Electronic Letters on Computer Vision and Image Analysis 25(3):46-64, 2026

LAVIA-MXID: A Multi-Anatomical Annotated X-Ray Dataset for Benchmarking Deep Learning-Based Compression

Amina Fettah*, Rafik Menassel+, Abdeljalil Gattal+, Yahia Menassel+, Abdelhak Gattal‡

* *Laboratory of Mathematics, Informatics and System (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria*
+ *Laboratoire de Vision et d'Intelligence Artificielle (LAVIA), Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria*
‡ *Public Hospital Establishment, EL-Aouinet, Tebessa, Algeria*

Received 15th of May 2025; revised 1 April 2026; accepted 18th of June 2026

Abstract

X-ray medical imaging datasets are crucial for accurate diagnosis and driving innovation in medical imaging technologies. The LAVIA-MXID extended dataset, curated by the Laboratory of Vision and Artificial Intelligence (LAVIA), addresses limitations of its predecessor, the MXID dataset. It offers increased diversity, nearly 10,000 annotated images, and features tailored for advanced medical imaging applications. This study presents a deep learning approach for medical image compression, focusing on recognizing peculiarities and problems in medical imaging pattern recognition, such as data set quality, labeling errors, and data imbalance. The LAVIA-MXID dataset, an enhanced version of the MXID dataset, features 9989 images managed by radiologists and medical experts. It is meticulously labelled, categorized into training, validation, and test subsets, and isolated and eliminated for high-quality annotation standards. It supports various image analysis tasks but does not include annotations for specific pathological conditions. The study analyzed 9,989 medical images from the LAVIA-MXID dataset using deep learning methods. The results showed marginal improvements in the extended dataset (version 2) but enhanced image quality with reduced reconstruction artifacts. The four models - Autoencoder (AE), Deep Convolutional Autoencoder (DCAE), Convolutional Neural Network (CNN), and Variational Autoencoder (VAE) - showed improvement in performance metrics like MSE, PSNR, and MS-SSIM. The proposed dataset and methods improve X-ray medical imaging diagnostic accuracy by training deep learning algorithms, addressing issues like dataset size, quality, labeling errors, and data imbalance. The novelty of the work lies in the development of the LAVIA-MXID dataset specifically for compressing medical images and the rigorous comparison of various deep learning models to establish the most appropriate model that preserves image quality upon compression. The LAVIA-MXID dataset, developed by LAVIA, offers advanced medical image analysis using deep learning paradigms, enhancing compression quality.

Key Words: X-ray medical imaging, LAVIA-MXID datasets, Medical x-ray imaging dataset (MXID), Image Compression, Image Analysis.

1 Introduction

The digitization of healthcare systems has led to an unprecedented increase in medical imaging data. Among these modalities, X-ray imaging remains one of the most widely used diagnostic tools due to its accessibility,

Correspondence to: r.menassel@univ-tebessa.dz

Recommended for acceptance by Angel D. Sappa
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

cost-effectiveness, and efficiency. X-ray medical imaging datasets consist of digital radiographic images that provide detailed visualization of internal anatomical structures such as bones, tissues, and organs, enabling clinicians to diagnose a wide range of conditions with high accuracy. However, the rapid growth in data volume introduces significant challenges in storage, transmission, and real-time processing, particularly in large-scale clinical environments. Comprehensive and high-quality X-ray datasets are essential for training machine learning and deep learning algorithms intended for clinical deployment[1]. These algorithms have demonstrated remarkable success in tasks such as detection, segmentation, and classification, significantly improving diagnostic accuracy and workflow efficiency. By leveraging large-scale datasets, artificial intelligence (AI) systems can identify complex patterns and subtle features that are often imperceptible to human observers, thereby enabling earlier disease detection and supporting personalized treatment strategies[2]. Furthermore, such datasets play a critical role in quality assurance by ensuring the proper functioning and calibration of imaging equipment. Despite these advances, an important research gap persists. While deep learning has achieved impressive results in high-level vision tasks, compression and efficient representation learning for medical images remain relatively underexplored [3]. This limitation is largely due to the absence of suitable datasets that simultaneously provide multi-anatomical diversity, high-quality annotations, and a benchmark-ready structure tailored for compression evaluation. Most publicly available X-ray datasets focus on single anatomical regions (e.g., chest), lack consistent annotation protocols, and are not designed for reproducible compression benchmarking. To address these limitations, this work introduces LAVIA-MXID, an extended multi-anatomical X-ray dataset curated by the Laboratory of Vision and Artificial Intelligence (LAVIA). The dataset comprises 9,989 annotated images spanning 18 distinct anatomical regions, significantly exceeding the anatomical diversity of widely used datasets such as JSRT and MURA. In addition to its scale, LAVIA-MXID incorporates clinically guided annotation protocols with multi-stage validation to ensure high-quality and reliable labeling. The dataset also includes gender labels, enabling fairness-aware model development—an aspect often overlooked in existing datasets. Beyond dataset construction, this study establishes a compression-focused evaluation framework designed to enable reproducible benchmarking of deep learning models. Unlike prior work that primarily emphasizes disease classification or segmentation, this research targets the underexplored problem of medical image compression. Specifically, we provide a statistically validated comparative analysis of four representative architectures: Autoencoder (AE), Convolutional Neural Network (CNN), Variational Autoencoder (VAE), and Deep Convolutional Autoencoder (DCAE). These models are evaluated in terms of their ability to achieve efficient representation while preserving structural fidelity—an essential requirement in medical imaging. The LAVIA-MXID dataset is designed to overcome several critical limitations of existing resources. Unlike datasets with restricted anatomical scope, it provides broad coverage across multiple body regions. It ensures standardized image quality through a consistent 1024×1024 PNG format, facilitating uniform training and evaluation. Additionally, it addresses common challenges such as data imbalance and annotation inconsistencies through careful curation and validation processes. Although explicit pathology annotations are not included. It is important to note that LAVIA-MXID is a general-purpose imaging dataset designed for technical evaluation tasks such as compression and representation learning, not for direct diagnostic or pathology-specific analysis without additional annotations. By offering rich, diverse, and well-structured visual data, LAVIA-MXID supports both clinical diagnostics and the development of advanced AI systems. It enables robust training and evaluation of deep learning models, contributes to improved diagnostic accuracy, and promotes standardization in medical image analysis. Moreover, it provides a foundation for future research in AI-driven radiology, particularly in areas where efficient data representation is critical. The primary contribution of this work is the introduction of the LAVIA-MXID dataset. The experimental evaluation is designed to establish baseline benchmarks using standard deep learning architectures, not to propose a new compression technique.

On the other hand, the main contributions of this work can be summarized as follows:

- Dataset Contribution: Introduction of LAVIA-MXID, a large-scale multi-anatomical X-ray dataset with 18 regions and 9,989 annotated images.
- Annotation Framework: Development of a clinically guided, multi-stage validated annotation protocol.

- Benchmark Contribution: Establishment of a compression-oriented evaluation framework for reproducible comparison.
- Experimental Insight: Comprehensive and statistically validated comparison of AE, CNN, VAE, and DCAE architectures for medical image compression.

The remainder of this paper is organized as follows. Section 2 reviews related work on medical imaging datasets, compression techniques, and evaluation metrics. Section 3 details the construction and characteristics of the LAVIA-MXID dataset. Section 4 presents the experimental setup and comparative analysis of deep learning models. Finally, Section 5 concludes the paper and outlines future research directions in AI-based medical image compression.

