MIND: Modality-Informed Knowledge Distillation Framework for Multimodal Clinical Prediction Tasks

Anonymous authors
Paper under double-blind review

Abstract

Multimodal learning leverages information across modalities to learn better feature representations for improved performance in fusion-based tasks. However, multimodal datasets, especially in medical settings, are typically smaller in size than their unimodal counterparts, which typically impedes the performance of multimodal models. The increase in number of modalities is often associated with an overall increase in the size of the multimodal network, which may be undesirable in medical use-cases. Alternatively, utilizing smaller unimodal encoders may lead to sub-optimal performance, especially in dealing with high-dimensional clinical data. In this paper, we propose the Modality-INformed knowledge Distillation (MIND) framework, a multimodal model compression framework based on knowledge distillation that transfers knowledge from ensembles of pre-trained deep neural networks of varying sizes into a smaller multimodal student. The teacher models consist of unimodal networks, allowing the student to learn diverse representations. MIND involves multi-head joint fusion models, compared to single-head models, thereby enabling the utilization of the unimodal encoders in case of missing modalities. As a result, MIND generates an optimized multimodal model, enhancing both multimodal and unimodal representations. It can also be leveraged to balance multimodal learning during training. We evaluate MIND on binary classification and multilabel clinical prediction tasks using clinical time series data and chest X-ray images extracted from publicly available datasets. In addition, we assess the generalizability of the MIND framework on three multimodal multiclass benchmark datasets. The experimental results demonstrate that MIND improves the performance of the smaller multimodal network across all five tasks, as well as fusion methods and multimodal network architectures, with respect to several state-of-the-art baselines.

1 Introduction

Healthcare decision-making, like most human-based active reasoning (Smith & Gasser, 2005), is multimodal. This means that clinical decisions are typically motivated by several sources of information or modalities, such as diagnostic imaging, clinical time-series data, or medical history, combined with clinical experience and expertise (Bate et al., 2012; Pauker & Kassirer, 1980; Pines et al., 2023). As a result, recent work has shown the promise of multimodal learning in modeling various clinical prediction tasks (Huang et al., 2020), especially in information fusion to improve downstream classification performance compared to relying on a unimodal network (Hayat et al., 2022; Yang et al., 2022; Wang et al., 2020a).

Despite recent advancements, the nature of the learning process in multimodal clinical prediction tasks remains unclear. Most work primarily focuses on achieving the best fusion performance (Huang et al., 2020), ignoring additional challenges of multimodal training such as optimizing modality utilization and modeling cross-modal interactions for effective training (Huang et al., 2022b; Wang et al., 2020b; Wu et al., 2022). For example, recent results show that multimodal deep neural networks are prone to overfitting due to their larger capacity (Wang et al., 2020b). They learn to rely primarily on one of the input modalities available, i.e., the fastest to learn from (Wu et al., 2022), and as different modalities may generalize and overfit at different rates, performing naive joint training typically leads to sub-optimal results (Wang et al., 2020b).

The methodological challenges of multimodal learning are coupled with other challenges when applied in healthcare. This includes the limited availability of large-scale multimodal clinical datasets as well as the sparse and heterogeneous nature of the input modalities, such as when fusing medical images and clinical time-series data. The increase in the number and heterogeneity of modalities entails the specification of modality-specific encoders, such as convolutional neural networks for images and recurrent neural networks for time-series data (Hayat et al., 2022). This increases the size of the overall multimodal network, which may hinder deployment for clinically relevant use cases such as privacy-preserving training (Baruch et al., 2022), secure inference (Lou et al., 2021) and resource-constrained devices (Zhao et al., 2018).

Our objective in this work is to improve the compression of multimodal networks, both in terms of size and predictive performance when dealing with small multimodal datasets. We study this in the context of joint fusion, where there are two modalities and their latent representations are aggregated before applying a fusion layer. To achieve this objective, we propose a training framework based on knowledge distillation that incurs minimal modifications to the joint fusion architecture and learning objective while significantly enhancing its predictive performance and reducing its size. We refer to this training framework as *Modality-INformed knowledge Distillation* (MIND). Specifically, we make the following contributions:

- We propose incorporating modality-specific supervision signals within the joint fusion architecture. This approach enhances unimodal encoder representations by minimizing the global model while focusing on both unimodal encoders, leading to better fusion (multimodal) performance.
- We propose to pre-train unimodal teachers and use them to distill knowledge into modality-specific signals, compressing the knowledge of larger models (or an ensemble of smaller ones) and enhancing the representations learned by the student model's modality encoders. We show that this approach significantly improves fusion performance with a more compact student network.
- We add additional weighting parameters to emphasize distillation learning. We empirically show that it improves both unimodal and fusion performance and that it also helps balance modality learning during multimodal training.

2 Background

Knowledge distillation. Knowledge distillation (KD) (Hinton et al., 2015), also referred to as model compression (Buciluă et al., 2006) or teacher-student networks (Gou et al., 2021), allows transferring the generalization capability of a large model into a typically smaller model. The most common technique is response-based distillation, which leverages the output class probabilities of the larger model as soft targets for training the smaller model (Hinton et al., 2015). In this offline setting, the aim of the student network is to mimic the performance of the teacher network by approaching the softened responses of the pre-trained teacher network. The main rationale is that the function learned by a large model can be approximated by a shallower model, which is computationally less expensive to train (Gou et al., 2021). Although increased network depth can improve learning, it is not always needed or desirable (Ba & Caruana, 2014). Other approaches involve the approximation of models' features and layers or instance relationships in online or self-distillation settings (Gou et al., 2021).

KD aims to transfer the knowledge and generalization of a large model to a smaller model (Hinton et al., 2015), usually via student-teacher networks (Gou et al., 2021). In a multiclass classification problem, the student network mimics the predictions of the teacher model during training (Gou et al., 2021) by minimizing:

$$\mathcal{L} = \underbrace{\mathcal{L}_S(y, p(z_s, \tau = 1))}_{\text{supervised learning}} + \underbrace{\mathcal{L}_{KD}(p(z_t, \tau), p(z_s, \tau))}_{\text{knowledge distillation}} \quad (1)$$

where $\mathcal{L}_S$ is the supervised learning loss applied to the student network (Cross-Entropy (CE) loss), z_s are the *logits* of the student, τ is a *temperature* parameter that regulates the importance of each target class, and p is the output probability obtained after applying softmax activation. $\mathcal{L}_{KD}$ is the KD loss, which quantifies the divergence between the teacher (z_t) and student outputs (z_s), either using the Kullback-Leibler divergence loss (KL) or CE. In a multiclass setting, $\tau = 1$ in the supervised loss directly obtains $\hat{y}$, while $\tau > 1$ KD generates the *soft targets* for the distillation loss.

Multimodal learning & knowledge distillation. Recent work leveraged KD for a variety of multimodal learning tasks, such as for generating a multimodal student from a unimodal teacher (Xue et al., 2021), addressing the limitations of CLIP (Dai et al., 2022), multimodal generation via cross-modal vision-language KD (Radford et al., 2021), and optimized action recognition using multimodal sensors via wavelet KD (Quan et al., 2023). Other studies focused on handling missing modalities and improving computational efficiency by shifting from multimodal to unimodal attention (Agarwal et al., 2021), or improving unlabeled image selection from a pre-trained vision-language model (Zhang et al., 2022). The findings of previous work highlight the versatility and effectiveness of KD in improving the performance of various multimodal learning tasks. In our work, we focus on KD in the context of multimodal fusion networks, leveraging it to compress knowledge from an ensemble of teachers.

Knowledge distillation in healthcare. KD has been utilized to compress deep neural networks for a wide range of clinical prediction tasks. For unimodal tasks, several studies explored KD for federated learning (Chen et al., 2022) and multi-institution collaboration (Huang et al., 2023), medical text classification (Huang et al., 2022a), electronic health record mining (Du & Hu, 2020), and multiclass classification for medical images (Yang et al., 2022; Ho & Gwak, 2020; Soni et al., 2019). The authors in Wang et al. (2020a); Dou et al. (2020) explored KD in the context of multimodal imaging data for multiclass classification, and multiclass cardiac structure and multi-organ segmentation, respectively. In another study (Dou et al., 2020), the authors proposed to learn shared representations among imaging modalities to distill knowledge for modality-specific segmentation, while in Wang et al. (2020a) they distilled the knowledge of a unimodal teacher into a multimodal student. Another work (Hu et al., 2020) investigated distilling knowledge from a multimodal teacher to a unimodal student network for clinical image segmentation. In general, the use of KD for multimodal clinical prediction tasks, and for multilabel classification in particular — a relevant problem for healthcare applications where multiple labels may be simultaneously present in a sample — has been under-explored. By contrast, our work investigates the application of KD for model compression on multimodal networks for heterogeneous medical modalities focusing on binary and multilabel classification tasks.

Training of multimodal networks. Joint training of multimodal fusion networks is a challenging task due to the tendency of these networks to overfit one of the input modalities (Wang et al., 2020b). As shown in recent work (Huang et al., 2022b), the different input modalities compete against each other during joint training, leading to sub-optimal joint training in which only a subset of the input modalities are learned efficiently while the rest remain unexplored and have a minor contribution (Wu et al., 2022). Several techniques have been proposed to generate a more balanced multimodal learning training (Huang et al., 2022a; Wu et al., 2022; Wang et al., 2020b; Peng et al., 2022) based on metrics to characterize the degree of modality overfitting (Wang et al., 2020b; Wu et al., 2022). In this work, we leverage the conditional utilization rate (Wu et al., 2022) to show that, as a by-product of our proposed framework, the weights of the distillation loss can be employed to emphasize learning from specific modalities, thus improving modality utilization and leading to a more balanced multimodal learning training. This enhances the multimodal and unimodal predictions of the model trained via the MIND framework.

