Data in Brief 61 (2025) 111761

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Mammography reporting dataset with BI-RADS system for natural language processing applications: Addressing public data gaps in Spanish

José Luis Vázquez Noguera[a], Alejandro Torres-Hurtado[b], Helena Gómez-Adorno[b,*], Julio César Mello-Román[a,c], Enrique Javier Fleitas-Alvarez[a], Federico Fernando Espinola Schulze[a,d], Miguel Garcia-Torres[a,e], Carlos Domingo Méndez Gaona[a], Pedro Esteban Gardel Sotomayor[f], Silvia Vázquez Noguera[a], Norma Elizabeth Zaracho Amarilla[g], Oxades Wilfrido Gamarra Esquivel[g]

[a] Universidad Americana, Asunción 1029, Paraguay
[b] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Coyoacán, 04510 México City, México
[c] Facultad Politécnica, Universidad Nacional de Asunción, San Lorenzo 111421, Paraguay
[d] Facultad de Ciencias Médicas, Universidad Nacional de Asunción, San Lorenzo 111421, Paraguay
[e] Division of Computer Science, Universidad Pablo de Olavide, ES-41013 Seville, Spain
[f] Universidad Católica Nuestra Señora de la Asunción, Campus Alto Paraná, Paraguay
[g] OXADES, Asunción, Paraguay

A R T I C L E   I N F O

Article history:
Received 25 March 2025
Revised 28 May 2025
Accepted 2 June 2025
Available online 7 June 2025

A B S T R A C T

Applying Natural Language Processing (NLP) to clinical reports is important for automating the analysis and classification of clinical data, improving diagnostic accuracy, and enhancing healthcare workflows. This article presents a dataset derived from mammography reports written in Spanish collected across multiple medical units operated by the Oxades company in Paraguay. The dataset contains 4357 records and

* Corresponding author.
E-mail address: helena.gomez@iimas.unam.mx (H. Gómez-Adorno).
Social media: @helen_py (H. Gómez-Adorno)

https://doi.org/10.1016/j.dib.2025.111761

Dataset link: Dataset of Spanish Mammographic Reports with BI-RADS Classification (Original data)

Keywords:
Mammography reports
Natural language processing
BI-RADS classification
Clinical data

15 variables, including the text of the complete report and also each of its sections separately (clinical observations, diagnostic conclusions, follow-up recommendations), and the BI-RADS (Breast Imaging Reporting and Data System) classification assigned to each one of the reports. Additionally, the dataset includes metadata such as report IDs, dates, and patient information such as age, patient reasons for the analysis, last menstruation period, type of hormonal therapy received, family history and number of children. To ensure patient confidentiality, all identifiable data was removed, and the dataset was structured using automated segmentation and manual verification to ensure quality and transparency. This dataset is an invaluable resource for both medical and AI research communities. It provides real-world data for developing and testing NLP algorithms and machine learning models, specifically for automating BI-RADS classification and analyzing mammography reports.

Specifications Table

Subject	Health Sciences, Medical Sciences & Pharmacology
Specific subject area	BI-RADS Classification Based on Mammographic Report Observations
Data format	Table
Type of data	CSV delimited by ";"
Data collection	Mammographic reports were collected from multiple clinics operated by the Oxades company, with ethical approval and patient anonymization. Data extraction involved segmenting reports into observations, conclusions, and recommendations, adding columns for report ID, complete text, date, and BI-RADS categorization from conclusions. Additional patient information, such as age, LMP, hormonal therapy, family history, reason and number of children, was included. Sensitive client data sections were excluded to protect privacy.
Data source location	Oxades Company: Álamo c/ Avda. Américo Picco N° 2610 Villa Elisa, Paraguay 1967 reports correspond to images acquired at Christian Medical Unit 1016 reports correspond to images acquired at Tava Capiatá Medical Unit 696 reports correspond to images acquired at Universitario Medical Unit 517 reports correspond to images acquired at Medicoop Medical Unit 114 reports correspond to images acquired at Santa Rita Medical Unit 47 reports correspond to images acquired at San Lucas Medical Unit
Data accessibility	Repository name: Zenodo Data identification number: 10.5281/zenodo.14827680 Direct URL to data: https://zenodo.org/records/14827680
Related research article	None.

