

Article

A Data Curation Framework for Unstructured Real-World Turkish Breast Imaging Reports

Seda Yıldırım ^{1,*}, Erkan Ülker ² and Necdet Poyraz ³

¹ Department of Computer Programming, Trade and Industry Vocational School, KTO Karatay University, 42020 Konya, Turkey

² Department of Computer Engineering, Faculty of Computer and Information Sciences, Konya Technical University, 42250 Konya, Turkey; eulker@ktun.edu.tr

³ Department of Radiology, Faculty of Medicine, Internal Medicine Sciences, Necmettin Erbakan University, 42090 Konya, Turkey; npoyraz@erbakan.edu.tr

* Correspondence: seda.yildirim@karatay.edu.tr

Abstract

Breast imaging reports are typically stored as unstructured free-text documents, which limits their use in clinical analytics, research, and downstream computational applications. These challenges are particularly pronounced in Turkish because of its agglutinative linguistic structure, orthographic variability, and the limited availability of standardized clinical corpora. This study presents a data curation framework for transforming heterogeneous real-world Turkish breast imaging reports into structured and machine-readable datasets, with a specific focus on the extraction and standardization of BI-RADS labels already recorded in routine clinical reports. The proposed pipeline integrates domain-specific preprocessing, text normalization, hybrid conclusion-section segmentation, multi-phase BI-RADS label extraction, duplicate and near-duplicate report handling, and modality-based separation within a transparent rule-guided workflow. The framework was applied to two real-world breast imaging datasets comprising 35,104 reports in Dataset 1 and 27,215 reports in Dataset 2 after overlap and duplicate control around the May 2023 transition period. The datasets included ultrasonography, mammography, and magnetic resonance imaging records. The framework achieved conclusion-section segmentation coverage rates of 96.71% and 99.94%, respectively, and BI-RADS label extraction coverage rates of 95.04% and 100.00%, respectively. These coverage values indicate the proportion of reports for which the rule-guided pipeline produced extractable outputs and should not be interpreted as precision, recall, F1-score, or accuracy against an independent gold-standard corpus. The curation process reduced report-level textual redundancy, improved structural consistency, and enabled systematic separation of report narratives from diagnostic assessment labels. Manual review of selected subsets was used as an internal plausibility check rather than as a substitute for independent radiologist-annotated gold-standard validation. By addressing the challenges of structuring free-text breast imaging reports in a low-resource language setting, this study provides a transparent and adaptable methodological basis for future clinical NLP, machine learning, and real-world healthcare analytics.

Keywords: data curation; BI-RADS; clinical natural language processing; Turkish breast imaging reports; information extraction

Academic Editor: Xianpeng Wang

Received: 15 June 2026

Revised: 10 July 2026

Accepted: 14 July 2026

Published: 17 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Breast imaging examinations such as ultrasonography (US), mammography (MG), and magnetic resonance imaging (MRI) routinely generate narrative radiology reports that are written in free-text form. These reports constitute the primary communication channel between radiologists and referring clinicians, containing critical diagnostic interpretations, follow-up recommendations, and clinically actionable findings. However, their unstructured nature limits their usability for large-scale analytics, automated information extraction, and downstream clinical decision-support research [1].

The continued reliance on free-text reporting introduces important challenges in consistency, standardization, and secondary data use. Variability in reporting style, ambiguous terminology, formatting inconsistencies, and incomplete documentation hinder systematic interpretation and reduce interoperability across healthcare systems [2]. As the volume of radiology data continues to grow, these limitations increasingly motivate the need for transparent and adaptable methods that can transform narrative reports into structured and machine-readable data resources.

In this context, data curation has become a critical prerequisite for reliable natural language processing (NLP), machine learning, and real-world evidence generation. Effective curation involves not only text cleaning but also semantic normalization, section segmentation, label extraction, duplicate control, and quality assurance. Nevertheless, many existing approaches remain task-specific, dataset-dependent, or narrowly optimized for English-language corpora, limiting their direct transferability to heterogeneous real-world clinical settings.

These challenges are particularly pronounced in Turkish, a morphologically rich and agglutinative language. Its complex word forms, spelling inconsistencies, frequent abbreviations, and limited availability of standardized clinical corpora make clinical text processing particularly challenging [3,4]. Although recent advances in transformer-based NLP models have shown promising performance for Turkish biomedical and clinical text processing [5,6], most studies focus on specific predictive tasks rather than on transparent end-to-end curation pipelines for radiology reports.

Breast imaging reports are particularly relevant for automated information extraction because they often contain standardized BI-RADS (Breast Imaging Reporting and Data System) categories that reflect malignancy risk and support clinical decision-making. Despite this structured diagnostic target, BI-RADS labels are often embedded in heterogeneous free-text narratives, inconsistently formatted, partially missing, or expressed using non-standard variants. As a result, their systematic extraction and standardization remain challenging in large-scale real-world settings. Furthermore, successful use in healthcare analytics requires not only label extraction but also transparent curation, standardization, quality control, and adaptability to institution-specific reporting conventions without requiring changes to the original report-writing process or clinical reporting interface.

To address these challenges, this study proposes a data curation framework for transforming unstructured Turkish breast imaging reports into structured and machine-readable datasets. The proposed framework integrates domain-specific preprocessing, terminology normalization, hybrid conclusion-section segmentation, multi-phase BI-RADS label extraction, duplicate and near-duplicate report handling, and modality-based separation in a transparent rule-guided workflow. The framework is applied to and internally assessed on two real-world datasets comprising 35,104 and 27,215 non-overlapping reports. By improving structural consistency, reducing report-level textual redundancy, and supporting the separation of report narratives from diagnostic assessment labels, the proposed approach provides a practical methodological basis for future NLP, machine learning, and healthcare analytics applications.

Contributions of This Study

The main contributions of this study can be summarized as follows:

- A transparent rule-guided data curation framework for transforming unstructured real-world Turkish breast imaging reports into structured and machine-readable datasets.
- An end-to-end workflow integrating preprocessing, conclusion-section segmentation, BI-RADS label extraction, duplicate and near-duplicate report handling, modality-based separation, and structured dataset construction.
- A multi-phase BI-RADS label extraction strategy designed to handle heterogeneous and inconsistently formatted free-text reports while explicitly distinguishing extraction coverage from independent annotation accuracy.
- Language-specific normalization and curation strategies tailored to the challenges of Turkish clinical text.
- Application and internal assessment of the framework on two real-world datasets comprising a total of 62,319 non-overlapping reports.

2. Literature Review

Medical text processing presents substantial challenges due to abbreviations, synonymous terminology, spelling variability, and domain-specific linguistic constructions [7]. These challenges are particularly evident in radiology reports, which are written as free-text narratives containing diagnostic findings, interpretations, and recommendations. Although some institutions employ partially structured templates, most real-world reports remain heterogeneous in format, limiting large-scale reuse for analytics and downstream computational applications [1,4].

Structured reporting has been recommended to improve the consistency and interoperability of radiology data and to make it more suitable for computer-assisted analysis. However, its adoption in routine clinical practice remains limited, as radiologists often require flexibility in narrative expression, efficient reporting workflows, and adaptation to institution-specific practices [8]. As a result, automated approaches that transform exported free-text radiology reports into structured datasets without requiring changes to the original report-writing process or clinical reporting interface have become increasingly important.

Early radiology NLP systems were primarily rule-based, offering transparency and strong domain control for tasks such as section detection, terminology mapping, and concept extraction. Nevertheless, purely rule-based systems often require intensive maintenance and may be sensitive to variation in reporting practices across institutions [2,3]. Later studies introduced hybrid approaches combining rules with statistical learning methods for extracting entities, measurements, and findings from radiology narratives [9]. In parallel, classical machine learning approaches based on feature representations such as bag-of-words, TF-IDF, and Word2Vec, combined with classifiers including support vector machines and random forests, have demonstrated moderate success in report classification and BI-RADS-related classification or extraction tasks [10,11].

More recently, deep learning and transformer-based approaches have further improved performance in report classification, entity recognition, and BI-RADS-related classification or extraction tasks. Contextual language models such as BERT and biomedical transformer variants have shown promise in clinical text understanding [12,13]. Several studies have also applied these methods to breast imaging reports for BI-RADS categorization and risk-oriented classification [10].

However, most of these studies are limited to English-language datasets and generally focus on specific predictive tasks rather than the full data preparation process. Despite

these methodological advances, several important limitations remain. First, while many studies prioritize model development, the data curation process, which is critical for ensuring data quality, consistency, and reusable structured outputs, has received comparatively less attention. Second, existing datasets are often small, single-centre, or limited to a single imaging modality, which restricts direct transferability and real-world applicability. Third, morphologically rich and low-resource languages such as Turkish remain underrepresented in the radiology NLP literature. In addition, practical use in clinical research and healthcare analytics requires data processing pipelines that are transparent, implementation-oriented, and adaptable to heterogeneous reporting formats without disrupting existing report-writing practices.

This gap is particularly relevant in Turkish radiology reporting, where complex word forms, spelling inconsistencies, frequent abbreviations, and the limited availability of standardized clinical corpora make automated processing especially challenging. Existing Turkish studies remain limited in both scale and scope, often focusing on task-specific objectives such as classification or de-identification rather than on transparent end-to-end curation pipelines for radiology reports. As summarized in Table 1, prior studies have predominantly focused on task-specific objectives such as classification, information extraction, benchmarking, summarization, dataset creation, de-identification, and prediction. This gap is also supported by a recent scoping review of NLP research on breast imaging reports, which showed that the literature is concentrated mainly on BI-RADS classification/annotation and information extraction tasks, while Turkish breast imaging NLP studies remain limited and reproducibility-related reporting issues persist [14]. By comparison, comprehensive curation pipelines that integrate preprocessing, normalization, section segmentation, BI-RADS label extraction, duplicate and near-duplicate report handling, and modality-based separation have received far less attention. Existing approaches also tend to focus on single-modality data or narrowly defined tasks rather than offering end-to-end solutions for heterogeneous radiology reports in real-world settings. Accordingly, the present study addresses this gap by proposing a transparent rule-guided data curation framework for Turkish breast imaging reports. The proposed approach aims to transform heterogeneous real-world narratives into structured and analysis-ready datasets that can support clinical NLP, machine learning, and real-world healthcare analytics.

Table 1. Comparison of representative studies in radiology NLP with respect to language, task, method, and BI-RADS relevance.

Study	Language	Task	Method	BI-RADS Related
Pons et al. (2016) [1]	English	Review	Rule-based/ML	No
Bozkurt et al. (2019) [9]	English	Information Extraction	Hybrid NLP	No
Boroumandzadeh et al. (2021) [8]	Turkish	Classification	ML-based	Yes
Hepsağ et al. (2024) [15]	Turkish	Classification	Transformer-based	Yes
López-Úbeda et al. (2024) [16]	English	Classification	ML/DL-based	Yes
Hussain et al. (2024) [17]	English	Dataset Creation	Dataset-based	Yes
Fanni et al. (2023) [10]	English	Report Structuring	NLP-based	No
Nguyen et al. (2020) [11]	English	Summarization	Hybrid ML	No
Çelikten et al. (2021) [18]	Turkish	Text Classification	Transformer-based	No
Dilmac et al. (2026) [5]	Turkish	De-identification	Transformer-based	No
Reichenpfader et al. (2024) [19]	English	Review	LLM-based	No
Zhang et al. (2023) [20]	English	EHR Processing	Deep Learning	No
This study	Turkish	Data Curation	Rule-guided pipeline	Yes

3. Data Curation Framework

Radiology reports are generated as free-text narratives containing imaging findings, quantitative measurements, anatomical descriptions, diagnostic interpretations, and follow-up recommendations. Although these reports are highly informative for clinical use, their unstructured nature limits their direct applicability in large-scale analytics, electronic health record integration, and automated clinical data processing.

The processing of Turkish clinical narratives presents additional challenges because of the agglutinative structure of the language, morphological variability, orthographic inconsistencies, and the frequent use of domain-specific terminology. In addition, the limited availability of large-scale standardized Turkish clinical datasets complicates the development of transparent and adaptable clinical NLP pipelines. These challenges require domain-aware, transparent, and implementation-oriented data curation frameworks.