3 Methods

Preliminaries. In a simple multimodal fusion task, we assume the presence of two input modalities represented by $\mathbf{x}_A$ and $\mathbf{x}_B$. The goal of the multimodal fusion task is to jointly learn from both modalities to predict a set of ground-truth labels denoted by $\mathbf{y}$. Due to heterogeneity, each modality is first processed by modality-specific encoders, such that $\mathbf{z}_A = f_A(\mathbf{x}_A)$ and $\mathbf{z}_B = f_B(\mathbf{x}_B)$, where $\mathbf{z}_A$ and $\mathbf{z}_B$ are latent feature representations. The latent representations are concatenated and processed by a fusion classification layer, such that $\hat{\mathbf{y}}_{AB} = g_{AB}(\mathbf{z}_A, \mathbf{z}_B)$. We then apply the binary cross-entropy (BCE) loss, denoted by $\mathcal{L}_{S_{AB}}(\mathbf{y}, \hat{\mathbf{y}}_{AB})$, and optimize f_A, f_B, and g_{AB} jointly, as shown in Figure 1 A.1. In this work, we focus on joint fusion, such that both encoders are trained from scratch (randomly initialized), although previous work primarily focuses

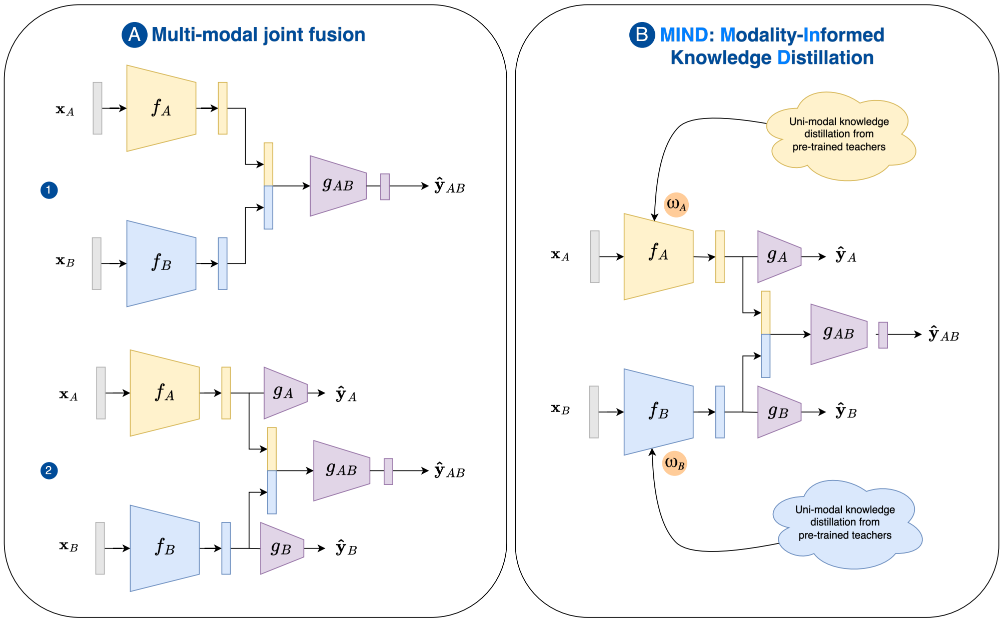
A Multi-modal joint fusion
1
2
$\mathbf{x}_A$
$\mathbf{x}_B$
f_A
f_B
g_{AB}
g_A
g_B
$\hat{\mathbf{y}}_{AB}$
$\hat{\mathbf{y}}_A$
$\hat{\mathbf{y}}_B$
B MIND: Modality-Informed Knowledge Distillation
Uni-modal knowledge distillation from pre-trained teachers
ω_A
ω_B
Uni-modal knowledge distillation from pre-trained teachers

Figure 1: **Architecture of multimodal fusion network.** A.1 Architecture for multimodal joint fusion as in recent work Hayat et al. (2022). A.2 Modified base architecture that incorporates modality-specific classification heads with the loss shown in Equation 2. B. Architecture of the MIND framework with the loss shown in Equation 5, which incorporates unimodal pre-trained teachers for KD to the unimodal encoders and modality weighting hyper-parameters (ω_A, ω_B) to emphasize encoder representation learning and improve modality utilization during multimodal training training .

on fine-tuning pre-trained unimodal encoders (Huang et al., 2020). We believe that learning with randomly initialized encoders is more comparable in the proposed setup.

Multimodal loss. We incorporate two additional loss terms, by introducing classification modules for each modality, such that $\hat{\mathbf{y}}_A = g_A(\mathbf{z}_A)$ and $\hat{\mathbf{y}}_B = g_B(\mathbf{z}_B)$. Hence, the overall supervised learning loss becomes:

$$\mathcal{L}_S = \mathcal{L}_{S_{AB}}(\mathbf{y}, \hat{\mathbf{y}}_{fusion}) + \mathcal{L}_{S_A}(\mathbf{y}, \hat{\mathbf{y}}_A) + \mathcal{L}_{S_B}(\mathbf{y}, \hat{\mathbf{y}}_B) \tag{2}$$

This modification enables the use of the unimodal classification heads independently in case of missing modalities at inference, as shown in Figure 1 A.2.

Unimodal teachers. We use KD to compress the knowledge of an ensemble of unimodal teacher models into a single and typically smaller student model. KD was originally conceived for multiclass classification problems, where different values of τ are used to generate the *soft targets*. In our work, we propose to use KD for a wider range of predictive tasks including binary and multilabel classification tasks. For multilabel classification, we modify the $\mathcal{L}_{KD}$ term in Equation 1 to:

$$\mathcal{L}_{\text{KD}}(\hat{\mathbf{y}}^s, \hat{\mathbf{y}}^t) = \frac{1}{N} \sum_{i=1}^{L} BCE(\hat{y}^{s,i}, \hat{y}^{t,i}), \tag{3}$$

where L is the set of possible labels in a multilabel classification problem, N is the sample size, and s and t denote the student and teacher networks, respectively.

Given an ensemble T of K teachers, we define a new average ensemble prediction for an i-th label (omitting i without loss of generality):

$$\hat{y}^T = \frac{1}{K} \sum_{k=1}^{K} \hat{y}^{t_k}, \tag{4}$$

such that the Ensemble Knowledge Distillation (EKD) loss is now denoted as $\mathcal{L}_{EKD}(\hat{\mathbf{y}}^s, \hat{\mathbf{y}}^T)$.

In a comprehensive multimodal setting, the goal of the classification model is to accurately classify multimodal but also unimodal samples, which is particularly relevant in healthcare applications where some modalities may be missing. To improve multimodal and unimodal encoder representations we introduce $\mathcal{L}_{EKD^U}$, which consists of an ensemble of unimodal teachers trained with sub-components of Equation 2 for each modality, i.e. $\mathcal{L}_{S_A}$ and $\mathcal{L}_{S_B}$. Hence, we introduce a new learning objective term per input modality to distill knowledge from unimodal teachers to the modality encoders of a smaller multimodal student. In the case of two modalities (A and B), two terms are introduced:

$$\underbrace{\mathcal{L}_{EKD^U_A}}_{\substack{\text{unimodal ensemble}\\ \text{knowledge distillation}\\ \text{for modality A}}} \quad , \quad \underbrace{\mathcal{L}_{EKD^U_B}}_{\substack{\text{unimodal ensemble}\\ \text{knowledge distillation}\\ \text{for modality B}}} \tag{5}$$

We note that goal of the multimodal student network is to learn the distribution $P(Y|X_A, X_B)$. To improve encoder representation and better learning of the multimodal distribution, we introduce the unimodal EKD components consisting of $\mathcal{L}_{EKD^U_A}$ and $\mathcal{L}_{EKD^U_B}$, which represent the distributions $P(Y|X_A)$ and $P(Y|X_B)$, respectively, learned by the unimodal teachers.

Characterization of modality overfit. The joint fusion training of multimodal networks is a challenging task in which the joint model shows excessive reliance on one of the modalities (modality overfitting). We adopt the *conditional utilization rate* ($\mathbf{u}$) (Wu et al., 2022) to characterize modality usage in multimodal deep neural networks. We modify $\mathbf{u}$ for joint fusion multimodal neural networks as follows:

$$\mathbf{u_A} = \frac{A(\hat{\mathbf{y}}_{AB}) - A(\hat{\mathbf{y}}_B)}{A(\hat{\mathbf{y}}_{AB})}, \mathbf{u_B} = \frac{A(\hat{\mathbf{y}}_{AB}) - A(\hat{\mathbf{y}}_A)}{A(\hat{\mathbf{y}}_{AB})}, \tag{6}$$

where $A(\cdot)$ denotes a classification accuracy metric, $\mathbf{u_A}$ computes the conditional utilization rate for modality A, and $\mathbf{u_B}$ for modality B. The conditional utilization rate measures the marginal contribution of each modality to the fusion model. Following Wu et al. (2022), d_{util} is defined as the difference between conditional utilization rates, $d_{util} = \mathbf{u_A} - \mathbf{u_B}$, allowing for the assessment of imbalanced modality usage by the multimodal fusion model. Specifically, $d_{util} \in \mathbb{R}\ s.t. -1 \leq d_{util} \leq 1$, with extreme values indicating imbalanced modality usage.

Balancing multimodal training. After identifying modality overfitting in the joint fusion model via d_{util}, we introduce weighting hyperparameters ω_A and ω_B to adjust the focus on distillation components, enhancing encoder representation and modality learning with the overall loss as:

$$\mathcal{L}_{\text{MIND}} = \underbrace{\mathcal{L}_{S_{AB}} + \mathcal{L}_{S_A} + \mathcal{L}_{S_B}}_{\text{supervision signals}} + \underbrace{\omega_A \times \mathcal{L}_{EKD^U_A} + \omega_B \times \mathcal{L}_{EKD^U_B}}_{\substack{\text{weighted unimodal ensemble}\\ \text{knowledge distillation}}} \tag{7}$$

A visual representation of the proposed MIND framework is provided in Figure 1 B. In particular, setting $\omega_A, \omega_B > 1$ minimizes both loss components (supervision signals and EKD), emphasizing distillation for unimodal encoder learning. Specifically, $\omega_A, \omega_B \gg 1$ prioritize knowledge distillation, while $\omega_A, \omega_B = 1$ treat all loss components equally. $\omega_A \gg \omega_B$ allows the learning to focus on modality A, while $\omega_A \ll \omega_B$ prioritizes modality B. This independent weighting can be leveraged to balance multimodal learning by emphasizing (larger weight) the less utilized modality. Setting $\omega_A, \omega_B = 0$ turns off distillation, reverting the learning objective to Equation 2.