1. Value of the Data

- The dataset provides a comprehensive collection of mammography reports, including BI-RADS categorization, which can be used for advancing research in automated classification and radiology report analysis.
- The data enables the development and testing of Natural Language Processing (NLP) algorithms for extracting relevant diagnostic information from mammography reports. Researchers can use this dataset to refine existing models or develop new ones for clinical applications. For example, the dataset can be used to train machine learning algorithms to develop models to classify mammography reports into BI-RADS categories based on clinical

observations, diagnostic conclusions, and follow-up recommendations [5,6]. Also, this dataset can be used to develop domain-specific NLP tools, such as automated report generation systems [7], which assist radiologists in generating structured reports based on raw text input to ensure the consistency and completeness of mammography reports.
- Due to the lack of publicly available, labeled mammography datasets, this collection helps bridge the gap, offering a standardized set for other researchers to compare methods and algorithms.
- By excluding sensitive personal data, the dataset provides a valuable resource that complies with privacy standards while still containing real and reliable medical information. The authenticity of the data, collected from actual mammography reports, ensures its high relevance and trustworthiness for healthcare research, unlike artificially generated datasets. This enhances the credibility and impact of studies in healthcare informatics.

2. Background

The research was motivated by the desire of two research teams, a Computer Vision Team from the National University of Asunción and a Natural Language Processing (NLP) Team from the National Autonomous University of Mexico, to use the dataset for developing AI models to assist in cancer diagnosis. The Oxades company was responsible for the Image Acquisition and Processing and Report Generation and Structuring, ensuring the quality and consistency of the data. The Computer Vision Team focused on applying machine learning models to mammographic images for diagnostic purposes, while the NLP Team handled Data Preparation, Anonymization, and Data Structuring, ensuring the dataset's usability for NLP applications. This collaboration aims to address the scarcity of publicly available datasets and advance AI-driven solutions for improving diagnostic accuracy in healthcare. The research drew significant inspiration from existing literature, particularly the systematic review by Casey et al. [2], which highlighted advancements and persistent challenges in the field of NLP applied to radiology reports.

A key barrier identified was the scarcity of publicly available datasets due to privacy, ethics and confidentiality concerns due to the nature of the data. Although many studies detail data collection processes and report important findings, they often refrain from publishing their data sets, making replicability and comparative analysis of algorithmic performance difficult, an example of this is the works of Kuling et al. [3] and López-Úbeda et al. [4] where it is mentioned that the use of external datasets are necessary to better generalize the results of their studies, however it is not possible due to the lack of existence of public datasets.

Addressing the scarcity of publicly available datasets in this domain enables replicability of previous studies and fosters the development of innovative diagnostic tools. Moreover, it supports advancements in evidence-based healthcare, assisting medical practitioners and researchers in improving diagnostic accuracy and enhancing patient care through AI-driven solutions.

3. Data Description

The dataset is available in CSV format (semicolon-delimited). It contains 4,357 reports and 15 variables, which capture key elements of the mammographic reports and patient data [1]. Below is a brief description of each variable:

- **ID_R:** Unique identifier for each record in the dataset.
- **Year:** The year when the mammography was conducted.
- **Month:** The month when the mammography was conducted.
- **Medical_Unit:** Indicates the folder from which each report originated. Each folder represents a different clinic where the mammography was performed.
- **Full_Report:** The full text of the mammographic report as recorded by the clinician.

Table 1
Missing values in dataset.

Column	Missing Values
ID_R	0
YEAR	0
MONTH	0
MEDICAL_UNIT	0
FULL_REPORT	0
LMP	135
HORMONAL_THERAPY	0
FAMILY_HISTORY	0
REASON	0
OBSERVATIONS	0
CONCLUSION	0
RECOMMENDATIONS	10
BI_RADS	0
AGE	15
CHILDREN	125

- **LMP:** Date of the last menstruation period of the patient.
- **Hormonal_Therapy:** Specifies the type of hormonal therapy received by the patient.
- **Familiy_History:** Indicates whether any family member has been previously diagnosed.
- **Reason:** The patient's reason for undergoing the analysis.
- **Observations:** The section within the report where detailed observations of the mammographic study are noted.
- **Conclusions:** A section in the report where a clinician assigns a BI-RADS category based on the observations.
- **Recommendations:** A section providing recommendations for the patient based on the findings of the mammography.
- **BI-RADS:** The BI-RADS category assigned to the report, which indicates the level of concern for malignancy.
- **Age:** The age of the patient at the time of the mammography.
- **Children:** The number of children the patient has.