In breast imaging NLP, free-text reports can be transformed into structured representations by identifying clinically relevant entities, normalizing terminology, extracting findings and their attributes, and detecting standardized assessment categories such as BI-RADS. Building on this general principle, the present study introduces a rule-guided data curation framework that integrates domain-specific preprocessing, terminology normalization, conclusion-section segmentation, BI-RADS label extraction, duplicate and near-duplicate report handling, and modality-based separation within a transparent workflow. The proposed framework is designed to transform heterogeneous Turkish breast imaging reports into structured, machine-readable representations while supporting the preservation of clinically relevant report-level information where possible and maintaining traceability across processing stages. An overview of the end-to-end workflow is presented in Figure 1, and the detailed architecture of the proposed framework is illustrated in Figure 2.

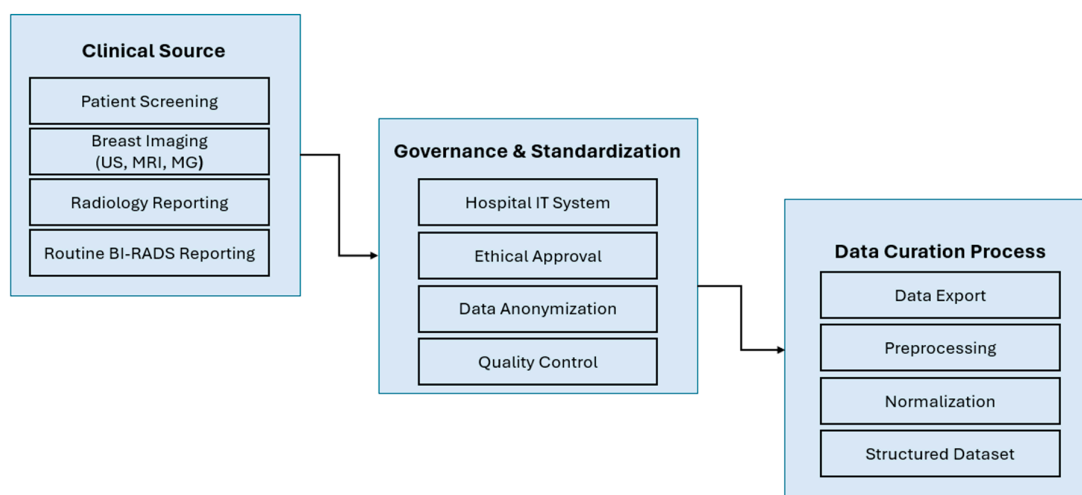


Figure 1. End-to-end data curation workflow for transforming exported unstructured Turkish breast imaging reports into structured, machine-readable report-level data. The workflow consists of three main stages: clinical source generation, governance and standardization, and data curation processing.

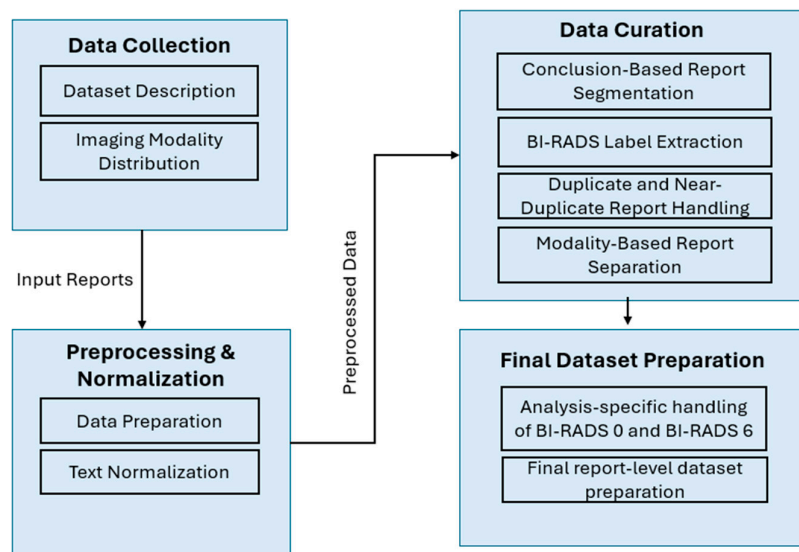


Figure 2. Detailed architecture of the proposed data curation framework for Turkish breast imaging reports. The workflow consists of four main stages: data collection, preprocessing and normalization, rule-guided data curation, and final report-level dataset preparation.

The stepwise implementation of the proposed rule-guided data curation pipeline is summarized in Algorithm 1.

Algorithm 1. Rule-guided data curation pipeline for Turkish breast imaging reports

Input: Retrospective report-level radiology report collection R

Output: Structured and curated report-level dataset D

1. Load retrospective report-level data exported from the hospital information system.
 2. Apply text preprocessing, including Unicode normalization, Turkish character standardization, spacing correction, punctuation normalization, and removal of non-informative repetitive template fragments.
 3. Detect modality-specific indicators for ultrasonography, mammography, and magnetic resonance imaging using predefined keyword patterns.
 4. Assign modality attributes, including *PrimaryModality*, *MultiModalityFlag*, and *AmbiguityFlag*.
 5. Segment the conclusion section using hybrid rule-based and fuzzy matching strategies.
 6. Extract BI-RADS expressions using multi-phase regular-expression and rule-based matching.
 7. Standardize extracted BI-RADS labels using a predefined two-level scheme consisting of coarse BI-RADS categories and optional BI-RADS 4 subcategory attributes.
 8. Assign Unknown labels and review flags for missing, ambiguous, inconsistent, or low-confidence cases.
 9. Identify exact duplicates and flag potential near-duplicates using identifier-based exact matching and similarity-based comparison.
 10. Generate modality-separated structured datasets and descriptive curation statistics.
 11. Return the final structured report-level dataset D .
-

3.1. Data Collection

3.1.1. Dataset Description

The dataset used in this study was retrospectively collected from the hospital information system following approval from the Necmettin Erbakan University Ethics Com-

mittee. All procedures were conducted in accordance with relevant ethical and data protection regulations. Data collection was performed in two temporally distinct phases. Dataset 1 comprised 35,104 radiology reports collected between 2 January 2020 and 8 May 2023. Dataset 2 initially comprised 27,254 reports collected between 8 May 2023 and 17 January 2025. Since both extraction periods included records from the boundary date, an overlap and consistency check was performed before the final analysis. This check identified 39 records requiring exclusion because of temporal overlap and/or export-level inconsistency. After their removal, Dataset 2 included 27,215 reports. The resulting non-overlapping report-level curation archive therefore consisted of 62,319 radiology reports.

The unit of analysis in this study was the radiology report rather than the individual patient. Each report contained three main fields available for research use: a patient/file identifier, the report date, and the free-text radiology report. The report-level characteristics of the two datasets after temporal overlap and consistency control are summarized in Table 2. Dataset 1 included 17,928 valid unique patient/file identifiers, with a mean of 1.96 reports per identifier, a median of 1 report per identifier, and a range of 1–18 reports per identifier. Dataset 2 included 11,910 valid unique patient/file identifiers, with a mean of 2.29 reports per identifier, a median of 2 reports per identifier, and a range of 1–11 reports per identifier. These values indicate that repeated reports for the same identifier were present in both datasets, which is expected in real-world breast imaging archives due to follow-up examinations, multimodal imaging, and longitudinal clinical assessment.

Table 2. Report-level characteristics of the two datasets after temporal overlap and consistency control.

Characteristic	Dataset1	Dataset2
Date range	2 January 2020–8 May 2023	8 May 2023–17 January 2025
Total reports	35,104	27,215
Valid unique patient/file identifiers	17,928	11,910
Mean reports per identifier	1.96	2.29
Median reports per identifier	1	2
Min–max reports per identifier	1–18	1–11

The inclusion criteria were routine breast imaging reports written in Turkish free-text format, recorded in the hospital information system within the predefined study periods, and related to ultrasonography, mammography, magnetic resonance imaging, or combined breast imaging examinations. No reports were excluded from the initial curation stage solely because of reporting style, formatting heterogeneity, linguistic variation, or missing BI-RADS expressions. Records without explicit BI-RADS labels, BI-RADS 0 and BI-RADS 6 categories, and potential near-duplicate reports were handled separately in later analysis steps according to the objective of each analysis.

The datasets consisted of routine clinical radiology reports generated during standard diagnostic workflows. The proposed framework did not require any change to the original report-writing process or clinical reporting interface; however, it transformed the exported report-level data through preprocessing, label extraction, standardization, and duplicate-control procedures for research use. Data access was conducted under ethics committee approval and institutional data protection procedures. The dataset provided to the research team contained only a patient/file identifier, report date, and free-text report. Direct personal identifiers such as patient names, national identification numbers, addresses, telephone numbers, and birth dates were not included in the analysis dataset. The patient/file identifier was used only for data organization, repeated-report assessment, temporal overlap control, and consistency checking. No attempts were made to re-identify patients or link the records to external data sources. Age, sex, and other demographic

variables were not available in the report-level dataset. Therefore, age range, age distribution, and other patient-level demographic characteristics could not be reported. This limits the assessment of demographic representativeness. Although repeated reports could be identified at the patient/file identifier level, lesion-level identifiers and detailed longitudinal clinical metadata were not available.

The distribution of routinely recorded BI-RADS categories in the two datasets is presented in Table 3. The clinical interpretation of BI-RADS categories is summarized in Table 4 based on the ACR BI-RADS framework [21], providing the clinical basis for subsequent analyses. The BI-RADS labels used in this study were not newly assigned by the research team. They were extracted from routinely generated clinical reports in which BI-RADS assessments had been recorded by expert radiologists as part of standard radiological reporting. Therefore, these routinely recorded BI-RADS assessments were treated as source expressions for rule-based extraction and standardization, not as independently annotated gold-standard labels. Accordingly, the reported extraction coverage reflects whether an explicit BI-RADS expression could be identified and standardized by the pipeline; it should not be interpreted as independent annotation accuracy. The automated methods proposed in this study were not designed to generate new BI-RADS labels, but to identify, extract, and standardize labels already present in free-text reports. Accordingly, the aim of the proposed framework was not clinical label generation, but the transformation of unstructured radiology reports into structured and machine-readable data. A marked decrease in MRI reports was observed in Dataset 2 compared with Dataset 1. This difference was not caused by an exclusion or preprocessing decision made by the research team; rather, it reflects the distribution of the real-world report archive provided through the hospital information system for the second period. Therefore, MRI-related findings in Dataset 2 should be interpreted cautiously due to the limited representation of MRI reports in that period.

Table 3. Distribution of routinely recorded BI-RADS categories in Dataset1 and Dataset2.

BI-RADS Category	Dataset1 (n, %)	Dataset2 (n, %)
BI-RADS 0	1366 (3.89)	2678 (9.84)
BI-RADS 1	9785 (27.87)	9741 (35.79)
BI-RADS 2	17,031 (48.52)	12,053 (44.29)
BI-RADS 3	3751 (10.69)	1827 (6.71)
BI-RADS 4, total	608 (1.73)	340 (1.25)
BI-RADS 4, unspecified	257 (0.73)	125 (0.46)
BI-RADS 4A	168 (0.48)	128 (0.47)
BI-RADS 4B	78 (0.22)	48 (0.18)
BI-RADS 4C	105 (0.30)	39 (0.14)
BI-RADS 5	539 (1.54)	254 (0.93)
BI-RADS 6	854 (2.43)	322 (1.18)
No explicit BI-RADS label	1170 (3.33)	0
Total	35,104 (100.0)	27,215 (100.0)

Note. BI-RADS 4A, 4B, and 4C are subcategories of BI-RADS 4 and are reported descriptively when explicitly stated in the report text. They are not counted as independent main BI-RADS categories in the primary coarse-label scheme.

Table 4. BI-RADS assessment categories and corresponding clinical interpretations based on the ACR BI-RADS framework [21].

BI-RADS Category	Clinical Meaning
BI-RADS 0	Incomplete—additional imaging evaluation and/or prior examinations are required.
BI-RADS 1	Negative—no abnormal findings.

