M3H: Multimodal Multitask Machine Learning for Healthcare

Dimitris Bertsimas
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA, 02139
dbertsim@mit.edu

Yu Ma *
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA, 02139
midsumer@mit.edu

Abstract

Recent breakthroughs in AI are poised to fundamentally enhance our study and understanding of healthcare. Developing an integrated many-to-many framework leveraging multimodal data for multiple tasks is essential to unifying modern medicine. We introduce M3H, an explainable Multimodal Multitask Machine Learning for Healthcare framework that consolidates learning from tabular, time-series, language, and vision data for supervised binary/multiclass classification, regression, and unsupervised clustering. M3H encompasses an unprecedented range of medical tasks and problem domains and consistently outperforms traditional single-task models by on average 11.6% across 40 disease diagnoses from 16 medical departments, three hospital operation forecasts, and one patient phenotyping task. It offers explainability through a proposed TIM score, shedding light on the dynamics of task learning interdependencies of the output space. The modular design of the framework ensures its generalizable data processing, task definition, and rapid model prototyping, applicable to both clinical and operational healthcare settings. Specifically, the model design features a novel lightweight attention mechanism balancing self-exploitation (learning source-task), and cross-exploration (learning cross-tasks) to ensure learning quality without overburdening computational resources. Its adaptable architecture supports easy customization and integration of new data modalities and tasks, establishing it as a robust, scalable solution for advancing AI-driven healthcare systems.

1 Introduction

The integration of Artificial Intelligence (AI) and Machine Learning (ML) has seen unprecedented promise to advance healthcare services and to fundamentally improve our understanding of medicine [1, 2]. Leveraging the increasingly accessible patient digital records, multimodal learning incorporates multiple modalities and sources of data input to provide holistic views of patient profiles [2-7]. However, beyond the integration of diverse inputs, a combination of outcomes is often further necessary to characterize patients comprehensively. Multitask learning, which leads to performance breakthroughs in large language models such as GPT-2 [8], and computer vision [9-11], is a natural extension under this premise to simultaneously learn multiple medical tasks to improve model performance across cardiology [12], psychiatry and psychology [13, 14], oncology [15], radiology [16] and other healthcare domains [17-19]. Specifically, in contrast with multiclass learning of mutually exclusive targets, multitask learning can simultaneously process multiple targets and thus provide better performance due to the sharing of common knowledge. Importantly, multimodal multitasking emulates existing collaborative efforts in clinical settings, where physicians and administrators across

*corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

multiple departments often integrate diverse sources of information to jointly navigate multiple complex medical decisions simultaneously.

However, it remains challenging to develop an integrative multimodal multitask machine learning framework that is consistently applicable across distinct healthcare domains and machine learning problem classes while maintaining efficiency in handling increasingly large healthcare datasets [20]. In particular, existing multimodal multitask medical models often primarily focus on image and vision tasks despite the majority of the medical knowledge still being encoded in tabular electronic health records. This calls for the need for the design of architectures that integrate more prominently studied machine learning problem classes on classification, regression, and clustering in these tabular settings. In addition, there is a lack of a rigorous understanding of why certain medical tasks should be combined into a single setting. Simply combining a set of tasks to achieve unified diagnostics, though appealing in concept, could unknowingly introduce tasks with conflicting objectives. Lastly, there is a lack of design of machine learning architectures that encourage simplicity and flexibility of individual components for practical implementation or adoption in other studies. Most existing frameworks are developed for specific studies, which exploit complex relationships between data and outputs through elaborate connections and repeats of hundred-level-layer neural networks. Such architectures are not feasible for many hospital's data and computing infrastructures and are difficult to adapt to environment-specific studies.

M3H addresses several challenges, including the difficulty of integration across multiple distinct machine learning problem classes into a single framework and the lack of explainability metrics to measure how and why combining certain tasks improves performance. In particular, the M3H framework complements and extends previous literature on important key topics and provides new perspectives on several topics.

- M3H represents the first integrated healthcare system to bridge beyond multi-disease diagnosis to hospital operations and patient phenotyping. In relation to this, it also represents the first step towards integrating not only clinical but also operational and biological dynamics of patient care, signaling a shift towards a holistic view across the healthcare continuum.
- M3H introduces the TIM score, an explainable metric measuring incremental performance benefits from training additional tasks in conjunction with the source task. While previous studies in this field rely on apriori assumptions about the quantitative and qualitative value of multitask learning for the target domain or rely on medical observations,
- M3H is designed to be particularly modular and flexible to allow easy substitution of each component with user-preferred model. It further develops a novel lightweight cross-task attention mechanism that explicitly models the learning between medical tasks by balancing self-exploitation (learning for the source task) and cross-exploration (learning from other tasks).

The rest of this paper is structured as follows. Section 2 outlines previous works in addressing multimodal multitask machine learning in healthcare settings. Section 3 details the architecture and technical details of the M3H framework. Section 4 describes the experimental set up using a large-scale intensive care unit (ICU) database. Section 5 demonstrates M3H's performance across a diverse set of medical and machine learning tasks. The explainability metric is characterized in Section 6, and managerial implications, limitations, and future works are discussed in Section 7.

2 Related Literature

2.1 Integrated Healthcare System

Medicine is not a standalone domain of study. On the quite opposite, medical departments rely heavily on the support and interactions of inter-departmental collaborations. Current studies in healthcare management largely rely on a single-disease prediction for a single medical diagnosis, treatment, or planning problem. These domain-specific models offer expert insights into a particular domain of healthcare and could benefit significantly from sharing knowledge with each other if jointly studied under a unifying framework. Some recent works on the integration of multiple clinical tasks across cardiology [12], psychiatry and psychology [13, 14], oncology [15], radiology [16] and other healthcare domains [17-19] all show promising potential for an integrated healthcare system for improved performance.

A further testimony to this critical direction towards integration is the rising interest in the medical field on the study of multi-disease [21], or multi-morbidity diagnosis [22]. Such approaches make heavy use of the underlining assumption that patient characteristics, as well as medical conditions, when studied holistically, provides a better understanding, and thus improves both performance as well as clinical understanding. Beyond providing insightful medical knowledge, these studies imply significant managerial benefits for the patients, caregivers, and even payers [23, 24]. A concrete illustration was demonstrated by the prediction of patient flow in a large hospital system [25-27], where predictions of multiple operational targets, including length of stay, mortality, and ICU admissions, are all studied to characterize the patient's condition. If integrated into a single framework, hospital systems could benefit from performance improvements, directly contributing to operational efficiency, and profits for the organization.

2.2 Medical Foundation Models

Our work is in line with recent literature on the development of medical foundation models [28-30], or in some cases referred to as the "generalist" models, in comparison to traditional "specialist" single-modality, single-task models. These methods were first extensively studied in computer vision, control, and natural language processing, which primarily aimed to combine machine learning tasks such as image segmentation, image detection, language translations into a single, cohesive framework. Our work differs from this line of work in two ways: 1) majority of these works primarily focus on vision and language task integrations, we instead heavily rely on electronic health records (EHR), which until remains the most used medical data, as an integral part of our analysis and base majority of our architecture design on these data. 2) Mere integration of all tasks into a single system, although appealing in concept, is not applicable in the medical practice. Instead, a more feasible approach is to offer a flexible framework where users can decide tasks to be integrated by leveraging their own expert knowledge of the healthcare system.

2.3 Explainability on the Outcome Space

Existing explainable multitask frameworks for learning biomedical tasks focus on explaining the contribution of input features on its corresponding outcome task [18]. However, multitask learning offers a unique perspective that lies in the interactions among the jointly learned tasks interact. With the introduction of several techniques on the integration of machine learning tasks, such as cross-stitch networks [31], and cross-task attention mechanism [32], we can effectively extract learned attention weights or cross-stitch tensors to map out the interactions between how tasks borrow other tasks' learned feature embeddings to improve its own learning process. However, these approaches are designed to measure the learning process contribution, not the end effect of model performance changes by specific combinations of tasks.

