IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 October 2024, accepted 4 November 2024, date of publication 13 November 2024,
date of current version 29 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3497668

RESEARCH ARTICLE

Robust Fusion of Time Series and Image Data for Improved Multimodal Clinical Prediction

ALI RASEKH, REZA HEIDARI*, AMIR HOSEIN HAJI MOHAMMAD REZAIE*, PARSA SHARIFI SEDEH*, ZAHRA AHMADI, PRASENJIT MITRA, AND WOLFGANG NEJDL

L3S Research Center, Leibniz University Hannover, 30167 Hannover, Germany

Corresponding author: Ali Rasekh (ali.rasekh@l3s.de)

*This research was conducted during a remote internship at L3S Research Center. Amir Hosein Haji Mohammad Rezaie and Parsa Sharifi Sedeh contributed equally to this work.

This work was supported in part by the Federal Ministry of Education and Research (BMBF), Germany, through the Project LeibnizKILabor, under Grant 01DD20003; and in part by the EU in the Connecting Europe Facility Program through the Project xAIM under Grant 98381272.

ABSTRACT With the increasing availability of diverse data types, particularly images and time series data from medical experiments, there is a growing demand for techniques designed to combine various modalities of data effectively. Our motivation comes from the important areas of predicting mortality and phenotyping where using different modalities of data could significantly improve our ability to predict. To tackle this challenge, we introduce a new method that uses two separate encoders, one for each type of data, allowing the model to understand complex patterns in both visual and time-based information. Apart from the technical challenges, our goal is to make the predictive model more robust in noisy conditions and perform better than current methods. We also deal with imbalanced datasets and use an uncertainty loss function, yielding improved results while simultaneously providing a principled means of modeling uncertainty. Additionally, we include attention mechanisms to fuse different modalities, allowing the model to focus on what's important for each task. We tested our approach using the comprehensive multimodal MIMIC dataset, combining MIMIC-IV and MIMIC-CXR datasets. Our experiments show that our method is effective in improving multimodal deep learning for clinical applications. The code for this work is publicly available at: https://github.com/AliRasekh/TSImageFusion

INDEX TERMS Multimodal learning, time series, attention mechanism, robustness, phenotyping.

I. INTRODUCTION

Artificial intelligence (AI) has become increasingly essential in medical fields, transforming healthcare by offering advanced capabilities in predicting mortality, identifying diseases, and conducting various diagnostic tasks. With the rise of deep learning techniques, AI has shown outstanding effectiveness and accuracy, especially in medical applications. Methods such as convolutional neural networks and recurrent neural networks have been found to achieve high diagnostic accuracies in classifying medical images and predicting disease progression from patient records. Multimodal learning, which has gained increasing attention, uses various data sources like electronic health records and medical images to strengthen predictive modeling and diagnostic abilities.

The integration of AI in medical practice brings several benefits. Firstly, it allows healthcare professionals to use large amounts of data to make quick and accurate decisions. For example, in predicting mortality, AI algorithms can analyze patient data such as vital signs, lab results, and medical images to identify signs of deteriorating health and take timely action. Secondly, AI supports personalized medicine by identifying patient groups with distinct characteristics, helping tailor treatment strategies. This personalized approach improves patient outcomes and reduces the risk of adverse reactions to treatments. Additionally, AI-powered diagnostic tools are highly sensitive and specific, aiding in early disease detection. Overall, integrating AI into medical

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.

practice holds great promise for improving patient care, simplifying processes, and ultimately saving lives.

However, seamlessly integrating different types of data, like medical images and time series data, brings significant challenges in realizing AI's full potential in healthcare. These challenges arise from varied aspects, such as the heterogeneity of data formats, the large size and high dimensionality of the data that require massive computational resources, the need for harmonizing unstructured text data like clinical notes, and the difficulty of aligning multimodal data due to differences in acquisition methods and temporal inconsistencies. Combining these diverse data types requires innovative approaches to address the complexities and variations within clinical datasets. For example, when diagnosing complex conditions like sepsis, integrating data from multiple sources such as physiological measurements and imaging studies is crucial for accurate diagnosis and timely intervention. Overcoming these integration challenges is essential for unlocking the transformative power of AI in healthcare and maximizing its impact on patient outcomes.

