International Telecommunication Union

ITU-T FG-AI4H Deliverable

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

15 September 2023

PRE-PUBLISHED VERSION

DEL10.12

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

ITU-T

International Telecommunication Union

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 solutions are being developed. The challenge artificial intelligence radiological solutions however face 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 generalise and properly handle edge cases.

This topic description document (TDD) 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 artificial intelligence radiological solution to be assessed on its ability to generalise across diverse geographical location, gender and age groups.

Keywords

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

Change Log

This document contains Version 1 of the Deliverable DEL10.12 on "*FG-AI4H Topic Description Document for the Topic Group on AI for radiology (TG-Radiology)*" approved on 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
E-mail:

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

CONTENTS

Page

1 Introduction 6
1.1 Document Structure 6
1.2 Status update for meeting [Meeting L] 6
1.3 Status update for meeting [Meeting M] 6
1.4 Status update for meeting [Meeting R] 7
1.5 Topic description 7
1.5.1 Impact of benchmarking 8
1.6 Ethical considerations 9
1.6.1 Overview 9
1.6.2 Reading race: AI recognises patient's racial identity in medical images 9
1.7 Existing AI solutions 11
1.7.1 Use case descriptors 11
1.7.2 Collected AI solutions and use cases 12
1.8 Imaging modalities 14
1.9 Existing work on benchmarking 20
1.10 Benchmarking overview 20
1.11 The NHS AI Lab - Call for AI driven COVID-19 models: Performance assessment using the national COVID-19 chest imaging database 21

2 AI4H Topic Group 21

3 Method 22
3.1 AI input data structure 22
3.1.1 Image conversion considerations 22
3.1.2 Image compression and other artifacts considerations 22
3.1.3 Lossless medical image compression for radiology 26
3.2 AI output data structure 31
3.3 Test data labels 32
3.4 Scores & metrics 32
3.4.1 Threshold metrics 32
3.4.2 Ranking metrics 38
3.4.3 Probability Metrics 39
3.5 Undisclosed test data set collection 40
3.6 Benchmarking methodology and architecture 41
3.6.1 Audit trial 41
3.6.2 Audit trial checklist 41
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 United States of America. 49

Page

3.6.4 Benchmarking solution 49
3.6.5 Evaluation metrics 51
3.6.6 Benchmark categorizations 52
3.6.7 Evaluation data 52
3.6.8 The panel of expert radiologists 53
3.6.9 Test radiologists 53
3.7 Evaluation data availability 53
3.8 Feasibility 53
3.9 Privacy and security 53
3.10 Impact 54
3.11 Reporting methodology 54
4 Results 54
5 Discussion 55
References 56
Annex A: Glossary 64
Annex B: Declaration of conflict of interest 65

List of Tables

Page

Table 1 – Reading race – Experiments, methods, and results 10
Table 2 – Imaging modalities 14
Table 3 – Image conversion considerations 22
Table 4 – Compression performance comparison for various compression methods 31
Table 5 – Draft audit verification checklist 43

List of Figures

Page

Figure 1 – Impact of the compression in the test dataset accuracy of the COVID-Next classifier 24
Figure 2 – Impact of the compression in the test accuracy of the brain tumour classifier 25
Figure 3 – Impact of the compression in the test accuracy of the brain tumour classifier 25
Figure 4 – Confusion matrix of the brain tumour classifier test accuracy of a JPEG compression scenario 26
Figure 4-bis – Multimedia representation phases for radiology images 27

Page

Figure 4-ter – Early example of vectorizing medical imagery [86] 28
Figure 4-5(a) – Illustration of Lossless Hilbert compression 28
Figure 4-5(b) – Illustration of Lossless Hilbert compression 29
Figure 4-6 – Example of interconnection network topologies [87] 30
Figure 4-7 – Processing time and storage requirement for various compression methods 31
Figure 5 – Model Results after trial audits using the benchmarking platform, health.aiaudit.org 41
Figure 6 – A prototype of the radiograph-agnostic precision evaluation platform 50
Figure 6-bis – The 'Location' category with its sub-categories and the metrics used 50
Figure 7 – Each sub-category would feature demographics intersection performances too 51
Figure 8 – The 'Gender' category 52

ITU-T FG-AI4H Deliverable 10.12

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, PET (positron emission tomography) scans, MRI (magnetic resonance imaging) scans and CT (computed tomography) 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 exams are performed each year in the United States alone. Research led by Elizabeth Kagan Arleo, MD, of Weill Cornell Medicine found that recommendation of annual screening starting at age 40 would result in a nearly 40 percent reduction in deaths due to breast cancer (Arleo et al, 2017). Simple radiological procedures like ultrasound can reduce the need for surgical interventions. And though clinical judgement may be sufficient, radiological procedures are necessary in confirming and properly evaluating the causes of many conditions and responses to treatments.

1.1 Document Structure

Overview of the whole document.

1.2 Status update for meeting [Meeting L]

Between the 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 on the 16th of April 2021. During the meeting, we discussed status updates and welcomed new members. We discussed open work streams within the topic group that our members can then lead and collaborate towards contributing to. Vincent Appiah, minoHealth AI Labs took the "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 under the chapter, "Existing work on benchmarking". Darlington Akogo, the Topic Driver, also wrote a summary on the work being done by the NHS AI Lab in benchmarking AI solutions for COVID-19. 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 JPEG compression ratios and PNG and contributed the results under "Image Compression Considerations".

1.3 Status update for meeting [Meeting M]

Towards Meeting M, Samori Issah, minoHealth AI Labs, contributed an overview for Ethical Considerations under AI for Radiology. Judy Wawira Gichoya, Emory University School of Medicine, contributed a section 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 tumor image classification task as well. The results further demonstrate the influence of the compression artifacts 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, F-Score, etc. for testing different compression and downsizing in a dataset. Darlington Akogo, minoHealth AI Labs expanded the list of evaluation metrics to include 10 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 [Meeting R]

Darlington Akogo contributed the section "**Audit Trial: minoHealth.ai: A Clinical Evaluation Of Deep Learning Systems For the Diagnosis of Pleural Effusion and Cardiomegaly In Ghana, Vietnam and the United States of America**", which covers the first AI clinical study in Africa, benchmarking the performance of AI for radiology systems against radiologists. Dominick Romano contributed the section "**Lossless Medical Image Compression for Radiology**", which covers techniques to compress medical images of different modalities. The section also contains benchmark tests on these different techniques.

1.5 Topic description

Challenges Facing Radiology

Though radiology is very important, there's a shortage of radiologists globally, especially in developing countries. Liberia, for example, only has about 2 radiologists (RAD-AID, 2017), whilst Ghana has 34 radiologists and Kenya has 200 radiologists (UCSF, 2015). And in the UK, only one-in-five trusts and health boards has sufficient number of interventional radiologists to run a safe 24/7 service to perform urgent procedures (Clinical Radiology UK Workforce Census Report, 2018) whilst their workload of reading and interpreting medical images has increased by 30% between 2012 and 2017. There's a need for scalable and accurate automated radiological systems. Deep Learning, especially Convolutional Neural Networks, 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 (A.I) and Deep Learning, due to growth in computing power and data, has led to advancements in Deep Convolutional Neural Networks, which has allowed for breakthrough research and applications in Radiology. Artificial Intelligence and Deep Learning holds a lot of potential in Radiology. Artificial Intelligence 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. A recent study found that the Fleiss' kappa measure of interrater reliability for detecting anterior cruciate ligament tear, meniscal tear, and abnormality were higher with model assistance than without it (Bien et al., 2018). Deep Learning has achieved performances comparable to humans and sometimes better. A recent study 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 (Liu et al., 2019). Deep Learning has performed as well as radiologists and sometimes better at detecting abnormalities like pneumonia, fibrosis, hernia, edema and pneumothorax in chest x-rays (Rajpurkar et. al, 2017). It has also been used to detect knee abnormalities via magnetic resonance (MR) imaging at near-human-level performance (Bien et. al, 2018). Researchers have also trained Deep Learning models that outperformed dermatologists at detecting skin cancer (Esteva et. al, 2017, Haenssle et. al, 2018).

Research Data

One key focus of deep learning radiological applications is breast cancer detection via mammograms. The CBIS-DDSM (Curated Breast Imaging Subset of Digital Database for Screening Mammography) is one of the key repositories publicly available. It contains 10,239 images and is 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 by the Stanford University School of Medicine. CheXpert contains 224,316 chest radiographs of 65,240 patients. It contains images for 12 different thoracic diseases including Atelectasis, Cardiomegaly, Enlarged Cardiomegaly, Consolidation, Edema, Lung Lesion, Lung Opacity, Pneumonia, Pneumothorax, Fracture, Pleural Effusion and Pleural Other. And it contains 2 other observations "No Finding" and "Support Devices", making 14 observations in total. The radiographs were collected from Stanford Hospital, between October 2002 and July 2017. Another publicly available chest radiograph dataset is MIMIC-CXR dataset by Massachusetts Institute of Technology (MIT). The dataset 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 dataset also contains 14 observations. There is also a chest x-ray dataset from the NIH Clinical Center that contains 100,000 x-rays from over 30,000 patients, including many with advanced lung disease. That leads to a total of 696,236 publicly available x-ray images for 12 thoracic conditions.

Challenges Facing AI in Radiology

The challenge however lasts in properly testing such systems and ensuring they work in all edge and diverse cases radiologists encounter. A study by Eric Oermann and colleagues found that, deep learning models that detected pneumonia on chest x-rays performed well on further data from sites they were trained on (AUC of 0.93–0.94) but significantly less on external data (AUC 0.75–0.89) (Zech et al., 2018). This demonstrates the challenge of assessing the generality and scalability of Deep Learning systems. Though the study by Liu and colleagues analysed 31,587 studies, only 69 studies provided enough data to construct contingency tables, enabling calculation of test accuracy. And out of that 69 studies, only 25 studies did out-of-sample external validations. And further, only 14 of such studies compared the models' performances to that of radiologists. They also realised 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 Artificial Intelligence solutions, their readiness to be deployed and the degree of autonomy they should be given.

1.5.1 Impact of benchmarking

There exists a large amount of publicly available medical image datasets online, and there have been a lot of research and development with such datasets. By developing frameworks that target these conditions first, we would make the standardized benchmarking platform immediately appealing to the A.I healthcare research and development community. This would also help speedup the deployment of AI solutions in Radiology globally. AI healthcare system developers and organisations usually have to go through the challenge of convincing health facilities to share their private data with them, such data unfortunately aren't always of high quality and they usually lack the broad demographic representations needed to truly assess how well an A.I system generalises. 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 A.I systems by ensuring they are tested in great detail across various groups. It'd 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

Artificial intelligence 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 tunable parameters. [3]

The complexity of these algorithms makes understanding the reasoning behind an AI model's decision very difficult. Thus, making auditing and debugging an AI model's decision 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 patients' data from healthcare institutions. A lot of discussions and concerns have sprung up around whether or not patients' consent is needed whenever their data is used in training an AI model. Some agree that the consent of patients is supposed to be requested while others argue that developing AI models for radiology is for the greater good and that no one's consent is needed to pursue the greater good.

There are also a lot of unanswered questions around data ownership and how profits derived from using patients' data will be shared [4]. Whoever is identified as an owner or part of the owners of a dataset deserves a share in the profit the dataset generates. So, if it is agreed that the data is owned by patients then they deserve a share in the profit an AI developer will make from a model that was trained on the patients' 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, skillset, etc). This will further exacerbate the existing inequality in society as only those with the required resources can harness the power of AI. [6]

With regards to liability, 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, or the radiologist who used the AI model, 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 recognises patient's racial identity in medical images

(Banerjee, I, et al, 2021) There are no known imaging biomarker correlates for racial identity, however, medical imaging artificial intelligence (AI) models produce racial disparities (Pierson, 2021;Seyyed-Kalantari, 2021). There is potential for discriminatory harm if we assume that AI models are agnostic to race – understanding the relationship between race and medical imaging AI

models is important (Tariq, 2020). We sought to answer how AI systems could produce disparities across racial groups and determine how AI could predict race from medical images.
In this study, we investigate a large number of publicly and privately available large-scale medical imaging datasets and find that self-reported race is trivially predictable by AI models trained with medical image pixel data alone as model inputs. We use standard deep learning methods for each of the image analysis experiments, training a variety of common models appropriate to the tasks. First, we show that AI models can predict self-reported race across multiple imaging modalities, various datasets, and diverse clinical tasks (A). The high level of performance persists during the external validation of these models across a range of academic centers and patient populations in the United States, as well as when models are optimized to perform clinically motivated tasks. We also perform ablations 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 (B). Finally, we show that the features learned appear to involve all regions of the image and frequency spectrum, suggesting that mitigation efforts will be challenging (C). A brief description of these experiments is included in Table 1.

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

Experiment	Description	Results
A.1 Detection of racial identity on chest XR	Resnet34 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 XR, cervical spine XR, 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	Densenet121 models to detect pathology on CXR/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	2 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 model	Average AUC of 0.95 across races.
B.5 Race detection using age and sex	2 models trained on split data (A1 method) – 5 age groups and male/female.	No significant deviation from A1.
C1 Frequency-domain imaging features	4 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]

Experiment	Description	Results
C2 Impact of image resolution and quality	3 new models created on modified datasets (A1 method) – various resolutions and 2 with image perturbations.	AUC - >0/95 for 160x160 resolution and 0.64 for 4x4 images; Average of 0.652 for perturbed.
C3 Anatomical localization	Produced saliency maps using grad-cam method, 5 radiologists perform qualitative evaluation. Mask regions of interest (ROI) 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 regions of interest.	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.
C4 Patch-based training	Train 2 new models (A1 methodology) on datasets – split images into 3x3 square cells of equal size remove 1 of 9 cells, only use 1 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. Our 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 are not relying 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.