BI-RADS 2	Benign finding.
BI-RADS 3	Probably benign—short-interval follow-up is recommended.
BI-RADS 4	Suspicious abnormality—biopsy should be considered.
BI-RADS 5	Highly suggestive of malignancy.
BI-RADS 6	Biopsy-proven malignancy.

3.1.2. Imaging Modality Distribution

The dataset consisted of breast imaging reports containing free-text narratives from ultrasonography (US), mammography (MG), and magnetic resonance imaging (MRI). In real-world clinical practice, reports may correspond to a single imaging modality or include findings from multiple modalities within the same document. To characterize the structural composition of the datasets, a rule-based modality analysis was performed using modality-specific keyword patterns. The keyword groups included common Turkish and abbreviated modality expressions, such as “ultrasonografi”, “ultrason”, and “US” for ultrasonography; “mamografi”, “mammo”, and “MG” for mammography; and “manyetik rezonans”, “MR”, “MRI”, and “MRG” for magnetic resonance imaging.

For each report, three attributes were derived:

- (i) the dominant imaging modality (*PrimaryModality*),
- (ii) the presence of multiple modalities (*MultiModalityFlag*), and
- (iii) the ambiguity of modality dominance (*AmbiguityFlag*).

To determine the dominant imaging modality, modality-specific indicators were first identified within each report. The *PrimaryModality* was then assigned based on structural and contextual evidence, including header-level indicators, modality-specific text segments, sentence-level evidence, and references to imaging studies mentioned within the report narrative. Reports containing indicators of more than one modality were marked using *MultiModalityFlag*. Reports in which competing modality signals could not be clearly resolved were labelled using *AmbiguityFlag*. This rule-based modality assignment was performed on the overlap-controlled report datasets before downstream BI-RADS-specific filtering in order to preserve real-world reporting variability. The modality assignment step was incorporated into the overall data curation pipeline summarized in Algorithm 1.

The comparative distribution of imaging modalities across both datasets is presented in Table 5. Ultrasonography (US) constituted the largest proportion of reports in both datasets, although its dominance was more pronounced in Dataset 1 than in Dataset 2. In contrast, mammography (MG) exhibited a higher relative proportion in Dataset 2 than in Dataset 1, indicating a more balanced US–MG distribution in the second dataset. Magnetic resonance imaging (MRI) remained limited in both datasets and was particularly underrepresented in Dataset 2. Reports with unidentified modality were rare in both datasets.

Structural characteristics of the datasets are summarized in Table 6. Most reports corresponded to single-modality examinations, accounting for 62.5% in Dataset 1 and 66.8% in Dataset 2, while multi-modality reports constituted 37.5% and 33.2%, respectively. In terms of rule-derived modality certainty, 80.0% of reports in Dataset 1 and 84.3% in Dataset 2 exhibited a clearly defined dominant modality, whereas 20.0% and 15.7% were classified as uncertain. These findings indicate a slightly higher degree of rule-derived modality clarity in the second dataset.

Table 5. Rule-derived distribution of primary imaging modalities across the two radiology report datasets.

Modality	Dataset1 (n, %)	Dataset2 (n, %)
Ultrasonography (US)	24,034 (68.5)	14,547 (53.4)
Mammography (MG)	10,032 (28.6)	12,627 (46.4)
Magnetic resonance imaging (MRI)	925 (2.6)	34 (0.1)
Unidentified modality	113 (0.3)	7 (<0.1)
Total	35,104 (100.0)	27,215 (100.0)

Table 6. Structural characteristics of reports according to rule-derived modality composition and modality certainty.

(a) Modality Composition		
Report Type	Dataset1 (n, %)	Dataset2 (n, %)
Single modality	21,944 (62.5)	18,193 (66.8)
Multi-modality	13,160 (37.5)	9022 (33.2)
Total	35,104 (100.0)	27,215 (100.0)
(b) Modality Certainty		
Report Type	Dataset1 (n, %)	Dataset2 (n, %)
Certain	28,094 (80.0)	22,947 (84.3)
Uncertain	7010 (20.0)	4268 (15.7)
Total	35,104 (100.0)	27,215 (100.0)

3.2. Preprocessing and Normalization

3.2.1. Data Preparation

A multi-stage data preparation pipeline was developed to transform raw report-level radiology data into a consistent format for subsequent rule-based information extraction. This stage included structural segmentation and preparation of report components. Analysis of the raw datasets showed that radiology reports typically contain both descriptive clinical findings and final assessment statements within the same document. In many reports, the segment preceding conclusion-like section headers, such as “SONUÇ”, “Sonuç”, or their spelling variants, corresponds to the clinical narrative, whereas the subsequent segment contains the final assessment statement, including BI-RADS labels when explicitly reported. Because conclusion headers were not always written in a uniform format, this separation was implemented as a rule-guided structural preparation step rather than as an independently validated clinical section annotation.

Based on this observation, each report was decomposed into two components:

- (i) the clinical narrative and
- (ii) the final assessment segment.

A synthetic example illustrating this structure is presented in Table 7. The example does not correspond to an individual patient record and was included only to illustrate the structural distinction between narrative findings and the final assessment segment. This structural separation enables a clear distinction between descriptive content and diagnostic assessment, supporting systematic downstream extraction and dataset structuring. The separated final assessment segment was subsequently used as the primary search region for BI-RADS label extraction, while the clinical narrative was retained as report text for downstream structured dataset construction. All preprocessing steps were implemented in Python 3.12.13 using standard data processing libraries, including Pandas 2.2.2, NumPy 2.0.2, and RapidFuzz 3.14.5.

Table 7. Synthetic example of a radiology report excerpt illustrating the unstructured nature of clinical narratives, translated from Turkish.

ID	Date	Clinical Report
ID-001	Example date	Bilateral breast ultrasound findings describe heterogeneous fibroglandular tissue and postoperative changes in one breast. No definite mass lesion or suspicious axillary lymphadenopathy is reported. The final assessment section includes a benign BI-RADS assessment. CONCLUSION: Postoperative benign findings. BI-RADS 2.

3.2.2. Text Normalization

Text normalization was performed to support consistent structural parsing and rule-based pattern matching across radiology reports. Unicode normalization was applied using the NFKC scheme, followed by standardization of case formatting and removal of redundant whitespace. In addition, domain-specific normalization rules were introduced to reduce spelling inconsistencies commonly observed in clinical narratives. These included the standardization of measurement expressions, for example, converting heterogeneous variants such as “15 x 7”, “15x7”, and “15X7” into a unified measurement format, and normalization of frequently used modality-related abbreviations. Advanced linguistic transformations, such as stemming or lemmatization, were intentionally avoided to preserve original clinical expressions and maintain traceability. This design choice helped preserve the original clinical wording throughout the curation process.

3.3. Data Curation

3.3.1. Conclusion-Based Report Segmentation

Radiology reports exhibit substantial variability in the representation of the conclusion keyword corresponding to “SONUÇ” in Turkish clinical reports, including case variations, missing characters, punctuation differences, and concatenated forms. Keyword variants were identified using fuzzy string matching based on *RapidFuzz* similarity scoring and incorporated into a dynamic regular expression framework. Segmentation was performed using a three-stage approach:

- (i) regex-based detection,
- (ii) fuzzy matching as a fallback mechanism, and
- (iii) full-text retention for unresolved cases.

This hybrid design was intended to improve segmentation coverage across heterogeneous reporting styles while preserving transparency and traceability in the segmentation process. Representative variants captured common spelling and formatting differences in the Turkish conclusion keyword, including forms such as “sonuç”, “sonuc”, “sonuç:”, “sonuc:”, and other punctuated or concatenated expressions. This procedure was used to separate reports, where possible, into a clinical narrative component and a conclusion segment that could contain BI-RADS-related assessment information. Algorithm 2 summarizes the proposed hybrid segmentation strategy for detecting conclusion sections under spelling and formatting variability.

The fuzzy matching threshold was set to 88 to balance sensitivity in detecting misspelled keywords and reducing false-positive section splits. This threshold was selected empirically during pipeline development and was applied consistently to both datasets. This design allowed unresolved cases to be retained rather than discarded during segmentation, thereby preserving report-level data completeness while limiting unresolved segmentation cases in heterogeneous reporting conditions. Additional implementation parameters and threshold settings are provided in Supplementary Table S1.

Algorithm 2. Identification and segmentation of the conclusion section**Input:** Radiology report R **Parameter:** $\tau_{fuzzy} = 88$ **Output:** *ClinicalNarrative*, *ConclusionSegment*

1. Normalize text R using case folding and whitespace normalization.
2. Apply regex patterns to detect conclusion keyword variants:
{"sonuç", "sonuc", "sonuç:", "sonuc:", "sonuç.", "sonuc."}
3. If a regex match found then:
4. Split R into:
5. *ClinicalNarrative* $\leftarrow$ text before keyword
6. *ConclusionSegment* $\leftarrow$ text after keyword
7. Else:
8. Apply fuzzy matching using *RapidFuzz* similarity scoring
9. If max similarity score $\geq \tau_{fuzzy}$ then:
10. Split R using approximate keyword location
11. Else:
12. *ClinicalNarrative* $\leftarrow R$
13. *ConclusionSegment* $\leftarrow$ NULL
14. Return *ClinicalNarrative*, *ConclusionSegment*

3.3.2. BI-RADS Label Extraction

BI-RADS expressions in radiology reports exhibit substantial variability, including concatenated, fragmented, hyphenated, and Roman numeral formats. This variability poses a challenge for extracting labels from real-world clinical reports. To address this issue, a multi-phase extraction strategy was developed to support the systematic identification and standardization of BI-RADS categories already recorded in the report text. This step was designed to extract source expressions already present in routine reports rather than to infer or assign new clinical BI-RADS categories. Following conclusion-based report segmentation, BI-RADS labels were extracted using a multi-phase regex-based framework designed to capture diverse expression patterns, including standard forms, spacing variations, and irregular textual representations.

The extraction process consisted of three sequential stages:

- (i) strict regex-based matching for standard BI-RADS expressions,
- (ii) flexible pattern matching to capture spelling and formatting variations, and
- (iii) fallback detection for irregular, fragmented, or non-standard representations.

This hierarchical design first searched for standard BI-RADS expressions and then applied broader matching rules when standard forms were not detected. To avoid ambiguity between main BI-RADS categories and explicitly stated BI-RADS 4 subcategories, a two-level labelling scheme was used. The main BI-RADS category variable represented the standardized category from 0 to 6. When BI-RADS 4 subcategories were explicitly stated in the report text, 4A, 4B, and 4C were normalized under the main BI-RADS 4 category, while the original subcategory information was retained in a separate optional BI-RADS 4 subcategory field. Reports in which no BI-RADS expression could be extracted were assigned to the "Unknown" category and marked using a *needs_review* flag to indicate unresolved or potentially ambiguous cases.

Algorithm 3 summarizes the proposed extraction framework for BI-RADS label identification under diverse spelling, formatting, and notation variations, including Roman numeral representations. The extracted main label denotes the standardized BI-RADS category, whereas the optional subcategory field preserves explicitly reported BI-RADS 4 subcategory information when available. This design supports consistent main-category analysis while retaining additional subcategory information for descriptive reporting

when explicitly present in the original report, without treating these subcategories as independent main BI-RADS classes. Representative rule categories, implementation parameters, and reproducibility settings for BI-RADS extraction are provided in Supplementary Table S1.

