Integrated multimodal artificial intelligence framework for healthcare applications

Luis R. Soenksen[1,4*], Yu Ma[2*], Cynthia Zeng[2*], Leonard D.J. Boussioux[2*], Kimberly Villalobos Carballo[2*], Liangyuan Na[2*], Holly M. Wiberg[2], Michael L. Li[2], Ignacio Fuentes[1], Dimitris Bertsimas[1,2,3] ‡

[1]Abdul Latif Jameel Clinic for Machine Learning in Health, MIT, Cambridge, MA 02139, USA. [2]Operations Research Center, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [3]Sloan School of Management, MIT, Cambridge, MA 02139, USA. [4]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA.
* These authors contributed equally to this work
‡ Corresponding author. Email: dbertsim@mit.edu

ABSTRACT:
Artificial intelligence (AI) systems hold great promise to improve healthcare over the next decades. Specifically, AI systems leveraging multiple data sources and input modalities are poised to become a viable method to deliver more accurate results and deployable pipelines across a wide range of applications. In this work, we propose and evaluate a unified Holistic AI in Medicine (HAIM) framework to facilitate the generation and testing of AI systems that leverage multimodal inputs. Our approach uses generalizable data pre-processing and machine learning modeling stages that can be readily adapted for research and deployment in healthcare environments. We evaluate our HAIM framework by training and characterizing 14,324 independent models based on MIMIC-IV-MM, a multimodal clinical database (N=34,537 samples) containing 7,279 unique hospitalizations and 6,485 patients, spanning all possible input combinations of 4 data modalities (i.e., tabular, time-series, text and images), 11 unique data sources and 12 predictive tasks. We show that this framework can consistently and robustly produce models that outperform similar single-source approaches across various healthcare demonstrations (by 6-33%), including 10 distinct chest pathology diagnoses, along with length-of-stay and 48-hour mortality predictions. We also quantify the contribution of each modality and data source using Shapley values, which demonstrates the heterogeneity in data type importance and the necessity of multimodal inputs across different healthcare-relevant tasks. The generalizable properties and flexibility of our Holistic AI in Medicine (HAIM) framework could offer a promising pathway for future multimodal predictive systems in clinical and operational healthcare settings.

KEYWORDS: artificial intelligence, machine learning, healthcare, multimodality, pipeline, predictive analytics, operations optimization

INTRODUCTION:

Artificial intelligence (AI) and machine learning (ML) systems are poised to become fundamental tools in next-generation clinical practice and healthcare operations.[1] Such anticipated utility, particularly in AI/ML systems aimed to improve clinical efficiency and patient outcomes, will require knowledge from multiple sources of data and various input modalities.[2-4] Multimodal architectures for AI/ML systems are attractive due to their ability to

emulate the input conditions that clinicians and healthcare administrators currently use to perform predictions and respond to their complex decision-making landscape.[2,5] A typical clinical practice uses a diverse set of information formats contained within the patient electronic health record (EHR) such as tabular data (e.g., age, demographics, procedures, history, billing codes), image data (e.g., photographs, x-rays, computerized-tomography scans, magnetic resonance imaging, pathology slides), time-series data (e.g., intermittent pulse oximetry, blood chemistry, respiratory analysis, electrocardiograms, ultra-sounds, in-vitro tests, wearable sensors), structured sequence data (e.g., genomics, proteomics, metabolomics) and unstructured sequence data (e.g., notes, forms, written reports, voice recordings, video) among other sources.[6] Recently, AI/ML models leveraging multiple data modalities have been demonstrated for the domains of cardiology[7-9], dermatology[10], gastroenterology[11], gynecology[12], hematology[13], immunology[14], nephrology[15], neurology[16,17], oncology[18-20], ophthalmology[21], psychiatry[22], radiology[23-25], public health[26] and healthcare operational analytics (i.e., mortality, length-of-stay, and discharge predictions)[27-30]. Furthermore, it has been shown that multimodality in most of these domains can increase the performance of AI/ML systems (accuracy: 1.2–27.7%) compared to single-modality approaches for the same task.[2] However, developing unified and scalable pipelines that can consistently be applied to train multimodal AI/ML systems that outperform their single-modality counterparts has remained challenging[2]. This motivates our development of our Holistic Artificial Intelligence in Medicine (HAIM) framework, a generalizable data pre-processing and machine learning pipeline (Fig.1) that can receive standard EHR information with multiple input data types (i.e., tabular data, images, time-series and text). Based on this framework, we build and test classification models for a range of typical target applications in clinical care and healthcare operations to illustrate the pipeline.

METHODS:

Dataset

For this work, we utilize the Medical Information Cart for Intensive Care (MIMIC)-IV[31,32], an openly accessible database that contains de-identified records of 383,220 individual patients admitted to the intensive care unit (ICU) or emergency department (ED) of Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA, USA, between 2008 and 2019 (inclusive). MIMIC-IV's most recent version (v1.0) improves on MIMIC-III[33] to provide public access to the EHR data of over 40,000 hospitalized patients based on BIDMC's MetaVision clinical information system. We selected MIMIC-IV due to its large-scale, detailed documentation, corroborated use in AI/ML applications[34], and prior evaluations in terms of AI/ML interpretability, fairness, and bias[35]. To augment BIDMC's MIMIC-IV v1.0, we used the MIMIC Chest X-ray (CXR) database v2.0.0[36] containing 377,110 radiology images with free-text reports representing 227,835 medical imaging events that can be matched to corresponding patients included in MIMIC-IV v1.0. Both data sources have been independently de-identified by deleting all personal health information (PHI), following the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements. After getting credentialled access from PhysioNet, we combined MIMIC-IV v1.0 and MIMIC-CXR-JPG v2.0.0 into a unified multimodal dataset (MIMIC-IV-MM) based on matched patient, admission, and imaging-study identifiers *(i.e., subject_id*, *stay_id*, *study_id* from MIMIC-IV and MIMIC-CXR-JPG databases). We used MIMIC-IV-MM throughout this study to test all the presented machine learning use-cases analyzing various combinations of structured patient information, time-series data, medical images, and unstructured text notes, as presented in the following sections.