We strongly recommend 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 the 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, we identified the following 9 descriptors that would be useful:

– 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
– Explainability and Interpretability framework

1.7.2 Collected AI solutions and use cases

minoHealth	
Descriptor	**Description**
Condition	Pneumonia, Hernia, Fibrosis, Atelectasis, Cardiomegaly, Enlarged Cardiomegaly, Consolidation, Edema, Lung Lesion, Lung Opacity, Pneumothorax, Fracture, Pleural Effusion and Pleural Other (14 different systems)
Medical imaging modality	Chest XRay
AI task/problem description	Image Classification
General algorithm description	Convolutional Neural Networks, 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	Sigmoid with range 0 - 1, 0 = Negative, 1 = Positive
Evaluation metrics	Accuracy Score, ROC curve & Area Under Curve Score
Explainability and Interpretability framework	Implementing LIME

minoHealth	
Descriptor	**Description**
Condition	Breast Cancer
Medical imaging modality	Mammograms
AI task/problem description	Image Classification
General algorithm description	Convolutional Neural Networks, 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 3 classes, Benign, Benign Without Callback and Malignant
Evaluation metrics	Accuracy Score, ROC curve & Area Under Curve Score
Explainability and Interpretability framework	Implementing LIME

Braid.Health	
Descriptor	**Description**
Condition	Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration,

	Mass, Nodule, Peural_Thickening, Pneumonia, Pneumothorax, Old Fracture, New Fracture, Scoliosis, Sternotomy, Enlarged Cardiomedistinum, Support Devices, Tuberculosis, Bronchiectasis, Foreign Body (22 conditions)
Medical imaging modality	Chest XRay
AI task/problem description	Image Classification
General algorithm description	Convolutional Neural Networks, 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, Area Under Curve ROC Score, Specificity at Sensitivity
Explainability and Interpretability framework	None currently

Braid.Health	
Descriptor	**Description**
Condition	Fracture, Dislocation, Edema, Arthritis, Osteoarthritis, Spur (6 conditions)
Medical imaging modality	Foot XRay
AI task/problem description	Image Classification
General algorithm description	Convolutional Neural Networks, 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, Area Under Curve ROC Score, Specificity at Sensitivity
Explainability and Interpretability framework	None Currently

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	XRay
AI task/problem description	Image Classification
General algorithm description	Convolutional Neural Networks, 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, Area Under Curve ROC Score, Specificity at Sensitivity
Explainability and Interpretability framework	None currently

1.8 Imaging modalities

Table 2 maps out 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 would have paragraphs dedicated to covering details using the pointers below:

– Description: Description of imaging modality

– Conditions: Conditions modalities are applied to

– Data structure: Data structure of images from modality
This would cover some details on the type of images generated from each modality. These details would include whether it's a single/multiple 2D image or 3D image, DICOM or some other format

– AI Applications: How AI is being 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,

Conventional radiography (plain x-rays)	
	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 [1,2].
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 [1,3].
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 [4]. – 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 edema, emphysema, pneumonia, and pediatric 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 convolutional neural network (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]. – 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

Conventional radiography (plain x-rays)	
	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 arterio-venous malformations, diagnosis of bleeding vessels, and assessment of vascularity of malignant tumors. [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 (FFR). [59]
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 radiations. [49][50]
Conditions	Mammography can be used as a screening tool or a diagnostic tool. – As a screening tool, mammography is used for the early detection of breast cancer.

Conventional radiography (plain x-rays)	
	– 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 (CT)	
Description	Computed Tomography (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 though less coregistration of CTs with other imaging means, helping to reduce patient fatigue and exposure; more may abound as this area of research, that is the combination of AI and CT scanning, is still new [83].
Single-photon emission computed tomography (SPECT)	
Description	Single photon emission computed tomography (SPECT) is a technique which allows nuclear medicine studies, which would otherwise be represented in planar images, to be rendered in three dimensions. Photons emitted by injected radiopharmaceuticals are detected by gamma cameras which rotate around the patient to provide spatial information on tissue distribution. The data is then reconstructed into three-dimensional 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.

Conventional radiography (plain x-rays)	
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 (MDP) bone scans which provide information about bone perfusion and turnover.
Data structure	
AI Applications	
Ultrasonography (US) and Doppler	
Description	Ultrasonography 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. Ultrasonography can be used to image the liver, kidney, heart, pancreas, etc. [41][42] Another common use case for ultrasonography is real-time imaging of developing fetuses 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 ultrasonography. These tasks include image classification, segmentation, detection, registration, biometric measurements and quality assessment. [44]
Magnetic resonance Imaging (MRI)	
Description	Magnetic resonance imaging is an imaging 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 the 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, the abdomen, etc.

Conventional radiography (plain x-rays)	
	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 artifacts in MRI scans [55] AI is also used to classify MRI scans as healthy or diseased. [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][46][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 kidney, lungs, heart, thyroid gland, and bone. [45]
Data structure	Nuclear images could be 2D images (scintigraphy) or 3D images (SPECT). Some modern nuclear imaging equipment are hybrid and allow for a fusion between CT and nuclear imaging. [45][47]
AI Applications	In nuclear imaging, AI is commonly used for radiomics. AI could potentially be used to detect artifacts and noise in nuclear images and correct them by applying the appropriate algorithm.
Positron emission tomography (PET)	
Description	Positron Emission Tomography (PET) is an imaging modality that uses a 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 improved 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	

Conventional radiography (plain x-rays)	
Description	Interventional Radiology (IR) is a means of radiology that uses current imaging methods, such as CTs,MRIs,,X-rays, PETs, 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 [IC] 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 methody combinations as described in the sections above.
AI Applications	AI has been used in IR to predict treatment 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 trialed, 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

– papers on existing attempts to benchmark solutions on the topic

– clinical evaluation attempts, RCT, etc.

– including existing numbers

1.10 Benchmarking overview

Artificial intelligence is considered to be one of the key driving forces of the 4th Industrial Revolution. This has led to the adoption of national AI strategies by many countries (Heumann & Zahn, 2021). However there is the lack of a consensus on how to measure the success of AI models. We therefore give a brief non-exhaustive list 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; Area under the curve(AUC), Accuracy, F1 score, Sensitivity and specificity (Park & Han, 2018). Model performance should be measured for both validation and test data. Benchmarking should also take into account the annotation of data. Is the data labeled, unlabeled or semi-labeled? This will determine what AI models and performance metrics to use. Appropriate models should also be used in AI-based solutions. A lot of factors should be considered when applying AI models; type of data, sample size, computational cost, etc. (Tang et al,2018). 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 are providing a platform that allows for AI solutions within the UK to be assessed based on 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 generalisability of AI models before they are used in clinical practice. (NHSX. "Performance Assessment Call - National COVID-19 Chest Image Database documentation.")

Before an AI system can be assessed on their platform, the AI developers would 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 AWS cloud-computing infrastructure provided by NHSX. The NCCID unseen dataset is then accessed in the form of an S3 bucket. The 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 the AI developers. To ensure Intellectual Property protections, all people involved in the AI model assessment, including external assessors will be bound to confidentiality by contractual agreements. Non-Disclosure Agreements (NDAs) are also used where need be.

At the end, the AI developer will receive 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 MHRA (Medicines and Healthcare products Regulatory Agency), the United Kingdom's healthcare products Regulatory Agency, for derogation of UKCA/CE marking or via standard conformance assessment processes. The UKCA (UK Conformity Assessed) 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 which previously required the CE marking.

2 AI4H Topic Group

- Topic Group structure
  - Subtopic 1
  - Subtopic 2
- Topic Group participation
- Tools/process of TG cooperation: Slack, Zoom, Google Docs, Github
- TG interaction with WG, FG: Work in DAISAM and DASH to test frameworks in Sandbox
- 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

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 wouldn't have to worry about the conversion.	– Easier for users in clinical settings – Conversion cannot be easily interfered. – Leaves little room for error on the part of users	– Requires further development of by manufacturers – Subjected to the quality of manufacturers' software development
Using a separate software. There's MicroDicom, a free windows tool, and a number of other tools that are either free or must be paid for.	– Easier for manufacturer since it requires no to little additional development – Can allow for reliance on already established and trusted high-quality tool – If offline, it can ensure data privacy better than an online 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
Using an online tool. There are also online free tools, like: https://www.onlineconverter.com/dicom-to-jpg	– 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 unauthorised access to data.

3.1.2 Image compression and other artifacts 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 is 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 we consider that:

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

– regular monitors don't 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 artifacts such as blocking and ringing. These artifacts may reduce an AI system's 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 the COVID-Next https://github.com/velebit-ai/COVID-Next-Pytorch, a COVID-19 classifier, inspired by the COVID-Net proposed by Wang et. al. (2020), based on ResNext50.

This model was trained using chest radiography with different resolutions, qualities and artifacts. The test accuracy of this model is 94.76%. However, if we compress the test dataset 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. We observe that it is possible to achieve significant bandwidth reduction with a negligible accuracy reduction.

Examining the cyan and red curves of Figure 1, one can see that the accuracy can be significantly reduced due to compression. In this case, the accuracy notably drops when the compression ratio goes lower than 0.10.

Despite the visual quality reduction due to the compression, the effect of the compression artifacts (blocking or ringing) is quite reduced due to a resize of the compressed image before feeding the COVID-Net.

In an extreme case, referring to the green (JPEG) and blue (JPEG 2000) curves in Figure 1, we resize the images in the dataset to 256x256 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 the COVID-Net as the image quality degrades. This image size was chosen due to the COVID-Net input architecture.

We conducted a similar test with a brain tumour image classifier available at: https://www.kaggle.com/preetviradiya/brian-tumor-dataset, 2021. The results are shown in Figures 2 and 3 where accuracy and F1 Score is 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 artifacts in medical image classification.

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

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

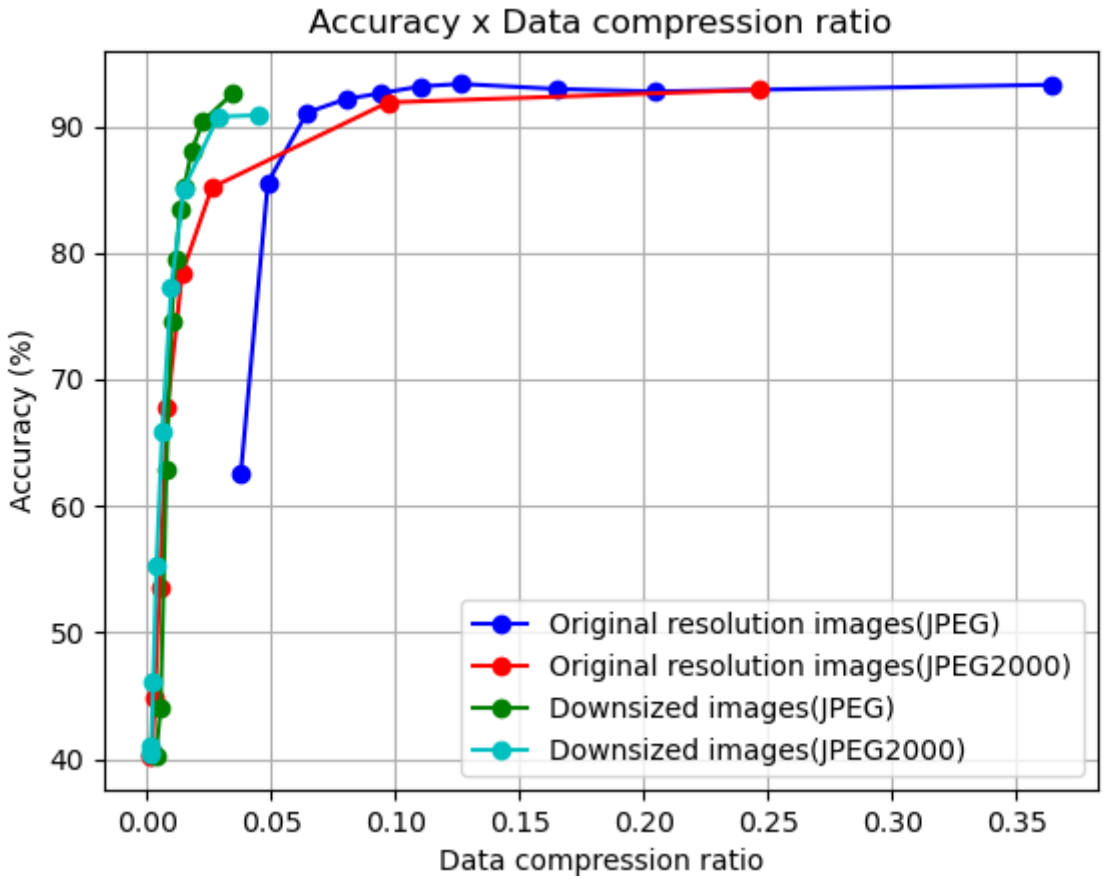


Figure 1 – Impact of the compression in the test dataset accuracy of the COVID-Next classifier

In blue (JPEG) and red (JPEG 2000) shows the case where dataset images were compressed with different compression rates. In green (Interpolative JPEG) and cyan (Interpolative JPEG 2000) the images were downsized to 256x256 pixels before compression. Without compressing the images (PNG) the accuracy is 94.76%, as shown in magenta.