This paper addresses these challenges in critical healthcare tasks such as predicting mortality and phenotyping. Our main goal is to design a robust multimodal framework capable of handling the complexities of clinical datasets effectively. Ultimately, we aim to contribute to improved and more informed healthcare decision-making.

The rich diversity of clinical data, characterized by its multimodal nature, requires innovative approaches to extract meaningful insights. Our research focuses on achieving the following key objectives:

1) **Enhanced Modality Fusion via Attention Mechanism:** We introduced an attention mechanism enabling dynamic allocation of attention across modalities, enhancing model flexibility and improving predictive accuracy. This underscores the importance of modality fusion in multimodal architectures.
2) **Uncertainty-Aware Multi-Task Learning with Uncertainty Loss:** Employing an uncertainty loss function for multi-task phenotype classification, our approach prioritizes simpler and more certain tasks, enhancing overall performance by adapting to complex and uncertain ones.
3) **Robustness in Noisy Environments:** Develop methods to ensure robust performance even in noisy settings commonly encountered in real-world hospital scenarios, where data may exhibit variability and imperfections.

Our research yields strong results, demonstrating the practical usefulness and resilience of our multimodal framework under challenging conditions, including noisy settings and data. We chose the MIMIC dataset [1], [2] for its diverse modalities and comprehensive nature, encompassing electronic health records, time series data, and chest X-ray images. The objectives of our study, including predicting mortality, identifying diseases, and labeling radiology images, align closely with the rich annotations and labels available within the MIMIC dataset. This consolidation facilitates a comprehensive understanding of patient health, allowing our model to learn correlations between temporal health records and visual representations. The resulting multimodal dataset is carefully preprocessed, aligning timestamps and standardizing imaging data, ensuring a coherent fusion of modalities for robust model training and evaluation.

Our thorough exploration of multimodal deep neural networks for clinical applications has revealed impactful high-level findings. The specialized encoders designed for images and time series data can successfully capture patterns within each modality, significantly enhancing the model's discriminative power. Additionally, the integration of attention mechanisms for modality fusion empowers our model to allocate attention dynamically based on task and modality relevance. This not only improves adaptability but also enhances interpretability across predicting mortality and phenotyping labels such as chronic kidney diseases, other liver diseases, and complications of surgical procedures or medical care. Another critical aspect, particularly relevant in multi-label classification, is the consideration of uncertainty and how to model it effectively. We have shown that the uncertainty loss function not only improves performance but also provides a principled means of modeling, specifically in identifying diseases. Our approach showcases strong results, demonstrating the practical usefulness and resilience of our multimodal framework, even under challenging conditions such as noisy settings.

In the following sections, we first review existing work in multimodal learning and machine learning methods in healthcare data (Section II). Then, we provide a comprehensive overview of our dataset, detailing its composition, preprocessing steps, and rationale for integrating the MIMIC-IV and MIMIC-CXR datasets, following the approach outlined in the MedFuse paper [3]. We detail the methodologies guiding our multimodal model training in (Section IV). Subsequent sections cover experimental outcomes and our findings, highlighting the robustness and superior performance of our approach compared to state-of-the-art methods (Section V), and conclude the paper by outlining future directions in multimodal deep learning for healthcare (Section VI).

II. RELATED WORK

The field of multimodal learning has seen significant attention for its potential to extract richer representations from heterogeneous data. An indicative example is the recent development in employing both MRI imaging data and clinical notes to better predict disease progression in neurodegenerative disorders such as Alzheimer's [4]. Our focus in this paper lies at the intersection of diverse endeavors seeking to capitalize on synergies among different data modalities. We explore two key domains in the following: Section II-A explores multimodal machine learning, emphasizing approaches integrating disparate data modalities for enhanced model performance. In Section II-B, we focus

on machine learning applications in healthcare, reviewing studies that have led computational methods in improving patient outcomes.

A. MULTIMODAL MACHINE LEARNING

Researchers have been exploring various methodologies to leverage multimodal data effectively in machine learning tasks. Rahim et al. [5], for instance, highlight the importance of integrating longitudinal MRI images with clinical data for disease progression prediction, while Niu et al. [6] combined time series and clinical data to improve mortality prediction. Soenksen et al. [7] take a step further by developing a platform capable of generating embeddings for various modalities, including images, text, and tabular data. These contributions have indeed advanced our understanding of the field. However, the simultaneous incorporation of image and time series data, which is a cornerstone in our research, remains unexplored in these studies. Our work addresses this by integrating both data types, thereby enhancing model performance.