3 M3H Architecture

3.1 Overview of the M3H framework

The M3H (Fig 1) framework is an end-to-end framework for integrating multimodal data feature extraction and multitask outcome learnings. To leverage the strong performance of existing state-of-the-art (SOTA) models, M3H first obtains fixed modality-specific embeddings through publicly available, pre-trained models, including ClincalBERT [33] for natural language and Densenet121-res224-chex [34] for images. These task-agnostic, not-trainable multimodal embeddings are then passed through further modality-specific learnable feedforward networks and then integrated into a shared-task learning module. In this module, we conduct (i) contrastive learning and (ii) shared-task learning, where the first aims to project embeddings from different modalities into a consistent embedding space, and the second serves as an over-arching tunnel that all tasks must contribute to learning and a proxy for a universal embedding that is relevant for all tasks. We then feed the shared-learned embedding to task-specific networks, which focuses on the learning of each individual task. Finally, these task-specific embeddings integrate knowledge from other task embeddings before making their final predictions via the cross-task attention mechanism. In multimodal multitask machine learning problems, it remains challenging to unify a diverse pool of outcomes due to the presence of different output spaces (continuous numeric, discrete categories). M3H integrates tasks of different medical domains and machine learning problem classes by unifying losses from each

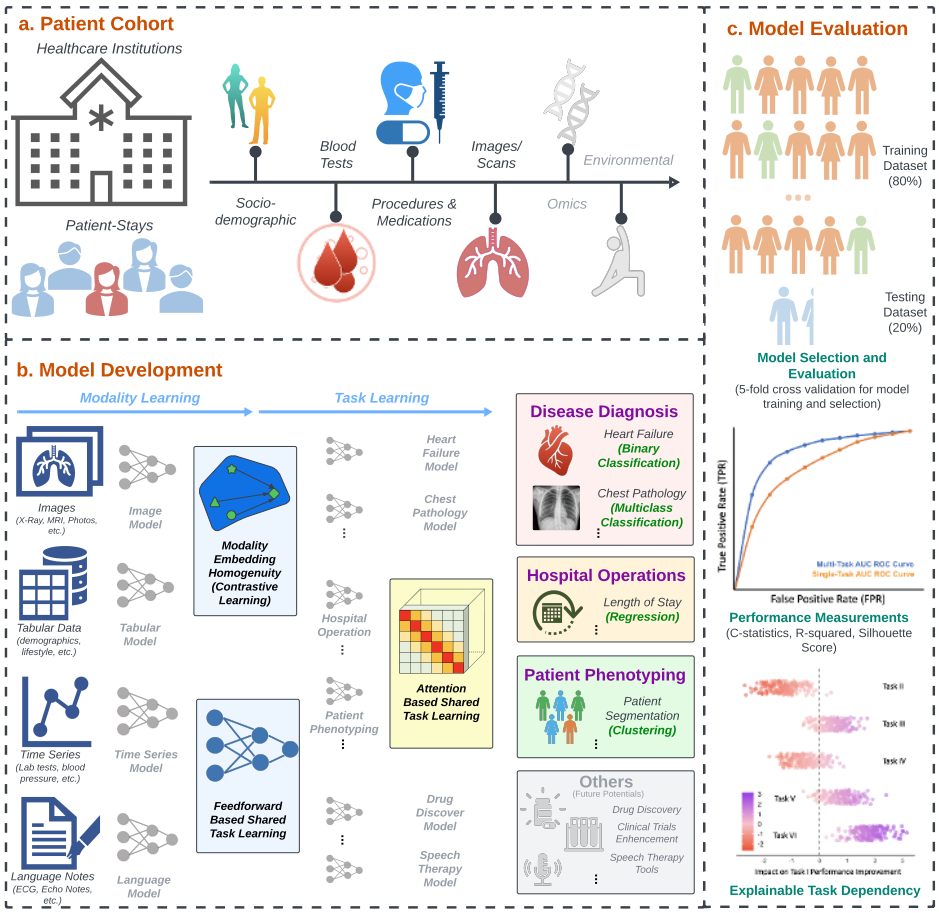


Figure 1: Multimodal Multitask Machine Learning for Healthcare (M3H) system.

sub-problem into a single objective function. Specifically, the overall loss is a combination of contrastive loss between multimodal inputs and aggregated problem class loss of jointly learned outcomes. During training, network updates by optimizing each individual loss sequentially.

3.2 Machine Learning Problem Class Architectures

The M3H framework assigns a pre-defined modality-specific feedforward network for each input modality and a task-specific network for each outcome task, which can be seen in Supplemental Figures A2-1 and A2-2. Specifically, as an unsupervised problem with unlabeled samples, clustering is uniquely challenging to incorporate into the M3H framework. Unlike the use of output layers to predict a predefined label, to effectively group patients into different phenotypes, we train an autoencoder that learns accurate low-dimensional latent space that can be then clustered into groups via traditional methods such as K-means clustering. We first concatenate all embeddings from all modalities into an aggregated embedding, this embedding is then fed during training into an autoencoder to compress the original feature space into low-dimension latent space and then re-expanded back to the original dimensions. A good quality latent space aims to achieve low reconstruction loss. The learned latent space is then clustered into 15 patient subgroups and evaluated for quality.

3.3 Cross-Task Attention for Knowledge Sharing

Integrating the learning of multiple tasks is of crucial importance for a successful M3H framework. Previous studies on related efforts include but not limited to multi-head attention mechanism, cross-stitch, and multilinear relationship network (MRN). Preliminary analyses in online Supplemental Materials show that these methods do not satisfy the quantitative or qualitative design requirements

for a high-performing, scalable objective of M3H, seen in Supplemental Materials Section A6. We thus develop below a novel cross-task attention mechanism to facilitate knowledge sharing among tasks. At its core, attention is constructed by variations of the key, query, and value vectors to capture interactions between its inputs. Attention mechanism is well-positioned to exploit dependencies between tasks: by projecting each task's embeddings as a token, we can leverage the attention mechanism to enable explicit task knowledge sharing. The overall architecture is demonstrated in Fig 2.

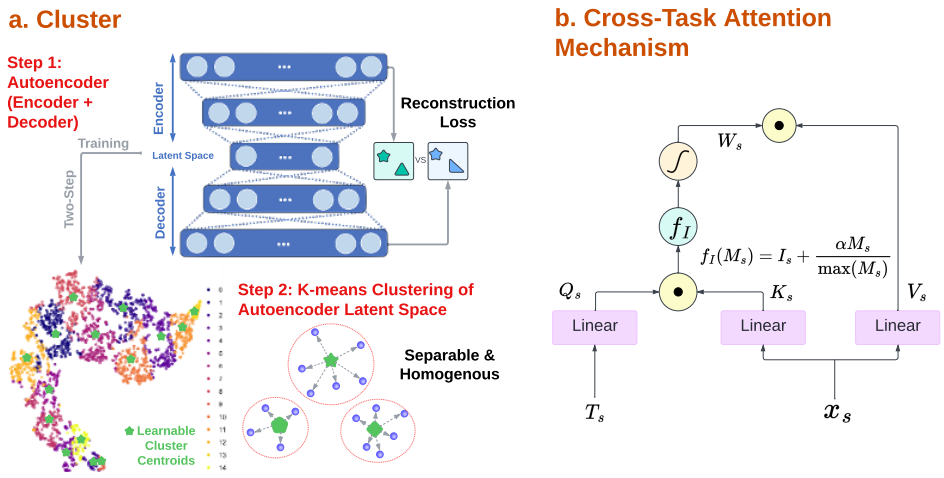


Figure 2: Architecture Design of the Clustering and Cross-Task Attention Mechanism.

Following traditional practice, input to obtain the key and value vectors is the original task embedding $x_s \in \mathbb{R}^{n_{batch} \times n_{tasks} \times n_{feature}}$ from joint learning of a specified set of task s, where n_{batch} refers to the number of samples passed through each learning iteration (batch size), $n_{feature}$ refers to the number of features generated to encode knowledge for each task, and n_{tasks} refers to the number of tasks in s. However, we aim to find a universal mapping between task tokens indicating the index of a task, with the embedding that best represents a task. This calls for a query vector that is independent of the batch update. To do so, we generate the query embedding via mapping of the task tokens vector T_s to task embeddings via a linear projection $Q_s = f(T_s) : N^{n_{tasks}} \rightarrow \mathbb{R}^{n_{tasks} \times n_{feature}}$. Finally, we apply linear projections to all embeddings to improve representation quality and obtain query vector $Q_s \in \mathbb{R}^{n_{tasks} \times n_{feature}}$, key vector $K_s \in \mathbb{R}^{n_{batch} \times n_{tasks} \times n_{feature}}$, and value vector $V_s \in \mathbb{R}^{n_{batch} \times n_{tasks} \times n_{feature}}$. The product of the computed query and key vectors, referred to as attention weight, indicates the relevance or emphasis put on a specific token in the value vector. We aim to find a balance between exploiting self-learning (reusing knowledge from the original task) while exploring cross-learning (incorporating knowledge from other unrelated tasks) in a controlled manner. This balance is achieved by adapting the initial attention weight $M_s \in \mathbb{R}^{n_{batch} \times n_{tasks} \times n_{tasks}}$ through the projection $W_s = \text{softmax}(I_s + \alpha M_s / \max(M_s))$ where $I_s \in \mathbb{R}^{n_{batch} \times n_{tasks} \times n_{tasks}}$ is the identity matrix encouraging self-learning, and α is the strength of exploration encouraging cross task learning. This attention weight is then applied to the values vector to obtain the final cross-learned task embeddings. The completed algorithm can be found in Algorithm **??**.

