ITUPublications

International Telecommunication Union
Standardization Sector

ITU-T Focus Group Deliverable

(09/2023)

Focus Group on Artificial Intelligence for Health

(FG-AI4H)

FG-AI4H DEL10.12

Topic Description Document for the Topic Group on AI for radiology (TG-Radiology)

ITU-T FG-AI4H Deliverable

DEL10.12 – FG-AI4H Topic Description Document for the Topic Group on AI for radiology (TG-Radiology)

Summary

Radiology has been essential to accurately diagnosing diseases and assessing responses to treatment. The challenge, however, lies in the shortage of radiologists globally. As a response to this, a number of artificial Intelligence (AI) solutions are in development. The challenge AI radiological solutions face, however, is the lack of a benchmarking and evaluation standard, and the difficulties of collecting diverse data to truly assess the ability of such systems to generalize and properly handle borderline cases.

This topic description document specifies a standardized benchmarking for AI-based symptom assessment. It covers all scientific, technical and administrative aspects relevant for setting up this benchmarking and describes a radiograph-agnostic platform and framework that would allow any AI radiological solution to be assessed on its ability to generalize across diverse geographical location, gender and age groups.

Keywords

Artificial intelligence, benchmarking, clinical relevance, data audit, data quality, ethics, health, overview, radiology, regulations, topic description, topic groups.

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Change Log

This document contains Version 1 of the Deliverable DEL10.12 entitled *FG-AI4H Topic Description Document for the Topic Group on AI for radiology (TG-Radiology)* approved 15 September 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

Editor: Darlington Ahiale Akogo
minoHealth AI Labs
Ghana
Tel: +233 50 404 9188
Email: darlington@gudra-studio.com

Contributors:

Vincent Appiah
minoHealth AI Labs
Ghana
E-mail: appiahv@rocketmail.com

Xavier Lewis-Palme
minoHealth AI Labs
United States of America
E-mail: xpalm001@odu.edu

Issah Abubakari Samori
minoHealth AI Labs
Ghana
E-mail: issahsamori@gmail.com

Camilo Sotomayor
IA·TRad Chile & Clinical Hospital University of Chile
Chile
E-mail: camilosotomayor@ug.uchile.cl

Andrew Murchison
Oxford University Hospitals NHS Foundation Trust
United Kingdom
E-mail: agmurchison@gmail.com

Benjamin Dabo Sarkodie
Euracare Advanced Diagnostic Center
Ghana
E-mail: bensarkodie@gmail.com

Judy Wawira Gichoya
Emory University School of Medicine
United States of America
E-mail: judywawira@emory.edu

Edson Mintsu Hung
Universidade de Brasília
Brazil
E-mail: mintsu@unb.br

Andrey O. O. dos Reis
Universidade de Brasília
Brazil
E-mail: xpalm001@odu.edu

Renam Castro da Silva
Universidade de Brasília
Brazil
E-mail: renam.silva@smt.ufrj.br

Saul Calderon Ramirez
De Montfort University (Metrics deliverables)
Brazil

Pierre Padilla-Huamantinco
Universidad Peruana Cayetano Heredia
E-mail: pierre.padilla.h@upch.pe

Dominick Romano
Drainpipe.io
United States of America
E-mail: dom@drainpipe.io

Alessandro Sabatelli
Braid.Health
United States of America
E-mail: alessandro@braid.health

Daniel Hasegan
Braid.Health
United States of America
E-mail: daniel@braid.health

Table of Contents

ITU-T FG-AI4H Deliverable

DEL10.12 – FG-AI4H Topic Description Document for the Topic Group on AI for radiology (TG-Radiology)

1 Introduction

An estimated 3.6 billion diagnostic medical examinations, such as x-rays, are performed worldwide every year. Advances in radiology technology have improved illness and injury diagnosis and treatments. These radiological procedures include x-rays, mammograms, ultrasound, positron emission tomography (PET) scans, magnetic resonance imaging (MRI) scans and computed tomography (CT) scans. They are used mainly in dealing with a broad range of non-communicable or chronic diseases. These are primarily cardiovascular diseases, cancer, chronic respiratory diseases and diabetes. Radiology has helped in the rapid non-invasive screening of conditions such as breast cancer, which reduces the mortality rate, especially with early detection. 33 Million screening mammography examinations are performed each year in the USA alone. Arleo et al. [89] found that recommended annual screening starting at age 40 results in a nearly 40% reduction in deaths due to breast cancer. Simple radiological procedures like ultrasound can reduce the need for surgical interventions. In addition, although clinical judgement may be sufficient, radiological procedures are necessary to confirm and properly evaluate the causes of many conditions and responses to treatment.

1.1 Document structure

Overview of the whole document.

1.2 Status update for meeting L

Between meeting K and L, the Topic Group on AI for Radiology onboarded three new members, Renam C. da Silva, Dominik Stosik and Bobby Bhartia. We also had a meeting 2021-04-16. During the meeting, we discussed status updates and welcomed new members. We discussed open work streams within the topic group to which our members can then lead and collaborate towards contributing. Vincent Appiah, minoHealth AI Labs, took *Existing work on benchmarking*. In contributing to this work stream, he reviewed published papers on benchmarking from regulators, clinicians, and AI developers. He then contributed a summary of these papers that appears in clause 1.9. Darlington Akogo, the Topic Driver, also summarized the work being done by the NHS AI Lab in benchmarking AI solutions for corona virus disease-2019 (COVID-19) that appears in clause 1.11. Edson Minstu, Renam C. da Silva, and Andrey O. O. dos Reis updated their experiments on assessing the effects of various compression techniques and ratio, and scaling on data validity during the AI model testing. They compared the performance of various Joint Photographic Experts Group (JPEG) compression ratios and portable network graphic (PNG), and contributed the results in clause 3.1.1.

1.3 Status update for meeting M

Towards meeting M, Samori Issah, minoHealth AI Labs, contributed an overview for ethical considerations under AI for radiology (clause 1.6). Judy Wawira Gichoya, Emory University School of Medicine, contributed clause 1.6.2 on a study conducted by her and her colleagues that demonstrated that AI models have unintended capacity to identify and differentiate between various races from the image data alone across various imaging modalities, even though there are no known imaging biomarker correlates for racial identity. They then highlight how this present biases and dangerous outcomes when such AI systems are deployed without oversight. They also share recommendations. Edson Minstu, Renam C. da Silva, and Andrey O. O. dos Reis, Universidade de Brasília, expanded their experiments to cover a brain tumour image classification task. The results further demonstrate the influence of the compression artefacts in medical image classification.

In order to evaluate image compression in the scenario, they developed a library that calculates a set of metrics, such as accuracy, sensitivity, specificity and *F*-score, for testing different compression and downsizing in a dataset. Darlington Akogo, minoHealth AI Labs expanded the list of evaluation metrics in clause 3.4 to include ten various metrics used for multi-label classification. This includes exact match ratio (EMR), Hamming loss, example-based accuracy, macro averaged accuracy, micro averaged accuracy, macro averaged precision, micro averaged precision, macro averaged recall, micro averaged recall and alpha evaluation score.

1.4 Status update for meeting R

Darlington Akogo contributed clause 3.6.3, which covers the first AI clinical study in Africa, benchmarking the performance of AI for radiology systems against radiologists. Dominick Romano contributed clause 3.1.3, which covers techniques to compress medical images of different modalities. Clause 3.1.3 also contains benchmark tests on these different techniques.

1.5 Topic description

Challenges facing radiology

Although radiology is very important, there is a shortage of radiologists globally, especially in developing countries. Liberia, for example, only has about two radiologists [99], while Ghana has 34 and Kenya 200. [101] In the UK, only one in five trusts and health boards has a sufficient number of interventional radiologists to run a safe 24/7 service to perform urgent procedures while their workload of reading and interpreting medical images has increased by 30% between 2012 and 2017. [93] There is a need for scalable and accurate automated radiological systems. Deep learning, especially in the form of convolutional neural networks (CNNs), is gaining wide attention for its ability to accurately analyse medical images, with the potential to help solve the shortage of radiologists.

Artificial intelligence in radiology

The re-emergence of artificial intelligence (AI) and deep learning, due to growth in computing power and data, has led to advancements in deep CNNs, which has allowed for breakthrough research and applications in radiology. AI and deep learning holds a lot of potential in radiology. AI can provide support to radiologists and alleviate radiologist fatigue. It can help in flagging patients who require urgent care to radiologists and physicians. Deep learning could also help increase interrater reliability among radiologists throughout their years in clinical practice. Bien et al. [90] found that Fleiss's kappa measure of interrater reliability for detecting anterior cruciate ligament tear, meniscal tear, and abnormality were higher with model assistance than without it. Deep learning has achieved performances comparable to humans and sometimes better. Liu et al. [117] analysed 14 research works done using deep learning to detect diseases via medical images, they found that on average, deep learning systems correctly detected a disease state 87% of the time – compared with 86% for healthcare professionals – and correctly gave the all-clear 93% of the time, compared with 91% for human experts. Deep learning has performed as well as radiologists and sometimes better at detecting abnormalities like pneumonia, fibrosis, hernia, oedema and pneumothorax in chest x-rays [100]. It has also been used to detect knee abnormalities via MRI at near-human-level performance. [90] Researchers have also trained deep learning models that outperformed dermatologists at detecting skin cancer. [88][94]

Research data

One key focus of deep learning radiological applications is breast cancer detection via mammograms. The curated breast imaging subset of digital database for screening mammography [91] is one key repository that is publicly available. It contains 10 239 images grouped under the labels: benign; benign without callback; and malignant. Another set of focus is the detection of thoracic conditions via chest x-rays. One publicly available chest x-ray dataset is CheXpert, [92] which contains

224 316 chest radiographs of 65 240 patients. It contains images for 12 different thoracic diseases including atelectasis, cardiomegaly, enlarged cardiomegaly, consolidation, oedema, lung lesion, lung opacity, pneumonia, pneumothorax, fracture, pleural effusion and pleural other. In addition, it contains two other observations. no finding and support devices, making 14 observations in total. The radiographs were collected at Stanford Hospital between 2002-10 and 2017-07. Another publicly available chest radiograph dataset is MIMIC-CXR, [96] which contains 371 920 chest x-rays associated with 227 943 imaging studies. Each imaging study contains a frontal view and a lateral view. MIMIC-CXR [96] dataset also contains 14 observations. There is also a chest x-ray dataset from the NIH Clinical Center [97] that contains 100 000 x-rays from over 30 000 patients, including many with advanced lung disease. The overall total is 696 236 publicly available x-ray images for 12 thoracic conditions.

Challenges facing AI in radiology

The challenge, however, remains in properly testing such systems and ensuring they work in all borderline and diverse cases radiologists encounter. Zech et al. [95] found that deep learning models that detected pneumonia on chest x-rays performed well on further data from sites they were trained on (area under curve (AUC) of 0.93–0.94) but significantly less on external data (AUC 0.75–0.89). This demonstrates the challenge of assessing the generality and scalability of deep learning systems. Though Liu et al. [117] analysed 31 587 studies, only 69 studies provided enough data to construct contingency tables, enabling calculation of test accuracy. In addition, out of those 69 studies, only 25 did out-of-sample external validations. Further, only 14 of such studies compared model performance to that of radiologists. Liu et al. [117] also realized the methodology and reporting of studies evaluating deep learning models is variable and often incomplete. This shows the need for standardization of evaluation frameworks and benchmarks for AI radiological systems. This is essential to assessing the quality of AI solutions, their readiness for deployment and the degree of autonomy they should be given.

1.5.1 Impact of benchmarking

A large number of publicly available medical image datasets exist online, and there has been a lot of research and development with such datasets. By developing frameworks that target these conditions first, the standardized benchmarking platform would be made immediately appealing to the AI healthcare research and development community. This would also help speed up the deployment of AI solutions in radiology globally. AI healthcare system developers and organizations usually have to go through the challenge of convincing health facilities to share their private data with them, such data unfortunately are not always of high quality and they usually lack the broad demographic representations needed to truly assess how well an AI system generalizes. A radiograph-agnostic benchmarking platform with data from various facilities across the globe, reviewed by a panel of experts to ensure quality and diversity, would drastically simplify the evaluation stage of such AI systems. The precision evaluation framework would help fight against demographically biased AI systems by ensuring they are tested in great detail across various groups. It would also help in the safe scaling of AI systems across different locations. The location sub-categorization of evaluation allows for geo-precision evaluation. Developers can tell how well their systems can perform within their country or first point of deployment, and should they intend to scale to neighbouring countries then eventually have it across the globe, they can tell how well their current version would perform at each point of such growth and scaling.

1.6 Ethical considerations

1.6.1 Overview

AI is the development of computer algorithms and models to perform tasks that require human-level intelligence. [1] The current trend of AI is based on machine-learning techniques that make intelligent predictions based on data. [2] A subset of machine learning algorithms, known as deep learning

algorithms, have powered most of the current advances in AI. Deep learning, as a subfield of machine learning, is the development of self-learning algorithms. These algorithms use artificial neural networks, which have millions of tuneable parameters. [3]

The complexity of these algorithms makes understanding the reasoning behind an AI model decision very difficult, thus making auditing and debugging of its process almost impossible. The ethical challenge here is that the biases AI models inherit from their training data and developers are reflected in their decisions. [4] Because these models lack transparency, it becomes difficult to correct the process that led to the biased decision. When these biased models are deployed, they reinforce the existing biases, and this can be detrimental. Studies have shown that AI models deployed in other fields have expressed biases against groups that were underrepresented in the training dataset. [5] A likely solution to the problem of bias is to train transparent algorithms on well-balanced datasets. Utilizing transparent and easily debuggable algorithms could, however, decrease the performance of these AI models. [4]

Another ethical dilemma worth considering is data ethics and data ownership. [4] Training AI models require huge amounts of data, so AI developers use patient data from healthcare institutions. A lot of discussions and concerns have sprung up around whether patient consent is needed whenever their data is used in training an AI model. Some agree that the consent of patients should be sought while others argue that developing AI models for radiology is for the greater good for which no-one's consent is needed.

There are also many unanswered questions around data ownership and how profits derived from using patient data will be shared. [4] Whoever is identified as a total or part owner of a dataset deserves a share in the profit the dataset generates. So, if it is agreed that patients own the data, then they deserve a share in the profit an AI developer will make from a model that was trained on the dataset.

Just like any technology, AI in its early stages might not be available to all people because of the uneven distribution of resources (including financial resources, computational resources and skillset). This will further exacerbate the existing inequality in society as only those with the required resources can harness the power of AI. [6]

An AI model cannot be held liable for a mistake, as some standards view an AI model as a tool. It becomes crucial to identify who is responsible for the mistakes of an AI model. Will the developer who designed the AI model, the radiologist who used it or the hospital that purchased it be responsible for any shortcomings on the path of the AI? Answering this question will force regulators to identify the key stakeholder in the AI pipeline and what their responsibilities are. [4][6]

In conclusion, AI can be a very powerful tool in the radiologist's toolbox but has a couple of ethical issues that have to be addressed first. These ethical issues have to be taken seriously (especially by regulators) in order to prepare the field of radiology for the fourth industrial revolution.

1.6.2 Reading race: AI recognizes patient's racial identity in medical images

There are no known imaging biomarker correlates for racial identity [102]; however, medical imaging AI models produce racial disparities [103]. There is potential for discriminatory harm if it is assumed that AI models are agnostic to race – understanding the relationship between race and medical imaging AI models is important [111]. An answer was sought on how AI systems could produce disparities across racial groups and determine how AI could predict race from medical images.

In this study, a large number of publicly and privately available large-scale medical imaging datasets are investigated and it was found that self-reported race is trivially predictable by AI models trained with medical image pixel data alone as model inputs. Standard deep learning methods are used for each of the image analysis experiments, training a variety of common models appropriate to the tasks. First, it is shown that AI models can predict self-reported race across multiple imaging modalities, various datasets and diverse clinical tasks (given prefix A in Table 1). The high level of performance persists during the external validation of these models across a range of academic centres and patient

populations in the USA, as well as when models are optimized to perform clinically motivated tasks. Ablations were also performed that demonstrate this detection is not due to trivial proxies, such as body habitus, age, tissue density or other potential imaging confounders for race such as the underlying disease distribution in the population (prefix B in Table 1). Finally, it is shown that the features learned appear to involve all regions of the image and frequency spectrum, suggesting that mitigation efforts will be challenging (prefix C in Table 1). Table 1 also lists brief descriptions of these experiments.

Table 1 – Reading race – Experiments, methods, and results

Experiment	Description	Results
A.1 Detection of racial identity on chest x-ray	Resnet34 [107] one-vs-all predict black, white, or Asian.	Average AUC across races of 0.974 internal validation, 0.949 external.
A.2 Detection of racial identity on hand x-ray, cervical spine x-ray, chest CT, and mammography images	Binary classification one-vs-all, black or white. For multi-slice, predictions at slice level aggregated at study level.	Average AUC per study of 0.915 internal and 0.885 external.
A.3 Train models for pathology detection and patient re-identification, evaluate on ability to predict race	DenseNet-121 [108] models to detect pathology on chest x-ray/re-identify unique patients. Removed final classifier and used model output as input on training to predict race.	Average AUC across races of 0.85.
B.1 Race detection using body habitus	Models predicting based on body mass index (BMI), presence of BMI data, and stratification of image data by body habitus.	AUC – BMI data 0.55, presence of BMI 0.52, and stratified by BMI [0.89, 0.98], [0.92, 0.99]
B.2 Tissue density analysis on mammograms	Multi-class logistic regression model to predict race black or white based on breast density and age, using one-vs-all approach.	AUC – density only 0.54, age and density 0.61.
B.3 Race detection using disease labels	Two models – predict only using disease labels and image classification only on images with 'no finding' labels.	AUC – disease labels 0.561, no finding 0.937 average across races.
B.4 Race detection using bone density	Remove bone density information by clipping bright pixels to 60% intensity, then train DenseNet-121 [108] model.	Average AUC of 0.95 across races.
B.5 Race detection using age and sex	Two models trained on split data (A1 method) – five age groups and male/female.	No significant deviation from A.1.
C.1 Frequency-domain imaging features	Four new models created on modified datasets (A1 method), low-pass filtered (LPF), high-pass filtered (HPF), bandpass filtered (BPF), notch filtered (NF).	AUC – LPF all results >0.5, >0.9 for LPF 50; HPF all results >0.5, >0.9 for HPF 100; BPF [0.75, 0.91]; NF [0.82, 0.91]
C.2 Impact of image resolution and quality	Three new models created on modified datasets (A1 method) – various resolutions and two with image perturbations.	AUC - >0/95 for 160 × 160 resolution and 0.64 for 4 × 4 images; average of 0.652 for perturbed.

Table 1 – Reading race – Experiments, methods, and results

Experiment	Description	Results
C.3 Anatomical localization	Produced saliency maps using grad-cam method, five radiologists perform qualitative evaluation. Mask regions of interest (ROIs) from maps, then test performance of A1 model on masked images. Segment lungs and train new model on lung only and lung removed images. Analysis of CT slice-by-slice error distribution for anatomical ROIs.	No finding of specific anatomical segment from qualitative evaluation or slice-by-slice CT errors. average AUC across races – masking ROI 0.82; non-lung 0.863; lung only 0.717.
C.4 Patch-based training	Train two new models (A1 methodology) on datasets – split images into 3 × 3 square cells of equal size remove one of nine cells, only use one cell.	Average AUC white vs others – cell removed 0.909; only one cell 0.796.

The result that deep learning models can trivially predict the self-reported race of patients from medical images alone is surprising, particularly as this task is not possible for human experts. This work confirms that model discriminatory performance for racial identity recognition generalizes across multiple different clinical environments, medical imaging modalities, and patient populations, suggesting that these models do not rely on hospital process variables or local idiosyncratic differences in how imaging studies are performed for patients with different racial identities. This capability is trivially learned and therefore likely to be present in many medical image analysis models, providing a direct vector for the reproduction or exacerbation of the racial disparities that already exist in medical practice.

Human oversight of AI models is of limited use to recognize and mitigate this problem. If an AI model relied on its ability to detect racial identity to make medical decisions, but in doing so misclassified all black patients, clinical radiologists (who do not typically have access to racial demographic information) would not be able to tell.

It is strongly recommended that all developers, regulators, and users who are involved with medical image analysis consider the use of deep learning models with extreme caution. In the setting of x-ray and CT imaging data, patient racial identity is readily learnable from the image data alone, generalizes to new settings, and may provide a direct mechanism to perpetuate or even worsen racial disparities that exist in current medical practice. Our findings indicates that future medical imaging AI work should emphasize explicit model performance audits based on racial identity, sex and age, and that medical imaging datasets should include the self-reported race of patients where possible to allow for further investigation and research into the human-hidden but model-decipherable information that these images appear to contain related to racial identity.