2 Related Works

This part emphasizes the work associated with medical imaging datasets and shows how these datasets are important in building and validating machine learning (ML) and deep learning (DL) algorithms in medical domains. The existence of many datasets has enabled researchers to formulate models and apply them in numerous clinical situations. Nevertheless, in-depth analysis of existing datasets is important in determining their shortcomings and areas that need to be further improved. Additionally, the existence of large X-ray images that have been described in detail helps to formulate trends and relationships across different populations, and, therefore, more tailored and precise treatment strategies. That dataset may prove to be very advantageous in early diagnoses of diseases resulting from timely interventions with better outcomes for patients. Additionally, a diverse dataset will ensure that AI algorithms do not show any bias toward any demographic group, thus ensuring fair and equal health care for everyone. All in all, developing such a dataset is very essential for the development of AI in health care and the improvement of patient care around the world. In fact, the NIH Chest X-ray dataset [5] has more than 100,000 frontal-view chest X-ray images, out of which you can find some images of different diseases, like pneumonia and lung nodules. It can be of great value for the training and testing of algorithms to automatically analyze anomalies related to the chest. From this new resource, MURA [6] is the public collection of 40,561 simple views of radiographs forming a musculoskeletal system, with materials from the chest, lower extremities, and upper extremities. It has been very useful in providing insights into deep learning-based model development in detecting musculoskeletal disorders. Yet MURA still serves very restricted applications: due to anatomical procurement limitations, it does not capture variations from other body parts. The JSRT dataset [7] has 247 images of chest X-rays and belongs to pulmonary disorders. Whereas it's small in size, it has been widely used for algorithm building on lung anomaly detection and classification. Another reference source is the RSNA Pneumonia dataset [8], which consists of 30,000 chest X-ray images dedicated to pneumonia detection to serve as a benchmark for developing algorithms in this area. ChestX-ray8 [9] has 108,948 images linking a radiological finding with such findings as cardiomegaly and pneumonia. On the other hand, the "Dataset for automatic Region-based Coronary Artery Disease Diagnostics Using X-Ray Angiography Images" [10], provides 1,500 annotated X-ray angiography images to support machine learning research in Coronary Artery Disease diagnosis. It provides dense structural and stenotic plaque annotations with easier task operation of lesion detection, vessel segmentation, and risk assessment. The data is divided into training, validation, and test datasets with PNG images and JSON labels. Expert cardiologists cross-checked high-quality labels for stenosis, which was ≥ 50 percent narrowing of ≥ 1.5 mm vessels. The dataset is publicly available under a CC0 license via Zenodo. Also, The CheXmask dataset [11] provides 657,566 well-annotated anatomical segmentation masks for CXR images, as a reaction to the lack of fine-grained anatomical annotations in medical image datasets. It consolidates five large public sources of CXR images—ChestX-ray8, CheXpert, MIMIC-CXR-JPG, Padchest, and VinDr-CXR. Segmentations were generated with HybridGNet model and confirmed through expert review and automatic quality assurance through the Reverse Classification Accuracy framework. Masks exist in original and uniform resolution using Run-Length Encoding. Original CXR images are unavailable due to privacy

limitations and must be acquired separately. CheXmask can be accessed free on PhysioNet. Releasing the Popov et al. (2024) [10] and CheXmask (2024) [11] datasets alongside the LAVIA-MXID dataset highlights the growing endeavor in the medical imaging community to provide accessible specialized, high-quality resources to the different demands of research. While Popov et al. [10] focus on coronary artery disease with fine lesion annotation on angiography images, and CheXmask [11] focuses on large-scale anatomical segmentation on chest X-rays, LAVIA-MXID complements these datasets by adding a broader anatomical spectrum of 18 body regions, comprising approximately 10,000 fully accessible, well-organized images. Presenting these datasets together highlights how LAVIA-MXID fills a key niche among current resources—it provides a balance between dataset size, variety, and ease of access that renders it particularly suitable for testing general-purpose models, particularly in tasks such as image compression, representation learning, and generative modeling. To this end, the inclusion of Popov et al. and CheXmask provides context and contrast that serve to underscore the usability and distinctive value of LAVIA-MXID within the medical imaging dataset landscape. In breast imaging, the ISIC Skin Lesion dataset [12], which contains 33,126 distinct images of different skin conditions, forms a baseline for algorithm development to evaluate several skin disorders, such as melanoma. The INbreast dataset [13], made of 410 mammographic images, also supports breast cancer screening and algorithm research. The others include lung imaging and multimodal imaging datasets, as listed above:

- The CAMELYON16 dataset [14], which includes digitized histopathology slides, is dedicated to breast cancer detection and serves as a benchmark for developing automated classification algorithms.
- The Open-I dataset [15] contains a diverse range of medical images, including X-rays, CT scans, and MRIs, making it a versatile tool for various medical imaging tasks.
- The LIDC-IDRI dataset [16] comprises 1,018 chest CT scans, focusing on lung nodule detection and classification.

Compared to existing X-ray image datasets, such as NIH ChestX-ray14, MURA, and JSRT, LAVIA-MXID has several distinguishing aspects. The first one is the inclusion of a boarder set of body parts, while the majority of public datasets cover a limited set of anatomical regions. Secondly, the LAVIA-MXID dataset includes gender classification and labeling for both male and female, which facilitates gender-based research, which is often limited in other datasets. Thirdly, all images were standardized and converted to PNG format to facilitate deep learning-based applications. On the other hand, LAVIA-MXID dataset lacks pathological labeling, but it offers a well-structured foundation for benchmarking tasks such as unsupervised learning, image compression, and enhancement—areas that are often under-represented in disease-centric datasets. The MXID [4] dataset comprises a total of 6869 X-ray images distributed over a number of different sections of the body, namely, lung, abdomen, skull, and extremities. As part of the research using diagnostic imaging for different medical conditions, this dataset is meant to assist in the examination of all insights related to conditions detected and diagnosed. The extended version, LAVIA-MXID, has a total of 9989 images, making it a larger and more diverse collection with high-quality annotations. The modest quantitative improvements between MXID and LAVIA-MXID are attributable to the similar imaging protocols used in both collections. The primary advantages of LAVIA-MXID—greater anatomical diversity and larger sample size—are expected to enhance model generalization and robustness, which may not be fully captured by pointwise distortion metrics such as MSE and PSNR. Existing datasets, however, are fraught with serious limitations:

- Restriction of coverage: A collection of datasets is developed focusing on a particular body region, such as the musculoskeletal system or the thorax, while leaving others, such as the abdomen and the extremities, completely uncovered.
- Class Imbalance: Some have rare conditions that are not adequately represented; this creates biased algorithms that may fail to generalize much.
- Annotation Quality: Inconsistent or incomplete annotations may compromise the training of an ML model, causing unreliability and lack of accuracy.

The above-mentioned limitations need to be resolved to advance the state-of-the-art diagnostic algorithms to accurately detect anomalies across the diverse body regions. Indeed, comprehensive datasets with underrepresented locations, balanced class distributions, and high-quality annotations would significantly improve the performance and generalizability of ML models. Medical image compression is critical for managing the large size of datasets while maintaining diagnostic quality. Compression techniques are classified as lossless and lossy:

- Lossless Compression: Reduces file size without any data loss, ensuring diagnostic integrity [17].
- Lossy Compression: Achieves greater size reduction by discarding some data, potentially degrading image quality compared to the original [1, 18].