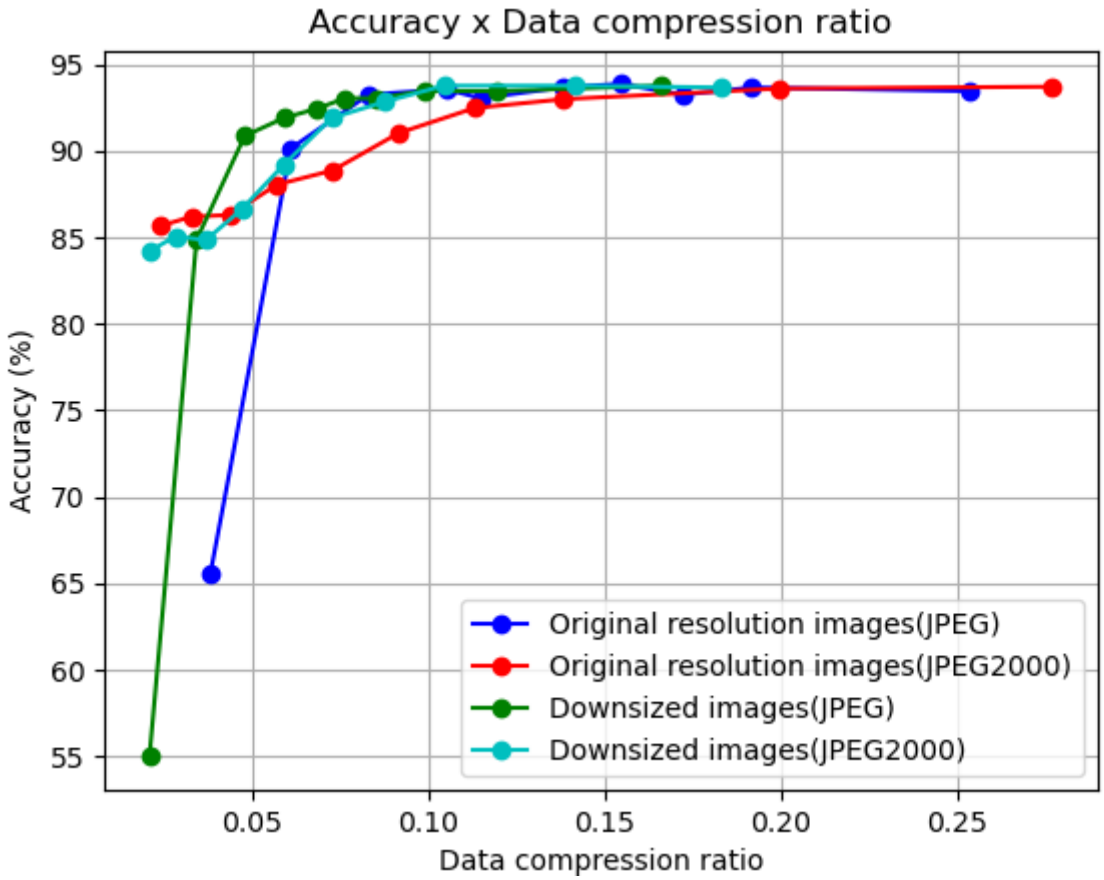


Figure 2 – Impact of the compression in the test accuracy of the brain tumour classifier

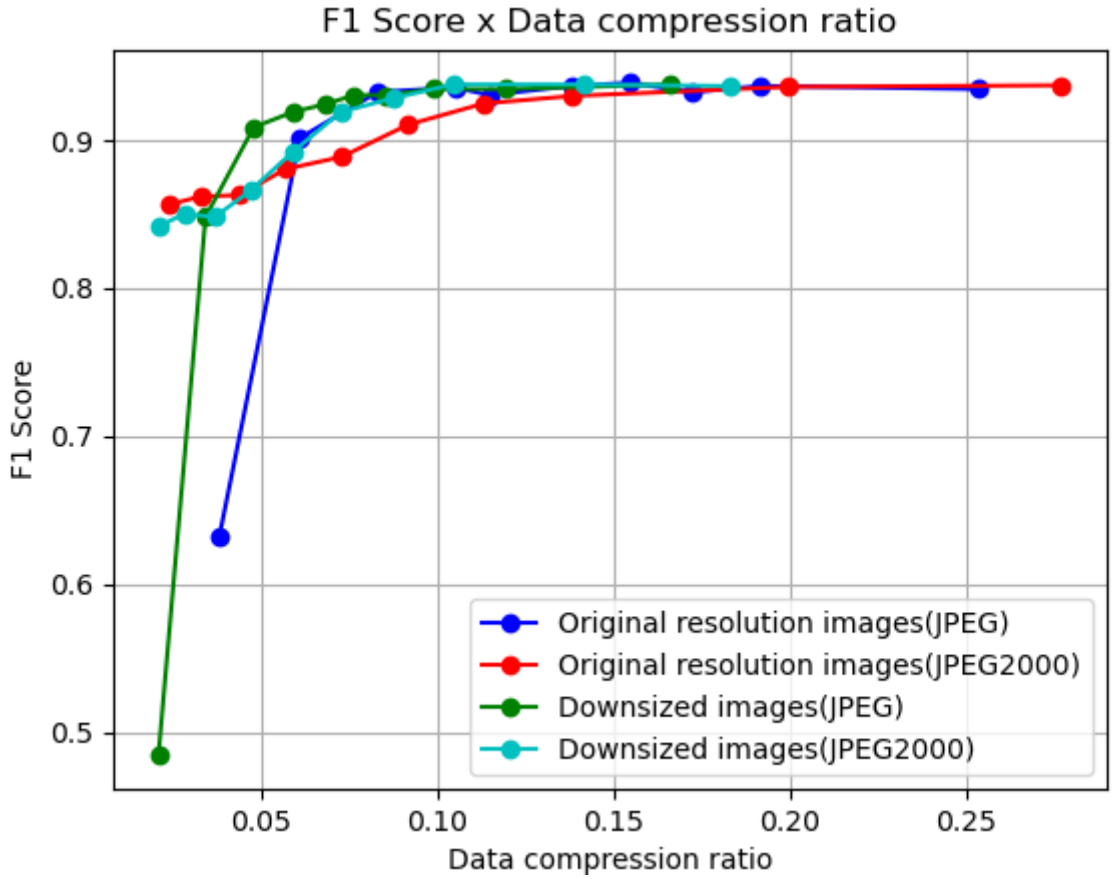


Figure 3 – Impact of the compression in the test accuracy of the brain tumour classifier

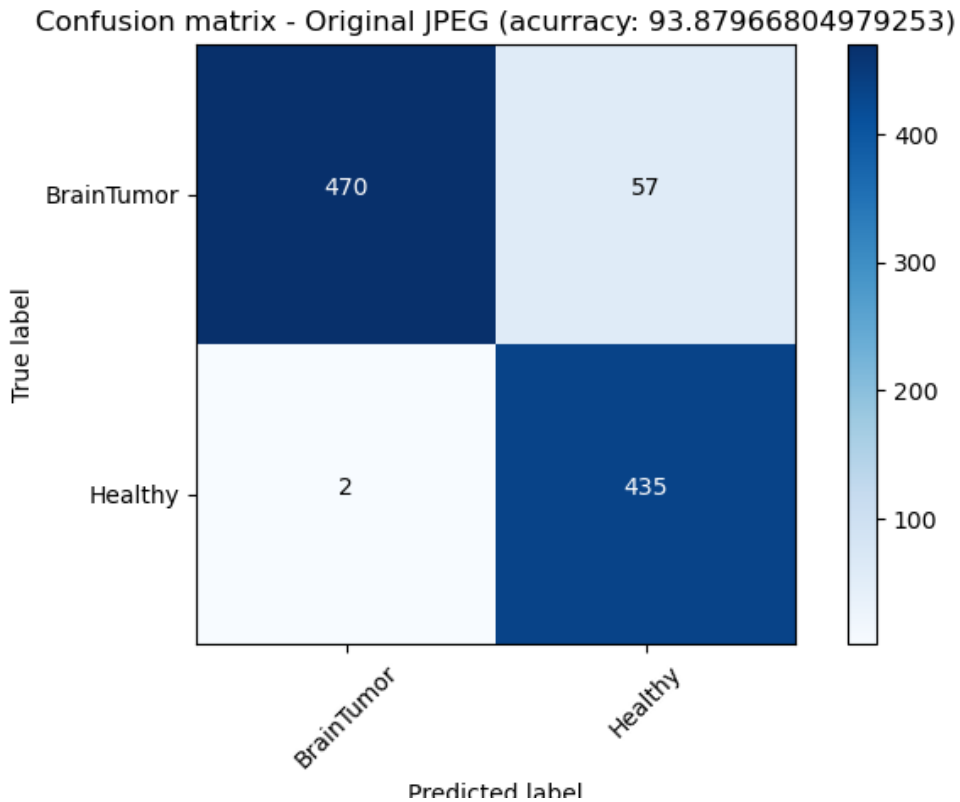


Figure 4 – Confusion matrix of the brain tumour classifier test accuracy of a JPEG compression scenario

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

Another artifact that may also be taken into consideration is the Moiré pattern. This kind of artifact may occur when a picture is taken from a screen. In this case, the pattern of the pixels in the screen is overlayed with the capturing pattern of a camera. As developers we must have to 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 Neuro Informatics files (NII) commonly used in CT and MRI is costly and time consuming. To load the media, and store it for the long term is extremely costly. To process the files, and transfer across systems is extremely time consuming. As more medical samples are accumulated and used to train AI Models, we must rethink how we store and process these files. We introduce a form of lossless Hilbert compression using neuro-symbolics to increase processing time, transfer, and training times for Medical Artificial Intelligence Models through pre-vectorization.

Representation Phases:

In working with multimedia, it is important to follow steps of standardization in which all new data which enters a system will be bound. This process diverges data by collecting the data points, converges the data, and allows for novel trends to emerge.

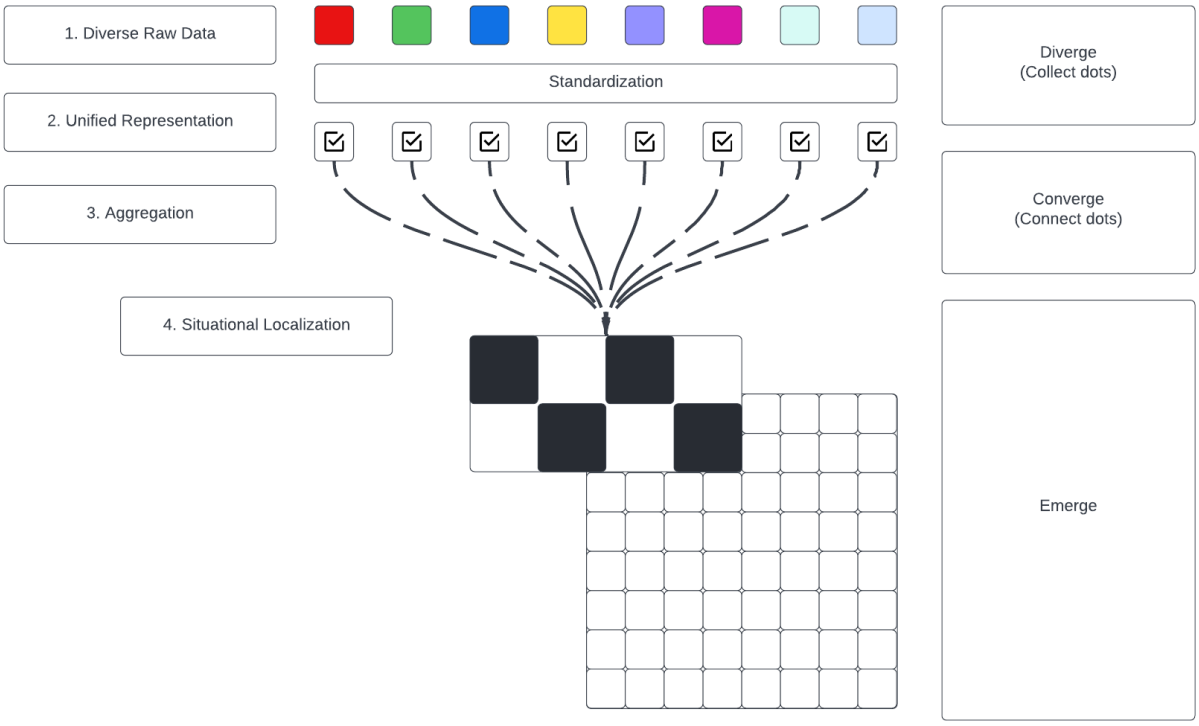


Figure 4-bis – Multimedia representation phases for radiology images

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's important to efficiently store the multimedia which is usually large in size, and sometimes sparse in situation. For training artificial intelligence models, this data must be vectorized in order to train ontologies of disease and diagnosis. In Figure 4-ter, we show 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.

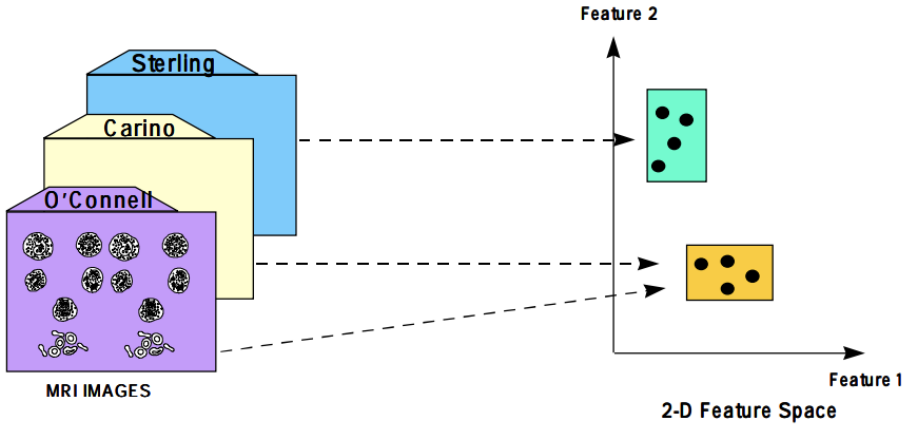


Figure 4-ter – Early example of vectorizing medical imagery [86]

Hilbert Symbolics:

We introduce a method of Lossless Hilbert compression using neuro-symbolics as an effective strategy for parallel computation of medical imagery, as illustrated in Figures 4-5(a) and (b). 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.

Where 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 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 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.