1.7 Existing AI solutions

1.7.1 Use case descriptors

To collect existing AI solutions and use cases, the following nine descriptors that would be useful have been identified:

– condition;
– medical imaging modality;
– AI task/problem description (e.g., image classification, image segmentation);
– general algorithm description (if shareable);
– project goal and current stage (if shareable);

– input structure and format;
– output structure and format;
– evaluation metrics;
– explicability and interpretability framework.

1.7.2 Collected AI solutions and use cases

1.7.2.1 minoHealth

Descriptor	Description
Condition	Pneumonia, hernia, fibrosis, atelectasis, cardiomegaly, enlarged cardiomegaly, consolidation, oedema, lung lesion, lung opacity, pneumothorax, fracture, pleural effusion and pleural other (14 different systems)
Medical imaging modality	Chest x-ray
AI task/problem description	Image classification
General algorithm description	CNNs, transfer learning
Project goal and current stage	Commercial, testing and piloting.
Input structure and format	Two dimensional (2D) image, JPEG (converted from digital imaging and communications in medicine (DICOM))
Output structure and format	Sigmoid with range 0 to 1 – 0: negative; 1: positive
Evaluation metrics	Accuracy score, receiver operating characteristic (ROC) curve and AUC score
Explicability and interpretability framework	Implementing lightweight interactive multimedia environment (LIME)

1.7.2.2 minoHealth

Descriptor	Description
Condition	Breast cancer
Medical imaging modality	Mammograms
AI task/problem description	Image classification
General algorithm description	CNNs, transfer learning
Project goal and current stage	Commercial, testing and piloting
Input structure and format	2D image, JPEG (converted from DICOM)
Output structure and format	Softmax with three classes, benign, benign without callback and malignant
Evaluation metrics	Accuracy score, ROC curve and AUC score
Explicability and interpretability framework	Implementing LIME

1.7.2.3 Braid.Health

Descriptor	Description
Condition	Atelectasis, cardiomegaly, consolidation, oedema, effusion, emphysema, fibrosis, hernia, infiltration, mass, nodule, pleural thickening, pneumonia, pneumothorax, old fracture, new fracture, scoliosis, sternotomy, enlarged cardiomedistinum, support devices, tuberculosis, bronchiectasis, foreign body (22 conditions)
Medical imaging modality	Chest x-rayay
AI task/problem description	Image classification
General algorithm description	CNNs, DenseNet-121, transfer learning, Bayesian optimization, strong augmentations
Project goal and current stage	Commercial, testing and piloting
Input structure and format	2D image, PNG (converted from DICOM)
Output structure and format	Calibrated score from 0.0 to 1.0 representing precision of data for the current distribution
Evaluation metrics	ROC curve, AUC ROC score, specificity at sensitivity
Explicability and interpretability framework	None currently

1.7.2.4 Braid.Health

Descriptor	Description
Condition	Fracture, dislocation, oedema, arthritis, osteoarthritis, spur (6 conditions)
Medical imaging modality	Foot x-ray
AI task/problem description	Image classification
General algorithm description	CNNs, DenseNet-121, transfer learning, Bayesian optimization, strong augmentations
Project goal and current stage	Commercial, testing and piloting
Input structure and format	2D image, PNG (converted from DICOM)
Output structure and format	Calibrated score from 0.0 to 1.0 representing precision of data for the current distribution
Evaluation metrics	ROC curve, AUC ROC score, specificity at sensitivity
Explicability and interpretability framework	None currently

1.7.2.5 minoHealth

Descriptor	Description
Condition	Chest_AP, Chest_LAT, Chest_PA, Foot_AP, Foot_LAT, Foot_OBL, Ankle_AP, Ankle_LAT, Ankle_OBL, Hand_LAT, Hand_OBL, Hand_PA, Knee_AP, Knee_LAT, Knee_OBL, Knee_SUNRISE, Wrist_LAT, Wrist_OBL, Wrist_PA, Wrist_SCAPHOID, Abdomen_AP, Abdomen_SUPINE, Finger_LAT, Finger_OBL, Finger_PA, Toe_AP, Toe_LAT, Toe_OBL, Shoulder_AP, Shoulder_EXTERNAL, Shoulder_INTERNAL, Shoulder_Y-VIEW, Elbow_AP, Elbow_LAT, Elbow_OBL, Forearm_AP, Forearm_LAT, Ribs_AP, Ribs_LOWER, Ribs_UPPER, Lumbar_Spine_AP, Lumbar_Spine_L5-S1, Lumbar_Spine_LAT, Cervical_Spine_AP, Cervical_Spine_LAT, Cervical_Spine_ODONTOID, Thoracic_Spine_AP, Thoracic_Spine_LAT, Thoracic_Spine_SWIMMERS, Clavicle_AP, Hip_AP, Hip_LAT, Pelvis_AP, Humerus_AP, Humerus_LAT, Unknown (56 classes)
Medical imaging modality	x-Ray
AI task/problem description	Image classification
General algorithm description	CNNs, DenseNet-121, transfer learning, Bayesian optimization, strong augmentations
Project goal and current stage	Commercial, testing and piloting
Input structure and format	2D image, PNG (converted from DICOM)
Output structure and format	Calibrated score from 0.0 to 1.0 representing precision of data for the current distribution
Evaluation metrics	ROC curve, AUC ROC score, specificity at sensitivity
Explicability and interpretability framework	None currently

1.8 Imaging modalities

Table 2 lists the various medical imaging modalities. The goal of this work is to identify each imaging modality, address how AI can be used with such modality towards diagnosis, triage, forecasts, prognosis or treatment of certain conditions.

Each modality has descriptions of the following details:

– Description: Description of imaging modality.

– Conditions: Conditions modalities are applied to.

– Data structure: Data structure of images from modality. This describes details of the type of images generated from each modality. These details include whether it is a single/multiple 2D image or three-dimensional (3D) image, DICOM or some other format.

– AI applications: How AI is used with modality.

Table 2 – Imaging modalities

Conventional radiography (plain x-rays)	
Description	Radiography is the use of x-rays to visualize the internal structures of a patient. x-Rays are a form of ionizing electromagnetic radiation, produced by an x-ray tube using a high voltage to accelerate the electrons produced by its cathode. The produced electrons interact with the anode, thus producing x-rays. The x-rays are passed through the body and captured behind the patient by a detector, film sensitive to x-rays or a digital detector. Different soft tissues attenuate x-ray photons differently, depending on tissue density; the denser the tissue, the whiter (more radiopaque) the image. The range of densities, from most to least dense, is represented by metal (white, or radiopaque), bone cortex (less white), muscle and fluid (grey), fat (darker grey), and air or gas (black, or radiolucent). This variance produces contrast within the image to give a 2D representation of all the structures within the patient. [112].
Conditions	Typically, conventional radiography is the first imaging method indicated to evaluate the extremities, chest, and sometimes the spine and abdomen. Chest: to assess lung pathology, e.g., atelectasis, pneumonia, pulmonary oedema, heart failure, solitary pulmonary nodule, lung masses, diffuse lung diseases, pleural diseases. Skeletal: to examine bone structure and diagnose fractures, dislocation or other bone pathology. Abdomen: can assess abdominal obstruction, free air or free fluid within the abdominal cavity. [113]
Data structure	Single/multiple 2D image.
AI applications	– Different AI approaches have been proposed to segment chest anatomical structures such as lungs, heart, and clavicle bones, for diagnostic purposes. [10] – AI has also been developed to classify normal and abnormal results from chest radiographs with major thoracic diseases including cardiomegaly, pulmonary malignant neoplasm, active tuberculosis, interstitial lung diseases, pneumothorax, pulmonary oedema, emphysema, pneumonia and paediatric pneumonia. [5–15] – For COVID-19 patients, new AI approaches focusing on detection, classification, segmentation, stratification and prognostication are showing encouraging results. [16–22] AI has been proposed to allow for lung disease severity staging. Deep-learning CNN accurately stages disease severity on portable chest x-ray of COVID-19 lung infection. [23] It has also been proposed that deep learning can thus help support the diagnosis of heart failure using chest x-ray images. [24] – Bone suppression techniques based on artificial intelligence have been developed to avoid overlooking lung nodules because of bones overlapping the lung fields. [25] – AI has been used for analysis and features extraction of spine x-ray images, which may allow prediction of high-risk populations with abnormal bone mineral density. [26] Application prospects have also been described in bone age assessment [14][27].

Table 2 – Imaging modalities

	– In the field of orthopaedics, an AI model can automatically measure Sharp's angle as observed on pelvic x-ray images to aid diagnosis of developmental dysplasia of the hip. [28] It has also been shown the utility of deep learning in detecting hip, pelvic and acetabular fractures with pelvic radiographs. [29] Collection, processing, and integration of pre-, intra-, and postoperative multimodal imaging data could be performed in a more efficient and accurate manner, which has been proposed could then be incorporated into robot-assisted orthopaedic surgery system, [30] as well as for numerous x-ray-guided procedures. [31]
Fluoroscopy	
Description	Fluoroscopy is a technique, usable as a standalone technique or in concert with others, that utilizes a continuous x-ray beam throughout a target in a subject's body to study both its structure and movement and can be applied to single organs or a system of them. [35-37]
Conditions	This modality is commonly applied to conditions that involve foreign bodies, obstruction or modification of fluid transport, or fractures. [35-37]
Data structure	Images generated through fluoroscopy can be produced in single-plane 2D images as well as multi-plane 3D images. [35-37]
AI applications	AI is being used to simplify and optimize presentation of imaging, as well as reduce radiation exposure to patients. [38] [39]
Angiography	
Description	Angiography is a medical imaging modality that focuses on imaging the inside of blood vessels and organs. In angiography, a contrast medium is injected into the blood vessel and the path of the tracer or contrast medium is imaged using x-ray. [57] [58]
Conditions	Some conditions angiography is applied to are: diagnosis of obstructive vascular disease, diagnosis of aneurysms, diagnosis of arteriovenous malformations, diagnosis of bleeding vessels, and assessment of vascularity of malignant tumours. [57]
Data structure	Angiograms can be 2D or 3D image files.
AI applications	AI is used in post processing tasks like segmentation. Also AI is used to perform certain calculations like calculating calcium score and fraction flow reserve. [59]

Table 2 – Imaging modalities

Mammography	
Description	Mammography is a medical imaging modality that uses low energy x-rays to image the human breast. Mammography is mostly used for early detection of breast cancer. Its mode of operation is very similar to that of the conventional x-ray machine, except that it employs low power radiation. [49][50]
Conditions	Mammography can be used as a tool for screening or diagnosis tool. – As a screening tool, mammography is used for the early detection of breast cancer. – As a diagnostic tool, mammography is used to investigate abnormal clinical findings in the breast, like breast lumps and nipple discharge. [50]
Data structure	Mammograms may be 2D or 3D image files. [50]
AI applications	AI, in combination to radiologists, is used to improve the accuracy of breast cancer screening. [51]
Computed tomography	
Description	CT also called computed axial tomography, is a non-invasive imaging method that uses x-rays, combined with computing to produce cross-sections of subjects, allowing for highly detailed models of patients or areas of interest to study; patients are sometimes given a contrasting material to improve image quality [72] [73].
Conditions	CTs are used in multiple diagnostic works and therapies, and have additional value in that full body scans are possible. [72][73] Examples of uses include disease diagnosis and prognosis, guidance of medical procedures, and treatment monitoring across a wide spectrum of disorders from problems with vasculature, bone fractures, investigations in oncology, psychiatry and more. [72-75]. It has even found use in investigating complications associated with COVID-19 within patients [76] [77].
Data structure	CT scans take numerous 2D images, and these can be used to make 3D representations, thus allowing 2D and 3D formats [72][84].
AI applications	Current AI uses extend from use of CT-images, but is also expanding through investigation of AI-Assisted smart tools to guide and upgrade the use of Ct scans through improved diagnosis, measurements, and prognoses. [78-82] It is believed that future uses can entail more comprehensive reconstructions of scanned areas and less radiation use through less coregistration of CTs with other imaging means, helping to reduce patient fatigue and exposure; more may result as this area of research, i.e., the combination of AI and CT scanning, is still new. [83]

Table 2 – Imaging modalities

Single-photon emission computed tomography	
Description	Single photon emission computed tomography (SPECT) is a technique that allows nuclear medicine studies, which would otherwise be represented in 2D, to be rendered in 3D. Photons emitted by injected radiopharmaceuticals are detected by gamma cameras that rotate around the patient to provide spatial information on tissue distribution. The data are then reconstructed into 3D images. SPECT can also be combined with conventional CT (SPECT-CT) to allow accurate attenuation correction for the purposes of reconstruction, and to provide additional anatomical information.
Conditions	The technique can theoretically be applied to any nuclear medicine studies, but it is not required in every situation. SPECT is commonly used in the context of technetium-99m sestamibi scans when evaluating the perfusion of the cardiac myocardium or the function of parathyroid glands. It is also used in the context of technetium methylene diphosphonate bone scans that provide information about bone perfusion and turnover.
Data structure	
AI applications	
Ultrasonography and Doppler	
Description	Ultrasonography (US) is an imaging modality that uses ultrasound (sound waves with frequencies greater than frequencies that are audible to the human ear) to create images of internal body parts. The ultrasound is sent into the body by a transducer and echoes from tissue interference are recorded to create an image of the structure under examination. [40]
Conditions	Ultrasound imaging is used to examine an organ whenever there is a symptom of pain, swelling or infection in that organ. US can be used to image the liver, kidney, heart, pancreas, etc. [41] [42] Another common use case for US is real-time imaging of developing foetuses in pregnant mothers.
Data structure	Sonograms may be stored as a single layer 2D image. Multiple 2D sonograms may also be projected into a 3D image. An additional time dimension can be added to a 3D sonogram to create a 4D sonogram. [43]
AI applications	AI is used to perform a wide range of tasks in US. These tasks include image classification, segmentation, detection, registration, biometric measurements and quality assessment. [44]

Table 2 – Imaging modalities

Magnetic resonance imaging	
Description	MRI is a modality that uses a strong magnetic field to create images of the internal structures of the body. The strong magnetic field forces protons of water molecules in the body to align with the field. When a radiofrequency current is passed through the patient, the alignment of the protons is disturbed. When the radiofrequency current is turned off, the protons return to equilibrium with the magnetic field and the MRI sensors detect the energy released by the protons as they return to equilibrium. Unlike CT or conventional x-ray, MRI does not employ any ionizable radiation, so it is safer and can be taken more frequently. [52] [53]
Conditions	MRI is suitable for imaging soft tissues like muscles, tendons, ligaments, brain, joints and the abdomen. MRI is also employed in image guided interventional procedures. [52] [54]
Data structure	MRI images can be 2D or 3D image files
AI applications	AI is used to correct artefacts in MRI scans. [55] AI is also used to classify MRI scans as depicting healthy or diseased tissue. [56]
Nuclear medicine imaging	
Description	Nuclear medicine imaging is an imaging modality that involves the injection or inhalation of small amounts of radioactive compounds (called radiotracers) into the body to visualize organs in the body. The radiotracers are organ specific and they emit gamma-rays when they arrive at the target organ. The emitted gamma-rays are captured and visualized using a gamma camera. Nuclear medicine imaging is considered as an "inside out" radiology, because it records radiations generated from the body rather than an external source like an x-ray. [45-47]
Conditions	This modality is applicable to conditions that require an assessment of the physiology of organs. Some organs that are commonly assessed using nuclear imaging are kidneys, lungs, heart, thyroid gland and bone. [45]
Data structure	Nuclear images can be 2D (scintigraphy) or 3D (SPECT). Some modern nuclear imaging equipment is hybrid and allows for a fusion between CT and nuclear imaging. [45] [47]
AI applications	In nuclear imaging, AI is commonly used for radiomics. AI can potentially be used to detect artefacts and noise in nuclear images and correct them by applying the appropriate algorithm.

Table 2 – Imaging modalities

Positron emission tomography	
Description	PET is an imaging modality that uses tracers or radioactive drugs to image the function of tissues of organs. [32]
Conditions	PET is used for diagnosis and staging in oncology, in addition to observing specific neurological and cardiovascular issues. [33]
Data structure	Images can come in 2D or 3D modalities. [34]
AI Applications	AI has been documented in use with PET for distinguishing between benign and malignant nodules, as well as detection and quantification of nodules [35] [60].Future developments may improve correlation of image features with clinical end points, correction of images, reduction of doses needed for reliable scans, guided use, and improved reconstructions [83] [85]. These together can result in savings and improved patient outcomes, with more to abound as research in this area is still new.
Interventional radiology	
Description	Interventional radiology (IR) is a means of radiology that uses current imaging methods, such as CT, MRI, x-rays, PET and ultrasound, led by teams of professionals to treat the source of diseases in a non-invasive or minimally invasive manner. A subset, interventional oncology is used to address cancer [61]
[Conditions	IR is used for diagnosis and guiding of treatment across cardiology, neurology, nephrology, oncology, and more. [61]
Data structure	Image modalities from IR depend on the imaging method combinations as described in previous entries.
AI applications	AI has been used in IR to predict outcomes for treatments like chemoembolization, incidents like a post-treatment stroke, or offer prognostic information on brain malformations [63-65]. Gesture capture, voice recognition, implement/tool guidance, and augmented reality have been employed to assist efforts across various tasks [66-69]. A smart assistant has been trialled, but more details await. [70] [71] Applications that improve features such as segmentation of subjects, improved lesion detection, prognostic information gathering, interpretation, reduction of waste, and improved cost-benefit analyses are imagined in the future of IR with AI. [62] [70] [71]

1.9 Existing work on benchmarking

Benchmarking work includes:

– papers on existing attempts to benchmark solutions on the topic;
– clinical evaluation attempts, randomized controlled trials, etc;
– existing numbers.

1.10 Benchmarking overview

AI is considered to be one of the key driving forces of the fourth industrial revolution. This has led to the adoption of national AI strategies by many countries. [110] However, there is the lack of a consensus on how to measure the success of AI models. A brief non-exhaustive list is therefore given of activities that could be performed as part of benchmarking AI models. Benchmarking may include measurement of the predictive performance of AI models. Several performance metrics have been proposed and a few are listed; AUC; accuracy; F_1-score; and sensitivity and specificity. [109] Model performance should be measured for both validation and test data. Benchmarking should also take into account the annotation of data – whether the data are labelled, unlabelled or semi-labelled. This will determine what AI models and performance metrics to use. Appropriate models should also be used in AI-based solutions. Many factors should be considered when applying AI models; type of data, sample size, computational cost, etc. [106] It is also important to assess the documentation of data analysis pipelines in order to determine the level of reproducibility of the methods.

1.11 The NHS AI Lab – Call for AI driven COVID-19 models: Performance assessment using the national COVID-19 chest imaging database

The NHS AI Lab created the National COVID-19 Chest Imaging Database (NCCID), currently with over 40 000 images. The majority of scans collected by the NCCID are chest x-rays and come from people with and without COVID-19. They provide a platform that allows for AI solutions within the UK to be assessed based on the NCCID dataset, in order to reduce the potential for bias and provide NHS commissioners and healthcare regulators with the evidence to judge the safety, efficacy and generalizability of AI models before they are used in clinical practice. [98]

Before an AI system can be assessed on their platform, AI developers have to fill an application form. They ask technical and clinical questions within the application form in order to understand the processes used in training and evaluating the AI system. Independent assessors with expertise in AI, technology and medicine are used to assess responses provided with a focus on NHS importance, technical feasibility, and financial viability. These external assessors prepare analysis plans, covering performance criteria and tailored to each AI solution. The AI system is then validated on the unseen NCCID dataset via an Amazon AWS cloud-computing infrastructure provided by the NHS Transformation Directorate (formerly NHSX) [114]. The NCCID unseen dataset is then accessed in the form of an S3 bucket. AI developers are never given access to the NCCID unseen dataset.

The whole process takes 12-16 weeks to complete, and is done at no cost to AI developers. To ensure intellectual property (IP) protections, all people involved in the AI model assessment, including external assessors will be bound to confidentiality by contractual agreements. Non-disclosure agreements are also used where need be.

At the end, an AI developer receives a written report with the assessment of the AI system against defined performance criteria. This covers model performance using metrics including sensitivity, specificity, as well as the clinical validity of the solution. The process is meant to be a validation study and does not qualify as a clinical investigation. However, this report can be used as evidence to support applications to the Medicines and Healthcare products Regulatory Agency, the United Kingdom's healthcare products regulatory agency, for derogation of UK Conformity Assessed (UKCA) or European Conformity (CE) marking or via standard conformance assessment processes. The UKCA marking is a new UK product marking that is used for goods being placed on the market in Great Britain (England, Wales and Scotland). It covers most goods that previously required the CE marking.