Patient-centric data representation

We generated the individual files containing patient-specific information for single hospital admissions by querying the aggregated multimodal dataset MIMIC-IV-MM. Every HAIM-EHR file contains the details of current and previous patient admissions, transfers, demographics, laboratory measurements, provider orders, microbiology cultures, medication administrations, prescriptions, procedure events, intravenous and fluid inputs, sensor outputs, measurement events, radiological images, radiological reports, electrocardiogram reports, echocardiogram reports, notes, hospital billing information (e.g., diagnosis and procedure-related codes), as well as other time-stamped and charted information. The samples therefore include all available patient data collected within a specific admission and stay with all prior information occurring before the discharge or death time stamp. We stored all the individual patient files in MIMIC-IV-MM as "pickle" python-language object structures for ease of processing in subsequent sampling and modeling tasks. The code to generate the aggregated MIMIC-IV-MM dataset from credentialled access to MIMIC-IV v1.0 and MIMIC-CXR-JPG v2.0.0 datasets is available at our GitHub repository (https://github.com/lrsoenksen/HAIM). In addition, the pre-processed pickle patient files of MIMIC-IV-MM are publicly accessible at the official PhysioNet repository (https://physionet.org/projects/MIMIC-IV-MM). A schematic of this patient-centric data representation as multimodal input for our HAIM framework is shown in Figure 1.

Patient data processing and multimodal feature extraction

We processed each HAIM-EHR patient file individually to generate fixed-dimensional vector embeddings for each of the possible input types, including all patient information from the time of admission until the selected inference event (e.g., time of imaging procedure for pathology diagnosis or end-of-day for 48-hour mortality predictions). The generated embeddings from input types include: tabular data such as demographics (E_{de}=demographics), structured time-series events (E_{ce}=chart events, E_{le}=laboratory events, E_{pe}=procedure events), unstructured free text (E_{radn}=radiological notes, E_{ecgn}=electrocardiogram notes, E_{econ}=echocardiogram notes), single-image vision (E_{vp}=visual probabilities, E_{vd}=visual dense-layer features) and multi-image vision (E_{vmp}=aggregated visual probabilities, E_{vmd}=aggregated visual dense-layer features). We then implemented fixed embedding extraction procedures based on standard data modalities (i.e., tabular data, time-series, text and images) to reduce its dependence on site-specific data architectures and allow for a consistent embedding format that may be applied to arbitrary ML pipelines.

We extracted the embeddings based on tabulated demographics data (E_{de}) solely by querying normalized numerical values from the patient record. We obtained time-series embeddings using time-stamped data from the structured patient chart, laboratory, and procedure event lists (i.e., E_c $E_{le,}$ E_{pe}, respectively). We selected a set of key clinical signals for each type of event list and constructed the corresponding time sequences from the time of patient admission to the time-stamp allowable for each individual feature (See Supplemental Table 1). The embeddings encode the signal length, maximum, minimum, mean, median, standard deviation, variance, number of peaks, and average time-series slope and piece-wise change over time of these metrics. The time-series signals for E_{ce} include: heart rate (HR), non-invasive systolic blood pressure (NBP_s), non-invasive diastolic blood pressure (NBP_d), respiratory rate (RR), oxygen

saturation by pulse oximetry (SpO_2), Glasgow coma scales (GCS) for verbal, eye, and motor response (GCS_V, GCS_E, GCS_M respectively). Moreover, time-series E_{le} include: glucose, potassium, sodium, chloride, creatinine, urea nitrogen, bicarbonate, anion gap, hemoglobin, hematocrit, magnesium, platelet count, phosphate, white blood cells (WBC), total calcium, mean corpuscular hemoglobin (MCH), red blood cells (RBC), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), red blood cell distribution width (RDW), platelet count, neutrophils, vancomycin. Lastly, time-series E_{pe} procedures include: foley catheter, peripherally inserted central catheter (PICC), intubation, peritoneal dialysis, bronchoscopy, electroencephalogram (EEG), dialysis with continuous renal replacement therapy, dialysis with catheter, removed chest tubes, and hemodialysis.

We obtained embeddings for the unstructured free text (E_{radn}, E_{ecgn}, and E_{econ}) by concatenating all available text from each of these types of notes as continuous strings and then by processing them using Clinical BERT[37], a transformer-based bidirectional encoder model pre-trained on a large corpus of biomedical and medical text. This transformer-based model generates a single 768-dimensional vector, or embedding, per unstructured text type. We split notes longer than the maximum input token size for Clinical BERT into the smallest number of processable text chunks to generate various embeddings sequentially, all of which are averaged to produce a single 768-dimensional output embedding for the entire text.

Finally, we processed vision data included in this work using a pre-trained Densenet121 convolutional neural network (CNN) fine-tuned on the X-ray CheXpert dataset[38] (i.e., Densenet121-res224-chex).[39] We selected this model because the availability of at least one time-stamped chest X-ray per patient file within the MIMIC-IV-MM database as its core visual component. Densenet121-res224-chex is part of TorchXRayVision, a unified library and repository of datasets and state-of-the-art pre-trained models for chest pathology classification using X-rays.[39] While other computer vision models pre-trained on large sets of medical imaging data may be utilized to extract embeddings within the HAIM framework, for this work, we used only the Densenet121-res224-chex as our canonical method to extract visual embeddings. We obtained the single-image embeddings per HAIM-EHR patient file by rescaling each image into 224x224 size using a standard interpolation method with resampling using pixel area relations, and then feeding it into the selected network to extract: a) output class probabilities and b) final dense-layer features. The output classes per image are the 18-dimensional diagnosis probability vector generated directly by Densenet121-res224-chex, which produces the embedding E_{vp}. The dense network features per image are the 1024-dimensional vector generated by extracting the outputs of the last dense layer of the model, which produces the embedding E_{vd}. Multi-image embeddings are also obtained by averaging the output class probabilities and dense-feature embeddings of all available images per HAIM-EHR patient file (e.g., X-ray studies with multiple planes and past X-ray studies). This produces an aggregated multi-image diagnosis probability embedding (E_{vmp}) and multi-image dense-layer embedding (E_{vmd}) per patient that considers all available X-rays and not only the most recent one.

The dimensionality of each of these embeddings is E_{de}=7, E_{ce}=99, E_{le}=242, E_{radn}=110, E_{ecgn}=768, E_{econ}=768, E_{dchn}=768, E_{vp}=18, E_{vd}=18 E_{vmp}=1024 and E_{vmd}=1024. Once all single-modality embeddings are generated, we concatenate them into a single multimodal fusion embedding per HAIM-EHR patient file, which constitutes the input for all downstream modeling

tasks in our HAIM framework. Since this deep patient representation in vector form can be made of fixed size within or across healthcare institutions (6405-dimensional for this work), it can be used as input for rapid iteration in the development of generic machine learning systems for relevant predictive analytics in healthcare.

Modeling

After we extracted all multimodal fusion embeddings for all HAIM-EHR patient files in the MIMIC-IV-MM database, we generated classification models across various clinical and operational tasks, including: a) chest pathology diagnosis, b) length-of-stay and c) 48-hour mortality predictions. For each of these modeling tasks, we split the available embeddings randomly into training (80%) and testing (20%) sets, stratifying by patient during our experiments to avoid data leakage of patient-level information from training to testing, and to ensure fair comparison of recorded predictive values. For the chest pathology diagnosis tasks, we applied an additional stratification by pathology to balance the target ratios. We then conducted experiments to compare the effect of all different combinations of input data modalities and sources using the extracted multimodal fusion embeddings.

