A Configuration-Driven Metadata Harmonisation Framework for Reproducible Multi-Database Mammography Research

Linda Blahová, Ivan Cimrák and Jozef Kostolný
University of Žilina
Žilina, Slovakia
linda.blahova, ivan.cimrak, jozef.kostolny@fri.uniza.sk

***Abstract*—Public mammography datasets allow extensive research, but integrating them directly is challenging due to variations in metadata schemas, label encodings, and annotation formats among different sources. This inconsistency forces researchers to repeatedly use scripts tailored to specific datasets, which increases the likelihood of errors and restricts reproducibility in the development of training and evaluation studies. We introduce a framework for harmonising metadata in mammography databases, which converts diverse dataset descriptions into a consistent lesion-level schema using database adapters and transformation pipelines. The framework standardises identifiers, imaging metadata, clinical labels, and localisation metadata, while maintaining traceability through clear domain identifiers. Dataset variants are created using stage-based pipelines specified in YAML configurations. These pipelines allow for unification, filtering, balancing, and addition of computed fields, without the need to change the underlying code. We demonstrate the framework by combining four mammography databases into a unified dataset that focuses on calcifications, combining 4,176 findings with both bitmap and bounding-box annotations. The resulting metadata output is reliable, can be reproduced, and is ready for use in subsequent preprocessing, modelling, or machine learning tasks.**

I. Introduction

The increasing availability of mammography datasets accessible to the public has enhanced research in automated breast imaging analysis [1]. However, conducting and replicating studies that involve multiple databases presents challenges due to variations in both imaging features and the organisation and language of metadata [2]. The differences in the images are visible even at patch level, where image appearance and contrast change across different sources (Fig. 1) [2]. The differences within the metadata are shown in detail in the following chapters.

A single idea, such as view position, laterality, or lesion type, may be represented by various terms and different value formats in different sources [3]. Diagnostic labels can be determined by pathology in one dataset and by radiological evaluation in another [1]. Additionally, the same label category might be represented using different scales or conventions for missing values, which can lead to inconsistencies. The metadata related to localisation introduces an extra level of diversity. Some datasets offer pixel-level masks (i.e. [4]), whereas others supply bounding boxes (i.e. [5]) or point annotations, which requires various approaches for subsequent processing.

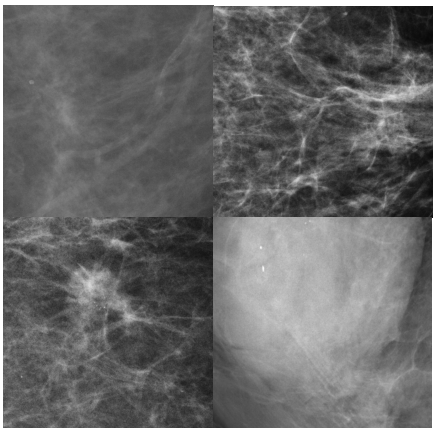

Fig. 1. Sample image demonstration of mammogram patches from all included databases

This variation results in the repetition of engineering tasks. Integrating datasets usually involves creating specific scripts for tasks such as aligning columns, standardising values, combining various metadata files, and making sure that identifiers and file paths remain uniform [6], [7]. These scripts tend to be challenging to manage, and not easily applicable to different projects. Consequently, preparing datasets often presents a significant challenge, and it is frequently the case that published experimental findings cannot be readily replicated without completely redoing the data preparation process [6], [7]. Currently the available tools for mammography metadata are limited. There is no direct and extensible framework that would focus directly on automatic metadata harmonisation, which could also be extensible to other databases, addressing the user needs.

This paper addresses the challenge by presenting a framework for metadata harmonisation that is driven by YAML configuration, aimed at supporting reproducible research in multi-database mammography studies. The main concept is to separate the logic that pertains to specific datasets from the fundamental processing code. A common standard schema outlines the intended representation, while dataset adapters established through configuration detail the methods for mapping and transforming raw metadata. In addition to this unified representation, different versions of the dataset can be created using clear transformation processes including unification, filtering, balancing, and computed fields. These processes can be combined based on the purpose of the dataset.

This method facilitates the creation of uniform datasets while also allowing for the addition of new databases.

The proposed framework is a single, unified system composed of several integrated components. It establishes one schema for lesion-level metadata and it uses configurable adapters to map diverse mammography databases into this schema without embedding dataset-specific logic into the core code. Built on top of this unified representation, the framework incorporates a YAML-based transformation pipeline that enables reproducible dataset creation through clear, sequential steps with intermediate outputs. As a demonstration, we harmonise metadata from four mammography databases using this framework and construct a calcification-focused dataset with consistent clinical labels and localisation metadata suitable for machine learning or data modelling tasks.