2 AI4H Topic Group

– Topic Group (TG) structure
  • Subtopic 1
  • Subtopic 2
– TG participation
– Tools/process of TG cooperation: Slack, Zoom, Google Docs, Github
– TG interaction with other FG-AI4H working groups: WG-DAISAM and WG-DASH to test frameworks in a sandbox environment
– Current topic group and topic status
– Contributors so far
– Next meetings
– Next steps for the work on this document.

3 Method

– Overview of the benchmarking.

3.1 AI input data structure

– possible inputs for benchmarking;
– ontologies, terminologies;
– data format.

3.1.1 Image conversion considerations

See Table 3.

Table 3 – Image conversion considerations

Conversion Approach	Advantages	Disadvantages
Integrating an automated conversion programme into AI software It is also possible to use Python tools pydicom and opencv-python to automate the process of converting DICOM to JPEG within the software platform, in that case, the users would not have to worry about the conversion.	– Easier for users in clinical settings. – Conversion cannot be easily interfered with. – Leaves little room for error on the part of users.	– Requires further development by manufacturers. – Subjected to the quality of manufacturer software development.
Using a separate software There is MicroDicom, a free Windows tool, and a number of others that are either free or require payment.	– Easier for manufacturer since it requires no to little additional development. – Can allow for reliance on already established and trusted high-quality tool.	– Requires additional procedures from users to use AI software. – Prone to errors and incorrect input data if misused. – Creates avenue for third party interference.

Table 3 – Image conversion considerations

Conversion Approach	Advantages	Disadvantages
	– If offline, it can ensure data privacy better than an online tool.	
Using an online tool There are also online free tools, like [118]	– Easier for manufacturer since it requires no to little additional development. – Can allow for reliance on already established and trusted high-quality tool.	– Requires additional procedures from users to use AI software. – Prone to errors and incorrect input data if misused. – Creates avenue for third party interference. – Can allow online tool manufacturers to have unauthorized access to data.

3.1.2 Image compression and other artefacts considerations

For use cases that require image conversions like DICOM to other formats before being used as input for an AI system, manufacturers should ensure input data integrity and quality are maintained. This is significant as DICOMs usually use 16 bit depth raw images and would be converted into 12 bit or even 8 bit depth images in JPEG, JPEG 2000 or PNG format.

This depth precision reduction may be negligible if it is considered that:

– the higher pixel depth cannot be perceived by the human eye;

– regular monitors do not use high-range depths;

– ground truths are usually made by physicians using regular monitors.

Another issue is related to the JPEG and JPEG 2000 image codec formats, which are lossy compression algorithms. These codecs, respectively, introduce compression artefacts such as blocking and ringing. These artefacts may reduce AI system performance and should also be taken into consideration in the system design.

In order to show the relevance of the compression in medical images in the performance of AI-based classification, we run a set of tests. Our baseline is COVID-Next, [119] a COVID-19 classifier, inspired by the COVID-Net proposed by Wang et. al., [16] based on ResNext50.

This model was trained using chest radiography with different resolutions, qualities and artefacts. The test accuracy of this model is 94.76%. However, if the test dataset is compressed with different quality parameters simulating a scenario where the image is compressed to reduce bandwidth before transmission to a classifier in the cloud for inference. It was observed that it is possible to achieve significant bandwidth reduction with a negligible accuracy reduction.

Examining the cyan and red curves in Figure 1, it is evident that the accuracy can be significantly reduced due to compression. In this case, accuracy notably drops when the compression ratio falls below 0.10.

Despite visual quality reduction due to compression, the effect of compression artefacts (blocking or ringing) is substantially reduced due to resizing of the compressed image before feeding to COVID-Net.

In an extreme case, referring to the green (JPEG) and blue (JPEG 2000) curves in Figure 1, the images in the dataset are resized to 256 × 256 pixels using a Lanczos-4 filter before performing the compression. In this scenario, the bitstream is outstandingly reduced, but the accuracy is significantly reduced, showing that severe compression is detrimental to COVID-Net as the image quality degrades. This image size was chosen due to the COVID-Net input architecture.

A similar test was conducted with a brain tumour image classifier in2021. [120] The results are shown in Figures 2 and 3 where accuracy and F_1-score are calculated for different compression ratios and different curves are obtained for each codec configuration.

The results show that, in both cases, there is a combination (between scaling and compression quality) where it is possible to achieve a large reduction in the transmission rate without impairing accuracy. The difference observed in the behaviour of the models can be associated with the amount of pre-compressed images present in the data.

These results cannot be extended to other cases, but can show the influence of the compression artefacts in medical image classification.

In order to evaluate image compression in the scenario, a library was developed that calculates a set of metrics, such as accuracy, sensitivity, specificity and *F*-score, for testing different compression and downsizing in a dataset.

Figure 4 shows an example of the confusion matrix for a given compression configuration. The library saves different matrices for each configuration parameter tested.

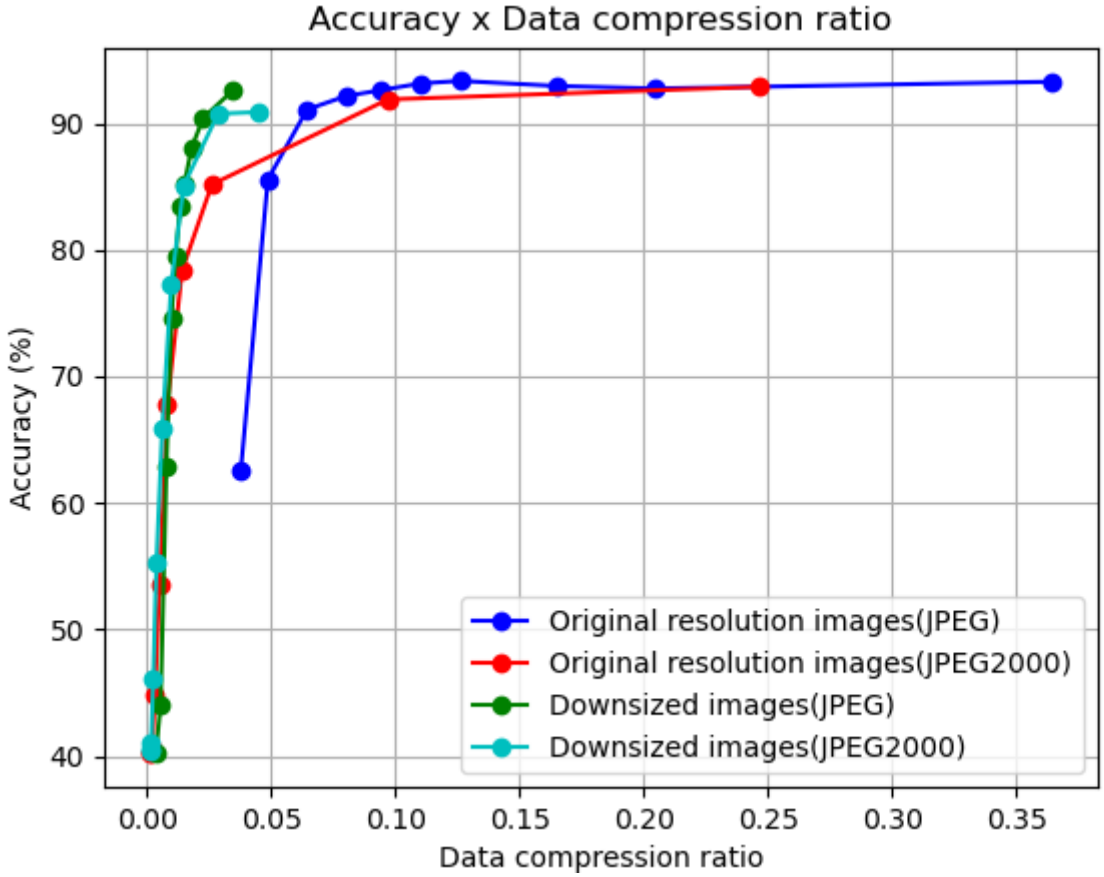


Figure 1 – Impact of the compression in the test dataset accuracy of the COVID-Next classifier

The blue (JPEG) and red (JPEG 2000) plots show cases where dataset images were compressed with different compression rates. In the green (Interpolative JPEG) and cyan (Interpolative JPEG 2000) plots, the images were downsized to 256 × 256 pixels before compression. Without compressing the images (PNG), the accuracy is 94.76%, as shown in magenta.

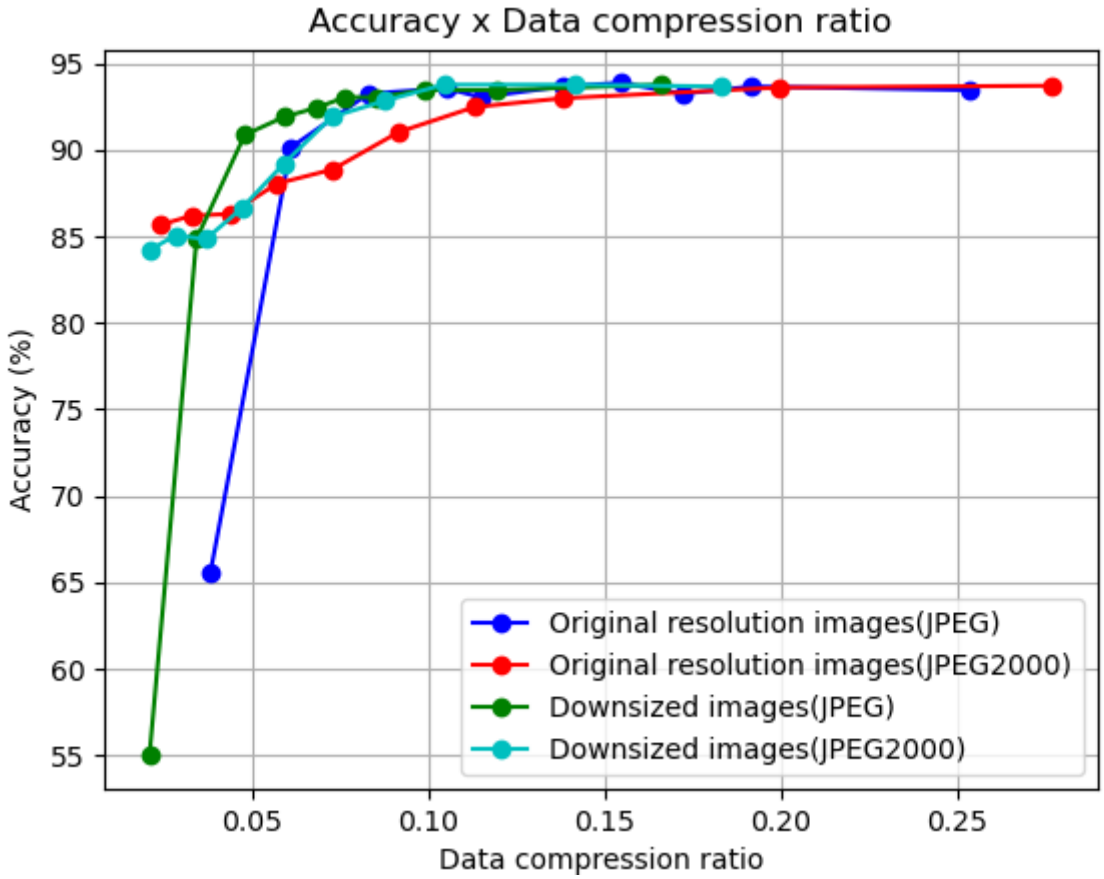


Figure 2 – Impact of the compression in the test accuracy of the brain tumour classifier

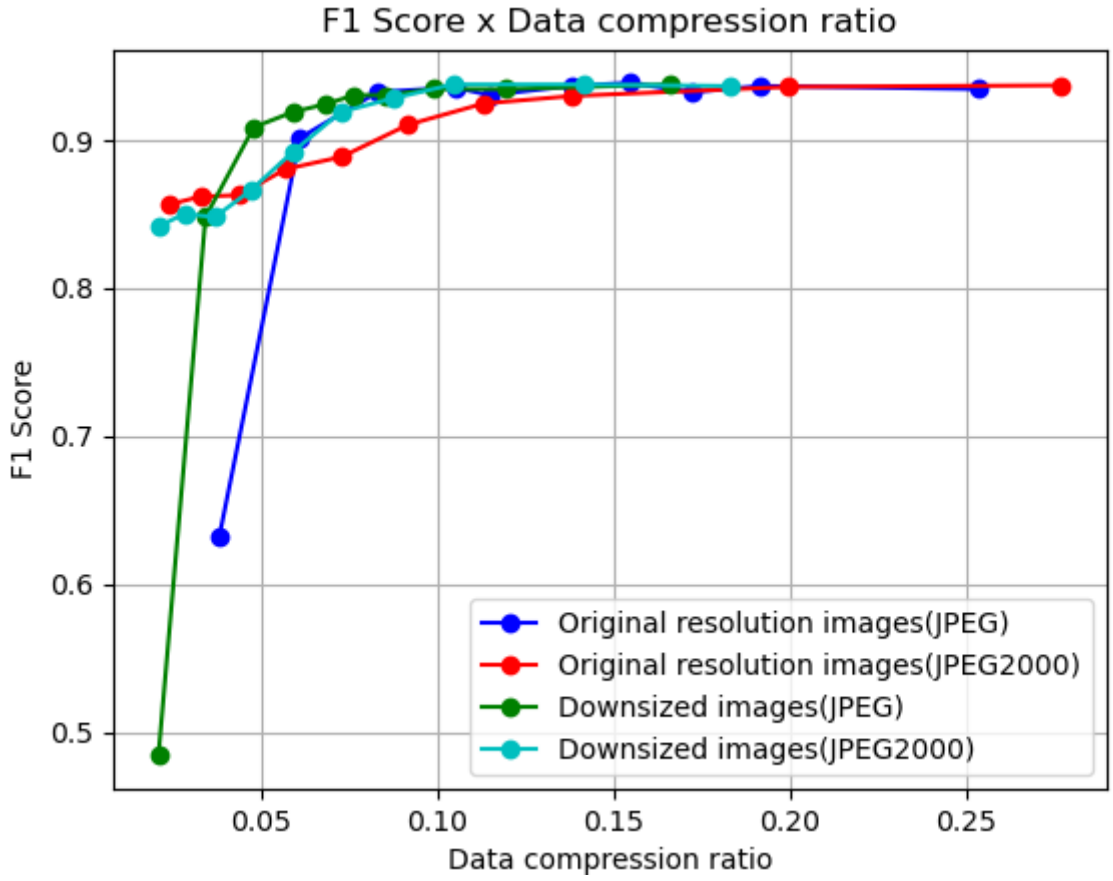


Figure 3 – Impact of the compression in the test accuracy of the brain tumour classifier

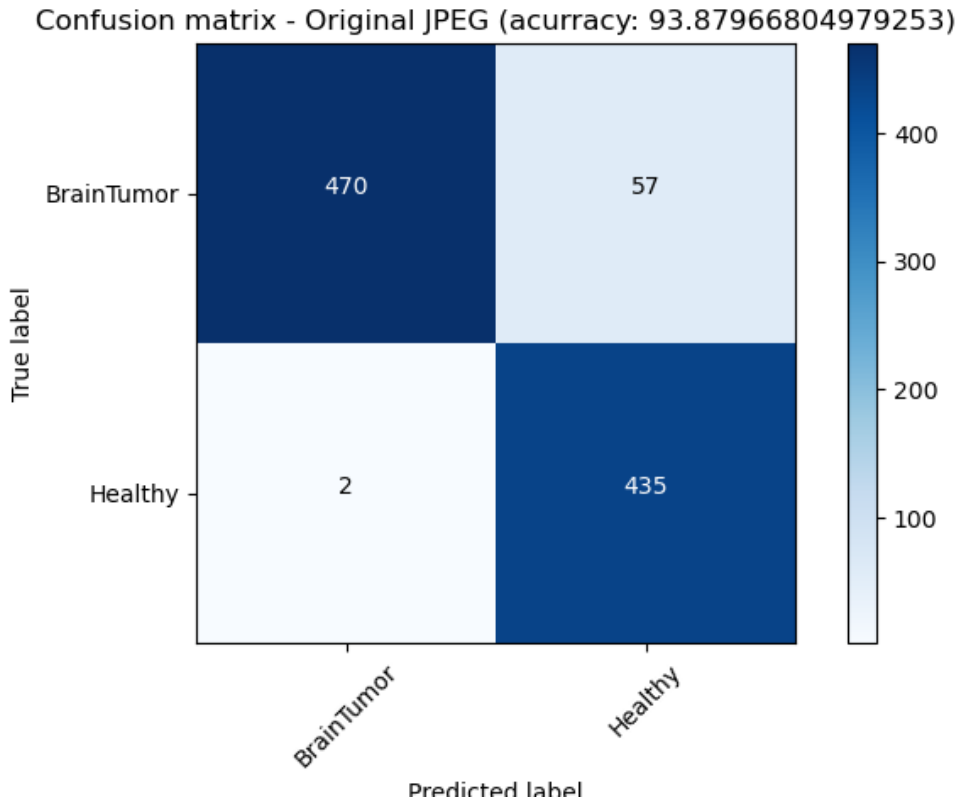


Figure 4 – Confusion matrix of the brain tumour classifier test accuracy of a JPEG compression scenario

Another artefact that may also be taken into consideration is the Moiré pattern. This kind of artefact can occur when a picture is taken from a screen. In this case, the pattern of the pixels in the screen is overlaid with the capturing pattern of a camera. Developers must consider that users may not use the AI solution properly and taking pictures may be a possible input of a proposed system.

3.1.3 Lossless medical image compression for radiology

Background

Loading, storing and visualizing large neuroinformatics files (NII) commonly used in CT and MRI are costly and time consuming. To process and transfer files across systems is extremely time consuming. As more medical samples are accumulated and used to train AI models, file storage and processing must be rethought. A form of lossless Hilbert compression is introduced using neuro-symbolics to decrease times for processing, transfer and training for medical AI models through pre-vectorization.

Representation phases

In working with multimedia, it is important to follow steps of standardization in which all new data that enter a system are bound. This process diverges data by collecting the data points, re-converges the data, and allows for novel trends to emerge. See Figure 5.

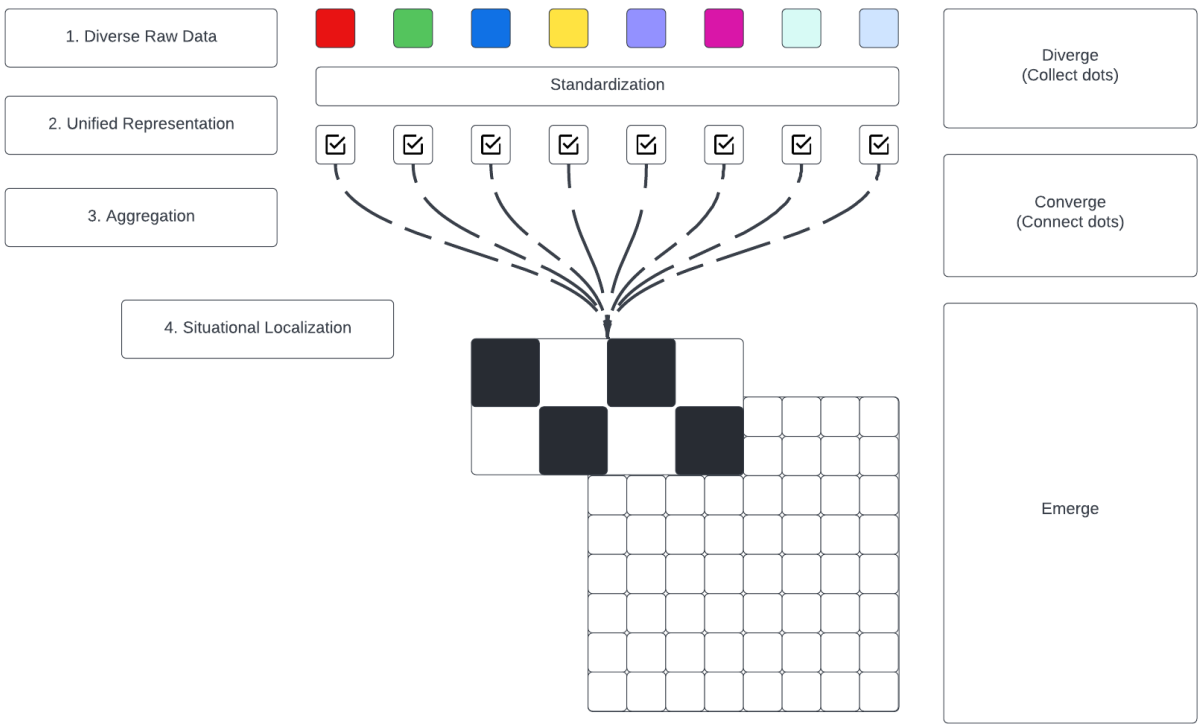


Figure 5 – Multimedia representation phases for radiology images

1) Diverse types of raw data and medical records enter a system.
2) The representation of diverse data is unified in representation by answering common questions of it. What is it? Where did it come from? When did it happen?
3) The data is then aggregated by following the same processing protocols.
4) The aggregation of this data enables situational localization in which converged points begin to emerge as trends.