3.1. Dataset statistics

In this section, we present various statistics related to the classes in our dataset:

Missing Values: Table 1 is a table displaying the missing values per column. The figure shows that the 'LMP' field has 135 records where the date of the last menstruation period was not included in the reports, 'Recommendations' field has 10 records where no follow-up recommendations were included in the reports. In the 'Age' field, 15 records lack patient age information. Lastly, the 'Children' field contains 125 records without data on the number of children for the respective patients. Except for these three fields, all other columns contain no missing values.

BI-RADS values: The dataset includes a distribution of BI-RADS categories as shown in Table 2. The most frequently assigned category is BI-RADS 2, with 2635 records indicating a likely benign result. BI-RADS 0 and 1 follow with 966 and 596 records respectively, representing cases needing additional imaging or negative findings. In contrast, higher BI-RADS categories that indicate increased suspicion of malignancy (BI-RADS 4, 5, and 6) are less represented, with only 52, 16, and 5 records each. This distribution highlights a class imbalance, particularly in categories associated with higher risk.

Records in years: The dataset contains records spanning five years, as detailed in Fig. 1. The highest number of records is from 2023, with 1505 entries, followed closely by 2022 with 1493 records. In contrast, 2019, 2020, 2021 and 2024 have fewer records because data collection for

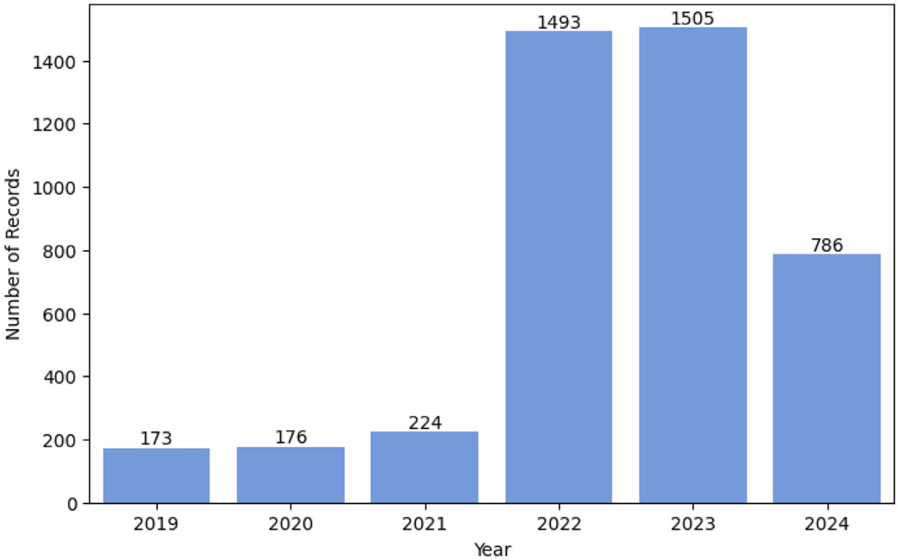

Fig. 1. Distribution of records by year.

Table 2
Number of records per BI-RADS Value.

BI-RADS Values	Number of records
0	966
1	596
2	2635
3	87
4	52
5	16
6	5

Table 3
Distribution of records by Month.

Month	Number of records
January	707
February	318
March	310
April	290
May	296
June	278
July	334
August	306
September	216
October	681
November	227
December	394

those years is available only to certain months, unlike the years 2022 and 2023 which include all 12 months.

Records in months: The distribution of records by month is summarized in Table 3. It can be observed that the months maintain a certain average of records except for 3, January, October and December, these have more because not all years within the dataset have records for the 12