3.4 Model Loss

The aggregated loss used to train to network is defined as a weighted average across all losses: $\ell_{total} = w_c \ell_{constrative} + \sum_{t \in B_s} w_t \ell_{binary,s} + \sum_{t \in M_s} w_t \ell_{multiclass} + \sum_{t \in R_s} w_t \ell_{regression} + \sum_{t \in C_s} w_t \ell_{cluster}$ where w_c, w_t refers to the weight assigned for contrastive loss and task t, B_s, M_s, R_s, C_s refers to the sets of tasks that are in the binary, multiclass, regression and cluster problem classes. Specifically, binary classification loss refers to binary cross entropy loss; multiclass classification loss refers to negative log-likelihood loss; regression loss refers to mean absolute error loss; cluster reconstruction loss refers to the mean squared error between encoder input and decoder output. The detailed definitions of each loss function can be found in Supplemental Materials Section A7. All weights have been initialized to 1.

Algorithm 1 Cross-task Learning Algorithm

Input:

s: Set of tasks to be jointly learned
N: Number of batches
α: Exploration strength parameter for cross-task learning
$x_s^i \in \mathbb{R}^{n_{\text{batch}} \times n_{\text{tasks}} \times n_{\text{feature}}}$: Task embedding tensor from batch i
$T_s \in \mathbb{R}^{1 \times n_{\text{tasks}}}$: Task tokens tensor of form $[0, 1, 2, \ldots, n_{\text{tasks}} - 1]$
$I_s \in \mathbb{R}^{n_{\text{batch}} \times n_{\text{tasks}} \times n_{\text{tasks}}}$: Identity tensor to encourage self-learning

Output:

$O_s \in \mathbb{R}^{n_{\text{batch}} \times n_{\text{tasks}} \times n_{\text{feature}}}$: Cross-learned embedding output.

Initialize Linear Transformations: Initialize linear transformations for queries, keys, and values: $W_Q, W_K, W_V \in \mathbb{R}^{n_{\text{feature}} \times n_{\text{feature}}}$, and task token linear transformation $W_T \in \mathbb{R}^{n_{\text{feature}} \times n_{\text{tasks}}}$.
for $i \leq N$ **do**
$Q_s \leftarrow W_T \cdot T_s$ ▷ Convert task tokens vector T_s into task embeddings
$Q_s \leftarrow W_Q \cdot Q_s$ ▷ Compute query vectors from task embeddings
$K_s \leftarrow W_K \cdot x_s^i$ ▷ Compute key vectors from task embeddings
$V_s \leftarrow W_V \cdot x_s^i$ ▷ Compute value vectors from task embeddings
$M_s \leftarrow Q_s \cdot K_s^\top$ ▷ Compute attention weight
$W_s \leftarrow \text{softmax}\left(I_s + \frac{\alpha M_s}{\max(M_s)}\right)$ ▷ Normalize by maximum weight entry, scale by exploration strength, and encourage self-learning
$O_s \leftarrow W_s \cdot V_s$ ▷ Output the cross-learned embeddings
end for

4 Experiment Setup on Large-scale Medical Database

4.1 Dataset and Patient Representation

HAIM-MIMIC-MM [3] is a patient-centric dataset derived from Medical Information Mart for Intensive Care IV (MIMIC-IV) [35], a public electronic health record database from Beth Israel Deaconess Medical Center containing de-identified records of all patients admitted to the intensive care unit (ICU) between 2008-2019. HAIM-MIMIC-MM offers access to contemporary, large-scale patient cohorts with modular constituent data organization, and most importantly, integrates multiple modalities of data inputs into a single database, ranging from demographics, chart events, laboratory events, procedure events, radiological notes, electrocardiogram notes, echo-cardiogram notes, as well as chest X-ray images. Specifically, HAIM-MIMIC-MM aggregates all available medical information of a patient's hospital admission-stay gathered before their expiration or discharge time. Fixed-size vector representation of data from four modalities: tabular (dimension of 6), time-series (dimension of 451), vision (dimension of 2084), and language (dimension of 2304), are extracted using pre-trained, state-of-the-arts models and combined into a comprehensive multimodal patient representation. Each sample within the HAIM-MIMIC-MM dataset corresponds to all prior patient information from the time of admission until an inference event, including the time of imaging procedure for pathology diagnosis, the 48-hour window for mortality prediction, or the end of hospital stay. Additional detailed discussions of the limitations and other characteristics of the dataset can be found in Supplemental Materials Section A5.

4.2 Model Training Pipeline

We initially explored various feedforward architectures for each modality-specific and task-specific network including different activation functions (ReLU, Sigmoid, Tanh), dropout layers, normalization layers, different optimizers (RMSprop, SGD, Momentum, Adam), and gradient clipping. The canonical architecture used in all following experiments was selected to support GPU optimization for computational efficiency (i.e., the number of filters in layers mostly are multiples of 64) and was shown to have a consistently superior performance during preliminary investigations. The rescaling coefficient in the cross-task attention mechanism α is set to be 0.1 as it is explored to be a stable point between performance stability and efficient learning. We first split the dataset into 80% training (n=10025) and 20% testing (n=2561) by stratifying on a patient level to ensure no data leakage between training and testing for all model training or validation processes. We then apply a 5-fold

cross-validation on the training set to select the best combinations of batch size (256, 512) and learning rate (0.0001, 0.0003). Specifically, within each run of 15 epochs, 4 of the 5 folds are used for model training, and the remaining one is used for validation. The average of all 5 validation scores across all tasks is computed for each hyperparameter combination, and a final model is trained on the entire training set with the hyperparameter with the highest average validation score. As the number of tasks included in joint learning grows, there is an exponentially growing number of potential possible task pair combinations. We consider the following task selection procedure to optimize our likelihood of locating the best-performing multitask model: given a set of tasks $s \cup i$ and its performance on i, we conduct experiments on all possible $s \cup i \cup j$ where j is a task not previously included. We only keep the best-performing top 3 pairs and repeat until no further improvements are observed. For computational efficiency, we restrict pairs experiments up to pairs of 3. A detailed step-by-step guide can be found in Supplemental Materials Section A3.

5 Experimental Results

5.1 Quantitative Performance Improvements Across Medical Tasks

We demonstrate the feasibility of the proposed M3H framework through its application to a pre-established and validated multimodal dataset. Across 16 disease groupings with 40 disease diagnoses, 3 hospital operations tasks (length of stay, general mortality, and hospital-acquired infection), and 1 patient phenotyping task, the M3H framework demonstrates consistent performance improvement over single-task models in Fig. 4. We report the percentage of improvement and its lower and upper bound accounting for standard deviation after applying bootstrapping on the out-of-sample test scores between the best-performing single-task models and best-performing multi-task models. The M3H framework improves performance scores in diagnosis (1% – 41.2%), in hospital operations (3.3% – 12.4%), and in patient phenotyping (62.7%). Specifically, the improvement across disease groupings or hospital functionalities include hospital operations (Δ_{AUROC} = 4.7 – 12.4%, $\Delta_{\text{R-squared}}$ = 3.3%), thoracic testing ($\Delta_{\text{Average AUROC}}$ = 2.4%), blood disorder (Δ_{AUROC} = 2.9%), cardiology (Δ_{AUROC} = 1.4 – 5.3%), critical care (Δ_{AUROC} = 1.6 – 4.3%), dermatology (Δ_{AUROC} = 36.2%), endocrinology (Δ_{AUROC} = 1 – 14.1%), gastroenterology and hepatology (Δ_{AUROC} = 3.8 –25.7%), infectious diseases (Δ_{AUROC} = 1.8 – 26.9%), internal medicine (Δ_{AUROC} = 4.4 – 4.8%), nephrology (Δ_{AUROC} = 3.2%), neurology (Δ_{AUROC} = 2.7 – 41.2%), oncology (Δ_{AUROC} = 1.7 – 15.0%), ophthalmology (Δ_{AUROC} = 9.9 – 24.9%), psychiatry and psychology (Δ_{AUROC} = 2.1 – 22.7%), pulmonology (Δ_{AUROC} = 2.4 – 7.4%), rheumatology (Δ_{AUROC} = 16.4%), and urology (Δ_{AUROC} = 30.1%). Multi-task models are also shown to have reduced variability with more narrow confidence intervals, implying their potential to generate more robust solutions on unseen datasets, as can be seen in Supplemental Figure A1.