Vectorizing medical imagery

In building databanks of medical samples and records, it is important to efficiently store the multimedia, which is usually large in size, and sometimes sparse in situation. For training AI models, these data must be vectorized in order to train ontologies of disease and diagnosis. Figure 6 shows early research found from the NIST Medical Databank which used MRI records as a basis for the example and diagram. This same process of vectorizing multimedia is still relevant to leading research across a range of disciplines today.

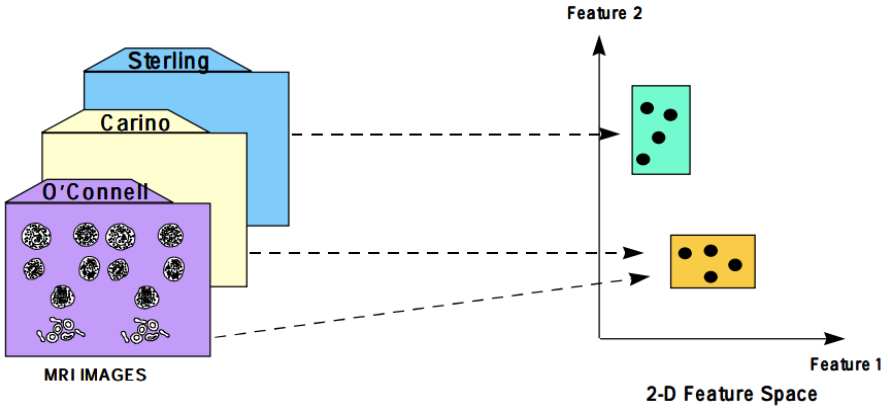


Figure 6 – Early example of vectorizing medical imagery (Source: [86])

Hilbert symbolics

A method of lossless Hilbert compression is introduced using neuro-symbolics as an effective strategy for parallel computation of medical imagery, as illustrated in Figures 7 and 8. An image, or slice, is broken down recursively across threads and systems into Hilbert spaces, which form the bounds for hash symbolics as unique floating-point signals.

Each space can be simultaneously processed as its representation is uniformly computed across multiple threads, nodes or systems to form a hierarchy of which each space originates. Each segment is processed down to the individual pixel, forming a high-resolution hash table of features within a slide or sequence of slides, which is calculated concurrently.

The computed features are bound to a vector index, using buckets which scale up or down with a given system's memory footprint. If a system is large and can handle a large memory footprint, then the bucket size may be larger; however, it is not required, as when buckets are full they simultaneously write to file, regardless of order, as the file can be read back and the contained features and positions are retained, therefore preserving the sanity of the data being ingested. See Figure 9.

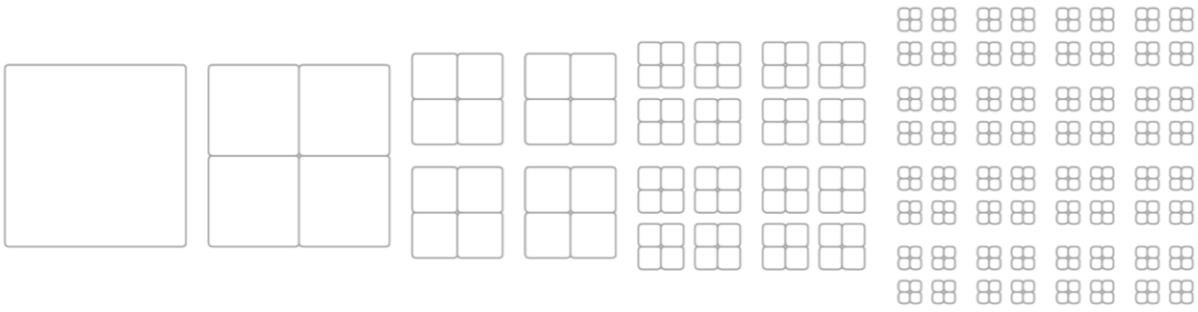


Figure 7 – Illustration of lossless Hilbert compression – partitioning

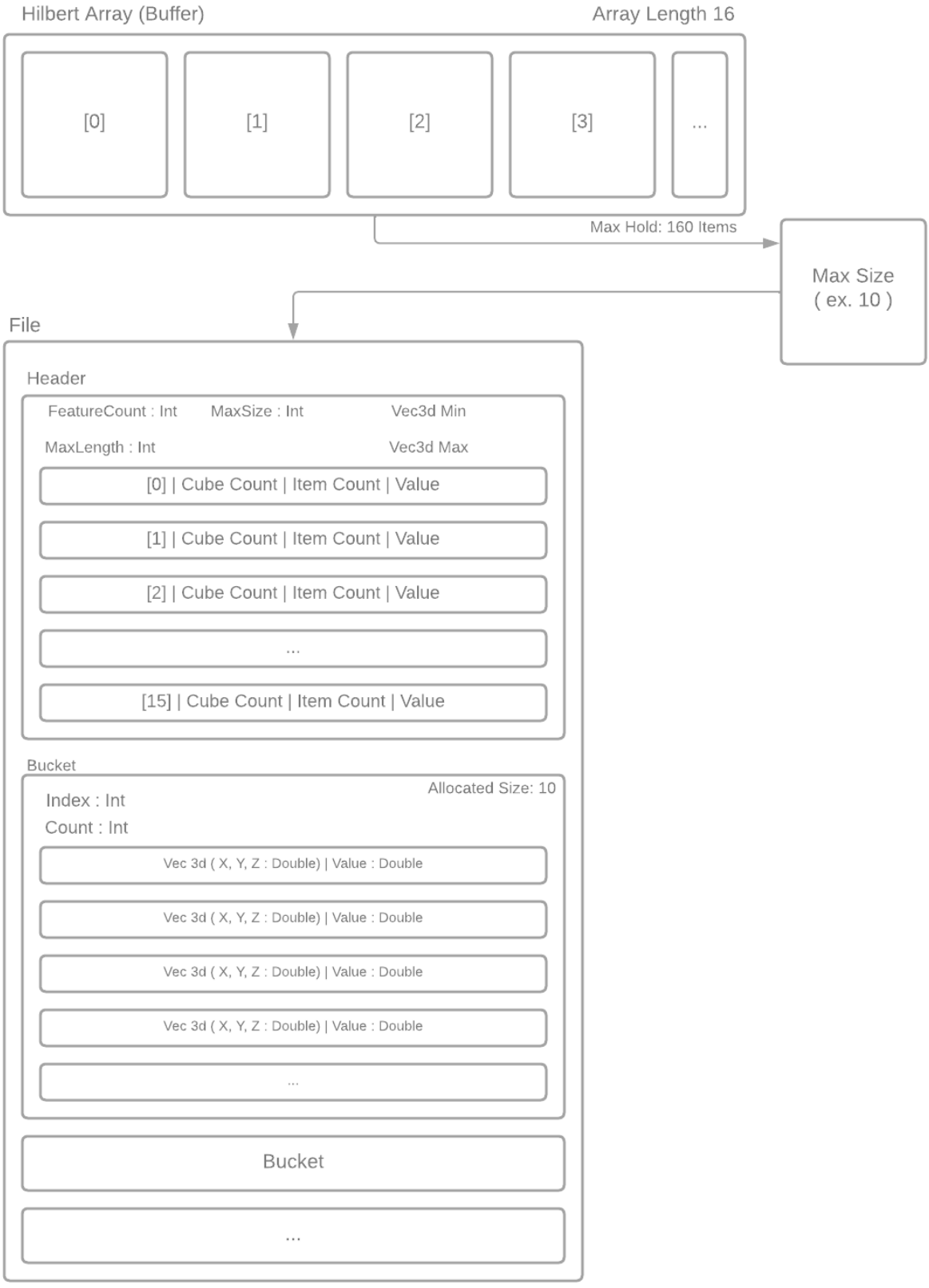


Figure 8 – Illustration of lossless Hilbert compression – flow

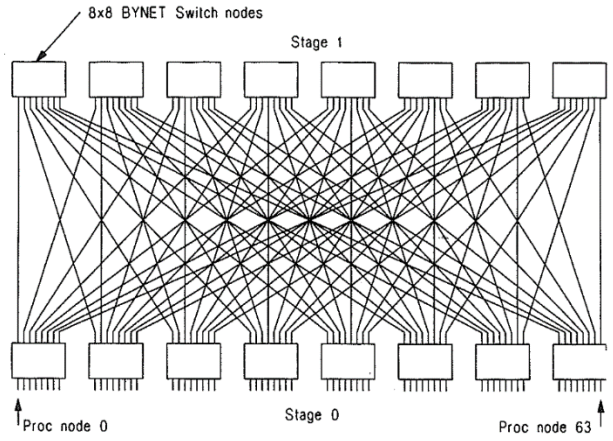
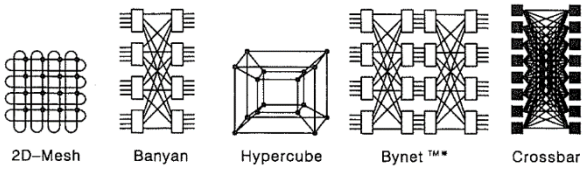


Figure 5. Bynet network topology.

8.4 INTERCONNECTION NETWORK TOPOLOGIES

The following illustrations and discussion[12,29,30,32] describe interconnection network topologies and performances. Figure 8 pictorially shows the topology for 2D-Mesh, Hypercube, Crossbar, BANYAN, and the Bynet.

Table 1 is a qualitative comparison of network topologies. Table 2 provides a quantitative practical comparison using 64 nodes as an example. SDC-OMEGA,[16] EDS-DELTA,[20,34] MESHNET,[35] and iPSC/860-PARAGON[36] are other interconnection networks that have been designed since the survey was written.[29] Our description above shows the thought processes and rationale behind the Bynet design choices.



* Bynet based on 8x8 switch nodes; picture uses 4x4 nodes.

Figure 8. Interconnection network topologies.

Figure 9 – Example of interconnection network topologies (Source: [87])

Early network topology is referenced as validation of such interconnected systems computing features in parallel. These early network topologies utilize computation formulas also found in neural network models. This was the basis for industrial supercomputers used in early iterations of the NIST Medical Databank for storing large quantities of medical samples.

Performance Benchmarks:

In testing benchmarks, the standard file size is shown of an .NII file containing a chest CT scan of a COVID positive patient and 512 slices. We compress this to .NII.GZ and .NII.BZ2, respectively, and the following results indicate a compression of 24.9% for GZIP and 51.5% for BZ2. The processing time for GZIP taking 3.811 s, and BZ2 taking 10.578 s.

We compare this with the process for compressing the same .NII file with Hilbert symbolics, which indicates an 87.4% compression rate taking 664.062 ms to process. The performance benefit being in ability to distribute the computation of each slice across 256 threads where each thread computes two slides. The total results of the benchmark are as shown in Table 4 and Figure 10.

Table 4 – Compression performance comparison for various compression methods

	Standard	GZIP	BZ2	Hilbert symbolics
Size (MB)	134.2	100.7	65	16.8
Time (s)		3.811 621 189	10.578 860 28	0.664 062 5

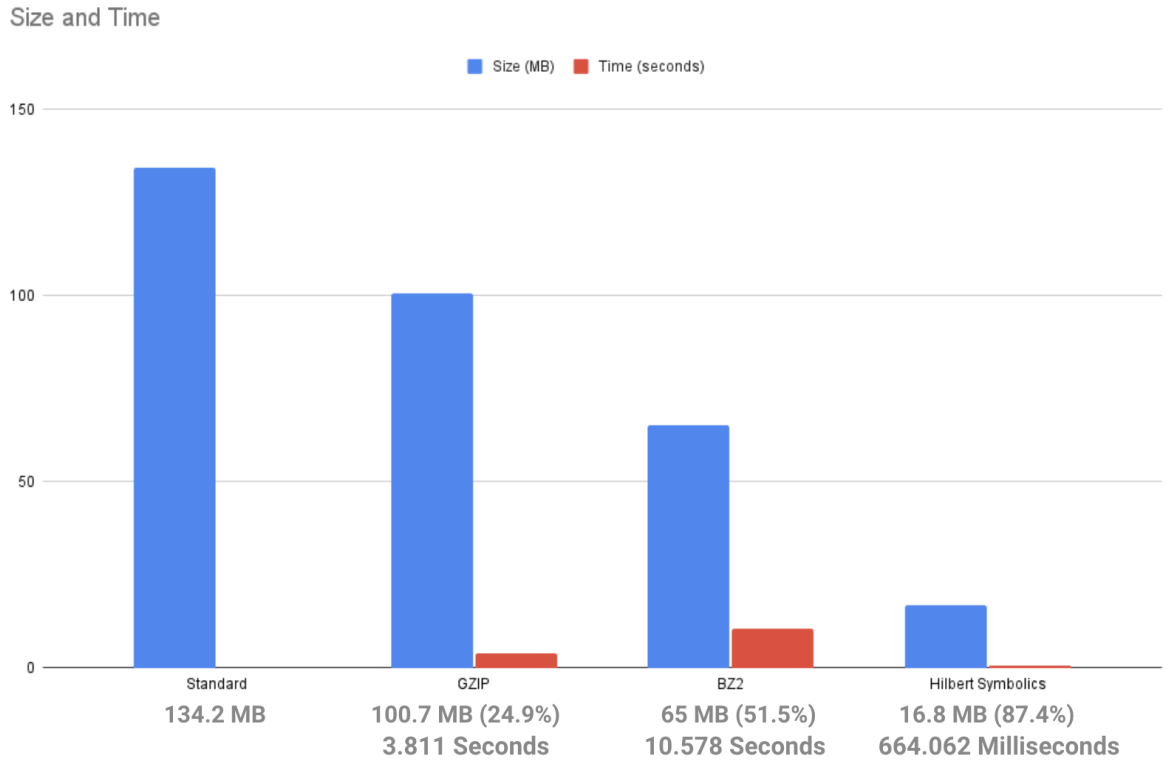


Figure 10 – Compression performance comparison for various compression methods

Further analysis is provided of the processing time, and of the storage requirements of resulting files, as illustrated in Figure 11.

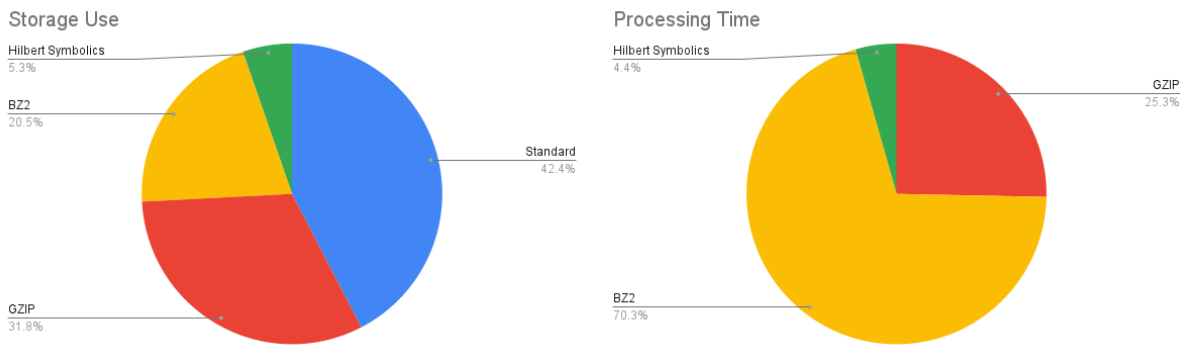


Figure 11 – Processing time and storage requirement for various compression methods

3.2 AI output data structure

– outputs to benchmark;

– ontologies, terminologies;

– data format.

3.3 Test data labels

– label types;

– ontologies, terminologies;

– data format.

3.4 Scores and metrics

The taxonomy used in grouping these evaluation metrics is that proposed by Ferri et al: [105]

– threshold;

– ranking;

– probability.

3.4.1 Threshold metrics

3.4.1.1 Accuracy metrics

Classification accuracy

This is the fraction of correct predictions of a model. It is not, however, suitable for imbalanced classification because a poorly fitted model that simply predicts the majority class would end up having a misleading high score.

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total predictions}}$$

Classification error

This measure is the inverse of classification accuracy. It is the fraction of incorrect predictions of a model. It is also not suitable for imbalance classification.

$$\text{Classification error} = \frac{\text{Incorrect predictions}}{\text{Total predictions}}$$

Patient level accuracy and image level accuracy

The patient level accuracy metric is determined as follows. For each patient, let N_t be the total number of images and N_c the number of images correctly classified, then patient score S can be defined as:

$$S = \frac{Nc}{Nt}$$

Therefore, the patient level accuracy can be calculated as

$$\text{Patient level accuracy} = \frac{\Sigma_{i=1}^{T} Si}{T}$$

Where T is the total number of patients.

The image level accuracy measures the rate of correctly classified images to the total number of images in the dataset. Let N be the total number of images in testing data and C the number of correctly classified images.

$$\text{Image level accuracy} = \frac{C}{N}$$

Pixel accuracy

In instance segmentation, pixel accuracy is used to evaluate the percentage of pixels in an image that were correctly classified. This is usually reported for each class separately and then across all classes.

This metric can be misleading in scenarios where the class representations are small within the image, as the measure will be biased in mainly reporting how well negative cases are identified.

Exact match ratio

The EMR metric extends the accuracy metric from single-label classification tasks to multi-label classification tasks. One of the drawbacks of EMR is that it does not account for partially correct labels.

Mathematically,

$$\text{EMR} = \frac{1}{n}\sum_{i=1}^{n} [I(y^{(i)} == \hat{y}^{(i)})]$$

Where:

n is the number of training examples

$y^{(i)}$ are the true labels for the i[th] training example

$\hat{y}^{(i)}$ are the predicted labels for the i[th] training example.

Example-based accuracy

This extends the accuracy metrics to multi-label classification. The overall accuracy is the average of accuracy across training instances.

Macro averaged accuracy

This extends the accuracy metric to multi-label classification. This metric computes the accuracy of individual class labels and then averages over all classes.

Mathematically,

$$\lambda - Accuracy\ (A^{j}_{Macro}) = \frac{\sum_{i=1}^{n} [y_j^{(i)} \wedge \hat{y}_j^{(i)}]}{\sum_{i=1}^{n} [y_j^{(i)} \vee \hat{y}_j^{(i)}]}$$

$$Accuracy_{Macro} = \frac{1}{k}\sum_{j=1}^{k} (\boldsymbol{A}^{j}_{Macro})$$

Where:

n is the number of training examples

$y_j^{(i)}$ are the true labels for the i[th] training example and j[th] class

$\hat{y}_j^{(i)}$ are the predicted labels for the i[th] training example and j[th] class

$\wedge$ is the logical AND operator

$\vee$ is the logical OR operator

k is the number of classes.

Micro averaged accuracy

This extends the accuracy metric to multi-label classification. This label based metric computes the accuracy globally over all instances and all class labels.

Mathematically,

$$Accuracy_{Micro} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} [y_j^{(i)} \wedge \hat{y}_j^{(i)}]}{\sum_{j=1}^{k}\sum_{i=1}^{n} [y_j^{(i)} \vee \hat{y}_j^{(i)}]}$$

Where:

n is the number of training examples

$y_j^{(i)}$ are the true labels for the i[th] training example and j[th] class

$\hat{y}_j^{(i)}$ are the predicted labels for the i[th] training example and j[th] class

$\wedge$ is the logical AND operator

$\vee$ is the logical OR operator

k is the number of classes.

3.4.1.2 Sensitivity-specificity metrics

Sensitivity

This is the true positive rate (TPR). It measures the proportion of positive samples correctly predicted by a model.

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

Specificity

This is the true negative rate. It measures the proportion of negative samples correctly predicted by a model.

$$\text{Specificity} = \frac{\text{True negative}}{\text{True positive} + \text{False negative}}$$

Geometric mean (*G*-mean)

The geometric mean metric is the square root of the product of the sensitivity (TPR) and specificity (true negative rate) scores of a model.

$$G\text{-mean} = \sqrt{\text{sensitivity} * \text{specificity}}$$

3.4.1.3 Precision-recall metrics

Precision

Precision is a metric that computes the fraction of true positive predictions among the outcomes that the model classified as positive.

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

Recall

Recall, also known as sensitivity, is the fraction of examples classified as positive, among all total numbers of positive examples. In other words, the number of true positives divided by the number of true positives plus false negatives.