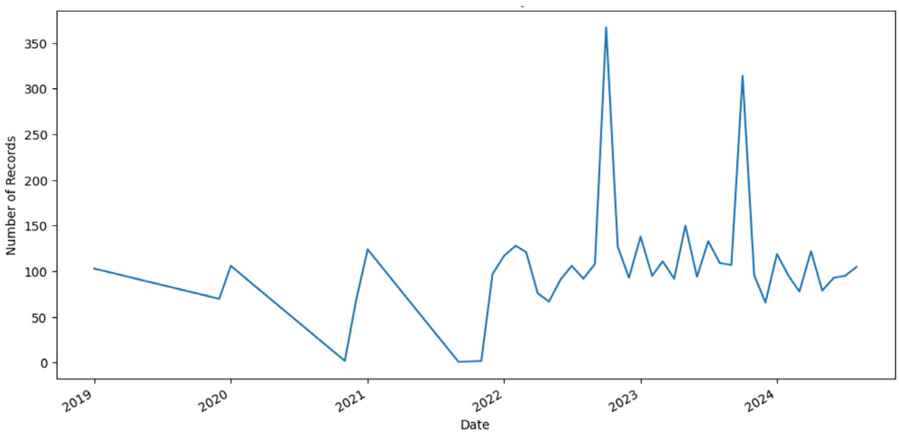

Fig. 2. Trend of records by date.

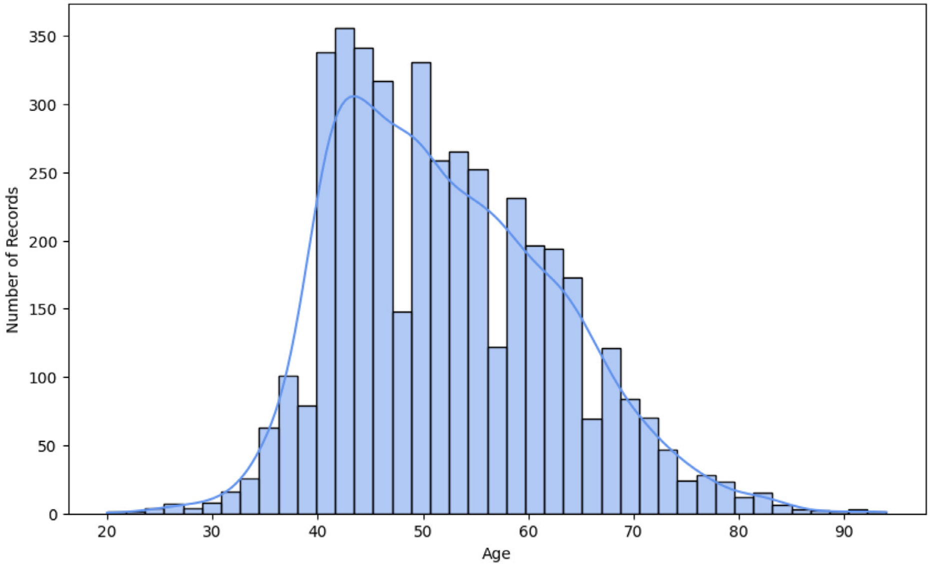

Fig. 3. Distribution of age

months, specifically the years 2019, 2020 and 2021 only have records in the months of January and December.

In addition, Fig. 2, illustrates how the number of records fluctuates over the five-year period. Notably, there are significant increases in the volume of reports as October approaches each year. This pattern aligns with heightened awareness and diagnostic activities during Breast Cancer Awareness Month, emphasizing the importance of early detection and prevention efforts.

Age distribution: Fig. 3 shows the distribution of patient ages in the dataset, covering a wide range from 23 to 94 years old. The age group with the highest frequency is between 40 and 50 years, peaking at age 42 with 158 records. There is a steady decline in records for patients older than 70 years, reflecting either lower screening rates or patient demographics. The dataset provides diverse age representation, contributing to comprehensive analyses of mammography report patterns across different life stages.

Table 4
Distribution of number of children.

Number of Children	Number of records
0	435
1	661
2	1288
3	997
4	416
5	189
6	103
7	46
8	39
9	18
10	19
11	10
12	7
13	3
15	2

Distribution of number of children: Table 4 presents the distribution of the number of children among the patients in the dataset. The largest group consists of individuals with 2 children, with 1288 records. This is followed by those with 3 children (997 records) and 1 child (661 records). Fewer individuals have higher numbers of children, with only a small number of patients reporting 13 or more children. This distribution highlights a typical family size, where most patients have between 0 to 5 children, with a significant drop in frequency for larger family sizes.

4. Experimental Design, Materials and Methods

This section describes the procedures used to acquire mammographic images and the organization of the dataset obtained from multiple clinics operated by the Oxades company in Paraguay.