5.2 Generalizability across Machine Learning Problem Classes

Supervised and unsupervised machine learning (ML) have unique modeling techniques tailored for each corresponding outcome and objectives. However, these distinctions of machine learning problem class should not pose barriers to integrating relevant tasks that can benefit from learning simultaneously. For binary classification, predicted major depressive disorder risk scores quantile, when compared against observed event rate, shows more consistency between female and male subgroups under the multitask setting, especially for low-risk patients (Fig. 3a); for multiclass classification, we observe reduced variability of ROC curve across different thorax conditions with higher averaged AUROC measure (Fig. 3b); for regression, multitask captures tail-predictions (extended length of stay) more closely than single-task (Fig. 3c); for clustering, post-UMAP (Uniform Manifold Approximation and Projection) processing demonstrates significantly more distinct boundaries between clusters and structural patterns in the multitask setting (Fig. 3d). Together, joint learning across machine learning problem classes improves both quantitative performance as well as qualitative understanding of the source tasks.

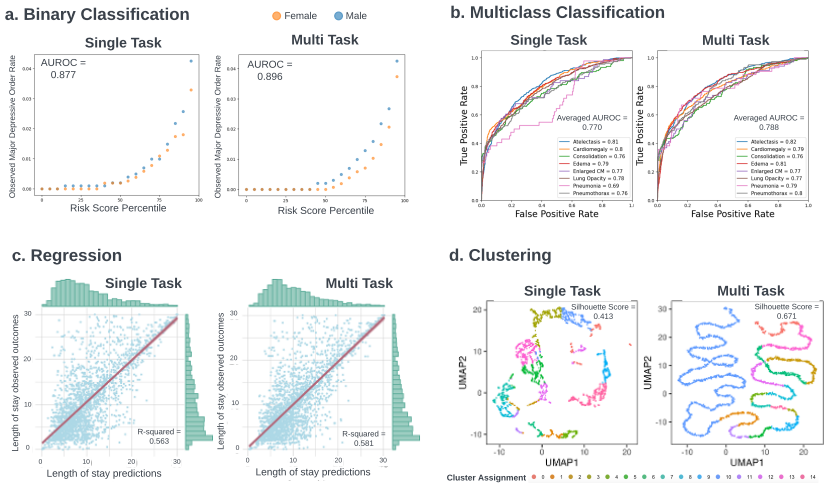


Figure 3: MultiTask outperforms SingleTask across the four machine learning problem classes.

6 Explainability

6.1 Task Interaction Measurement (TIM) Score Formulation

Explainability of how the input contributes to the output can be done through established methods such as SHAP, which is demonstrated in Supplemental Materials Section A4. The M3H framework also provides explainability of task-dependency by computing a task interaction measurement (TIM) score, which measures how joint training of additional tasks affects the performance of the source task. It helps identify tasks that should be trained together to improve performance and provide qualitative medical insights into how different medical domains interacts and potentially connects. The score is computed as the difference in performance scores between joint learning of task pairs and source-task learning. Given M as the number of all possible tasks, S as a set of tasks that do not contain either task i or task j, and $\widetilde{f}_x(S \cup i)$ as a function of the performance score of task i given features x and joint learning all tasks belonging to S and task i, we define TIM as:

$$\delta_i j = \frac{1}{2^{M-2}} \sum_{S \subseteq \setminus \{i,j\}} \widetilde{f}_x(S \cup \{i,j\}) - \widetilde{f}_x(S \cup \{i\}))$$

As the number of all possible tasks grows, this score requires an exponentially increasing number of all potential task combinations of S. In practice, to avoid computational hurdles, we can either sample a subset of potential S to obtain an approximation of the true TIM score or restrict the number of task pair sizes to be small (i.e., smaller than 5).

6.2 Task Interdependency Understanding

We show in Fig. 4 that using the proposed task interaction measurement (TIM) score, we can quantify both the positive and negative contribution of additional tasks on a source task. Notably, consistent with previous findings, the additional joint learning of infectious diseases helps improve the forecast of length of stay, and inflammatory bowel disease learning contributing to bipolar disorder risk prediction. We compare multitask models of all pairwise task combinations of size 2 (restricted to a small number to ensure computational efficiency) against single-task models using only the source task across various medical domains. We remark that the heatmap is not symmetric, showing that the direction of task interdependencies matters, as the effect of task A on task B may differ from the effect of task B on task A. This asymmetry highlights the complex nature of task relationships and their varying impacts depending on the direction of the interdependency and suggests that when designing multi-task models, it is important to clarify the rank of objectives when multiple tasks are jointly learned. Overall, the TIM score helps understand whether a particular task combination improves individual learning by sharing knowledge, impairs learning by competing between conflictive objectives, and can provide qualitative insights to better understand under-investigated medical outcome connections.

Figure 4: Heatmap of TIM Scores Across Medical Tasks to Illustrate Task Interactions.

7 Conclusions and Limitations

7.1 Implications to Practice

M3H can be readily adopted in production for hospital systems, especially in resource-constrained settings. By leveraging a modular architecture, M3H is adaptable to each system's specific patient cohorts, available data modalities, and targeted medical tasks of interest by retraining the network on these new cohorts. This versatility facilitates user-defined modifications, replacements, and extensions, ensuring a tailored application in diverse hospital environments. M3H is also developed to be easily implementable with standard data storage or computational infrastructures available in most hospital settings, but rigorous validation must be conducted to avoid potential model instability and negative predictions that could contribute worsen patient care. It has been packaged and tested on standard PCs as an executable software, and preliminary testing showed its feasibility in these local systems across both the Linux and Mac operating systems. Particularly in resource-constrained hospital systems, where information technology (IT) departments lack the capacity to manage huge-scale models, M3H offers a scalable alternative to democratize the use of such AI systems. Once validated, these prototypes can be implemented both in the clinical care delivery and the administrative operation management routines.

7.2 Limitations and Discussions

M3H could benefit from the use of additional data modalities and medical tasks such as omics and wearable device signals. Other well-studied medical tasks, such as image segmentation, and language understanding can also be included to the framework. M3H framework should further investigate robustness towards data perturbation, as well as numerical instability inherited by most deep learning architectures via the use of methods such as distributionally robust optimization. Furthermore, designing multidisease is a nonlinear, combinatorial problem that can be challenged by the curse of dimensionality as the number of possible combinations explode. Some recent works have been done to explore Pareto optimal disease-combinations, which is a promising direction to explore. Lastly, an interesting usage of the M3H framework is in connection with the predict-then-optimization literature. With more accurate performance in the prediction phase, it is possible that we can simultaneously improve multiple downstream optimization problems for better operational efficiency and recover analytical insights across medical departments.

References

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

[2] Yu, K., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, **2**(10), 719-731.

[3] Soenksen, L. R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., Wiberg, H. M., Li, M. L., Fuentes, I., & Bertsimas, D. (2022). Integrated multimodal artificial intelligence framework for healthcare applications. *Npj Digital Medicine*, **5**(1):1-10.

[4] 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 Dig. Med.* ***3***, 1–9 (2020).

[5] Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, **28**(9), 1773-1784.

[6] Baltrusaitis, T., Ahuja C., and Morency L-P. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 2 (February 2019), 423–443.

[7] Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database (Oxford)*. 2020 Jan 1;2020:baaa010.

[8] Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever. "Language Models are Unsupervised Multitask Learners." (2019).

[9] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*

[10] He, K., Gkioxari, G., Dollár, P., & Girshick, R.B. (2017). Mask R-CNN. 2017 *IEEE International Conference on Computer Vision (ICCV)*, 2980-2988.

[11] Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., & De Freitas, N. (2022). A Generalist Agent.

[12] Torres-Soto, J., Ashley, E.A. Multi-task deep learning for cardiac rhythm detection in wearable devices. *npj Digit. Med.***3**, 116 (2020).

[13] Tseng, V.WS., Sano, A., Ben-Zeev, D. et al. Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. *Sci Rep* **10**, 15100 (2020).

[14] Lee, M.H., Kim, N., Yoo, J. et al. Multitask fMRI and machine learning approach improve prediction of differential brain activity pattern in patients with insomnia disorder. *Sci Rep* **11**, 9402 (2021).

