprocessed data. Finally, the BioGPT classifier achieved mean sensitivity and accuracy scores of 0.45 and 0.74 on unprocessed data, and 0.60 and 0.80 on preprocessed data.

From Table 2, it can be observed that BioGPT performs the best among all models on BI-RADS correct category classification in terms of mean sensitivity and XGB performed best in terms of accuracy. On the other hand, SVM and RF performed worse than any other model.

Discussion

Numerous studies have been conducted on information extraction from radiological reports [30, 33–41] but only a few studies have released datasets on radiological reports and baselines [42–46]. Furthermore, a few studies have focused on structured information extraction from breast imaging-based radiological reports [12, 47–49], however, the data has not been released publicly. It has been shown in the literature that AI based breast cancer diagnosis can be greatly enhanced by the information extracted from the patient radiological reports, however, lack of public dataset has restricted further research into this domain. In order to address this gap, we have curated and make the radiological report based dataset available to drive further research in this direction.

In this work, we compared ML, DL and LLMs-based architectures. We evaluated the model's performance using mean sensitivity and accuracy. We observed that the LLM i.e., BioGPT model achieved best mean sensitivity of 0.60, which is 6% higher than the second best classifier BERT. In terms of accuracy, XGB performed best compared to all other models with an accuracy of 0.86.

In this work, we explored two word embedding techniques with ML-based classifiers to extract relevant information. We also limit our work to only extracting BI-RADS category classification. In future, we plan to extend the dataset and add associated mammography images while applying vision-language models to analyze the performance. Also, we plan to classify various other factors necessary for breast cancer diagnosis and prognosis, such as benign and malignant cases, age of the patient, family history of the cancer, risk of the cancer, recurrence, breast density and more.

Our benchmark will help and encourage the scientific community to work on extracting relevant information by applying SOTA ML, DL and LLMs architectures that will lead to rapid improvement in radiological language processing.

Table 2 This table presents detailed results from seven machine learning models and three deep learning models. The first section outlines the performance of the machine learning models, reporting mean sensitivity and accuracy using two word embedding techniques: Word2Vec and TF-IDF. Confidence intervals are provided alongside the sensitivity and accuracy scores, with results shown separately for both unprocessed and preprocessed data. The second section details the performance of the deep learning models, also reporting sensitivity, accuracy, and their corresponding confidence intervals for both preprocessed and unprocessed data

Machine learning methods								
Word embs.	**Word2Vec**				**TF-IDF**			
Model	**U-Data (mSen.)**	**P-Data (mSen.)**	**U-Data (mAcc.)**	**P-Data (mAcc.)**	**U-Data (mSen.)**	**P-Data (mSen.)**	**U-Data (mAcc.)**	**P-Data (mAcc.)**
KNN	0.33	0.30	0.78	0.78	0.34	0.36	0.81	0.80
	(0.238-0.422)	(0.210-0.389)	(0.755-0.806)	(0.757-0.808)	(0.247-0.433)	(0.266-0.454)	(0.792-0.840)	(0.782-0.830)
SVM	0.20	0.20	0.77	0.77	0.44	0.42	0.85	0.85
	(0.122-0.278)	(0.122-0.278)	(0.750-0.801)	(0.750-0.801)	(0.343-0.537)	(0.323-0.517)	(0.837-0.880)	(0.835-0.878)
NB	0.20	0.24	0.76	0.70	0.20	0.22	0.78	0.79
	(0.122-0.278)	(0.156-0.324)	(0.738-0.791)	(0.673-0.729)	(0.247-0.433)	(0.266-0.454)	(0.787-0.835)	(0.789-0.837)
RF	0.20	0.20	0.77	0.77	0.34	0.36	0.81	0.81
	(0.122-0.278)	(0.122-0.278)	(0.750-0.801)	(0.750-0.801)	(0.247-0.433)	(0.266-0.454)	(0.787-0.835)	(0.789-0.837)
AdaBoost	0.33	0.35	0.65	0.61	0.33	0.33	0.41	0.74
	(0.238-0.422)	(0.256-0.443)	(0.627-0.686)	(0.589-0.649)	(0.238-0.422)	(0.238-0.422)	(0.389-0.450)	(0.715-0.769)
GB	0.30	0.31	0.79	0.79	0.43	0.45	0.84	0.85
	(0.210-0.389)	(0.219-0.400)	(0.773-0.823)	(0.774-0.824)	(0.333-0.527)	(0.352-0.547)	(0.818-0.863)	(0.829-0.872)
XGB	0.33	0.33	0.80	0.80	0.49	0.52	0.85	0.86
	(0.238-0.422)	(0.238-0.422)	(0.778-0.828)	(0.784-0.832)	(0.392-0.588)	(0.422-0.618)	(0.838-0.881)	(0.840-0.883)
Deep learning methods								
Model		**U-Data(mSen)**		**P-Data(mSen)**		**U-Data(Accuracy)**		**P-Data(Accuracy)**
LSTM		0.42		0.53		0.70		0.78
		(0.346-0.490)		(0.455-0.619)		(0.673-0.730)		(0.753-0.805)
BERT		0.40		0.54		0.72		0.79
		(0.331-0.459)		(0.477-0.607)		(0.692-0.746)		(0.768-0.819)
BioGPT		0.45		**0.60**		0.74		0.80
		(0.235-0.669)		(0.391-0.811)		(0.710-0.764)		(0.772-0.822)