Algorithm 3. BI-RADS Label Extraction

Input: Report text R

Output: Main BI-RADS label L_{main} , optional BI-RADS 4 subcategory L_{sub} , *needs_review* flag

1. Normalize text R using case folding, whitespace normalization, and punctuation harmonization.

2. Apply strict regex patterns for standard BI-RADS forms:

Examples: "BI-RADS 2", "BIRADS 4", "BI-RADS-5"

3. If strict match found then:

4. Extract candidate label L_{raw}

5. Else:

6. Apply flexible regex patterns for spelling and formatting variations:

Examples: "bi rads 3", "bi-rads4", "birads:2"

7. If flexible match found then:

8. Extract candidate label L_{raw}

9. Else:

10. Apply fallback detection for irregular or fragmented expressions:

Examples: "B I R A D S", "BIRADS", Roman numeral forms

11. If fallback match found then:

12. Extract candidate label L_{raw}

13. Else:

14. $L_{main} \leftarrow \text{Unknown}$

15. $L_{sub} \leftarrow \text{NULL}$

16. $needs_review \leftarrow 1$

17. Return L_{main} , L_{sub} , $needs_review$

18. If L_{raw} is expressed in Roman numerals then:

19. Convert L_{raw} to the corresponding numeric category

20. If $L_{raw} \in \{4A, 4B, 4C\}$ then:

21. $L_{main} \leftarrow 4$

22. $L_{sub} \leftarrow L_{raw}$

23. Else:

24. $L_{main} \leftarrow \text{numeric BI-RADS category}$

25. $L_{sub} \leftarrow \text{NULL}$

26. $needs_review \leftarrow 0$

27. Return L_{main} , L_{sub} , $needs_review$

3.3.3. Duplicate and Near-Duplicate Report Handling

Exact duplicate reports were removed, whereas potential near-duplicate reports were identified and handled conservatively using a two-stage similarity-based approach to reduce report-level textual redundancy. Exact duplicates were eliminated based on strict equality of the patient/file identifier, normalized report text, and extracted main BI-RADS label where available. For near-duplicate detection, textual similarity was computed using the *token_set_ratio* metric from the *RapidFuzz* library, which is robust to word reordering and partial textual overlap. Report pairs with similarity scores of 95% or higher were flagged as candidate near-duplicates.

The similarity threshold of 95% was determined empirically through preliminary inspection of highly similar report pairs. Lower threshold values, such as 85–90%, tended

to group reports with similar template structures but different textual details, whereas higher thresholds above 95% missed near-duplicate report texts with minor wording or formatting variations. The selected threshold was therefore used as a conservative criterion to prioritize highly similar report texts while reducing the risk of grouping textually distinct reports.

To reduce inappropriate comparisons and improve computational efficiency, near-duplicate comparisons were performed within the same patient/file identifier and were further restricted to records sharing the same extracted main BI-RADS category. No report was excluded solely based on text length. Overall, this design supports controlled report-level textual redundancy reduction; however, it should not be interpreted as a definitive assessment of clinical redundancy because lesion-level identifiers, procedure-level identifiers, and detailed longitudinal clinical metadata were not available.

To further assess the text-level consistency of the near-duplicate handling step, near-duplicate candidates selected for removal were logged together with their retained paired reports, similarity scores, BI-RADS labels, and removal reasons. A random sample of these candidate pairs from each dataset was then manually inspected to determine whether the automated keep/drop decision was textually reasonable and whether potential information loss could occur. This manual inspection was performed as a text-level validation step and was not intended to establish clinical diagnostic equivalence between reports. The results of the duplicate and near-duplicate handling procedure and the manual text-level review are reported in the internal validation section.

3.3.4. Modality-Based Report Separation

To address modality-related ambiguity in heterogeneous breast imaging reports, a modality-based separation strategy was implemented as part of the curation workflow. Reports were analyzed using predefined modality-specific indicators for ultrasonography, mammography, and magnetic resonance imaging, and modality-related evidence within each report was identified to support modality-aware structuring.

The primary imaging modality was assigned using rule-based structural and contextual evidence, including header-level indicators, modality-specific text segments, sentence-level mentions, and references to externally performed imaging studies. Mentions appearing in structurally salient sections of the report were given greater importance, whereas references to external examinations were down-weighted to prioritize the modality most likely to correspond to the primary diagnostic report content.

Following this process, each report was enriched with structured modality metadata, including *PrimaryModality*, *MultiModalityFlag*, and *AmbiguityFlag*. This representation supports more systematic modality-aware interpretation of report narratives by aligning extracted information with the most relevant imaging modality and by explicitly marking reports with multiple or ambiguous modality signals. It also enables the curated dataset to distinguish between single-modality, multi-modality, and ambiguous reports, which are common in real-world clinical reporting.

This modality separation step was rule-derived and should be interpreted as a curation-oriented structuring procedure rather than as an independently validated modality annotation. The resulting modality attributes were used for descriptive characterization and modality-aware dataset organization.

3.4. Final Dataset Preparation

To improve analytical consistency and reduce report-level textual redundancy, a multi-stage data curation strategy was applied. This process included analysis-specific handling of BI-RADS categories, duplicate and near-duplicate report handling, and construction of structured report-level outputs. In BI-RADS-based analyses, the inclusion of

categories depends on the clinical objective and data consistency requirements. BI-RADS 0 and BI-RADS 6 are commonly treated separately because of their distinct clinical implications. BI-RADS 0 represents an incomplete assessment requiring additional imaging and therefore does not provide a definitive diagnostic assessment. In contrast, BI-RADS 6 corresponds to biopsy-proven malignancy and reflects a confirmed diagnosis rather than a primary diagnostic interpretation.

Previous studies have adopted different strategies for handling BI-RADS categories depending on the target task, dataset composition, and clinical objective [15,16,19,22]. Building on this rationale, BI-RADS 0 and BI-RADS 6 were retained in the descriptive report-level curation archive but excluded from the task-specific BI-RADS 1–5 analysis-ready subset. This decision was intended to reduce clinical heterogeneity in the subset used for subsequent analyses focused on primary radiological assessment categories.

This handling is particularly relevant for automated NLP-based analysis because BI-RADS 0 cases represent incomplete assessments requiring additional imaging, whereas BI-RADS 6 cases correspond to biopsy-proven malignancies and may introduce post-diagnostic bias. Therefore, retaining these categories in the task-specific BI-RADS 1–5 subset would not be fully aligned with the objective of constructing an analysis-ready dataset focused on primary radiological assessment. Accordingly, BI-RADS 0 and BI-RADS 6 were retained for descriptive reporting but were not included in the BI-RADS 1–5 analysis-ready subset.

Following preprocessing and label extraction, structured report-level outputs were constructed using three core components:

- (i) record identifier (ID),
- (ii) clinical narrative (TEXT), and
- (iii) BI-RADS label (CLASS).

Entries labelled as “Unknown” or marked with a `needs_review` flag were excluded from the BI-RADS 1–5 analysis-ready subset to preserve label consistency. Consequently, subsequent BI-RADS-based analyses were conducted on consistently structured records with extractable BI-RADS labels. After completion of the extraction and curation pipeline, each report was transformed into a structured representation that separates clinical narrative content from BI-RADS labels. As illustrated in Table 8, this process produced a structured and machine-readable report-level format for subsequent analysis. No report was excluded solely based on text length; exclusions were based on unresolved BI-RADS extraction, review flags, duplicate or near-duplicate status, and predefined task-specific clinical category criteria.

Table 8. Synthetic example of a structured radiology report after preprocessing, translated from Turkish.

ID	Clinical Narrative	BI-RADS Category
ID-001	Bilateral breast ultrasound findings describe a suspicious mass lesion in the right breast with irregular margins and associated suspicious axillary lymph node findings. The final diagnostic assessment indicates a high-suspicion BI-RADS category.	BI-RADS 5

3.5. Implementation Details and Reproducibility Settings

Implementation details and reproducibility settings were predefined and applied consistently to both datasets. The pipeline was implemented in Python using Pandas and NumPy for data handling, Python regular expressions for rule-based extraction, and *RapidFuzz* for fuzzy string matching. The fuzzy threshold for conclusion-section segmentation was set to 88, and the near-duplicate similarity threshold was set to 95 using *token_set_ratio*. Pairwise near-duplicate comparisons were restricted to records with the

same patient/file identifier and the same extracted BI-RADS label in order to reduce inappropriate comparisons between clinically unrelated reports. Text pairs exceeding 1200 characters were skipped during pairwise near-duplicate comparison to reduce computational overhead; however, no report was removed solely because of text length. Detailed configuration parameters, software versions, rule categories, platform information, and reproducibility settings are provided in Supplementary Table S1.

4. Results

4.1. Dataset Overview

Two real-world report-level datasets from different time periods, comprising 35,104 and 27,215 radiology reports, respectively, were used in this study. In Dataset 1, explicit BI-RADS categories were present in the original hospital-exported reports for 33,934 reports (96.67%), whereas 1170 reports (3.33%) did not contain an explicit BI-RADS label. In Dataset 2, explicit BI-RADS categories were present in all 27,215 original reports. These values refer to the original report-level BI-RADS distribution before the full curation pipeline was applied.

The distributions of routinely recorded BI-RADS categories extracted from the two datasets are presented in Table 3. In both datasets, BI-RADS 1 and BI-RADS 2 categories dominated the distribution, reflecting the class imbalance commonly observed in real-world breast imaging report archives. In Dataset 1, BI-RADS 2 was the most frequent category (48.52%), followed by BI-RADS 1 (27.87%) and BI-RADS 3 (10.69%). In Dataset 2, BI-RADS 2 also constituted the largest category (44.29%), followed by BI-RADS 1 (35.79%) and BI-RADS 0 (9.84%). Higher-risk categories, including BI-RADS 4 and BI-RADS 5, were relatively infrequent in both datasets. This distribution provides important context for interpreting the subsequent curation results and highlights the naturally imbalanced structure of real-world breast imaging report data.

4.2. Case Study Analysis

4.2.1. Comparative Dataset Characteristics

The transformation process resulted in structurally organized report-level datasets in which descriptive findings and assessment labels were distinguished. This structuring provided a more consistent representation of report narratives across both datasets. The word-frequency, word-cloud, and TF-IDF analyses in this subsection were performed on the original report text before explicit BI-RADS label removal and were intended only to describe source-report terminology. Therefore, the high frequency of the token “BI-RADS” in Table 9 and Figure 3 should not be interpreted as residual label leakage in the final label-separated analysis text.

Exploratory lexical analysis showed that the original report texts contained frequently occurring breast-imaging-related terminology. Word frequency analysis identified common terms such as “BI-RADS”, “breast”, and “tissue” in both datasets, reflecting terminology commonly used in routine breast imaging reports (Table 9). Dataset 1 was additionally characterized by the prominence of terms such as “lesion”, whereas Dataset 2 showed relatively higher frequencies for terms such as “type” and “skin”, indicating minor dataset-specific lexical differences. Overall, these findings suggest substantial consistency in frequently used breast-imaging terminology across both datasets.

4	Tissue (Doku)	82,185	Tissue (Doku)	67,879
5	Skin (Cilt)	79,841	Skin (Cilt)	65,057

Note. Turkish tokens are presented with English translations in parentheses where applicable. Token frequencies were calculated on the original report text before BI-RADS label removal and are presented only as an exploratory lexical summary.

4.2.2. BI-RADS Label Extraction Results Across Datasets

To examine the effect of preprocessing and BI-RADS label extraction across both datasets, comparisons between raw and curated BI-RADS distributions are provided separately for Dataset 1 and Dataset 2 in Tables 10 and 11. The curated distributions reflect pipeline-derived standardized labels obtained from routinely recorded BI-RADS expressions in the report text. They should therefore be interpreted as rule-based curation outputs rather than as independently corrected gold-standard labels. Quantitative evaluation results for the segmentation and BI-RADS label extraction processes are presented in Section 4.3.

The observed changes reflect the effect of rule-based label standardization and structural harmonization across both datasets. Ambiguous, incomplete, or inconsistently formatted BI-RADS expressions were assigned to standardized categories when sufficient textual evidence was available. To avoid ambiguity between main BI-RADS categories and explicitly stated BI-RADS 4 subcategories, BI-RADS labels were represented using a two-level scheme. The main BI-RADS category variable retained the standard categories from 0 to 6, while 4A, 4B, and 4C were stored as optional BI-RADS 4 subcategories when explicitly available in the report text. Accordingly, BI-RADS 4 subcategories were not treated as independent main classes beyond BI-RADS 4 but were reported separately for descriptive purposes.