In the context of the LAVIA-MXID and MXID datasets, both lossless and lossy compression techniques are evaluated to balance file size reduction and image quality. Ensuring pixel-level accuracy is essential to maintain diagnostic value, particularly for X-ray images where small anomalies can have significant clinical implications. Table 1 shows some X-ray images datasets (publicly available) that provide a unique zone of the human body, unlike our dataset that represents a multi-body part.

3 Overview of LAVIA-MXID dataset

The LAVIA-MXID dataset is an advancement compared to the MXID dataset revised with 6869 X-ray images in that this enhanced version is accomplished with greater diversity and a more enriched collection of medical images. It now includes 9989 images selected and managed by experienced radiologists and medical experts. Every image in the LAVIA-MXID dataset is meticulously labelled with its corresponding body region, thus allowing for precise evaluation of various medical analysis techniques. Since it has highly detailed annotations, the dataset can even be applied for a wide range of machine-learning and deep-learning applications.

3.1 Dataset collection and preparation

The dataset preparation methodology has been primarily designed to achieve balanced representation across different body parts, voiding bias in model creation and evaluation. Further, images are categorized into training, validation, and test subsets, thereby confirming the sturdiness of the framework for development and testing purposes of diagnostic algorithms.

The dataset collected consists of 9,989 images from AOUINET Hospital in a region in Tebessa City, during approximately 2 years. Each medical image in the dataset has a resolution of 1024×1024 pixels and was recorded for patients undergoing diagnostic imaging acquisition procedures with different acquisition devices. Two-dimensional medical images were obtained using CR plates. Such images are crucial for patients placed in the treatment area in order to enable the design of the lung block configuration. The tabletop CR 15-X digitizer is based on proven Agfa technology. Figure 1 demonstrates the general process of the dataset construction.

Healthcare technology with a modular but durable architecture that integrates cost-effectiveness with consistently superior picture quality, customizable speed, and a user-customized workflow. This CR digitizer is adaptable and capable of managing a wide array of digital radiography applications. The CR 15-X digitizer is compatible with an NX workstation, a DRYSTAR 5302 digital thermal printer, and/or an Operating System software package. The CR 15-X employs the specialized MUSICA image processing technology to guarantee superior picture quality. MUSICA autonomously evaluates the attributes of each picture and optimizes processing settings irrespective of the anatomical region and variations in X-ray dosage. Consequently, we achieved a minimal quantity of redundant photos or post-processing.

MUSICA offers the same picture presentation capabilities as computed radiography (CR) when used with direct radiography (DR) equipment, hence enhancing diagnostic confidence. The MUSICA method is entirely

Table 1: Summary of Medical Imaging Datasets [4].

Dataset	Clinical Context	Images	Body Parts	Source	Availability
LAVIA-MXID	Multiple body regions	9,989	Skull, lungs, shoulders, knees, etc.	AOUINET Hospital	Available upon request
MXID Dataset [4]	Multiple body regions	6,869	Skull, lungs, shoulders, knees, etc.	AOUINET Hospital	Publicly available
NIH Chest X-ray [5]	Chest conditions	100,000	Chest	National Institutes of Health	Publicly available
MURA [6]	Musculoskeletal disorders	40,561	Elbow, wrist, shoulder, etc.	Stanford University	Publicly available
JSRT [7]	Pulmonary diseases	247	Chest	Japanese Society of Radiological Technology	Limited access
RSNA Pneumonia [8]	Pneumonia detection	30,000	Chest	Radiological Society of North America	Publicly available
ChestX-ray8 [9]	Chest conditions	108,948	Chest	National Institutes of Health	Publicly available
Dataset for Automatic Region-Based Coronary Artery Disease Diagnostics Using X-Ray Angiography Images [10]	Coronary artery disease diagnosis via X-ray angiography	1,500	Coronary arteries (heart)	Popov, Amanturdieva, Zhaksylyk et al.	Freely available (CC0 license via Zenodo)
CheXmask [11]	Anatomical segmentation of chest X-rays	657,566 masks	Chest (lungs, ribs, diaphragm, etc.)	Derived from ChestX-ray8, CheXpert, MIMIC-CXR-JPG, PadChest, and VinDr-CXR	Freely available (via PhysioNet; masks only)
ISIC Skin Lesion [12]	Skin lesions	33,126	Skin	International Skin Imaging Collaboration	Publicly available
INbreast [13]	Breast cancer screening	410	Breast	University of São Paulo	Limited access
CAMELYON-16 [14]	Breast cancer detection	23,916	Breast	Radboud University Medical Center	Limited access
Open-I [15]	Multimodal imaging	7,470	Multiple regions	National Library of Medicine	Publicly available
LIDC-IDRI [16]	Lung nodule detection and classification	1,018	Lungs	National Institutes of Health	Limited access

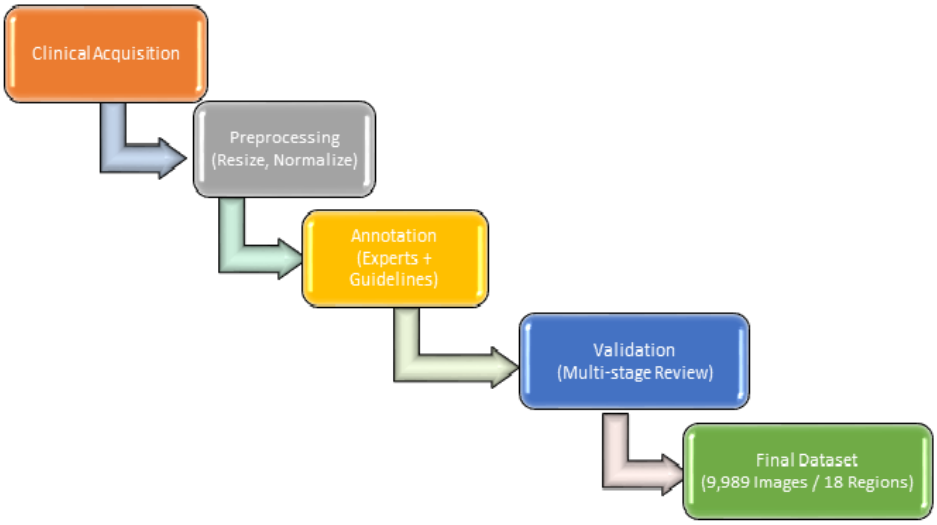


Figure 1: Process of data acquisition and construction.

automated, eliminating the need for calibration and preprocessing, therefore obviating the requirement for an extended duration of human training.

The whole dataset had an anonymization process done by removing patients' names and any other personally identifiable information to protect the confidentiality of patients. In quality control, the first part of the dataset was evaluated, and 152 corrupted or incomplete images were removed from it to guarantee reliability and validity in the dataset. This rigid stance in identifying and weeding out flawed images preserves the overall integrity of the dataset, minimizes potential biases, and increases the reliability of any successive research and experimental efforts.

A rigorous procedure was then applied to isolate and eliminate duplicate images in order to maintain quality and minimize duplication within the dataset. We isolated and eliminated duplicate images systematically, ensuring that in the dataset, there was only one copy of each unique image. More than a hundred irrelevant photos were filtered out from the main and side collections of the gathered datasets. Figure 2 illustrates the detailed preparation workflow for the LAVIA-MXID dataset, emphasizing its methodical organization and high-quality annotation standards. This makes the LAVIA-MXID a valuable resource for advancing AI-driven innovations in medical imaging. The medical body parts are collected and curated by experienced radiologists and medical experts. All the images are clearly labelled with their corresponding body regions, which will allow very accurate measurement of the different techniques of medical analysis.