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

***F*-measure**

The *F*-measure provides a way to combine precision and recall into a single score. It is the harmonic mean of two fractions. It is sometimes called the *F*- or F_1-score. It is the most popular metric for working with imbalanced datasets.

$$F - \text{measure} = \frac{(2 * \text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

F_β-measure

F_β measure is an abstraction of F-measure score. A coefficient called beta is used to control the calculation of the harmonic mean of the precision and recall.

$$F_\beta\text{-measure} = \frac{((1 + \beta^2) * \text{Precision} * \text{Recall})}{(\beta^2 * \text{Precision} + \text{Recall})}$$

Matthews correlation coefficient (MCC)

The **Matthews correlation coefficient** (MCC) or phi coefficient is a measure of the quality of binary (two-class) classifications. MCC according to Chiccoand Jurman [104] is more informative than F_1-score and accuracy score in evaluating binary classification problems, because it produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset.

$$\text{MCC} = \sqrt{\frac{x^2}{n}}$$

where $n$is the total number of observations.

MCC can also be calculated directly from the confusion matrix as:

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FN})}}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives.

Macro averaged precision

This extends the precision metric to multi-label classification. This metric computes the precision of individual class labels and then averages over all classes.

Mathematically,

$$\lambda - Precision\ (P^j_{Macro}) = \frac{\sum_{i=1}^{n} [y_j^{(i)} \wedge \hat{y}_j^{(i)}]}{\sum_{i=1}^{n} [\hat{y}_j^{(i)}]}$$

$$Precision_{Macro} = \frac{1}{k}\sum_{j=1}^{k}(P^j_{Macro})$$

Where:

n	is the number of training examples
$y_j^{(i)}$	are the true labels for the i[th] training example and j[th] class
$\hat{y}_j^{(i)}$	are the predicted labels for the i[th] training example and j[th] class
$\wedge$	is the logical AND operator
P^j_{Macro}	is the precision for label/class k
k	is the number of classes.

Micro averaged precision

This extends the precision metric to multi-label classification. This label-based metric computes the precision globally over all instances and all class labels.

Mathematically,

$$Precision_{Micro} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}[y_j^{(i)} \wedge \hat{y}_j^{(i)}]}{\sum_{j=1}^{k}\sum_{i=1}^{n}\hat{y}_j^{(i)}}$$

Where:

n	is the number of training examples
$y_j^{(i)}$	are the true labels for the i[th] training example and j[th] class
$\hat{y}_j^{(i)}$	are the predicted labels for the i[th] training example and j[th] class
$\wedge$	is the logical AND operator
k	is the number of classes

Macro averaged recall

This extends the precision metric to multi-label classification. This metric computes the precision of individual class labels and then averages over all classes.

Mathematically,

$$\lambda - Recall\,(R_{Macro}^{j}) = \frac{\sum_{i=1}^{n}[y_j^{(i)} \wedge \hat{y}_j^{(i)}]}{\sum_{i=1}^{n}[\,y_j^{(i)}]}$$

$$Recall_{Macro} = \frac{1}{k}\sum_{j=1}^{k}(R_{Macro}^{j})$$

Where:

n	is the number of training examples
$y_j^{(i)}$	are the true labels for the i[th] training example and j[th] class
$\hat{y}_j^{(i)}$	are the predicted labels for the i[th] training example and j[th] class
$\wedge$	is the logical AND operator
R_{macro}^{j}	is the Recall for label/class k
k	is the number of classes

Micro averaged recall

This extends the precision metric to multi-label classification. This label-based metric computes the precision globally over all instances and all class labels.

Mathematically,

$$Recall_{Micro} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}[y_j^{(i)} \wedge \hat{y}_j^{(i)}]}{\sum_{j=1}^{k}\sum_{i=1}^{n}y_j^{(i)}}$$

Where:

n is the number of training examples

$y_j^{(i)}$ are the true labels for the i[th] training example and j[th] class

$\hat{y}_j^{(i)}$ are the predicted labels for the i[th] training example and j[th] class

$\wedge$ is the logical AND operator

k is the number of classes

Negative predictive value

The negative predictive value (NPV) is a metric that computes the fraction of true negative predictions among the outcomes that the model classified as negative.

This is useful for use cases where the false negative predictions are costly.

$$\text{NPV} = \frac{\text{True negative}}{\text{True negative} + \text{False negative}}$$

3.4.2 Ranking metrics

Receiver operating characteristic curve

The ROC curve is a graphical plot used to summarize the diagnostic ability of a classification model. It is created by plotting the TPR (sensitivity) against the false positive rate (FPR, 1 − specificity). It was created primarily for binary classification, but it can be generalized for multiclass classification. The AUC can be calculated and used as a single score to summarize the performance of a model.

Precision-recall curve

The precision-recall curve is also a graphical plot used to summarize the diagnostic ability of a classification model. ROC curves can be misleading with an imbalanced dataset, especially when the number of negative samples is small. A poorly fitted model that simply predicts positive can end with a high AUC score, which would be misleading. In such a scenario, the precision-recall curve and AUC could be used. It is created by plotting the precision score against the recall score (sensitivity).

Average precision

Average precision (AP) is the area under the precision-recall curve (AUC-PR). Precision recall curves are not monotonically decreasing curves, so they are often made so using interpolation methods. Some of the interpolation methods used include 11-point interpolation method and all-point interpolation method.

Mean average precision

Average precision is calculated individually for each class. In an objection task with many classes, mean average precision (mAP) is the average of all the AP values over all the classes. mAP is defined as:

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_i$$

where N is the number of classes.

3.4.3 Probability metrics

Logarithmic loss or cross-entropy

Cross-entropy is a measure of the difference between two probability distributions. A lower score implies a better model, with 0.0 being the best. Log-loss is defined as:

$$\text{cross entropy} = -\sum_{i}^{C} t_i \log(s_i)$$

where t_i and s_i are the ground truth and the model's score for each class i in C.

Brier score

The Brier score is calculated as the mean squared error between the expected probabilities for the positive class (e.g., 1.0) and the predicted probabilities. It ranges between 0.0 and 1.0.

$$\text{Brier score} = \frac{1}{N}\sum_{i}^{N}(g_i - p_i)^2$$

where expected values are p_i and the predicted values are g_i.

Brier skill score

In order to more appropriately compare the Brier score of different models, the Brier score can be scaled against a reference, such as the score of no skill model.

$$\text{Brier skill score} = 1 - \left(\frac{\text{Brier score}}{\text{Brier score reference}}\right)$$

Intersection over union

Intersection over union (IoU) evaluates the intersection between the predicted bounding box of an object detection model, and the ground truth bounding box. It is calculated as the area of overlap between the ground truth bounding box (*gt*) and the predicted bounding box (*pb*), divided by the area of the union of *gt* and *pb.* IoU metric ranges from 0 and 1 with 0 meaning no overlap and 1 implying a perfect overlap between *gt* and *pb.*

$$\text{IoU} = \frac{\text{area}(gt \cap pb)}{\text{area}(gt \cup pb)}$$

Hamming loss

Hamming loss is used to calculate the proportion of incorrectly predicted labels to the total number of labels. When applied to multi-label classification, it is used to calculate the number of false positives and false negative per instance and then average it over the total number of training samples.

Mathematically,

$$\text{Hamming Loss} = \frac{1}{nL}\sum_{i=1}^{n}\sum_{j=1}^{L}[I(y_j^{(i)} \neq \hat{y}_j^{(i)})]$$

Where:

n is the number of training examples

$y_j^{(i)}$ are the true labels for the i[th] training example and j[th] class

$\hat{y}_j^{(i)}$ are the predicted labels for the i[th] training example and j[th] class.

α-Evaluation score

Alpha evaluation score is a generalized form of the Jaccard similarity for evaluating each multi-label prediction. The α-evaluation score provides a flexible way to evaluate multi-label classification results for both aggressive as well as conservation tasks.

Mathematically,

$$\alpha-\text{ evaluation score } = \left(1 - \frac{|\beta \boldsymbol{M_x} + \gamma \boldsymbol{F_x}|}{|\boldsymbol{Y_x} \vee \boldsymbol{P_x}|}\right)^{\alpha}$$

$$(\alpha \geq 0, 0 \leq \beta, \gamma \leq 1, \beta = 1 | \gamma = 1)$$

Where:

$\boldsymbol{M_x}$ is the number of missed labels/ false negatives

$\boldsymbol{F_x}$ is the number of false positives

$\boldsymbol{Y_x}$ is the number of positive samples in the true labels (TP+FN)

$\boldsymbol{P_x}$ is the number of positive samples in the predicted labels (TP+FP)

$\vee$ i s the logical OR operator

3.5 Undisclosed test data set collection

Undisclosed test data was provided by Vasantha Kumar Venugopal. The use case was the diagnosis of COVID-19 via chest x-ray. The dataset contained 917 cases, with 436 Real-Time Reverse Transcription Polymerase (RTPCR) confirmed positive cases, and 481 COVID negative cases. The dataset was collected from Mahajan Imaging in India.

– raw data acquisition/acceptance;
– test data source(s): availability, reliability;
– labelling process/acceptance;
– bias documentation process;
– quality control mechanisms;
– discussion of the necessary size of the test data set for relevant benchmarking results;
– specific data governance derived by general data governance document [115].

3.6 Benchmarking methodology and architecture

– technical architecture;
– hosting;
– possibility of an online benchmarking on a public test dataset;
– protocol for performing the benchmarking (who does what when etc.);
– AI submission procedure including considerations on contracts, rights, intellectual property, etc.

3.6.1 Audit trial

An audit trial was conducted using the undisclosed test data for the diagnosis of COVID-19 via chest x-ray. The machine learning auditing platform from the Open Code Initiative; health.aiaudit.org was used. This platform will automate the assessment of AI systems. See Figure 12.

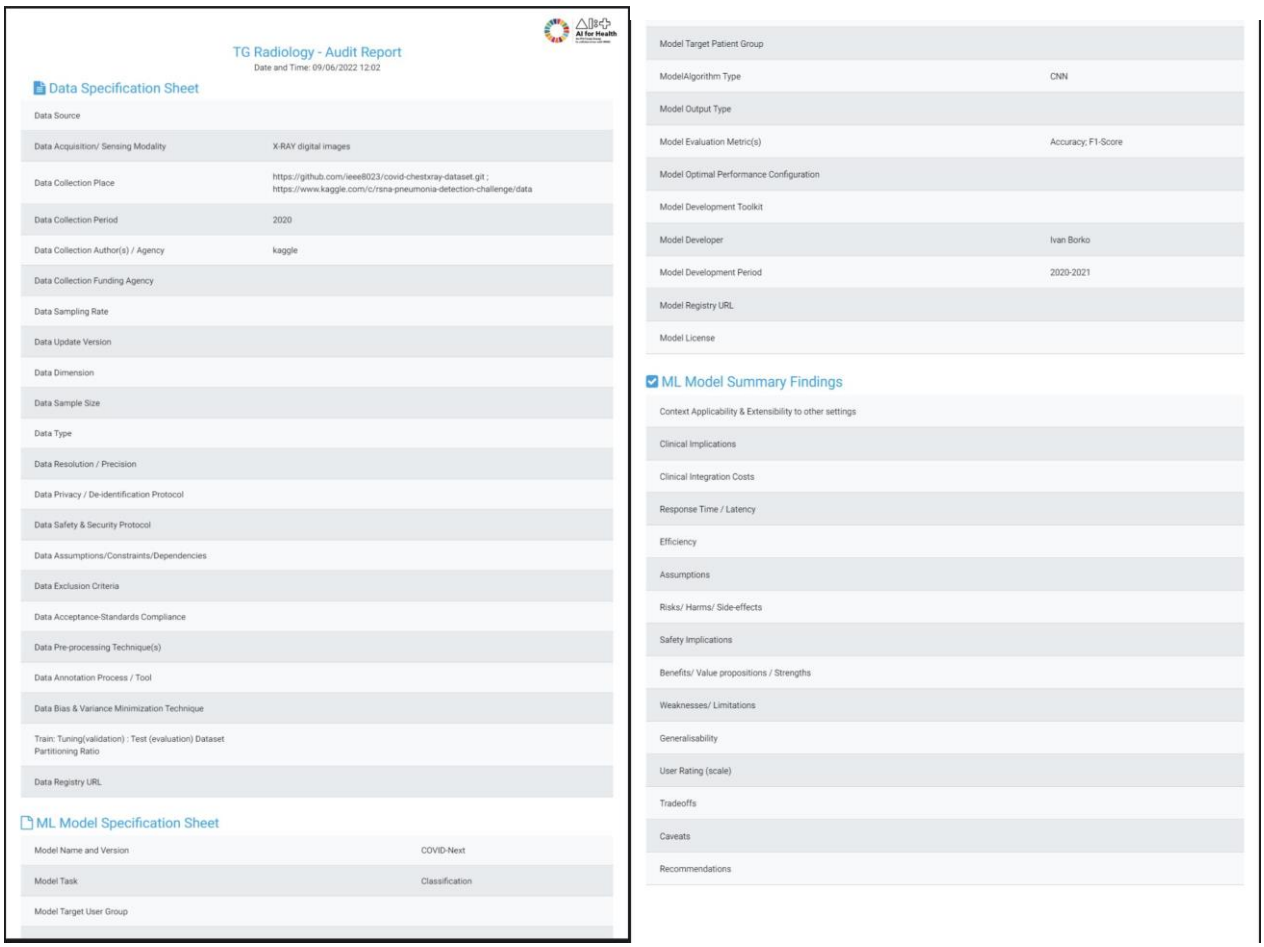


Figure 12 – Model results after trial audits using the benchmarking platform, health.aiaudit.org

3.6.2 Audit trial checklist

An audit checklist was adapted from the FG, as part of the audit trial.

Descriptions of items on the checklist follow.

Working draft

Table 5 consists of a minimum viable set of audit verification checklist items. This checklist is basically derived from the FG-AI4H standardized model survey questionnaire [116]. It is evident from Table 5 that the checklist items are categorized on the basis of their respective Machine Learning for Health (ML4H) lifecycle stage, the applicable assessment criteria and the assessment type they signify.

NOTE – Each audit team is free to expand, extend and modify the existing set of checklist items based on their TG or use-case specific considerations and relevance.

Task description

1. Please perform an expert review of the given checklist items and possibly try to provide expert assessment feedback based on the following questions:

NOTE – All your expert responses can be marked directly on to the editable working document in the 'Remarks' column of the table.

a) Is the given set of checklist items comprehensive enough and does it cover all the relevant ML4H lifecycle requirements (ML technology, clinical, regulatory and ethical). If "No", please indicate the missing aspects.

b) From the given set, are there any checklist items that are conflicting or ambiguous to the determining context and hence need further clarification, correction, modification or substitution? If "Yes", please indicate them.

c) From the given set, are there any checklist items that identify as not applicable or not valid to a particular TG or use case? If "Yes", please indicate them along with the respective exclusion criteria.

d) Are any additional checklist items proposed? If "Yes", please indicate them along with the respective inclusion criteria.

2. Based on expert assessment, please in column 6 assign a significance level or conformance priority to each of the checklist items listed. A first level criterion could be to assess the expected conformance significance of a particular checklist item with respect to the applicable ML4H regulations, laws, standards, guidelines and best practices.

The significance level may be assigned a categorical label from among the following four types: mandatory; preferred; conditional; or optional based on its TG or use-case specific significance.

Purpose

This set of verification checklists is reviewed, finalized, vetted and approved by the audit experts. This approved set of checklists then serves as a questionnaire for TG use case developers to fill in their response. The response or results are verified (with the help of quantitative and qualitative records, proofs or evidence) and validated (by applicable test cases) for conformity assessment to generate a final audit report.

NOTE – Since this set of checklists serves as a common interface to both use case developers and the audit experts for the process of designing the check list or questionnaire, both parties (audit experts and TG or domain experts) are encouraged to collaborate on this so that there is consensus and less confusion at the real audit time.

Table 5 – Draft audit verification checklist

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute or metric	Significance level	Verification and validation record or proof
Planning	Regulatory assessment	Qualitative	Product name and version	Intended use or product specification	Mandatory	
Planning	Regulatory assessment	Qualitative	Target clinical intervention area of the product, e.g., – prevention – screening – diagnosis – treatment – triage – prognosis – other.	Intended use or product specification	Mandatory	
Planning	Regulatory assessment	Qualitative	Primary product function – primary function – secondary function (if applicable), e.g., – classification – prognosis – matching – labelling – detection – segmentation – recommendation – data modelling – other.	Intended use or product specification	Mandatory	
Planning	Regulatory assessment	Qualitative	Product category – software-as-a-medical device (SaMD) – software-as-a-medical service (SaMS) – software-in-a-medical device (SiMD) – mobile medical applications (MMA) – medical device data systems (MDDS) – other.	Intended use or product specification	Preferred	

Table 5 – Draft audit verification checklist

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute or metric	Significance level	Verification and validation record or proof
Planning	Regulatory assessment	Qualitative	Product user group – primary – secondary(if applicable)	Intended use or product specification	Mandatory	
Planning	Regulatory assessment	Qualitative	Product operational mode – fully automatic – semi-automatic	Intended use or product specification	Mandatory	
Planning	Regulatory assessment	Qualitative	Product autonomy level (based on IMDRF - risk acceptance criteria & criticality of the clinical use case or any other standard control baselines for clinical system level risk assessment)	Intended use or product specification	Mandatory	
Data collection	Technical validation	Qualitative	From where and when was the training dataset collected? Place: Time Period:	– Social representation bias – Historical data bias	Preferred	
Data collection	Technical validation	Quantitative	How many total data samples does the source dataset contain?	Sampling bias	Preferred	
Data collection	Technical validation	Quantitative	Did you encounter any missing data in the source dataset? If yes, please specify affected variables, missing fraction relative to all entries.	Sampling bias	Preferred	
Data collection	Technical validation	Quantitative	Whether the data acquisition modality, the data inclusion and the data exclusion criteria were properly validated to find if there is any mismatch between 'reported' sample size and 'actual 'reproduced' sample size?	Data reproducibility		
Data collection	Regulatory assessment	Qualitative	Does the data identify any subpopulations Or Does the dataset contain confidential/personal information? (age-group, gender, ethnicity, religion, etc.)? If yes, specify the type	Data privacy	Mandatory	
Data collection	Regulatory assessment	Qualitative	Did you obtain consent from individuals who are represented in this data to use their information for this purpose? If "Yes", were they provided with any mechanism to revoke their consent in the future or for specific uses?	Data privacy and protection Patient safety	Mandatory	

Table 5 – Draft audit verification checklist

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute or metric	Significance level	Verification and validation record or proof
Data collection	Regulatory assessment	Qualitative	Whether any due diligence and processes were followed in conformance to institutional review and ethical review policies when input datasets were de-identified or anonymized? Or were any exemptions obtained under special conditions?	Data privacy and protection	Mandatory	
Data preparation	Technical validation	Quantitative	How many instances of each label class were present in the training dataset? (e.g., proportionate sample size of different classes)	Sampling bias	Preferred	
Data preparation	Technical validation	Quantitative	If ground truth annotation was used as the basis for data labelling quality control, how did you evaluate the quality of ground truth annotation?	Data labelling bias	Mandatory	
Data preparation	Technical validation	Quantitative	For data labelling, how were the perceptual errors and biases accounted for? Was inter-annotator reliability measured as part of a quality check and what is its specification?	Data labelling bias	Preferred	
Data preparation	Technical validation	Quantitative	By which proportion did you split the preprocessed data samples into a training set, the validation (tuning) set and the test set?	Data bias leading to ML model under- or over-fitting	Mandatory	
Data preparation	Technical validation	Qualitative	Do you ensure that there is no patient sample overlap among the training, the validation (tuning) and the test datasets?	Sampling bias	Mandatory	
Data preparation	Regulatory assessment	Qualitative	Is it possible to identify individuals from the dataset? Were the datasets de-identified or anonymized? (Yes or No)	Data privacy	Mandatory	
Data preparation	Regulatory assessment	Qualitative	Type and level of de identification used like complaint removal under the US Health Insurance Portability and Accountability Act of 1996 of private DICOM elements, image cropping to avoid identification from reconstructed images, etc.	Data privacy	Mandatory	
Data preparation	Regulatory assessment	Qualitative	How do you justify the selection of ground truth?	Data labelling quality	Preferred	