Conclusion

In conclusion, we curated and annotated a new breast imaging-based radiological reports dataset. This dataset consists of 5,046 radiological reports. These reports are based on mammography, DBT and breast US. We also compared the baseline performance of ML, DL and LLMs architectures on the dataset. This study used TF-IDF and word2vec as word embedding techniques. The BioGPT classifier with preprocessed text performed better with a mean sensitivity of 0.60, compared to all the other classifiers using TF-IDF and word2vec word embedding techniques. In terms of accuracy, XGB outperformed all the other classifiers by achieving a score of 0.86. Our work provides baselines on TECRR dataset for the researchers and clinicians for further investigation. It can improve the classification accuracy using different ML, DL and LLMs based techniques.

Abbreviations

TF-IDF	Term Frequency Inverse Document Frequency
GB	Gradient Boosting
MRI	Magnetic resonance imaging
US	Ultrasound
DBT	Digital breast tomosynthesis
DL	Deep learning
ML	Machine learning
NLP	Natural language processing
DM	Digital mammography
BI-RADS	Breast Imaging Reporting and Data System
ACR	American College of Radiology

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02717-7.

Supplementary Material 1

Acknowledgements

The authors would like to thank the Tecnológico de Monterrey and CONACYT for supporting their studies.

Authors' contributions

Sadam Hussain: Methodology, Conceptualization, Writing—Original Draft, Writing-review and Editing. Usman Naseem: Writing-Review and Editing, formatting. Mansoor Ali: Validation, Writing-Review and Editing. Daly Betzabeth Avendaño Avalos: Data collection, Annotation Validation, Writing-Review and Editing. Servando Cardona-Huerta: Data collection, Annotation, Validation, Writing-Review and Editing. Beatriz Alejandra Bosques Palomo: Data Preprocessing, Writing-Review and Editing. Jose Gerardo Tamez-Peña:

Writing—Review and Editing, Validation, Conceptualization, Supervision. All authors reviewed the manuscript.

Funding
Not applicable.

Data availability
All papers are available on publisher websites. All data generated or analyzed during this study are included in this published article. The dataset underlying this article will be shared on reasonable request to the corresponding author.

Declarations

Ethics approval and consent to participate
This retrospective study received approval from the local institutional ethical committee (protocol number: P000542-MIRAI-MODIFICADO-CEIC-CR002). The local institutional ethical committee also provided a waiver for written informed consent. A comprehensive retrospective review of the digital mammography database was conducted at TecSalud, a private hospital-based breast cancer center in Mexico. The study identified 58,321 consecutive mammography examinations (both diagnostic and screening examinations) in 37,916 women who voluntarily underwent mammography from January 2014 to June 2021. Subsequently, this was narrowed down to 20,297 screening mammography examinations from January 2014 to December 2016 (to ensure that all included women had at least 5 years of follow-up) in 13,028 women. Women were excluded if they were younger than 40 years, had an incomplete mammography.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Engineering and Sciences, Tecnológico de Monterrey, Monterrey 64849, Nuevo Leon, Mexico. [2]School of Computing, Macquarie University, Sydney 2109, NSW, Australia. [3]School of Medicine, Tecnológico de Monterrey, Monterrey 64849, Nuevo Leon, Mexico.

Received: 7 September 2023 Accepted: 10 October 2024
Published online: 24 October 2024