[15] Fu, S., Lai, H., Li, Q., Liu, Y., Zhang, J., Huang, J., Chen, X., Duan, C., Li, X., Wang, T., He, X., Yan, J., Lu, L., & Huang, M. (2021). Multi-task deep learning network to predict future macrovascular invasion in hepatocellular carcinoma. In *eClinicalMedicine* (Vol. 42, p. 101201). Elsevier BV.

[16] Jin, C., Yu, H., Ke, J. et al. Predicting treatment response from longitudinal images using multi-task deep learning. *Nat Commun* **12**, 1851 (2021).

[17] Eyuboglu, S., Angus, G., Patel, B.N. et al. Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body FDG-PET/CT. *Nat Commun* **12**, 1880 (2021).

[18] Wang, X., Cheng, Y., Yang, Y. et al. Multitask joint strategies of self-supervised representation learning on biomedical networks for drug discovery. *Nat Mach Intell* **5**, 445–456 (2023).

[19] Tang, X., Zhang, J., He, Y. et al. Explainable multi-task learning for multi-modality biological data analysis. *Nat Commun* **14**, 2546 (2023).

[20] A. Ahmed, R. Xi, M. Hou, S. A. Shah and S. Hameed, "Harnessing Big Data Analytics for Healthcare: A Comprehensive Review of Frameworks, Implications, Applications, and Impacts," in *IEEE Access*, vol. 11, pp. 112891-112928, 2023.

[21] Buergel, T., Steinfeldt, J., et al(2022). Metabolomic profiles predict individual multidisease outcomes. *Nature Medicine*, **28**(11), 2309-2320.

[22] Zhao, Y., Zhuang, Z., Li, Y., Xiao, et al. (2024). Elevated blood remnant cholesterol and triglycerides are causally related to the risks of cardiometabolic multimorbidity. *Nature Communications*, **15**(1):1-9.

[23] Bish, D. R., Bish, E. K., & El Hajj, H. (2024). Disease Bundling or Specimen Bundling? Cost- and Capacity-Efficient Strategies for Multidisease Testing with Genetic Assays. *Manufacturing & Service Operations Management*, **26**(1):95–116. Institute for Operations Research and the Management Sciences (INFORMS).

[24] Apergi, L.A., Bjarnadóttir, M.V., Baras, J.S., & Golden, B.L. (2023). Cost Patterns of Multiple Chronic Conditions: A Novel Modeling Approach Using a Condition Hierarchy. *INFORMS Journal on Data Science*. Institute for Operations Research and the Management Sciences (INFORMS).

[25] Bertsimas, D., & Pauphilet, J. (2023). Hospital-Wide Inpatient Flow Optimization. *Management Science*. Institute for Operations Research and the Management Sciences (INFORMS).

[26] Bertsimas, D., Pauphilet, J., Stevens, J., & Tandon, M. (2022). Predicting Inpatient Flow at a Major Hospital Using Interpretable Analytics. *Manufacturing & Service Operations Management*, **24**(6):2809–2824. Institute for Operations Research and the Management Sciences (INFORMS).

[27] Na, L., Carballo, K. V., Pauphilet, J., Kombert, D., Castiglione, et al. (2023). Patient Outcome Predictions Improve Operations at a Large Hospital Network.

[28] Tu, T., Azizi, S., Driess, D., Schaekermann, et al. (2023). Towards Generalist Biomedical AI.

[29] Yang, L., Xu, S., Sellergren, A., et al. (2024). Advancing Multimodal Medical Capabilities of Gemini.

[30] Mo, S., & Liang, P. P. (2024). MultiMed: Massively Multimodal and Multitask Medical Understanding.

[31] Misra I., Shrivastava A., Gupta A., and Hebert M. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016.

[32] Liu, S., Johns, E., & Davison, A. J. (2018). End-to-End Multi-Task Learning with Attention.

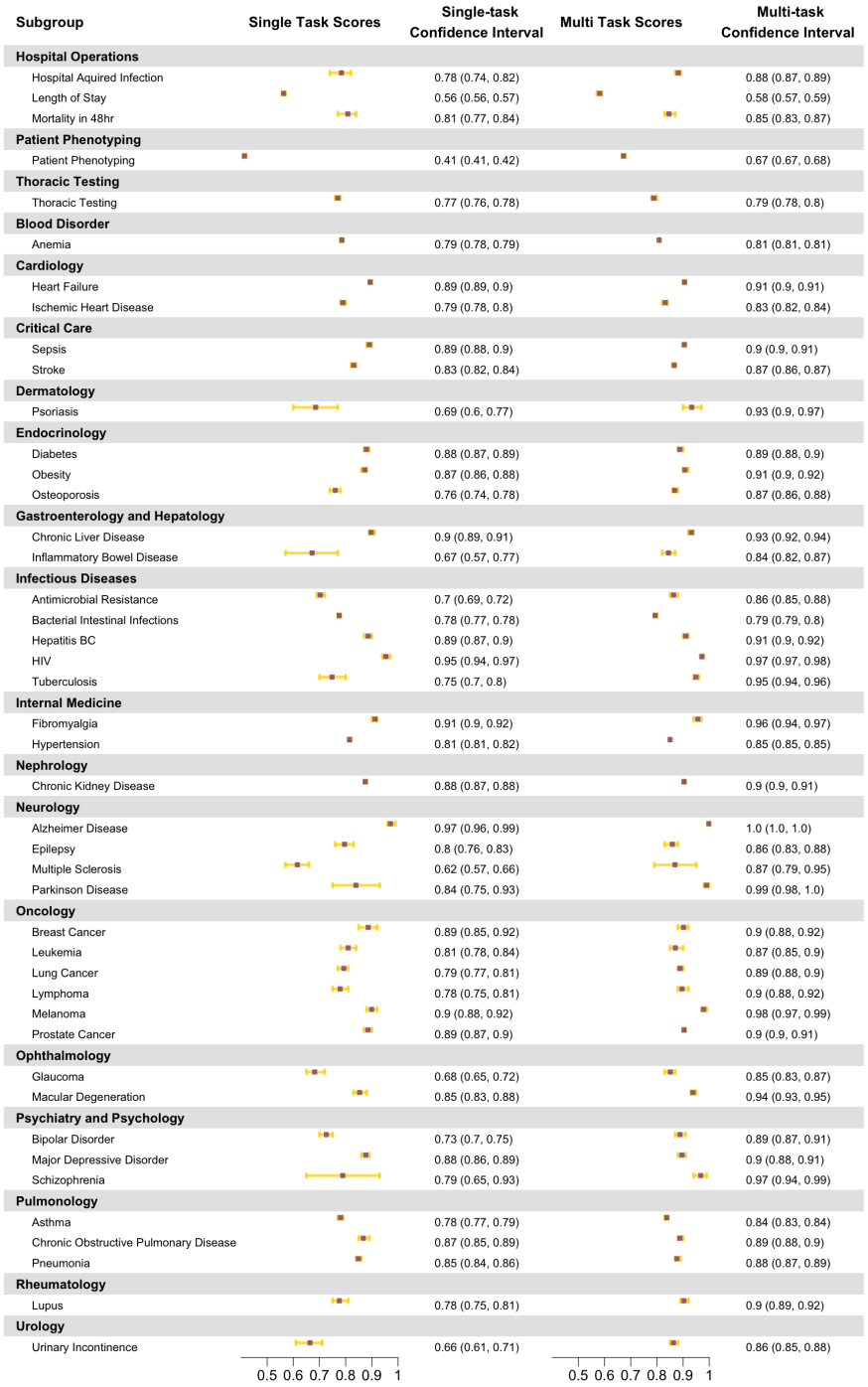
[33] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W., Jin, D., Naumann, T., & McDermott, M. B. (2019). Publicly Available Clinical BERT Embeddings.

[34] Cohen, J. P., Viviano, J. D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M. P., Chaudhari, A., Brooks, R., Hashir, M., & Bertrand, H. (2021). TorchXRayVision: A library of chest X-ray datasets and models.

[35] Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10(1), 1-9.