Table 5 – Draft audit verification checklist

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute or metric	Significance level	Verification and validation record or proof
Data preparation	Technical validation	Qualitative	Is the prevalence of the real world disease types or conditions reflected in the configuration of train datasets? (e.g., relative frequency of disease and non-disease types in the dataset)	Data bias	Preferred	
Model training	Technical validation	Qualitative	Have you evaluated the influence of particular input data features that positively affects the model performance scores?	Model performance	Mandatory	
Model tuning	Clinical evaluation	Quantitative	Are decision thresholds being used for classification? If yes, specify the thresholds and the thresholding rule. Can you also state the clinical significance of the selected operating threshold, if any?	Technical vs clinical accuracy equivalence	Mandatory	
Model tuning	Regulatory assessment	Qualitative	Is your ML model optimized for a specific local or clinical setting (e.g., a specific clinical department, country)?	Model generalizability	Mandatory	
Model tuning	Technical validation	Qualitative	Does your use case give high importance to the most prevalent output class types and thus optimize the model performance? Alternatively, does your use case give equal prominence to each output class type?	Model optimization	Preferred	
Model evaluation	Clinical evaluation	Qualitative	Were patients and clinicians involved or consulted during the ML algorithm selection stage, algorithm development stage or algorithm acceptance and adoption stage?	Model explicability	Mandatory	
Model evaluation	Technical validation	Quantitative	Are there output classes or disease types for which the ML model performed worse than others? Provide the confusion matrix results.	Model performance	Mandatory	
Model evaluation	Technical validation	Quantitative	Is there an interpretability-performance trade-off observed. If yes, provide the comparative analysis results.	Model interpretability and model performance tradeoff	Preferred	
Model evaluation	Technical validation	Quantitative	Specify the guarantees and limits of the performance metrics used for model evaluation	Model performance	Preferred	
Model evaluation	Technical validation	Quantitative	Specify the guarantees and limits of the gold or reference standard against which the performance metrics are evaluated	Model performance	Preferred	

Table 5 – Draft audit verification checklist

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute or metric	Significance level	Verification and validation record or proof
Model evaluation	Clinical evaluation	Qualitative	Specify the selection criteria of the performance metrics used for model evaluation. – clinical significance – optimization – specialization – generalization – other.	Technical accuracy vs clinical effectiveness equivalence	Mandatory	
Model evaluation	Clinical evaluation	Quantitative	Whether any comparative analysis was done over the model safety risks with that of the alternative technologies (both ML and non-ML based)	Patient safety	Preferred	
Model evaluation	Clinical evaluation	Qualitative	Have you used any methods that are specific or agnostic to the model for interpretability?	Model interpretability	Mandatory	
Model evaluation	Technical validation	Quantitative	Have you estimated the risk probabilities associated with model performance variability when tested against the following conditions: – non-specified use environment – non-specified hardware and software configurations – patients of different age, sex, race, co-morbidities – patients with different severity of disease type – other.	Model uncertainty and robustness	Mandatory	
Model usage or deployment	Clinical evaluation	Quantitative	Specify the computational efficiency of the model in terms of the response time	Clinical efficiency	Preferred	
Model usage or deployment	Clinical evaluation	Qualitative	How does the ML model adoption reduce the overall clinical practice cost (or enhance the clinical practice savings)? – faster patient diagnosis or treatment – percentage reduction in clinician cognitive workload – degree of automation or semi-automation introduced – degree of smartness or intelligence augmentation – new knowledge discovery – enabling replacement or redefinition of existing gold standard – other.	Clinical integration	Mandatory	

Table 5 – Draft audit verification checklist

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute or metric	Significance level	Verification and validation record or proof
Model usage or deployment	Clinical evaluation	Qualitative	What is the care quality impact delivered by the ML model? – early detection and lowering of disease severity levels – increased coverage under screening programs – workflow efficiency – reliability and reproducibility of outcomes – increased accessibility – increased patient and clinician satisfaction – other	Clinical effectiveness	Mandatory	
Model usage or deployment	Clinical evaluation	Qualitative	How does the model fit into the intended health intervention workflow? – autonomous tool – assistive tool – augmentative tool – add-on unit to existing system/workflow – replacement unit for existing – system/workflow component – new stand-alone application – other	Clinical integration	Mandatory	
Model usage or deployment	Clinical evaluation	Qualitative	Have you estimated the risk probabilities associated with the potential hazards and harms as a consequence of a model not meeting the expected or desired performance specification? In addition, have you specified the values or ranges for performance metrics in order to avoid unacceptable risks?	Patient safety	Mandatory	
Model usage or deployment	Clinical evaluation	Qualitative	Was input data feature importance validated for its significance in the clinical setting by the clinician or specialist? Which features were ranked as most important?	Model interpretability	Preferred	
Model usage or deployment	Clinical evaluation	Qualitative	Did the model fail to address any relevant clinically important findings?	Clinical effectiveness	Preferred	
Model usage or deployment	Clinical evaluation	Quantitative	Is there a comparative analysis done on the patient outcomes for (1) patients on whom the ML model is applied versus (2) patients on whom the ML model is not applied?	Clinical effectiveness	Mandatory	

Table 5 – Draft audit verification checklist

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute or metric	Significance level	Verification and validation record or proof
Model usage or deployment	Regulatory assessment	Qualitative	Whether any safety control measures were incorporated to deal with unintended consequences (if any) of ML model intervention in the clinical setting?	Operating environment risks or patient safety	Mandatory	
Model maintenance and versioning	Regulatory assessment	Qualitative	Is the ML model maintained as (a) a static system or (b) a continuously learning system? I	Model maintainability	Mandatory	
Model maintenance and versioning	Regulatory assessment	Quantitative	If the ML model is attributed to a continuous learning system , specify the algorithm change or update cycle	Model maintainability	Mandatory	
Model maintenance and versioning	Regulatory assessment	Quantitative	Has there been a proper plan for test data quality and correctness assessment after model deployment (i.e., concept drift, training/test data distribution mismatch, etc.)?	Model maintainability	Mandatory	

3.6.3 Audit trial: minoHealth.ai: A clinical evaluation of deep learning systems for the diagnosis of pleural effusion and cardiomegaly in Ghana, Vietnam and the USA

Background: A rapid and accurate diagnosis of cardiomegaly and pleural effusion is of the utmost importance to reduce mortality and medical costs. AI has shown promise in diagnosing medical conditions. The aim of this study is to evaluate how well AI systems, developed by minoHealth AI Labs, will perform at diagnosing cardiomegaly and pleural effusion, using chest x-rays from Ghana, Vietnam and the USA, and how well AI systems will perform when compared with radiologists working in Ghana.

Method: The evaluation dataset used in this study contained 100 images randomly selected from three datasets. Twenty images were selected from the VinBig Data Chest x-ray dataset [48], another 21 images were selected from the CheXpert [92] dataset and 59 images were selected from the Euracare dataset, an in-house dataset collected by minoHealth AI Labs from Euracare Advanced Diagnostics and Heart Centre, Accra, Ghana. The deep learning models were further tested on a larger Ghanaian dataset containing 561 samples. Two AI systems were then evaluated on the evaluation dataset, while the same chest x-ray images within the evaluation dataset to were given to four radiologists, with 5-20 years experience, to diagnose independently.

Results: For cardiomegaly, minoHealth.ai systems scored AUC-ROC values of 0.90 and 0.97, while the AUC-ROC of individual radiologists ranged from 0.77 to 0.87. For pleural effusion, the minoHealth.ai systems scored 0.97 and 0.91, whereas individual radiologists scored between 0.75 and 0.86. On both conditions, the best performing AI model outperforms the best performing radiologist by about 10%. We also evaluate the specificity, sensitivity, NPV and positive predictive value between the minoHealth.ai systems and radiologists.

Conclusion: In regions like Sub-Saharan Africa, where radiologists are scarce and are also overloaded with other clinical responsibilities, solutions like the minoHealth.ai systems will be of great utility. These solutions can achieve the performance of multiple radiologists working together to complement the efforts of radiologists and ease the burden on them.

3.6.4 Benchmarking solution

A radiograph-agnostic benchmarking platform and framework are proposed that would allow for the evaluation of AI radiological systems for various conditions and serve as a standard. This would require registered developers and organizations seeking to evaluate their AI system to download the test images and a comma-separated values (CSV) file with two columns: ID – containing the unique Identification of each test image; and class – that would be left blank in order to be populated by the outputs of an AI system. Developers are then to submit the fully populated CSV file, which would then provide outputs of the model to be evaluated with the true labels. Tutorial scripts in popular machine learning libraries and frameworks would be provided to developers on how to correctly get model outputs to be populated in the CSV file.

See Figures 13 and 14.

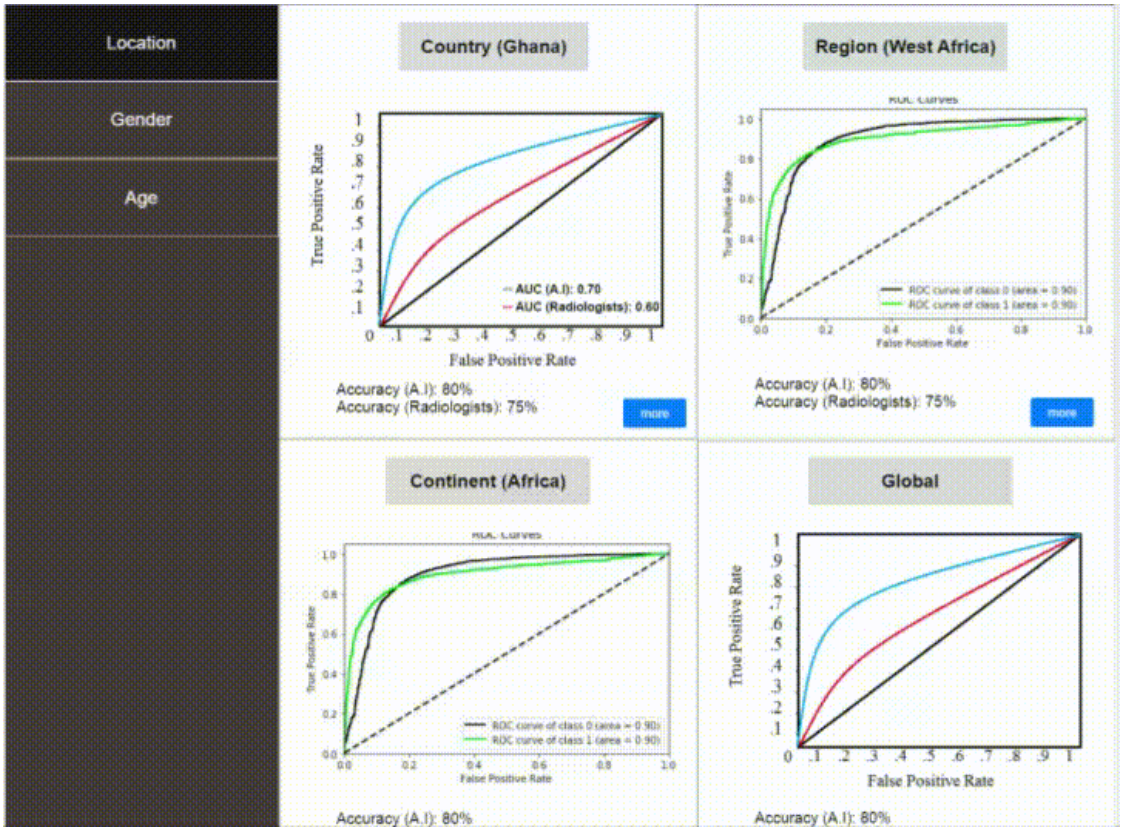


Figure 13 – A prototype of the radiograph-agnostic precision evaluation platform

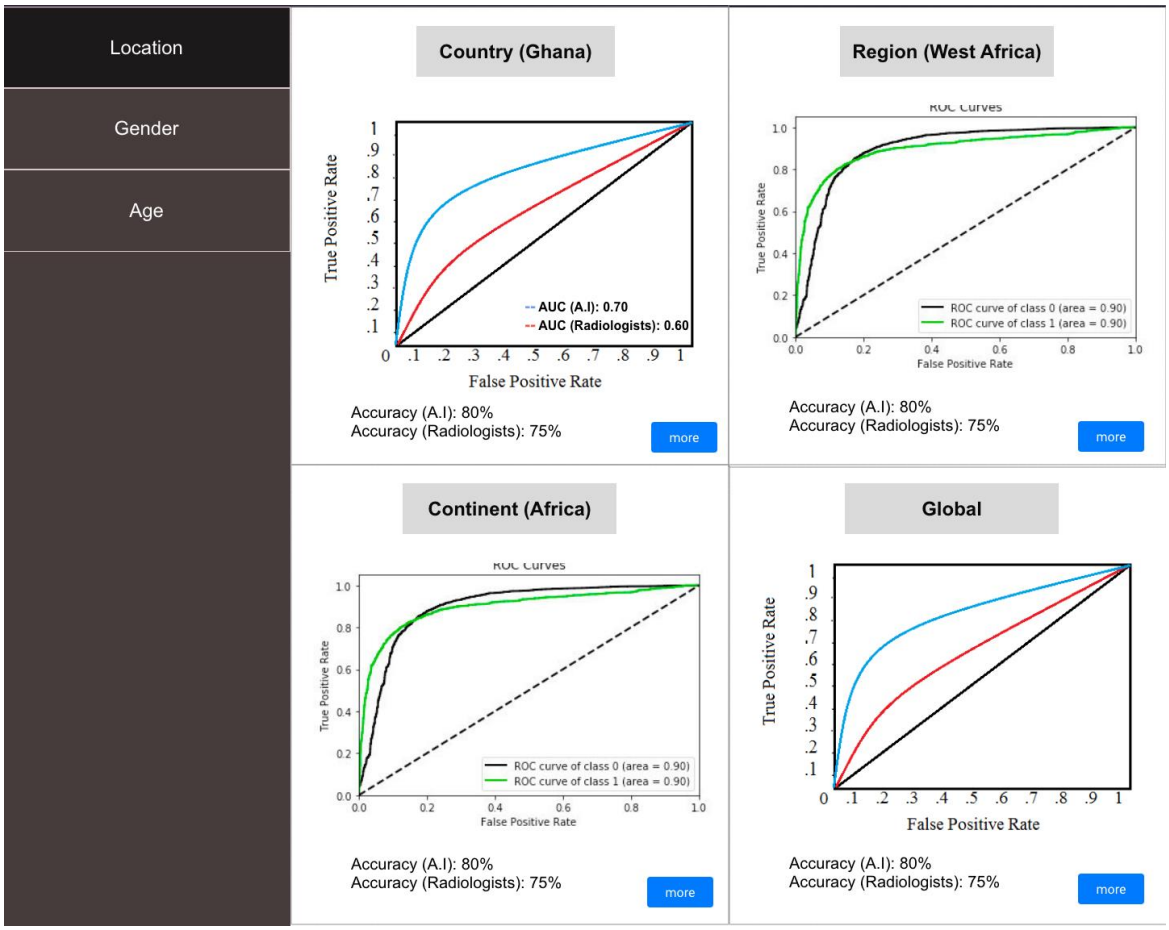


Figure 14 – The location category with its sub-categories and the metrics used

3.6.5 Evaluation metrics

All our supported condition tests on the platform would be image classification tasks and therefore evaluation metrics would be used for classification. Some of the conditions and tests would be binary while others would be multi-class classification tasks, therefore metrics would be used that are suitable for both. As shown in Figures 1 and 2, the evaluation metrics to be used would be the ROC curve, its AUC score and the accuracy score. The ROC curve and AUC score would help to identify the model's TPR (sensitivity) and its FPR (1 − specificity). Though originally for binary classification, the ROC curve and AUC score can be generalized to multi-class classification.

The performance of an AI system would be compared with radiologists using the various metrics. This would help developers see how well their models perform compared to the current popular approach, standalone radiologists. Benchmarking vis-à-vis radiologists would also help in assessing the level of autonomy that should be given to each AI system. See Figure 15.

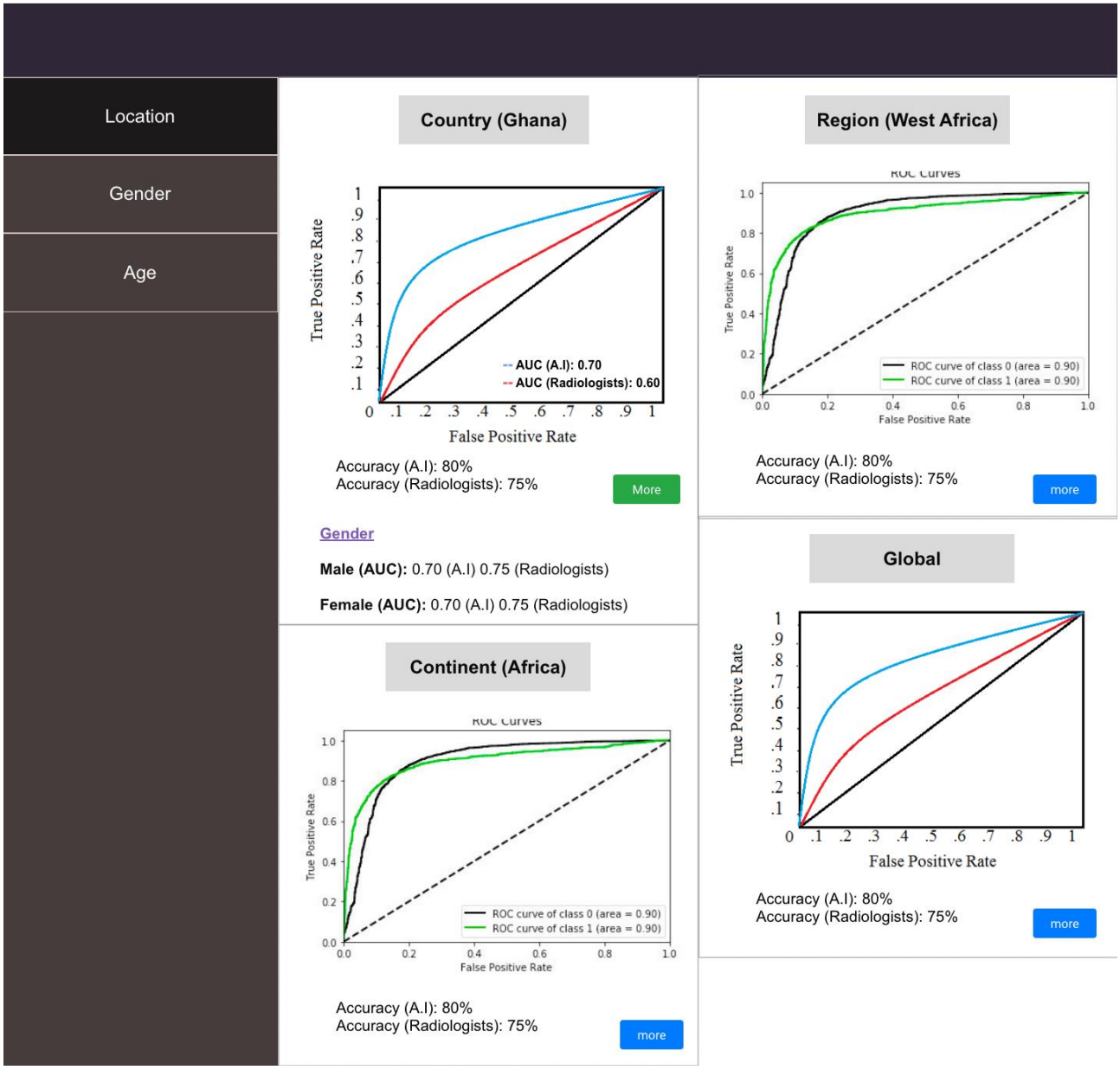


Figure 15 – Each sub-category would feature demographics intersection performances too

3.6.6 Benchmark categorizations

The evaluation results would be divided into location, gender and age, as shown in Figure 16. Under location, the performance of the AI model would be shown under the sub-categories: country; continent; region; and global. The country sub-category shows the performance of the AI system within the nation in which it was developed. The continent sub-category would show how well the model performs on data from the continent in which it was developed; this would help the developers know how well they can scale the current version of their AI system. Region specifically focuses on the performance of the AI system within the sub-continental region in which it was developed (e.g., West Africa, South East Asia, Northern Europe). This would help the developers see how ready their AI system is to be deployed in neighbouring countries. Finally, global shows how well the model performs on data from across the world, showing its ability to truly generalize. Each sub-category under location would also feature an AUC score for each gender and age group, as shown in Figures 1 and 3. This would allow developers to tell specifically how well their AI system generalizes across gender and age within each geographical area.

Under Gender, there would be two main sub-categories, male and female, as shown in Figures 16 and 4. This would show how well the AI system performs on radiographs of male and female patients. Each of the two sub-categories would also feature AUC scores for various age groups. This would show how well the AI system performs on male and female patients of different age groups. Conditions that, however, only affect one gender would not feature in the gender category.

The age category would feature various age groups as sub-categories. Age groups that are not featured within certain datasets and conditions would not be shown for those specific conditions. Similar to the other categories, AI system performance on each of the age groups would be shown and it would also feature male and female AUC score under each age group.