Overall, the MIND framework produces a smaller, optimized version of the original multimodal model. This compressed model can make predictions with both multimodal (both modalities present) and unimodal (single modality) inputs. Through the introduced weighted unimodal ensemble knowledge distillation, the MIND framework enhances and balances modality representation learning, significantly improving predictive performance for both multimodal and unimodal inputs. We provide a pseudo-code implementation of the

MIND framework for the training of an enhanced, compact multimodal student network based on the MIND framework for two modalities in Algorithm 1.

Algorithm 1 MIND Framework

```
Require: Multimodal training dataset $\mathcal{D}_{AB_{train}}$, Multimodal validation dataset $\mathcal{D}_{AB_{val}}$, Modality A training dataset $\mathcal{D}_{A_{train}}$, Modality A validation dataset $\mathcal{D}_{A_{val}}$, Modality B training dataset $\mathcal{D}_{B_{train}}$, Modality A training validation $\mathcal{D}_{B_{train}}$, Multimodal Model $M_{AB}$, Loss function $\mathcal{L}_{\text{MIND}}$, Number of epochs $N$, Optimizer $O$, Weighting coefficients $w_A$, $w_B$
Ensure: Trained multimodal model $M_{AB}$
  /* Train $T$ unimodal teacher models per modality */
  /* Create ensemble of unimodal teachers per modality */
  $E_A = [T_0^A, ..., T_T^A]$
  $E_B = [T_0^B, ..., T_T^B]$
  /* Train multimodal student model $M_{AB}$ */
  Initialize model parameters and add classification heads to modality encoders ($M_A$, $M_B$)
  for epoch = 1 to $N$ do
      Shuffle the training data $\mathcal{D}_{AB_{train}}$
      for minibatch $b$ in $\mathcal{D}_{AB_{train}}$ do
          Forward pass: Compute predictions with all classification heads $\hat{y}_{AB}$, $\hat{y}_A$, $\hat{y}_B$, and teacher ensembles $\hat{y}_{E_A}$, $\hat{y}_{E_B}$ for minibatch $b$ using models $M$, $M_A$, $M_B$, $E_A$ and $E_B$
          Compute loss components $\mathcal{L}_{S_{AB}}(\hat{y}_{AB}, y)$, $\mathcal{L}_{S_A}(\hat{y}_A, y)$, $\mathcal{L}_{S_B}(\hat{y}_B, y)$, $\mathcal{L}_{EKD^U_A}(\hat{y}_A, \hat{y}_{E_A})$ and $\mathcal{L}_{EKD^U_B}(\hat{y}_B, \hat{y}_{E_B})$ for minibatch $b$ following Equation 2 and Equation 4
          Compute $\mathcal{L}_{\text{MIND}}(\mathcal{L}_{S_{AB}}, \mathcal{L}_{S_A}, \mathcal{L}_{S_B}, \mathcal{L}_{EKD^U_A}, \mathcal{L}_{EKD^U_B}, w_A, w_B)$ for minibatch $b$ following Equation 7
          Backward pass: Compute gradients of the loss with respect to model $M_{AB}$ parameters
          Update model parameters using optimizer $O$
      end for
  end for=0
```

4 Experiments

Datasets. To evaluate our approach, we focus on two clinical prediction tasks: clinical conditions prediction ($L = 25$, Equation 3), and in-hospital mortality prediction after a 48-hour ICU stay ($L = 1$, Equation 3), using chest X-ray images and clinical time-series data. We use chest X-ray images from MIMIC-CXR (Johnson et al., 2019), where each image ($\mathbf{x}_A$) is replicated across three channels, yielding $\mathbf{x}_A \in \mathbb{R}^{224\times224\times3}$. Associated clinical time-series data ($\mathbf{x}_B$) is obtained from MIMIC-IV (Johnson et al., 2023), with dimensions $\mathbf{x}_B \in \mathbb{R}^{t\times76}$, where t represents the number of time-steps based on the patient's ICU stay duration, and 76 denotes the number of pre-processed features per time-step. We follow the dataset splits and multimodal architecture from Hayat et al. (2022), where f_A is parameterized by a ResNet-34, f_B is parameterized as a two-layer LSTM and the fusion model g_{AB} is parameterized as an LSTM layer, referred to as MedFuse. We use the multimodal dataset for training the MIND framework and baselines, containing samples with both modalities present. For the clinical conditions task, the dataset comprises 7,728 training, 877 validation, and 2,161 test samples. For in-hospital mortality prediction, it consists of 4,885 training, 540 validation, and 1,373 test samples. For training unimodal models, we utilize all available data per modality. Specifically, the CXR dataset comprises 124,671 training, 15,282 validation, and 36,625 test samples, while the EHR dataset contains 42,628 training, 4,802 validation, and 11,914 test samples.

Ensembles. In the MIND framework, we first pre-train the unimodal teachers with available unimodal datasets. To ensure diversity, we train multiple models for each modality following the architectures in Hayat et al. (2022). For chest X-ray, we train ResNet-34, ResNet-18, and ResNet-10 models. Similarly, for clinical time-series data, we train 2-layer, 3-layer, and 4-layer LSTM networks. The top three performing models from each modality are then combined to create ensembles. These ensembles of unimodal teachers are utilized to distill knowledge, as described in Equation 7, into a smaller randomly initialized multimodal student model. This student model consists of a ResNet-10 and a 2-layer LSTM serving as encoders for chest X-ray and time-series data, respectively.

Table 1: **Multimodal fusion performance results.** Performance results on the multimodal test set of the MIND model and all baselines (MedFuse, MedFuse-3H, TS, MKE-CXR, MKE-EHR, and UME). Best results are shown in bold.

Model	Clinical Conditions		In-hospital Mortality	
	AUROC	**AUPRC**	**AUROC**	**AUPRC**
MedFuse (Hayat et al., 2022)	0.748 (0.721, 0.774)	0.447 (0.402, 0.495)	0.816 (0.785, 0.847)	0.468 (0.398, 0.546)
MedFuse-3H	0.750 (0.724, 0.777)	0.452 (0.410, 0.500)	0.820 (0.787, 0.851)	0.461 (0.392, 0.541)
MKE-CXR (Xue et al., 2021)	0.711 (0.681, 0.739)	0.412 (0.372, 0.458)	0.707 (0.667, 0.747)	0.308 (0.255, 0.377)
MKE-EHR (Xue et al., 2021)	0.744 (0.717, 0.770)	0.447 (0.405, 0.495)	0.827 (0.795, 0.856)	0.491 (0.422, 0.567)
UME (Du et al., 2023)	0.767 (0.741, 0.793)	0.489 (0.450, 0.544)	0.818 (0.787, 0.849)	0.491 (0.419, 0.569)
TS (Wang et al., 2020a)	0.768 (0.742, 0.794)	0.483 (0.439, 0.532)	0.828 (0.797, 0.857)	0.483 (0.414, 0.555)
MIND (Ours)	**0.782** (0.757, 0.807)	**0.506** (0.460, 0.556)	**0.844** (0.815, 0.872)	**0.505** (0.433, 0.587)

Baselines. We compare the performance of MIND with the original multimodal model, MedFuse, trained with randomly initialized encoders. The MIND model is three times smaller (in terms of learnable parameters) than MedFuse, as shown in Table A1. Further, we modify the original MedFuse architecture to adopt the loss introduced in Equation 2, denoted as MedFuse-3H. We also compare MIND to other related KD baselines, including TS (Wang et al., 2020a), MKE (Xue et al., 2021), and UME (Du et al., 2023). For MKE, we evaluate two models: MKE-CXR and MKE-EHR. Further details about the baseline models can be found in Appendix A.4.

Hyperparameter Tuning. All models, including baselines, undergo training for a maximum of 300 epochs with early stopping (40 epochs). We use a batch size of 16 and the Adam optimizer across all experiments. Hyperparameter tuning involves optimizing the learning rate and weighting parameters for MIND and all baselines. Further details on hyperparameter tuning are provided in Appendix A. We select the best models based on the checkpoint yielding the highest Area Under the Receiving Operating Characteristic (AUROC) on the validation set and present the results on the test set in terms of AUROC and Area Under the Precision-Recall Curve (AUPRC). We also report 95% confidence intervals with 1,000 iterations using the bootstrapping method (Efron & Tibshirani, 1994). For reproducibility, our code is included in Appendix A.5 as we aim to make it publicly available, following the pseudo-code implementation of the MIND framework in Algorithm 1.

We conduct our experiments on a shared high-performance computing cluster equipped with Nvidia A100 GPUs. For the clinical conditions task, training models requires less than 20 hours (120-150 epochs), while for in-hospital mortality prediction, models are trained in less than 2 hours.

5 Results

5.1 Multimodal fusion performance results

Table 1 compares the performance of multimodal baselines with our proposed model on the test set. The MIND framework achieves the highest AUROC and AUPRC across both tasks, outperforming all baselines by over 1.4% AUROC and 2.3% AUPRC for clinical conditions prediction, and 1.6% AUROC and 1.4% AUPRC for in-hospital mortality prediction. Notably, MIND surpasses MedFuse and MedFuse-3H by over 3.4% AUROC and 5.4% AUPRC for clinical conditions, and by over 2.4% AUROC and 3.7% AUPRC for in-hospital mortality. Validation set results are in Appendix B.1. The results show that the MIND framework excels in compressing knowledge from unimodal teacher models while enabling unimodal predictions when modalities are missing — an advantage absent in the compared frameworks.