II. Related work

Several recent works have created unified or large benchmark datasets for mammography, such as Mammo-Bench [8] and the breast-cancer VQA benchmark dataset in study [9]. These datasets are useful, but they have clear limits. They are released as large but fixed, task-specific datasets, so researchers cannot easily extend them, change the metadata, or add new sources when their project requires something different. For example, Mammo-Bench provides one large dataset, but it does not allow adjustments such as adding new databases or modifying label definitions [8]. The VQA benchmark is designed mainly as dataset with large application for question-answering tasks, but it does not offer detailed information on how to reproduce the metadata harmonisation across all these datasets or support an integration of additional mammography datasets [9].

A related line of work is MammoClean [10], which focuses on standardising image appearance and metadata across datasets to reduce bias and improve consistency. While MammoClean supports harmonisation, its main goal is bias correction and dataset-level standardisation, not a flexible or configurable approach to dataset construction. Therefore, it cannot easily produce different dataset variants or incorporate new sources without modifying code.

Since mammography does not currently have directly comparable tools, to demonstrate the functionality of similar tools for metadata harmonisation used on different medical data we could refer to MAMS framework for single-cell data [11] or a tool for automated harmonisation using LLMs [12].

Since mammography research often needs different subsets, different lesion types, or different label mappings, fixed datasets are not enough. Researchers still end up writing their own scripts to filter data, unify metadata, or balance classes for training. This leads to inconsistencies and makes it hard to reproduce results. Our framework resolves these limitations through a scalable, configurable architecture detailed in this study.

III. Methods: Configuration-driven harmonisation and transformation pipelines

The proposed approach is designed as a configuration-driven framework that converts varied mammography metadata into a unified representation. The main goal is to separate dataset-specific handling (schemas, naming conventions, value encodings and file structures) from the core dataset creation logic, so that new sources can be integrated without changing the core implementation. Fig. 2 provides a high-level overview of the framework, showing how database adapters, the unified schema and the transformation pipeline interact to produce reproducible dataset variants.

A. Unified schema and design principles

The structure is designed with a unified lesion-level schema that represents each finding as a single entry in a unified metadata table. The framework includes four categories of fields. Initially, identifier fields provide information about the origin of the sample and enable consistent connections, including patient, study, and image identifiers, along with optional finding identifiers when applicable. Additionally, imaging metadata, including laterality and view position, is represented using a standardised encoding system [3]. Third, clinical label fields record target variables like pathology and BI-RADS assessment when they are accessible, along with additional metadata such as density, if available. Fourth, localisation metadata provides a uniform representation of lesion annotations across different datasets by including a clear field for the type of region of interest (ROI).

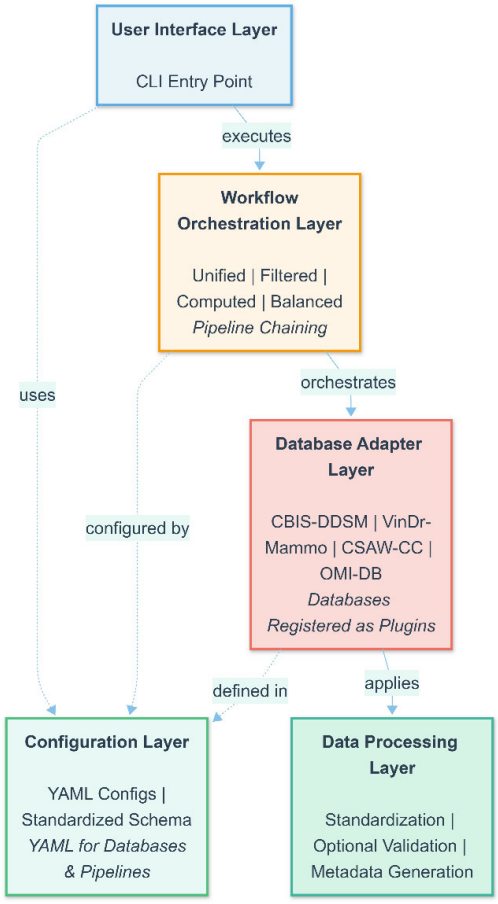

Fig. 2. Configuration-driven metadata harmonisation framework high-level architecture

A practical design requirement is that the schema should accommodate both pixel-level masks and bounding boxes. To accomplish this, each sample contains the ROI_type along with the relevant annotation payload – a mask path for bitmap

annotations or the coordinates (x1, y1, x2, y2) for bounding boxes. When a dataset includes various types of localisations, the framework consistently determines which type has higher priority. For instance, binary masks may be prioritised due to their ability to offer better localisation accuracy. Furthermore, every sample contains a specific field for the database (domain), which indicates the origin of the dataset. This prevents the loss of origin information during the merging process and enables subsequent processing even on domain level. [3]

B. Adapter-based mapping and value normalisation

Every database is connected via a database adapter that is specified through configuration. The adapter outlines the relationship between raw metadata columns and the unified schema, as well as the process by which values are standardised into a uniform format. The mapping layer addresses variations in naming conventions (for example, “LEFT,” “L,” and “Left”) and translates labels specific to datasets into standardised categories. When databases require particular preprocessing steps, adapters facilitate the implementation of specific logic either prior to or following the mapping process. This can include actions such as merging several metadata files into a single table or transforming density formats into a consistent scale. This design maintains a general approach to the main processing code while addressing the unique aspects of the dataset through configuration adjustments. An overview of how key fields from the individual source databases are mapped into the unified schema is summarised in Fig. 3.