In Dataset 1, the increase in the Unknown category reflects the conservative flagging of reports with unresolved, conflicting, or low-confidence BI-RADS expressions after normalization and extraction. In Dataset 2, an increase in BI-RADS 0 cases was observed after curation because conclusion-based extraction resulted in additional assignment of BI-RADS 0 when explicit incomplete-assessment expressions were identified by the rules. Minor decreases in BI-RADS 1–3 categories reflect the effect of standardizing uncertain or inconsistently represented labels during preprocessing and extraction.

Overall, the proposed pipeline maintained the broad class-imbalance pattern observed in the original reports while providing a more standardized report-level representation of BI-RADS categories across both datasets. The BI-RADS 4 subcategories shown in Tables 10 and 11 were identified only when explicitly stated in the report text and are reported as descriptive subcategory information under the main BI-RADS 4 category.

Table 10. Comparison of raw and pipeline-curated BI-RADS distributions in Dataset1.

BI-RADS Category	Raw Data (n)	Raw (%)	Curated Data (n)	Curated (%)
BI-RADS 0	1366	3.89	2195	6.25
BI-RADS 1	9785	27.87	9706	27.65
BI-RADS 2	17,031	48.52	16,052	45.73
BI-RADS 3	3751	10.69	3660	10.43
BI-RADS 4, total	608	1.73	562	1.60
BI-RADS 4, unspecified	257	0.73	278	0.79
BI-RADS 4A	168	0.48	145	0.41
BI-RADS 4B	78	0.22	42	0.12
BI-RADS 4C	105	0.30	97	0.28
BI-RADS 5	539	1.54	497	1.42
BI-RADS 6	854	2.43	691	1.97

Unknown	1170	3.33	1741	4.96
Total	35,104	100.0	35,104	100.0

Note. BI-RADS 4A, 4B, and 4C are subcategories of BI-RADS 4 and are reported descriptively when explicitly stated in the report text. They are not counted as independent main BI-RADS categories in the primary coarse-label scheme.

Table 11. Comparison of raw and pipeline-curated BI-RADS distributions in Dataset2.

BI-RADS Category	Raw Data (n)	Raw (%)	Curated Data (n)	Curated (%)
BI-RADS 0	2678	9.84	3117	11.45
BI-RADS 1	9741	35.79	9740	35.79
BI-RADS 2	12,053	44.29	11,634	42.75
BI-RADS 3	1827	6.71	1810	6.65
BI-RADS 4, total	340	1.25	337	1.24
BI-RADS 4, unspecified	125	0.46	163	0.60
BI-RADS 4A	128	0.47	118	0.43
BI-RADS 4B	48	0.18	17	0.06
BI-RADS 4C	39	0.14	39	0.14
BI-RADS 5	254	0.93	253	0.93
BI-RADS 6	322	1.18	324	1.19
Unknown	0	0.00	0	0.00
Total	27,215	100.0	27,215	100.0

Note. BI-RADS 4A, 4B, and 4C are subcategories of BI-RADS 4 and are reported descriptively when explicitly stated in the report text. They are not counted as independent main BI-RADS categories in the primary coarse-label scheme.

4.2.3. Task-Specific Dataset Filtering Results Across Datasets

The datasets underwent a data curation and filtering process to reduce category-level heterogeneity and support the construction of structured task-specific analysis-ready subsets for subsequent clinical NLP and machine learning analyses.

As part of this process, BI-RADS 0 and BI-RADS 6 categories were retained in the descriptive report-level curation archive but excluded from the task-specific BI-RADS 1–5 analysis-ready subset. BI-RADS labels were identified using the proposed regex-based extraction framework applied to free-text BI-RADS expressions recorded in the report text. The exclusion of BI-RADS 0 and BI-RADS 6 was performed only for the task-specific BI-RADS 1–5 subset in order to reduce category-level clinical heterogeneity. BI-RADS 0 reports typically represent incomplete assessments requiring additional imaging evaluation, whereas BI-RADS 6 reports correspond to biopsy-proven malignancies and may contain post-diagnostic, treatment-oriented, or follow-up-oriented language patterns that differ from primary diagnostic reporting.

Following this task-specific filtering stage, both datasets were restricted to BI-RADS categories 1–5. BI-RADS 4 entries without explicit subcategory information were retained as “BI-RADS 4, unspecified”, whereas explicitly reported 4A, 4B, and 4C labels were retained as descriptive subcategories under the main BI-RADS 4 category. The resulting BI-RADS class distributions after category-based filtering are presented in Table 12.

Table 12. BI-RADS distribution in the task-specific analysis-ready subset after category-based filtering (BI-RADS 1–5, with BI-RADS 4 subcategories reported separately when explicitly available).

BI-RADS Category	Dataset1 (n)	Dataset2 (n)
BI-RADS 1	9706	9740
BI-RADS 2	16,052	11,634
BI-RADS 3	3660	1810

BI-RADS 4, total	562	337
BI-RADS 4, unspecified	278	163
BI-RADS 4A	145	118
BI-RADS 4B	42	17
BI-RADS 4C	97	39
BI-RADS 5	497	253
Total	30,477	23,774

Note. BI-RADS 4A, 4B, and 4C are subcategories of BI-RADS 4 and are reported descriptively when explicitly stated in the report text. They are not counted as independent main BI-RADS categories in the primary coarse-label scheme.

4.2.4. Duplicate and Near-Duplicate Report Handling Across Datasets

To reduce report-level textual redundancy, a two-stage duplicate and near-duplicate handling strategy was applied. Prior to duplicate detection, light text normalization was performed, including whitespace standardization and removal of repetitive template-like expressions where applicable, in order to stabilize textual representations.

In the first stage, exact duplicate records were identified based on strict equality of the patient/file identifier, normalized report text, and extracted BI-RADS label. This step removed 6720 exact duplicate records from Dataset 1 and 9316 exact duplicate records from Dataset 2. In addition, 34 records in Dataset 1 and 22 records in Dataset 2 were removed because they had the same patient/file identifier and identical normalized report text.

In the second stage, potential near-duplicate records were identified within records belonging to the same patient/file identifier. For Dataset 1, 3758 patient/file identifiers containing multiple reports were screened for candidate near-duplicates, whereas the analysis covered 2194 such identifiers in Dataset 2. Textual similarity was computed using the *token_set_ratio* metric from the *RapidFuzz* library, and report pairs with similarity scores $\geq 95\%$ were considered candidate near-duplicates. Pairwise comparisons were restricted to reports sharing the same extracted BI-RADS label. Very long report pairs were skipped during pairwise comparison to avoid excessive computational overhead; however, no report was removed solely based on text length. Pairs with different BI-RADS labels were not compared for near-duplicate handling. Because duplicate and near-duplicate handling was the most computationally intensive component of the pipeline, its runtime was recorded separately. This step required approximately 0.2 min (~12 s) for each dataset under the reported configuration.

The duplicate and near-duplicate handling results are summarized in Table 13. Following this process, the redundancy-reduced BI-RADS 1–5 subset decreased from 30,477 to 19,793 records in Dataset 1 and from 23,774 to 12,597 records in Dataset 2. Candidate near-duplicate records selected for removal were logged with their retained paired reports, similarity scores, BI-RADS labels, and removal reasons. In addition, a random sample of 200 candidate near-duplicate pairs from each dataset was prepared for manual text-level inspection.

Table 13. Summary of duplicate and near-duplicate handling in the task-specific BI-RADS 1–5 subsets across datasets.

Indicator	Dataset1 (n)	Dataset2 (n)
Initial BI-RADS 1–5 records	30,477	23,774
Exact duplicate records removed	6720	9316
Same ID + same normalized text records removed	34	22
Patient/file identifiers screened for near-duplicates	3758	2194
Candidate near-duplicate records removed	3930	1839

Long-text pairs skipped during comparison	93	25
Pairs skipped due to different BI-RADS labels	3017	706
Final records after duplicate/near-duplicate handling	19,793	12,597
Manual inspection sample	200 pairs	200 pairs
Duplicate and near-duplicate handling runtime	0.2 min (~12 s)	0.2 min (~12 s)

Note. Near-duplicate removal represents a redundancy-reduced report-level subset based on textual similarity and should not be interpreted as confirmed clinical redundancy. Pairwise comparisons were restricted to reports with the same patient/file identifier and the same extracted BI-RADS label. Text pairs exceeding 1200 characters were skipped during pairwise comparison; however, no report was removed solely because of text length. Runtime values refer only to the duplicate and near-duplicate handling step. Near-duplicate detection was performed using *RapidFuzz* 'token_set_ratio' with a similarity threshold of 95% or higher.

Manual inspection was performed to evaluate the text-level appropriateness of the automated near-duplicate handling decisions. This validation did not aim to determine radiological diagnostic equivalence; instead, each retained–removed report pair was reviewed for textual redundancy, modality consistency, measurement or localization differences, comparison-date or follow-up-related wording, and the presence of additional findings. The manual review categories were not intended as mutually exclusive clinical annotations; rather, they were used to identify whether high textual similarity could still coexist with potentially information-bearing differences. The results of this manual text-level risk assessment are summarized in Table 14.

For Dataset 1, 200 candidate near-duplicate pairs were sampled, and after excluding duplicate entries within the manual inspection file, 192 pairs were evaluated. Of these, 108 pairs (56.3%) were judged to be textually duplicate or near-duplicate, while 172 pairs (89.6%) were considered potentially information-bearing or clinically distinct at the text-review level. The automated keep/drop decision was considered textually acceptable for 20 pairs (10.4%) and was flagged as requiring caution for 172 pairs (89.6%).

For Dataset 2, 200 candidate near-duplicate pairs were sampled, and 191 pairs remained after excluding duplicate entries. Among these, 113 pairs (59.2%) were judged to be textually duplicate or near-duplicate, whereas 164 pairs (85.9%) were considered potentially information-bearing or clinically distinct at the text-review level. The automated keep/drop decision was considered textually acceptable for 27 pairs (14.1%) and was flagged as requiring caution for 164 pairs (85.9%).

These findings indicate that although the token_set_ratio ≥ 95 threshold identified highly similar report pairs, high textual similarity alone should not be interpreted as clinical equivalence, particularly when modality structure, measurements, comparison dates, follow-up-related wording, additional findings, or localization details differ. Therefore, the near-duplicate handling step should be interpreted as report-level textual redundancy reduction rather than as a definitive assessment of clinical redundancy. The resulting BI-RADS 1–5 class distributions after duplicate and near-duplicate handling are presented in Table 15.

Although report date information was available for dataset description and temporal overlap control, temporal intervals were not incorporated as a primary criterion in the duplicate or near-duplicate handling decision. Therefore, highly similar reports from the same patient/file identifier but different dates may include longitudinal follow-up examinations. For this reason, the deduplication step should not be interpreted as a longitudinal clinical follow-up analysis.

Table 14. Manual text-level risk assessment of candidate near-duplicate pairs selected for removal.

Indicator	Dataset1	Dataset2
Initial manual inspection sample	200 pairs	200 pairs
Final evaluated pairs after excluding duplicate entries	192	191
Textually duplicate/near-duplicate, <i>n</i> (%)	108 (56.3%)	113 (59.2%)
Potentially information-bearing or clinically distinct, <i>n</i> (%)	172 (89.6%)	164 (85.9%)
Automated keep/drop decision acceptable, <i>n</i> (%)	20 (10.4%)	27 (14.1%)
Automated keep/drop decision requiring caution, <i>n</i> (%)	172 (89.6%)	164 (85.9%)

Note. Manual review categories were used for text-level risk assessment and were not intended to establish radiological diagnostic equivalence. Categories are not necessarily mutually exclusive because a report pair may be textually similar while still containing potentially information-bearing differences.

Table 15. BI-RADS distribution in the redundancy-reduced BI-RADS 1–5 subset after category-based filtering and duplicate/near-duplicate handling.

BI-RADS Category	Dataset1 (<i>n</i>)	Dataset2 (<i>n</i>)
BI-RADS 1	8413	5446
BI-RADS 2	7944	5417
BI-RADS 3	2613	1336
BI-RADS 4, total	451	239
BI-RADS 4, unspecified	215	122
BI-RADS 4A	117	81
BI-RADS 4B	36	12
BI-RADS 4C	83	24
BI-RADS 5	372	159
Total	19,793	12,597

Note. BI-RADS 4A, 4B, and 4C are subcategories of BI-RADS 4 and are reported descriptively when explicitly stated in the report text. They are not counted as independent main BI-RADS classes beyond the BI-RADS 4 total.