Figure 3 illustrates some samples of the streamlined LAVIA-MXID Collection. The dataset is very carefully divided into training, validation, and test sets in a way that the representation of body parts is appropriate in all subsets so that models could be created and evaluated without bias.

The dataset was reviewed by radiologists and experts to verify the accuracy of image classifications, ensuring that each image was accurately classified to its corresponding anatomical region by identifying and correcting the misclassifications to maintain the dataset's integrity.

4 Dataset labeling

To ensure the dataset's quality and accuracy, a rigorous label verification process was undertaken. Each image and its corresponding label were reviewed to confirm their accuracy. Any discrepancies were corrected, and the causes of errors were documented. The process was systematic and thorough, continuing until all images were verified.

Initially, the dataset was screened using an automated program to detect labeling mismatches by comparing



Figure 2: Detailed Preparation Process for the Proposed LAVIA-MXID Dataset [4].



Figure 3: Sample Images Representing the Dataset [4].

image names and assigned labels. After this preliminary check, detailed manual inspections were conducted to resolve errors and refine the dataset structure as needed.

To ensure consistency and facilitate efficient management of the dataset, a standardized naming convention was implemented for each image. This systematic approach comprised the following components:

- **PatientID:** A unique identifier assigned to each patient, allowing for easy retrieval of records and reference to multiple X-ray images of the same individual. For example, images are labelled as P00001 for Patient 01 through P05318 for Patient 05318.
- **TypeID:** A categorical descriptor used to classify images based on the body part or health-related condition. Each body part is assigned a specific code: 01 → Abdomen, 02 → Dental, 03 → Forearm, 04 → Ankles, 05 → Cervical Spinal Column, 06 → Lumbar Spine, 07 → Elbow, 08 → Skull, 09 → Fingers, 10 → Upper Arm, 11 → Femur, 12 → Knee, 13 → Leg, 14 → Toes, 15 → Pelvic-Basin, 16 → Feet, 17 → Wrist, 18 → Lung.
- **Image Count:** The total number of images captured for a given patient, which may vary from 1 to 12 or more, depending on the patient and the clinical requirements.
- **Gender:** This demographic attribute is encoded as: 0 → Male, 1 → Female.
- **Serial Number:** A unique numerical code assigned to each image to ensure sequential organization and traceability within the dataset. For instance, images are numbered from 00001 through 09960.

This structured labeling system enhances the dataset's usability for research, analysis, and algorithm development. The careful annotation, supported by radiologists and medical experts, ensures that the dataset meets high standards of reliability and quality.

The dataset includes metadata annotations indicating the anatomical region and the patient's gender. It does not contain pathology or disease-specific labels.

The LAVIA-MXID dataset is constructed for the different image analysis tasks, not only for the image compression task, classified based on body part classes, and verified for consistency. Nevertheless, it does not include annotations for specific pathological conditions due to the support of boarder range of image analysis tasks rather than being limited to pathology-specific studies.

Medical imaging employs various techniques to capture diagnostic images of the human body, including X-ray imaging, which is widely used to visualize internal structures. Our LAVIA-MXID dataset has been meticulously classified into 18 distinct body parts and distinguished males from females, as shown in Figure 4 and Figure 5, including the lungs, abdomen, teeth, skull, femur, upper arm, knee, fingers, cervical spinal column, lumbar spine, ankles, feet, wrist, forearm, elbow, pelvic basin, toes, and pelvis. Each X-ray image in the dataset not only captures detailed anatomical structures but also provides insights into diagnosing a wide range of medical conditions. For instance, chest and limb bone X-rays play a crucial role in identifying specific health issues related to these areas.

The dataset exhibits variation in sample counts across anatomical regions and a moderate imbalance between male and female samples. Such distributions may influence model learning and generalization. Future work could explore class-balancing strategies (e.g., resampling, weighted loss functions) to mitigate potential biases.

Each image in the dataset is also labelled based on the gender of the individual, classified as either Male (denoted as "0") or Female (denoted as "1"). Table 2 showcases a comparative summary of the number of samples in our LAVIA-MXID dataset alongside the MXID dataset, categorized by body part and gender distribution.

Although the lack of the pathological annotations in LAVIA-MXID dataset, its design supports several medical imaging tasks, such as classification or detection. In this study, the main focus was on image compression emphasizing image quality preservation for more accurate diagnosis, which is crucial for clinical interpretation. Future improvements may incorporate detailed pathology annotations for more applicability across diagnostic domain datasets for gender classification and body part-based analysis.

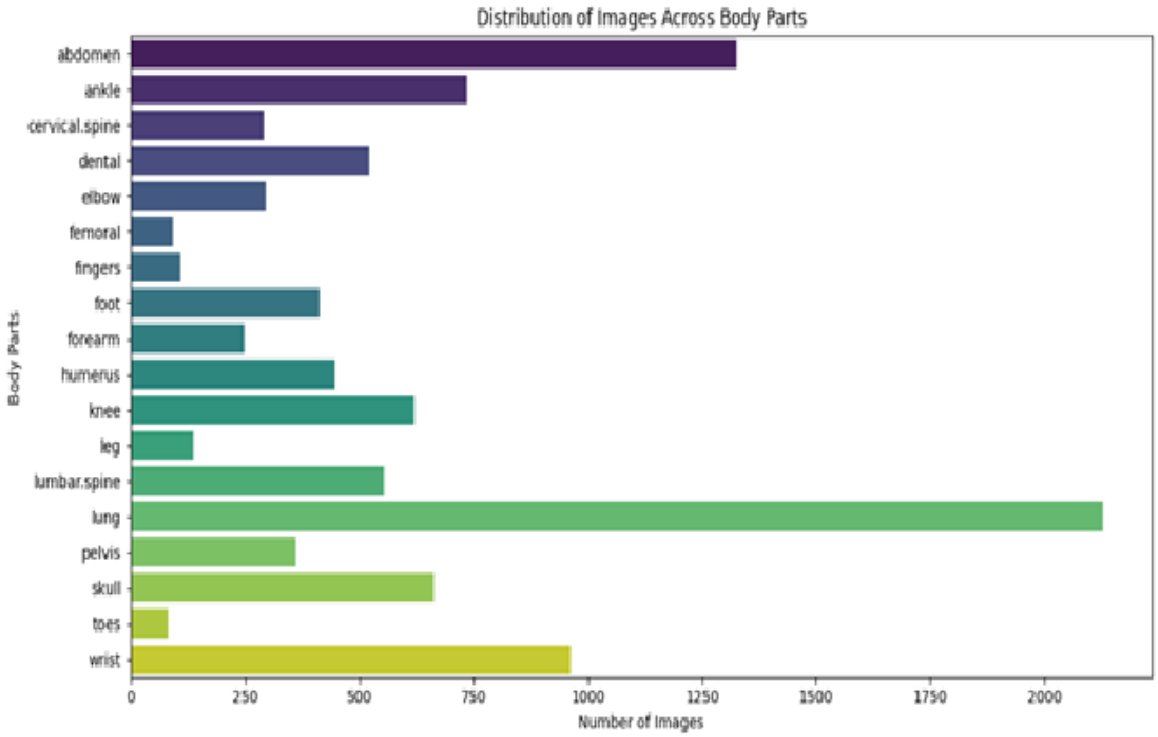


Figure 4: MXID dataset distribution across the 18 body parts classes.