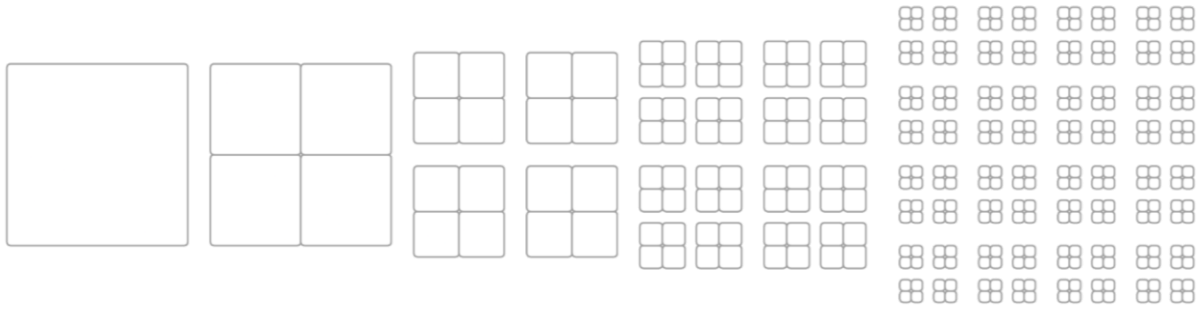


Figure 4-5(a) – Illustration of Lossless Hilbert compression

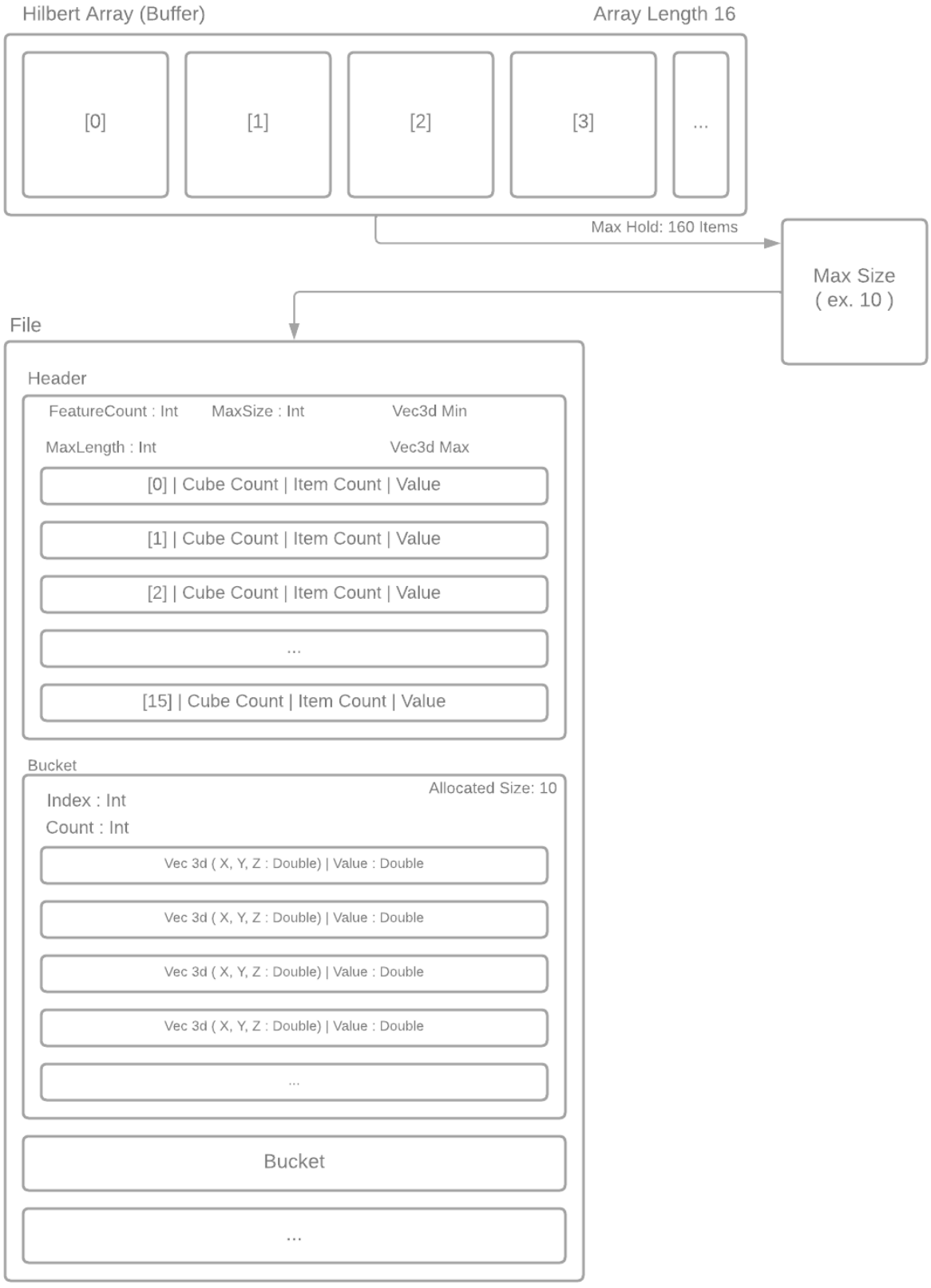


Figure 4-5(b) – Illustration of Lossless Hilbert compression

NIST Medical Databank

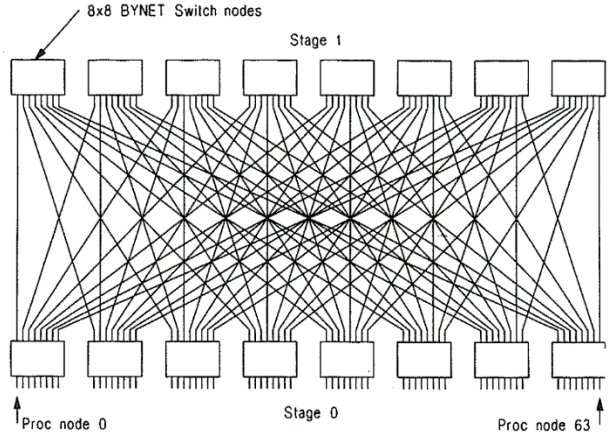


Figure 5. Bynet network topology.

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.

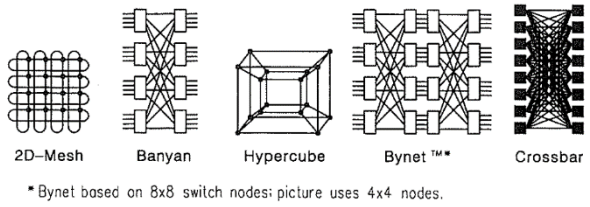


Figure 8. Interconnection network topologies.

Figure 4-6 – Example of interconnection network topologies [87]

We reference early network topology 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.

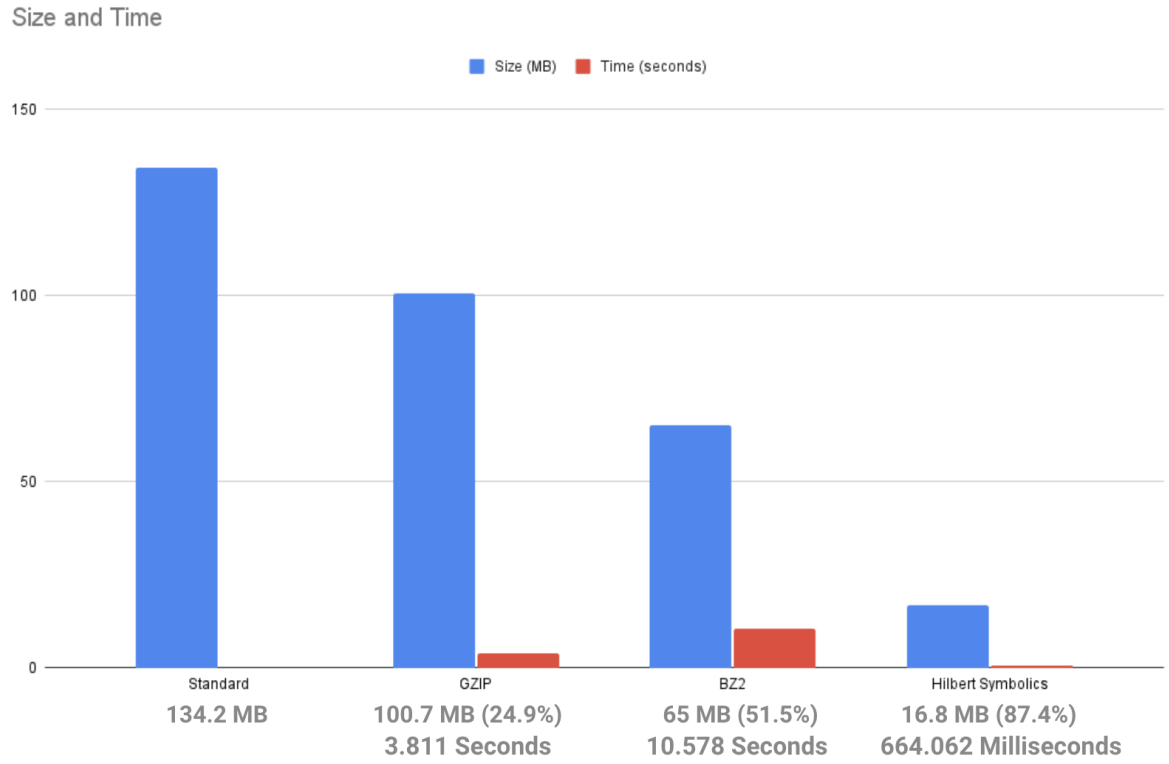
Performance Benchmarks:

In testing benchmarks we show the standard file size 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 results shown below indicate a compression of 24.9% for GZIP and 51.5% for BZ2. The processing time for GZIP taking 3.811 seconds, and BZ2 taking 10.578 seconds.

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 Milliseconds to process. The performance benefit being in ability to distribute the computation of each slice across 256 threads where each thread computes 2 slides. The total results of the benchmark is as shown in Table 4.

Table 4 – Compression performance comparison for various compression methods

	Standard	GZIP	BZ2	Hilbert Symbolics
Size (MB)	134.2	100.7	65	16.8
Time (seconds)		3.811621189	10.57886028	0.6640625



We provide further analysis of the processing time, and of the storage requirements of resulting files, as illustrated in Figure 4-7.

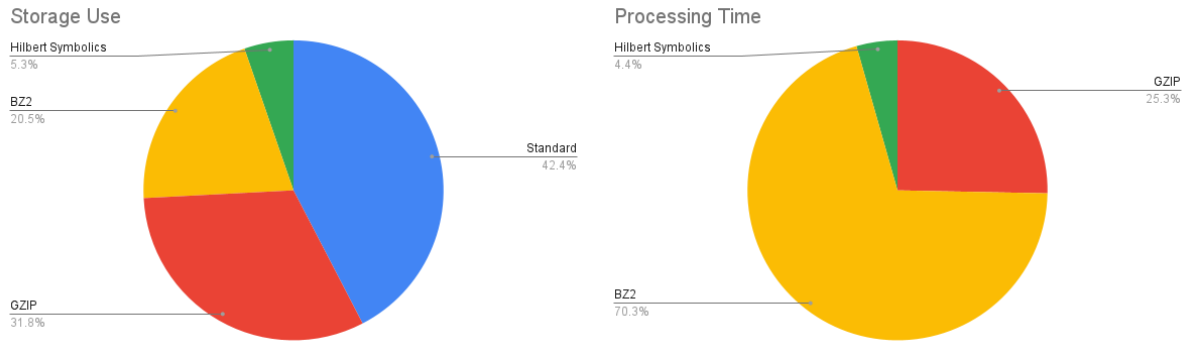


Figure 4-7 – Processing time and storage requirement for various compression methods

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 & metrics

The taxonomy used in grouping these evaluation metrics is that which was proposed by Cesar Ferri, et al. in their 2008 paper titled "An Experimental Comparison Of Performance Measures For Classification."

– Threshold Metrics

– Ranking Metrics

– Probability Metrics.

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 however not suitable for imbalanced classification because a poorly fitted model that simply predicts the majority class would end up having a misleading high score.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$$

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.

$$Classification\ Error = \frac{Incorrect\ Predictions}{Total\ Predictions}$$

Patient Level Accuracy & Image Level Accuracy

The patient level accuracy metric is defined as follows. For each patient, let *Nt* be the total number of images and *Nc* 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

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

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.

$$Image\ level\ Accuracy = \frac{C}{N}$$

Pixel Accuracy

In instance segmentation, pixel accuracy is used to evaluate the percent of pixels in an image which 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 you identify negative cases.

Exact Match Ratio (EMR)

The Exact Ratio 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,

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

Where,

$n \implies$ Number of training examples

$y^{(i)} \implies$ true labels for the ith training example

$\hat{y}^{(i)} \implies$ predicted labels for the ith training example

Example-Based Accuracy

This extends the Accuracy metrics to multilabel classification. The overall accuracy is the average of accuracy across training instances.

Macro Averaged Accuracy

This extends the Accuracy metric to multilabel classification. This metric computes the Accuracy of individual class labels and then averages over all classes.

Mathematically,

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

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

Where,

$n \implies$ Number of training examples

$y_j^{(i)} \implies$ true labels for the ith training example and jth class

$\hat{y}_j^{(i)} \implies$ predicted labels for the ith training example and jth class

$\wedge \implies$ logical AND operator

$\vee \implies$ logical OR operator

$k \implies$ Number of classes

Micro Averaged Accuracy

This extends the Accuracy metric to multilabel classification. This Label based metric computes the Accuracy globally over all instances and all class labels.

Mathematically,

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

Where,

$n \implies$ Number of training examples

$y_j^{(i)} \implies$ true labels for the ith training example and jth class

$\hat{y}_j^{(i)} \implies$ predicted labels for the ith training example and jth class

$\wedge \implies$ logical AND operator

$\vee \implies$ logical OR operator

$k \implies$ Number of classes

3.4.1.2 Sensitivity-specificity metrics

Sensitivity

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

$$Sensitivity = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative}$$

Specificity

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

$$Specificity = \frac{True\ Negative}{True\ Positive + False\ Negative}$$

Geometric mean (G-Mean)

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

$$G - Mean = \sqrt{sensitivity * specificity}$$

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.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

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.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

F-Measure

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 score or F1 score. It is the most popular metric for working with imbalanced datasets.

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

Fbeta-Measure

Fbeta 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.

$$Fbeta - Measure = \frac{((1 + beta^2) * Precision * Recall)}{(beta^2 * Precision + Recall)}$$

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 Chicco [6] is more informative than F1 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.

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

where $n$is the total number of observations.

MCC could also be calculated directly from the confusion matrix as;

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

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 multilabel classification. This metric computes the Precision of individual class labels and then averages over all classes.