Figure 2 shows the label-wise AUROC performance of the MIND model with TS, the best baseline. MIND surpasses TS in overall multimodal AUROC and AUPRC metrics and shows superior AUROC across all individual and group labels: 0.777 vs. 0.763 for acute conditions, 0.813 vs. 0.797 for mixed conditions, and

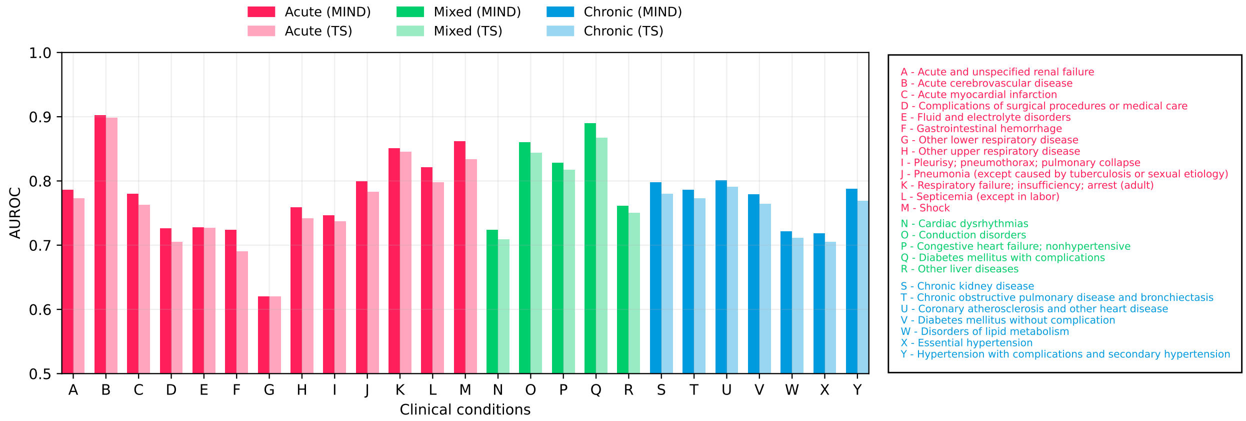

Figure 2: **Label-wise AUROC Performance**. Comparison of AUROC performance between the MIND model and the best baseline (TS) for each clinical condition label and group (acute, mixed, and chronic).

0.770 vs. 0.756 for chronic conditions. Similarly, for AUPRC (Figure B1, Appendix B.2), MIND scores 0.448 vs. 0.422 for acute conditions, 0.592 vs. 0.567 for mixed, and 0.551 vs. 0.534 for chronic conditions.

5.2 Quality of unimodal representations

We also evaluate the quality of the unimodal representations. The test set results are summarized in Table 2 and Table 3 for the clinical conditions and in-hospital mortality tasks, respectively. Additional results on the validation set are provided in Appendix B.3. For clinical conditions prediction using X-ray images, the MIND model achieves a superior AUROC (0.709 vs. 0.663) and AUPRC (0.409 vs. 0.349) compared to MedFuse-3H. Similarly, for time-series data, the MIND model outperforms MedFuse-3H with an AUROC of 0.746 vs. 0.715 and an AUPRC of 0.451 vs. 0.404. In the in-hospital mortality task, the MIND model improves chest X-ray model performance by over 10% compared to MedFuse-3H and maintains similar performance for clinical time series. The results show that the MIND model not only enhances multimodal performance, but also ensures unimodal performance is on par with unimodal models trained on much larger datasets (124,671 for CXR and 42,628 for EHR vs. 7,728 paired CXR-EHR samples).

These results highlight the MIND framework's ability to significantly enhance multimodal and unimodal performance. The unimodal encoders can be used when modalities are missing, achieving performance comparable to unimodal encoders trained on larger datasets. Unlike related work, our multimodal compression framework uniquely incorporates the use of unimodal encoders.

5.3 Ablation study I: sensitivity analysis

We conduct ablation studies to understand the impact of each component in Equation 7. To evaluate the benefits of ensembling, we experiment using both single-teacher KD and EKD (three models per modality ensemble). Specifically, we consider six settings:

1. Supervised learning as in previous work (Hayat et al., 2022), without knowledge distillation.
2. Supervised learning with modified loss (Equation 2), without knowledge distillation.
3. Supervised learning with unweighted unimodal single-teacher KD. Specifically, we set $\omega_A, \omega_B = 1$ in Equation 7 and use a single unimodal model as modality teacher.
4. Supervised learning with weighted unimodal single-teacher KD. We apply Equation 7 but we use a single unimodal model as modality teacher.

Table 2: **Unimodal performance results for the clinical conditions task.** Evaluation of unimodal and multimodal models per modality on the multimodal test set, including the number of training samples per model. Note that unimodal models are trained on much larger datasets compared to the fusion models, which are trained on the multimodal set. The best results are highlighted in bold.

Model	**Training set**	**Chest X-Ray images**		**Clinical time series**	
		AUROC	**AUPRC**	**AUROC**	**AUPRC**
Multimodal					
MedFuse-3H	7,728	0.663 (0.633, 0.693)	0.349 (0.313, 0.391)	0.715 (0.687, 0.743)	0.404 (0.365, 0.449)
MIND (Ours)	7,728	0.709 (0.680, 0.738)	0.409 (0.370, 0.455)	**0.746** (0.719, 0.772)	**0.451** (0.408, 0.499)
Unimodal					
ResNet-34	124,671	0.704 (0.674, 0.733)	0.400 (0.361, 0.445)	-	-
ResNet-10	124,671	**0.710** (0.681, 0.740)	**0.413** (0.373, 0.459)	-	-
2-layer LSTM	42,628	-	-	0.744 (0.717, 0.771)	0.448 (0.406, 0.496)
4-layer LSTM	42,628	-	-	0.742 (0.715, 0.768)	0.447 (0.404, 0.495)

Table 3: **Unimodal performance results for the in-hospital mortality prediction task.** Evaluation of unimodal and multimodal models per modality on the multimodal test set, including the number of training samples per model. Note that unimodal models are trained on much larger datasets compared to the fusion models, which are trained on the multimodal set. The best results are highlighted in bold.

Model	**Training set**	**Chest X-ray images**		**Clinical time series**	
		AUROC	**AUPRC**	**AUROC**	**AUPRC**
Multimodal					
MedFuse-3H	4,885	0.572 (0.529, 0.617)	0.201 (0.164, 0.251)	0.827 (0.794, 0.857)	0.480 (0.406, 0.555)
MIND (Ours)	4,885	**0.690** (0.648, 0.732)	**0.290** (0.236, 0.358)	**0.830** (0.795, 0.859)	**0.502** (0.427, 0.577)
Unimodal					
ResNet-34	124,671	0.681 (0.638, 0.724)	0.276 (0.226, 0.346)	-	-
ResNet-10	124,671	0.682 (0.643, 0.721)	0.259 (0.214, 0.318)	-	-
2-layer LSTM	42,628	-	-	**0.830** (0.801, 0.858)	0.489 (0.425, 0.569)
4-layer LSTM	42,628	-	-	0.823 (0.790, 0.853)	0.495 (0.418, 0.569)

5. Supervised learning with unweighted unimodal EKD. We set $\omega_A, \omega_B = 1$ in Equation 7 and use an ensemble of three teacher models as teachers.

6. MIND: supervised learning with weighted unimodal EKD (Equation 7).

Table 4 shows the MIND setting results for the clinical conditions task, and Table B4 presents results for the in-hospital mortality task, both in terms of AUROC and AUPRC on the test set. Validation set results for both tasks are in Appendix B.4. We observe that adding supervised loss terms for both modalities slightly improves AUROC (0.743 to 0.748) and AUPRC (0.440 to 0.449). Incorporating the unimodal unweighted distillation component further improves AUROC and AUPRC to 0.761 and 0.472, respectively, with ensembling boosting these to 0.766 and 0.476. Adding weighting coefficients to the distillation components significantly enhances performance on both multimodal and unimodal predictions. This demonstrates the benefits of distilling knowledge from pre-trained unimodal networks and the importance of the weighting parameters in Equation 7. All these factors collectively enhance the multimodal model, enabling unimodal predictions and achieving superior overall performance.

Table 4: **Ablation study on MIND for multimodal and unimodal performance in the clinical conditions task**. Fusion and unimodal performance on the multimodal test set for MIND with different loss components. Each setting indicates the components used in the modified loss for training and the type of knowledge distillation performed. The best results are shown in bold.

$\mathcal{L}_{S_{AB}}$	$\mathcal{L}_{S_{A/B}}$	$\mathcal{L}_{KD^U_{A/B}}$	$\mathcal{L}_{EKD^U_{A/B}}$	$\omega_{A/B}$	Fusion AUROC	Fusion AUPRC	Chest X-Ray AUROC	Chest X-Ray AUPRC	Time Series AUROC	Time Series AUPRC
✓					0.743 (0.716, 0.769)	0.440 (0.398, 0.489)	-	-	-	-
✓	✓				0.748 (0.721, 0.774)	0.449 (0.407, 0.498)	0.674 (0.644, 0.705)	0.364 (0.328, 0.407)	0.713 (0.685, 0.740)	0.405 (0.366, 0.450)
✓	✓	✓			0.761 (0.735, 0.787)	0.472 (0.428, 0.520)	0.686 (0.657, 0.716)	0.381 (0.344, 0.423)	0.730 (0.702, 0.756)	0.427 (0.386, 0.474)
✓	✓		✓		0.766 (0.740, 0.791)	0.476 (0.432, 0.525)	0.689 (0.659, 0.718)	0.383 (0.345, 0.427)	0.734 (0.706, 0.761)	0.428 (0.386, 0.474)
✓	✓	✓		✓	0.778 (0.753, 0.803)	0.498 (0.453, 0.548)	0.695 (0.665, 0.725)	0.394 (0.356, 0.438)	0.738 (0.711, 0.764)	0.437 (0.396, 0.484)
✓	✓		✓	✓	**0.782** (0.757, 0.807)	**0.506** (0.460, 0.556)	**0.709** (0.680, 0.738)	**0.409** (0.370, 0.455)	**0.746** (0.719, 0.772)	**0.451** (0.408, 0.499)

