Contents

A Broader Impact   **24**

B Compute and Environment Configuration   **24**

C Platform Repository   **24**

D Author Statement   **24**

E Datasheet for Datasets   **25**

    E.1 Motivation 25

    E.2 Composition 25

    E.3 Collection process 26

    E.4 Preprocessing/cleaning/labeling 27

    E.5 Uses 27

    E.6 Distribution 27

    E.7 Maintenance 28

F Training Task – Lung Opacity Detection   **29**

    F.1 Task Description 29

    F.2 License and Ethics 29

    F.3 Access and Preprocessing 29

    F.4 Data Samples 29

G Training Task – COVID-19 Detection   **30**

    G.1 Task Description 30

    G.2 License and Ethics 30

    G.3 Access and Preprocessing 30

    G.4 Data Samples 30

H Training Task – ECG Abnormal Detection   **31**

    H.1 Task Description 31

    H.2 License and Ethics 31

    H.3 Access and Preprocessing 31

    H.4 Data Samples 31

I Training Task – Mortality Prediction   **32**

    I.1 Task Description 32

    I.2 License and Ethics 32

    I.3 Access and Preprocessing 32

22

947   I.4   Data Samples   32

948 **J   Validation Task – Enlarged Cardiomendiastinum Detection**   **34**

949   J.1   Task Description   34

950   J.2   Access and Preprocessing   34

951   J.3   Data Samples   34

952 **K   Validation Task – Sepsis Prediction**   **35**

953   K.1   Task Description   35

954   K.2   License and Ethics   35

955   K.3   Access and Preprocessing   35

956   K.4   Data Samples   35

957 **L   Validation Task – MedVQA**   **36**

958   L.1   Task Description   36

959   L.2   License and Ethics   36

960   L.3   Access and Preprocessing   36

961   L.4   Data Samples   36

962 **M   Validation Task – Signal Noise Clarification**   **37**

963   M.1   Task Description   37

964   M.2   License and Ethics   37

965   M.3   Access and Preprocessing   37

966   M.4   Data Samples   37

967 **N   Implementation Details**   **38**

968   N.1   Model Details   38

969   N.2   Optimizer Hyperparameters   38

970   N.3   Task Prompts   38

971   N.4   Baselines   38

23

A Broader Impact

The Federated Medical Knowledge Injection (FMKI) platform introduces a transformative approach in healthcare AI, addressing critical issues of data privacy and accessibility by leveraging federated learning to inject medical knowledge into foundation models. This method not only complies with stringent health regulations, thereby protecting patient confidentiality, but also enhances the scalability and adaptability of medical foundation models. By enabling these models to utilize diverse, multi-modal medical data without direct data sharing, FMKI significantly broadens the potential applications of AI in healthcare, offering improved diagnostic accuracy and personalized treatment options. Furthermore, the platform facilitates equitable technology access, allowing institutions with varying resources to participate in and benefit from cutting-edge medical AI developments. This innovative approach not only promises to improve global healthcare outcomes but also sets new benchmarks in the ethical development and deployment of AI technologies in sensitive sectors.

B Compute and Environment Configuration

All experiments are conducted on an NVIDIA A100 with CUDA version 12.0, running on a Ubuntu 20.04.6 LTS server. More details can be found in the GitHub repository.

C Platform Repository

We have established a GitHub repository, available at `https://github.com/psudslab/FEDMEKI`. This repository includes resources for data processing, baselines, environmental setup, our proposed platform, and sample execution scripts. All the details have been documented at the ReadMe file[3]. We are committed to continuously updating this repository with additional modalities, datasets, and tasks.

D Author Statement

As authors of this repository and article, we bear all responsibility in case of violation of rights and licenses. We have added a disclaimer on the repository to invite original dataset creators to open issues regarding any license-related matters.

---

[3]`https://github.com/psudslab/FEDMEKI/blob/main/README.md`

24

E Datasheet for Datasets

E.1 Motivation

- **For what purpose was the dataset created?**
  This work investigates a novel yet practical task – scaling existing medical foundation models by injecting diverse medical knowledge with distributed private medical data. However, no available datasets are suitable for evaluation. Thus, we curated a new multi-site, multi-modal, and multi-task dataset, including five training tasks and three validation tasks and covering six commonly used medical modalities.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
  The authors of this paper.

- **Who funded the creation of the dataset? If there is an associated grant, please provide the name of the granter and the grant name and number.**
  This work is partially supported by the National Science Foundation under Grant No. 2238275, 2333790, 2348541, and the National Institutes of Health under Grant No. R01AG077016.

E.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**
  The FEDMEKI data suit contains medical images and the corresponding annotations for the pneumonia detection and COVID-19 detection tasks; ECG signals and the labels for the ECG abnormal detection task; 48 clinical features for the mortality prediction and sepsis prediction tasks; medical text (questions, candidate answers, document collections, ground truths) for the MedQA task; ECG signals, questions and answers for the signal noise clarification task; and medical images, questions, and answers for the MedVQA task.

- **How many instances are there in total (of each type, if appropriate)?**
  The Lung Opacity Detection task has 18,406 samples, the ECG Abnormal Detection task has 21,797 samples, and the Mortality Prediction task has 38,129 samples. The COVID-19 Detection task has 13,808 samples. The MedVQA, Signal Noise Clarification and Sepsis Prediction tasks each contain 1,000 samples. Additionally, the Enlarged Cardiomediastinum Detection task has 234 samples. Detailed information about the data can be found in Table 1.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
  The ECG Abnormal Detection task includes all available samples from its corresponding database. The Lung Opacity Prediction, COVID-19 Detection, and Mortality Prediction tasks encompass all data samples with available binary labels, making them subsets of the original dataset. For validation tasks without a predefined test set or with an excessively large test set, we randomly selected 1,000 samples for testing. These tasks include MedVQA, Signal Noise Clarification, and Sepsis Prediction. For the Enlarged Cardiomediastinum Detection task, the original database provided a small test set of 234 samples, which we have retained.