Mathematically,

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

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

Where,

$n \implies$ Number of training examples

$y_j^{(i)} \implies$ true labels for the ith training example and jth class

$\hat{y}_j^{(i)} \implies$ predicted labels for the ith training example and jth class

$\wedge \implies$ logical AND operator

$P^j_{macro} \implies$ Precsion for label/class k

$k \implies$ Number of classes

Micro Averaged Precision

This extends the Precision metric to multilabel classification. This Label based metric computes the Precision globally over all instances and all class labels.

Mathematically,

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

Where,

$n \implies$ Number of training examples

$y_j^{(i)} \implies$ true labels for the ith training example and jth class

$\hat{y}_j^{(i)} \implies$ predicted labels for the ith training example and jth class

$\wedge \implies$ logical AND operator

$\vee \implies$ logical OR operator

$k \implies$ Number of classes

Macro Averaged Recall

This extends the Precision metric to multilabel classification. This metric computes the Precision of individual class labels and then averages over all classes.

Mathematically,

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

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

Where,

$n \implies$ Number of training examples

$y_j^{(i)} \implies$ true labels for the ith training example and jth class

$\hat{y}_j^{(i)} \implies$ predicted labels for the ith training example and jth class

$\wedge \implies$ logical AND operator

$R_{macro}^{j} \implies$ Recall for label/class k

$k \implies$ Number of classes

Micro Averaged Recall

This extends the Precision metric to multilabel classification. This Label based metric computes the Precision globally over all instances and all class labels.

Mathematically,

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

Where,

$n \implies$ Number of training examples

$y_j^{(i)} \implies$ true labels for the ith training example and jth class

$\hat{y}_j^{(i)} \implies$ predicted labels for the ith training example and jth class

$\wedge \implies$ logical AND operator

$\vee \implies$ logical OR operator

$k \implies$ Number of classes

Negative Predictive Value (NPV)

The negative predictive value 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.

$$NPV = \frac{True\ Negative}{True\ Negative + False\ Negative}$$

3.4.2 Ranking metrics

Receiver Operating Characteristic (ROC) Curve

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

Precision-Recall Curve

Precision-Recall curve is also a graphical plot used to summarise the diagnostic ability of a classification model. ROC curves can be misleading with an imbalanced dataset, especially when the 'negative' samples are 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 area under the curve could be used. It is created by plotting the precision score against the recall score (sensitivity).

Average Precision (AP)

It 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 (mAP)

Average Precision is calculated individually for each class. In an objection task with many classes, mAP is the average of all the AP values over all the classes. mAP is defined as;

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

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 groundtruth 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. Src It ranges between 0.0 and 1.0.

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

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.

$$\text{BrierSkillScore} = 1 - \left(\frac{Brier\ Score}{Brier\ Score\ reference}\right)$$

Intersection Over Union (IoU)

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*.

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

Hamming Loss

Hamming Loss is used to calculate the proportion of incorrectly predicted labels to the total number of labels. When applied to multilabel 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,

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

Where,

$n \implies$ Number of training examples

$y_j^{(i)} \implies$ true labels for the ith training example and jth class

$\hat{y}_j^{(i)} \implies$ predicted labels for the ith training example and jth 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,

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

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

Where,

$M_x \implies$ Number of Missed labels / False Negatives

$F_x \implies$ Number of False Positives

$Y_x \implies$ TP + FN

$P_x \implies$ TP + FP

$\vee \implies$ 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 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 (currently C-004)

3.6 Benchmarking methodology and architecture

– technical architecture

– hosting (IIC, etc.)

– possibility of an online benchmarking on a public test dataset

– protocol for performing the benchmarking (who does what when etc.)

– AI submission procedure including contracts, rights, IP etc. considerations

3.6.1 Audit trial

We conducted an audit trial using the undisclosed test data for the diagnosis of COVID-19 via Chest X-Ray. We used the machine learning auditing platform from the Open Code Initiative; health.aiaudit.org. This platform will automate the assessment of AI systems.

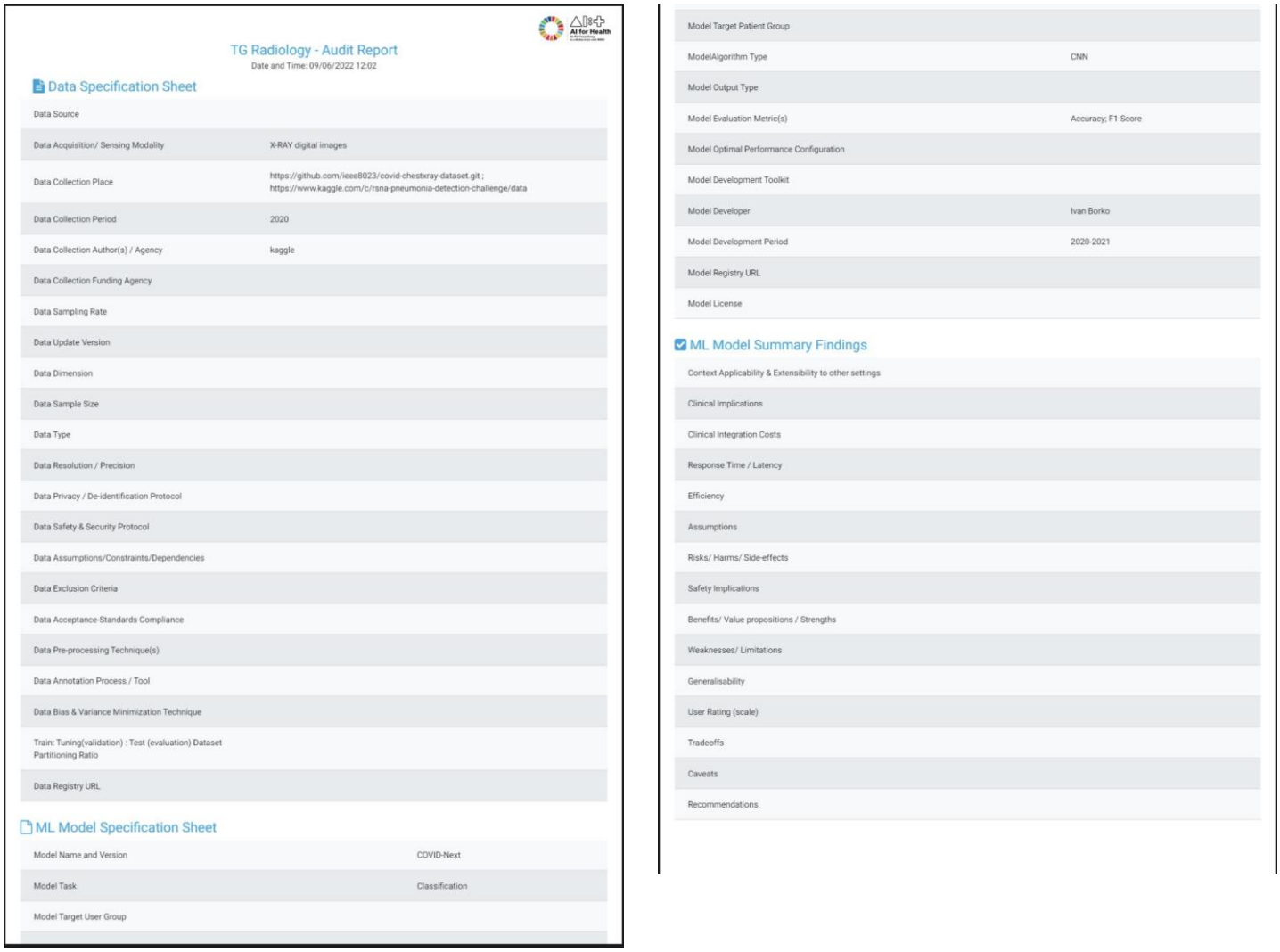


Figure 5 – Model Results after trial audits using the benchmarking platform, health.aiaudit.org

3.6.2 Audit trial checklist

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

The checklist is below.

Working Draft:

Table 1.0 consists of a minimum viable set of audit verification checklist items. This checklist is basically derived from the FG-AI4H standardized model survey questionnaire: Reference document J-038 on FG-AI4H server (2020). URL https://extranet.itu.int/sites/itu-t/focusgroups/ai4h. You can see from the table that the checklist items are categorized on the basis of their respective ML4H lifecycle stage, the applicable assessment criteria, 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 / use-case specific considerations and relevance.

Task Description:

1. Please perform an expert review of the given checklist items and may try to provide your 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 whether it covers all the relevant ML4H lifecycle requirements (ML technology, Clinical, Regulatory and Ethical requirements). If 'NO', please indicate the missing aspects

   b) From the given set, are there any checklist items that you find conflicting or ambiguous to the defining context and hence needs further clarification, correction, modification or substitution? If 'Yes', please indicate them

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

   d) Would you like to propose any additional checklist items? If 'Yes', please indicate them along with the respective inclusion criteria

2. Based on your expert assessment, please assign a 'significance level' / 'conformance priority' to each of the checklist items listed under column-6 titled 'Significance Level'. A first level criteria 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 4 types: 'mandatory', 'preferred', 'conditional' or 'optional' based on its TG / use-case specific significance

   Purpose: This set of verification checklists are reviewed, finalized, vetted and approved by the audit experts. Then this approved set of checklists is served as a 'questionnaire' to the TG Use Case developers to fill in their response. The response / results are verified(with the help of quantitative and qualitative records / proofs/ evidence) and validated (by applicable test cases) for conformity assessment to generate an audit report finally.

   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 checklist / questionnaire, we truly encourage both parties (audit experts and TG / domain experts) to collaboratively work 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 / metric	Significance level	Remarks	Verification & validation record / proof
Planning	Regulatory assessment	Qualitative	Product name and version	Intended use / 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 / Product specification	Mandatory	Clinical validation setting should fit to the intended use of the model	
Planning	Regulatory assessment	Qualitative	Primary product function − Primary function − Secondary function (if applicable) e.g. – Classification – Prognosis – Matching – Labeling – Detection – Segmentation – Recommendation – Data Modeling – Other	Intended use / 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 / Product specification	Preferred		
Planning	Regulatory assessment	Qualitative	Primary product user group − Primary user group − Secondary user group (if applicable)	Intended use / Product specification	Mandatory		
Planning	Regulatory assessment	Qualitative	Product operational mode − fully automatic − semi-automatic	Intended use / Product specification	Mandatory		

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute / metric	Significance level	Remarks	Verification & validation record / proof
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 / Product specification	Mandatory	Saul: Mandatory Preferred	
Data collection	Technical validation	Qualitative	Where and when was the training dataset collected from? Place: Time Period:	− Social representation bias − Historical data bias	Preferred	Saul: Mandatory Vasanth: Preferred, should not be mandatory	
Data collection	Technical validation	Quantitative	How many total data samples does the source dataset contain?	Sampling bias	Preferred	Saul: Which is the "original" dataset? Would not be better to refer to it as "source" dataset? Please clarify	
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	Saul: idem. Vasanth : 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		Saul: Whats the "reproduced" sample? Is it the target dataset?	
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	Saul: 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 & protection Patient safety	Mandatory	Saul: Mandatory	
Data collection	Regulatory assessment	Qualitative	Whether any due diligence and processes were followed in	Data privacy & protection	Mandatory	Saul: Preferred	

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute / metric	Significance level	Remarks	Verification & validation record / proof
			conformance to institutional review and ethical review policies when input datasets were de-identified / anonymised? Or were any exemptions obtained under special conditions?				
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	Saul: Preferred Vasanth: Preferred/ Mandatory	
Data preparation	Technical validation	Quantitative	If ground truth annotation was used as the basis for data labeling quality control, how did you evaluate the quality of ground truth annotation?	Data labeling bias	Mandatory	Saul: Mandatory	
Data preparation	Technical validation	Quantitative	For data labeling, 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 labeling bias	Preferred	Saul: Preferred, a little bit overlapping with the previous point.	
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-fitting / over-fitting	Mandatory	Saul: 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	Saul: Preferred	
Data preparation	Regulatory assessment	Qualitative	Is it possible to identify individuals from the dataset? Were the datasets de-identified / anonymised ? (Yes / No)	Data privacy	Mandatory	Saul: Mandatory	
Data preparation	Regulatory assessment	Qualitative	Type and level of de identification used like HIPAA complaint removal of private DICOM elements, image cropping to avoid identification from reconstructed images etc	data privacy	Mandatory	Preferred	
Data preparation	Regulatory assessment	Qualitative	How do you justify the selection of ground truth?	Data labeling quality	Preferred	Saul: Preferred	
Data preparation	Technical validation	Qualitative	Is the prevalence of the real world disease types/conditions reflected in the configuration of train datasets? (e.g. relative frequency of disease and non-disease types in the dataset)	Data bias	Preferred	Saul: 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	Saul: Preferred	

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute / metric	Significance level	Remarks	Verification & validation record / proof
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 accuracy' Vs 'Clinical accuracy' equivalence	Mandatory	Saul: Preferred Vasanth: 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, etc.)?	Model generalizability	Mandatory	Saul: Preferred. Instead of optimized, I would use "fine-tuned"	
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? Or does your use case give equal prominence to each output class type?	Model optimization	Preferred	Saul: 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 explainability	Mandatory	Saul: 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	Saul: Mandatory	
Model evaluation	Technical validation	Quantitative	Is there an interpretability-performance trade-off observed. If yes, provide the comparative analysis results.	Model interpretability &Model performance tradeoff	Preferred	Saul: Preferred	
Model evaluation	Technical validation	Quantitative	Specify the guarantees and limits of the performance metrics used for model evaluation	Model performance	Preferred	Saul: Preferred Vasanth: Mandatory	
Model evaluation	Technical validation	Quantitative	Specify the guarantees and limits of the 'gold standard' or 'reference standard' against which the performance metrics are evaluated	Model performance	Preferred	Saul: Preferred Vasanth: Mandatory	
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	Saul: Preferred	