Tasks of interest

Chest pathology diagnosis prediction: Early detection of certain pathologies in CT scans and other diagnostic imaging modalities enables clinicians to focus on early intervention rather than delayed treatment for advanced stages of relevant pathologies. Within this task of interest, we chose to target the prediction of 10 common thorax-level pathologies (i.e., fractures, lung lesions, enlarged cardio mediastinum, consolidation, pneumonia, lung opacities, atelectasis, pneumothorax, edema and cardiomegaly) that can be typically assessed by radiologists through chest X-ray, we demonstrate that HAIM outperforms image-only approaches. The ground-truth values for each chest pathology included in MIMIC-IV-MM are derived from MIMIC-CXR-JPG v2.0.0, where radiology notes were processed to determine if each of these pathologies was explicitly confirmed as present (value=1), explicitly confirmed as absent (value=0), inconclusive in the study (value=-1), or not explored (no value). We only selected samples with 0 or 1 values, removing the rest from the training and testing data. Thus, for this specific task, we utilized the multimodal fusion embeddings as input and the ground-truth chest pathology MIMIC-IV-MM values as the output target to predict. From these embeddings, we only excluded the unstructured radiology notes component (E_{rad}) from the allowable input to avoid potential overfitting or misrepresentations of real predictive value. We trained and tested independent binary classification models for each target chest pathology and input source combination as described in the general model training setup section. Final sample sizes for each pathology diagnosis task are: Fracture (N=557), Lung Lesion (N=930), Enlarged cardio mediastinum (N=3,206), Consolidation (N=4,465), Pneumonia (N=7,225), Lung opacity (N=14,136), Atelectasis (N=15,213), Pneumothorax (N=17,159), Edema (N=17,182) and Cardiomegaly (N=18,571).

Length-of-Stay prediction: Projected patient length-of-stay plays a vital role for both patients and hospital systems in making informed medical and economic decisions. An accurate prediction of patient stay enhances patient satisfaction of pre-admission expectations, hospital resource allocations, and doctors' ability to make more effective treatment planning.[40] Particularly, predicting next 48-hrs discharges is critical for physicians to identify and prioritize patients who

are ready for discharge and for case management teams to accelerate discharge preparations, which ultimately reduces patient burden and direct operating costs in healthcare systems.[41] To demonstrate the HAIM framework for healthcare operations tasks, we predicted whether or not a patient will be discharged without expiration during the next 48-hrs as a binary classification problem: discharged alive ≤ 48-hrs (1) or otherwise (0). In case of patient death, we set the class label to 0. Each sample in this predictive task corresponds to a single patient-admission EHR time point where an X-ray image was obtained (N=45,050).

48-hour mortality prediction: Due to its time and outcome-critical environments, clinicians in ICU units often need to make rapid evaluations of patient conditions to inform treatment plans.[42] However, current standards of estimating patient severity, such as the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) score, fail to incorporate medical characteristics beyond acute physiology.[43] Accurate mortality prediction can give clinicians advanced warnings of possible deteriorations and share the burdens of making information-heavy decisions.[42] To further demonstrate the versatility of the HAIM framework, we also built models to predict the probability that a patient will expire during the next 48-hrs as a binary classification problem: expired ≤ 48-hrs (1) or otherwise (0). In the case of a patient whose hospital exit status is not expiration, we set the class label to 0. It should be noted that a patient can acquire different target class labels at different time points during their stay due to changes in status and proximity to the discharge or time of death. Similar to the length-of-stay modeling, each sample in this predictive task corresponds to a single patient-admission EHR time point where an X-ray image was obtained (N=45,050).

General model training setup

We initially explored seven machine learning architectures, including logistic regression, classification and regression trees (CART), random forest, multi-layer perceptron (MLP), gradient boosted trees (XGBoost), gradient boosting machines (LightGBM), as well as attentive tabular networks TabNet to heuristically decide on the best model choice for follow-up experiments. Since XGBoost supports fast computations for large-scale experiments and consistently outperformed other architectures during preliminary observations, we selected this canonical architecture for all further tests. Our XGBoost-based modeling experiments were conducted using every possible combination of input embeddings, extracted as described in previous sections, from the allowable 11 data sources (i.e., E_{de}, E_{ce}, E_{pe}, E_{le}, E_{ecgn}, E_{econ}, E_{radn}, E_{vp}, E_{vd}, E_{vmp} and E_{vmd}) and 4 modalities (i.e., tabular, time-series, text and images). In this process, we concatenated each data stream permutation to produce fusion embeddings and train this canonical architecture using single-modality (N_{1M}=52), double-modality (N_{2M}=392), triple-modality (N_{3M}=972) and quadruple-modality (N_{4M}=630) inputs. This led to the generation of 2,047 models (per predictive task) for the cases of length-of-stay and 48-hour mortality. As previously mentioned, in the case of chest pathology diagnosis, the embeddings corresponding to all radiology notes (E_{radn}) are not included as part of the input fusion embeddings to allow for fair comparison with the output target, which was originally determined from examining notes in MIMIC-CXR-JPG. This reduced the total number of possible models per chest pathology diagnosis task to 1023 (N_{1M}=26, N_{2M}=196, N_{3M}=486, N_{4M}=315). Since there are 10 chest pathologies, defined as binary classification problems for our experiments, this leads to a total of 1,023*10=10,230 models for chest pathology diagnosis prediction.

All defined models (N_{Models}=14,324) were trained and tested to evaluate the advantage of multimodal predictive systems, based on the HAIM framework, as compared to single modality ones for the aforementioned clinical and operational tasks. We capture average trends of model performance by reporting the average area under the receiver operating characteristic (AUROC) curve on the testing set (20%) over five consecutives iterations of randomized train-test data splitting and model training. Within each training loop, the hyperparameter combinations of individual XGBoost models were selected using a 5-fold cross-validated grid search on the training set (80%). This XGBoost tuning process selected the maximum depth of tree (5-8), number of estimators (200 or 300) and the learning rate (0.05, 0.1, 0.3) according to the parameter value combination leading to the highest observed AUROC within the training loop. The average performance metrics of all these models by number of data sources and modalities can be found in Figure 2. We conducted all embedding generation and computational experiments using a parallelization strategy under MIT's Supercloud server (https://supercloud.mit.edu) with 30GB RAM and 1 NVIDIA Tesla V100 Volta graphics processing unit per instance. A high-level schematic representation of the HAIM framework from data sourcing to model benchmarking can be found in Figure 3.