- **What data does each instance consist of?**
  The Lung Opacity Prediction, COVID-19 Detection, and Enlarged Cardiomediastinum Detection tasks involve radiological images, while the ECG Abnormal Detection task involves 12-channel, 10-second ECG signals. The Mortality Prediction and Sepsis Prediction tasks cover temporal features involving vital signs, lab events, and input/output data. The Signal Noise Clarification task includes signal-text pairs, while the MedVQA task comprises image-text pairs.

- **Is there a label or target associated with each instance?**
  The answer (label) is provided for each instance.

25

- **Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.**
  No.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**
  No.

- **Are there any errors, sources of noise, or redundancies in the dataset?**
  Questions are created by filling the slots in the templates with pre-defined values and records from the database. Thus, some questions can be grammatically incorrect but not critical (e.g., verb tense).

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**
  The proposed dataset depends on several open-source databases: RSNA [17], COVQU [18], PTB-XL [19], MIMIC-III [23], CheXpert [21], VQA-RAD [11], and ECG-QA [22].

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**
  No.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
  No.

- **Does the dataset relate to people?**
  Yes.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?**
  No.

- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**
  No. The source datasets are already de-identified.

E.3 Collection process

- **How was the data associated with each instance acquired?**
  We directly used the original data instance to curate our own dataset.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?**
  We mainly used Python scripts to collect, process and label the data.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
  The random sampling involved in this study relies on specific seed (42), thus becomes deterministic.

- **Who was involved in the data collection process (e.g., students, crowd workers, contractors), and how were they compensated (e.g., how much were crowd workers paid)?**
  The data collection process was fully performed by the study's authors.

- **Over what timeframe was the data collected?**
  N/A

26

- **Were any ethical review processes conducted (e.g., by an institutional review board)?**
  N/A.

- **Does the dataset relate to people?**
  Yes.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
  All data are collected through open-source database without interaction with individuals.

- **Were the individuals in question notified about the data collection?**
  N/A.

- **Did the individuals in question consent to the collection and use of their data?**
  N/A.

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
  N/A.

- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**
  The dataset does not have individual-specific information.

E.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
  Yes. The preprocessing on MIMIC-III data follows existing work [20].

- **Was the "raw" data saved in addition to the preprocess/cleaned/labeled data (e.g., to support unanticipated future uses)?**
  N/A.

- **Is the software that was used to preprocess/clean/label the data available?**
  Preprocessing, cleaning, and labeling are done via Python.

E.5 Uses

- **Has the dataset been used for any tasks already?**
  No.

- **Is there a repository that links to any or all papers or systems that use the dataset?**
  No.

- **What (other) tasks could the dataset be used for?**
  While the dataset is curated for research on federated medical knowledge injection problem, other studies concerning developing centralized medical foundation model can also leverage the dataset.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
  N/A.

- **Are there tasks for which the dataset should not be used?**
  N/A.

E.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**
  No.

27

- **How will the dataset be distributed?**
  The preprocessing code is available at `https://github.com/psudslab/`
  `FEDMEKI/tree/main/data_preprocess`. Users can download corresponding
  dataset and utilize the preprocessing scripts for generating the final dataset used in this study.

- **Will the dataset be distributed under a copyright or other intellectual property (IP)
  license, and/or under applicable terms of use (ToU)?**
  The dataset is released under MIT License.

- **Have any third parties imposed IP-based or other restrictions on the data associated
  with the instances?**
  No.

- **Do any export controls or other regulatory restrictions apply to the dataset or to
  individual instances?**
  No.

E.7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**
  The authors of this paper.

- **How can the owner/curator/manager of the dataset be contacted(e.g., email address)?**
  Contact the first authors (jqwang@psu.edu and xcwang@psu.edu).

- **Is there an erratum?**
  No.

- **Will the dataset be updated (e.g., to correct labeling erros, add new instances, delete
  instances)?**
  If any corrections are required, our plan is to upload an updated version of the dataset with
  comprehensive explanations for the changes. Furthermore, as we broaden our QA scope, we
  will consistently update the dataset with new QA templates/instances.

- **If the dataset relates to people, are there applicable limits on the retention of the data
  associated with the instances (e.g., were the individuals in question told that their data
  would be retained for a fixed period of time and then deleted)?**
  N/A

- **Will older versions of the dataset continue to be supported/hosted/maintained?**
  Primarily, we plan to maintain only the most recent version of the dataset. However, under
  certain circumstances, such as significant updates to our dataset or the need for validation
  of previous research work using older versions, we will exceptionally preserve previous
  versions of the dataset for up to one year.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mecha-
  nism for them to do so?**
  Contact the authors of this paper.

28

F Training Task – Lung Opacity Detection

F.1 Task Description

In the United States, pneumonia keeps the ailment on the list of top 10 causes of death in the country. The task is to locate lung opacities on chest radiographs. In this challenge [17], 18,406 images are annotated as either Lung Opacity or Normal, providing a basis for extracting the binary classification task. The task is to develop an algorithm to detect visual indicators of pneumonia in medical images. Specifically, the algorithm needs to identify and localize lung opacities in chest radiographs.

F.2 License and Ethics

This dataset is permitted to access and utilize these de-identified imaging datasets and annotations for academic research, educational purposes, or other commercial or non-commercial uses, provided you adhere to the appropriate citations.

F.3 Access and Preprocessing

The resource is available to access via the official website at `https://www.rsna.org/rsnai/ai-image-challenge/rsna-pneumonia-detection-challenge-2018` and Kaggle at `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/overview`. It includes dataset description, annotations, and mapping from RSNA image dataset to original NIH dataset. The data is organized as a set of patient IDs with corresponding image class annotations, including "No Lung Opacity/Not Normal," "Normal," and "Lung Opacity." We collected images labeled as either "Normal" or "Lung Opacity" and formulated the problem as a binary classification task. The code for preprocessing is available at `https://github.com/psudslab/FEDMEKI/tree/main/data_preprocess`.