4.1. Image acquisition and processing

The mammographic images used in this study were acquired at the diagnostic imaging service following strict quality protocols to ensure optimal visualization of breast structures and diagnostic precision. The imaging studies were performed using three different mammography machines and subsequently digitized using a CR Carestream system, which holds FDA-approved mammographic certification. This guarantees the fidelity of the images during the conversion from analog to digital, meeting international quality standards.

The images were captured by licensed radiology specialists trained in mammographic image acquisition, ensuring adherence to positioning protocols and image quality standards. Each study included the two fundamental mammographic projections:

- **Cranio-Caudal (CC):** Captures the breast from a top-down perspective, providing a superior-inferior view.
- **Medio-Lateral Oblique (MLO):** Captures the breast at an approximately 45° angle, allowing better visualization of the axillary region and deep tissues.

Each breast was examined using both projections, resulting in a total of four images per patient. Once obtained, the images were digitized using the CR Carestream system, ensuring that the quality remained consistent with the original scan. The digitized images were then processed and stored in a PACS (Picture Archiving and Communication System) for analysis and interpretation by diagnostic imaging specialists.

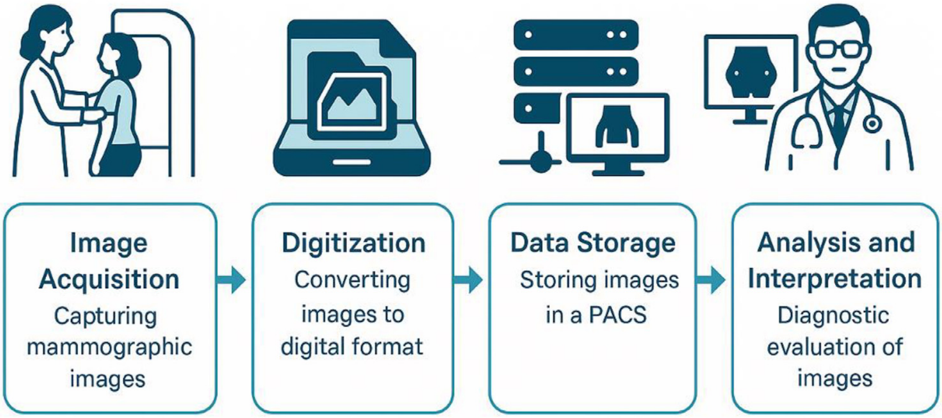

Fig. 4. Diagram of the image acquisition and processing.

Fig. 4 represents the key steps involved in the process, from the acquisition of mammographic images to their digitization, storage, and subsequent analysis.

4.2. Report generation and structuring

After the acquisition and processing of mammographic images, structured radiology reports were generated to provide a detailed clinical interpretation. Each report was prepared by a certified radiologist, following the BI-RADS (Breast Imaging Reporting and Data System) standard, ensuring consistency and reliability in diagnostic conclusions.

4.3. Report standardization

Each radiology report followed a structured format, consisting of the following sections:

- **Clinical Observations:** Detailed description of the mammographic images.
- **Diagnostic Conclusions**: BI-RADS category assigned by the radiologist based on the evaluation.
- **Follow-up Recommendations:** Suggestions based on the BI-RADS classification, including the need for additional studies or biopsies in specific cases.

4.4. Data preparation and anonymization

To ensure patient confidentiality, all identifiable personal data, such as names, dates, and medical histories, were excluded from the dataset. Only non-sensitive information such as patient age and number of children was retained. The anonymization process involved excluding the introductory section of each report, which contained personal data, as well as the final section containing the data and the signature of the treating physician, which resulted in the complete anonymity of the mammographic studies.

4.5. Data structuring and quality control

Data quality was ensured using Python scripts to segment and structure the reports into relevant sections automatically. The dataset was processed and structured using Python scripts, which automatically segmented the reports into relevant sections, including: Clinical observations, Diagnostic conclusions, Follow-up recommendations and BI-RADS classification.

Manual verification of a representative sample was performed to confirm the consistency and accuracy of the segmentation. Additionally, values with missing data were identified and documented to improve dataset transparency.

Finally, the dataset was provided in a standardized tabular format (CSV, delimited by ";"), ensuring compatibility with analytical tools and ease of use for researchers and AI applications.