RESULTS:

We demonstrate the feasibility and versatility of the HAIM framework on a compiled multimodal version of the MIMIC-IV-MM dataset, which includes a total of 34,537 samples involving 7,279 hospitalization stays and 6,485 unique patients. We summarize the general characteristics of MIMIC-IV-MM (i.e., number of samples and features) in Table 1. Qualitatively, our HAIM framework appears to improve on previous work in this field[30] by including scalable patient-centric data pre-processing and enabling standardized feature extraction stages that allow for rapid prototyping, testing, and deployment of predictive models based on user-defined prediction targets. Our HAIM framework displays consistent improvement on average AUROC (Fig. 2A color gradient) across all models as the number of modalities and data sources increases. Furthermore, the trend of reducing AUROC standard deviation (SD) values also appears to follow from increasing the number of modalities and data sources (Fig. 2A greyscale gradient). We also report Receiver Operating Characteristic (ROC) curves for the best found single-modality predictive models (Fig. 2C) as compared with typical multimodal predictive models based on the HAIM framework (Fig. 2B). All 14,324 individual model AUROCs (10,230 for chest diagnosis prediction tasks, 2,047 for length-of-stay and 2,047 mortality prediction) are shown along with their respective standard deviations in Supplemental FigS1A-D. These results suggest that our HAIM framework can consistently improve predictive analytics for various applications in healthcare as compared with single-modality analytics. Quantitatively, Figure 3A-B shows that our HAIM framework produces models with multi-source and multimodality input combinations that improve from average performance of canonical single-source (and by extension single-modality) systems for chest x-ray pathology prediction (Δ_{AUROC}: 6-22%), length-of-stay (Δ_{AUROC}: 8-20%) and 48-hour mortality (Δ_{AUROC}: 11-33%). Specifically, for chest pathology prediction, the minimum per task improvements include: Fracture (Δ_{AUROC}=6%), Lung Lesion (Δ_{AUROC}=7%), Enlarged Cardio mediastinum (Δ_{AUROC}=9%), Consolidation (Δ_{AUROC}=10%), Pneumonia (Δ_{AUROC}=8%), Atelectasis (Δ_{AUROC}=6%), Lung Opacity (Δ_{AUROC}=7%), Pneumothorax (Δ_{AUROC}=8%), Edema (Δ_{AUROC}=10%) and Cardiomegaly (Δ_{AUROC}=10%). Furthermore, the average percent improvement of all multimodal HAIM predictive systems is 9-28% across all evaluated tasks (Fig. 3A). All AUROC-related results

displayed in Figure 2A and Figure 3A-B are grouped and ordered by number of modalities (range=1 to 4, encompassing tabular, time-series, text, and images), number of data sources (range=1 to 11, including each individual data source in MIMIC-IV-MM) and sample size (N) for ease of analysis. To understand how each data source and modality contributes to the final performance, we calculate Shapley values of each of the 11 sources and 4 modalities as it contributes to the final AUROC performance.[44] Since our demonstrated predictive tasks are treated as binary classification problems, we assumed that the AUROC of a model with no data source is 0.5, and the AUROC of the model of a particular modality is the average AUROC of the models of all sources that belong to such modality. Aggregated Shapley values for all data modalities per predictive task are reported in Figure 3C, while Shapley values for all data sources per predictive task are shown in Supplemental Figure S2. Different tasks exhibit distinct distributions of aggregated Shapley values across data modalities and sources. In particular, we observe that vision data contributed most to the model performance for the chest pathology diagnosis tasks, but for predicting length-of-stay and 48-hour mortality, the patient's historical time-series records proved to be most relevant. Nevertheless, we see that across all tasks, every modality contributed positively to the predictive capacity of the models. These observations attest to the value of using multimodal inputs and HAIM-like frameworks that may facilitate the generation of these patient representation to model diverse clinical tasks more accurately. A high-level schematic of the developed HAIM pipeline for training and evaluation of models throughout this work is described in Figure 3D. The general process of MIMIC-IV-MM database preparation, as well as embedding extraction and fusion that serves as input for this pipeline can be found in Figure 1.

DISCUSSION:

Inferring latent features from rich and heterogeneous multimodal EHR information could provide clinicians, administrators, and researchers with unprecedented opportunities to develop better pathology detection systems, actionable healthcare analytics, and recommendation engines for precision medicine. Our results directly illustrated that different data modes are more useful for different tasks, and thus a multimodal approach to construct a comprehensive pipeline for AI/ML in healthcare. In addition to, leveraging multimodal inputs, our HAIM framework attempts to solve several bottleneck challenges in this kind of AI/ML pipeline for healthcare in a more unified and robust way than previous implementations, including the possibility of working with tabular and non-tabular data of unknown sparsity from multiple standardized and unstandardized heterogeneous data formats. The use of fusion embeddings obtained directly from individual patient files suggests that a HAIM framework can potentially facilitate the definition, testing, and deployment of AI/ML models that may be useful for managing complex clinical situations and day-to-day practice in healthcare systems. More specifically, if implemented across a large number of predictive tasks while using the same patient embeddings, this approach could potentially help accelerate the advent of scalable predictive systems to improve patient outcomes and quality of care. From these observations, we highlight three main contributions of this work: A) Our presented HAIM framework shows the potential for processing a diverse range of data types, while still offering the flexibility to include new sources and other specialized embedding extraction techniques; B) Based on our extensive testing, we observe a consistent and robust trend in the capacity of multimodal HAIM-like AI systems to improve upon state-of-the-art single-modality approaches for a variety of healthcare tasks; and C) Through the utilization of aggregated Shapley values we are able to quantitatively establish

the importance and heterogeneity of different data sources and modalities across a large number experiments in different healthcare tasks. Thus, demonstrating the potentials of learning from multiple data sources and modalities, and underscoring the need to collect holistic patient data that facilitates the application of multimodal machine in the healthcare domain. We envision a broad utility for the HAIM framework and its subprocesses focusing on driving cost-effective AI/ML activities for clinical and non-clinical operations. We hope that our HAIM framework can help reduce the time required to develop relevant AI/ML systems, while providing efficient utilization of human, economic and digital resources in a more timely and unified approach than the current methods used in healthcare organizations.

Acknowledgments: We thank the PhysioNet team from the MIT Laboratory for Computational Physiology for providing our researchers with credentialled access to the MIMIC-IV and MIMIC-CXR-JPG datasets and for their support in guiding multimodal data interrogation and consolidation. We specially thank Leo A. Celi and Sicheng Hao for their support on MIMIC-IV data review, as well as the Harvard TH Chan School of Public Health, Harvard Medical School, the Institute for Medical Engineering and Science at MIT and the Beth Israel Deaconess Medical Centre for their continued support of this work. We thank the MIT Supercloud for their support and help in setting up a workspace as well as offering technical advice throughout the project. Finally, we thank Eli Pivo for providing feedback and support on computational experiments to our work.