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute / metric	Significance level	Remarks	Verification & validation record / proof
Model evaluation	Clinical evaluation	Quantitative	Whether any comparative analysis was done over the model safety risks with that of the alternative technologies (both ML based and Non-ML based)	Patient safety	Preferred	Saul: Mandatory	
Model evaluation	Clinical evaluation	Qualitative	Have you used any model-specific or model agnostic methods for model interpretability?	Model interpretability	Mandatory	Saul: preferred	
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	Saul: Mandatory	
Model usage /deployment	Clinical evaluation	Quantitative	Specify the computational efficiency of the model in terms of the response time	Clinical efficiency	Preferred	Saul: Preferred	
Model usage /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 / treatment – percentage reduction in clinician cognitive workload – degree of automation / semi-automation introduced – degree of smartness/intelligence augmentation – new knowledge discovery – enabling replacement or redefinition of existing gold standard – other.....	Clinical integration	Mandatory	Saul: Mandatory	
Model usage /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	Clinical effectiveness	Mandatory	Saul: Preferred	

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute / metric	Significance level	Remarks	Verification & validation record / proof
			– other……				
Model usage /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	Saul: Mandatory	
Model usage /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? And have you specified the values or ranges for performance metrics in order to avoid unacceptable risks?	Patient safety	Mandatory	Saul: Mandatory	
Model usage /deployment	Clinical evaluation	Qualitative	Was 'input data feature 'importance validated for its significance in the clinical setting by the clinician/ specialist? Which of the features were ranked as the most important ones?	Model interpretability	Preferred	Saul: Preferred	
Model usage /deployment	Clinical evaluation	Qualitative	Did the model fail to address any relevant clinically important findings?	Clinical effectiveness	Preferred	Saul: Preferred	
Model usage /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	Saul: Mandatory	
Model usage /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 / Patient safety	Mandatory	Saul: Mandatory	
Model maintenance & versioning	Regulatory assessment	Qualitative	s the ML model maintained as (a) a static system or (b) a continuously learning system? I	Model maintainability	Mandatory	Saul: Mandatory	
Model maintenance & versioning	Regulatory assessment	Quantitative	If the ML model is attributed to a continuous learning system , specify the algorithm change / update cycle	Model maintainability	Mandatory	Saul: Preferred	
Saul: Model maintenanc	Regulatory assessment	Quantitative	Has there been a proper plan for test data quality and correctness assessment after model deployment	Model maintainability	Mandatory	Saul: Mandatory	

ML4H process lifecycle stage	Assessment criteria	Assessment type	Audit verification checklist item	Assessment attribute / metric	Significance level	Remarks	Verification & validation record / proof
e & versioning			(i.e., concept drift, training/test data distribution mismatch, etc.)?				

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 United States of America.

Background: A rapid and accurate diagnosis of cardiomegaly and pleural effusion is of the utmost importance to reduce mortality and medical costs. Artificial Intelligence has shown promise in diagnosing medical conditions. With this study, we seek to evaluate how well Artificial Intelligence (AI) systems, developed my 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 (20) images were selected from the VinBig Data Chest X-ray dataset, another twenty-one (21) images were selected from the Chexpert dataset, and fifty nine (59) images were selected from the Euracare dataset, an in-house dataset collected by minoHealth AI Labs from Euracare Advanced Diagnostics and Heart Centre, a top-tier health institution in Accra, Ghana. The Deep Learning models were further tested on a larger Ghanaian dataset containing five hundred and sixty one (561) samples. Two AI systems were then evaluated on the evaluation dataset, whilst we also gave the same chest x-ray images within the evaluation dataset to 4 radiologists, with 5 - 20 years experience, to diagnose independently.

Results: For cardiomegaly, minoHealth.ai systems scored Area under the Receiver operating characteristic Curve (AUC-ROC) of 0.9 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, negative predictive value (NPV), and positive predictive value (PPV) 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

We are proposing a radiograph-agnostic benchmarking platform and framework that would allow for the evaluation of AI radiological systems for various conditions and serve as a standard. This would require registered developers and organisations seeking to evaluate their A.I system to download the test images and a csv file with two columns; 'ID', containing the unique Identification of each test image and 'Class' which would be left blank in order to be populated by the outputs of an A.I system. Developers are then to submit the fully populated csv file, which would then provide the model's outputs 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 your model's outputs to be populated in the CSV file.

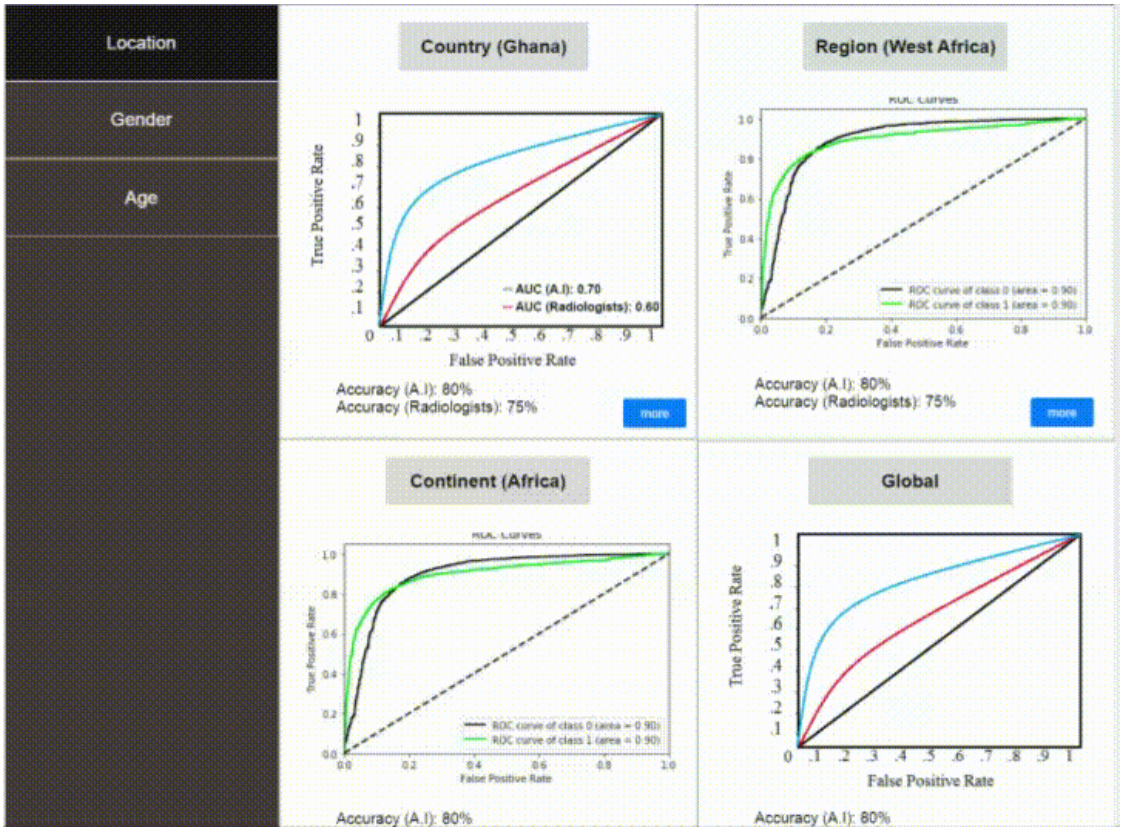


Figure 6 – A prototype of the radiograph-agnostic precision evaluation platform

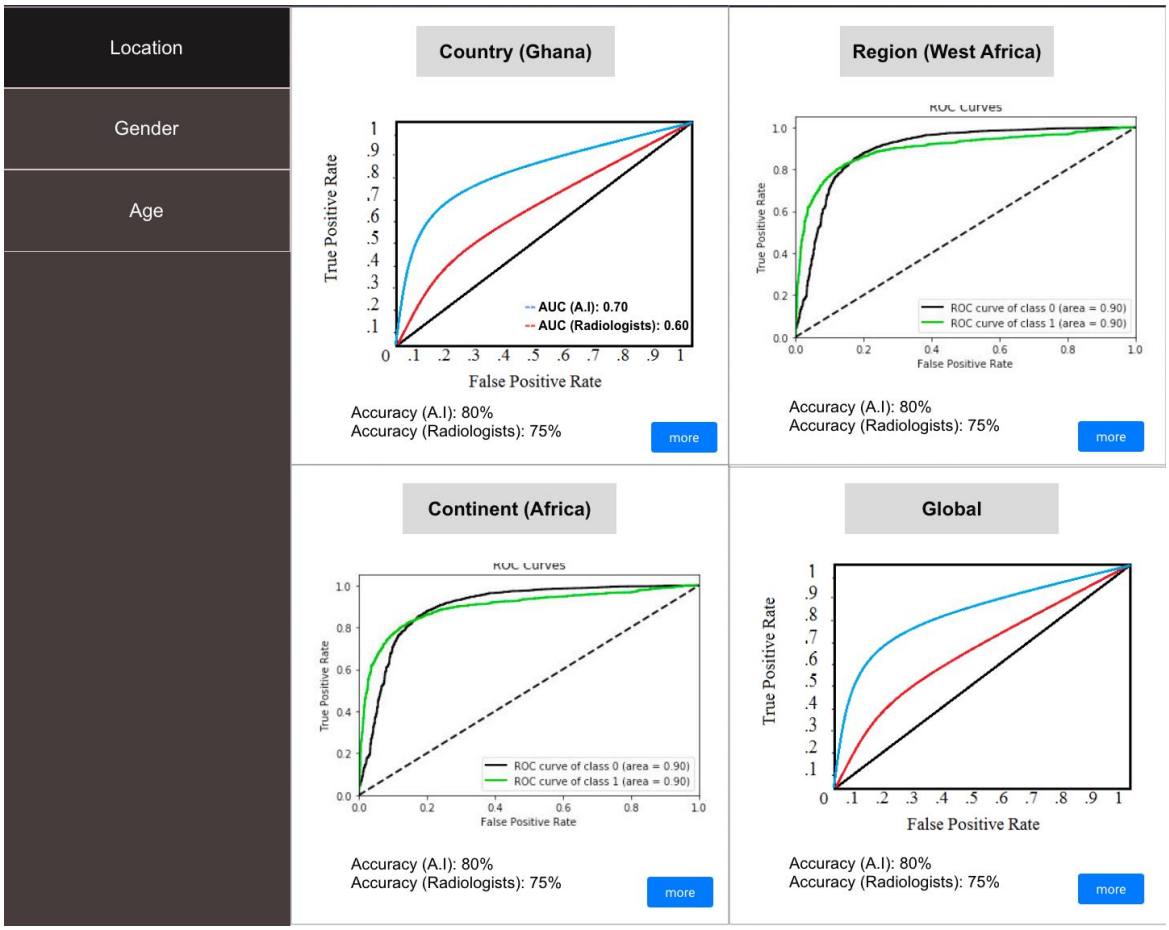


Figure 6-bis – The 'Location' category with its sub-categories and the metrics used

3.6.5 Evaluation metrics

All our supported condition tests on the platform would be image classification tasks and therefore we would be using evaluation metrics for classification. Some of the conditions and tests would be binary classification tasks while others would be multi-class classification, therefore we would be using metrics that can be used for both types of classification. As shown in Figure 1 and Figure 2, the evaluation metrics to be used would be the Receiver Operating Characteristic (ROC) curve, its Area Under the Curve (AUC) score and the Accuracy Score. The ROC curve and AUC score would help us identify the model's true positive rate (TPR) (Sensitivity) and its false positive rate (FPR) (1 - Specificity). Though originally for binary classification, the ROC curve and AUC score can be generalised to multi-class classification.

The performance of an A.I 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 each A.I system.

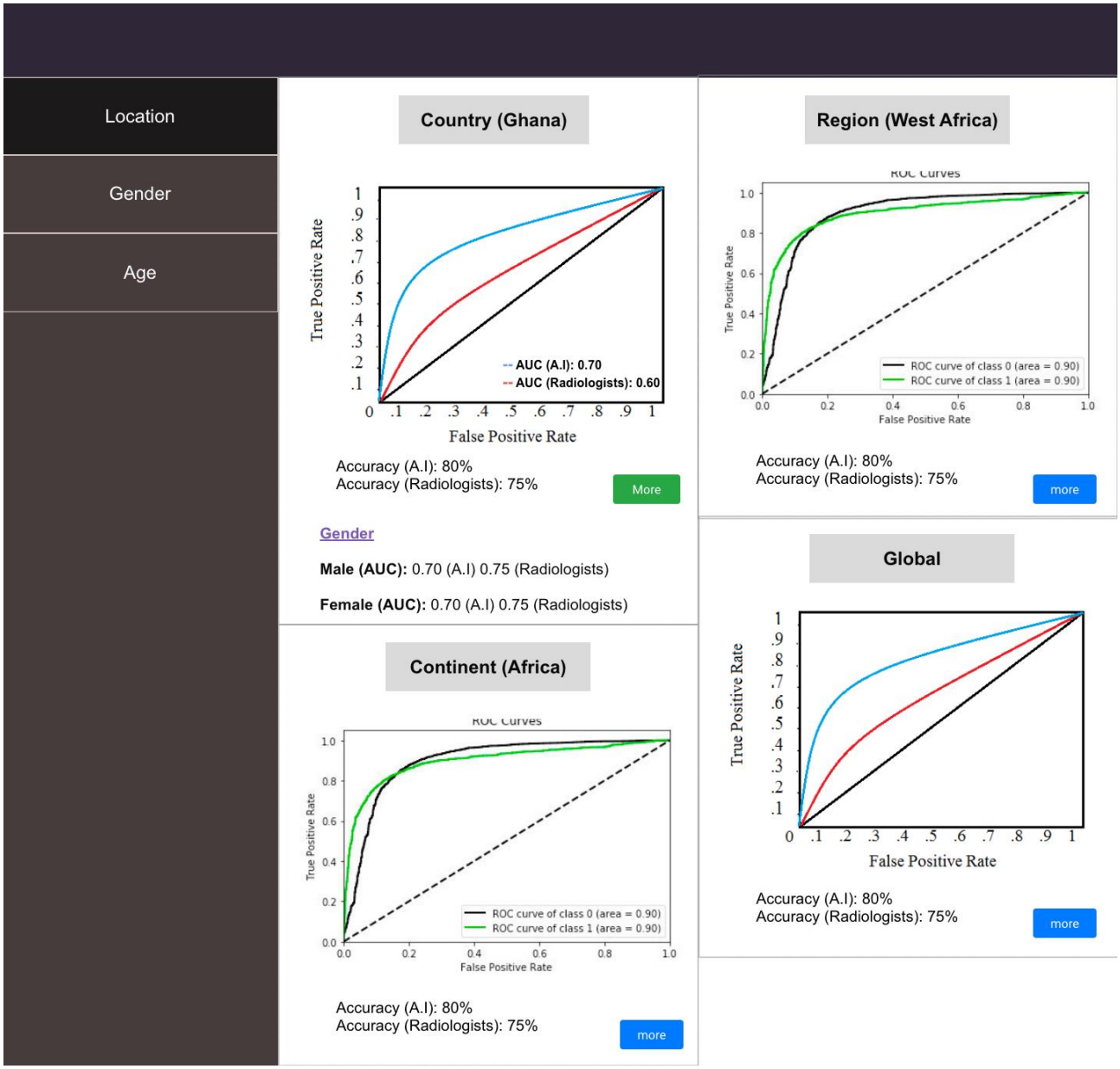


Figure 7 – Each sub-category would feature demographics intersection performances too

3.6.6 Benchmark categorizations

The evaluation results would be divided into Location, Gender and Age, as shown in Figure 1. 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 A.I system within the very nation it was developed. The 'Continent' sub-category would show how well the model performs on data from the continent it was developed in, this would help the developers know how well they can scale the current version of their A.I system. 'Region' specifically focuses on the performance of the AI system within the sub-continental region 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. And finally, 'Global' shows how well the model performs on data from across the world, showing its ability to truly generalise. Each of the subcategories 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 within each geographical area, how well their AI system generalises across gender and age.