4.2.5. Structural and Modality-Based Curation Results Across Datasets

To provide an integrated descriptive summary of the proposed data curation framework, the overall quantitative and structural effects of the pipeline were analysed across both datasets. Table 16 summarizes the dataset characteristics before and after task-specific curation steps, including total report counts, included BI-RADS categories, modality composition, ambiguity rates, and modality-specific report distributions.

Across both datasets, the curation pipeline reduced report-level textual redundancy and produced more structurally organized report-level datasets. Following category-based filtering and duplicate and near-duplicate handling, the total number of reports decreased from 35,104 to 19,793 in Dataset 1 and from 27,215 to 12,597 in Dataset 2, corresponding to overall reduction rates of 43.6% and 53.7%, respectively. These reductions reflect the combined effect of task-specific BI-RADS 1–5 filtering and redundancy reduction rather than exclusion from the original descriptive report-level archive.

The curation process also changed the modality-level composition of the task-specific subsets. After filtering and duplicate/near-duplicate handling, the proportion of single-modality reports increased from 62.5% to 82.7% in Dataset 1 and from 66.8% to 85.0% in Dataset 2, whereas the proportion of multi-modality reports decreased from 37.5% to 17.3% and from 33.2% to 15.0%, respectively. Similarly, rule-derived ambiguous modality assignments decreased from 20.0% to 15.5% in Dataset 1 and from 15.7% to 13.4% in Dataset 2. Reports with unclassified modality remained negligible after curation in both datasets.

In terms of modality-specific distributions, ultrasonography (US) and mammography (MG) remained the dominant imaging modalities after filtering and duplicate/near-duplicate handling. In Dataset 1, the BI-RADS 1–5 analysis-ready subset included 11,973 US reports, 7594 MG reports, and 219 MRI reports. In Dataset 2, the corresponding counts were 7557 US reports, 5029 MG reports, and 6 MRI reports. The limited number of MRI reports, particularly in Dataset 2, indicates that MRI-specific findings should be interpreted cautiously.

Overall, these findings show that the proposed pipeline reduced report-level textual redundancy, decreased rule-derived modality ambiguity, and generated modality-oriented structured report datasets for subsequent clinical NLP and structured information extraction tasks. These results should be interpreted as structural and descriptive effects of the curation process rather than as evidence of independent clinical accuracy or externally validated modality-classification results.

Table 16. Overall quantitative and structural changes after task-specific category-based filtering and duplicate/near-duplicate handling across Dataset1 and Dataset2.

Indicator	Dataset1 Before	Dataset1 After	Dataset2 Before	Dataset2 After
Total number of reports	35,104	19,793	27,215	12,597
Reduction rate (%)	N/A	43.6%	N/A	53.7%
BI-RADS categories included	0–6	1–5	0–6	1–5
Single-modality reports (%)	62.5 (<i>n</i> = 21,944)	82.7 (<i>n</i> = 16,362)	66.8 (<i>n</i> = 18,193)	85.0 (<i>n</i> = 10,704)
Multi-modality reports (%)	37.5 (<i>n</i> = 13,160)	17.3 (<i>n</i> = 3431)	33.2 (<i>n</i> = 9022)	15.0 (<i>n</i> = 1893)
Ambiguous modality (%)	20.0 (<i>n</i> = 7010)	15.5 (<i>n</i> = 3077)	15.7 (<i>n</i> = 4268)	13.4 (<i>n</i> = 1688)
Unclassified modality (%)	0.3 (<i>n</i> = 113)	<0.1 (<i>n</i> = 7)	<0.1 (<i>n</i> = 7)	<0.1 (<i>n</i> = 5)
US reports (<i>n</i>)	24,034	11,973	14,547	7557
MG reports (<i>n</i>)	10,032	7594	12,627	5029
MRI reports (<i>n</i>)	925	219	34	6

Note. Values are presented as percentages with counts in parentheses where applicable. The “after” columns represent the BI-RADS 1–5 analysis-ready subsets after category-based filtering and duplicate/near-duplicate handling. These values describe structural changes in the curated subsets and should not be interpreted as independent clinical validation metrics.

4.3. Evaluation of the Data Curation Pipeline

4.3.1. Conclusion Section Segmentation Coverage and Internal Validation

Conclusion-section segmentation was evaluated by quantifying whether the proposed rule-guided procedure could detect a conclusion marker and split each report into a clinical narrative component and a conclusion segment. These values should be interpreted as segmentation coverage rather than independent accuracy estimates, because exhaustive manual annotation of conclusion boundaries for the full datasets was not available.

In Dataset 1, the conclusion marker was detected, and the report was split into 33,949 reports, including 33,948 reports detected by dynamic regex and one additional report resolved by fuzzy matching. This corresponded to a conclusion-section segmentation coverage of 96.71%. A total of 1155 reports (3.29%) remained unmatched, mainly because of missing or structurally inconsistent conclusion markers in the source texts. In Dataset 2, the conclusion marker was detected in 27,200 reports, including 27,199 regex-based detections and one fuzzy-matching fallback case, corresponding to a segmentation coverage of 99.94%. Only 15 reports (0.06%) remained unmatched in Dataset 2. Detailed segmentation coverage statistics for both datasets are presented in Table 17. These findings indicate that most reports contained recognizable conclusion-section patterns that could be captured

by the proposed rule-guided segmentation strategy. The higher coverage observed in Dataset 2 may reflect greater consistency in reporting structure during the later period.

To assess the internal consistency of the segmentation outputs, manual validation was conducted on enlarged sample subsets from both datasets. The validation samples were stratified according to segmentation outcome and were reviewed after duplicate entries were removed from the validation files. For Dataset 1, 290 non-duplicate records were reviewed, including 259 regex-based cases, one fuzzy-matching case, and 30 no-match cases. Manual review indicated that all 290 records were appropriately segmented or appropriately classified according to the system output, corresponding to 100.0% agreement within the reviewed sample. For Dataset 2, 294 non-duplicate records were reviewed, including 284 regex-based cases, one fuzzy-matching case, and 9 no-match cases. Manual review confirmed appropriate segmentation for 285 records, including all regex-based and fuzzy-matching cases, whereas 9 records were judged as incorrect; all incorrect cases belonged to the no-match group. These incorrect no-match cases indicate that some conclusion-like boundaries may remain undetected under heterogeneous reporting conditions. This corresponded to 96.94% agreement within the reviewed Dataset 2 sample.

To examine the effect of the fuzzy matching threshold on conclusion-section segmentation, a sensitivity analysis was performed using threshold values of 80, 85, 88, 90, and 95. As shown in Table 18, regex-based detections remained unchanged across all threshold values in both datasets. Fuzzy fallback detections were rare, involving only one report in each dataset at the selected threshold of 88. Increasing the threshold to 95 in Dataset 1 and to 90 or higher in Dataset 2 changed only one report from fuzzy fallback to no match. These results indicate that the overall segmentation coverage was driven primarily by explicit regex-detected conclusion markers and was minimally affected by the selected fuzzy matching threshold.

Together, the enlarged manual validation samples and threshold sensitivity analysis support the internal consistency of the conclusion-section segmentation procedure in the reviewed records. However, because manual validation was conducted on selected samples rather than on a fully annotated independent gold-standard corpus, these findings should be interpreted as evidence of internal consistency in the reviewed records, not as dataset-wide segmentation coverage.

Table 17. Conclusion-section segmentation coverage outcomes of the proposed hybrid strategy across the two datasets.

Splitting Method	Dataset1 (n, %)	Dataset2 (n, %)
Dynamic Regex Detection	33,948 (96.71)	27,199 (99.94)
Fuzzy Matching (Fallback)	1 (<0.01)	1 (<0.01)
No Match	1155 (3.29)	15 (0.06)
Total	35,104 (100.0)	27,215 (100.0)

Table 18. Sensitivity analysis of fuzzy matching thresholds for conclusion-section segmentation.

Dataset	Fuzzy Threshold	Regex Detection n (%)	Fuzzy Fallback n (%)	No Match n (%)	Total
Dataset 1	80	33,948 (96.71)	1 (<0.01)	1155 (3.29)	35,104
Dataset 1	85	33,948 (96.71)	1 (<0.01)	1155 (3.29)	35,104
Dataset 1	88	33,948 (96.71)	1 (<0.01)	1155 (3.29)	35,104
Dataset 1	90	33,948 (96.71)	1 (<0.01)	1155 (3.29)	35,104
Dataset 1	95	33,948 (96.71)	0 (0.00)	1156 (3.29)	35,104
Dataset 2	80	27,199 (99.94)	1 (<0.01)	15 (0.06)	27,215
Dataset 2	85	27,199 (99.94)	1 (<0.01)	15 (0.06)	27,215
Dataset 2	88	27,199 (99.94)	1 (<0.01)	15 (0.06)	27,215
Dataset 2	90	27,199 (99.94)	0 (0.00)	16 (0.06)	27,215

Dataset 2	95	27,199 (99.94)	0 (0.00)	16 (0.06)	27,215
-----------	----	----------------	----------	-----------	--------

4.3.2. BI-RADS Label Extraction Coverage and Internal Validation

The BI-RADS label extraction module was quantitatively evaluated in terms of extraction coverage across both datasets. These values indicate whether a BI-RADS expression could be detected and standardized by the proposed rule-guided framework; therefore, they should not be interpreted as independent label extraction consistency estimates in the absence of a fully annotated external gold-standard corpus.

In Dataset 1, the proposed multi-phase extraction framework identified BI-RADS expressions in 33,347 reports through dynamic regex detection, corresponding to a direct extraction coverage of 94.99%. An additional 16 reports (0.05%) were resolved through auxiliary spaced-pattern matching, while 1741 reports (4.96%) remained unmatched because of missing labels, incomplete documentation, conflicting expressions, low-confidence cases, or irregular notation. This curation-derived count differs from the original BI-RADS distribution reported in Section 4.1 because Section 4.1 describes the hospital-exported raw report distribution, whereas the present section reports the output of the standardized extraction and curation procedure. During curation, reports with missing, conflicting, incomplete, or low-confidence BI-RADS expressions were assigned to the Unknown/no-match group, even when a BI-RADS-like expression had been present in the original report distribution.

In Dataset 2, explicit BI-RADS expressions were extracted and standardized from all 27,215 reports using the dynamic regex layer, resulting in complete extractable BI-RADS coverage for this dataset. Detailed BI-RADS label extraction coverage statistics are presented in Table 19.

These findings indicate that the primary dynamic regex layer captured most extractable BI-RADS expressions in the two retrospective report datasets, while auxiliary extraction rules contributed only a small number of additional matches. The complete extraction coverage observed in Dataset 2 may reflect greater consistency in BI-RADS documentation during the later reporting period.

To assess the internal consistency of the BI-RADS extraction outputs, manual validation was conducted on enlarged sampled subsets after removing duplicate entries from the validation files. For Dataset 1, 290 non-duplicate records were reviewed, including 237 dynamic-regex cases, 14 auxiliary spaced-pattern cases, and 39 no-match cases. Manual review agreed with the BI-RADS extraction outcome in 276 of 290 records, corresponding to 95.17% agreement within the reviewed Dataset 1 sample. A total of 14 records were judged incorrect; among these, 8 belonged to the dynamic-regex group, and 6 belonged to the no-match group.

For Dataset 2, 294 non-duplicate records were reviewed. All records in this validation sample belonged to the dynamic-regex extraction group. Manual review agreed with the BI-RADS extraction outcome in 289 of 294 records, corresponding to 98.30% agreement within the reviewed Dataset 2 sample. The remaining 5 incorrect records also belonged to the dynamic-regex group.

Together, these manual validation findings support the internal consistency of the BI-RADS extraction procedure in the reviewed samples. However, because validation was performed on selected samples rather than on an independently annotated full gold-standard corpus, these findings should be interpreted as internal validation results rather than dataset-wide BI-RADS extraction accuracy estimates.