In the area of medical imaging with multimodal learning, a few notable studies have demonstrated promising results. Nie et al. [8] employed a multi-channel 3D CNN to integrate various MR imaging modalities, achieving high accuracy in predicting overall survival (OS) time for high-grade glioma patients. Despite their achievements, a key concern is the singular focus on one type of MRI data potentially reducing the ability to fully capture the complex patterns of the medical problems. Similarly, Srinivas and Sasibhushana Rao [9] combined multiple imaging types to create a two-stage learning method. Yet, their techniques might be limited by an over-reliance on specific imaging types, not considering time-series and leaving unaddressed the problem of model uncertainty. Our work, by contrast, applies separate encoders for distinct data types, allowing the better recognition of complex patterns in both visual and time-based data, and also addresses the challenges related to imbalanced data and model uncertainty.

Further pushing the boundaries in the field of medical imaging, Muduli et al. [10] designed a deep CNN model by integrating mammograms and ultrasound images. Similarly, Khan et al. [11] proposed a multimodal deep neural network to improve multi-class diagnosis of malignant liver conditions. These studies primarily focus on singular tasks. Distinctly, our work also explores a multi-task setting, broadening the scope of applications by simultaneously predicting multiple clinical outcomes from our extensive multimodal MIMIC dataset.

Certain research efforts have concentrated on predicting specific clinical outcomes using multimodal data. Sun et al. [12] proposed a multimodal neural network integrating multi-dimensional data to predict breast cancer prognosis, achieving superior performance compared to pre-existing methods. Furthering this concept, Joo et al. [13] developed a deep learning model to predict the pathologic complete response (pCR) to neoadjuvant chemotherapy (NAC) in breast cancer patients. They integrated high-dimensional MR images and clinical information, significantly outperforming models that employed only one type of data. These studies illustrate the potential of multimodal deep learning in health prognostics and individualized healthcare. While these studies make significant advances in their respective areas, they are primarily focused on the prediction of specific outcomes in breast cancer patients, and do not consider the multi-task setting in medical domain.

B. MACHINE LEARNING IN HEALTHCARE

Despite promising advancements in applying machine learning to healthcare, significant limitations persist, including the need for extensive labeled data, interpretability challenges, and potential biases. This section examines these issues and current research efforts aimed at addressing them to ensure ethical and practical implications are carefully considered.

Zeng et al. [14] amalgamate structured clinical data with clinical narratives for predicting distant recurrences in breast cancer, while Harerimana et al. [15] encourage health informaticians to utilize deep learning in medical settings because of its potential. Daberdaku et al. [16] and Zikos et al. [17] embrace EHR data, with the former using it to navigate missing longitudinal clinical data and the latter augmenting the Healthcare Cost and Utilization Project tools. Finally, Guo et al. [18] advocate for the adoption of machine learning for the prediction of mortality among patients with liver cirrhosis. Although these studies underline the applicability of EHR and clinical information in tackling various healthcare challenges, they each fall short in addressing multimodal data integration, which is one our of contributions.

Several studies have been conducted in the field of disease prognosis and diagnosis. The study by Jeon et al. [19] anticipates changes in blood glucose concentration. Mir and Dhage [20] utilize machine learning for diabetes prediction, while Gogi and Vijayalakshmi [21] enhance prognosis methods for liver diseases using machine learning. Further, Maity and Das [22] contribute to the diagnosis and prognosis of Alzheimer's disease and breast cancer, respectively. Niu et al. [23], and Suvon et al. [24], focused on sizable, detailed health record data to predict disease. These studies inform our work by highlighting the potential of machine learning in disease prognosis. Our research addresses under-explored areas in their work under-explored areas in their work, specifically by leveraging multimodal data integration and applying uncertainty modeling for more accurate predictions.

Zhu et al. [25] developed a fall detection framework, while Balbi et al. [26] highlighted the role of chest X-rays in predicting the severity of COVID-19 patients. Rahane [27] used image processing and machine learning techniques for lung cancer detection while Jacenkow et al. [28] focused on the influence of textual information in radiology reports on image classification tasks. Bezirganyan et al. [29] introduced

M2-Mixer for multimodal classification tasks. While these studies highlight the significance of image data in healthcare prediction tasks, they focus primarily on single-modal data. Our paper addresses this limitation by introducing a method that employs separate encoders for each modality, providing a mechanism for efficient multimodal fusion.