To illustrate the heterogeneity that the adapters must handle, we show representative excerpts of the original metadata tables from each source database. These examples highlight typical differences in column naming, missing-value conventions, and how lesion attributes and localisation are represented. Some sources provide a single consolidated table, while others require merging multiple files using shared identifiers. Fig. 4–7 provide these raw metadata examples for CBIS-DDSM [4], OMI-DB [5], CSAW-CC [13] and VinDr-Mammo [14], respectively, and motivate why explicit mapping and value normalisation are needed before further processing.

C. Transformation pipelines and intermediate outputs

In addition to the unified representation, variations of the dataset can be generated through transformation processes defined in YAML pipeline configurations. A pipeline consists of a series of stages, with each stage taking a table as input and generating a modified table as output. The framework includes four main types of transformations – unification, filtering, balancing, and computed fields. Unification combines one or more datasets using adapters, resulting in a single, integrated table. Filtering involves choosing specific rows according to established criteria, such as selecting only those with calcifications, samples that have valid local annotations, or samples that have defined labels i.e. BI-RADS. Balancing allows for adjustable sampling methods that can be dependent on another variable, such as group membership, for instance, by only balancing the training set. Computation transformations produce derived fields, including risk categories based on BI-RADS. Outputs generated at each stage are stored to ensure that the process of creating the dataset is transparent and can be easily reviewed. The four supported transformation types and their role in the overall pipeline are summarised in Fig. 8.

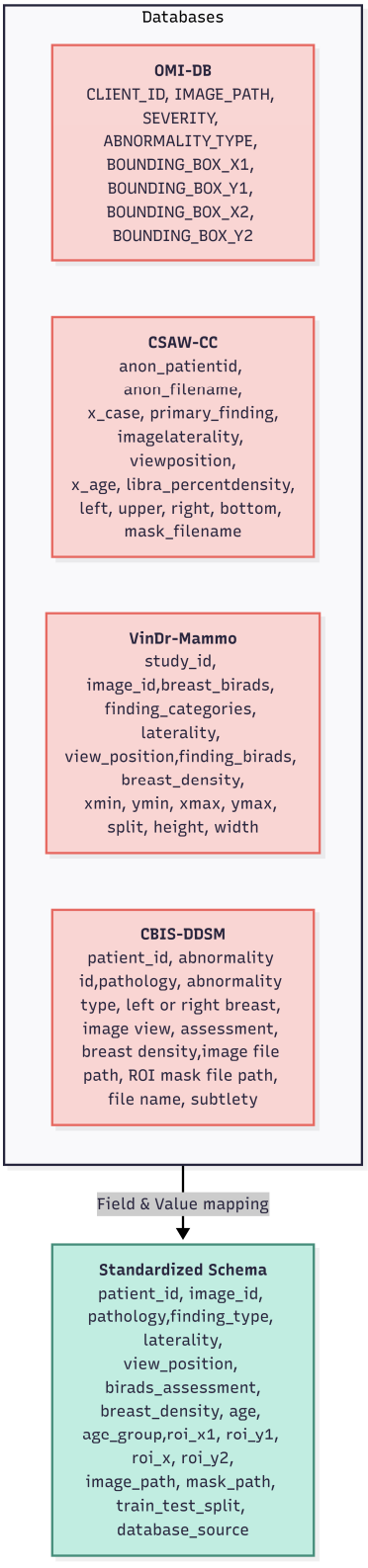

Fig. 3. Mapping for the source databases into the unified metadata schema

patient_id	breast density	left or right breast
P_00038	2	LEFT
P_00038	2	LEFT
P_00041	1	LEFT
P_00041	1	LEFT
P_00077	2	LEFT
image view	**abnormality id**	**abnormality type**
CC	1	calcification
MLO	1	calcification
CC	2	calcification
MLO	2	calcification
CC	1	calcification
assessment	**pathology**	**subtlety**
4	BENIGN	2
4	BENIGN	2
2	BENIGN_WITHOUT_CALLBACK	5
2	BENIGN_WITHOUT_CALLBACK	5
2	BENIGN_WITHOUT_CALLBACK	3

Fig. 4. Example of CBIS-DDSM metadata fields before harmonisation

CLIENT_ID	ABNORMALITY_TYPE	SEVERITY
demd6867	MASS	Malignant
demd6867	MASS	Malignant
demd5702	MASS	Malignant
demd4595	CALC	Malignant
demd4595	CALC	Malignant
BOUNDING_BOX_X1	**BOUNDING_BOX_Y1**	**BOUNDING_BOX_X2**
93	2301	266
232	2133	407
4454	3068	4605
1731	795	1877
1577	2078	1861