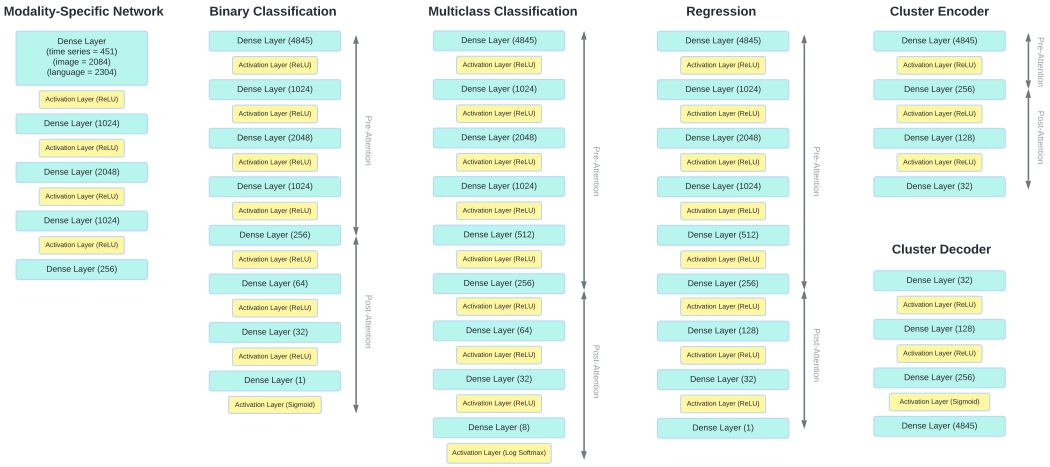
A Appendix / supplemental material

A.1 Computational Results

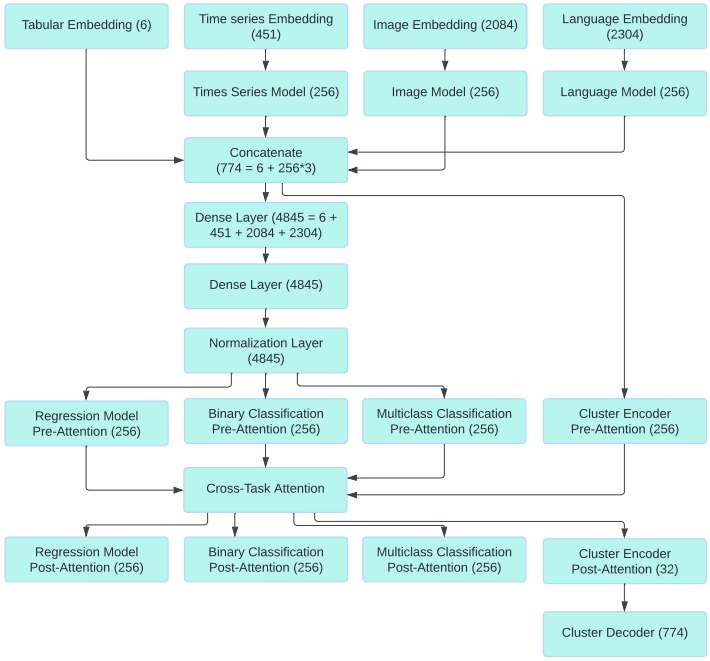


Supplemental Figure A1 Comparison of the Performance of Single-task and Multi-task Models Across Important Healthcare Tasks.

A.2 Architecture Details



Supplemental Figure A2-1. Modality-specific and task-specific network architectures.



Supplemental Figure A2-2. Overall Pipeline of the M3H Architecture.

A.3 Step-by-Step Data Integration and Modeling Procedure

Algorithm A3 End-to-End Data Integration and Modeling Pipeline

Input: Tabular data $X^{tabular}$, Time-series data $X^{time-series}$, Image data X^{vision}, Language data $X^{language}$, Feature extractor f_i of modality i, Outcome vector y_k for task $k \in \mathcal{K}$, $\hat{k} \in \hat{\mathcal{K}}$ indicates a set of task combinations. $\mathcal{L}_{\hat{k}}$ as the aggregated loss function of each task combination $\hat{k}$. $p \in \mathcal{P}$ is the set of hyperparameter combinations. $\epsilon = 10^{-6}$ to avoid numerical precision error during computation,
Output: Trained model and evaluation scores

Step 0 – Data pre-processing and cleaning

- Impute missing values for all modalities, where here x is a generic data entry:

$$x = \begin{cases} 0 & \text{if } x \text{ is numerical or image data} \\ \text{"" (empty string)} & \text{if } x \text{ is text data} \end{cases}$$

- Rescale image size:

$$X^{vision} \leftarrow \text{resize}(X^{vision}, 224 \times 224)$$

Step 1 – Embedding generation of each modality, an example of difference sources with the same modality is EKG notes vs. radiology notes:

$$E^i_j = f_i(X^i_j) \quad \forall i \in \{tabular, \textit{time-series}, vision, language\}, \\ \forall j \in \{\textit{different sources in each modality}\}$$

Step 2 – Concatenate embeddings of all sources of the same modality into a single flattened vector:

$$E^i = \text{vec}(E^i_1, E^i_2, \cdots, E^i_n) \quad \forall i \in \{tabular, \textit{time-series}, vision, language\}$$

Step 3 – Data Normalization

$$E^i = \frac{E^i - \text{mean}(E^i)}{\text{STD}(E^i) + \epsilon} \quad \forall i \in \{tabular, \textit{time-series}, vision, language\}$$

Step 4 – Structure input data with outcomes for a task combination

$$E_{\hat{k}} = \text{vec}(E^{tabular}, E^{\textit{time-series}}, E^{vision}, E^{language}, y_1, y_2, \cdots, y_k)$$

Step 5 – Model Training, Validation and Evaluation

For task combination $\hat{k}$ in the set of all prediction tasks $\hat{\mathcal{K}}$:

- Split data into train and test datasets with fixed seed.

$$E^{train}_{\hat{k}}, E^{test}_{\hat{k}}, y^{train}_{\hat{k}}, y^{test}_{\hat{k}} \leftarrow \textit{train\_test\_split}(E_{\hat{k}})$$

- Perform 5-fold cross-validation with grid search to select the best parameter combination $p^* \in \mathcal{P}$ on the training data that has the best cumulative performance across all tasks inside the task combination $\hat{k}$.

$$M3H^*_{\hat{k}} \leftarrow \underset{M3H_p \forall p \in \mathcal{P}}{\text{argmin}} \mathcal{L}_{\hat{k}}(M3H_p(E^{train}_{\hat{k}}), y^{train}_{\hat{k}})$$

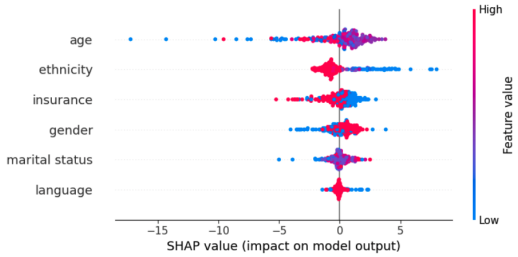
- Evaluate the optimal M3H model on the test set data:

$$\text{test\_set\_score} = M3H^*_{\hat{k}}(E^{test}_{\hat{k}}, y^{test}_{\hat{k}})$$

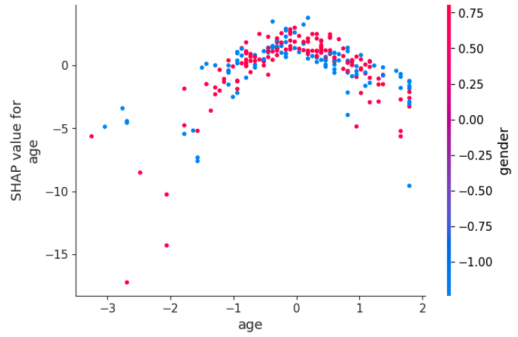
For each potential number of task combinations (i.e., single task = 1, 3-combined multitask = 3) and each task k, report the best model performance for each task.

A.4 Explainability of Input Space: by SHAP

We demonstrate below that by using SHAP values, we can effectively understand the magnitude and directionality of each input clinical variable's contribution to the outcome prediction and thus provide actionable insights for the physicians. Specifically, we analyze an M3H-framework- trained multi-task model between diabetes and heart failure and study the effects of tabular features on diabetes outcomes. We sampled 100 patients and studied their mean-standard deviation normalized tabular features using two types of analysis: feature importance and feature interaction. We observe that patients with lower age are less likely to have diabetes (blue dots for age have mostly negative SHAP values).



Supplemental Figure A4-1. SHAP feature importance plot: each dot indicates a single sample among the 100 test set samples. Higher values of the feature are indicated in red, and lower values in blue. The most important feature is ranked at the top, followed by other features. A higher SHAP value (right-hand side of the axis) indicates a higher likelihood of a positive outcome (has diabetes), and a lower SHAP value indicates a negative outcome (does not have diabetes).



Supplemental Figure A4-2. The SHAP interaction plot demonstrates the nonlinear interactions between features on the outcome prediction captured by the M3H model. Age impacts the risk of diabetes differently depending on the patient's gender.