Türkmen [30] developed Turkish biomedical language models to enhance the classification performance in clinical contexts. Niu et al. [23] proposed an attention mechanism with Clinical-BERT for disease risk prediction using textual inputs. Jacenkow et al. [28] investigated the influence of textual information on image classification tasks in radiology reports. While these papers underline the importance of text-based data, they address their processing in isolation without discussing its integration with other data modalities. In contrast, our work fills this gap by proposing a robust method capable of handling both visual and time series data effectively and fusing the information for improved prognostic performance.

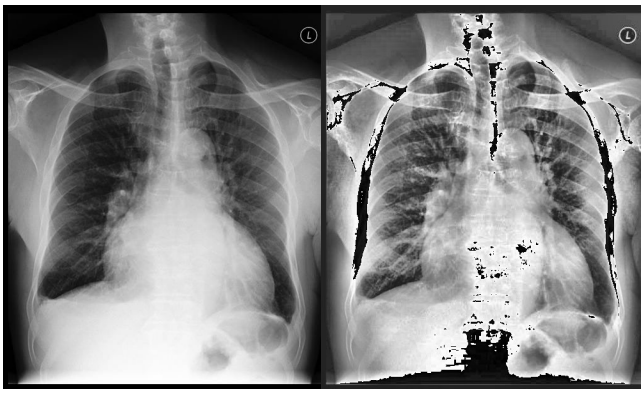
The paper MedFuse [3] is the most related work to ours. This research introduces a novel LSTM-based fusion module designed to integrate uni-modal and multimodal input, addressing challenges in multimodal fusion where data collection is asynchronous. MedFuse demonstrates significant performance improvements in in-hospital mortality prediction and phenotype classification tasks by using clinical time series data and chest X-ray images from the MIMIC-IV and MIMIC-CXR datasets. Unlike other fusion approaches, MedFuse treats multimodal representations as a sequence of uni-modal representations, performing even with partially paired data. However, the selected architecture is not the best-performing and robust solution. Their model also does not consider the uncertainty of the different tasks in multi-task phenotyping classification. Our proposed approach, including our architecture, attention mechanism, preprocessing methods, and the incorporation of uncertainty modeling through a loss function, enabled us to outperform their method in both noisy and noise-free environments.

III. MULTIMODAL DATA PREPARATION

This section describes the process of preparing our multimodal dataset. First, we will discuss the details of the MIMIC dataset (Section III-A), including its overall structure and relevant information for our analysis. Next, we will explore the specifics of the MIMIC-CXR subset (Section III-B), which focuses on chest X-ray images within the MIMIC dataset. Finally, we will review the MIMIC-IV dataset (Section III-C), which provides additional information relevant to our study. At the end (Section III-D), we'll discuss how the data from MIMIC-IV and MIMIC-CXR are combined to generate the multimodal dataset.

A. MIMIC DATASET

The Medical Information Mart for Intensive Care (MIMIC) is a large, freely available database containing healthcare data. MIMIC-IV (v2.2) [2], the latest version released in January 2023, incorporates data from patients admitted between the years 2008 and 2019. It improves upon numerous aspects of its predecessors by adopting a modular data organization approach, highlighting data provenance. This section provides a high-level overview of the MIMIC-IV dataset, highlighting its structure and relevant information for our analysis. MIMIC-IV's rich data allows researchers to explore a wide range of healthcare topics, including patient demographics, diagnoses, procedures, medications, laboratory measurements, vital signs, and even information from the online medical record system (e.g., height, and weight). Importantly, the data is deidentified using a strict protocol to protect patient privacy while still enabling valuable medical research.



(a) (b)

FIGURE 1. The effect of CLAHE augmentation: (a) original chest X-ray image, (b) CLAHE augmented image. CLAHE significantly improves the visibility of inner body parts, showcasing intricate details such as the kidney on the right side of the image. Additionally, it enhances the depiction of bone density, providing clearer insights. Such refined details play a crucial role in mortality prediction and phenotype classification.