F.4 Data Samples

We provide a random data sample from the dataset and visualize it in Figure 2.

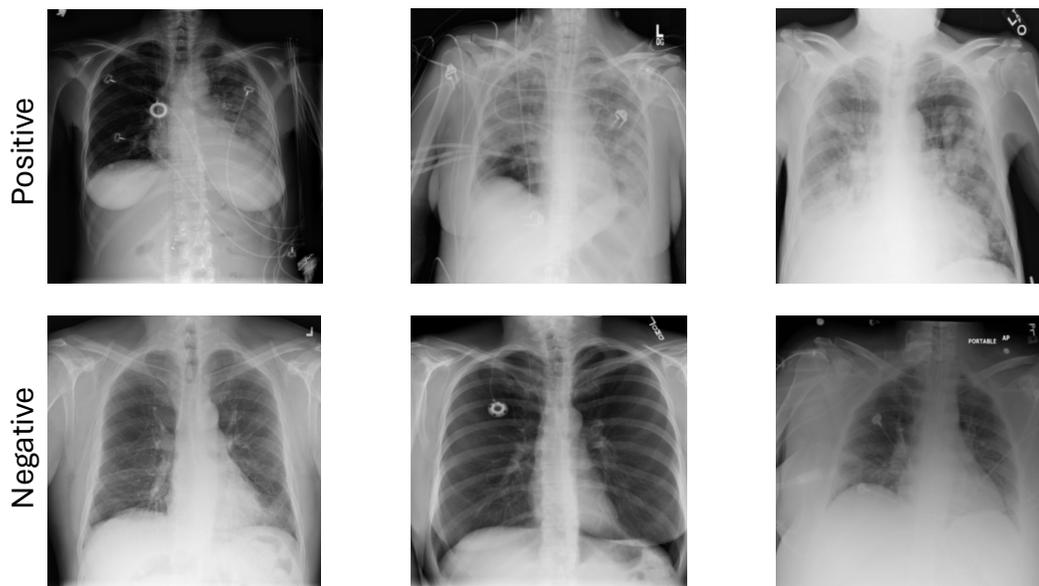



Figure 2: Data sample of lung opacity detection.

29

G   Training Task – COVID-19 Detection

G.1   Task Description

This task challenges the model to assess whether X-ray images display symptoms of Covid-19, thereby evaluating the model's proficiency in interpreting medical imagery. For this purpose, we employ the COVQU dataset [18].

G.2   License and Ethics

The licensing and ethical compliance adhere to the regulations established by the original datasets.

G.3   Access and Preprocessing

This dataset can be accessed via the link at `https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database`. This dataset, featuring COVID-19, normal, and other lung infection categories, is being released incrementally. The initial release comprised 219 COVID-19, 1,341 normal, and 1,345 viral pneumonia chest X-ray (CXR) images. The first update expanded the COVID-19 category to include 1,200 CXR images. In the second update, the collection was further enlarged to include 3,616 COVID-19 positive cases, along with 10,192 normal, 6,012 lung opacity (non-COVID lung infection), and 1,345 viral pneumonia images, complete with corresponding lung masks. We selected normal and COVID-19 positive images to formulate this task as a binary classification problem.

G.4   Data Samples

We provide a random data sample from the dataset and visualize it in Figure 3.

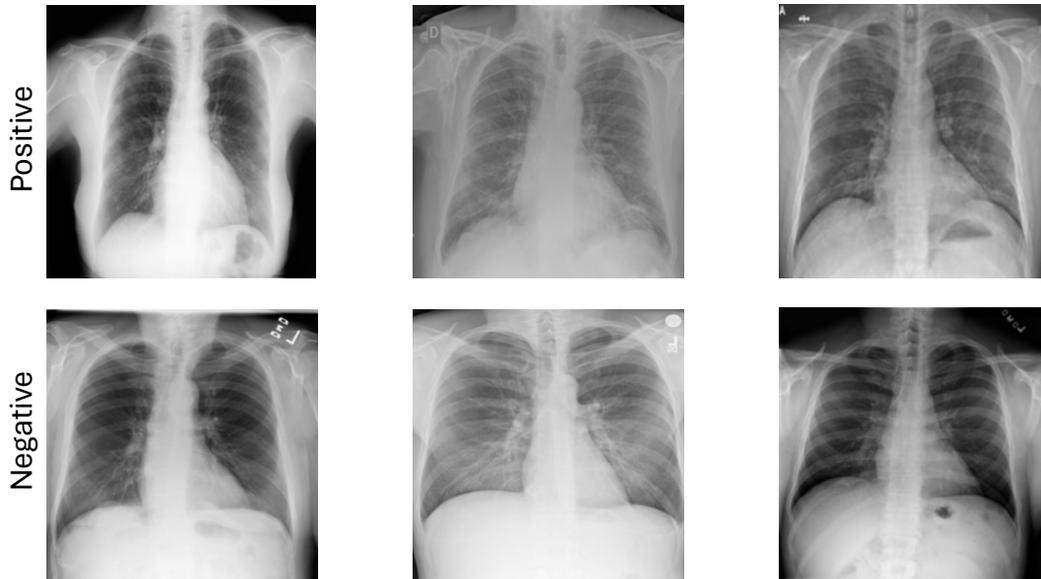



Figure 3: Data sample of Covid-19 detection.

H Training Task – ECG Abnormal Detection

H.1 Task Description

Electrocardiography (ECG) is a crucial diagnostic tool for assessing a patient's cardiac condition, and automatic ECG interpretation algorithms offer significant support to medical personnel given the volume of ECGs routinely performed. The task involves analyzing the PTB-XL ECG dataset to develop and evaluate automatic ECG interpretation algorithms. We offer training and test set splits to facilitate algorithm comparability and include extensive metadata on demographics, infarction characteristics, diagnostic likelihoods, and signal properties, making it a comprehensive resource for training and evaluating automatic ECG interpretation algorithms.