Fig. 5. Sample subset of the provided OMI-DB metadata

image_id	laterality	view_position	breast_birads
d8125545210c8e1b...	L	CC	BI-RADS 2
290c658f4e75a3f8...	L	MLO	BI-RADS 2
cd0fc7bc53ac632a...	R	CC	BI-RADS 2
71638b1e853799f2...	R	MLO	BI-RADS 2
dd9ce3288c0773e0...	L	CC	BI-RADS 1
breast_density	**split**	**finding_categories**	**finding_birads**
DENSITY C	training	['Mass']	BI-RADS 4
DENSITY C	training	['Mass']	BI-RADS 4
DENSITY C	training	['Global Asymmetry']	BI-RADS 3
DENSITY C	training	['Global Asymmetry']	BI-RADS 3
DENSITY C	training	['Architectural Distortion']	BI-RADS 4

Fig. 6. Examples of raw VinDr-Mammo metadata

anon_patientid	exam_year	x_age	x_case
2	2015	1	1
2	2015	1	1
2	2015	1	1
2	2015	1	1
4	2012	1	0
x_cancer_laterality	**x_type**	**rad_timing**	**rad_r1**
Left	3	2	0
Left	3	2	0
Left	3	2	0
Left	3	2	0
NA			0
left	**upper**	**right**	**bottom**
3156	2204	3157	2205
384	1124	566	1291
156	2666	157	2667
43	2384	44	2385
34	2352	139	2459

Fig. 7. Examples of raw CSAW-CC metadata from multifile sources

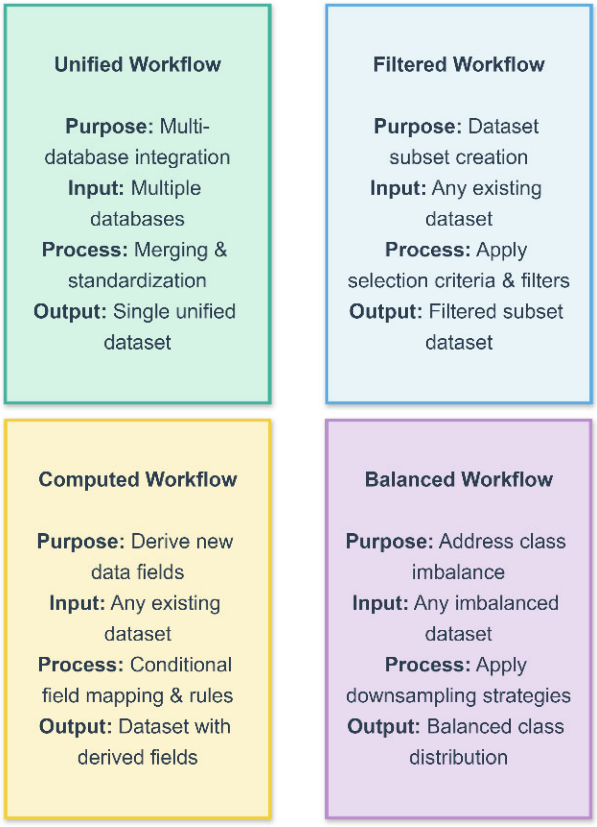

Fig. 8. Available workflows that can be used in the pipeline

D. Example: Calcification dataset pipeline (YAML)

The following YAML snippet demonstrates a typical pipeline that combines various databases, filters for calcifications with accurate localisation and balances the training split by pathology. This aligns with the example outlined in the framework design but is presented as a configuration for an executable pipeline.

```yaml
# Global pipeline configuration
workflow_type: pipeline
# Pipeline configuration
pipeline:
  # Stage 1: Unify databases
  - workflow_type: unified_dataset
    stage_name: "Unify Mammography Databases"
    # Database configuration - All major databases
enabled
    databases:
      cbis_ddsm:
        enabled: true
      vindr:
        enabled: true
      csaw_cc:
        enabled: true
      omi_db:
        enabled: true
    # Output for unified dataset (intermediate)
    output:
      path: "../outputs/dataset_4/step1_unified_base.csv"
      format: csv
      encoding: utf-8
  # Stage 2: Filter for calcifications first
  - workflow_type: filtered_dataset
    stage_name: "Apply Calcification Filtering"
    # Filters for calcification records
    filters:
      finding_type:
        operator: in
        values: [CALCIFICATION]
      pathology:
        operator: in
        values: [BENIGN, MALIGNANT]
    # Intermediate output
    output:
      path: "../outputs/dataset_4/step2_filtered.csv"
      format: csv
      encoding: utf-8
  # Stage 3: Apply pathology balancing
  - workflow_type: balanced_dataset
    stage_name: "Balance Pathology Classes"
    # Balancing strategy configuration
    balancing_strategy:
      # Use pathology as the primary balancing field
      primary_field: pathology
      target_distribution:
        type: equal
        # Pathology classes to balance
        classes: ['BENIGN', 'MALIGNANT']
      sampling_method: downsample_to_minimum
      preserve_proportions:
        database_source: true
        train_test_split: true
        balance_test_split: false
        balance_validation_split: false
    # Final balanced output
    output:
      directory: "../outputs/dataset_4"
      filename: "dataset_4_pathology_balanced.csv"
      format: csv
```