B. MIMIC-CXR

MIMIC-CXR [1] is a valuable subset of MIMIC specifically focusing on chest X-ray images. This publicly available resource provides a rich dataset for researchers in medical image analysis. It contains 377,110 de-identified chest radiographs, including both frontal and lateral views captured during patients' hospital admissions. In this research paper, the multimodal dataset utilized combines MIMIC-CXR with MIMIC-IV, integrating chest X-ray images from MIMIC-CXR as the second modality. This fusion allows for a comprehensive dataset including both time series data and image modalities.

C. MIMIC-IV

MIMIC-IV database represents a significant advancement in medical data resources, building upon the success of MIMIC-III. Incorporating contemporary data from 2008 to 2019, MIMIC-IV is sourced from two in-hospital database systems, a custom hospital-wide Electronic Health Record (EHR) and an ICU-specific clinical information system. The latest version of this dataset includes the information of 299,712

patients, 431,231 admissions, and 73,181 ICU stays. This dataset is used to run experiments on the proposed model for two tasks; namely in-hospital mortality prediction, and phenotyping task. The latter includes 25 binary labels for predicting a range of diseases which can be categorized into groups of acute, mixed, and chronic diseases. For instance, the *chronic kidney diseases* label in the chronic group is considered for patients who have long-term damage to the kidneys, often progressive and irreversible, leading to impaired kidney function. As an acute disease, *complications of surgical procedures or medical care* are adverse events or complications arising from surgical or medical treatments and *other liver diseases* label determines the various liver disorders not classified under specific categories, including conditions like fatty liver disease and hepatitis. MIMIC-IV serves as a valuable resource for driving advancements in clinical informatics, epidemiology, and machine learning to improve patient care and outcomes. In our study, we have used this dataset to get the time series modality for our research.

D. GENERATING MULTIMODAL DATASET

As previously mentioned, we employ the MIMIC-IV dataset to assess and benchmark our models. To generate this dataset, we adopt the data extraction and preprocessing approach proposed in MedFuse. In the original dataset, each patient report may contain zero or more chest X-ray images taken during the patient's hospital stay, in addition to time series data. Our data generation process involves selecting the last captured image for each patient report and combining it with the associated time series data to create a sample. Using only the last image for each patient not only accelerates the training process and results in a more agile network, but also reduces resource consumption. Additionally, not all patients have multiple chest X-rays; some may only have a single image without a comprehensive healthcare report. Also, the latest image provides the most relevant information for the current state of the patient, while older images may introduce noise and hinder prediction accuracy. Table 1 presents a statistical analysis of the time difference, in hours, between the last CXR image taken and the recorded time series data for each patient. On average, the time series data was recorded 23 hours after the last X-ray image. Also, the maximum period length is 48 hours showcasing the maximum period for recording data for in-hospital mortality prediction task.

We utilize consistent dataset settings for reporting our results. Employing the patient identifier from the clinical time series data, we randomly partition the dataset into 70% for training, 10% for validation, and 20% for the test set. In our notation, we denote the clinical time series data as EHR and the chest X-ray images as CXR. The dataset is categorized into (EHR + CXR)PARTIAL, containing paired and partially paired samples (i.e., samples with missing chest X-rays), and (EHR + CXR)PAIRED, containing data samples where both modalities are present.

The training set for patient phenotyping, (EHR + CXR)PARTIAL, includes a total of 42,628 samples. Out of these, 7,756 samples are from the (EHR + CXR)PAIRED training set. The remaining samples in the (EHR + CXR)PARTIAL training set are made up of time series data for each sample. Chest X-ray images are extracted from MIMIC-CXR and split based on a random patient split. Images from the training set are then transferred to either the validation or test set if associated with patients in the validation or test splits of the clinical time series data. This procedure results in 325,188 images in the training set, 15,282 images in the validation set, and 36,625 images in the test set.

E. PREPROCESSING

We use the MedFuse data extraction and preprocessing procedure along with another preprocessing method for images known as CLAHE [31]. For chest X-ray images, a consistent set of transformations is applied during both pre-training and fine-tuning across all experiments and tasks. Specifically, each image is resized to 256 $\times$ 256 pixels, undergoes a random horizontal flip, and experiences various random affine transformations, including rotation, scaling, shearing, and translation. Subsequently, a random crop is applied to achieve an image size of 224 $\times$ 224 pixels. During the validation and testing phases, image resizing to 256 $\times$ 256 and a center crop to 224 $\times$ 224 pixels are performed.