H.2 License and Ethics

The Institutional Ethics Committee approved the publication of the anonymous data in an open-access database (PTB-2020-1).

H.3 Access and Preprocessing

The dataset can be directly downloaded with granted permission at `https://physionet.org/content/ptb-xl/1.0.3/` or via the terminal by `wget -r -N -c -np https://physionet.org/files/ptb-xl/1.0.3/`. Raw signal data was recorded in a proprietary compressed format, encompassing the standard set of 12 leads (I, II, III, AVL, AVR, AVF, V1, ..., V6) with reference electrodes on the right arm. Corresponding metadata, including age, sex, weight, and height, was systematically gathered in a database. Each ECG record includes a report, either generated by a cardiologist or automatically by the ECG device, which was then translated into a standardized set of SCP-ECG statements (scp_codes). For the relevant metadata, it is saved as one row per record identified by ecg_id. Totally, there are 28 columns categorized into identifiers, general metadata, ECG statements, signal metadata, and cross-validation folds. Additional details such as the heart's axis and stages of infarction (if applicable) were also documented. To ensure privacy and compliance with HIPAA standards, all personal information, including names of cardiologists and nurses, recording locations, and patient ages (with ages over 89 years reported within a 300-year range), was pseudonymized.

H.4 Data Samples

We provide a random data sample from the dataset and visualize it in Figure 4. Here, a positive result indicates the presence of an abnormality in the ECG signal, while a negative result represents a normal signal. All signals are 12-channel, derived from the standard set of 12 leads (I, II, III, aVL, aVR, aVF, V1, ..., V6) with reference electrodes on the right arm.

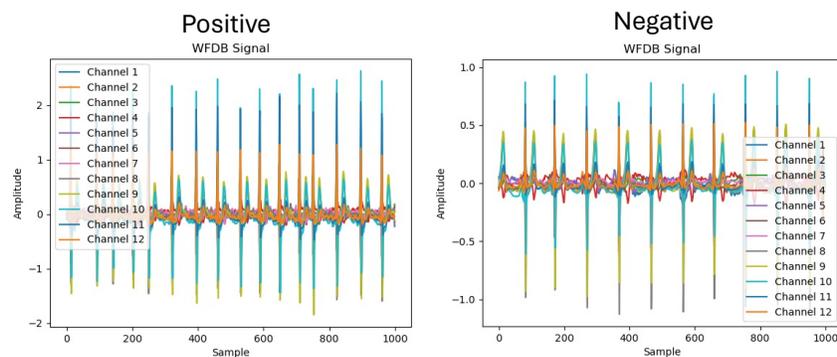



Figure 4: Data sample of ECG abnormal detection.

I Training Task – Mortality Prediction

I.1 Task Description

The data source is MIMIC-III(Medical Information Mart for Intensive Care III), which is a substantial, anonymous, and publicly accessible repository of medical records. Each entry in the dataset contains ICD-9 codes that categorize the diagnoses and procedures conducted. In our work, we use the processed dataset to conduct the mortality prediction task.

I.2 License and Ethics

The dataset is available for non-profit use in accordance with the license at `https://www.physionet.org/content/mimiciii/view-license/1.4/`.

I.3 Access and Preprocessing

MIMIC-III can be accessed as a credentialed user on PhysioNet with an approved application at `https://mimic.mit.edu/`. In our experiment, we follow the ICU-oriented preprocessing pipeline [72] to process the data and follow the feature extraction pipeline [20] to extract dynamic features. Features extracted from this MIMIC-III database are listed in Table 5.

I.4 Data Samples

We provide a random data sample from the dataset and visualize it in Figure 5.

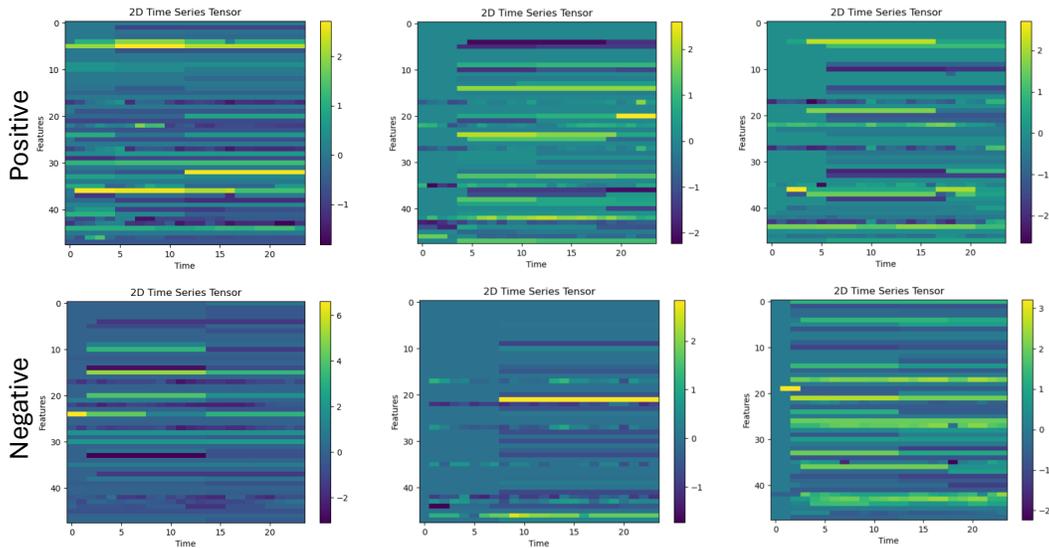



Figure 5: Data sample of mortality prediction.

Table 5: Clinical concepts extracted from MIMIC-III database [23]. The information is based on Table 15 provided in [20].