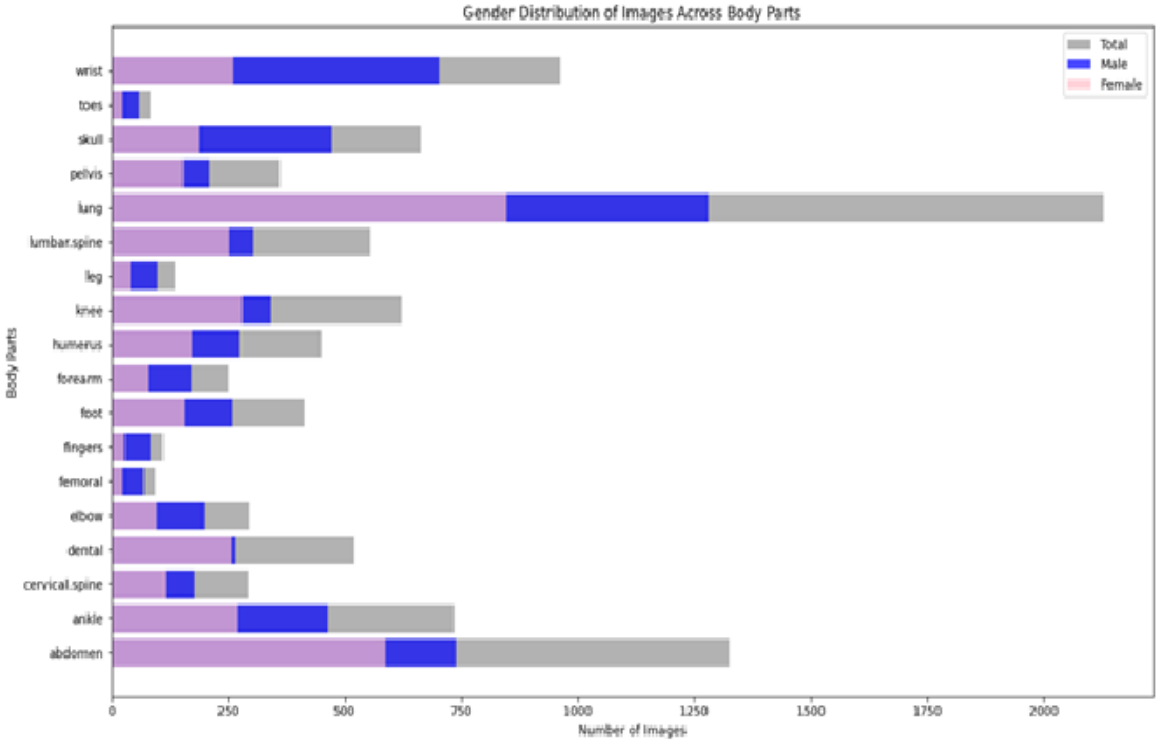


Figure 5: MXID dataset distribution across the 18 body part classes-based gender classification.

Table 2: Number of sample images from the proposed LAVIA-MXID dataset compared to MXID dataset [4], divided up by body part and gender.

N°	Body part	Dataset	Male	Female	Total	N°	Body part	Dataset	Male	Female	Total
1	Abdomen	LAVIA-MXID	738	587	1325	2	Dental	LAVIA-MXID	265	256	521
		MXID	588	476	1064			MXID	115	133	248
3	Forearm	LAVIA-MXID	171	78	249	4	Ankle	LAVIA-MXID	463	272	735
		MXID	127	52	179			MXID	191	112	303
5	Cervical Spine	LAVIA-MXID	177	115	292	6	Lumbar Spine	LAVIA-MXID	303	252	555
		MXID	75	59	134			MXID	152	116	268
7	Elbow	LAVIA-MXID	198	97	295	8	Skull	LAVIA-MXID	473	189	662
		MXID	120	68	188			MXID	406	172	578
9	Fingers	LAVIA-MXID	83	27	110	10	Upper arm	LAVIA-MXID	275	173	448
		MXID	61	19	80			MXID	187	112	299
11	Femoral	LAVIA-MXID	69	24	93	12	Knee	LAVIA-MXID	341	280	621
		MXID	49	16	65			MXID	211	189	400
13	Leg	LAVIA-MXID	98	39	137	14	Toes	LAVIA-MXID	59	23	82
		MXID	79	24	103			MXID	39	11	50
15	Pelvic-Basin	LAVIA-MXID	209	152	361	16	Feet	LAVIA-MXID	259	154	413
		MXID	135	93	228			MXID	256	153	409
17	Wrist	LAVIA-MXID	702	260	962	18	Lung	LAVIA-MXID	1280	848	2128
		MXID	455	168	623			MXID	984	666	1650

To enhance the development and evaluation of deep learning models, the LAVIA-MXID dataset underwent automated multiclass splitting into training, testing, and validation subsets. This split was carefully designed to maintain proportional representation across all class labels (body parts and genders). The allocation of the dataset is as follows:

- 60% for the training set,
- 20% for the testing set, and
- 20% for the validation set.

his balanced approach ensures that each subset remains representative of the overall dataset, minimizing bias and improving the robustness of model training and evaluation. The detailed breakdown of the dataset's distribution, including training, testing, and validation sets for each body part, is presented in Table 3.

5 Dataset for patient identification

Medical imaging provides immense potential for enhancing diagnostic accuracy by offering a comprehensive view of a patient's health when images from multiple body regions are available. Such multiregional imaging allows healthcare providers to gain deeper insights, facilitating a holistic assessment of a patient's condition. By examining radiographs from various anatomical regions, patterns, relationships, and trends can be identified, further aiding in diagnosis and the formulation of personalized treatment strategies.

The LAVIA-MXID dataset comprises imaging data from 5,318 unique patients, creating a rich, multibody-part dataset that can be applied to machine learning and deep learning techniques for improved medical diagnosis. This dataset supports the development of evidence-based decision-making processes for personalized healthcare.

Table 3: LAVIA-MXID dataset distribution into training, testing, and validation sets, compared with the MXID dataset [4].

Body Part	**MXID Dataset**				**LAVIA-MXID Dataset**			
	Train	**Test**	**Validation**	**Total**	**Train**	**Test**	**Validation**	**Total**
ABDOMEN	659	307	98	1064	913	272	140	1325
DENTAL	150	71	27	248	352	111	58	521
FOREARM	106	56	17	179	165	65	19	249
ANKLES	181	93	29	303	520	148	67	735
CERVICAL SPINAL COLUMN	79	43	12	134	210	55	27	292
LUMBAR SPINE	158	81	29	268	395	103	57	555
ELBOW	122	55	11	188	209	48	38	295
SKULL	338	171	69	578	448	134	80	662
FINGERS	48	23	10	81	73	24	13	110
UPPER ARM	188	81	30	299	316	88	44	448
FEMUR	37	26	2	65	72	11	10	93
KNEE	232	131	36	399	437	120	64	621
LEG	67	27	9	103	102	24	11	137
TOES	29	13	7	49	56	17	9	82
PELVIC-BASIN	138	58	32	228	259	67	35	361
FEET	254	112	44	410	307	54	52	413
WRIST	383	176	64	623	682	188	92	962
LUNG	963	517	170	1650	1537	403	188	2128
Total	**4132**	**2041**	**696**	**6869**	**7053**	**1932**	**1004**	**9989**

To ensure systematic organization and identification, all image files in the dataset are labelled using a consistent naming convention: PatientID_TypeID-ImageCount_Gender_SerialNumber. This naming scheme encodes the following information for each image.