To ensure fair comparisons and showcase the efficacy of multimodal learning, we utilize a consistent set of 17 clinical variables. Among these, five are categorical: capillary refill rate, glasgow coma scale eye opening, glasgow coma scale motor response, glasgow coma scale verbal response, and glasgow coma scale total. The remaining 12 are continuous variables: diastolic blood pressure, fraction of inspired oxygen, glucose, heart rate, height, mean blood pressure, oxygen saturation, respiratory rate, systolic blood pressure, temperature, weight, and pH. The input for all tasks is regularly sampled every two hours, with discretization and standardization of clinical variables following established protocols, as detailed in prior work.

After data preprocessing and one-hot encoding of categorical features, we obtain a vector representation of size 76 at each time step in the clinical time series data. For a given instance, the representation is denoted as $x_{ehr} \in \mathbb{R}^{t \times 76}$, where the value of t is dependent on the specific instance and task.

In addition, we explored a data preprocessing method known as CLAHE contrast enhancement. Developed as an extension of traditional histogram equalization, CLAHE provides a dynamic and localized approach to contrast enhancement. The method involves dividing an image into small, non-overlapping tiles and independently applying histogram equalization to each tile. This adaptive approach ensures that contrast enhancement is tailored to the unique characteristics of local regions, preventing the over-amplification of noise in homogeneous areas. We applied the CLAHE method to

TABLE 1. Statistical analysis of time differences between the last CXR image and time series data for each patient (in hours).

Mean	Std	Minimum	25% Percentile	50% Percentile	75% Percentile	Maximum
23.10	15.00	0.00	8.00	25.00	36.00	48.00

all images to enhance image quality, thereby facilitating the extraction of more information by the model.

In Figure 1 the impact of CLAHE on the image is visible. One of the primary benefits of CLAHE in chest X-ray imaging is its ability to enhance the visibility of lung parenchyma. In addition to lung pathology, CLAHE enhances the visualization of thoracic skeletal structures, including the ribs, vertebrae, and mediastinum. This improved depiction of bony anatomy is invaluable for detecting fractures, degenerative changes, and mediastinal masses, thereby assisting in the diagnosis of conditions ranging from traumatic injuries to neoplastic processes.

IV. PROPOSED FRAMEWORK

In the following sections, we will describe our proposed framework for multimodal healthcare analysis and prediction. We will explore the architecture of our model including the modality-specific encoders and the attention mechanism used for modality fusion (Section IV-A). Then we will go into more detail about our attention-based model in Section IV-B. Finally, we will explain the motives for using the uncertainty loss function and its advantages (Section IV-C).

Algorithm 1 Training of Model

Initialize model parameters: $\theta_{\text{enc\_ehr}}$, $\theta_{\text{enc\_cxr}}$, $\theta_{\text{projection}}$, θ_{fusion}, θ_{classify},σ_c
for each epoch **do**
  for each batch $(x_{\text{ehr}}, x_{\text{cxr}}, y)$ **do**
    $f_{\text{ehr}} \leftarrow \text{Enc}_{\text{ehr}}(x_{\text{ehr}}; \theta_{\text{enc\_ehr}})$
    $f'_{\text{cxr}} \leftarrow \text{Enc}_{\text{cxr}}(x_{\text{cxr}}; \theta_{\text{enc\_cxr}})$
    $f_{\text{cxr}} \leftarrow \text{projection}(f'_{\text{cxr}}; \theta_{\text{projection}})$
    $f \leftarrow [f_{\text{ehr}}, f_{\text{cxr}}]$
    $z \leftarrow \text{Fuse}(f; \theta_{\text{fusion}})$
    $\hat{y} \leftarrow \text{Classify}(z; \theta_{\text{classify}})$
    $L_c \leftarrow \ell(y, \hat{y})/\sigma_c^2 + \log(\sigma_c^2)$
    Backpropagate L_c and update $\theta_{\text{enc\_ehr}}$, $\theta_{\text{enc\_cxr}}$, $\theta_{\text{projection}}$, θ_{fusion}, σ_c
  end for
end for

A. MODEL ARCHITECTURE

Our proposed model shown in Figure 2 consists of two major parts: modality-specific encoders and a multimodal Transformer encoder [32] as our modality fusion network. We use an image encoder (e.g., a ResNet-34 model [33]) to extract features from our image modality and an LSTM network [34] to extract latent feature representations from our time series modality. If a modality is missing in our sample we input a zero matrix in its place to the respective unimodal