As an additional post-curation quality-control step, the curated clinical narrative text field was screened for residual BI-RADS-related expressions to assess the risk of direct label leakage before downstream use. This analysis was performed after structured BI-RADS labels had been extracted and assigned to the label field. Regular-expression pat-

terns were used to detect residual explicit label expressions, including common variants such as “BI-RADS”, “BIRADS”, “BI RADS”, fragmented forms, and category-attached expressions such as “BI-RADS 4A”. In Dataset 1, this check was applied to 33,363 reports with extractable BI-RADS labels after excluding the unmatched/Unknown cases. Residual BI-RADS-related expressions were detected in 42 reports, corresponding to 0.13% of the analyzed Dataset 1 subset. In Dataset 2, the analysis was applied to all 27,215 reports, since all reports had extractable BI-RADS labels. Residual BI-RADS-related expressions were detected in 25 reports, corresponding to 0.09% of Dataset 2. These cases were flagged for manual inspection and correction before downstream use. This leakage-control step does not replace independent gold-standard validation; rather, it provides an additional internal safeguard by showing that explicit BI-RADS category mentions were rarely retained in the clinical input text.

Table 19. Quantitative BI-RADS label extraction coverage outcomes of the proposed layered framework across the two datasets.

Extraction Method	Dataset1 (n, %)	Dataset2 (n, %)
Dynamic Regex Detection	33,347 (94.99)	27,215 (100.0)
Auxiliary (Spaced Pattern)	16 (0.05)	0
No Match	1741 (4.96)	0
Total	35,104 (100.0)	27,215 (100.0)

4.3.3. Modality-Based Separation Results and Internal Validation

After completion of the data curation pipeline, imaging modality distributions were analysed to describe the structural composition of the curated datasets and to assess the internal consistency of the rule-derived modality-based separation procedure. In Dataset 1, the curated dataset was predominantly composed of ultrasonography (US) and mammography (MG) reports, accounting for 60.5% and 38.4% of all records, respectively. Magnetic resonance imaging (MRI) reports represented a relatively small proportion (1.1%), whereas unidentified or unclassified cases were negligible (<0.1%).

A comparable distribution was observed in Dataset 2, where US and MG again remained the dominant modalities, representing 60.0% and 39.9% of reports, respectively. MRI and unclassified cases likewise constituted only a minimal fraction of the dataset. Report structure was further evaluated in terms of modality composition. In Dataset 1, 82.7% of reports corresponded to single-modality narratives, whereas 17.3% contained references to multiple imaging modalities. Similarly, Dataset 2 consisted of 85.0% single-modality and 15.0% multi-modality reports.

The certainty of rule-derived primary modality assignment was also examined as an indicator of modality-assignment consistency. In Dataset 1, 84.5% of reports were assigned a clear primary modality, while 15.5% were flagged as ambiguous. In Dataset 2, 86.6% of reports were assigned a clear primary modality, whereas 13.4% were flagged as ambiguous. Detailed modality distribution and structural statistics for both datasets are presented in Tables 20 and 21.

To further assess the internal consistency of modality-based separation, manual validation was conducted on sampled records from both datasets. This validation focused on whether modality-related textual cues and report-level modality assignments were consistent with the original report text. It did not involve clinical reinterpretation of imaging findings or reassignment of diagnostic labels.

In Dataset 1, 238 reports were manually reviewed. Among 159 reports marked as single modality, the assigned primary modality was consistent with manual review in 154 cases (96.86%). Among 79 reports marked as multi-modality, manual inspection agreed with the multi-modality assignment in 74 cases (93.67%). Overall, 228 of 238 manually

reviewed reports (95.80%) were consistent with the expected report-level modality assignment.

For Dataset 2, manual validation was repeated on the updated final curated dataset. A total of 255 sampled records were reviewed, including 125 records marked as single modality and 130 records marked as multi-modality. Among the 125 records marked as single modality, all were consistent with single-modality status based on manual inspection of the report text. The *PrimaryModality* assignment was consistent with the modality-related textual structure in 122 of these records, whereas 3 records were assigned as Unknown despite containing ultrasonography-related textual content.

Among the 130 records marked as multi-modality in Dataset 2, 127 were consistent with the presence of more than one performed imaging modality or modality-specific report section. The remaining 3 records represented false-positive multi-modality assignments. These mismatches were mainly caused by recommendation or reference phrases, such as “USG önerilir,” “MRG’de tarifi lenen,” or “MR ile değerlendirilebilir,” which were interpreted by the rule-based procedure as evidence of an additional modality, although they did not represent a separately performed imaging examination. Overall, the single/multi-modality assignment was consistent with manual inspection in 252 of 255 reviewed Dataset 2 records, corresponding to 98.82% agreement within the reviewed sample.

These findings indicate that the proposed modality-based separation procedure produced internally consistent report-level modality patterns across the reviewed samples. The predominance of US and MG reports, together with the high proportion of single-modality and clearly assigned records, shows that the curated datasets were organized into modality-oriented report-level subsets. However, these results should be interpreted as internal validation findings based on reviewed samples rather than as dataset-wide independent accuracy estimates for modality classification.

Table 20. Rule-derived imaging modality distribution after data curation.

Modality	Dataset1 (n, %)	Dataset2 (n, %)
Ultrasonography (US)	11,973 (60.5)	7557 (60.0)
Mammography (MG)	7594 (38.4)	5029 (39.9)
Magnetic resonance imaging (MRI)	219 (1.1)	6 (<0.1)
Unidentified modality	7 (<0.1)	5 (<0.1)
Total	19,793 (100.0)	12,597 (100.0)

Note. Modality assignments are rule-derived curation outputs and should not be interpreted as independently annotated modality labels.

Table 21. Rule-derived distribution of report types after data curation: (a) modality composition and (b) modality certainty.

(a) Modality Composition		
Report Type	Dataset1 (n, %)	Dataset2 (n, %)
Single modality	16,362 (82.66)	10,704 (85.0)
Multi-modality	3431 (17.34)	1893 (15.0)
Total	19,793 (100.0)	12,597 (100.0)
(b) Modality Certainty		
Report Type	Dataset1 (n, %)	Dataset2 (n, %)
Certain	16,716 (84.45)	10,909 (86.6)
Uncertain	3077 (15.55)	1688 (13.4)
Total	19,793 (100.0)	12,597 (100.0)

Note. Modality assignments are rule-derived curation outputs and should not be interpreted as independently annotated modality labels.

5. Discussion

The data curation procedures applied in this study resulted in substantial quantitative and structural changes in the two report-level datasets. As shown in Tables 10 and 11, the preprocessing and BI-RADS extraction steps changed the distribution of several categories by standardizing inconsistently formatted, incomplete, or previously unresolved BI-RADS expressions. In Dataset1, the number of BI-RADS 0 cases increased from 1366 to 2195, while the Unknown category increased from 1170 to 1741, reflecting the conservative handling of unresolved, conflicting, or low-confidence reports after normalization and extraction. In Dataset2, BI-RADS 0 increased from 2680 to 3119, indicating that conclusion-based extraction resulted in additional assignment of incomplete assessment statements when explicit BI-RADS 0 expressions were identified by the rules.

The extraction of BI-RADS 4A, 4B, and 4C information after curation further illustrates the value of retaining explicitly stated subcategory information. These subcategories were not treated as independent main BI-RADS classes; rather, they were retained as optional descriptive subcategory attributes under the main BI-RADS 4 category when explicitly available in the report text. This two-level representation allowed the curated outputs to preserve additional descriptive BI-RADS 4 information while maintaining a consistent primary coarse-label scheme.

In addition to the quantitative distributional changes presented in Tables 10 and 11, supplementary lexical analysis showed that clinically relevant terms followed interpretable patterns across BI-RADS categories. Higher-risk categories were associated with suspicious and intervention-related terminology, whereas lower-risk categories were dominated by benign or general descriptive expressions. Because the word-frequency, word-cloud, and TF-IDF analyses were performed on the original report text before explicit BI-RADS label removal, these findings should be interpreted as exploratory lexical characterization rather than as evidence of downstream classification results or independent validation. Nevertheless, the observed lexical patterns suggest that the source report narratives contained clinically meaningful risk-related terminology (Figure S1).

After task-specific exclusion of BI-RADS 0 and BI-RADS 6, duplicate and near-duplicate handling, and modality-based separation, the resulting BI-RADS 1–5 analysis-ready subsets became more structurally organized for subsequent analysis. The reduction in dataset size should be interpreted as the combined effect of category-based filtering and report-level textual redundancy reduction, rather than as a definitive removal of clinically redundant information. This distinction is important because high textual similarity does not necessarily imply clinical equivalence, particularly in longitudinal breast imaging records where follow-up examinations may contain similar template structures but clinically relevant differences.

The overall quantitative and structural changes associated with the proposed curation pipeline are summarized in Table 16. Across both datasets, substantial reductions in dataset size were observed after task-specific filtering and duplicate/near-duplicate handling, with reduction rates of 43.6% for Dataset1 and 53.7% for Dataset2. At the same time, rule-derived structural clarity increased. In Dataset1, the proportion of single-modality reports increased from 62.5% to 82.7%, while multi-modality reports decreased from 37.5% to 17.3%. Similarly, Dataset2 showed an increase in single-modality reports from 66.8% to 85.0%, accompanied by a reduction in multi-modality reports from 33.2% to 15.0%. A comparable change was observed in rule-derived modality certainty. Ambiguous modality assignments decreased from 20.0% to 15.5% in Dataset1 and from 15.7% to 13.4% in Dataset2. These findings indicate that the proposed modality-based separation framework improved the structural organization of heterogeneous report narratives; however, they should not be interpreted as independent modality-classification accuracy.

After modality-based separation, ultrasonography (US) reports constituted the largest subset in both datasets, followed by mammography (MG), as shown in Table 20. This pattern may reflect the real-world composition of the hospital report archive and routine breast imaging workflows, where US and MG are more frequently represented in screening, follow-up, and diagnostic assessment. The relatively small MRI subset, particularly in Dataset #2, limits the interpretability of MRI-specific findings and should be considered when using the curated outputs for modality-specific downstream analyses.

5.1. Clinical and Research Implications

The proposed framework has several important implications for clinical research and healthcare analytics. First, structured report-level datasets can support large-scale retrospective analysis of radiology reports and may reduce the need for fully manual chart review in preliminary data preparation. Such resources can support epidemiological studies, institutional audits, quality monitoring, and real-world evidence generation, provided that their limitations and curation assumptions are taken into account.

Second, structured radiology corpora can provide a useful foundation for downstream computational applications, including BI-RADS classification, report recommendation systems, longitudinal follow-up research, and future clinical decision-support studies. By separating narrative findings from diagnostic assessment labels and reducing report-level textual redundancy, the proposed framework creates a more consistent data environment for subsequent NLP and machine learning studies. However, downstream use should consider that the labels are pipeline-extracted from routine reports and were not validated against an independently annotated full gold-standard corpus.

Third, the proposed framework is particularly relevant for low-resource languages such as Turkish, where publicly available annotated clinical corpora remain limited. By transforming heterogeneous free-text reports into structured report-level datasets, the framework helps reduce a major barrier to clinical NLP research in underrepresented languages and supports the development of adaptable language-specific medical NLP resources.

5.2. Limitations

Despite its strengths, the proposed approach has several limitations. First, the study was conducted using data derived from a single clinical institution, and reporting conventions may differ across hospitals, radiology departments, or radiology information systems. External validation on multi-center datasets obtained from different institutions would strengthen the transferability of the proposed framework.

Second, the datasets were available as retrospective report-level data containing only a patient/file identifier, report date, and free-text report. Patient-level demographic variables such as age and sex, lesion-level identifiers, clinical outcomes, and detailed longitudinal metadata were not available. Therefore, demographic representativeness, lesion-level continuity, and longitudinal clinical trajectories could not be evaluated.

Third, the BI-RADS labels used in this study were extracted from routinely generated clinical reports and were not independently assigned by the research team. Although manual review of selected samples supported the internal consistency of the extraction outputs, a fully annotated radiologist-reviewed gold-standard corpus was not available. Therefore, the reported extraction coverage and internal validation results should not be interpreted as dataset-wide precision, recall, F1-score, or independent label accuracy. In addition, the manual validation samples represented selected subsets rather than exhaustive annotation of the full datasets; therefore, they should be interpreted as internal consistency checks rather than definitive dataset-wide validation metrics.