A sample file in the dataset (illustrated in Fig. 6) is labelled as follows:

- Patient Identifier: P00051 (indicating the 51st patient).
- Type Identifiers: 06, 15, 17, 03, 08, 18, 01, 12 (indicating multiple body parts, including Lumbar Spine, Pelvic Basin, Wrist, Forearm, Skull, Lung, Abdomen, and Knee).
- Image Count: 10 images captured during a single session.
- Gender: 0 (denoting a male patient).
- Serial Number: A unique image identifier within the range (e.g., 00075 to 00085).

This specific patient, P00051, underwent X-ray imaging for eight different body parts, as represented in the dataset. Such comprehensive multiregional imaging facilitates a deeper understanding of the patient's overall health and supports the detection of systemic health patterns across different anatomical structures.

Figure 6 demonstrates how a single patient may have multiple associated X-ray images captured from distinct body parts. This standardized dataset structure not only organizes multiregional imaging data effectively but also ensures its compatibility with machine learning frameworks for advanced analytics in medical imaging.

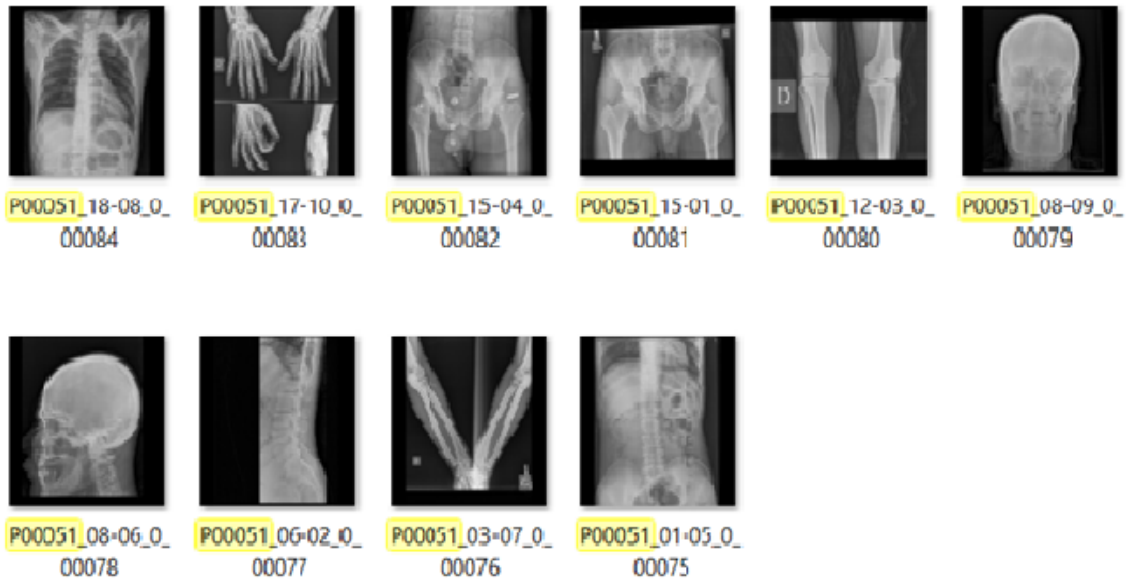


Figure 6: Patient Identification - Demonstrating the concept that a single patient may be associated with diverse images captured from distinct body parts [4].

5.1 Ethical Approval and Patient Consent

The data used in this study were collected in compliance with institutional ethical standards. Data collection was approved by the Institutional Review Board of Larbi Tebessi University. All patient data were anonymized prior to use, and informed consent was obtained from all participants.

The dataset used in this study is retrospective and fully anonymized, and patient consent was obtained and anonymization protocols strictly followed.

6 Experimental Results

To evaluate the performance of the proposed lossy image compression methods, a comprehensive assessment methodology incorporating four key metrics was employed: PSNR (Peak Signal-to-Noise Ratio), MSE (Mean Squared Error), MS-SSIM (Multi-Scale Structural Similarity Index). These metrics provide a thorough evaluation of the quality and efficiency of the compression techniques.

While the current study focuses on distortion-based metrics (cited above), we acknowledge that bitrate analysis and comparison with traditional compression standards (e.g., JPEG, JPEG2000) represent important directions for future work.

6.1 Implementation Details (Hyperparameters Configuration)

Table 4 presents the hyperparameter configurations employed across the deep learning models evaluated in this study. These include the initial learning rate, batch size, number of epochs, total trainable parameters, choice of optimizer, loss function, and activation functions. The selection of these parameters plays a crucial role in enhancing model performance. The table details the specific configurations adopted for Autoencoders (AE), Convolutional Neural Networks (CNN), Variational Autoencoders (VAE), Deep Convolutional Autoencoders (DCAE), and Convolutional Autoencoders (CAE).

It's important to notice that the evaluated architectures in this study—AE, CNN, VAE, and DCAE—are standard deep learning models. They are presented to establish baseline performance benchmarks on the LAVIA-MXID dataset, rather than to propose a new compression method.

Table 4: Hyperparameter Settings across AE, CNN, VAE, DCAE, and CAE models [19].

Technique Used	Initial Learning Rate	Batch Size	Epochs	Trainable Parameters	Optimizer	Error Function	Activation Function
Autoencoder (AE)	0.001	8	Fixed in 50	16859456	Adam Optimizer	Mean Square Error (MSE)	RELU and Sigmoid
Convolutional Neural Network (CNN)	0.001	4	Fixed in 50	1008449	Adam Optimizer	Mean Square Error (MSE)	ReLU
Variational Autoencoder (VAE)	0.001	8	Fixed in 100	–	Adam Optimizer	Mean Square Error (MSE)	–
Deep Convolutional Autoencoder (DCAE)	0.001	4	Fixed in 100	12156609	Adam Optimizer	Mean Square Error (MSE)	ReLU and PReLU

The experiments were conducted using Python 3.7.16, TensorFlow 2.7.0, and the Keras API, implemented on an NVIDIA GPU system with 16 GB of RAM. During training, the best-performing outputs and the optimized weights and biases of the enhanced DCAE model were preserved using Keras callbacks such as `ModelCheckpoint`. To highlight the autoencoder's ability to learn meaningful latent representations from medical images, the model was trained under different conditions, such as 50 and 100 epochs, while maintaining consistent training parameters. This allowed for an analysis of performance progression over time. The subsequent section presents the reconstruction results obtained using the model, initially trained on our carefully curated X-ray dataset.

6.2 Experimentation

A total of 9,989 medical images from the proposed LAVIA-MXID data set, in PNG format (256×256 pixels), were used to assess compression methods. Due to limited GPU memory, a batch size of four was used and the images were normalized to a pixel value range of [0, 1].

All images were resized to 256×256 pixels to ensure computational feasibility and consistent input across models. However, this reduction may result in the loss of fine anatomical details, and the findings should be interpreted with this limitation in mind. Future work will explore higher-resolution evaluations.

Several learning methods were tested using deep learning-based autoencoders. The performance of these methods was evaluated using the metrics mentioned above, and the results are summarized in Table 5.

Table 5: Result of deep learning method for image compression using our LAVIA MXID dataset compared to MXID dataset [4].