Funding: This work was supported by the Abdul Latif Jameel Clinic for Machine Learning in Health (L.R.S., D.B. and I.F). H.W. is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 174530. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Author contributions: L.R.S., Y.M., C.Z., L.D.J.B. planned and performed experiments, wrote code, analyzed the data, and wrote the manuscript. K.V.C., L.N., H.M.W., M.L.L. performed experiments, wrote code, analyzed the data, and edited the manuscript. I.F. contributed on research design and edited the manuscript. D.B. directed overall research and edited the manuscript.

Competing interests: Authors declare no competing interests.

Materials availability & correspondence: Our aggregated MIMIC-IV-MM dataset can be found at the official PhysioNet repository (https://physionet.org/projects/MIMIC-IV-MM) All code needed to evaluate the conclusions of this work can be found in the present manuscript, the Supplemental Materials or our GitHub repository (https://github.com/lrsoenksen/HAIM). Correspondence and requests for any materials should be addressed to D.B.

Additional information:
Supplemental information is available for this paper at: www.nature.com/###########
Table S1
Figure S1

Figure S2

Tables

Characteristic	MIMIC-IV-MM
# Samples	34537
# Demographic Features	6
# Chart Event Features	9
# Laboratory Event Features	23
# Procedure Event Features	10
# X-ray Features	1
# Text Note Features	3

Table 1. General characteristics of the MIMIC-IV-MM database. MIMIC-IV-MM is a combination of MIMIC-IV and MIMIC Chest X-ray filtered to only include patients that have at least one chest X-ray performed with the goal of validating multimodal predictive analytics in healthcare operations. Number of samples and quantities of features are described. Demographic features correspond only to a tabular data modality, while chart, laboratory and procedure events correspond to time series. X-ray features correspond to types of medical images, while text note features correspond to the test in radiology, electrocardiogram and echocardiogram natural language reports.

Figures

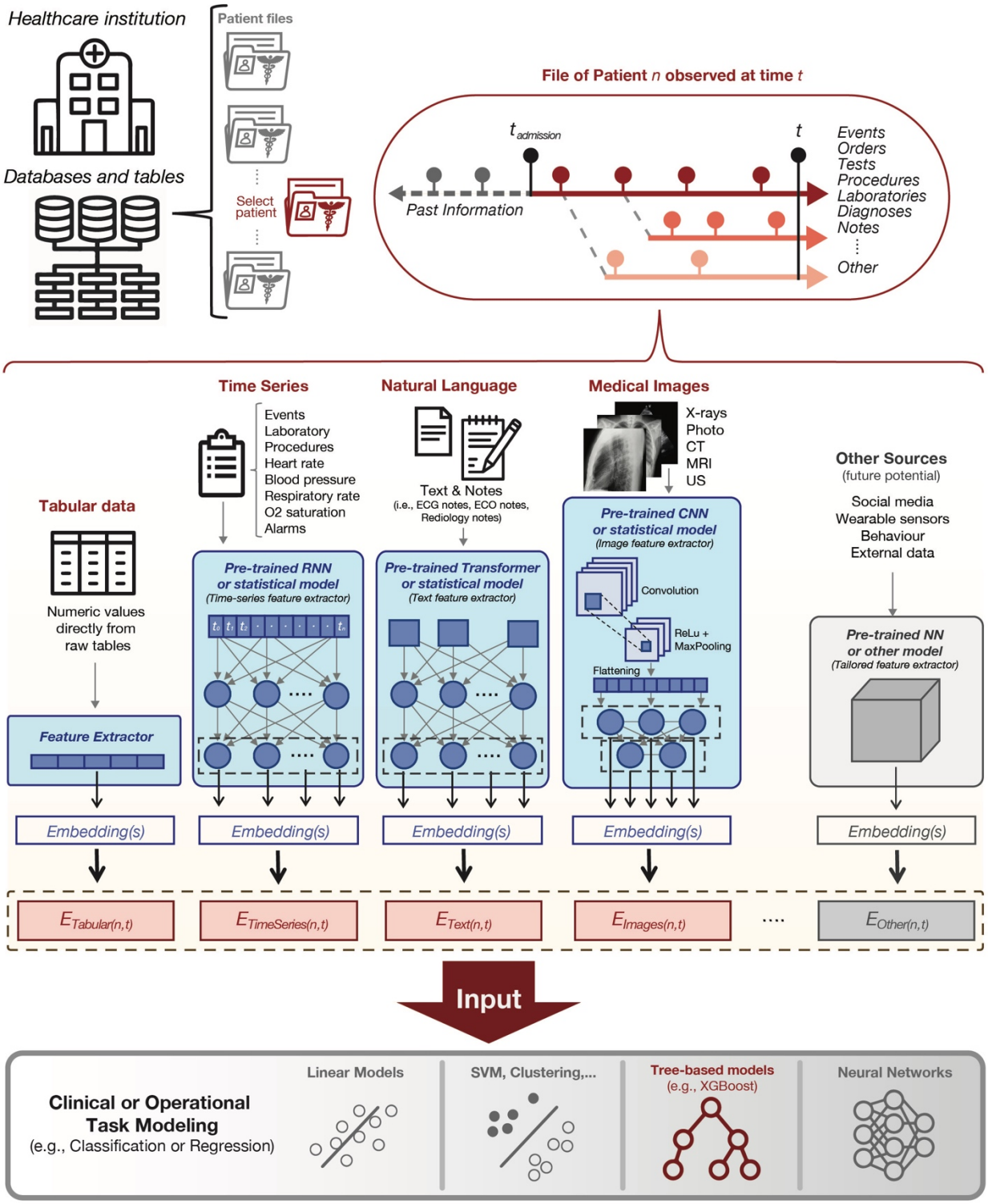

Fig. 1 Holistic Artificial Intelligence in Medicine (HAIM) framework. Under this framework, databases and tables sourced from specific healthcare institutions such as MIMIC-IV-MM

combined from MIMIC-IV and MIMIC-CXR-JPG for this work are processed to generate individual patient files. These files contain past and present multimodal patient information from the moment of admission. For processing under the HAIM framework, every data modality is fed to independent embedding generating streams. In this work, tabular data is minimally processed using simple transformations or normalizations to produce encodings or embedding-like categorical numerical values ($E_{Tabular(n,t)}$, where n=unique stay/hospitalization/patient and t=sampling time). Selected time-series are processed by generating statistical metrics on each of the time-dependent signals to produce embeddings representative of their trends ($E_{TimeSeries(n,t)}$) from the moment of admission until the sampling time. Natural language inputs such as notes are processed using a pre-trained transformer neural network to generate text embeddings of fixed size ($E_{Text(n,t)}$). Furthermore, image inputs such as X-rays are processed using a pre-trained convolutional neural network to also extract fixed-size embeddings out of the model output probability vectors and dense features ($E_{Images(n,t)}$). While not done in this work, thanks to the modularity of the embedding extraction process in the HAIM framework, other pre-trained models or systems could be added to generate embeddings from other types of data sources if needed ($E_{Other(n,t)}$). All generated embeddings are concatenated to generate a fusion embedding, which can be used to train, test, and deploy models for predictive analytics in healthcare operations. For this work, we tested and utilized only XGBoost as a canonical type of architecture for building the downstream predictive models based on fusion embeddings. CNN=Convolutional Neural Network, CT=Computerized Tomography, ECG=Electrocardiogram, ECO=Echocardiogram, MRI=Magnetic Resonance Imaging, NN=Neural Network, O2=Oxygen, ReLU=Rectified Linear Unit, RNN=Recurrent Neural Network, US=Ultrasound.