This mechanism allows production of different datasets by making changes only to the YAML file. For instance, balancing can be removed to keep all training data, filtering can be adjusted for specific view positions, or additional computed columns can be created.

E. Full pipeline used in the study

The process of creating a dataset can be outlined in four main stages. Initially, unification imports CBIS-DDSM, OMI-DB, CSAW-CC, and VinDr-Mammo using dataset adapters and creates a single harmonised table with consistent field names and values. Secondly, the filtering process identifies calcification findings and excludes samples that lack localisation metadata (binary mask or bounding box). In the pipeline used in our study, dataset balancing is skipped at this stage, as it is handled in later processing by on-the-fly data augmentation. Additionally, we use column computation to produce one derived target – a binary risk label (Low/High) based on BI-RADS categories when they are available. This approach marks any empty values as 'UNKNOWN' for datasets that do not include BI-RADS assessments. This organised procedure demonstrates dataset creation in mammography research, where different versions of datasets need to be created consistently and reliably for various experimental conditions.

IV. Experimental Demonstration: Harmonised Calcification Dataset Across Four Databases

A. Dataset composition and label availability

We illustrate the framework by creating a unified dataset that focuses on calcification, utilising data from four mammography databases. The combined dataset includes 4,176 instances of calcification findings. The per-database composition of the unified calcification dataset is summarised in Table 1.

TABLE I. Unified dataset composition by database

Database	Number of Samples
CBIS-DDSM	1,401
OMI-DB	2,229
CSAW-CC	144
VinDr-Mammo	402

The dataset focuses on individual lesions and contains only samples that have localisation information appropriate for patch-based analysis. In the combined dataset, the localisation annotations consist of both bitmap masks and bounding boxes: 1,545 samples utilise bitmap masks, while 2,631 samples employ bounding-box localisation. This shows the natural variety in public datasets and highlights the necessity for a consistent way to represent location in metadata.

In terms of diagnostic labels, pathology is accessible for many sources, though not consistently. The combined dataset includes 2,525 cases of malignant findings and 1,651 cases of benign findings, based on the pathology definitions applied for the unified representation. Ambiguous categories have been addressed explicitly where necessary. BI-RADS assessments are inherently included in CBIS-DDSM and VinDr-Mammo. Furthermore, radiological evaluation from different standard is linked from OMI-DB samples and mapped into BI-RADS.

B. Output schema consistency and provenance

After standardisation, all samples have a uniform set of fields that detail their origin, imaging information, labels, and location. The clear database field maintains the origin of the source for later analysis specific to each domain and for managing processing that is particular to each domain. The clear ROI_type field allows for the management of bitmap and bounding-box annotations through one unified downstream interface, eliminating the need for dataset-specific variations in later processes. A representative sample of the harmonised metadata output is shown in Fig. 9.

database_source	**patient_id**	**image_id**
csaw_cc	11184	11184_20990909_L_MLO_1
csaw_cc	11184	11184_20990909_L_CC_1
csaw_cc	11266	11266_20990909_R_MLO_1
omi_db	demd4595	1.2.826.0.1.3680043...9258.0
omi_db	demd4595	1.2.826.0.1.3680043...9266.0
breast_density	**birads_assessment**	**pathology**
2	UNKNOWN	MALIGNANT
2	UNKNOWN	MALIGNANT
2	UNKNOWN	MALIGNANT
UNKNOWN	UNKNOWN	MALIGNANT
UNKNOWN	UNKNOWN	MALIGNANT
finding_type	**train_test_split**	**age_group**
CALCIFICATION	TRAIN	1
CALCIFICATION	TRAIN	1
CALCIFICATION	TEST	1
CALCIFICATION	TRAIN	UNKNOWN
CALCIFICATION	TRAIN	UNKNOWN
roi_type	**roi_x1**	**roi_y1**
bitmap	804.0	1304.0
bitmap	329.0	894.0
bitmap	2409.0	1533.0
bounding_box	1731.0	795.0
bounding_box	1577.0	2078.0
roi_x2	**roi_y2**	**composite_id**
1056.0	1601.0	csaw_cc_11184_..._MLO_1
649.0	1154.0	csaw_cc_11184_..._CC_1
2508.0	1602.0	csaw_cc_11266_..._MLO_1
1877.0	885.0	omi_db_demd4595_...
1861.0	2165.0	omi_db_demd4595_...