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin. 2015;65(2):87–108. https://doi.org/10.3322/caac.21262.
2. Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. N Engl J Med. 2005;353(17):1784–92. https://doi.org/10.1056/nejmoa050518.
3. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med. 2009;151(10):716. https://doi.org/10.7326/0003-4819-151-10-200911170-00008.
4. Oeffinger KC, Fontham ETH, Etzioni R, Herzig A, Michaelson JS, Shih YCT, et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. JAMA. 2015;314(15):1599–614.
5. Sickles EA, D'Orsi CJ. Einleitung. In: ACR BI-RADS®-Atlas der Mammadiagnostik. Springer Berlin Heidelberg; 2016. pp. 475–480. https://doi.org/10.1007/978-3-662-48818-8_15.
6. Tariq A, Assen MV, Cecco CND, Banerjee I. Bridging the Gap between Structured and Free-form Radiology Reporting: A Case-study on Coronary CT Angiography. ACM Trans Comput Healthc. 2021;3(1):1–20. https://doi.org/10.1145/3474831.
7. Cury RC, Abbara S, Achenbach S, Agatston AS, Berman DS, Budoff MJ, et al. Coronary Artery Disease - Reporting and Data System (CAD-RADS): An Expert Consensus Document of SCCT, ACR and NASCI: Endorsed by the ACC. JACC Cardiovasc Imaging. 2016;9(9):1099–113.
8. Reiner BI. The Challenges, Opportunities, and Imperative of Structured Reporting in Medical Imaging. J Digit Imaging Off J Soc Comput Appl Radiol. 2009;22:562–8.
9. Sevenster M, van Ommering R, Qian Y. Automatically Correlating Clinical Findings and Body Locations in Radiology Reports Using MedLEE. J Digit Imaging. 2011;25(2):240–9. https://doi.org/10.1007/s10278-011-9411-0.
10. Ip IK, Mortele KJ, Prevedello LM, Khorasani R. Repeat Abdominal Imaging Examinations in a Tertiary Care Hospital. Am J Med. 2012;125(2):155–61. https://doi.org/10.1016/j.amjmed.2011.03.031.
11. Cheng LTE, Zheng J, Savova GK, Erickson BJ. Discerning Tumor Status from Unstructured MRI Reports—Completeness of Information in Existing Reports and Utility of Automated Natural Language Processing. J Digit Imaging. 2009;23(2):119–32. https://doi.org/10.1007/s10278-009-9215-7.
12. Bozkurt S, Lipson JA, Senol U, Rubin DL. Automatic abstraction of imaging observations with their characteristics from mammography reports. J Am Med Inform Assoc. 2014;22(e1):e81–92. https://doi.org/10.1136/amiajnl-2014-003009.
13. Percha B, Nassif H, Lipson J, Burnside E, Rubin D. Automatic classification of mammography reports by BI-RADS breast tissue composition class. J Am Med Inform Assoc. 2012;19(5):913–6. https://doi.org/10.1136/amiajnl-2011-000607.
14. Morioka C, Meng F, Taira R, Sayre J, Zimmerman P, Ishimitsu D, et al. Automatic Classification of Ultrasound Screening Examinations of the Abdominal Aorta. J Digit Imaging. 2016;29(6):742–8. https://doi.org/10.1007/s10278-016-9889-6.
15. Solti I, Cooke CR, Xia F, Wurfel MM. Automated classification of radiology reports for acute lung injury: Comparison of keyword and machine learning based natural language processing approaches. In: 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop. IEEE; 2009. https://doi.org/10.1109/bibmw.2009.5332081.
16. Zuccon G. Automatic Classification of Free-Text Radiology Reports to Identify Limb Fractures using Machine Learning and the SNOMED CT Ontology. AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science; 2013.
17. Boumaraf S, Liu X, Ferkous C, Ma X. A New Computer-Aided Diagnosis System with Modified Genetic Feature Selection for BI-RADS Classification of Breast Masses in Mammograms. BioMed Res Int. 2020;2020(1):7695207.
18. Saslow D, Boetes C, Burke W, Harms SE, Leach MO, Lehman CD, et al. American cancer society guidelines for breast screening with MRI as an adjunct to mammography. CA Cancer J Clin. 2007;57:75–89.
19. An JY, Unsdorfer KML, Weinreb JC. BI-RADS, C-RADS, CAD-RADS, LI-RADS, Lung-RADS, NI-RADS, O-RADS, PI-RADS, TI-RADS: Reporting and Data Systems. Radiological Society of North America (RSNA); 2019. https://doi.org/10.1148/rg.2019190087.pres.
20. Burnside ES, Sickles EA, Bassett LW, Rubin DL, Lee CH, Ikeda DM, et al. The ACR BI-RADS® Experience: Learning From History. J Am Coll Radiol. 2009;6(12):851–60. https://doi.org/10.1016/j.jacr.2009.07.023.
21. D'Orsi C. Breast Imaging Reporting and Data System (BI-RADS). Lee CI, Lehman CD, Bassett LW, editors. Oxford University Press; 2018. https://doi.org/10.1093/med/9780190270261.003.0005.
22. of Radiology AC, et al. ACR BI-RADS® atlas of breast diagnostics: guidelines for diagnosis, recommendations for action and monitoring. Springer-Verlag; 2016.
23. Niknejad M, Weerakkody Y. Breast imaging-reporting and data system (BI-RADS). Radiopaedia.org; 2010. https://doi.org/10.53347/rid-10003.
24. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA, et al. ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston: American College of Radiology; 2013.
25. Jones KS. A statistical interpretation of term specificity and its application in retrieval. J Doc. 1972;28(1):11–21.
26. Mikolov T, Chen K, Corrado GS, Dean J. Efficient estimation of word representations in vector space. In International Conference on Learning Representations; 2013. https://openreview.net/forum?id=idpCdOWtqXd60.
27. Hochreiter S. Long Short-term Memory. Neural Computation MIT-Press; 1997.
28. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran