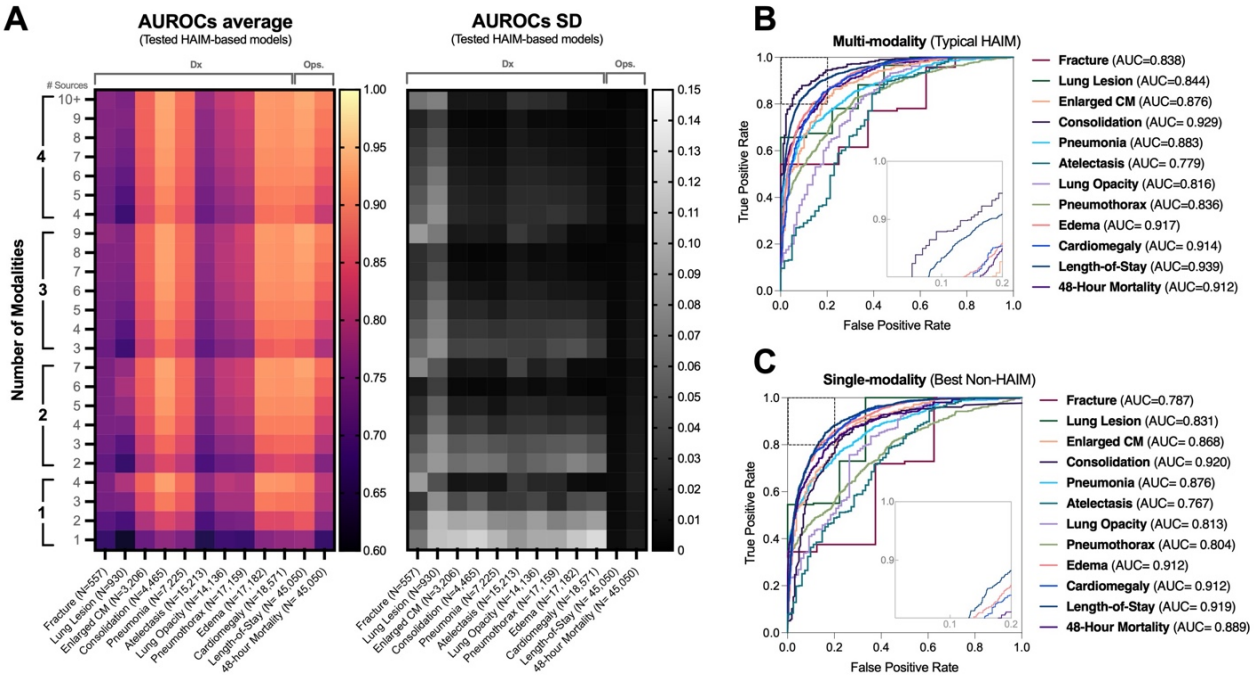

Fig. 2 Performance of the multimodal HAIM framework on various demonstrations for healthcare operations. A) Average and standard deviation values of the area under the receiver operating characteristic (AUROC) for all demonstrations including pathology diagnosis (i.e., lung lesions, fractures, atelectasis, lung opacities, pneumothorax, enlarged cardio mediastinum, cardiomegaly, pneumonia, consolidation, and edema), as well as length-of-stay and 48-hour mortality prediction. Number of modalities refers to the coverage among tabular, time-series, text, and image data. Number of sources refers to the coverage among available input data sources (10 for pathology diagnosis, while 11 for length-of-stay and 48-hour mortality prediction). Thus, the position (Modality=2, Sources=3) corresponds to the average AUROC of all models across all input combinations covering any 2 modalities using any 3 input sources. Increasing gradients on average AUROC appear to follow from increasing the number of modalities and number of sources across all evaluated tasks. Decreasing gradients on AUROC standard deviations follow from less variability in performance as a higher number of modalities and data sources is used. B) Receiver operating characteristic (ROC) curves for typical HAIM model across all use-cases exhibiting input multimodal. C) ROC curves for a best performing model with single-modality inputs across same use-cases. Consistent averaged improvements across all tasks are observed in multimodality as compared to single-modality systems. AUROC=Area under the curve, AUROC=Area under the receiver operating characteristic curve, CM=Cardiomediastinum. Dx=Diagnosis, HAIM=Holistic Artificial Intelligence in Medicine, Ops=Operations, SD=Standard deviation.