Feature	RICU	Unit
Blood pressure (systolic)	sbp	mmHg
Blood pressure (diastolic)	dbp	mmHg
Heart rate	hr	beats/minute
Mean arterial pressure	map	mmHg
Oxygen saturation	o2sat	%
Respiratory rate	resp	breaths/minute
Temperature	temp	∘C
Albumin	alb	g/dL
Alkaline phosphatase	alp	IU/L
Alanine aminotransferase	alt	IU/L
Aspartate aminotransferase	ast	IU/L
Base excess	be	mmol/L
Bicarbonate	bicar	mmol/L
Bilirubin (total)	bili	mg/dL
Bilirubin (direct)	bili_dir	mg/dL
Band form neutrophils	bnd	%
Blood urea nitrogen	bun	mg/dL
Calcium	ca	mg/dL
Calcium ionized	cai	mmol/L
Creatinine	crea	mg/dL
Creatinine kinase	ck	IU/L
Creatinine kinase MB	ckmb	ng/mL
Chloride	cl	mmol/L
CO2 partial pressure	pco2	mmHg
C-reactive protein	crp	mg/L
Fibrinogen	fgn	mg/dL
Glucose	glu	mg/dL
Haemoglobin	hgb	g/dL
International normalised ratio (INR)	inr_pt	-
Lactate	lact	mmol/L
Lymphocytes	lymph	%
Mean cell haemoglobin	mch	pg
Mean corpuscular haemoglobin concentration	mchc	%
Mean corpuscular volume	mcv	fL
Methaemoglobin	methb	%
Magnesium	mg	mg/dL
Neutrophils	neut	%
O2 partial pressure	po2	mmHg
Partial thromboplastin time	ptt	sec
pH of blood	ph	-
Phosphate	phos	mg/dL
Platelets	plt	1,000 / μL
Potassium	k	mmol/L
Sodium	na	mmol/L
Troponin T	tnt	ng/mL
White blood cells	wbc	1,000 / μL
Fraction of inspired oxygen	fio2	%
Urine output	urine	mL

33

1267 # J Validation Task – Enlarged Cardiomendiastinum Detection

1268 ## J.1 Task Description

This task is designed to evaluate the probability of an enlarged cardiomediastinum by using medical images from clinical assessments. It serves to gauge the model's ability to interpret radiographs effectively. The data for this task are sourced from the CheXpert Dataset [21]. CheXpert is a collection of 224,316 chest radiographs from 65,240 patients who underwent radiographic examinations at Stanford Health Care from October 2002 to July 2017. These images were gathered from both inpatient and outpatient centers and include the associated radiology reports.

J.2 Access and Preprocessing

This dataset can be accessed via the link at `https://aimi.stanford.edu/chexpert-chest-x-rays` and downloaded via the link at `https://stanfordaimi.azurewebsites.net/datasets/8cbd9ed4-2eb9-4565-affc-111cf4f7ebe2`. The training set comprises 224,316 high-quality X-ray images from 65,240 patients, annotated to automatically identify 14 different observations from radiology reports, reflecting the inherent uncertainties of radiographic interpretation. The validation set includes 234 images from 200 patients, each manually annotated by three board-certified radiologists. The test set, which remains unreleased to the public and is held by the organizers for final assessment, contains images from 500 patients annotated through the consensus of five board-certified radiologists. CheXpert images have an average resolution of 2828x2320 pixels.

J.3 Data Samples

We provide a random data sample from the dataset and visualize it in Figure 6.

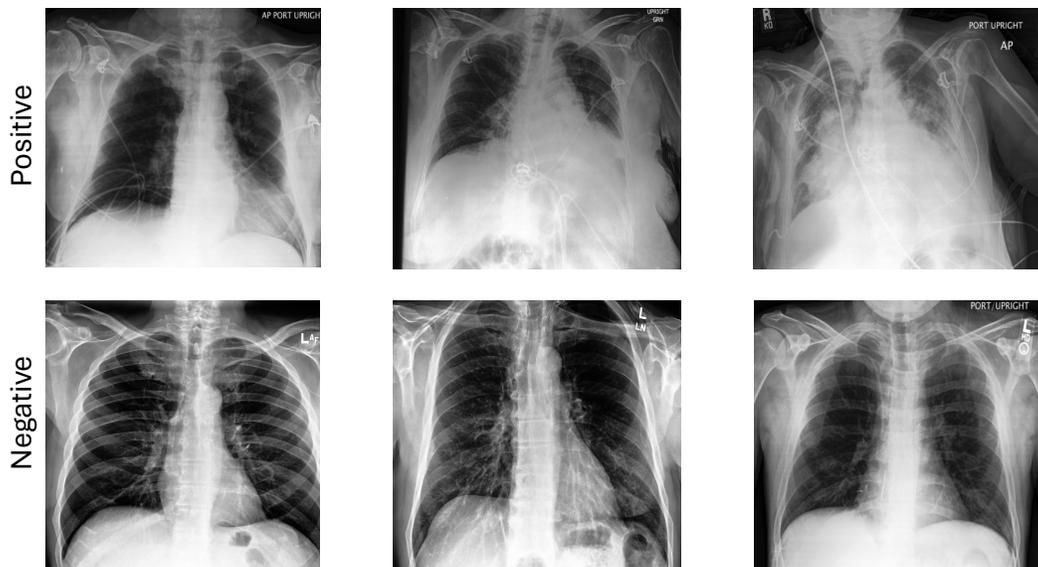



Figure 6: Data sample of enlarged cardiomendiastinum detection.

34

K Validation Task – Sepsis Prediction

K.1 Task Description

This task focuses on predicting the likelihood of sepsis during ICU stays, assessing the model's ability to analyze various clinical data, including lab events, diagnoses, and prescriptions. For this research, we utilize the MIMIC-III database, extracting features and cohorts through a well-established preprocessing pipeline [20].

K.2 License and Ethics

This dataset is governed by the license available at the following URL: `https://www.physionet.org/content/mimiciii/view-license/1.4/`.

K.3 Access and Preprocessing

MIMIC-III dataset can be accessed with the approved permission via `https://mimic.mit.edu/`. A random sampling strategy is applied to select a subset with 1,000 samples for testing.