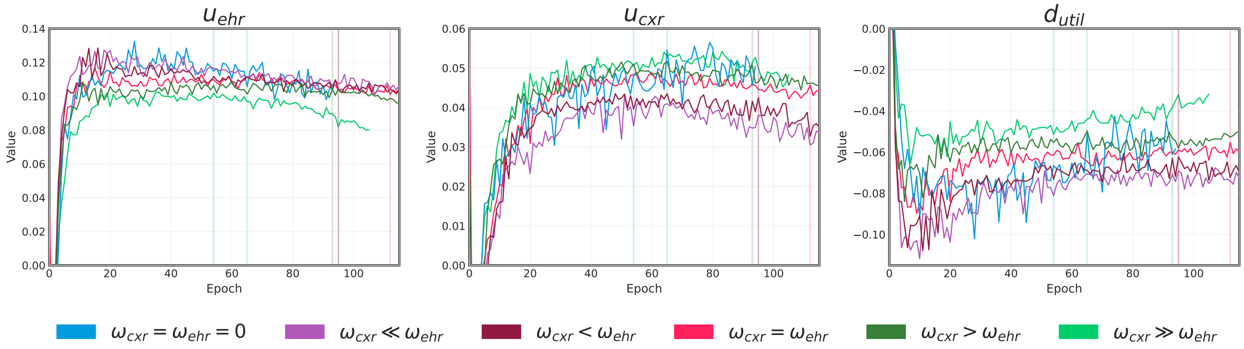

Figure 3: **Characterization of modality utilization in multimodal learning training for the clinical conditions task.** The line graphs depict the conditional utilization rate (u_{cxr}, u_{ehr}) and their difference ($d_{util} = u_{cxr} - u_{ehr}$) per epoch during training for the four settings. Vertical lines indicate the epoch with the best AUROC performance for each model on the validation set.

5.4 Ablation study II: balancing multimodal learning with w_A and w_B

We conduct hyperparameter sensitivity analysis for the weighting parameters ω_A and ω_B introduced in Equation 7. These parameters help the model focus on unimodal representation learning, thereby improving unimodal prediction quality, as shown in Section 5.2. We evaluate their impact on enhancing multimodal learning balance using the conditional utilization rate per modality (u_{cxr}, u_{ehr}) and their difference (d_{util}). Using validation AUROC as the accuracy metric, we test six settings: (i) $\omega_{cxr} = \omega_{ehr} = 0$, (ii) $\omega_{cxr} \ll \omega_{ehr}$, (iii) $\omega_{cxr} < \omega_{ehr}$, (iv) $\omega_{cxr} = \omega_{ehr}$, (v) $\omega_{cxr} > \omega_{ehr}$, and (vi) $\omega_{cxr} \gg \omega_{ehr}$.

Figure 3 shows the conditional utilization rates and their differences for the clinical conditions task. In the setting $\omega_{cxr} = \omega_{ehr} = 0$, the fusion model tends to overfit the EHR modality, with utilization rate differences up to 10%. Assigning significantly different weights to the modalities, as in the $\omega_{cxr} \gg \omega_{ehr}$ and $\omega_{cxr} \ll \omega_{ehr}$ settings, increases the conditional utilization rate of the modality with the larger weight and decreases it for the other, affecting the balance of utilization rates during training. This effect can be leveraged to achieve more balanced multimodal learning, as seen in the $\omega_{cxr} \gg \omega_{ehr}$ setting, which emphasizes the under-utilized CXR modality. Here, the EHR modality utilization never exceeds 10%, and the difference between the conditional utilization rates remains around 5%, the smallest observed. A smaller absolute value of d_{util} indicates more balanced multimodal learning. Additionally, in this setting, the utilization is consistent with a smoother u_{cxr} line. This shows that assigning a larger weight to the under-utilized modality results in more balanced multimodal learning, reducing overfit of the dominant modality and increasing the use of the weaker modality. However, it's important to note that a more balanced multimodal training does not always equate to better fusion performance but is related to improved unimodal encoder performance.

5.5 Additional Results on Multimodal Benchmark Datasets

The results in Section 5.1 demonstrate the utility of the MIND framework in clinical settings, which is the primary motivation for our study, particularly for multilabel and binary clinical tasks, using a state-of-the-art architecture as the base model. To evaluate the generalization and applicability of the MIND framework across diverse multimodal datasets, tasks, and architectures, we conducted additional validation on three multimodal, multiclass benchmark datasets: CREMA-D (Cao et al., 2014), S-MNIST (Khacef et al., 2019), and LUMA (Bezirganyan et al., 2024). We report the best accuracy results for each model in Table 5. For these experiments, the feature encoders consist of ResNet-3/6 models, with the output from each encoder concatenated and used as input for the fusion model, which is a linear layer. Additional implementation details such as dataset decriptions, multimodal architectures employed and their sizes can be found in Appendix C.

Table 5: **Multimodal fusion performance results on multiclass classification tasks.** Performance results on the multimodal test set of the MIND model and all baselines (Vanilla, Vanilla-3H, TS, MKE-CXR, MKE-EHR, and UME) for the multiclass classification tasks defined by the CREMA-D, S-MNIST and LUMA datasets. Best results are shown in bold.

Model	**CREMA-D**			**S-MNIST**			**LUMA**		
	Audio-Video	**Audio**	**Video**	**Audio-Image**	**Image**	**Audio**	**Audio-Image**	**Audio**	**Image**
Vanilla	0.599	0.482	0.273	0.970	0.967	0.417	0.847	0.385	0.280
Vanilla-3H	0.616	0.545	0.315	0.966	0.981	0.604	0.870	0.539	0.599
MKE-AUDIO (Xue et al., 2021)	0.612	0.555	0.235	0.664	0.515	0.521	0.715	0.456	0.346
MKE-VISUAL (Xue et al., 2021)	0.601	0.442	0.382	0.966	0.965	0.379	0.815	0.429	0.428
UME (Du et al., 2023)	0.620	**0.561**	0.450	0.898	**0.998**	0.699	0.903	**0.759**	0.606
TS (Wang et al., 2020a)	0.625	0.528	0.274	0.970	0.930	0.473	0.894	0.610	0.509
MIND (Ours)	**0.641**	0.552	**0.461**	**0.982**	0.981	**0.772**	**0.921**	0.721	**0.677**

As shown in Table 5, the MIND model consistently outperforms all multimodal baselines across all datasets. Additionally, the performance of the unimodal encoders either surpasses or is comparable to the baselines. Note that UME does not involve any multimodal training and only averages the unimodal models' predictions to compute the multimodal prediction. We observe that the MIND model improves multimodal prediction by an average of 1.5% and boosts the performance of the weaker modality by approximately 5.2% across tasks.

Overall, the results in Table 1 and Table 5 show that the MIND model significantly outperforms all baseline models across datasets, tasks and architectures. These results prove the benefits of MIND over all other multimodal training methods for multimodal and unimodal encoder performance for binary, multiclass and multilabel tasks. In particular, we have validated our proposed framework on 4 multimodal datasets (MIMIC-IV-CXR, CREMA-D, S-MNIST and LUMA), 5 tasks, 4 architectures and 2 fusion methods. MIND consistently outperforms all multimodal training baselines.

6 Discussion

Recent multimodal learning research highlights the challenges of learning from multiple modalities (Wu et al., 2022; Wang et al., 2020b). While increasing model size is sometimes feasible, it is not always desirable. Additionally, multimodal models often overfit to one modality, under-utilizing others. In this work, we propose MIND, a simple yet effective KD framework that distills knowledge from pre-trained unimodal teachers to a smaller multimodal student, enhancing both multimodal and unimodal predictive power. Our experimental results demonstrate significant benefits from this approach, likely due to unimodal encoders leveraging larger training datasets and learning better representations than those trained on the comparably smaller multimodal clinical datasets. To further improve ensemble impact, we hypothesize that stronger teachers and potentially a larger number of teachers are needed (Hinton et al., 2015). Note that our framework is architecture-agnostic and can be easily adapted to similar settings in other application domains.

Compared to existing work, only Buciluǎ et al. (2006) combine ensemble learning and knowledge distillation to train compact networks. Their approach involves using an ensemble to label a large unlabeled dataset, followed by training a smaller model on this labeled set. Despite methodological differences, MIND similarly avoids the need for a large ensemble of classifiers, which demands significant resources during inference. In addition, the MIND framework can be used to address imbalanced multimodal learning during training and leverages modality encoders to handle missing modalities, achieving unimodal performance on par with powerful unimodal models.

The introduction of the new learning objective (Equation 7) in our proposed approach does not compromise its applicability to real-world scenarios related to its scalability as the number of modalities, M, increases. For two modalities, as defined in Equation 7, the loss consists of five components (three supervised learning losses and two weighted unimodal ensemble knowledge distillation losses). For $M = 3$, the proposed loss would contain seven components (four supervised losses and three weighted unimodal ensemble knowledge distillation losses), incorporating one additional supervised loss and one knowledge distillation loss for the new modality. Generalizing to M modalities, the proposed loss would have $2M + 1$ loss terms, i.e., $\mathcal{O}(M)$, making our approach (linearly) scalable for practical settings.

We customize the framework for multilabel classification, addressing a less explored area in knowledge distillation, which typically focuses on multiclass classification. multilabel classification is common in clinical prediction tasks, where multiple labels may be present in a sample. We also evaluate our framework on a binary classification task. We demonstrate its effectiveness in two relevant clinical tasks: multilabel clinical conditions prediction and binary in-hospital mortality prediction using the publicly available MIMIC-CXR and MIMIC-EHR datasets. In addition, we evaluate its performance on multimodal multiclass benchmark datasets, thus providing a comprehensive validation of our proposed framework across multilabel, multiclass and binary classification tasks. Our code is publicly available to ensure reproducibility and facilitate future comparisons (See Appendix A.5).

Limitations. Since our evaluation is limited to two clinical tasks, further experiments with additional clinical tasks and datasets are needed. However, the scarcity of real-world multimodal medical datasets with images and clinical time-series data (heterogeneous modalities) poses a significant challenge. In addition, while our solution can help balance multimodal learning, it cannot ensure complete balance if one modality dominates entirely. In such cases, any multimodal approach may be inefficient. In future work, we aim to explore larger ensembles and employ stronger teachers. Despite the benefits of a compressed multimodal network during deployment, significant computational resources are still needed during training. Finally, considering the societal impact, while multimodal networks may improve overall performance in the test cohort, further work is required to assess its fairness and generalizability in patient subcohorts and instance-level analysis.