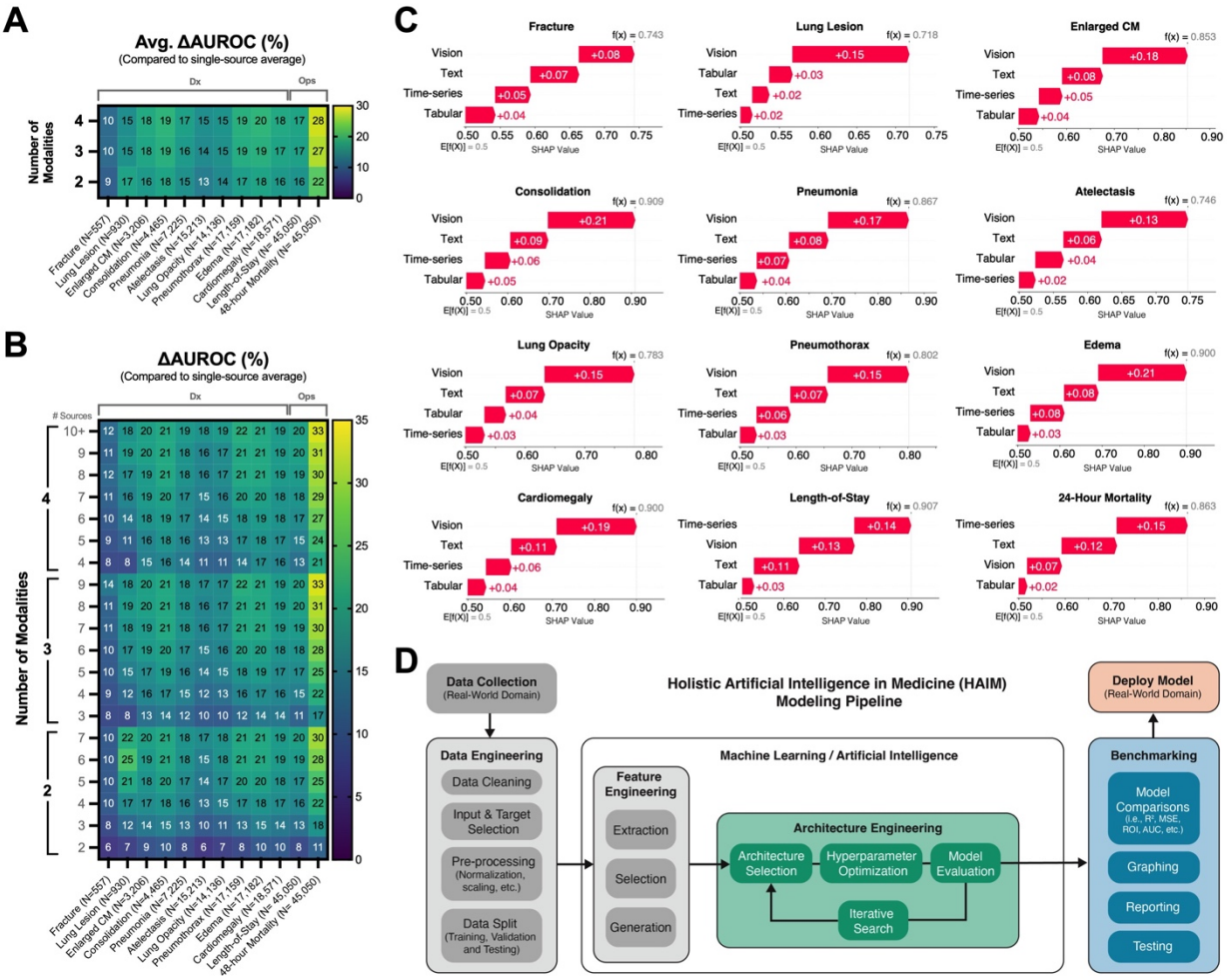

Fig. 3 Multimodal HAIM framework is a flexible and robust method to improve predictive capacity for healthcare machine learning systems as compared to a single-modality approaches. A) Average percent change of area under the receiver operating characteristic curve (Avg. ΔAUROC) for all tested multimodality HAIM models as compared to their single-source single-modality counterparts. While different models exhibit varying degrees of improvement, all tested models show positive Avg. ΔAUROC percentages. Number of modalities refers to the coverage among tabular, time-series, text, and image data. Number of sources refers to the coverage among available input data sources (10 for pathology diagnosis, 11 for length-of-stay and 48-hour mortality prediction). Thus, the position (Modality=2, Sources=3) corresponds to the average AUROC of all models across all input combinations covering any 2 modalities using any 3 input sources. B) Expanded Avg. ΔAUROC percentages for the all tested multimodality HAIM ordered by number of used modalities (i.e., tabular, time-series, text or images) as well as number of used data sources. C) Waterfall plots of aggregated Shapley values for independent data modalities per predictive task. While Shapley values for all data modalities appear to be positively contributing to the predictive capacity of all models, different tasks exhibit distinct distributions of aggregated Shapley values. D) High-level schematic of the HAIM pipeline developed to support the presented work. After data collection or sourcing (MIMIC-IV-MM for

this work), a process of feature selection and embedding extraction is applied to feed fusion embeddings into a process of iterative architecture engineering (model and hyperparameter selection). After particular models are selected and trained, they can be benchmarked to test and report results. This process concludes by the selection of a model for deployment in a use-case scenario.

References:

1 Topol, E. *Deep medicine: how artificial intelligence can make healthcare human again*. (Hachette UK, 2019).

2 Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine* **3**, 1-9 (2020).

3 Gietzelt, M., Löpprich, M., Karmen, C. & Ganzinger, M. Models and data sources used in systems medicine. *Methods of information in medicine* **55**, 107-113 (2016).

4 Boonn, W. W. & Langlotz, C. P. Radiologist use of and perceived need for patient data access. *Journal of digital imaging* **22**, 357-362 (2009).

5 Wang, W. & Krishnan, E. Big data and clinicians: a review on the state of the science. *JMIR medical informatics* **2**, e1 (2014).

6 Sun, W. *et al.* Data processing and text mining technologies on electronic medical records: a review. *Journal of healthcare engineering* **2018** (2018).

7 Agrawal, S. *et al.* Selection of 51 predictors from 13,782 candidate multimodal features using machine learning improves coronary artery disease prediction. *Patterns*, 100364 (2021).

8 Bagheri, A. *et al.* Multimodal learning for cardiovascular risk prediction using EHR data. *arXiv preprint arXiv:2008.11979* (2020).

9 Li, P., Hu, Y. & Liu, Z.-P. Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods. *Biomedical Signal Processing and Control* **66**, 102474 (2021).

10 Liu, Y. *et al.* A deep learning system for differential diagnosis of skin diseases. *Nature medicine* **26**, 900-908 (2020).

11 Stidham, R. W. Artificial Intelligence for Understanding Imaging, Text, and Data in Gastroenterology. *Gastroenterology & Hepatology* **16**, 341 (2020).

12 Paquette, A. G., Hood, L., Price, N. D. & Sadovsky, Y. Deep Phenotyping During Pregnancy for Delivery of Predictive and Preventive Medicine. *Science translational medicine* **12** (2020).

13 Purwar, S., Tripathi, R. K., Ranjan, R. & Saxena, R. Detection of microcytic hypochromia using cbc and blood film features extracted from convolution neural network by different classifiers. *Multimedia Tools and Applications* **79**, 4573-4595 (2020).

14 Hügle, M., Kalweit, G., Hügle, T. & Boedecker, J. in *Explainable AI in Healthcare and Medicine* 79-92 (Springer, 2021).