A.5 Characteristics of HAIM-MIMIC-MM

A.5.1 Limitations:

HAIM-MIMIC-IV was developed from the MIMIC-IV database, and several inherent biases and limitations should be addressed. The cohort is collected from a single-care hospital in Boston and focuses on intensive-care unit patients. This could potentially restrict the demographics and clinical conditions of the patients to this specific geographical location and hospital setting. We also note that MIMIC-IV has recording errors, missing values, and other inconsistencies that are universal to all medical datasets and could pose a challenge for model development.

A.5.2 Embedding Dimensionality and Corresponding Clinical Variables:

The embeddings used as input data for M3H come from the multimodal database HAIM-MIMIC-MM, where the dimensionality of the features is explained and summarized in the paper's original supplemental tables 1 and 2, which are included below for reference. The size of time-series embedding is computed as the number of raw features multiplied by 11 unique features extracted: maximum, minimum, mean, variance, average piece-wise change over time, average absolute piece-wise change over time, maximum absolute piece-wise change over time, sum of absolute piece-wise change over time, change from end-beginning magnitude, number of peaks, and slope of the original time series sequence. There are three categories: chart event ($9 \times 11 = 99$ features), lab event ($22 \times 11 = 242$ features), and procedure event ($10 \times 11 = 110$ features). The size of note embedding comes from the output shape of the pre-trained model ClinicalBERT, which is 768. Similarly, the size of vision embeddings comes from the output shape of the pre-trained model Densenet121-res224-chex, which is 1024 (the dimension of the second to last layer of the model), and 18 (the output/last layer dimension).

A.5.3 Missing Data:

We also include here a table of the missing value distribution of the HAIM-MIMIC-MM dataset reported in the original paper (originally Supplemental Table 3) and how it was handled in that integration procedure.

#	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
9	Motor GCS response	Hemoglobin	Chest tube removed
10		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 A5-1. 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.

#	Data Modalities	#	Data Sources
1	Tabular	1	Demographics (E_{de})
2	Time-series	2	Chart events (E_{ce})
		3	Laboratory events (E_{le})
		4	Procedure events (E_{pe})
3	Text	5	Radiological notes (E_{radn})
		6	Electrocardiogram notes (E_{ecgn})
		7	Echocardiogram notes (E_{econ})
4	Images	8	Visual probabilities (E_{vp})
		9	Visual dense-layer feature (E_{vd})
		10	Aggregated visual probabilities (E_{vmp})
		11	Aggregated visual dense-layer features (E_{vmd})

Supplemental Table A5-2. List of different data modalities and data sources used to test the HAIM framework based on the MIMIC-IV-MM database. There are a total of four data modalities and eleven data sources. All data sources correspond to only one data modality. Thus, a model trained on a single data modality can have as little as 1 data source and many as 4 different data sources (of the same kind) as inputs. Double, triple and quadruple modality models can have a number of data sources ranging from [2 to 7], [3 to 9] and [4 to 11], respectively.

Feature Name	Missing %	Source	Handling
anchor_age	0.0	Demographics	N/A
gender_int	0.0	Demographics	N/A
ethnicity_int	0.0	Demographics	N/A
marital_status_int	0.0	Demographics	N/A
language_int	0.0	Demographics	N/A
insurance_int	0.0	Demographics	N/A
Foley Catheter	82.6	Procedure	Fill with 0
PICC Line	63.7	Procedure	Fill with 0
Intubation	75.3	Procedure	Fill with 0
Peritoneal Dialysis	99.7	Procedure	Fill with 0
Bronchoscopy	81.5	Procedure	Fill with 0
EEG	91.5	Procedure	Fill with 0
Dialysis - CRRT	93.1	Procedure	Fill with 0
Dialysis Catheter	88.9	Procedure	Fill with 0
Chest Tube Removed	93.1	Procedure	Fill with 0
Hemodialysis	92.9	Procedure	Fill with 0
Glucose	4.4	Lab	Fill with 0
Sodium	4.7	Lab	Fill with 0
Potassium	4.7	Lab	Fill with 0
Chloride	4.7	Lab	Fill with 0
Creatinine	4.7	Lab	Fill with 0
Urea Nitrogen	4.7	Lab	Fill with 0
Bicarbonate	4.7	Lab	Fill with 0
Anion Gap	4.7	Lab	Fill with 0
Hemoglobin	4.7	Lab	Fill with 0
Hematocrit	4.8	Lab	Fill with 0
Magnesium	5.4	Lab	Fill with 0
Platelet Count	9.8	Lab	Fill with 0

Feature Name	Missing %	Source	Handling
Phosphate	6.0	Lab	Fill with 0
White Blood Cells	4.9	Lab	Fill with 0
Calcium, Total	6.0	Lab	Fill with 0
MCH	4.9	Lab	Fill with 0
Red Blood Cells	4.9	Lab	Fill with 0
MCHC	4.9	Lab	Fill with 0
MCV	4.9	Lab	Fill with 0
RDW	4.9	Lab	Fill with 0
Neutrophils	36.9	Lab	Fill with 0
Vancomycin	60.0	Lab	Fill with 0
Heart Rate	19.5	Chart	Fill with 0
Non-Invasive Blood Pressure systolic	23.4	Chart	Fill with 0
Non-Invasive Blood Pressure diastolic	23.4	Chart	Fill with 0
Non-Invasive Blood Pressure mean	23.3	Chart	Fill with 0
Respiratory Rate	19.5	Chart	Fill with 0
O_2 saturation pulse oximetry	19.6	Chart	Fill with 0
GCS - Verbal Response	20.8	Chart	Fill with 0
GCS - Eye Opening	20.7	Chart	Fill with 0
GCS - Motor Response	20.8	Chart	Fill with 0
Electrocardiogram Notes	11.2	Notes	Empty String
Echocardiogram Notes	30.5	Notes	Empty String
Radiology Notes	0.1	Notes	Empty String

Supplemental Table A5-3. List of missing data percentages by individual variables and handling strategy. Individual variables (i.e., feature name) within key MIMIC-IV-MM data source groups are shown. The strategy for missing value handling used in our tests is as follows: 1) We exclude patients with no available X-rays from our selection cohort; 2) Time-series features are imputed with 0 if there is no measurement at any timestamp; 3) Text embeddings are generated from an empty string if there is no note available; 4) There were no missing values for demographics data.

A.6 Multitask Comparison

We implemented three methods using a universal problem setting of N tasks with feature dimension of d, with input features $X = \{x_j\}_{j=1}^{N}$ where $x_j \in \mathbb{R}^d$. We do not include reshaping operations or the batch size dimension in the description to capture only the mathematical essence of the implementations.

Multi-head attention:

- Initialize linear transformation matrices:
  - $W^Q, W^K, W^V \in \mathbb{R}^{d \times d}$ as query, key, value transformations
  - $W^O \in \mathbb{R}^{d \times d}$ as the output transformation
  - $H = 4$ as the number of heads
  - $d_H = d/H$ as the dimension per head
- Apply linear transformation and projection on input features:
  - $Q = XW^Q, K = XW^K, V = XW^V$
- Scaled dot-product to obtain attention weight ($\sqrt{d_h}$ is used to stabilize gradient):
  - $A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)$
- Apply attention weights to obtain output:
  - $O = (AV)W^O$

Cross-stitch:

- Initialize task interaction matrix:
  - $\{T_{ij}\}_{i \neq j}^{1:N}$ where $T_{ij} \in \mathbb{R}^{2 \times 2}$
- Apply interaction matrix:
  - $z_{ij} = T_{ij} \cdot [x_i, x_j] \quad \forall (i, j)$
- Aggregation:
  - $z_i = \sum_{j=1}^{N} z_{ij} \quad \forall\, i = 1, \ldots, N$
- Output learned features:
  - $\{z_i\}_{i=1}^{N} \quad \forall\, i = 1, \ldots, N$

  For n tasks, this requires $\frac{n(n-1)}{2}$ weight matrices of size 2×2.

Multilinear relationship network (MRN):

- Initialize linear transformation matrices:
  - $\{T_{ij}\}_{i,j=1}^{N}$ where $T_{ij} \in \mathbb{R}^{d \times d}$
- Apply linear transformation and projection on input features:
  - $z_{ij} = T_{ij} \cdot [x_i, x_j] \quad \forall (i, j)$
- Aggregation:
  - $z_i = \sum_{j=1}^{N} z_{ij} \quad \forall\, i = 1, \ldots, N$
- Output learned features:
  - $\{z_i\}_{i=1}^{N} \quad \forall\, i = 1, \ldots, N$

  For n tasks, this requires $\frac{n^2}{2}$ weight matrices of size $d \times d$.