C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics; 2019. pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
29. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23(6):bbac409.
30. Banerjee I, Madhavan S, Goldman RE, Rubin D. Intelligent Word Embeddings of Free-Text Radiology Reports. AMIA Annual Symposium proceedings AMIA Symposium. 2017;2017:411–20.
31. Farouk M. Sentence semantic similarity based on Word Embedding and WordNet. 2018 13th International Conference on Computer Engineering and Systems (ICCES); 2018. p. 33–7. https://ieeexplore.ieee.org/document/8639211.
32. Lyu SY, Zhang Y, Zhang MW, Zhang BS, Gao LB, Bai LT, et al. Diagnostic value of artificial intelligence automatic detection systems for breast BI-RADS 4 nodules. World J Clin Cases. 2022;10(2):518.
33. Jnawali K, Arbabshirani MR, Ulloa AE, Rao N, Patel AA. Automatic Classification of Radiological Report for Intracranial Hemorrhage. In: 2019 IEEE 13th International Conference on Semantic Computing (ICSC). IEEE; 2019. https://doi.org/10.1109/icosc.2019.8665578.
34. Kłos M, Żyłkowski J, Spinczyk D. Automatic Classification of Text Documents Presenting Radiology Examinations. In: Advances in Intelligent Systems and Computing. Springer International Publishing; 2018. pp. 495–505. https://doi.org/10.1007/978-3-319-91211-0_43.
35. Semi-Supervised Deshmukh N, Approach Natural Language Processing, for Fine-Grained Classification of Medical Reports. In: 2019 IEEE MIT Undergraduate Research Technology Conference (URTC). IEEE; 2019. https://doi.org/10.1109/urtc49097.2019.9660430.
36. Kim C, Zhu V, Obeid J, Lenert L. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. PLoS ONE. 2019;14(2):e0212778. https://doi.org/10.1371/journal.pone.0212778.
37. Garg R, Oh E, Naidech A, Kording K, Prabhakaran S. Automating Ischemic Stroke Subtype Classification Using Machine Learning and Natural Language Processing. J Stroke Cerebrovasc Dis. 2019;28(7):2045–51. https://doi.org/10.1016/j.jstrokecerebrovasdis.2019.02.004.
38. Shin B, Chokshi FH, Lee T, Choi JD. Classification of radiology reports using neural attention models. In: 2017 International Joint Conference on Neural Networks (IJCNN). IEEE; 2017. https://doi.org/10.1109/ijcnn.2017.7966408.
39. Wheater E, Mair G, Sudlow C, Alex B, Grover C, Whiteley W. A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records. BMC Med Inform Decis Mak. 2019;19(1). https://doi.org/10.1186/s12911-019-0908-7.
40. Gorinski PJ, Wu H, Grover C, Tobin R, Talbot C, Whalley HC, et al. Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning Approaches. 2019. arXiv:1903.03985.
41. Alex B, Grover C, Tobin R, Sudlow C, Mair G, Whiteley W. Text mining brain imaging reports. J Biomed Semant. 2019;10(S1). https://doi.org/10.1186/s13326-019-0211-7.
42. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, ying Deng C. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019;6:317.
43. Jain S, Agrawal A, Saporta A, Truong S, Duong D, Bui T, et al. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. 2021. arXiv:2106.14463.
44. Bustos A, Pertusa A, Salinas JM, de la Iglesia-Vayá M. PadChest: a large chest x-ray image dataset with multi-label annotated reports. Med Image Anal. 2019;66:101797.
45. Nguyen HQ, Lam K, Le LT, Pham H, Tran DQ, Nguyen DB, et al. VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. Sci Data. 2020;9:429.
46. Datta S, Roberts K. A dataset of chest X-ray reports annotated with Spatial Role Labeling annotations. Data Brief. 2020;32:106056.
47. Patel TA, Puppala M, Ogunti RO, Ensor JE, He T, Shewale JB, et al. Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. Cancer. 2016;123(1):114–21. https://doi.org/10.1002/cncr.30245.
48. Miao S, Xu T, Wu Y, Xie H, Wang J, Jing S, et al. Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches. Int J Med Inform. 2018;119:17–21. https://doi.org/10.1016/j.ijmedinf.2018.08.009.
49. Banerjee I, Bozkurt S, Alkim E, Sagreiya H, Kurian AW, Rubin DL. Automatic inference of BI-RADS final assessment categories from narrative mammography report findings. J Biomed Inform. 2019;92:103137. https://doi.org/10.1016/j.jbi.2019.103137.

Publisher's Note