K.4 Data Samples

We provide a random data sample from the dataset and visualize it in Figure 7. It has clinical features only, including lab events and vital signs. Although the data feature space of both mortality prediction and sepsis prediction is the same, the feature distributions are significantly different. That is why the zero-shot inference on this task performs worse than the mortality prediction, as shown in Table 4.

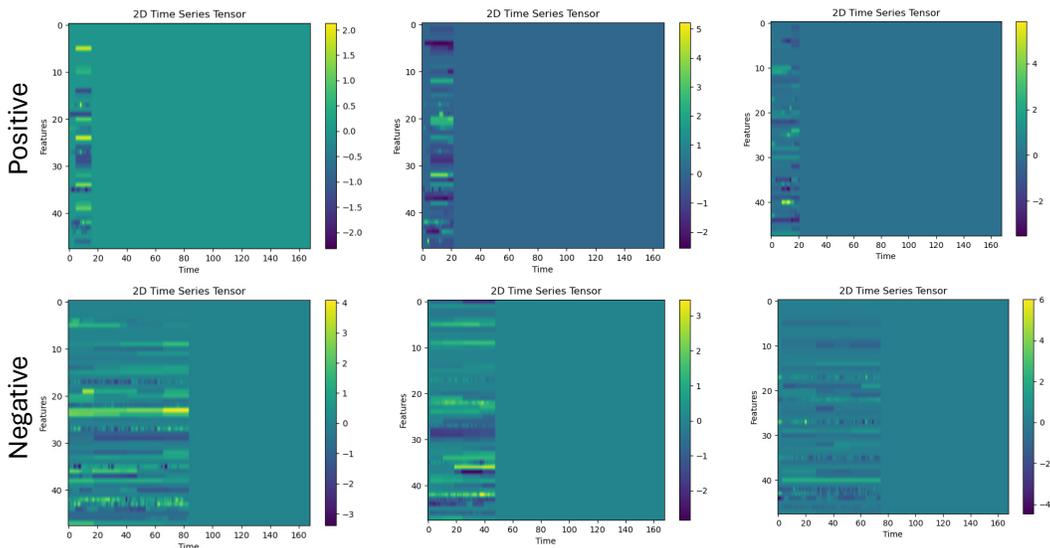



Figure 7: Data sample of sepsis prediction.

L   Validation Task – MedVQA

L.1   Task Description

The SLAKE dataset is designed for Medical Visual Question Answering (Med-VQA), integrating detailed visual and textual annotations with a medical knowledge base. It features semantic segmentation masks and object detection bounding boxes for each radiology image. SLAKE includes both basic clinical and complex compositional questions, and is uniquely bilingual in English and Chinese. It expands coverage to more body parts and introduces new question types related to shape and knowledge graphs, with comparative data provided against the VQA-RAD dataset. In this task, the model uses both visual context and verbal questions as inputs, requiring answers that integrate textual questions and visual context. This task tests the model's ability to align text and image modalities in the medical domain.

L.2   License and Ethics

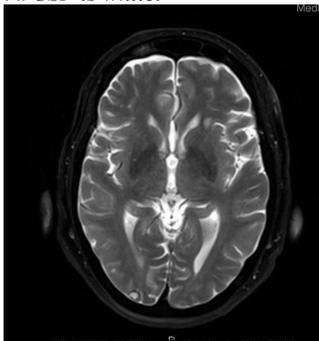
Ethical approval was not required as confirmed by the license attached with the open access data in [12].

L.3   Access and Preprocessing

This dataset can be accessed at `https://www.med-vqa.com/slake/`. For the image part, 642 images, including 12 diseases and 39 organs, were in the format of CTs and MRIs. With the help of a constructed knowledge graph, it covers questions with ten different content types and semantic labels proposed by doctors. We randomly select 1000 samples from the test dataset.

L.4   Data Samples

We provide a random data sample from the dataset. Question-answering pair and corresponding image (Figure 9) are listed below. This MedVQA dataset contains different types of images except for chest X-ray images, which are different from the ones we used in the model training. In addition, the trained FEDMEKI does not use any medical question-answering training tasks. Therefore, its performance of this "new" task is limited, as shown in Table 4.

Q: *How can you tell this is a T2 weighted image?*
A: *CSF is white.*

Q: *Is the heart enlarged?*
A: *No.*

Q: *What it causing the widening?*
A: *Mass*

Figure 8: Data sample of MedVQA task.

36

M Validation Task – Signal Noise Clarification

M.1 Task Description

This task is dedicated to precisely characterizing noise in ECG signals through a question-and-answer format.

M.2 License and Ethics

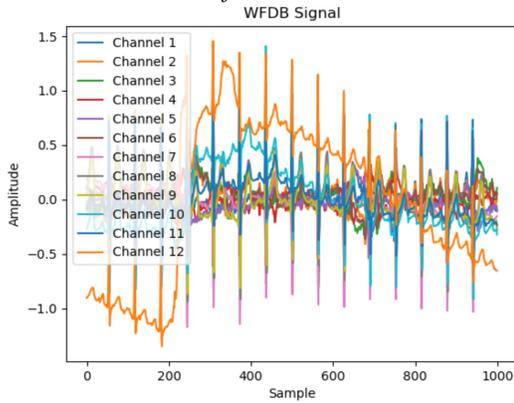
The Institutional Ethics Committee approved the publication of the anonymous data in an open-access database (PTB-2020-1).

M.3 Access and Preprocessing

It utilizes data from an established ECG question answering dataset [22] and a related ECG database [19], which can be accessed via `https://physionet.org/content/ptb-xl/1.0.3/` and `https://github.com/Jwoo5/ecg-qa/tree/master/ecgqa/ptbxl`. The ECG signals used in this task consist of 12 channels and have a duration of 10 seconds, mirroring the parameters used in the ECG Abnormal Detection task. We randomly sample 1,000 ECG-question pairs as the validation data.