Fig. 9. Sample from subset of harmonised metadata of the unified dataset

C. Reproducibility and traceability

The main practical result of the framework is that the process of creating the dataset can be repeated simply by using the available defined configurations. Saving intermediate outputs after each transformation stage allows for the validation of each step and enables the exact regeneration of the dataset, including any decisions related to filtering or balancing. This is especially important when datasets or label mappings change over time, or when several researchers need to create similar dataset versions for various studies. To allow the usage of this framework and dataset reproducibility, the framework is available at the GitHub in https://github.com/lindajblahova/mammography_metadata_harmonization.

To verify that the framework produces consistent and analysable outputs, we conducted a rigorous validation by comparing the descriptive statistics of the unified dataset against the baseline statistics of the raw databases. This step is crucial to ensure that the mapping and transformation pipelines did not introduce data corruption or loss. By using a previously developed database analysis tool [15], we verified that the distribution of key clinical features, such as breast density and pathology, remained aligned between the source data and the harmonised output. For example, the breast density distributions shown in Fig. 10 confirm that the categorical mapping in the adapters correctly reflects the original data proportions across all four domains. Some additional checks include the composition of the unified dataset by database (Fig. 11) and database proportions across train and test splits in the used dataset, which is useful for further usage of this dataset in machine learning (Fig. 12). These checks serve as structural validation, proving that the framework successfully unifies heterogeneous sources into a single, reliable schema while maintaining the integrity of the medical information.

V. Discussion, Conclusions and Future Work

The main practical result of the study is the framework that allows reproducible, configurable and extensible dataset creation. Saving intermediate outputs after each transformation stage allows the validation of each step and enables the exact regeneration of the dataset, including any decisions related to filtering or balancing. This is important when several researchers need to create different dataset versions for various studies while having the same baseline but different research objective.

A limitation of metadata harmonisation is that it cannot completely address the fundamental differences among various datasets. For instance, some databases may provide fields that are not included in other databases, which cannot be fixed by this framework. For such result, a standard on the provided metadata for mammography databases should be defined by medical organisations on the international level. Therefore, the framework aims to unify the available data and show the missing values. This ensures consistency across research groups while reducing the need for separate processing of specific databases. Since the framework is configurable for both databases and pipelines, it can be easily extended to new databases, simply by implementing the YAML mapping configuration and database adapter, or YAML configuration for different output dataset, i.e. for mass-

only dataset that combines only the CSAW-CC and VinDr-Mammo databases. This architectural design promotes careful reuse of diverse data while ensuring that the process of creating the dataset is consistent and can be verified.

While this study establishes the structural foundation for metadata harmonisation, several paths for future research remain to enhance the framework's utility and performance. Primarily, we intend to conduct downstream machine learning experiments to provide a quantitative evaluation of how this harmonised multi-database approach impacts model preparation and training processes compared to multi-database training without metadata preprocessing. To address scalability for even larger imaging cohorts, future iterations will focus on code optimisation to reduce computational overhead during the transformation and filtering stages, ensuring efficient dataset generation. Furthermore, we aim to perform benchmarking against existing tools to compare efficiency and mapping accuracy. Finally, exploring the integration of Large Language Models (LLMs) could automate the initial mapping of raw metadata columns, further reducing the manual configuration effort and facilitating the rapid inclusion of new international mammography sources.

In summary, using configuration-driven metadata harmonisation offers a useful foundation for scalable research in mammography. By reducing the engineering workload and enhancing reproducibility, the framework allows research teams to concentrate more on modelling and evaluation while ensuring clear and restorable dataset definitions.

Additionally, the core of the framework is CSV oriented but independent of the context, therefore the target purpose could be easily transformed into different CSV data processing. It could be used i.e. to process data from fluorescent spectroscopy [16], simply by defining corresponding YAML configurations, database adapters, own standard schema and pipelines for the specific purpose. All core logic for CSV loading, processing, merging, and workflows would remain the same.