15 Tomašev, N. *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116-119 (2019).

16 Ieracitano, C., Mammone, N., Hussain, A. & Morabito, F. C. A novel multi-modal machine learning based approach for automatic classification of EEG recordings in dementia. *Neural Networks* **123**, 176-190 (2020).
17 Prashanth, R., Roy, S. D., Mandal, P. K. & Ghosh, S. High-accuracy detection of early Parkinson's disease through multimodal features and machine learning. *International journal of medical informatics* **90**, 13-21 (2016).
18 Hyun, S. H., Ahn, M. S., Koh, Y. W. & Lee, S. J. A machine-learning approach using PET-based radiomics to predict the histological subtypes of lung cancer. *Clinical nuclear medicine* **44**, 956-960 (2019).
19 Yala, A., Lehman, C., Schuster, T., Portnoi, T. & Barzilay, R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**, 60-66 (2019).
20 Reda, I. *et al.* Deep learning role in early diagnosis of prostate cancer. *Technology in cancer research & treatment* **17**, 1533034618775530 (2018).
21 An, G. *et al.* Comparison of machine-learning classification models for glaucoma management. *Journal of healthcare engineering* **2018** (2018).
22 Patel, M. J. *et al.* Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. *International journal of geriatric psychiatry* **30**, 1056-1067 (2015).
23 Huang, S.-C., Pareek, A., Zamanian, R., Banerjee, I. & Lungren, M. P. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports* **10**, 1-9 (2020).
24 Tiulpin, A. *et al.* Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Scientific reports* **9**, 1-11 (2019).
25 Wu, J. *et al.* Radiological tumour classification across imaging modality and histology. *Nature Machine Intelligence* **3**, 787-798 (2021).
26 Mei, X. *et al.* Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nature medicine* **26**, 1224-1228 (2020).
27 Bardak, B. & Tan, M. Improving clinical outcome predictions using convolution over medical entities with multimodal learning. *Artificial Intelligence in Medicine* **117**, 102112 (2021).
28 Jin, M. *et al.* Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276* (2018).
29 Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* **1**, 1-10 (2018).
30 Li, Y. *et al.* Inferring multimodal latent topics from electronic health records. *Nature communications* **11**, 1-17 (2020).
31 Johnson, A. *et al.* MIMIC-IV (version 1.0). *PhysioNet*, doi:https://doi.org/10.13026/s6n6-xd98. (2021).
32 Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* **101**, e215-e220 (2000).
33 Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Scientific data* **3**, 1-9 (2016).
34 Royalty, J. P. *Machine Learning Time-to-Event Mortality Prediction in MIMIC-IV Critical Care Database*, (2021).

35 Meng, C., Trinh, L., Xu, N. & Liu, Y. MIMIC-IF: Interpretability and Fairness Evaluation of Deep Learning Models on MIMIC-IV Dataset. *arXiv preprint arXiv:2102.06761* (2021).
36 Johnson, A. E. *et al.* MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019).
37 Alsentzer, E. *et al.* Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).
38 Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI conference on artificial intelligence* **33**, 590-597 (2019).
39 Cohen, J. P. *et al.* TorchXRayVision: A library of chest X-ray datasets and models. *arXiv preprint arXiv:2111.00595* (2021).
40 Bertsimas, D., Pauphilet, J., Stevens, J. & Tandon, M. Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management* (2021).
41 Zhu, T., Luo, L., Zhang, X., Shi, Y. & Shen, W. Time-series approaches for forecasting the number of hospital daily discharged inpatients. *IEEE journal of biomedical and health informatics* **21**, 515-526 (2015).
42 Awad, A., Bader-El-Den, M., McNicholas, J. & Briggs, J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *International journal of medical informatics* **108**, 185-195 (2017).
43 Awad, A., Bader–El–Den, M. & McNicholas, J. Patient length of stay and mortality prediction: a survey. *Health services management research* **30**, 105-120 (2017).
44 Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* **41**, 647-665 (2014).

SUPPLEMENTAL MATERIAL

Integrated multimodal artificial intelligence framework for healthcare applications

Luis R. Soenksen[1,4*], Yu Ma[2*], Cynthia Zeng[2*], Leonard D.J. Boussioux[2*], Kimberly M Villalobos[2*], Liangyuan Na[2*], Holly Mika Wiberg[2], Michael L. Li[2], Ignacio Fuentes[1], Dimitris Bertsimas[1,2,3] ‡

[1]Abdul Latif Jameel Clinic for Machine Learning in Health, MIT, Cambridge, MA 02139, USA. [2]Operations Research Center, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [3]Sloan School of Management, MIT, Cambridge, MA 02139, USA. [4]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA.
* These authors contributed equally to this work
‡ Corresponding author. Email: dbertsim@mit.edu

Supplemental Tables

#	Chart events	Laboratory events	Procedure events
1	Heart rate	Glucose	Foley Catheter
2	Non-invasive systolic blood pressure	Potassium	PICC Line
3	Non-invasive blood diastolic pressure	Sodium	Intubation
4	Non-invasive nominal blood pressure	Chloride	Peritoneal dialysis
5	Respiratory rate	Creatinine	Bronchoscopy
6	O_2 saturation by pulse oximetry	Urea nitrogen	EEG
7	Verbal GCS response	Bicarbonate	Dialysis CRRT
8	Eye opening GCS response	Anion gap	Dialysis catheter
8	Motor GCS response	Hemoglobin	Chest tube removed
9		Hematocrit	Hemodialysis
11		Magnesium	
12		Platelet count	
13		Phosphate	
14		White Blood Cells	
15		Total calcium	
16		MCH	
17		Red Blood Cells	
18		MCHC	
19		MCV	
20		RDW	
21		Platelet count	
22		Neutrophils	
23		Vancomycin	

Supplemental Table 1. Patient signals in MIMIC-IV-MM by type of event used as time-series for embedding extraction. Nine time-dependent signals were derived from procedures, twenty-three were derived from laboratories, and eight were derived from information included in the patient chart. CRRT=Continuous renal replacement therapy, EEG=Electroencephalogram, GCS=Glasgow Coma Scale, MCH=Mean corpuscular hemoglobin, MCHC=Mean corpuscular hemoglobin concentration, PICC=Peripherally inserted central catheter, RDW=Red blood cell distribution width.

Supplemental Figures

Supplemental Figure S1. A) Values of area under the receiver operating characteristic (AUROC) curves and B) Standard deviations, for all models trained for the pathology diagnosis tasks (i.e., lung lesions, fractures, atelectasis, lung opacities, pneumothorax, enlarged cardio mediastinum, cardiomegaly, pneumonia, consolidation, and edema), ordered by individual combinations of the used 10 input sources (total=1,023 models). C) Values of AUROC curves and D) Standard deviations for length-of-stay and 48-hour mortality prediction tasks), ordered by individual combinations of the used 11 inputs sources (total=2,047 models). Additional input source in length-of-stay and 48-hour mortality prediction tasks corresponds to radiology notes, which were not used in the pathology diagnosis tasks to prevent overfitting or misrepresentation of predictive capacity of trained models.

Supplemental Figure S2. Waterfall plots of aggregated Shapley values for independent data sources per predictive task. Shapley values for different tasks exhibit distinct distributions of aggregated Shapley values across input data sources, with mostly positively contributing effects towards predictive capacity (red values pointing right), with the exception of a small number of Shapley values with marginal negative values (blue values pointing left).