7 Conclusion

Multimodal learning offers potential enhancements for clinical prediction tasks by leveraging cross-modality interactions. However, integrating multiple modalities can lead to increased network complexity and size, posing challenges for resource-constrained applications. Additionally, multimodal clinical data present hurdles such as modality heterogeneity and missing data. Overall, our work addresses these challenges through weighted ensemble knowledge distillation, offering a promising approach to enhance fusion networks for real-world multimodal clinical data. Finally, we demonstrate the generalizability of our proposed approach by validating it across three multimodal benchmark datasets, alongside two clinical tasks, two fusion methods, and various multimodal network architectures.

References

Dhruv Agarwal, Tanay Agrawal, Laura M Ferrari, and François Bremond. From multimodal to unimodal attention in transformers using knowledge distillation. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–8. IEEE, 2021.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014.

Moran Baruch, Nir Drucker, Lev Greenberg, and Guy Moshkowich. A methodology for training homomorphic encryption friendly neural networks. In *Applied Cryptography and Network Security Workshops: ACNS 2022 Satellite Workshops, AIBlock, AIHWS, AIoTS, CIMSS, Cloud S&P, SCI, SecMT, SiMLA, Rome, Italy, June 20–23, 2022, Proceedings*, pp. 536–553. Springer, 2022.

Louise Bate, Andrew Hutchinson, Jonathan Underhill, and Neal Maskrey. How clinical decisions are made. *British journal of clinical pharmacology*, 74(4):614–620, 2012.

Grigor Bezirganyan, Sana Sellami, Laure Berti-Équille, and Sébastien Fournier. Luma: A benchmark dataset for learning from uncertain and multimodal data. *arXiv preprint arXiv:2406.09864*, 2024.

Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Yiqiang Chen, Wang Lu, Xin Qin, Jindong Wang, and Xing Xie. Metafed: Federated learning among federations with cyclic knowledge distillation for personalized healthcare. *arXiv preprint arXiv:2206.08516*, 2022.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*, 2022.

Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. *IEEE transactions on medical imaging*, 39(7):2415–2425, 2020.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pp. 8632–8656. PMLR, 2023.

Hongbin George Du and Yanke Hu. Squeezebiobert: Biobert distillation for healthcare natural language processing. In *Computational Data and Social Networks: 9th International Conference, CSoNet 2020, Dallas, TX, USA, December 11–13, 2020, Proceedings 9*, pp. 193–201. Springer, 2020.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. *arXiv preprint arXiv:2207.07027*, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Thi Kieu Khanh Ho and Jeonghwan Gwak. Utilizing knowledge distillation in deep learning for classification of chest x-ray abnormalities. *IEEE Access*, 8:160749–160761, 2020.

Minhao Hu, Matthis Maillard, Ya Zhang, Tommaso Ciceri, Giammarco La Barbera, Isabelle Bloch, and Pietro Gori. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pp. 772–781. Springer, 2020.

Chung-ju Huang, Leye Wang, and Xiao Han. Vertical federated knowledge transfer via representation distillation for healthcare collaboration networks. In *Proceedings of the ACM Web Conference 2023*, pp. 4188–4199, 2023.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):136, 2020.

Szu-Chi Huang, Cheng-Fu Cao, Po-Hsun Liao, Lung-Hao Lee, Po-Lei Lee, and Kuo-Kai Shyu. Enhancing chinese multi-label text classification performance with response-based knowledge distillation. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pp. 25–31, 2022a.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International Conference on Machine Learning*, pp. 9226–9259. PMLR, 2022b.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Benjamin Moody, Brian Gow, Li-wei H Lehman, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Lyes Khacef, Laurent Rodriguez, and Benoit Miramond. Written and spoken digits database for multimodal learning. 2019.

Qian Lou, Yilin Shen, Hongxia Jin, and Lei Jiang. Safenet: A secure, accurate and fast neural network inference. In *International Conference on Learning Representations*, 2021.

Stephen G Pauker and Jerome P Kassirer. The threshold approach to clinical decision making. *New England Journal of Medicine*, 302(20):1109–1117, 1980.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022.

Jesse M Pines, Ali S Raja, Fernanda Bellolio, and Christopher R Carpenter. *Evidence-based emergency care: diagnostic testing and clinical decision rules.* John Wiley & Sons, 2023.

Zhenzhen Quan, Qingshan Chen, Moyan Zhang, Weifeng Hu, Qiang Zhao, Jiangang Hou, Yujun Li, and Zhi Liu. Mawkdn: A multimodal fusion wavelet knowledge distillation approach based on cross-view attention for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005.

Ravi Soni, Jiahui Guan, Gopal Avinash, and V Ratna Saripalli. Hmc: a hybrid reinforcement learning based model compression for healthcare applications. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pp. 146–151. IEEE, 2019.

Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1828–1838, 2020a.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020b.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 854–863, 2021.

Litao Yang, Deval Mehta, Dwarikanath Mahapatra, and Zongyuan Ge. Leukocyte classification using multimodal architecture enhanced by knowledge distillation. In *Medical Optical Imaging and Virtual Microscopy Image Analysis: First International Workshop, MOVI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, pp. 63–72. Springer, 2022.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. Unims: A unified framework for multimodal summarization with knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11757–11764, 2022.

Zhuoran Zhao, Kamyar Mirzazad Barijough, and Andreas Gerstlauer. Deepthings: Distributed adaptive deep learning inference on resource-constrained iot edge clusters. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(11):2348–2359, 2018.

A Implementation details

A.1 Model size

Table A1 shows the size of the different model architectures. MedFuse (Hayat et al., 2022) is the base architecture of our work and it is provided as a reference. While originally proposed using ResNet-34 and 2-layer LSTM encoders, for the sake of comparison, we adopt the same configuration for MIND (ResNet-10 and 2-layer LSTM) and all baselines. Note that the MIND model is 3 times smaller than the original architecture, while including unimodal classification capabilities.

Table A1: **Size of multimodal models.** We provide a summary of the number of trainable parameters for each multimodal architecture. We use ResNet-10 and the two-layer LSTM as the student network within proposed MIND framework and for all baselines.

Model	**No. of trainable parameters**
ResNet-34 and two-layer LSTM (Hayat et al., 2022)	23.9 M
ResNet-10 and two-layer LSTM (MIND)	7.5 M

A.2 Dataset

In our experimental setup, we use two publicly available datasets for two clinical prediction tasks: (i) multilabel clinical conditions prediction and (ii) in-hospital mortality prediction. For modality A, we use chest X-ray images (CXR) data extracted from the MIMIC-CXR dataset (Johnson et al., 2019). For modality B, we use electronic health record (EHR) time-series data extracted from the MIMIC-IV dataset (Johnson et al., 2023). We link the modalities following Hayat et al. (2022). The dataset is composed of all clinical time-series that have an associated chest X-ray image, thus both modalities are present for each sample. We summarize the dataset in Table A.2.

Table A2: **Summary of dataset.** Size, shape and composition of the datasets used to train, validate and evaluate the models. The chest X-ray images, where a single image $\mathbf{x}_A \in \mathbb{R}^{224\times224\times3}$, are represented by dataset $\mathcal{D}_A$, and the clinical time-series data, such that a sample $\mathbf{x}_B \in \mathbb{R}^{t\times76}$ where t is the number of time-steps based on patient's stay in the intensive care unit and 76 is the number of pre-processed features per time-step, are represented by $\mathcal{D}_B$.

Dataset	**Training**	**Validation**	**Test**	**Shape, composition**
CXR ($\mathcal{D}_A$)	124671	8813	20747	$\mathbf{x}_A \in \mathbb{R}^{224\times224\times3}$
EHR ($\mathcal{D}_B$)	42628	4802	11914	$\mathbf{x}_B \in \mathbb{R}^{t\times76}$
Used (multimodal)	7756	882	2166	$\mathcal{D}_A \cap \mathcal{D}_B$

A.3 MIND architecture & implementation

We use MedFuse (Hayat et al., 2022) as the multimodal baseline for our experimental setup. The original implementation of MedFuse uses an LSTM-based fusion module that processes a sequence of modality representations provided by a ResNet-34 encoder for the chest X-ray images and a 2-layer LSTM encoder for the clinical time-series. In the MIND framework, we reduce the size of the image encoder to a ResNet-10 model aiming for model compression. In our experiments, we adhere to the hyperparameters used in the original MedFuse implementation for the randomly initialized multimodal fusion network. For the MIND model and all baselines, we perform random hyper-parameter tuning of the learning rates, and weighting coefficients where applicable. We make a minimum of 50 runs per model on each task. For the unimodal baselines, we randomly select ten different learning rates and used the best performing model based on the epoch with the best AUROC on the validation set. All hyperparameters are summarized in Table A3.

Table A3: **Model hyperparameters.** Description of the hyperparameters used in our experimental setup. We follow the ones suggested by Hayat et al. (2022) for the multimodal model with randomly initialized encoders and baselines. For the unimodal encoders, we evaluate different learning rates and select the best one based on the epoch with best AUROC on the validation set.

Hyper-parameter	**Value/s**
Batch size	16
Drop out	0.3
Epochs	300
Early stopping	Patience of 40
Learning rate - multimodal/baselines	$[1 \times 10^{-3}, ..., 1 \times 10^{-5}]$
Learning rate - unimodal CXR	$[1 \times 10^{-4}, ..., 1 \times 10^{-6}]$
Learning rate - unimodal EHR	$[1 \times 10^{-4}, ..., 1 \times 10^{-6}]$

A.4 Multimodal Baselines

Aligning with our proposed response-based knowledge distillation framework, MIND, we adapt for multilabel classification the baselines following their corresponding papers, including MedFuse (Hayat et al., 2022), MedFuse-3H, MKE (Xue et al., 2021), UME (Du et al., 2023) and TS (Wang et al., 2020a). In our MIND framework and for all baselines, we save the best model checkpoint in terms of AUROC on the validation set during training and load the saved model to evaluate the performance on the test set. Table A4 provides an overview of the loss functions proposed by each baseline for multimodal training.