This concept of precision evaluation is to precisely assess how well an AI system generalizes across demographics.

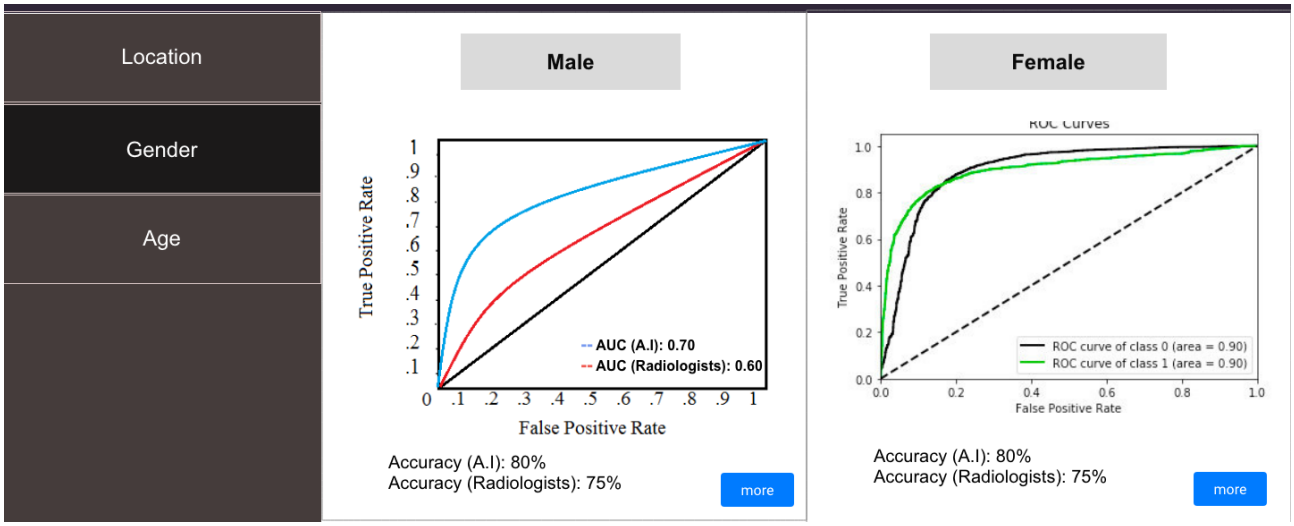


Figure 16 – The gender category

3.6.7 Evaluation data

The goal is to ensure a proportional amount of the diverse demographics and their intersections. With diverse evaluation data, the generality of an AI system can truly be assessed. The platform would be open to facilities to register, and submit images and demographical data. Facilities with approved images would be credited with contributing to the set-up of such dataset. This would hopefully serve as an incentive to facilities to contribute more data to the platform. Submitted radiographs should be accompanied by a CSV file with information about patient gender, age and imaging facility location. This would allow for the proposed precision evaluation framework.

3.6.8 Panel of expert radiologists

To ensure quality, submitted images and data would be reviewed by a panel of expert radiologists. This panel of expert radiologists would also ensure that borderline cases and diversity are represented in each evaluation set. The panel would be open to qualified radiologists to join and participate in. Each evaluation set and condition would have its own panel of expert radiologists. Radiologists who are part of the panel would be credited on the platform for the evaluation sets to which they contribute. This would also hopefully serve as an incentive for more radiologists to join the panel of expert radiologists.

3.6.9 Test radiologists

Beyond the panel of expert radiologists, ideally radiologists from different parts of the world would be available who would be asked to classify the test images without access to their true labels. The goal would be to get as many testing radiologists as possible from each continent, region or possibly country. These radiologists would also be ideally given test images from within their region. This would allow a comparison of AI system performance on test images within each location sub-category with radiologists also within such geographical regions. This would more appropriately help to estimate how well an AI system performs when compared with the level of performance of standalone radiologists within each specific region.

3.7 Evaluation data availability

minoHealth AI Labs is currently working with institutions in Ghana, including Christian Health Association of Ghana (CHAG), National Catholic Health Service (NCHS), Euracare Advanced Diagnostic Center and Paradise Diagnostic Center in order to collect mammograms and chest radiographs. Some of those data can be made available to the benchmarking platform. With the collaboration of various members and organizations affiliated with FG-AI4H, more radiographs can be collected from around the world. Also as explained earlier, the platform would be open to registered facilities to contribute data.

3.8 Feasibility

Though the proposed radiograph-agnostic framework and platform has several moving parts and complexities, it is possible to modularize it and build with different levels of complexity. It is also possible for the categories and sub-categories to adjust based on the number and diversity of samples as well as the radiologists available. If the evaluation data for a particular condition is not large enough to support all four location sub-categories, it can be limited to just region or continent and global. If there were not enough test radiologists within a specific country where an AI system was developed, the regional, continental or global average performance of radiologists would be used. A similar principle can apply to the sub-categories of gender and age. Implementation of the platform would be started with chest x-rays for 12 different thoracic diseases supported in MIMIC-CXR, [96] CheXpert [92] and NIH CXR 8 [97] datasets.

3.9 Privacy and security

Anonymized data can be de-anonymized using techniques like linkage attacks, which involve combining data from multiple sources in order to form a whole picture about targets. It is then possible to use the demographics data (date of birth, gender and location) of an anonymized patient whose medical image is available and cross-reference with public voter lists in order to identify who the patient is. This is because there are very few individuals likely to have the same data of birth and gender, and live in the same location. To prevent linkage attacks, developers and testing radiologists are only given access to test images without demographics data. To further defend against this attack, date of birth is abstracted to just age (in years) of the patient when they were imaged, and location to country. To add additional security measures as long as the panel of expert radiologists has access to such demographics data, variations of differential privacy can be explored.

Also, we are ensuring a secure system by demanding that developers and organizations that require a standardized evaluation of their AI systems register before they would be allowed to. The registration process can include an in-person assessment by their local World Health Organization (WHO) or International Telecommunication Union (ITU) branch office, just to ensure they are a valid institution, start-up or developer. A moderate fee can be charged for the registration, which could then serve as funds to support the maintenance of the platform. Equally, health facilities seeking to donate medical images and data must register and be assessed. In addition, even the images and data they submit to the platform would be evaluated before being added to the system. All radiologists, both in the expert panel and the testing community would have to register and be verified before being allowed to contribute to the platform.

In order to not infringe upon their IP rights, AI developers and organizations would not be required to submit their AI system itself, but only the output (CSV files) of their AI system, which would then be used for its evaluation.

3.10 Impact

There is a large number of publicly available medical image datasets online, into which there has been a lot of research and development. By developing frameworks that target these conditions first, the standardized benchmarking platform would be made immediately appealing to the AI healthcare research and development community. This would also help speed up the deployment of AI solutions in radiology globally. AI healthcare system developers and organizations usually have to go through the challenge of convincing health facilities to share their private data with them; such data unfortunately are not always of high quality and they usually lack the broad demographic representations needed to truly assess how well an AI system generalizes. A radiograph-agnostic benchmarking platform with data from various facilities across the globe, reviewed by a panel of experts to ensure quality and diversity, would drastically simplify the evaluation stage of such AI systems. The precision evaluation framework would help fight against demographically biased AI systems by ensuring they are tested in great detail across various groups. It would also help in the safe scaling of AI systems across different locations. The location sub-categorization of evaluation allows for geo-precision evaluation. Developers can tell how well their systems can perform within their country or first-point of deployment, and should they intend to scale to neighbouring countries then eventually have it across the globe, they can tell how well their current version would perform at each point of such growth and scaling.

3.11 Reporting methodology

– Report publication in papers or as part of ITU documents
– Online reporting
– Public vs private leaderboards
– Approval like a credit check for sharing with selected stakeholders
– Report structure including an example
– Frequency of benchmarking.

4 Results

– Insert here the reports of the different benchmarking runs.

5 Discussion

– Discussion of the insights from executing the benchmarking on
    • external feedback on the whole topic and its benchmarking
    • technical architecture

- data acquisition
- benchmarking process
- benchmarking results
- field implementation success stories.

References

[1] Kersting K. Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines. *Front Big Data*. 2018;**1**:6. DOI: 10.3389/fdata.2018.00006.

[2] Fielding AH. An introduction to machine learning methods. In; Fielding AH, editor. *Machine learning methods for ecological applications*,pp. 1–35. Boston, MA: Springer. 1999. DOI: 10.1007/978-1-4615-5289-5_1.

[3] Schmidhuber J. Deep learning in neural networks: An overview. *Neural Networks*. 2015;**61**:85–117. DOI: 10.1016/j.neunet.2014.09.003.

[4] Brady AP, Neri E. Artificial intelligence in radiology − Ethical considerations. *Diagnostics (Basel)*. 2020;**10**(4):231. DOI: 10.3390/diagnostics10040231.

[5] Akinci D'Antonoli T. Ethical considerations for artificial intelligence: An overview of the current radiology landscape. *Diagn Interv Radiol*. 2020;**26**(5):504–11. DOI: 10.5152/dir.2020.19279.

[6] Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, Langer SG, Kitts AB, Birch J, Shields WF, van den Hoven van Genderen R, Kotter E, Gichoya JW, Cook TS, Morgan MB, Tang A, Safdar NM, Kohli M. Ethics of artificial intelligence in radiology : Summary of the Joint European and North American Multisociety statement. *Can Assoc Radiol J*. 2019;**70**(4):329-34. DOI: 10.1016/j.carj.2019.08.010.

[7] Parveen NR, Sathik MM. Detection of pneumonia in chest x-ray images. *J x-Ray Sci Technol*. 2011;**19**:423–8. DOI: 10.3233/XST-2011-0304.

[8] Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, Dong J, Prasadha MK, Pei J, Ting MYL, Zhu J, Li C, Hewett S, Dong J, Ziyar I, Shi A, Zhang R, Zheng L, Hou R, Shi W, Fu X, Duan Y, Huu VAN, Wen C, Zhang ED, Zhang CL, Li O, Wang X, Singer MA, Sun X, Xu J, Tafreshi A, Lewis MA, Xia H, Zhang K. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;**172**:1122-1131.e9. DOI: 10.1016/j.cell.2018.02.010.

[9] Kitamura G, Deible C. Retraining an open-source pneumothorax detecting machine learning algorithm for improved performance to medical images. *Clin Imaging*. 2020;**61**:15–9. DOI: 10.1016/j.clinimag.2020.01.008.

[10] Filice RW, Stein A, Wu CC, Arteaga VA, Borstelmann S, Gaddikeri R, Galperin-Aizenberg M, Gill RR, Godoy MC, Hobbs SB, Jeudy J, Lakhani PC, Laroia A, Nayak SM, Parekh MR, Prasanna P, Shah P, Vummidi D, Yaddanapudi K, Shih G. Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest x-ray dataset. *J Digit Imaging*. 2020;**33**:490–6. DOI: 10.1007/s10278-019-00299-9.

[11] Qin C, Yao D, Shi Y, Song Z. Computer-aided detection in chest radiography based on artificial intelligence: A survey. *Biomed Eng Online*. 2018;**17**:113. DOI: 10.1186/s12938-018-0544-y.

[12] Kumar A, Wang YY, Liu KC, Tsai IC, Huang CC, Hung N. Distinguishing normal and pulmonary edema chest x-ray using Gabor filter and SVM. *2014 IEEE International Symposium on Bioelectronics and Bioinformatics (IEEE ISBB 2014)*, pp. 1-4. DOI: 10.1109/ISBB.2014.6820918.

[13] Mohd Noor N, Mohd Rijal O, Yunus A, Mahayiddin AA, Gan CP, Ong EL, et al. Texture-based statistical detection and discrimination of some respiratory diseases using chest radiograph. In: *Advances in medical diagnostic technology. Lecture notes in bioengineering*, pp. 75–97. Singapore: Springer,; 2014. DOI: 10.1007/978-981-4585-72-9_4.

[14] Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, et al. Fully automated deep learning system for bone age assessment. *J Digit Imaging*. 2017;**30**:427–41. DOI: 10.1007/s10278-017-9955-8.

[15] Cicero M, Bilbily A, Colak E, Dowdell T, Gray B, Perampaladas K, Barfett J. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Invest Radiol*. 2017;**52**:281–7. DOI: 10.1097/RLI.0000000000000341.

[16] Wang L, Lin ZQ, Wong A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. *Sci Rep.* 2020;**10**:19549. DOI: 10.1038/s41598-020-76550-z.

[17] Apostolopoulos ID, Mpesiana TA. Covid-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Phys Eng Sci Med.* 2020;**43**:635–40. DOI: 10.1007/s13246-020-00865-4.

[18] Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Anal Appl.* 2021;**24**(3):1207-20. DOI: 10.1007/s10044-021-00984-y.

[19] Afshar P, Heidarian S, Naderkhani F, Oikonomou A, Plataniotis KN, Mohammadi A. COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images. *Pattern Recog Lett*. 2020;**138**:638-43. DOI: 10.1016/j.patrec.2020.09.010.

[20] Brunese L, Mercaldo F, Reginelli A, Santone A. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. *Comput Meth Prog Biomed*. 2020;**196**:105608. DOI: 10.1016/j.cmpb.2020.105608.

[21] Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput Biol Med*. 2020;**121**:103792. DOI: 10.1016/j.compbiomed.2020.103792.

[22] Hwang EJ, Park CM. Clinical implementation of deep learning in thoracic radiology: Potential applications and challenges. *Korean J Radiol*. 2020;**21**:511–25. DOI: 10.3348/kjr.2019.0821.

[23] Zhu J, Shen B, Abbasi A, Hoshmand-Kochi M, Li H, Duong TQ. Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs. *PLoS One*. 2020;**15**:e0236621. DOI: 10.1371/journal.pone.0236621.

[24] Matsumoto T, Kodera S, Shinohara H, Ieki H, Yamaguchi T, Higashikuni Y, Kiyosue A, Ito K, Ando J, Takimoto E, Akazawa H, Morita H, Komuro I. Diagnosing heart failure from chest X-ray images using deep learning. *Int Heart J*. 2020;**61**:781–6. DOI: 10.1536/ihj.19-714.

[25] Matsubara N, Teramoto A, Saito K, Fujita H. Bone suppression for chest X-ray image using a convolutional neural filter. *Phys Eng Sci Med*. 2020;**43**:97–108. DOI: 10.1007/s13246-019-00822-w.

[26] Lee S, Choe EK, Kang HY, Yoon JW, Kim HS. The exploration of feature extraction and machine learning for predicting bone density from simple spine X-ray images in a Korean population. *Skeletal Radiol*. 2020;**49**:613–8. DOI: 10.1007/s00256-019-03342-6.

[27] Hu TH, Wan L, Liu TA, Wang MW, Chen T, Wang YH. Advantages and application prospects of deep learning in image recognition and bone age assessment [in Chinese]. *Fa Yi Xue Za Zhi* [Journal of Forensic Medicine] 2017;**33**:629-634**.** DOI: 10.3969/j.issn.1004-5619.2017.06.013.

[28] Li Q, Zhong L, Huang H, Liu H, Qin Y, Wang Y, Zhou Z, Liu H, Yang W, Qin M, Wang J, Wang Y, Zhou T, Wang D, Wang J, Xu M, Huang Y. Auxiliary diagnosis of developmental dysplasia of the hip by automated detection of Sharp's angle on standardized anteroposterior pelvic radiographs. *Medicine (Baltimore)*. 2019;**98**: e18500. DOI: 10.1097/MD.0000000000018500.

[29] Kitamura G. Deep learning evaluation of pelvic radiographs for position, hardware presence, and fracture detection. *Eur J Radiol*. 2020;**130**:109139. DOI: 10.1016/j.ejrad.2020.109139.

[30] Zheng G, Nolte LP. Computer-aided orthopaedic surgery: State-of-the-art and future perspectives. *Adv Exp Med Biol*. 2018;**1093**:1–20. DOI: 10.1007/978-981-13-1396-7_1.

[31] Unberath M, Zaech JN, Gao C, Bier B, Goldmann F, Lee SC, Fotouhi J, Taylor R, Armand M, Navab N. Enabling machine learning in X-ray-based procedures via realistic simulation of image formation. *Int J Comput Assist Radiol Surg*. 2019;**14**:1517–28. DOI: 10.1007/s11548-019-02011-2.

[32] Vaquero JJ, Kinahan P. Positron emission tomography: Current challenges and opportunities for technological advances in clinical and preclinical imaging systems. *Annu Rev Biomed Eng*. 2015;**17**:385-414. DOI: 10.1146/annurev-bioeng-071114-040723.

[33] Mawlawi O, Podoloff DA, Kohlmyer S, Williams JJ, Stearns CW, Culp RF, Macapinlac H; National Electrical Manufacturers Association. Performance characteristics of a newly developed PET/CT scanner using NEMA standards in 2D and 3D modes. *J Nucl Med*. 2004;**45**(10):1734-42.

[34] Shiraishi J, Li Q, Appelbaum D, Doi K. Computer-aided diagnosis and artificial intelligence in clinical imaging. *Semin Nucl Med*. 2011;**41**(6):449-62. DOI: 10.1053/j.semnuclmed.2011.06.004.

[35] Imaging Modalities. (2016, December 02). https://web.archive.org/web/20220119070712/http://www.who.int/diagnostic_imaging/imaging_modalities/en/. Web archive date 2022-01-19

[36] Clarke C. *Overview of imaging modalities*. Nottingham: Radiology Café, 2011-23. Available [viewed 2024-05-14] at: https://www.radiologycafe.com/medical-students/radiology-basics/imaging-modalities

[37] Johns Hopkins Medicine. *Fluoroscopy procedure*. Baltimore, MD: Johns Hopkins University, 2024. Available [viewed 2024-05-14] at: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/fluoroscopy-procedure

[38] Weese J, Penney GP, Desmedt P, Buzug TM, Hill DL, Hawkes DJ. Voxel-based 2-D/3-D registration of fluoroscopy images and CT scans for image-guided surgery. *IEEE Trans Inf Technol Biomed*. 1997;**1**(4):284-93. DOI:10.1109/4233.681173

[39] Bang JY, Hough M, Hawes RH, Varadarajulu S. Use of artificial intelligence to reduce radiation exposure at fluoroscopy-guided endoscopic procedures. *Am J Gastroenterol*. 2020;**115**(4):555-61. doi:10.14309/ajg.0000000000000565

[40] Evirgen Ş, Kamburoğlu K. Review on the applications of ultrasonography in dentomaxillofacial region. *World J Radiol*. 2016;**8**(1):50-8. DOI: 10.4329/wjr.v8.i1.50

[41] Healthline. *Ultrasound*. San Francisco, CA: Healthline Media, 2016. Available [viewed 2024-05-14] at: https://www.healthline.com/health/ultrasound

[42] RadiologyInfo.org. *General ultrasound*. Oak Brook, IL: Radiological Society of North America, 2022. Available [viewed 2024-05-14] at: https://www.radiologyinfo.org/en/info.cfm?pg=genus

[43] UTC. *4 Types of ultrasound imaging*. Ultrasound Technician Center, 2012-23." Available [viewed 2024-05-14] at: https://www.ultrasoundtechniciancenter.org/ultrasound-knowledge/medical-ultrasound-imaging-types.html

[44] Liu S, Wang Y, Yang X, Lei B, Liu L, Li SX, Ni D, Wang T. Deep learning in medical ultrasound analysis: A review. *Engineering*. 2019;**5**(2):261-75. DOI: 10.1016/j.eng.2018.11.020

[45] "Nuclear Medicine - WHO."https://web.archive.org/web/20220202201223/http://www.who.int/diagnostic_imaging/imaging_modalities/dim_nuclearmed/en/. Web archive date 2022-02-02.

[46] RadiologyInfo.org. *General nuclear medicine*, Oak Brook, IL: Radiological Society of North America, 2022. Available [viewed 2024-05-14] at: https://www.radiologyinfo.org/en/info.cfm?pg=gennuclear

[47] Wikipedia. *Nuclear medicine*. San Francisco, CA: Wikimedia Foundation, 2024. Available [viewed 2024-05-14] at: https://en.wikipedia.org/wiki/Nuclear_medicine. Accessed 17 Sep. 2020

[48] Nguyen, H.Q., Lam, K., Le, L.T. et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Sci Data **9**, 429 (2022). DOI: 10.1038/s41597-022-01498-w (Accessed 2025-02-17)

[49] Wikipedia. *Mammography*. San Francisco, CA: Wikimedia Foundation, 2024. Available [viewed 2024-05-14] at: https://en.wikipedia.org/wiki/Mammography.