Fourth, although the rule-guided extraction and cleaning procedures achieved high coverage in the present datasets, highly heterogeneous narratives, uncommon reporting styles, or previously unseen terminology may still introduce segmentation or label extraction ambiguities. Some design choices, including fuzzy matching thresholds, similarity thresholds for near-duplicate detection, and heuristic rules for modality assignments, were empirically selected and may require recalibration when transferred to other institutions or other low-resource languages.

Fifth, duplicate and near-duplicate handling was performed primarily as a report-level textual redundancy reduction step. High textual similarity should not be interpreted as clinical equivalence, especially in longitudinal breast imaging data where follow-up reports may share template-like wording while still containing clinically meaningful differences. Because lesion-level identifiers and detailed longitudinal clinical metadata were not available, the near-duplicate handling step should not be interpreted as a validated clinical redundancy assessment. In addition, independent radiologist adjudication and inter-annotator agreement for duplicate/non-duplicate status were not available.

Sixth, this study did not include a formal comparison against a simple keyword-based baseline or a machine-learning extraction baseline. The primary aim was to develop and internally evaluate a transparent data curation framework rather than to benchmark competing extraction models. Future studies should compare rule-guided, keyword-based, and data-driven extraction strategies to quantify the incremental contribution of each curation component.

Finally, although runtime was recorded for the duplicate and near-duplicate handling step, the multi-stage pipeline may still require additional computational optimization and more detailed runtime and memory profiling before deployment in substantially larger institutional repositories.

Overall, the findings suggest that transparent data curation is not merely a preprocessing step but an important foundation for clinical NLP research. Structured report-level datasets can support reproducible model development, healthcare analytics, and future real-world evidence studies when their curation assumptions, coverage-based evaluation design, and validation limitations are explicitly considered.

6. Conclusions

This study presented a report-level data curation framework for transforming unstructured Turkish breast imaging reports into structured, machine-readable, and analytically usable clinical datasets, with a specific focus on BI-RADS label extraction from routinely generated report text. By integrating domain-aware preprocessing, terminology normalization, hybrid conclusion-section segmentation, BI-RADS label extraction, duplicate and near-duplicate report handling, and modality-based stratification, the proposed framework addresses key challenges associated with heterogeneous real-world clinical text.

The stepwise reduction in dataset size reflected the cumulative effect of task-specific category-based filtering and report-level textual redundancy reduction. In Dataset1, the initial overlap-controlled archive contained 35,104 reports. After excluding BI-RADS 0, BI-RADS 6, and unresolved or Unknown records, the BI-RADS 1–5 analysis-ready subset consisted of 30,477 reports. Following duplicate and near-duplicate handling, the redundancy-reduced BI-RADS 1–5 subset included 19,793 reports. Similarly, Dataset2 contained 27,215 reports after temporal overlap and consistency control, was reduced to 23,774 reports after category-based filtering, and included 12,597 reports after duplicate and near-duplicate handling. These reductions should be interpreted as task-specific curation and redundancy-reduction effects rather than as definitive removal of clinically redundant information.

The results show that large collections of free-text radiology reports can be systematically transformed into structured report-level datasets without requiring changes to routine report-writing practices or the clinical reporting interface. Across both temporal datasets, the pipeline improved structural organization, supported standardized extraction of routinely recorded BI-RADS expressions, reduced report-level textual redundancy, and enabled modality-oriented dataset organization. Manual validation of selected samples further supported the internal consistency of the conclusion segmentation, BI-RADS extraction, and modality-based separation procedures.

Importantly, the study demonstrates that clinically meaningful information embedded in narrative radiology reports can be organized through transparent rule-guided curation strategies. This is particularly relevant for low-resource languages such as Turkish, where curated clinical text resources remain limited.

Beyond its methodological contribution, the proposed framework has practical implications for retrospective clinical research, radiology quality assessment, healthcare analytics, and the development of downstream clinical NLP applications, including BI-RADS classification models and longitudinal report analysis. However, downstream use should consider that the reported values represent extraction coverage, internal consistency, and structural curation outputs rather than independent gold-standard accuracy.

Overall, the transition from free-text radiology narratives to reusable structured report-level data is feasible within an institution-specific retrospective data setting. The proposed framework provides a transparent and adaptable methodological basis for data-driven radiology research and future clinical NLP studies.

7. Future Directions

Several promising research directions emerge from this study. First, external validation across multiple institutions is needed to assess the transferability of the proposed framework under different reporting conventions, radiology information systems, and regional documentation practices.

Second, future studies should develop independently annotated radiologist-reviewed reference corpora, including conclusion boundaries, BI-RADS labels, modality assignments, laterality, and clinically relevant findings. Such resources would allow conventional precision, recall, F1-score, and inter-annotator agreement analyses and would strengthen validation beyond coverage-based and internal consistency assessments.

Third, future studies may link curated radiology report data with corresponding imaging data to support multimodal curation strategies, richer patient-level representations, and more comprehensive clinical modelling. Such integration may enable joint analysis of textual findings, BI-RADS labels, and imaging features.

Fourth, machine learning and large language model approaches may be incorporated to complement rule-guided pipelines. These methods may support more flexible information extraction, detection of inconsistent or ambiguous labels, and reduction of manual review burden, particularly when combined with transparent rule-based validation mechanisms.

Finally, adaptive knowledge integration mechanisms capable of periodically updating terminology, reporting standards, and clinical guideline references may improve the long-term sustainability of automated clinical NLP systems.

Overall, these directions may further improve the transferability, adaptability, and practical utility of data curation frameworks for radiology and healthcare text analytics.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16147189/s1>, Figure S1: TF-IDF-weighted and log-scaled

distributions of clinically relevant terms across BI-RADS categories; Table S1: Main implementation parameters and reproducibility settings of the data curation pipeline.

Author Contributions: All authors contributed to the study conception, design, methodology, data interpretation, and writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Necmettin Erbakan University, Meram Faculty of Medicine (Approval No. 13549, Date: 7 April 2022).

Informed Consent Statement: Patient consent was waived due to the retrospective nature of the study and the use of clinical data.

Data Availability Statement: The data used in this study are not publicly available due to privacy and ethical restrictions. The data may be made available from the corresponding author upon reasonable request and with permission from the relevant institution.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pons, E.; Braun, L.M.; Hunink, M.M.; Kors, J.A. Natural language processing in radiology: A systematic review. *Radiology* **2016**, *279*, 329–343. <https://doi.org/10.1148/radiol.16142770>.
2. Kahn, C.E.; Langlotz, C.P.; Burnside, E.S.; Carrino, J.A.; Channin, D.S.; Hovsepian, D.M.; Rubin, D.L. Toward best practices in radiology reporting. *Radiology* **2009**, *252*, 852–856. <https://doi.org/10.1148/radiol.2523081992>.
3. Kohli, M.; Morrison, J.J.; Wawira, J.; Morgan, M.B.; Hostetter, J.; Genereaux, B.; Hussain, M.; Langer, S.G. Creation and Curation of the Society of Imaging Informatics in Medicine Hackathon Dataset. *J. Digit. Imaging* **2018**, *31*, 9–12. <https://doi.org/10.1007/s10278-017-0003-5>.
4. Marcus, D.S.; Olsen, T.R.; Ramaratnam, M.; Buckner, R.L. The extensible neuroimaging archive toolkit (XNAT). *Neuroinformatics* **2007**, *5*, 11–34. <https://doi.org/10.1385/NI:5:1:11>.
5. Dilmac, F.; Alpkocak, A. De-identification of electronic health records using deep learning and transformers. *Appl. Sci.* **2026**, *16*, 1692. <https://doi.org/10.3390/app16041692>.
6. Eser, M.; Bilgin, M. Irony and sarcasm detection in Turkish texts: A comparative study of transformer-based models and ensemble learning. *Appl. Sci.* **2025**, *15*, 12498. <https://doi.org/10.3390/app152312498>.
7. Bodenreider, O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270. <https://doi.org/10.1093/nar/gkh061>.
8. Boroumandzadeh, M.; Parvinnia, E. Automated classification of BI-RADS in textual mammography reports. *Turk. J. Electr. Eng. Comput. Sci.* **2021**, *29*, 632–647. <https://doi.org/10.3906/elk-2002-31>.
9. Bozkurt, S.; Alkim, E.; Banerjee, I.; Rubin, D.L. Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. *J. Digit. Imaging* **2019**, *32*, 544–553. <https://doi.org/10.1007/s10278-019-00237-9>.
10. Fanni, S.C.; Romei, C.; Ferrando, G.; Volpi, F.; D’Amore, C.A.; Bedini, C.; Ubbiali, S.; Valentino, S.; Neri, E. Natural language processing to convert unstructured COVID-19 chest CT reports into structured reports. *Eur. J. Radiol. Open* **2023**, *11*, 100512. <https://doi.org/10.1016/j.ejro.2023.100512>.
11. Nguyen, E.; Theodorakopoulos, D.; Pathak, S.; Geerdink, J.; Vijlbrief, O.; Van Keulen, M.; Seifert, C. A hybrid text classification and language generation model for automated summarization of Dutch breast cancer radiology reports. In Proceedings of the 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI), Atlanta, GA, USA, 1–3 December 2020; pp. 72–81. <https://doi.org/10.1109/CogMI50398.2020.00019>.
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.

13. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model. *Bioinformatics* **2020**, *36*, 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>.
14. Yıldırım, S.; Ülker, E.; Poyraz, N. Natural language processing in breast imaging reports: A scoping review with implications for low-resource clinical languages. *Appl. Sci.* **2026**, *16*, 5847. <https://doi.org/10.3390/app16125847>.
15. Hepsağ, P.U.; Özel, S.A.; Dalcı, K.; Yazıcı, A. Using BERT models for breast cancer diagnosis from Turkish radiology reports. *Lang. Resour. Eval.* **2024**, *58*, 981–1012. <https://doi.org/10.1007/s10579-023-09669-w>.
16. López-Úbeda, P.; Martín-Noguerol, T.; Luna, A. Automatic classification and prioritisation of actionable BI-RADS categories. *Clin. Radiol.* **2024**, *79*, e1–e7. <https://doi.org/10.1016/j.crad.2023.09.009>.
17. Hussain, S.; Naseem, U.; Ali, M.; Avalos, D.B.A.; Cardona-Huerta, S.; Bosques Palomo, B.A.; Tamez-Peña, J.G. TECRR: A benchmark dataset of radiological reports for BI-RADS classification. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 310. <https://doi.org/10.1186/s12911-024-02717-7>.
18. Çelikten, A.; Bulut, H. Turkish medical text classification using BERT. In Proceedings of the Signal Processing and Communications Applications Conference (SIU), Istanbul, Turkey, 9–11 June 2021; pp. 1–4.
19. Reichenpfader, D.; Müller, H.; Denecke, K. LLM-based approaches for radiology report information extraction: A review. *npj Digit. Med.* **2024**, *7*, 222. <https://doi.org/10.1038/s41746-024-01219-0>.
20. Zhang, T.; Tan, T.; Wang, X.; Gao, Y.; Han, L.; Balkenende, L.; D'Angelo, A.; Bao, L.; Horlings, H.M.; Teuwen, J.; et al. Radio-LOGIC: A healthcare model for EHR processing. *Cell Rep. Med.* **2023**, *4*, 101131. <https://doi.org/10.1016/j.xcrm.2023.101131>.
21. American College of Radiology. *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*, 5th ed.; American College of Radiology: Reston, VA, USA, 2013.
22. Wu, N.; Phang, J.; Park, J.; Shen, Y.; Huang, Z.; Zorin, M.; Jastrzębski, S.; Févry, T.; Katsnelson, J.; Kim, E.; et al. Deep neural networks improve radiologists' performance. *IEEE Trans. Med. Imaging* **2020**, *39*, 1184–1194. <https://doi.org/10.1109/TMI.2019.2945514>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.