Algorithm 2 Inference of Model

Set model to evaluation mode
for each test sample $(x_{\text{ehr}}, x_{\text{cxr}})$ **do**
  $f_{\text{ehr}} \leftarrow \text{Enc}_{\text{ehr}}(x_{\text{ehr}})$
  $f'_{\text{cxr}} \leftarrow \text{Enc}_{\text{cxr}}(x_{\text{cxr}})$
  $f_{\text{cxr}} \leftarrow \text{Projection}(f'_{\text{cxr}})$
  $f \leftarrow [f_{\text{ehr}}, f_{\text{cxr}}]$
  $z \leftarrow \text{Fuse}(f)$
  $\hat{y} \leftarrow \text{Classify}(z)$
end for

encoder. We then utilize a projection layer to project the image embeddings to the time series embedding dimension. We then concatenate these feature representations and feed them to a Transformer encoder before feeding them to a final linear classifier to predict the labels.

In the first part, we pre-train each of our encoders with the unimodal data to independently extract meaningful representations from each of our modalities. An LSTM architecture works best for extracting feature embeddings from our time series data due to the quantity of available data and its consecutive nature. We use the Adam optimizer [35] to optimize our Binary Cross-Entropy losses and pre-train our modality-specific encoders.

We use a ResNet-34 model as the backbone for our image encoder and we set the output dimension of its classifier layer to be equal to the number of labels in our specific task. For our time series backbone, we use an LSTM network with $N = 2$ layers stacked on top of each other with a hidden dimension of $d = 256$ and a dropout layer with a dropout probability of $p = 0.3$. We also utilize a linear layer as the final classifier for our LSTM network.

In the second part, we remove the classification heads from our unimodal encoders and use latent feature embeddings f'_{cxr} and f_{ehr}. We feed f'_{cxr} to a fully connected projection layer to get f_{cxr} that has the same dimensionality as f_{ehr}. We then concatenate f_{ehr} and f_{cxr} to create the sequence f_{fused} that consists of our unimodal feature embeddings:

$$f_{\text{fused}} = [f_{\text{ehr}}, \text{projection}(f'_{\text{cxr}})]. \quad (1)$$

We use a Transformer Encoder without positional embeddings with a linear layer on top of that as our modality fusion network to resolve the issue of modality bias in our baselines. In models that use an LSTM network for fusion a major problem is the order of modality embeddings in the input sequence. Therefore, the sequence order may create a bias towards the modalities that come first in the input sequence, and changing it may vary results significantly when in reality the modalities do not possess a specific ordering.

By using an attention-based network for fusion our model learns the importance of every modality in each of our tasks and thus performs better than state-of-the-art LSTM or MLP architectures. Finally, we optimize the multi-task uncertainty loss (3) introduced in Kendall et al. [36] with an Adam optimizer to fine-tune our network. The steps of the training and inference processes for our model are outlined in algorithms 1 and 2, respectively.

B. ATTENTION-BASED MULTIMODAL FUSION

We feed our concatenated unimodal feature embedding f_{fused} to a Transformer Encoder network with $L = 2$ encoder layers stacked on top of one another, each of them having $h = 8$ heads and a feedforward dimension of $c = 1024$. We then feed the output of this network to a linear classifier to get the final predictions $\hat{y}$. Finally, we optimize the following multi-task uncertainty loss function with an Adam optimizer to fine-tune our model:

$$L(\hat{y}, y) = \text{UncertaintyLoss}(\hat{y}, y). \tag{2}$$

where y is the model ground truth and $\hat{y}$ is the model prediction.

Our attention-based model can learn the importance of each modality for each task simultaneously and find meaningful relations between the latent features of different modalities. This allows our model to deeply integrate different modalities to better understand our task and achieve more precise results.

C. UNCERTAINTY LOSS

A key issue with our baseline models is their dependency on the relative weights of each task's loss. For instance, in phenotype classification we have 25 different labels (each label corresponds to a separate task) and usually, the same weight is given to each of their losses, or the weights are manually selected. Giving the same weights to the losses of different tasks could negatively impact our model's performance due to the separate nature of each task. Manually choosing the relative weights is also a time-consuming ordeal that should be done for every single classification problem separately and requires vast resources.