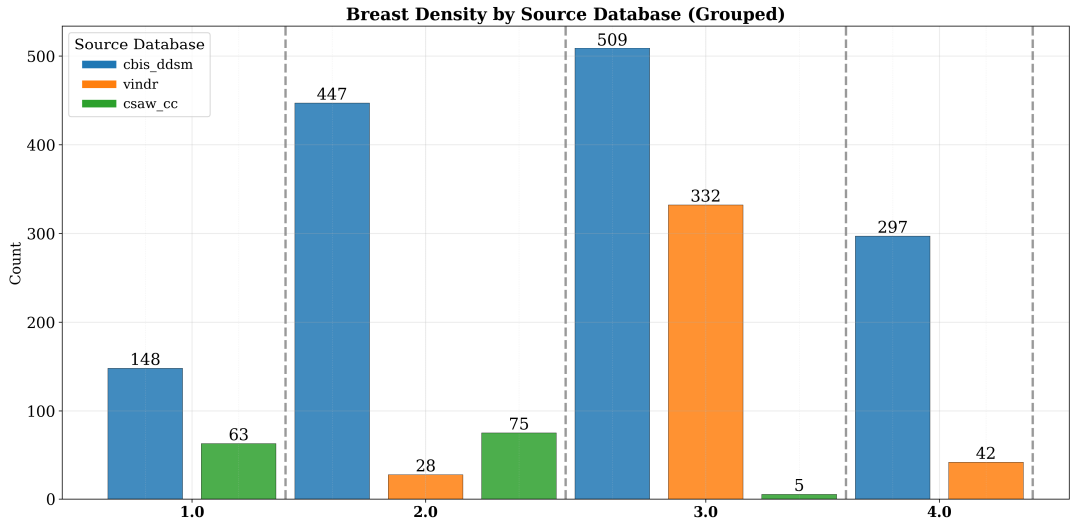

Fig. 10. Distribution of breast density grouped by source database

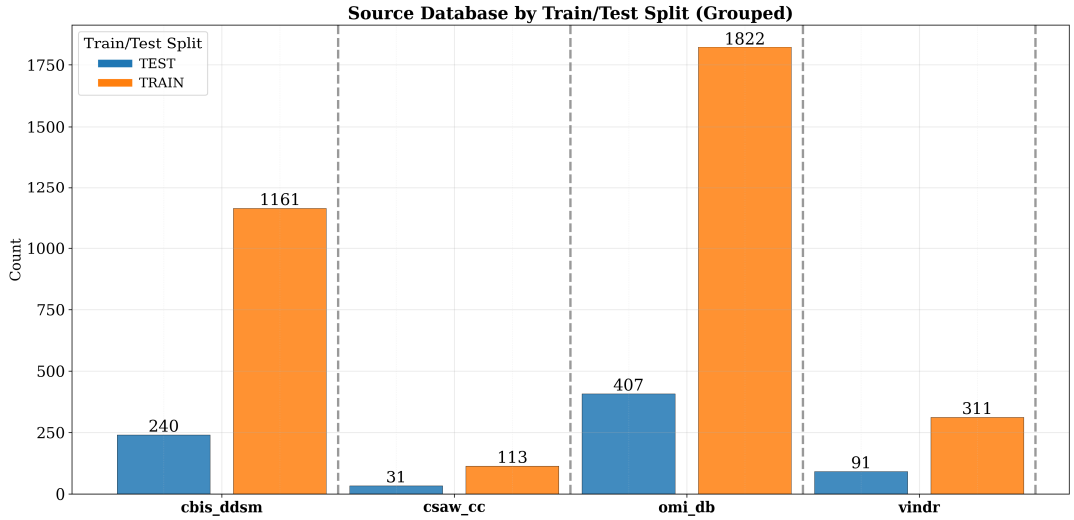

Fig. 11. Distribution of source database grouped by source database

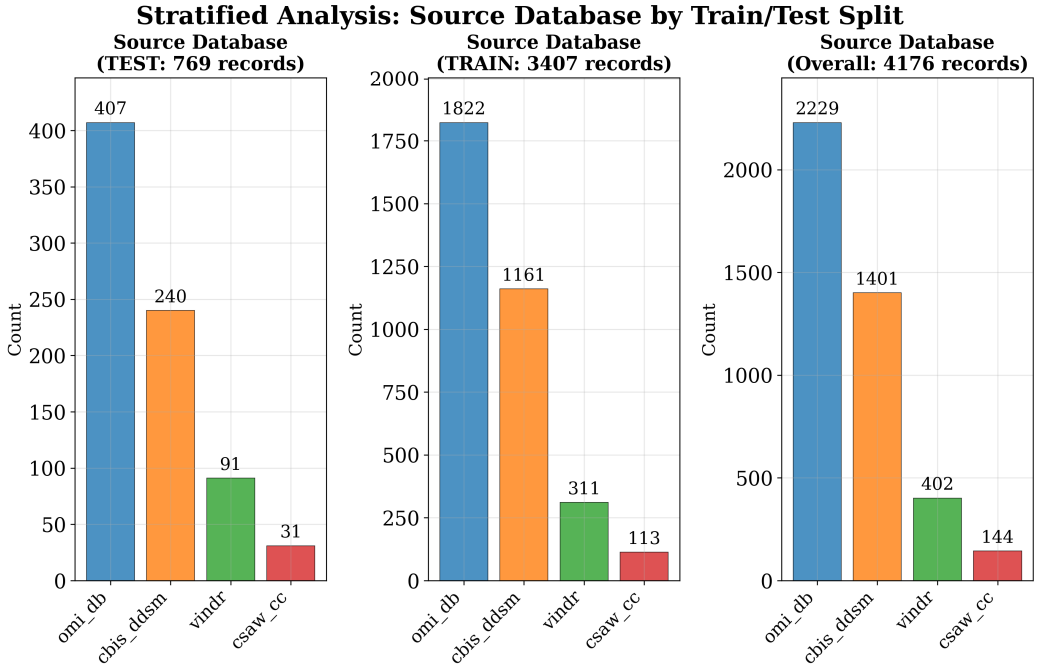

Fig. 12. Distribution of source databases across the train, test and overall dataset

ACKNOWLEDGMENT

Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I05-03-V02-00078.