MedFuse (Hayat et al., 2022) serves as the base architecture for all models and experiments in our work. It is a typical multimodal architecture proposed for medical tasks, featuring two modality encoders: a ResNet-34 for CXR images and a 2-layer LSTM for time series data. The encoder outputs are concatenated and passed through an LSTM layer before the final multimodal classification head. To incorporate their methodology, we utilize the open-source code available at https://github.com/nyuad-cai/MedFuse. We perform learning rate tuning in our experiments while adopting the default settings for the other hyper-parameters.

MedFuse-3H. Like most multimodal networks, the original MedFuse architecture does not include a classification head on the modality encoders, resulting in an exclusively multimodal output. In our MIND framework, the first step is to modify the multimodal network architecture to leverage the encoders for cases with missing modalities, avoiding the need for zero or mean imputation. We extend the original MedFuse architecture to include three classification heads (3H), naming it MedFuse-3H, and use it as a baseline for our approach. We perform learning rate tuning in our experiments while adopting the default MedFuse settings for other hyper-parameters.

TS (Wang et al., 2020a) proposes a response-based knowledge distillation framework where teacher models trained with large datasets, including samples with missing modalities, transfer knowledge via soft labels to a multimodal student model. The multimodal student is trained using the soft labels from the unimodal teachers along with the supervision loss (one-hot-encoded labels) as shown in Table A4. Originally proposed for multiclass classification, we adapt it for our multilabel classification framework. We perform random hyperparameter tuning of the learning rate and the α, β distillation coefficients. Following(Wang et al., 2020a), $\alpha, \beta \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. We keep the other hyperparameters consistent across models and baselines, using the default MedFuse settings.

Multimodal Knowledge Expansion (MKE) (Xue et al., 2021) proposes a response-based knowledge distillation framework composed of a unimodal teacher and a multimodal student. Originally designed to work with labeled and unlabeled data and evaluated on binary classification, multiclass classification, and segmentation tasks, we adapt it for our multilabel classification tasks framework. While it proposes the loss described in Table A4, following the instructions from the authors, we train the multimodal student network using a method equivalent to the equation in the table. Since it proposes knowledge distillation (KD) from a single unimodal teacher to a multimodal student, we split this baseline into two: **MKE-EHR**, where we use the best EHR unimodal model as the teacher, and **MKE-CXR**, where we use the best CXR unimodal model as the teacher. We perform learning rate tuning in our experiments while adopting the default MedFuse settings for other hyperparameters.

Uni-modal Ensemble (UME) (Du et al., 2023) proposes the aggregation of predictions by independently trained unimodal models for multimodal samples on a given task. The prediction for a multimodal instance is given by weighting the predictions of an ensemble of unimodal models (one model per modality). Following Du et al. (2023), we average the unimodal models' predictions to provide the final prediction. Originally proposed for multiclass classification, we adapt it to our multilabel classification framework. No further training is needed using this framework apart from the independent training of the unimodal models.

Table A4: **Comparison of loss functions.** Using our previously introduced notation, we provide a summary of the loss functions proposed by the MIND framework and all baseline models.

Model	**Loss function**
MedFuse (Hayat et al., 2022)	$\mathcal{L}_{S_{AB}}$
MedFuse-3H	$\mathcal{L}_{S_{AB}} + \mathcal{L}_{S_A} + \mathcal{L}_{S_B}$
MKE-CXR (Xue et al., 2021)	$\mathcal{L}_{KD_A} + \gamma \times \mathcal{L}_{reg}$
MKE-EHR (Xue et al., 2021)	$\mathcal{L}_{KD_B} + \gamma \times \mathcal{L}_{reg}$
UME (Du et al., 2023)	$\mathcal{L}_{S_A}$ and $\mathcal{L}_{S_B}$ (independently trained unimodal models)
TS (Wang et al., 2020a)	$\mathcal{L}_{S_{AB}} + \alpha \times \mathcal{L}_{KD^U_{AB}} + \beta \times \mathcal{L}_{KD^U_{AB}}$
MIND (ours)	$\mathcal{L}_{S_{AB}} + \mathcal{L}_{S_A} + \mathcal{L}_{S_B} + \omega_A \times \mathcal{L}_{EKD^U_A} + \omega_B \times \mathcal{L}_{EKD^U_B}$

A.5 Code for reproducibility

For reproducibility of our results, we make our code available at: https://anonymous.4open.science/r/MIND_Framework/

B Additional results on clinical tasks

B.1 Multimodal performance results on the validation set

Table B1 provides performance results of the MIND model and all baselines on the multimodal validations set for both clinical tasks.

Table B1: **Multimodal fusion performance results.** Performance results on the multimodal validation set of the MIND model and all baselines (MedFuse, MedFuse-3H, TS, MKE-CXR, MKE-EHR, and UME). Best results are shown in bold.

Model	**Clinical Conditions**		**In-hospital Mortality**	
	AUROC	**AUPRC**	**AUROC**	**AUPRC**
MedFuse (Hayat et al., 2022)	0.757 (0.716, 0.795)	0.464 (0.397, 0.540)	0.833 (0.778, 0.883)	0.503 (0.395, 0.619)
MedFuse-3H	0.760 (0.719, 0.798)	0.460 (0.393, 0.536)	0.835 (0.782, 0.884)	0.438 (0.348, 0.566)
MKE-CXR (Xue et al., 2021)	0.718 (0.673, 0.761)	0.419 (0.360, 0.492)	0.727 (0.664, 0.791)	0.314 (0.239, 0.437)
MKE-EHR (Xue et al., 2021)	0.753 (0.713, 0.791)	0.456 (0.391, 0.531)	0.850 (0.807, 0.899)	0.544 (0.433, 0.672)
UME (Du et al., 2023)	0.775 (0.735, 0.813)	0.492 (0.425, 0.570)	0.850 (0.797, 0.900)	0.543 (0.423, 0.679)
TS (Wang et al., 2020a)	0.778 (0.738, 0.815)	0.490 (0.423, 0.568)	0.846 (0.795, 0.890)	0.563 (0.449, 0.667)
MIND (Ours)	**0.790** (0.752, 0.826)	**0.516** (0.446, 0.593)	**0.866** (0.817, 0.908)	**0.572** (0.455, 0.692)

B.2 Label-wise AUPRC performance

Figure B1 compares the label-wise AUPRC performance of the MIND model with TS, the best baseline.

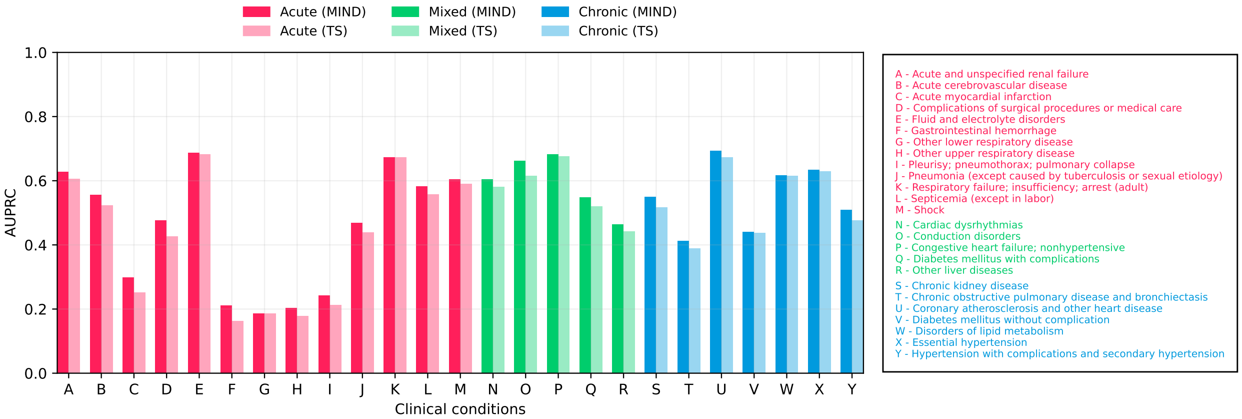

Figure B1: **Label-wise AUPRC Performance**. Comparison of AUPRC performance between the MIND model and the best baseline (TS) for each clinical condition label and group (acute, mixed, and chronic).

B.3 Unimodal performance results on the validation set

Table B2 reports the unimodal performance results for the clinical conditions prediction task on the validation set. Table B3 reports the unimodal performance results for the in-hospital mortality prediction task on the validation set. Note that unimodal models are trained on much larger datasets compared to the fusion models, which are trained on the multimodal set.

Table B2: **Unimodal performance results for the clinical conditions task on the validation set.** Evaluation of unimodal and multimodal models per modality on the multimodal validation set, including the number of training samples per model.

MODEL	TRAINING SET	CHEST X-RAY IMAGES AUROC	CHEST X-RAY IMAGES AUPRC	CLINICAL TIME SERIES AUROC	CLINICAL TIME SERIES AUPRC
Multimodal					
MedFuse-3H	7,728	0.679 (0.632, 0.724)	0.371 (0.316, 0.439)	0.713 (0.669, 0.754)	0.401 (0.343, 0.472)
MIND (Ours)	7,728	0.710 (0.665, 0.754)	0.412 (0.353, 0.484)	0.754 (0.714, 0.792)	0.459 (0.394, 0.534)
Unimodal					
ResNet-34	124,671	0.714 (0.700, 0.727)	0.442 (0.422, 0.464)	-	-
ResNet-10	124,671	0.719 (0.705, 0.732)	0.450 (0.430, 0.472)	-	-
2-layer LSTM	42,628	-	-	0.762 (0.743, 0.780)	0.420 (0.388, 0.456)
4-layer LSTM	42,628	-	-	0.759 (0.740, 0.777)	0.418 (0.386, 0.454)

B.4 Ablation studies

Table B4 provides sensitivity analysis (Ablation study I, Section 5.3) for the in-hospital mortality task on the test set with different loss components. Each setting indicates the components used in the modified loss for training and the type of knowledge distillation performed. Similarly, Table B5 provides the results on the validation set for ablation study I for the prediction of clinical conditions task while Table B6 provides the results on for the in-hospital mortality prediction task.