M.4 Data Samples

We provide a random data sample from the dataset. Question-answering pair and corresponding signal (Figure 9) are listed below. Although we have a training task on ECG, the ECG abnormal detection task is different from this one. This task aims to answer the noise types of ECG signals according to the input ECG and the question. We can see that the ECG signals in Figure 9 are quite different from the ones in Figure 4, which increases the difficulty of this task significantly.

Q: *What kind of noises does this ECG show in lead aVL?*
A: *baseline drift, static noise.*

Q: *Does this ECG show baseline drift?*
A: *Yes.*

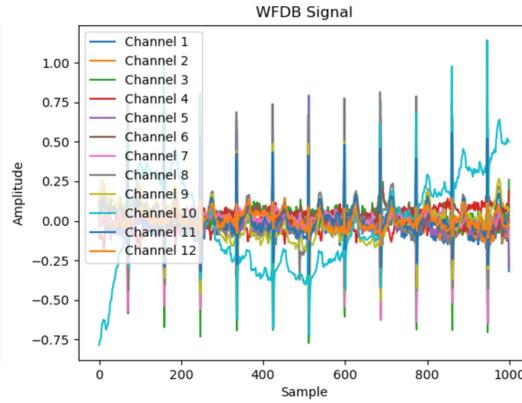



Figure 9: Data sample of signal noise clarification.

N Implementation Details

N.1 Model Details

For each local model $\mathbf{W}_n$ deployed in client C_n, we implement their modality-specific encoders and task-specific decoders. Details about encoders for different modalities can be found in Table N. As all training tasks can be categorized as binary classification, we use MLPs as their task-specific decoders, where decoders for different tasks do not share any parameter.

Table 6: Details of modality-specific encoders.

Modality	Encoder	# of Parameters
Image	Deit-tiny [73]	7.8M
Signal	CNN [74]	4.1M
Diagnosis Code	Transformer [75]	3.7M
Procedure Code	Transformer [75]	3.7M
Drug	Transformer [75]	3.7M
Lab Results	Transformer [75]	3.7M

N.2 Optimizer Hyperparameters

We leverage Adam optimizer [76] for optimizing both local model $\mathbf{W}_n$ and foundation model $\mathcal{F}$. The number of communication rounds is set to 10. For local model $\mathbf{W}_n$, we find the learning rate of 1e-4 for local models achieves a decent convergence, while the learning rate for the foundation model is configured to 5e-4. The batch size of both the foundation model and local models is set to 64.

N.3 Task Prompts

Task prompts for classification tasks are listed in Table 7. For MedVQA and signal noise clarification, prompts are questions themselves.

Table 7: Task Prompts.

Task Name	Prompt
Lung Opacity Detection	Assess this CT image: should it be classified as lung opacity?
Covid-19 Detection	Based on this image, is the patient COVID-19 positive?
ECG Abnormal Detection	Is the given ECG abnormal?
Mortality Prediction	Based on these clinical features, will mortality occur in this patient?
Enlarged Cardiomendiastinum Detection	Does this image show evidence of enlarged cardiomediastinum?
Sepsis Prediction	Based on these clinical features, will sepsis occur in this patient?

N.4 Baselines

To better understand the benchmarks used in the experiments, we use visualizations to demonstrate each approach clearly.

For single-task evaluation, we use FedAvg$_s$/FedProx$_s$ (Figure 10), FedAvg$_s^+$/FedProx$_s^+$ (Figure 11), FedAvg$_s^*$/FedProx$_s^*$, and FedAvg$_s^{\mathcal{F}}$/FedProx$_s^{\mathcal{F}}$ (Figure 12). When training single tasks, we only use each task data as the model input. For multi-task training, we also have eight baselines that are shown from Figure 13 to Figure 15. These models will train all the task data together.

38

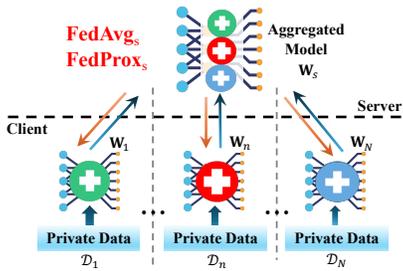



Figure 10: **FedAvg**$_s$ or **FedProx**$_s$

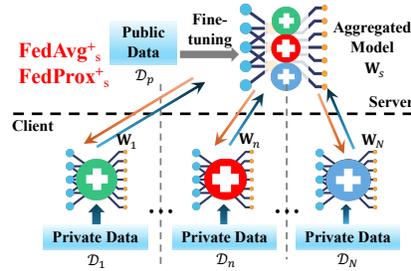



Figure 11: **FedAvg**$_s^+$ and **FedProx**$_s^+$

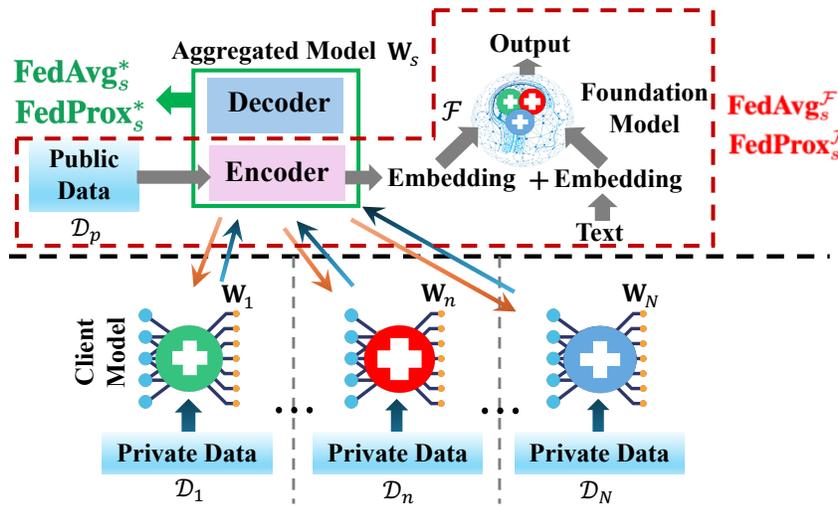



Figure 12: **FedAvg**$_s^{\mathcal{F}}$/**FedProx**$_s^{\mathcal{F}}$ (red dot line) and **FedAvg**$_s^*$/**FedProx**$_s^*$ (green line).