Method	MXID Dataset				LAVIA-MXID Dataset			
	Epochs	MSE	PSNR	MS-SSIM	Epochs	MSE	PSNR	MS-SSIM
AE	35	0.0140	37.61	0.61	48	0.0120	38.38	0.63
VAE	25	–	21.64	0.75	29	–	21.00	0.73
CNN	21	0.0060	41.43	0.77	35	0.0060	41.00	0.74
DCAE	30	0.0002	46.78	0.99	74	0.0001	47.00	0.99

The comparative analysis of the results across several epochs of the different models, namely, AE, VAE, CNN, and DCAE, on the MXID version 1 and 2 demonstrates a marginal improvement with the extended dataset (version 2); however, it reflects an enhanced image quality with a reduction in terms of reconstruction artifacts. The following section provides a detailed comparative discussion across the four models, namely Autoencoder (AE), Deep Convolutional Autoencoder (DCAE), Convolutional Neural Network (CNN), and Variational Autoencoder (VAE). In addition, analyzing the models' performance based on the evaluation metrics, including MSE, PSNR, and MS-SSIM, Figure 7 highlights the consistent improvement in image quality for AE and DCAE on the LAVIA-MXID dataset, with DCAE ach0ieving the highest reconstruction quality overall. Also, the discussion examines the impact of dataset versions on models' performance.

While the results in Table 5 show significant improvement with models such as DCAE, negligible improvement was observed with others like AE and CNN. This is attributed to architectural limitations lighter models just lack the depth and capacity to learn the complex spatial features in X-ray images. Also, VAE, although a strong generative model, axiomatically produces somewhat blurred output because of its probabilistic decoding procedure. DCAE, with deeper convolutional layers, is capable of preserving more detailed structural information and achieving better performance in MSE, PSNR, and MS-SSIM.

6.3 Model-Specific Comparisons

For the auto-encoder, there exists an improvement from MXID to LAVIA-MXID dataset, although the model may struggle with complex feature extraction in comparison to the DCAE model. Moreover, the DCAE model presented an improved result in PSNR and MS-SSIM due to its deep architecture on the LAVIA-MXID dataset. On the other hand, the CNN model didn't perform as well as the AE and DCAE models, representing a decent result in MSE and PSNR values. Conversely, this model has the potential to yield improved results for feature extraction and classification tasks. Finally, VAE is effective when dedicated for probabilistic modeling; it also has a decent result in PSNR and MS-SSIM values, indicating a subtle artifact, due to the variability introduced by the stochastic properties of the model's outputs.

The results demonstrate that the DCAE model outperformed the AE, VAE, and CNN in terms of PSNR and MS-SSIM, which indicates higher maintenance of structural details quality. Due to its deep convolutional architecture, the DCAE model extracts features and reconstructs the image efficiently, which proves that the model is well suited for medical image compression. Moving to the AE model, its simpler architecture outperformed the VAE model's in terms of PSNR and MS-SSIM, suggesting that even basic architectures can benefit from dataset expansion. For the CNN model, it performed higher than AE and VAE models, while it performed moderately with slightly lower MSE and PSNR values than DCAE, indicating less robustness in

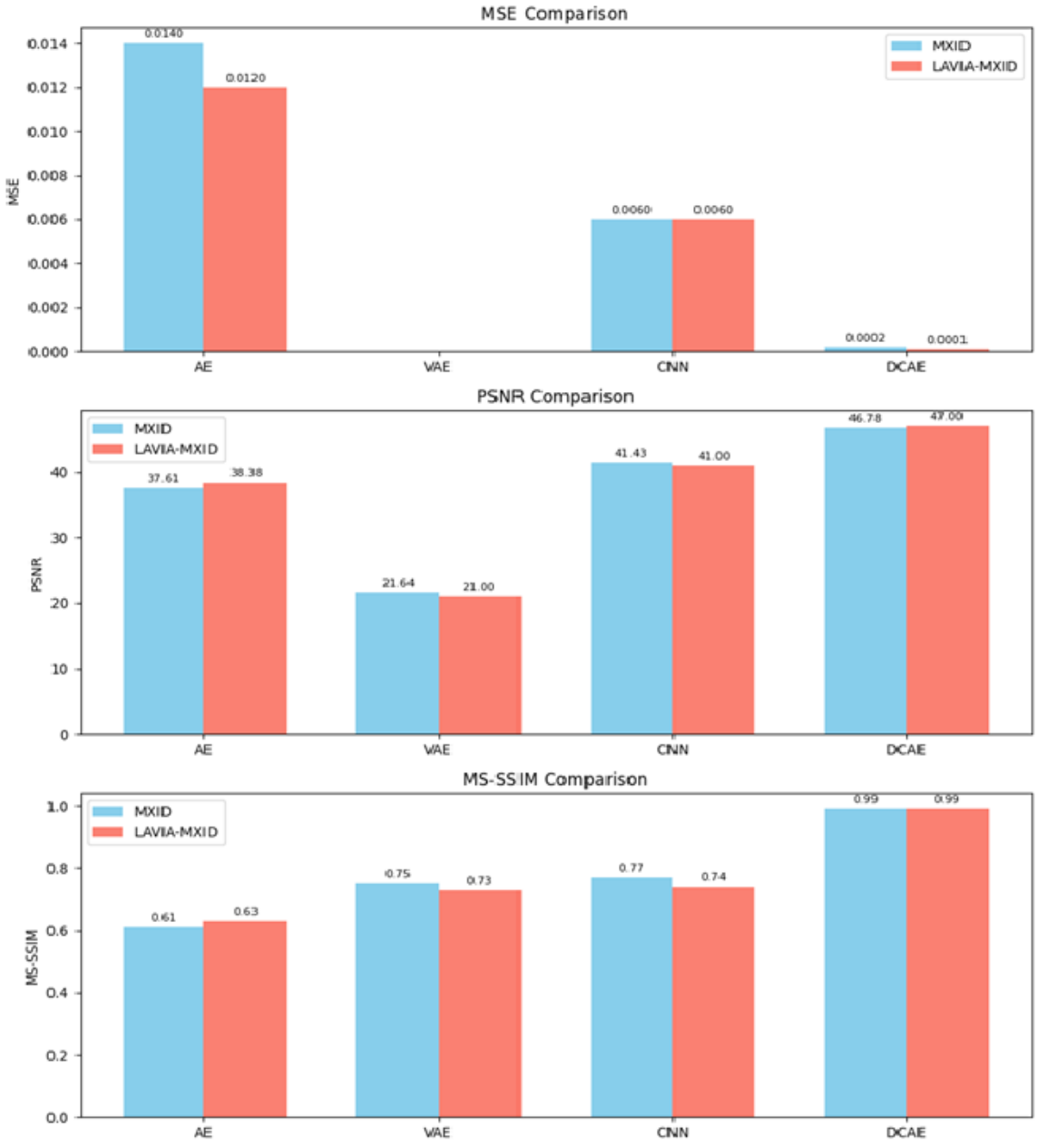


Figure 7: Comparison of the performance of four compression models—Autoencoder (AE), Variational Autoencoder (VAE), Convolutional Neural Network (CNN), and Deep Convolutional Autoencoder (DCAE)—on the MXID dataset and the LAVIA-MXID dataset. Evaluation metrics employed are Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Multi-Scale Structural Similarity Index (MS-SSIM).

preserving structural details. The VAE model exhibited reasonable PSNR and MS-SSIM, as demonstrated in Figure 7, as a result of the stochastic nature of its latent space, which introduced subtle artifacts impacting structural similarity. As a conclusion, the experiments confirm that deep learning models specifically DCAE revealed a high-quality reconstructed image quality when trained on a sufficient and well-preprocessed dataset.

In contrast to hybrid approach proposed by Al-Khafaji & Ramaha [20], where traditional image processing was supplemented with deep learning for better texture recovery, our approach relies solely on compression using convolutional autoencoders with low network complexity and facilitating end-to-end deep learning-optimized optimization. While their model yields very high PSNR (from 50.36 dB), our approach optimizes compression rate against reconstruction quality for very heterogeneous radiological data from LAVIA-MXID. Moreover, our method was trained and validated on a publicly available dataset with almost 10,000 images across 18 body regions, unlike some of the related works which used limited access datasets or single-modality images.