Funded by the European Union NextGenerationEU

RECOVERY AND RESILIENCE PLAN

REFERENCES

[1] C. R. Taylor, N. Monga, C. Johnson, J. R. Hawley, and M. Patel, "Artificial Intelligence Applications in Breast Imaging: Current Status and Future Directions", *Diagnostics*, vol. 13(12), 2023, 2041.

[2] S. Seoni, A. Shahini, K. M. Meiburger, F. Marzola, G. Rotunno, U. R. Acharya, F. Molinari, and M. Salvi, "All you need is data preparation: A systematic review of image harmonization techniques in Multi-center/device studies for medical support systems", *Comput Methods Programs Biomed*, vol. 250, 2024, 108200.

[3] N. Sourlos, R. Vliegenthart, J. Santinha, M. E. Klontzas, R. Cuocolo, M. Huisman, and P. M. A. van Ooijen, "Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology", *Insights Imaging*, vol. 15, no. 1, 2024, pp. 1–12.

[4] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research", *Sci. Data*, vol. 4, 2017.

[5] M. D. Halling-Brown, L. M. Warren, D. Ward, E. Lewis, A. Mackenzie, M. G. Wallis, L. S. Wilkinson, R. M. Given-Wilson, R. McAvinchey, and K. C. Young, "OPTIMAM Mammography Image Database: A LargeScale Resource of Mammography Images and Clinical Data", *Radiol. Artif. Intell.*, vol. 3, no. 1, 2020, e200103.

[6] J. Mongan, L. Moy, and C. E. Kahn, "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers", *Radiol. Artif. Intell.*, vol. 2, no. 2, 2020.

[7] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future", *npj Digital Medicine*, vol. 5, no. 1, 2022, pp. 48.

[8] G. Bhole, S. Suba, and N. Parekh, "Mammo-Bench: A Large-scale Benchmark Dataset of Mammography Images", *medRxiv*, 2025, Web: https://doi.org/10.1101/2025.01.31.25321510.

[9] J. Zhu, F. Huang, Q. Luo, and H. Chen, "A Benchmark for Breast Cancer Screening and Diagnosis in Mammogram Visual Question Answering", *Nat. Commun.*, vol. 16, no. 1, 2025, 11683.

[10] Y. Zafari, H. Pan, G. Durak, U. Bagci, E. A. Rashed, and M. Mabrok, "MammoClean: Toward Reproducible and Bias-Aware AI in Mammography through Dataset Harmonization", 2025, Web: https://www.arxiv.org/pdf/2511.02400.

[11] Y. Wang, I. Sarfraz, W. K. Teh, A. Sokolov, B. R. Herb, H. H. Creasy, I. Virshup, R. Dries, K. Degatano, A. Mahurkar, D. J. Schnell, P. Madrigal, J. Hilton, N. Gehlenborg, T. Tickle, and J. D. Campbell, "Matrix and analysis metadata standards (MAMS) to facilitate harmonization and reproducibility of single-cell data", *bioRxiv*, 2023.

[12] K. Higashi, Z. Nakagawa, T. Yamada, and H. Mori, "Automated Harmonization and Large-Scale Integration of Heterogeneous Biomedical Sample Metadata Using Large Language Models", *bioRxiv*, 2024.

[13] F. Strand, "CSAW-CC (mammography) – a dataset for AI research to improve screening, diagnostics and prognostics of breast cancer (Version 1)", *Karolinska Institutet*, 2022, Web: https://doi.org/10.5878/45vm-t798.

[14] H. T. Nguyen, H. Q. Nguyen, H. H. Pham, K. Lam, L. T. Le, M. Dao, and V. Vu, "VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography", *Scientific Data*, vol. 10, no. 1, 2023, pp. 1–8.

[15] L. Blahová, J. Kostolný, and I. Cimrák, "Cross-Database Analysis of Mammography Data via a Configuration-Driven Framework with Automated Visualization", *Symposium on Applied Machine Intelligence and Informatics (SAMI 2026)*, Stará Lesná, Slovakia, 2026.

[16] M. Švecová, L. Blahová, J. Kostolný, A. Birková, P. Urdzík, M. Mareková, and K. Dubayová, "Enhancing endometrial cancer detection: Blood serum intrinsic fluorescence data processing and machine learning application", Talanta, vol. 283, 2025, 127083.