Under 'Gender', there would be two main sub-categories, Male and Female, as shown in Figure 1 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 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, an AI system's performance on each of the age groups would be shown and it'd 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 generalises across demographics.

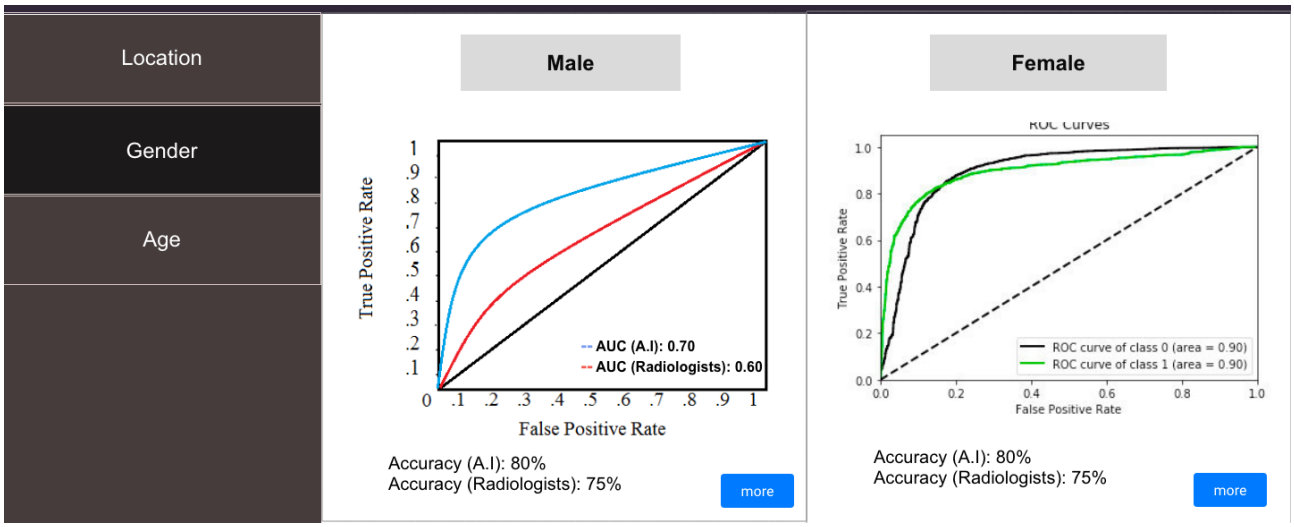


Figure 8 – The 'Gender' category

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 incentive to facilities to contribute more data to the platform. Submitted

radiographs should be accompanied by a csv file with information about the patient's gender, age and imaging facility's location. This would allow for the proposed Precision Evaluation framework.

3.6.8 The 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 edge 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 they contribute to. 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, we would ideally have radiologists from different parts of the world 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 us to compare an A.I system's performance on test images within each of the 'Location' sub-category with radiologists also within such geographical regions. This would more appropriately help us 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 that data can be made available to the benchmarking platform. With the collaboration of various members and organisations affiliated with FG-AI4H, we can collect more radiographs 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's possible to modularise it and build with different levels of complexities. It is also possible for the categories and subcategories to adjust based on the number and diversity of samples as well as radiologists available. If the evaluation data for a particular condition isn't large enough to support all four subcategories of 'Location', it can be limited to just 'Region' or 'Continent' and 'Global'. If there weren't 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 across. The same can apply to the sub-categories of Gender and Age. We would also start implementing the platform with chest x-rays for 12 different thoracic diseases supported in MIMIC-CXR, CheXpert and NIH Chest XRay datasets.

3.9 Privacy and security

Anonymised data can be de-anonymised using techniques like linkage attacks. Linkage attacks 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 anonymised 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, the developers and testing radiologists are only given access to test images without demographics data. To further

defend against this attack, we are abstracting 'Date of Birth' to just the Age (in years) of the patient when they were imaged, and we can abstract the location to just 'Country'. To add additional security measures as far as the panel of expert radiologists has access to such demographics data, we can explore variations of Differential Privacy.

Also, we are ensuring a secure system by demanding that developers and organisations that require a standardised evaluation of their A.I systems register before they'd be allowed to. The registration process can include an in-person assessment by their local World Health Organisation (W.H.O) or ITU branch office, just to ensure they are a valid institution, startup 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. And even the images and data they submit to the platform would be evaluated before being added to the system. All radiologists, both in the 'panel of expert radiologists' and the 'testing radiologists' would have to register and be verified before being allowed to contribute to the platform.

In order to not infringe upon the Intellectual Properties (IP) rights of AI developers and organisations, they would not be required to submit their A.I system itself. They are only supposed to submit the outputs (csv file) of their AI system, which would then be used for the evaluation of their system.

3.10 Impact

There exists a large amount of publicly available medical image datasets online, and there have been a lot of research and development with such datasets. By developing frameworks that target these conditions first, we would make the standardized benchmarking platform immediately appealing to the A.I healthcare research and development community. This would also help speedup the deployment of AI solutions in Radiology globally. AI healthcare system developers and organisations usually have to go through the challenge of convincing health facilities to share their private data with them, such data unfortunately aren't always of high quality and they usually lack the broad demographic representations needed to truly assess how well an A.I system generalises. 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 A.I systems by ensuring they are tested in great detail across various groups. It'd 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 leaderboards vs. private leaderboards

– Credit-Check like on approved 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] K. Kersting, "Machine Learning and Artificial Intelligence: Two Fellow Travelers on the Quest for Intelligent Behavior in Machines," Front. Big Data, vol. 1, p. 6, Nov. 2018.

[2] A. H. Fielding, "An introduction to machine learning methods," in Machine Learning Methods for Ecological Applications, Springer US, 1999, pp. 1–35.

[3] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85–117, Jan. 2015.

[4] E. Considerations, "Artificial Intelligence in Radiology — Ethical Considerations," pp. 1–9, 2020.

[5] T. Akinci, "current radiology landscape," no. 5, pp. 504–511, 2020.

[6] J. R. Geis, P. Adrian, F. C. C. Wu, and J. Spencer, "Ethics of Artificial Intelligence in Radiology : Summary of the Joint European and North American Multisociety Statement," 2019.

[7] Parveen NRS, Sathik MM. Detection of Pneumonia in chest X-ray images. Journal of X-Ray Science and Technology 2011;19:423–8. https://doi.org/10.3233/XST-2011-0304.

[8] Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 2018;172:1122-1131.e9. https://doi.org/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. Clinical Imaging 2020;61:15–9. https://doi.org/10.1016/j.clinimag.2020.01.008.

[10] Filice RW, Stein A, Wu CC, Arteaga VA, Borstelmann S, Gaddikeri R, et al. Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset. Journal of Digital Imaging 2020;33:490–6. https://doi.org/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. BioMedical Engineering Online 2018;17:113. https://doi.org/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, IEEE Computer Society; 2014. https://doi.org/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. Lecture Notes in Bioengineering, Springer, Singapore; 2014, p. 75–97. https://doi.org/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. Journal of Digital Imaging 2017;30:427–41. https://doi.org/10.1007/s10278-017-9955-8.

[15] Cicero M, Bilbily A, Colak E, Dowdell T, Gray B, Perampaladas K, et al. Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs. Investigative Radiology 2017;52:281–7. https://doi.org/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. n.d.

[17] Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks 2020;43:635–40. https://doi.org/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. n.d.

[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. n.d.

[20] Brunese L, Mercaldo F, Reginelli A, Santone A. Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays. Computer Methods and Programs in Biomedicine 2020;196:105608. https://doi.org/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. Computers in Biology and Medicine 2020;121:103792. https://doi.org/10.1016/j.compbiomed.2020.103792.

[22] Hwang EJ, Park CM. Clinical implementation of deep learning in thoracic radiology: Potential applications and challenges. Korean Journal of Radiology 2020;21:511–25. https://doi.org/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. https://doi.org/10.1371/journal.pone.0236621.

[24] Matsumoto T, Kodera S, Shinohara H, Ieki H, Yamaguchi T, Higashikuni Y, et al. Diagnosing Heart Failure from Chest X-Ray Images Using Deep Learning. International Heart Journal 2020;61:781–6. https://doi.org/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. Australasian Physical and Engineering Sciences in Medicine 2019. https://doi.org/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 Radiology 2020;49:613–8. https://doi.org/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. Journal of Forensic Medicine 2017;33. https://doi.org/10.3969/j.issn.1004-5619.2017.06.013.

[28] Li Q, Zhong L, Huang H, Liu H, Qin Y, Wang Y, et al. Auxiliary diagnosis of developmental dysplasia of the hip by automated detection of Sharp's angle on standardized anteroposterior pelvic radiographs. Medicine (United States) 2019;98. https://doi.org/10.1097/MD.0000000000018500.

[29] Kitamura G. Deep learning evaluation of pelvic radiographs for position, hardware presence, and fracture detection. European Journal of Radiology 2020;130. https://doi.org/10.1016/j.ejrad.2020.109139.

[30] Zheng G, Nolte LP. Computer-aided orthopaedic surgery: State-of-the-art and future perspectives. Advances in Experimental Medicine and Biology, vol. 1093, Springer New York LLC; 2018, p. 1–20. https://doi.org/10.1007/978-981-13-1396-7_1.

[31] Unberath M, Zaech JN, Gao C, Bier B, Goldmann F, Lee SC, et al. Enabling machine learning in X-ray-based procedures via realistic simulation of image formation. International Journal of Computer Assisted Radiology and Surgery 2019;14:1517–28. https://doi.org/10.1007/s11548-019-02011-2.

[32] Vaquero, J. J., & Kinahan, P. (2015). Positron Emission Tomography: Current Challenges and Opportunities for Technological Advances in Clinical and Preclinical Imaging Systems. Annual review of biomedical engineering, 17, 385–414. https://doi.org/10.1146/annurev-bioeng-071114-040723

[33] Mawlawi, O., Podoloff, D. A., Kohlmyer, S., Williams, J. J., Stearns, C. W., Culp, R. F., Macapinlac, H., & National Electrical Manufacturers Association (2004). Performance characteristics of a newly developed PET/CT scanner using NEMA standards in 2D and 3D modes. Journal of nuclear medicine: official publication, Society of Nuclear Medicine, 45(10), 1734–1742.

[34] Shiraishi, J., Li, Q., Appelbaum, D., & Doi, K. (2011, November). Computer-aided diagnosis and artificial intelligence in clinical imaging. In Seminars in nuclear medicine (Vol. 41, No. 6, pp. 449-462). WB Saunders.

[35] Imaging Modalities. (2016, December 02). Retrieved July 27, 2020, from https://www.who.int/diagnostic_imaging/imaging_modalities/en/

[36] Sia, M. (n.d.). Radiology basics - Imaging modalities. Retrieved July 27, 2020, from https://www.radiologycafe.com/medical-students/radiology-basics/imaging-modalities

[37] Fluoroscopy Procedure. (n.d.). Retrieved July 27, 2020, from https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/fluoroscopy-procedure

[38] Weese, J., Penney, G., Desmedt, P., Buzug, T., Hill, D., & Hawkes, D. (1997). Voxel-based 2-D/3-D registration of fluoroscopy images and CT scans for image-guided surgery. IEEE Transactions on Information Technology in Biomedicine, 1(4), 284-293. doi:10.1109/4233.681173

[39] Bang, J. Y., Hough, M., Hawes, R. H., & Varadarajulu, S. (2020). Use of Artificial Intelligence to Reduce Radiation Exposure at Fluoroscopy-Guided Endoscopic Procedures. The American Journal of Gastroenterology, 115(4), 555-561. doi:10.14309/ajg.0000000000000565

[40] "Review on the applications of ultrasonography in ... - NCBI." 28 Jan. 2016, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4731348/. Accessed 16 Sep. 2020

[41] "Ultrasound: Purpose, Procedure, and Preparation - Healthline." https://www.healthline.com/health/ultrasound. Accessed 16 Sep. 2020.

[42] "Ultrasound (Sonography) - RadiologyInfo.org." https://www.radiologyinfo.org/en/info.cfm?pg=genus. Accessed 16 Sep. 2020.

[43] "4 Types of Ultrasound Imaging • Ultrasound Technician." https://www.ultrasoundtechniciancenter.org/ultrasound-knowledge/medical-ultrasound-imaging-types.html. Accessed 17 Sep. 2020.

[44] "Deep Learning in Medical Ultrasound Analysis: A Review ...." https://www.sciencedirect.com/science/article/pii/S2095809918301887. Accessed 17 Sep. 2020.

[45] "Nuclear Medicine - WHO." https://www.who.int/diagnostic_imaging/imaging_modalities/dim_nuclearmed/en/. Accessed 17 Sep. 2020.