In contrast, the quantum-enhanced model by Balasubramani et al. [21] offers impressive compression ratios, especially for MRI and CT-SCAN images. However, the method is constrained by the current limitations in quantum hardware and does not report performance on X-ray datasets comparable to ours. Our model offers practical applicability using conventional hardware, making it more feasible for real-world clinical deployment in resource-limited environments.

The industrial-focused study by Hejazi et al. [22] addresses compression in XCT data using VQ-VAE and D-CNN architectures. Although it introduces a novel image quality metric (Mean Squared Laplacian Error), its relevance to clinical diagnostic preservation is limited. Our method, by contrast, directly targets diagnostic X-ray images, prioritizing clinical interpretability and diagnostic feature retention.

In brief, while previous work shows new approaches to image compression, our research differs in the use of a deep learning approach for multi-class medical X-ray compression with training on a heterogeneous and curated clinical dataset (LAVIA-MXID). Our assessments show competitive performance in image quality preservation and compression efficiency, in addition to enabling broader generalization across anatomical regions.

6.4 Dataset's Version Comparisons

The extended dataset provided a more varied samples that enabled models to have improved generalization on all metrics. However, the marginal improvements in VAE and CNN may indicate that the dataset V1 has been sufficiently representative for these models. On the other hand, DCAE and AE achieved a slight improvement on LAVIA-MXID dataset in all metrics. For the CNN and VAE demonstrated limited progress compared to AE and DCAE, indicating potential limitations in terms of model's architectures or minimal performance enhancements despite the addition of more data.

While the proposed DCAE of [4] presented promising results using the LAVIA-MXID dataset, this dataset lacks pathology annotations, limiting the ability to assess the impact of compression on disease-specific characteristics. Moreover, all the images were resized to a fixed resolution of 256×256, which may lead to a loss of fine details, specifically in high-resolution diagnostic contexts. Lastly, compression performance was tested under limited GPU resources, which constrained batch size and training scalability. Future work will attempt to overcome these limitations through the incorporation of multiple image types, pathology annotations, and dynamic resolution management, along with the use of more advanced architectures like GANs for further optimization.

7 Conclusion

Datasets related to X-ray medical imaging are critical for accurate diagnosis of conditions and require huge and good-quality X-ray sets to train deep learning algorithms that could be clinically useful. Such algorithms enhance image interpretation, have sped up the process, and make the diagnosis more reliable.

Special focus is given to peculiarity and problems during medical imaging pattern recognition: scale of dataset, quality of dataset, errors in labeling the dataset, data imbalance, etc. Solving these problems is very

crucial for making the application of deep learning algorithms in medicine a productive and robust business.

The LAVIA-MXID extended dataset from the Laboratory of Vision and Artificial Intelligence (LAVIA) is one of these types of advances, offering solutions for the deficiencies of the MXID dataset. This research shows especially deep learning approaches for medical image compression, which aims to provide solutions for handling the complexities of medical imaging data.

This study introduces the LAVIA-MXID dataset, a comprehensive medical imaging dataset designed for diverse applications, including image compression. By training multiple deep learning models, including a novel Deep Convolutional Autoencoder (DCAE), we achieved superior results in compression quality. Overall, the novelty of this work lies in the applied methodology and the dataset design, which establish a new benchmark for evaluating medical image compression using deep learning models. Although the LAVIA-MXID dataset lacks pathology annotations it covers high-quality imaging across diverse anatomical regions, which is crucial for the different imaging tasks. It's to know that **original images were acquired at 1024×1024 pixels. For experimental consistency and computational efficiency, all images were downsampled to 256×256 pixels. This preprocessing step may reduce the visibility of fine anatomical structures, and results should be interpreted accordingly.** To extend its utility to disease-specific applications, future work will entail the incorporation of pathological labeling through expert annotation campaigns; this improvement will make the dataset useful for clinical diagnosis research for specific chronic and acute conditions.

The proposed dataset and methods outperform existing approaches and emphasize the importance of systematic dataset structuring in medical research. Future work will focus on addressing the underperformance of certain models by:

- The design of a hybrid technique combining the strength of multiple learning paradigms.
- The incorporation of attention mechanisms to focus on diagnostically important regions for higher compression quality.
- To improve image reconstruction, Generative Adversarial Networks (GANs) will be explored.

References

[1] A. Fettah, R. Menassel, and A. Gattal, "Machine Learning for Medical Image Analysis: A Survey," *Lecture Notes in Networks and Systems*, vol. 713 LNNS, pp. 148–164, 2023.

[2] R. Najjar, "Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging," *Diagnostics*, 2023.

[3] S. Hussain et al., "Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review," 2022.

[4] A. Fettah, R. Menassel, A. Gattal, and A. Gattal, "Convolutional Autoencoder-Based medical image compression using a novel annotated medical X-ray imaging dataset," *Biomed Signal Process Control*, vol. 94, p. 106238, 2024.

[5] R.W. Filice et al., "Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset," 2019.

[6] P. Rajpurkar et al., "MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs," 2017.

[7] J. Shiraishi et al., "Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule: Receiver Operating Characteristic Pulmonary Nodules," 2000, pp. 71–74.

[8] I. Pan, A. Cadrin-chênevert, and P.M. Cheng, "Tackling the Radiological Society of North America Pneumonia," 2019, pp. 1–7.

[9] X. Wang et al., "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *CVPR*, 2017.

[10] M. Popov et al., "Dataset for Automatic Region-based Coronary Artery Disease Diagnostics Using X-Ray Angiography Images," *Sci Data*, vol. 11, 2024.

[11] N. Gaggion et al., "CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images," *Sci Data*, vol. 11, 2024.

[12] R. Ali, R.C. Hardie, M.S. De Silva, and T.M. Kebede, "Skin Lesion Segmentation and Classification for ISIC 2018 by Combining Deep CNN and Handcrafted Features," 2018.

[13] I.C. Moreira et al., "INbreast: Toward a Full-field Digital Mammographic Database," *Acad Radiol*, vol. 19, pp. 236–248, 2012.

[14] M. Jaber, "A deep learning-based iterative digital pathology annotation tool," 2021.

[15] D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," *JAMIA*, vol. 23, pp. 304–310, 2016.

[16] S.G. Armato et al., "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans," *Med Phys*, vol. 38, pp. 915–931, 2011.

[17] M.R. Lone, "A high speed and memory efficient algorithm for perceptually-lossless volumetric medical image compression," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, pp. 2964–2974, 2022.

[18] C. Tensmeyer and T. Martinez, "Historical Document Image Binarization: A Review," *SN Comput Sci*, vol. 1, p. 173, 2020.

[19] A. Fettah, Thesis, Echahid Cheikh Larbi Tebessi University, 2024.

[20] M. Al-Khafaji and N.T.A. Ramaha, "Hybrid deep learning architecture for scalable and high-quality image compression," *Sci Rep*, vol. 15, 2025.

[21] B. Subbiyan et al., "A Quantum-Enhanced Artificial Neural Network Model for Efficient Medical Image Compression," *IEEE Access*, vol. 13, pp. 31809–31828, 2025.

[22] B. Hejazi, K. Chand, T. Fritsch, and G. Bruno, "D-CNN and VQ-VAE Autoencoders for Compression and Denoising of Industrial X-ray Computed Tomography Images," 2025.