Specifically, we conduct experiments in the original dataset on 10 different combinations of multi-tasks that comprehensively evaluate multitask strategies across all four types of machine learning problem classes. The choice of diabetes and heart failure is arbitrary.

- Length of stay (regression), patient phenotyping (clustering)

- Length of stay (regression), thoracic testing (multiclass classification)
- Thoracic testing (multiclass classification), patient phenotyping (clustering)
- Diabetes (binary classification), length of stay (regression)
- Diabetes (binary classification), patient phenotyping (clustering)
- Diabetes (binary classification), thoracic testing (multiclass classification)
- Heart failure (binary classification), length of stay (regression)
- Heart failure (binary classification), patient phenotyping (clustering)
- Heart failure (binary classification), thoracic testing (multiclass classification)
- Diabetes (binary classification), Heart failure (binary classification)

We observe that cross-task attention has a clear advantage in the majority of the cases across all three strategies, with cross-stitch being a close competitor in these 2-tasks experiments (but with qualitative disadvantages discussed below).

Machine Learning Problem Class	**Cross-task (M3H)**	**Multi-Head Attention**	**Multilinear Relationship Network**	**Cross Stitch**
Regression	0.567	0.562 (-0.88%)	0.431 (-23.99%)	0.565 (-0.35%)
Clustering	0.405	0.521 (+28.64%)	0.176 (-56.54%)	0.487 (+20.25%)
Multiclass	0.755	0.715 (-5.30%)	0.595 (-21.19%)	0.755 (+0%)
Binary (diabetes)	0.873	0.824 (-5.61%)	0.873 (+0%)	0.869 (-0.46%)
Binary (heart failure)	0.881	0.864 (-1.93%)	0.896 (+1.7%)	0.888 (+0.79%)

Supplemental Table A6. Comparison of machine learning problem classes across different models. The values represent performance metrics and percentage differences from the baseline (Cross-task M3H).

Beyond the quantitative advantage of the proposed cross-task framework, we would also like to emphasize the qualitative advantage of the chosen framework over existing methods:

- **Interpretability**: Available multimodal multi-task foundation models heavily rely on complex architectures, for example, with repeated use of multi-head attention mechanisms tens or hundreds of times to achieve good performance guarantees. Even with known visualization efforts to interpret these architectures, in practice, these attention weights are almost very often not interpretable and non-sensible. This is why we opted for such a model structure design. As reviewer 2 later correctly pointed out, the existing style of complex architecture makes it very difficult to obtain clinician trust in hospital settings precisely because of such lack of interpretability. Instead, in our case, we apply a single cross-task attention with one single channel and a clean 2D attention weight to explicitly model how self-attention and cross-attention interact. Such design allows for future analysis of interpretability a lot more easily.
- **Lightweight design for deployment**: Existing architectures, such as Google's Med-PaLM 2 (released March 2023), contain 540 billion parameters and can be estimated usually to need months to train with commercial-grade GPUs (such as Nvidia Volta V100) with heavy RAM memory requirements (at least 1000GB if not parallelized). Although lighter-weight versions of these models exist, they remain in the billion-level parameters and pose a significant implementation challenge for hospitals if they wish to host in-house models for data privacy reasons. M3H, on the other hand, can be offered as a much lighter solution to avoid these issues.

Similarly, all three of the compared multiclass methods require significantly more complex network structures. Multi-head model (in our case with 4 heads) requires 4 additional channels to integrate the data from separate heads; cross-stitch models would require significantly more weight matrices as the number of co-learned tasks increases, MRN models will require even more parameters, as they require a linear transformation of each combination of task pairs.

A.7 Loss Function Definition

Binary cross entropy loss (Binary classification):
Given $x \in \mathbb{R}^d$ as an input feature of dimension d, $y \in \{0,1\}^d$ as the binary outcomes, $\hat{y} = \sigma(w^T x + b)$ is the predicted outcome from the M3H framework. Here $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function, w is the weight matrix, b is the bias vector, the loss function is defined as: $l_{\text{binary}}(y, \hat{y}) = -(y \log(\hat{y}) + (1-y)\log(1-\hat{y}))$.

Negative log-likelihood loss (Multiclass classification):
Given $x \in \mathbb{R}^d$ as an input feature of dimension d, $y \in \{1, 2, ..., K\}^d$ as the multiclass outcomes from K classes, $\hat{y} = \sigma(w^T x + b)$ is the predicted outcome from the M3H framework. Here: $\sigma(z) = z - \log\left(\sum_{k=1}^{K} e^{z_k}\right)$ is the log-softmax function, w is the weight matrix, b is the bias vector, the loss function is defined as: $l_{\text{multiclass}}(y, \hat{y}) = -\log(\hat{y})$.

Mean absolute error (Regression):
Given $x \in \mathbb{R}^d$ as an input feature of dimension d, $y \in \mathbb{R}^d$ as the regression outcomes, $\hat{y} = w^T x + b$ is the predicted outcome from the M3H framework. Here w is the weight matrix, b is the bias vector, the loss function is defined as: $l_{\text{regression}}(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|$.

Mean squared error (Clustering):
Given $x \in \mathbb{R}^d$ as an encoder input of dimension d, $\hat{x} \in \mathbb{R}^d$ as the decoder output of the same dimension, here w is the weight matrix, b is the bias vector, the loss function is defined as: $l_{\text{clustering}}(x, \hat{x}) = \frac{1}{d}\sum_{i=1}^{d} (\hat{x}_i - x_i)^2$.

Contrastive Learning:
This learning aims to project embeddings of different modalities into the same embedding space by contrasting positive pairs (modalities from the same samples) and negative pairs (modalities from dissimilar samples). In the M3H framework, because of the small dimension of the tabular features (6) in comparison to the rest of the three modalities, we only apply contrastive learning among time series, vision, and language data inputs. The formulation is as follows:

Given $\hat{N}$ as the number of permutations between the N samples' three modalities (or N choose 2), and given E_i and E_j as pairs of embedding vectors from different modalities, y_i as the label for the pair of (i, j), where they are either from the same sample (1) or different samples (0). We define a positive margin $p = 0$, and a negative margin $n = 1$. Specifically, for positive pairs, the loss is only computed if $|E_i - E_j| > p$, which aims to decrease positive pairs' distance to 0, and for negative pairs, we only compute the loss when $|E_i - E_j| < n$, which aims to push the distance to be close to 1. The contrastive loss is computed as follows: $L = \frac{1}{N}\sum_{i=1}^{N} \left(y_i \max(0, |E_i - E_j| - p)^2 + (1-y_i)\max(0, n - |E_i - E_j|)^2\right)$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have provided detailed outline of how experiments are conducted and the improvements of our model in comparison to the nominal single-task models both in the introduction and abstract. We have also highlighted key findings and technical novelties introduced in the later sections as well.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a limitation section at the end of the paper detailing the several possibilities for limitations, ranging from data, to inclusion of other tasks. We have also provided a practical implication section to reflect rigorously how the framework should be adopted in new data and system settings, and what are the potential remedies to deal of potential challenges.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

   Justification: Our paper is focused on model architecture design and thus does not have theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper clearly outlined the architectures, parameters, data cohort availability, processing steps to ensure that all necessary data is needed for the reproduction of the computational results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Unfortuantely due to privacy reasons we are unable to release the code. But readers are encouraged to contact the authors to its access.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper outlined all the necessary details for where to obtain the data cohort, how to preprocess the dataset, how to split the train, validation, and test sets, the hyperparameters chosen in the paper in order to reproduce all the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars and confidence intervals are provided for the main computational results in the supplemental figure. Details on how these results are computed are also included in the methods.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes] ,

   Justification: The paper's section on practical implementation details the computational resources needed to run the model, as well as an estimated time to run on a local computer. This is also accompanied by the operating system details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] ,

   Justification: The content of this paper complies with NeuRIPS code of ethics, and was conducted with the hopes to advance our understanding of medicine to better patient care.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: In the implication to practice section, we discuss thoroughly both the negative and positive implications of the introduced model and its impact if applied to hospital systems.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: In the practical implementation section, we discuss how the model should be validated and evaluated prior to its implementation. This includes safeguards against for example data perturbations to ensure that the model does not negatively impact patient outcome predictions.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The data and models used for this paper are properly introduced, elaborated, and cited by the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes] .

Justification: The provided new model has been thoroughly outlined by the paper with detailed instructions on its training parameters and architectures, data used, as well as evaluations.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes] .

Justification: This study is conducted using a licensed, but publicly available dataset of human subjects. The details of this cohort has been thoroughly discussed in the patient cohort section.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA] .

Justification: IRB approval was not needed due to the use of a public national database.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.