Table B3: **Unimodal performance results for the in-hospital mortality task on the validation set.** Evaluation of unimodal and multimodal models per modality on the multimodal validation set, including the number of training samples per model.

MODEL	**TRAINING SET**	**CHEST X-RAY IMAGES**		**CLINICAL TIME SERIES**	
		AUROC	**AUPRC**	**AUROC**	**AUPRC**
Multimodal					
MedFuse-3H	4,885	0.691 (0.629, 0.749)	0.257 (0.189, 0.356)	0.828 (0.775, 0.876)	0.445 (0.338, 0.580)
MIND (Ours)	4,885	0.712 (0.650, 0.775)	0.277 (0.208, 0.375)	0.858 (0.811, 0.898)	0.548 (0.432, 0.678)
Unimodal					
ResNet-34	124,671	0.732 (0.715, 0.749)	0.183 (0.167, 0.205)	-	-
ResNet-10	124,671	0.725 (0.709, 0.742)	0.177 (0.160, 0.196)	-	-
2-layer LSTM	42,628	-	-	0.870 (0.846, 0.891)	0.528 (0.467, 0.595)
4-layer LSTM	42,628	-	-	0.870 (0.846, 0.892)	0.536 (0.473, 0.601)

Table B4: **Ablation study on MIND for multimodal and unimodal performance in the in-hospital mortality prediction task**. Fusion and unimodal performance on the multimodal test set for MIND with different loss components.

$\mathcal{L}_{S_{AB}}$	$\mathcal{L}_{S_{A/B}}$	$\mathcal{L}_{KD^U_{A/B}}$	$\mathcal{L}_{EKD^U_{A/B}}$	$\omega_{A/B}$	**Fusion**		**Chest X-Ray**		**Time Series**	
					AUROC	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
✓					0.816 (0.785, 0.847)	0.468 (0.398, 0.546)	-	-	-	-
✓	✓				0.820 (0.787, 0.851)	0.461 (0.392, 0.541)	0.669 (0.626, 0.713)	0.276 (0.232, 0.339)	0.813 (0.783, 0.843)	0.442 (0.372, 0.523)
✓	✓	✓			0.823 (0.792, 0.852)	0.454 (0.389, 0.537)	0.661 (0.618, 0.704)	0.268 (0.219, 0.333)	0.818 (0.788, 0.847)	0.455 (0.381, 0.532)
✓	✓		✓		0.832 (0.803, 0.860)	0.482 (0.412, 0.563)	0.661 (0.613, 0.705)	0.279 (0.232, 0.348)	0.827 (0.794, 0.857)	0.478 (0.402, 0.558)
✓	✓	✓		✓	0.835 (0.804, 0.864)	0.494 (0.421, 0.574)	0.682 (0.641, 0.721)	0.289 (0.237, 0.359)	0.827 (0.797, 0.856)	0.484 (0.410, 0.560)
✓	✓		✓	✓	0.844 (0.815, 0.872)	0.505 (0.433, 0.587)	0.689 (0.652, 0.726)	0.290 (0.233, 0.354)	0.830 (0.801, 0.859)	0.502 (0.426, 0.577)

Table B5: **Ablation study on MIND for multimodal and unimodal performance in the clinical conditions task**. Fusion and unimodal performance on the multimodal validation set for MIND with different loss components.

$\mathcal{L}_{S_{AB}}$	$\mathcal{L}_{S_{A/B}}$	$\mathcal{L}_{KD^U_{A/B}}$	$\mathcal{L}_{EKD^U_{A/B}}$	$\omega_{A/B}$	**Fusion**		**Chest X-Ray**		**Time Series**	
					AUROC	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
✓					0.755 (0.716, 0.794)	0.454 (0.390, 0.529)	-	-	-	-
✓	✓				0.757 (0.715, 0.796)	0.468 (0.399, 0.546)	0.677 (0.629, 0.722)	0.370 (0.316, 0.438)	0.718 (0.674, 0.760)	0.417 (0.356, 0.489)
✓	✓	✓			0.773 (0.734, 0.811)	0.486 (0.416, 0.564)	0.695 (0.648, 0.740)	0.391 (0.333, 0.461)	0.739 (0.698, 0.778)	0.435 (0.371, 0.511)
✓	✓		✓		0.779 (0.738, 0.816)	0.489 (0.421, 0.567)	0.699 (0.652, 0.744)	0.397 (0.337, 0.468)	0.742 (0.700, 0.781)	0.437 (0.375, 0.510)
✓	✓	✓		✓	0.784 (0.745, 0.820)	0.502 (0.432, 0.580)	0.700 (0.654, 0.745)	0.400 (0.341, 0.470)	0.747 (0.707, 0.786)	0.445 (0.380, 0.520)
✓	✓		✓	✓	0.790 (0.752, 0.826)	0.516 (0.446, 0.593)	0.710 (0.665, 0.754)	0.412 (0.353, 0.484)	0.754 (0.714, 0.792)	0.459 (0.394, 0.534)

Table B6: **Ablation study on MIND for multimodal and unimodal performance in the in-hospital mortality task**. Fusion and unimodal performance on the multimodal validation set for MIND with different loss components.

$\mathcal{L}_{S_{AB}}$	$\mathcal{L}_{S_{A/B}}$	$\mathcal{L}_{KD^U_{A/B}}$	$\mathcal{L}_{EKD^U_{A/B}}$	$\omega_{A/B}$	**Fusion**		**Chest X-Ray**		**Time Series**	
					AUROC	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
✓					0.833 (0.778, 0.883)	0.503 (0.395, 0.619)	-	-	-	-
✓	✓				0.836 (0.783, 0.885)	0.478 (0.388, 0.606)	0.691 (0.629, 0.749)	0.257 (0.189, 0.356)	0.828 0.775, 0.876)	0.445 (0.338, 0.580)
✓	✓	✓			0.839 (0.784, 0.887)	0.467 (0.355, 0.606)	0.670 (0.599, 0.735)	0.248 (0.187, 0.343)	0.832 (0.779, 0.881)	0.467 (0.363, 0.602)
✓	✓		✓		0.842 (0.790, 0.893)	0.488 (0.381, 0.615)	0.678 (0.605, 0.745)	0.280 (0.201, 0.395)	0.837 (0.786, 0.880)	0.475 (0.372, 0.605)
✓	✓	✓		✓	0.858 v(0.810, 0.901)	0.545 (0.428, 0.674)	0.698 (0.636, 0.760)	0.278 (0.205, 0.388)	0.850 (0.802, 0.892)	0.543 (0.423, 0.665)
✓	✓		✓	✓	0.866 (0.817, 0.908)	0.572 (0.455, 0.692)	0.712 (0.650, 0.775)	0.280 (0.211, 0.378)	0.858 (0.811, 0.898)	0.548 (0.432, 0.678)

C Additional results on multimodal multiclass benchmark datasets

C.1 Datasets

CREMA-D (Cao et al., 2014) is an audio-visual dataset for speech emotion recognition. It includes 7,442 short video clips from 91 actors speaking a selection of 12 sentences. The utterances express, with varying degrees of intensity, one of six common emotions: anger, happiness, disgust, fear, neutral, and sadness. The training set includes 5,210 samples, while both the validation and test sets include 1,116 samples.

S-MNIST (Khacef et al., 2019) is an audio-visual dataset designed for multimodal fusion classification. It was created by pairing the original MNIST handwritten digits database with a spoken digits database extracted from Google Speech Commands. To construct the training, validation and test sets, we randomly sampled 8,000 and 2,000 instances from the original training dataset (10 classes, grayscale digits 0-9) as our training and validation sets. We randomly sampled 2,000 instances from the original test set to construct our test set.

LUMA (Bezirganyan et al., 2024) is a multimodal dataset designed for benchmarking multimodal learning. The image modality includes images from a 50-class subset from CIFAR-10 and CIFAR-100 datasets while the audio modality contains utterances of the class labels. The dataset is imbalanced, with the most prevalent 22 classes having 1,500 paired samples or more. To construct our dataset, we generate an evenly distributed sample of the population, containing 1,500 paired instances per class. The resulting dataset including 33,000 paired instances is randomly split into training, validation and test sets, with 25,000, 4,000, and 4,000 audio-video samples, respectively.

C.2 Architectures and Implementation

We use a multimodal architecture composed of Resnet variants as modality encoders for both modalities. The output of the modality encoders is concatenated and passes through a linear layer that computes the final prediction. In particular, the vanilla multimodal architectures and their sizes are provided as follows:

- For the S-MNIST dataset, we use Resnet-10 models, trained with the whole training set, as unimodal teachers and Resnet-3 models as modality encoders for the compressed model, trained with a fraction of the original dataset (8,000 randomly selected samples). The teacher models have $\approx$ 4.9 M parameters while the whole MIND model has only $\approx$ 151 k (32.5x size reduction).
- For the CREMA-D dataset, we use Resnet-18 models as unimodal teachers and Resnet-6 models as modality encoders for the compressed model. All models are trained with the full training dataset. The teacher models have $\approx$ 11 M parameters while the whole MIND model has only $\approx$ 612 k (18x size reduction).
- For the LUMA dataset, we use Resnet-10 models as unimodal teachers and Resnet-3 models as modality encoders for the compressed model. All models are trained with the whole training dataset. The teacher models have $\approx$ 4.9 M parameters while the whole MIND model has only $\approx$ 151 k (32.5x size reduction).

For the MIND model and all baselines, we perform random hyper-parameter tuning of the learning rates, and weighting coefficients where applicable. We make a minimum of 50 runs per model on each task. For the unimodal baselines, we randomly select ten different learning rates and used the best performing model based on the epoch with the best accuracy on the validation set. In general, we use a batch size of 64, and train the models between 50-100 epochs. The learning rates used are in the ranges described in Table A3.