[46] "Nuclear Medicine, General - RadiologyInfo.org." https://www.radiologyinfo.org/en/info.cfm?pg=gennuclear. Accessed 17 Sep. 2020.

[47] "Nuclear medicine - Wikipedia." https://en.wikipedia.org/wiki/Nuclear_medicine. Accessed 17 Sep. 2020

[48] "Artificial intelligence and radiomics in nuclear medicine ...." 15 Nov. 2019, https://link.springer.com/article/10.1007/s00259-019-04593-0. Accessed 17 Sep. 2020

[49] "Mammography - Wikipedia." https://en.wikipedia.org/wiki/Mammography. Accessed 17 Sep. 2020.

[50] "Mammography (Mammogram) - RadiologyInfo.org." https://www.radiologyinfo.org/en/info.cfm?pg=mammo. Accessed 17 Sep. 2020.

[51] "Study: AI improves radiologists' readings of mammograms ...." 2 Mar. 2020, https://newsroom.uw.edu/news/study-ai-improves-radiologists-readings-mammograms. Accessed 17 Sep. 2020.

[52] "Magnetic Resonance Imaging (MRI)." https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri. Accessed 18 Sep. 2020.

[53] "Magnetic resonance imaging - Wikipedia." https://en.wikipedia.org/wiki/Magnetic_resonance_imaging. Accessed 18 Sep. 2020.

[54] "Magnetic resonance imaging - WHO." https://www.who.int/diagnostic_imaging/imaging_modalities/dim_magresimaging/en/. Accessed 18 Sep. 2020.

[55] "Artificial intelligence enhances MRI scans | National Institutes ...." 10 Apr. 2018, https://www.nih.gov/news-events/nih-research-matters/artificial-intelligence-enhances-mri-scans. Accessed 18 Sep. 2020.

[56] "Artificial intelligence in radiology - NCBI - NIH." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6268174/. Accessed 18 Sep. 2020.

[57] "Angiography - WHO." https://www.who.int/diagnostic_imaging/imaging_modalities/dim_angiography/en/. Accessed 18 Sep. 2020.

[58] "Angiography - Wikipedia." https://en.wikipedia.org/wiki/Angiography. Accessed 18 Sep. 2020.

[59] "Artificial intelligence in cardiovascular imaging: state of the art ...." 9 Aug. 2019, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6712136/. Accessed 18 Sep. 2020.

[60] Sharif, M. S., & Amira, A. (2009, November). An intelligent system for PET tumour detection and quantification. In 2009 16th IEEE International Conference on Image Processing (ICIP) (pp. 2625-2628). IEEE.

[61] Lakhan, S. E., Kaplan, A., Laird, C., & Leiter, Y. (2009). The interventionalism of medicine: interventional radiology, cardiology, and neuroradiology. International Archives of Medicine, 2(1), 27.

[62] Iezzi, R., Goldberg, S. N., Merlino, B., Posa, A., Valentini, V., & Manfredi, R. (2019). Artificial Intelligence in Interventional Radiology: A Literature Review and Future Perspectives. Journal of oncology, 2019.

[63] Abajian, A., Murali, N., Savic, L. J., Laage-Gaupp, F. M., Nezami, N., Duncan, J. S., ... & Chapiro, J. (2018). Predicting treatment response to intra-arterial therapies for hepatocellular carcinoma with the use of supervised machine learning—an artificial intelligence concept. Journal of Vascular and Interventional Radiology, 29(6), 850-857.

[64] Asadi, H., Dowling, R., Yan, B., & Mitchell, P. (2014). Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. PloS one, 9(2), e88225.

[65] Asadi, H., Kok, H. K., Looby, S., Brennan, P., O'Hare, A., & Thornton, J. (2016). Outcomes and complications after endovascular treatment of brain arteriovenous malformations: a prognostication attempt using artificial intelligence. World neurosurgery, 96, 562-569.

[66] Wachs, J. P., Stern, H. I., Edan, Y., Gillam, M., Handler, J., Feied, C., & Smith, M. (2008). A gesture-based tool for sterile browsing of radiology images. Journal of the American Medical Informatics Association, 15(3), 321-323.

[67] El-Shallaly, G. E. H., Mohammed, B., Muhtaseb, M. S., Hamouda, A. H., & Nassar, A. H. M. (2005). Voice recognition interfaces (VRI) optimize the utilization of theatre staff and time during laparoscopic cholecystectomy. Minimally Invasive Therapy & Allied Technologies, 14(6), 369-371.

[68] Herniczek, S. K., Lasso, A., Ungi, T., & Fichtinger, G. (2014, March). Feasibility of a touch-free user interface for ultrasound snapshot-guided nephrostomy. In Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling (Vol. 9036, p. 90362F). International Society for Optics and Photonics.

[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., ... & Taylor, A. (2019). 03: 45 PM Abstract No. 38 The development of a machine learning smart speaker application for device sizing in interventional radiology. Journal of Vascular and Interventional Radiology, 30(3), S20.

[71] Letzen, B., Wang, C. J., & Chapiro, J. (2019). The Role of Artificial Intelligence in Interventional Oncology: A Primer. Journal of vascular and interventional radiology: JVIR, 30(1), 38.

[72] What is Computed Tomography? (2019, December 5). Retrieved September 19, 2020, from https://www.fda.gov/radiation-emitting-products/medical-x-ray-imaging/what-computed-tomography

[73] Scott C. Litin, M. (2020, March 11). Could CT scans cause cancer? Retrieved September 19, 2020, from https://www.mayoclinic.org/tests-procedures/ct-scan/expert-answers/ct-scans/faq-20057860

[74] Other Information Resources Related to Whole-Body CT Screening. (2019, June 14). Retrieved September 19, 2020, from https://www.fda.gov/radiation-emitting-products/medical-x-ray-imaging/other-information-resources-related-whole-body-ct-screening

[75] Ghanem, M. H. (1986). CT scan in psychiatry: A review of the literature. L'Encéphale: Revue de psychiatrie clinique biologique et thérapeutique.

[76] Li, W., Cui, H., Li, K., Fang, Y., & Li, S. (2020). Chest computed tomography in children with COVID-19 respiratory infection. Pediatric radiology, 1-4.

[77] Grillet, F., Behr, J., Calame, P., Aubry, S., & Delabrousse, E. (2020). Acute pulmonary embolism associated with COVID-19 pneumonia detected by pulmonary CT angiography. Radiology.

[78] Ahsan, M. M., Gupta, K. D., Islam, M. M., Sen, S., Rahman, M., & Hossain, M. S. (2020). 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 preprint arXiv:2007.12525.

[79] Zhao, J., Zhang, Y., He, X., & Xie, P. (2020). COVID-CT-Dataset: a CT scan dataset about COVID-19. arXiv preprint arXiv:2003.13865.

[80] Chassagnon, G., Vakalopoulou, M., Battistella, E., Christodoulidis, S., Hoang-Thi, T. N., Dangeard, S., ... & Hajj, S. E. (2020). AI-Driven CT-based quantification, staging and short-term outcome prediction of COVID-19 pneumonia. arXiv preprint arXiv:2004.12852.

[81] Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., ... & Ye, L. (2020). Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. Cell.

[82] Venugopal, V. K., Vaidhya, K., Murugavel, M., Chunduru, A., Mahajan, V., Vaidya, S., ... & Mahajan, H. (2020). Unboxing AI-Radiological Insights Into a Deep Neural Network for Lung Nodule Characterization. Academic Radiology, 27(1), 88-95.

[83] Seifert, R., Weber, M., Kocakavuk, E., Rischpler, C., & Kersting, D. (2020, September). AI and Machine Learning in Nuclear Medicine: Future Perspectives. In Seminars in Nuclear Medicine. WB Saunders.

[84] Gallo, M., Spigolon, L., Bejko, J., Gerosa, G., & Bottio, T. (2020). How to evaluate the outflow tract of LVAD after minimally invasive implantation by 3D CT-scan. Artificial Organs.

[85] Le, V., Frye, S., Botkin, C., Christopher, K., Gulaka, P., Sterkel, B., ... & Osman, M. (2020). Effect of PET Scan with Count Reduction Using AI-Based Processing Techniques on Image Quality. Journal of Nuclear Medicine, 61(supplement 1), 3095-3095.

[86] F. Cariño and W. Sterling, Parallel Strategies and Concepts for a Petabyte Multimedia Database Computer, *IEEE Parallel Database Techniques*, 1998.

[87] Felipe Cariño Jr., Warren Sterling, and Pekka Kostamaa, Industrial Database Supercomputer Exegesis: The DBC/1012, The NCR 3700, The Ynet, and The Bynet, Chapter 8, pp 9 11, PARLE Lecture Notes, 1992

[88] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature volume 542, pages 115–118

[89] Arleo, Elizabeth & Hendrick, R. Edward & Helvie, Mark & Sickles, Edward. (2017). Comparison of recommendations for screening mammography using CISNET models. Cancer. 123. 10.1002/cncr.30842.

[90] Bien N, Rajpurkar P, Ball RL, Irvin J, Park AK, Jones E, et al. AI-assisted diagnosis for knee MR: Development and retrospective validation. PLoS Med. 2018;15(11):e1002699. https://doi.org/10.1371/journal.pmed.1002699

[91] CBIS-DDSM. https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM

[92] CheXpert. https://stanfordmlgroup.github.io/competitions/chexpert/. https://arxiv.org/abs/1901.07031

[93] Clinical Radiology UK Workforce Census Report 2018 https://www.rcr.ac.uk/publication/clinical-radiology-uk-workforce-census-report-2018

[94] H A Haenssle, C Fink, R Schneiderbauer, F Toberer, T Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, L Thomas, A Enk, L Uhlmann, Reader study level-I and level-II Groups, Man against machine: diagnostic performance of a deep learning convolutional neural

network for dermoscopic melanoma recognition in comparison to 58 dermatologists, Annals of Oncology, Volume 29, Issue 8, August 2018, Pages 1836–1842, https://doi.org/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] MIMIC-CXR Dataset: https://archive.physionet.org/physiobank/database/mimiccxr/ https://arxiv.org/abs/1901.07042

[97] NIH Chest XRay: https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community. https://nihcc.app.box.com/v/ChestXray-NIHCC

[98] NHSX. Performance Assessment Call - National COVID-19 Chest Image Database documentation. https://nhsx.github.io/covid-chest-imaging-database/AI_Performance_Assessment.html

[99] RAD-AID in Liberia. https://www.rad-aid.org/countries/africa/liberia/

[100] Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of CheXNeXt to practicing radiologists. PLoS Med. 2018;15(11):e1002686. https://doi.org/10.1371/journal.pmed.1002686

[101] UCSF: Digital X-Ray On-The-Go in Kenya. https://radiology.ucsf.edu/blog/digital-x-ray-go-kenya

[102] Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

[103] Chicco, D. (2017). "Ten quick tips for machine learning in computational biology". BioData Mining. 10 (35): 35. doi:10.1186/s13040-017-0155

[104] Chicco, D., Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6 (2020). https://doi.org/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] An Tang, Roger Tam, Alexandre Cadrin-Chênevert, Will Guest, Jaron Chong, Joseph Barfett, Leonid Chepelev, Robyn Cairns, J. Ross Mitchell, Mark D. Cicero, Manuel Gaudreau Poudrette, Jacob L. Jaremko, Caroline Reinhold, Benoit Gallix, Bruce Gray, Raym Geis, Timothy O'Connell, Paul Babyn, David Koff, Darren Ferguson, Sheldon Derkatch, Alexander Bilbily, Wael Shabana,

[107] Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology,

[108] Canadian Association of Radiologists Journal, Volume 69, Issue 2, 2018,, Pages 120-135, ISSN 0846-5371, https://doi.org/10.1016/j.carj.2018.02.002.

[109] Park, SH & Han K. (2018). Technology for Medical Diagnosis and Prediction. Radiology, Vol 283 No. 3

[110] Heumann, S. and Zahn, N. Benchmarking National AI Strategies. Available at https://www.stiftung-nv.de/sites/default/files/benchmarking_ai_strategies.pdf. Accessed May 4, 2021

[111] WHO | World Health Organization n.d. https://www.who.int/diagnostic_imaging/imaging_modalities/dim_plain-radiography/en/ (accessed August 1, 2020).

[112] X-rays | Radiology Reference Article | Radiopaedia.org n.d. https://radiopaedia.org/articles/x-rays-1?lang=us (accessed August 1, 2020).

[113] Conventional Radiography - Special Subjects - Merck Manuals Professional Edition n.d. https://www.merckmanuals.com/professional/special-subjects/principles-of-radiologic-imaging/conventional-radiography (accessed August 1, 2020).

[114] Arsalan M, Owais M, Mahmood T, Choi J, Park KR. Artificial Intelligence-Based Diagnosis of Cardiac and Related Diseases. Journal of Clinical Medicine 2020;9:871. https://doi.org/10.3390/jcm9030871.

[115] Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. JAMA Network Open 2019;2:e191095. https://doi.org/10.1001/jamanetworkopen.2019.1095.

[116] O S, M S, UJ M, DU J. An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare. Journal of Healthcare Engineering 2019;2019. https://doi.org/10.1155/2019/4180949.

Annex A:
Glossary

This section lists all the relevant abbreviations, acronyms and uncommon terms used in the document.

Acronym/Term	Expansion	Comment
AI	Artificial intelligence	
AI4H	Artificial intelligence for health	
AI-MD	AI based medical device	
API	Application programming interface	
CfTGP	Call for topic group participation	
DEL	Deliverable	
FDA	Food and Drug administration	
FGAI4H	Focus Group on AI for Health	
GDP	Gross domestic product	
GDPR	General Data Protection Regulation	
IMDRF	International Medical Device Regulators Forum	
IP	Intellectual property	
ISO	International Standardization Organization	
ITU	International Telecommunication Union	
LMIC	Low-and middle-income countries	
MDR	Medical Device Regulation	
PII	Personal identifiable information	
SaMD	Software as a medical device	
TDD	Topic Description Document	Document specifying the standardized benchmarking for a topic on which the FG AI4H Topic Group works. This document is the TDD for the Topic Group MCH
TG	Topic Group	
WG	Working Group	
WHO	World Health Organization	

Annex B:
Declaration of conflict of interest

No declarations made by the contributors to this document