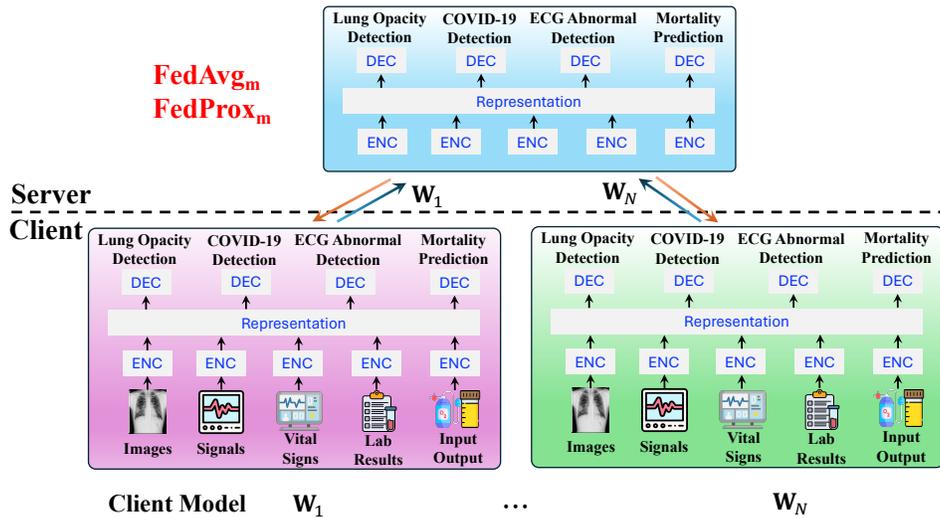



Figure 13: **FedAvg**$_m$ or **FedProx**$_m$

39

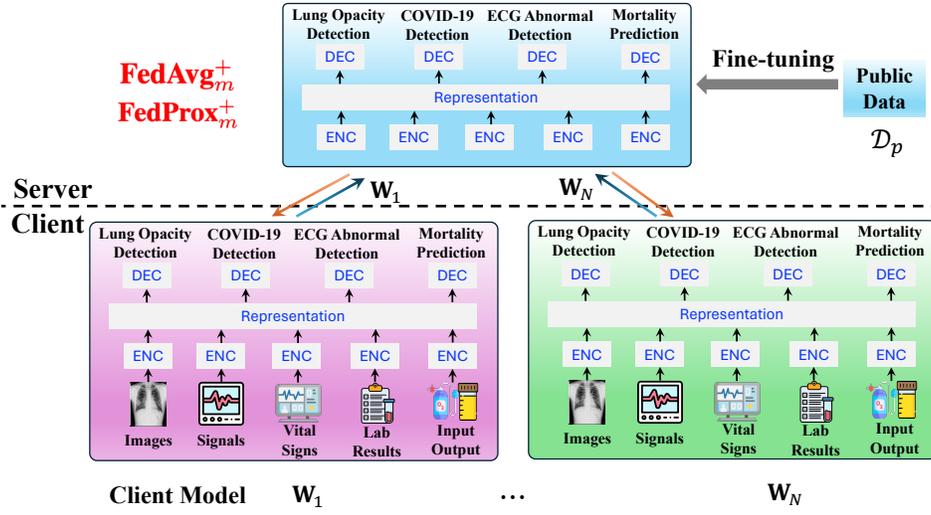



Figure 14: $\mathbf{FedAvg}_m^+$ and $\mathbf{FedProx}_m^+$

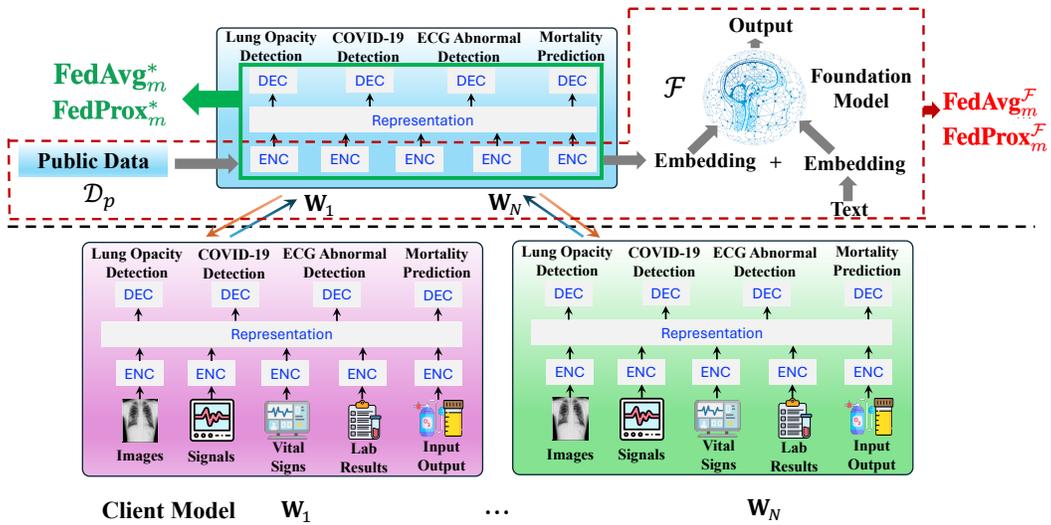



Figure 15: $\mathbf{FedAvg}_m^{\mathcal{F}}$/$\mathbf{FedProx}_m^{\mathcal{F}}$ (red dot line) and $\mathbf{FedAvg}_m^*$/$\mathbf{FedProx}_m^*$ (green line).

40