[50] RadiologyInfo.org. *Mammography*. Oak Brook, IL: Radiological Society of North America, 2023. Available [viewed 2024-05-14] at: https://www.radiologyinfo.org/en/info.cfm?pg=mammo

[51] UW Medicine: Newsroom (2020). *Study: AI improves radiologists' readings of mammograms*. Seattle, WA: University of Washington. Available [viewed 2024-05-14] at: https://newsroom.uw.edu/news/study-ai-improves-radiologists-readings-mammograms

[52] National Institute of Biomedical Imaging and Bioengineering. *Magnetic resonance imaging (MRI)*. Bethesda, MD: National Institutes of Health. Available [viewed 2024-05-14] at: https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri

[53] Wikipedia. *Magnetic resonance imaging*. San Francisco, CA: Wikimedia Foundation, 2024. Available [viewed 2024-05-14] at: https://en.wikipedia.org/wiki/Magnetic_resonance_imaging

[54] Magnetic resonance imaging - WHO. https://web.archive.org/web/20220119083521/http://www.who.int/diagnostic_imaging/imaging_modalities/dim_magresimaging/en/. Web archive date 2022-01-19.

[55] Piazza P. *Artificial intelligence enhances MRI scans*. Bethesda, MD: National Institutes of Health. Available [viewed 2024-05-14] at: https://www.nih.gov/news-events/nih-research-matters/artificial-intelligence-enhances-mri-scans

[56] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;**18**(8):500-510. DOI: 10.1038/s41568-018-0016-5

[57] Angiography - WHO. https://web.archive.org/web/20181030225852/http://www.who.int/diagnostic_imaging/imaging_modalities/dim_angiography/en/. Web archive date 2018-10-30.

[58] Wikipedia. *Angiography*. San Francisco, CA: Wikimedia Foundation, 2024. Available [viewed 2024-05-14] at: https://en.wikipedia.org/wiki/Angiography

[59] Siegersma KR, Leiner T, Chew DP, Appelman Y, Hofstra L, Verjans JW. Artificial intelligence in cardiovascular imaging: state of the art and implications for the imaging cardiologist. *Neth Heart J*. 2019;**27**(9):403-13. DOI: 10.1007/s12471-019-01311-1.

[60] Sharif MS, Amira A. An intelligent system for PET tumour detection and quantification. In: *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 2625-2628. New York, NY: IEEE. DOI: 10.1109/ICIP.2009.5414100.

[61] Lakhan SE, Kaplan A, Laird C, Leiter Y. The interventionalism of medicine: Interventional radiology, cardiology, and neuroradiology. *Int Arch Med*. 2009;**2**(1):27. DOI: 10.1186/1755-7682-2-27.

[62] Iezzi R, Goldberg SN, Merlino B, Posa A, Valentini V, Manfredi R. Artificial intelligence in interventional radiology: A literature review and future perspectives. J Oncol. 2019;**2019**:6153041. DOI: 10.1155/2019/6153041.

[63] Abajian A, Murali N, Savic LJ, Laage-Gaupp FM, Nezami N, Duncan JS, Schlachter T, Lin M, Geschwind JF, Chapiro J. Predicting treatment response to intra-arterial therapies for hepatocellular carcinoma with the use of supervised machine learning – An artificial intelligence concept. *J Vasc Interv Radiol*. 2018;**29**(6):850-7. DOI: 10.1016/j.jvir.2018.01.769.

[64] Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PloS One*. 2014;**9**(2), e88225. DOI: 10.1371/journal.pone.0088225.

[65] Asadi H, Kok HK, Looby S, Brennan P, O'Hare A, Thornton J. Outcomes and complications after endovascular treatment of brain arteriovenous malformations: A prognostication attempt using artificial intelligence. *World Neurosurg*. 2016;**96**:562-9. DOI: 10.1016/j.wneu.2016.09.086.

[66] Wachs JP, Stern HI, Edan Y, Gillam M, Handler J, Feied C, Smith M. A gesture-based tool for sterile browsing of radiology images. *J Am Med Inform Assoc*. 2008;**15**(3):321-3. DOI: 10.1197/jamia.M241. Erratum in*: J Am Med Inform Assoc*. 2009;**16**(3):284.

[67] El‐Shallaly GE, Mohammed B, Muhtaseb MS, Hamouda AH, Nassar AH. Voice recognition interfaces (VRI) optimize the utilization of theatre staff and time during laparoscopic cholecystectomy. *Minim Invasive Ther Allied Technol*. 2005;**14**(6):369-71. DOI: 10.1080/13645700500381685.

[68] Herniczek SK, Lasso A, Ungi T, Fichtinger G. Feasibility of a touch-free user interface for ultrasound snapshot-guided nephrostomy. In: *Proceedings Volume 9036, Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*, p. 90362F. Bellingham, WA: Society of Photo-Optical Instrumentation Engineers. DOI: 10.1117/12.2043564.

[69] Solbiati, M., Passera, K. M., Rotilio, A., Oliva, F., Marre, I., Goldberg, S. N., ... & Solbiati, L. (2018). Augmented reality for interventional oncology: proof-of-concept study of a novel high-end guidance system platform. European radiology experimental, 2(1), 18.

[70] Seals K, Al-Hakim R, Mulligan P, Lehrman E, Fidelman N, Kolli K, Kohlbrenner R, Kohi M, Taylor, A. (2019). 03:45 PM Abstract No. 38. The development of a machine learning smart speaker application for device sizing in interventional radiology. *J Vasc Interv Radiol*. 2019;**30**(3)Suppl: S20. DOI: 10.1016/j.jvir.2018.12.077.

[71] Letzen B, Wang CJ, Chapiro J. The role of artificial intelligence in interventional oncology: A primer. *J Vasc Interv Radiol*. 2019;**30**(1):38-41.e1. DOI: 10.1016/j.jvir.2018.08.032.

[72] FDA. *What is computed tomography?* Silver Spring, MD: US Food and Drug Administration, 2020. Available [viewed 2024-05-14] at: https://www.fda.gov/radiation-emitting-products/medical-x-ray-imaging/what-computed-tomography

[73] Berrington de Gonzalez A, Pasqual E, Veiga L. Epidemiological studies of CT scans and cancer risk: the state of the science. Br J Radiol. 2021 Oct 1;94(1126):20210471. doi: 10.1259/bjr.20210471. Erratum in: Br J Radiol. 2022 Jul; 95(1135):20210471c.

[74] FDA. *Other information resources related to whole-body CT screening*. Silver Spring, MD: US Food and Drug Administration, 2019. Available [viewed 2024-05-14] at: https://www.fda.gov/radiation-emitting-products/medical-x-ray-imaging/other-information-resources-related-whole-body-ct-screening

[75] Ghanem MH. CT scan in psychiatry: A review of the literature. *Encéphale*. 1986;**12**:3-12.

[76] Li W, Cui H, Li K, Fang Y, Li S. Chest computed tomography in children with COVID-19 respiratory infection. *Pediatr Radiol*. 2020;**50**(6):796-9. DOI: 10.1007/s00247-020-04656-7.

[77] Grillet F, Behr J, Calame P, Aubry S, Delabrousse E. Acute pulmonary embolism associated with COVID-19 pneumonia detected by pulmonary CT angiography. *Radiology*. 2020;**296**(3):E186-8. DOI: 10.1148/radiol.2020201544.

[78] Ahsan MM, Gupta KD, Islam MM, Sen S, Rahman M, Hossain MS. Study of different deep learning approach with explainable AI for screening patients with COVID-19 symptoms: Using CT scan and chest x-ray image dataset. arXiv:2007.12525 [preprint], 2020. DOI: 10.48550/arXiv.2007.12525.

[79] Yang X, He X, Zhao J, Zhang Y, Zhang S, Xie P. COVID-CT-Dataset: a CT scan dataset about COVID-19. arXiv:2003.13865 [preprint], 2020. DOI: 10.48550/arXiv.2003.13865.

[80] Chassagnon G, Vakalopoulou M, Battistella E, Christodoulidis S, Hoang-Thi TN, Dangeard S, Deutsch E, Andre F, Guillo E, Halm N, El Hajj S, Bompard F, Neveu S, Hani C, Saab I, Campredon A, Koulakian H, Bennani S, Freche G, Barat M, Lombard A, Fournier L, Monnier H, Grand T, Gregory J, Nguyen Y, Khalil A, Mahdjoub E, Brillet PY, Tran Ba S, Bousson V, Mekki A, Carlier RY, Revel MP, Paragios N.. AI-driven CT-based quantification, staging and outcome prediction of COVID-19 pneumonia. *Med Image Anal*. 2021;**67**:101860. DOI: 10.1016/j.media.2020.101860.

[81] Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, Zha Y, Liang W, Wang C, Wang K, Ye L, Gao M, Zhou Z, Li L, Wang J, Yang Z, Cai H, Xu J, Yang L, Cai W, Xu W, Wu S, Zhang W, Jiang S, Zheng L, Zhang X, Wang L, Lu L, Li J, Yin H, Wang W, Li O, Zhang C, Liang L, Wu T, Deng R, Wei K, Zhou Y, Chen T, Lau JY, Fok M, He J, Lin T, Li W, Wang G. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell*. 2020;**181**(6):1423-33.e11. DOI: 10.1016/j.cell.2020.04.045.

[82] Venugopal VK, Vaidhya K, Murugavel M, Chunduru A, Mahajan V, Vaidya S, Mahra D, Rangasai A, Mahajan H. Unboxing AI – Radiological insights into a deep neural network for lung nodule characterization. *Acad Radiol*. 2020;**27**(1):88-95. DOI: 10.1016/j.acra.2019.09.015.

[83] Seifert R, Weber M, Kocakavuk E, Rischpler C, Kersting D. Artificial intelligence and machine learning in nuclear medicine: Future perspectives. *Semin Nucl Med*. 2021;**51**(2):170-7. DOI: 10.1053/j.semnuclmed.2020.08.003.

[84] Gallo M, Spigolon L, Bejko J, Gerosa G, Bottio T. How to evaluate the outflow tract of LVAD after minimally invasive implantation by 3D CT-scan. *Artif Organs*. 2020;**44**(12):1306-9. DOI: 10.1111/aor.13777.

[85] Le V, Frye S, Botkin C, Christopher K, Gulaka P, Sterkel B, Frye R, Muzaffar R, Osman M. Effect of PET scan with count reduction using AI-based processing techniques on image quality. *J Nucl Med*. 2020;**61**(Suppl 1):3095.

[86] Cariño F, Sterling W. Parallel strategies and concepts for a petabyte multimedia database computer. *IEEE Parallel Database Techniques*, 1998.

[87] Cariño F, Sterling W, Kostamaa P. Industrial database supercomputer exegesis: The DBC/1012, the NCR 3700, the Ynet, and the Bynet. In: *Emerging trends in database and knowledge-base machines: The application of parallel architectures to smart information systems*, pp. 139–57. Washington, DC: IEEE Computer Society Press, 1995.

[88] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;**542**(7639):115–118. DOI: 10.1038/nature21056. Erratum in: *Nature*. 2017;**546**(7660):686.

[89] Arleo EK, Hendrick RE, Helvie MA, Sickles EA. Comparison of recommendations for screening mammography using CISNET models. *Cancer*. 2017;**123**:3673-80. DOI: 10.1002/cncr.30842.

[90] Bien N, Rajpurkar P, Ball RL, Irvin J, Park AK, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation. *PLoS Med*. 2018;**15**(11):e1002699. DOI: 10.1371/journal.pmed.1002699

[91] National Cancer Institute. *CBIS-DDSM: Curated breast imaging subset of digital database for screening mammography*. Bethesda, MD: Cancer Imaging Archive, 2024. Available [viewed 2024-05-12] at: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=22516629

[92] Stanford ML Group. *CheXpert: A large chest x-ray dataset and competition*. Stanford, CA: Stanford ML Available [viewed 2024-05-12] at: https://stanfordmlgroup.github.io/competitions/chexpert/

[93] Royal College of Radiologists. *Clinical radiology: UK workforce census report 2018*. London: Royal College of Radiologists, 2019. Available [viewed 2024-05-12] from: https://www.rcr.ac.uk/news-policy/policy-reports-initiatives/clinical-radiology-census-reports/

[94] Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, Kalloo A, Hassen ABH, Thomas L, Enk A, Uhlmann L; Reader study level-I and level-II Groups; Alt C, Arenbergerova M, Bakos R, Baltzer A, Bertlich I, Blum A, Bokor-Billmann T, Bowling J, Braghiroli N, Braun R, Buder-Bakhaya K, Buhl T, Cabo H, Cabrijan L, Cevic N, Classen A, Deltgen D, Fink C, Georgieva I, Hakim-Meibodi LE, Hanner S, Hartmann F, Hartmann J, Haus G, Hoxha E, Karls R, Koga H, Kreusch J, Lallas A, Majenka P, Marghoob A, Massone C, Mekokishvili L, Mestel D, Meyer V, Neuberger A, Nielsen K, Oliviero M, Pampena R, Paoli J, Pawlik E, Rao B, Rendon A, Russo T, Sadek A, Samhaber K, Schneiderbauer R, Schweizer A, Toberer F, Trennheuser L, Vlahova L, Wald A, Winkler J, Wölbing P, Zalaudek I. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists, *Ann Oncol*. 2018;**29**(8):1836–42, DOI: 10.1093/annonc/mdy166

[95] John R. Zech ,Marcus A. Badgeley ,Manway Liu,Anthony B. Costa,Joseph J. Titano,Eric Karl Oermann (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. https://doi.org/10.1371/journal.pmed.1002683

[96] Johnson A, Pollard T, Mark R, Berkowitz S, Horng S. *MIMIC-CXR Dataset*. Version 2.0.0 Boston, MA: MIT Laboratory for Computational Physiology. Available [viewed 2024-05-12] at: https://physionet.org/content/mimic-cxr/2.0.0/

[97] NIH Clinical Center. CXR 8. Available [viewed 2024-05-16] at: https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community. Accessible [viewed 2024-05-16] from: NIH. *NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community*. Bethesda, MD: National Institutes of Health. Available [viewed 2024-05-16] at: https://nihcc.app.box.com/v/ChestXray-NIHCC

[98] NHS England. *National COVID-19 Chest Image Database (NCCID)*. Available [viewed 2024-05-15] at: https://transform.england.nhs.uk/covid-19-response/data-and-covid-19/national-covid-19-chest-imaging-database-nccid/

[99] RAD-AID. *Volunteer with RAD-AID Liberia*. Chevy Chase, MD: RAD-AID International, 2024. Available [viewed 2024-05-12] at: https://www.rad-aid.org/countries/africa/liberia/

[100] Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz CP, Patel BN, Yeom KW, Shpanskaya K, Blankenberg FG, Seekins J, Amrhein TJ, Mong DA, Halabi SS, Zucker EJ, Ng AY, Lungren MP. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med*. 2018;**15**(11):e1002686. DOI: 10.1371/journal.pmed.1002686

[101] UCSF Department of Radiology and Biomedical Imaging. *Digital x-ray on-the-go in Kenya*. San Francisco, CA: Regents of the University of California. 2015. Available [viewed 2024-05-12] at: https://radiology.ucsf.edu/blog/digital-x-ray-go-kenya

[102] Banerjee I et al., *Reading race: AI recognizes patient's racial identity in medical images*, Pre-print, https://arxiv.org/pdf/2107.10356

[103] Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A. *et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations*. *Nat Med* **27**, 2176–2182 (2021). https://doi.org/10.1038/s41591-021-01595-0

[104] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F_1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;**21**:6. DOI: 10.1186/s12864-019-6413-7

[105] C.Ferri, J.Hernández-Orallo, R.Modroiu. An experimental comparison of performance measures for classification. Pattern Recognition Letters 30, 1, (2009), Pages 27-38

[106] Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, Chepelev L, Cairns R, Mitchell JR, Cicero MD, Poudrette MG, Jaremko JL, Reinhold C, Gallix B, Gray B, Geis R; Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group. Canadian Association of Radiologists white paper on artificial intelligence in radiology. *Can Assoc Radiol J*. 2018;**69**(2):120-35. DOI: 10.1016/j.carj.2018.02.002

[107] ResNet-34 from Deep Residual Learning for Image Recognition. https://pytorch.org/vision/main/models/generated/torchvision.models.resnet34.html (visited 2025-02-19)

[108] Densenet-121 model from Densely Connected Convolutional Networks, https://pytorch.org/vision/main/models/generated/torchvision.models.densenet121.html

[109] Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*. 2018;**286**(3):800-9. DOI: 10.1148/radiol.2017171920

[110] Heumann S, Zahn N. *Benchmarking national AI strategies*. Berlin: Stiftung Neue Verantwortung, 2018. 38 pp. Available [viewed 2024-05-15] at: https://www.stiftung-nv.de/sites/default/files/benchmarking_ai_strategies.pdf

[111] Amara Tariq, Saptarshi Purkayastha, Geetha Priya Padmanaban, Elizabeth Krupinski, Hari Trivedi, Imon Banerjee, Judy Wawira Gichoya, *Current clinical applications of artificial intelligence in radiology and their best supporting evidence*. Journal of the American College of Radiology, Volume 17, Issue 11, 2020, pp.1371-1381, https://doi.org/10.1016/j.jacr.2020.08.018.

[112] Bell DJ. *X-rays*. Radiopaedia.org, 2022. Available [viewed 2024-05-17] at: https://radiopaedia.org/articles/x-rays-1?lang=us

[113] Mafraji MA. Conventional radiography. In: *MSD manual – Professional version*, 2023. Available [viewed 2024-05-17] at: https://www.msdmanuals.com/professional/special-subjects/principles-of-radiologic-imaging/conventional-radiography.

[114] National Health Services Transformation Directorate, *NHSX moves on* (4 February 2022) https://transform.england.nhs.uk/blogs/nhsx-moves-on/ (visited 2025-02-19)

[115] ITU-T Focus Group on AI for Health Deliverable 5.5 (2023), *Data handling*, https://handle.itu.int/11.1002/plink/1685437902

[116] ITU-T Focus Group on AI for Health Document J-038 (2020), WG-DAISAM Metrics and Measures Paper Questionnaire, https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/all/FGAI4H-J-038.docx.

[117] Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit Health*. 2019;**1**(6):e271-e297. DOI: 10.1016/S2589-7500(19)30123-2. Erratum in: *Lancet Digit Health*. 2019;**1**(7):e334.

[118] Online Converter. *DICOM to JPG*. OnlineConverter.com, Internet. Available [viewed 2024-05-16] at: https://www.onlineconverter.com/dicom-to-jpg

[119] GitHub. *velebit-ai/COVID-Next-Pytorch*. San Francisco, CA: GitHub, Internet. Available [viewed 2024-05-15] at: https://github.com/velebit-ai/COVID-Next-Pytorch

[120] Viradiya P. *Brian tumor dataset*. San Francisco, CA: Kaggle, Internet. Available [viewed 2024-05-15] at: https://www.kaggle.com/preetviradiya/brian-tumor-dataset

Annex A

Glossary

Table A.1 lists all the relevant abbreviations and acronyms used in the document.

Table A.1 –Abbreviations and acronyms used in this document

Acronym/Term	Expansion	Comment
2D	two Dimensional	
3D	three Dimensional	
AI	Artificial intelligence	
AI4H	Artificial intelligence for health	
AI-MD	AI based medical device	
API	Application programming interface	
AUC	Area Under the Curve	
BMI	Body Mass Index	
BPF	Bandpass Filtered	
CE	European Conformity	
CfTGP	Call for topic group participation	
COVID-19	Corona Virus Disease-2019	
CSV	Comma-Separated Values	
CT	Computed Tomography	
DEL	Deliverable	
DICOM	Digital Imaging and Communications in Medicine	
EMR	Exact Match Ratio	
FDA	Food and Drug administration	
FGAI4H	Focus Group on AI for Health	
GDP	Gross domestic product	
GDPR	General Data Protection Regulation	
HPF	High-Pass Filtered	
IMDRF	International Medical Device Regulators Forum	
IP	Intellectual Property	
IR	Interventional Radiology	
ISO	International Organization for Standardization	
ITU	International Telecommunication Union	
JPEG	Joint Photographic Experts Group	
LIME	Lightweight Interactive Multimedia Environment	
LMIC	Low-and middle-income countries	

Table A.1 –Abbreviations and acronyms used in this document

Acronym/Term	Expansion	Comment
LPF	Low-Pass Filtered	
MDR	Medical Device Regulation	
ML4H	Machine Learning for Health	
MRI	Magnetic Resonance Imaging	
NCCID	National COVID-19 Chest Imaging Database	
NF	Notch Filtered	
NPV	Negative Predictive Value	
PET	Positron Emission Tomography	
PII	Personal identifiable information	
PNG	Portable Network Graphic	
ROI	Region Of Interest	
SPECT	Single Photon Emission Computed Tomography	
TG	Topic Group	
TPR	True Positive Rate	
UKCA	UK Conformity Assessed	
US	Ultrasonography	
WG	Working Group	
WHO	World Health Organization	

Annex B

Declaration of conflict of interest

No declarations made by the contributors to this document.