Limitations

A significant limitation of this dataset is the class imbalance within the BI-RADS categories, particularly those representing higher risk classifications. While BI-RADS categories 5 and 6, which indicate probable or confirmed malignancy, have only 9 records each, category 2, representing benign findings, contains over 2,000 records. This distribution reflects real-world clinical scenarios where high-risk diagnoses are relatively rare but presents a challenge for machine learning models, which may become biased toward predicting the more common benign cases.

The imbalance may limit the dataset's immediate usability for tasks requiring balanced classification across all BI-RADS categories without additional data processing techniques, such as class reweighting or oversampling. Despite this, the dataset remains a valuable resource for studying the characteristics of mammographic findings and developing natural language processing models tailored to radiology reports. Further contributions to the dataset could help mitigate these limitations and support more robust algorithm development.

Ethics Statement

The Ethics Committee for Research (CEI) of the Faculty of Medical Sciences, National University of Asunción, approved this study under resolution CD No. 1173-00-2023, with approval code TI 384/2024, Version 2.This study adheres to the ethical standards established by the committee. All data used remain confidential, and no personally identifiable information was included in the research.

CRediT Author Statement

José Luis Vázquez Noguera: Investigation, Funding acquisition, Resources, Writing – original draft. **Alejandro Torres-Hurtado:** Data curation, Investigation, Visualization, Writing – original draft. **Helena Gómez-Adorno**: Project administration, Data curation, Investigation, Visualization, Writing – original draft. **Julio César Mello-Román**: Investigation, Writing – original draft, Writing – review & editing. **Enrique Javier Fleitas-Alvarez:** Data curation. **Federico Fernando Espinola Schulze:** Validation, Writing – review & editing. **Miguel Garcia-Torres:** Investigation. **Carlos Domingo Méndez Gaona**: Writing – review & editing. **Pedro Esteban Gardel Sotomayor:** Validation, Writing – review & editing. **Silvia Vázquez Noguera:** Writing – review & editing. **Norma Elizabeth Zaracho Amarilla**: Data curation, Validation. **Oxades Wilfrido Gamarra Esquivel**: Data curation, Writing – review & editing.

Data Availability

Dataset of Spanish Mammographic Reports with BI-RADS Classification (Original data) (ZENODO).

Acknowledgments

This research was funded by the CONACYT, Paraguay, under Grant PINV01-922. Helena Gómez-Adorno also thanks the support of PAPIIT project IN104424. During the preparation of this work, the authors used AI tools to improve readability and language. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2025.111761.

References

[1] J.L. Vázquez Noguera, H. Gómez Adorno, A. Torres Hurtado, J.C. Mello Román, E.J. Fleitas Alvarez, F.F. Espinola Schulze, M. Garcia Torres, C.D. Méndez Gaona, P.E. Gardel Sotomayor, N.E. Zaracho Amarilla, O.W. Gamorra Esquivel, Dataset of Spanish mammographic reports with BI-RADS classification (Version V1) [Data set], Zenodo (2025), doi:10.5281/zenodo.14827680.
[2] A. Casey, E. Davidson, M. Poon, et al., A systematic review of natural language processing applied to radiology reports, BMS Med. Inform. Decis. Mak. 21 (2021) 179, doi:10.1186/s12911-021-01533-7.
[3] G. Kuling, B. Curpen, A.L. Martel, BI-RADS BERT and using section segmentation to understand radiology reports, J. Imaging 8 (5) (2022) 131, doi:10.3390/jimaging8050131.
[4] P. López-Úbeda, T. Martín-Noguerol, A. Luna, Automatic classification and prioritisation of actionable BI-RADS categories using natural language processing models, Clin Radiol 79 (2024) e1–e7, doi:10.1016/j.crad.2023.09.009.
[5] M. Boroumandzadeh, E. Parvinnia, Automated classification of BI-RADS in textual mammography reports, Turk. J. Electr. Eng. Comput. Sci. 29 (2) (2021) 632–647.
[6] B. Percha, Machine Learning Approaches to Automatic BI-RADS Classification of Mammography Reports, Department of Biomedical Informatics SU,(Ed.), Standford University, 2010.
[7] P. Sloan, P. Clatworthy, E. Simpson, M. Mirmehdi, Automated radiology report generation: a review of recent advances, IEEE Rev. Biomed. Eng. 18 (2024) 368–387.