The multi-task uncertainty loss shown in Figure 3 largely resolves this issue by weighing multiple losses and considering the homoscedastic uncertainty of each task. This method learns the weighing parameter σ for each of the losses during the training process. The multi-task uncertainty loss function for classification is the following:

$$\text{UncertaintyLoss}(\hat{y}, y) = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \text{BCE}(\hat{y}, y) + log\sigma_i^2, \tag{3}$$

where $\text{BCE}(\hat{y}, y)$ is the Binary Cross-Entropy loss and N is the number of tasks.

The term $log\sigma_i^2$ is included to constrain the value of the weight parameters. Without this term, the optimal solution to minimizing the uncertainty loss would simply involve setting all σ_i values to be as large as possible, which would not provide meaningful or accurate results.

By using this loss to fine-tune our model and learn its parameters while simultaneously learning the value of σ for each task, we were able to boost the performance of our model in phenotype classification and achieve better results.

V. EXPERIMENTS AND RESULTS

In this section, we explain our experiments. Starting from the complex architecture described in the method section, we explain how we put it into practice with a detailed demonstration of our experimental setup and baseline models. We experimented with our method on two important medical tasks: in-hospital mortality prediction and phenotype classification. After analyzing these tasks, we break down the different parts of our approach to see how each contributes. Additionally, we investigate uncertainties and test the model's robustness.

A. EXPERIMENTAL SETUP

In this study, we establish our experimental framework using the MIMIC-IV and the MIMIC-CXR datasets. Our choice of MIMIC is based on its extensive scale, comprehensive documentation and standardized formatting. Our experiments focus on two key objectives:

1) Predicting the binary in-hospital mortality label after the first 48 hours of a patient's ICU stay.
2) Classify a set of 25 phenotype labels for the patients during their ICU stays.

We train our proposed network separately for each task and evaluate the results. We use a batch size of 16 for our data loader. For each task, we first pre-train our ResNet-34 encoder with images from the MIMIC-CXR dataset, and pre-train the LSTM encoder network with time series data from the MIMIC-IV dataset.

For the phenotyping task, we use learning rates of 5×10^{-4} and 1×10^{-4} for pre-training the image and time series encoders, respectively and for the in-hospital mortality task, we use learning rates of 5×10^{-4} and 3×10^{-5} for pre-training the image and time series encoders. Then we jointly fine-tune the fusion module with the encoders, using a learning rate of 7×10^{-5} for the phenotyping task and 1×10^{-4} for the in-hospital mortality task. We report final results on test sets and compute 95% confidence intervals with 1000 iterations via the bootstrap method [37].

B. BASELINES

MedFuse [3] is a distinctive approach involving the utilization of a fusion module based on an LSTM architecture. This module effectively combines information from both image and time series data. Fine-tuning of the MedFuse model follows a two-step pre-training process. Initially, the image encoder is pre-trained on 14 radiology labels, focusing on the detection of a specific disease in unpaired chest X-rays. Simultaneously, the LSTM is pre-trained using unpaired time

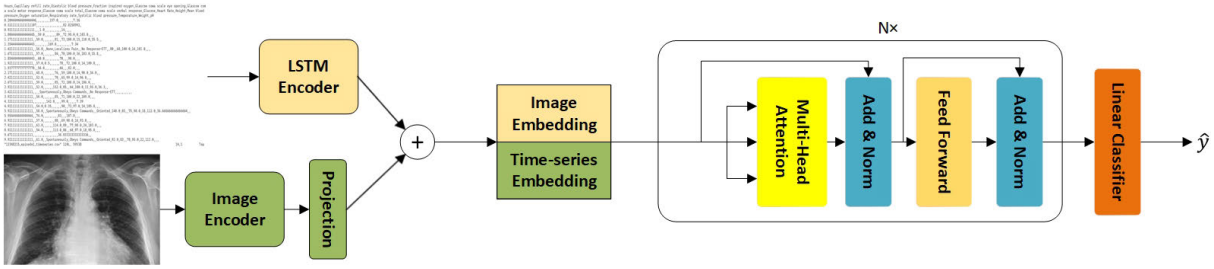


FIGURE 2. Model architecture consisting of modality-specific encoders and a multilayer transformer encoder as our multimodal fusion network.

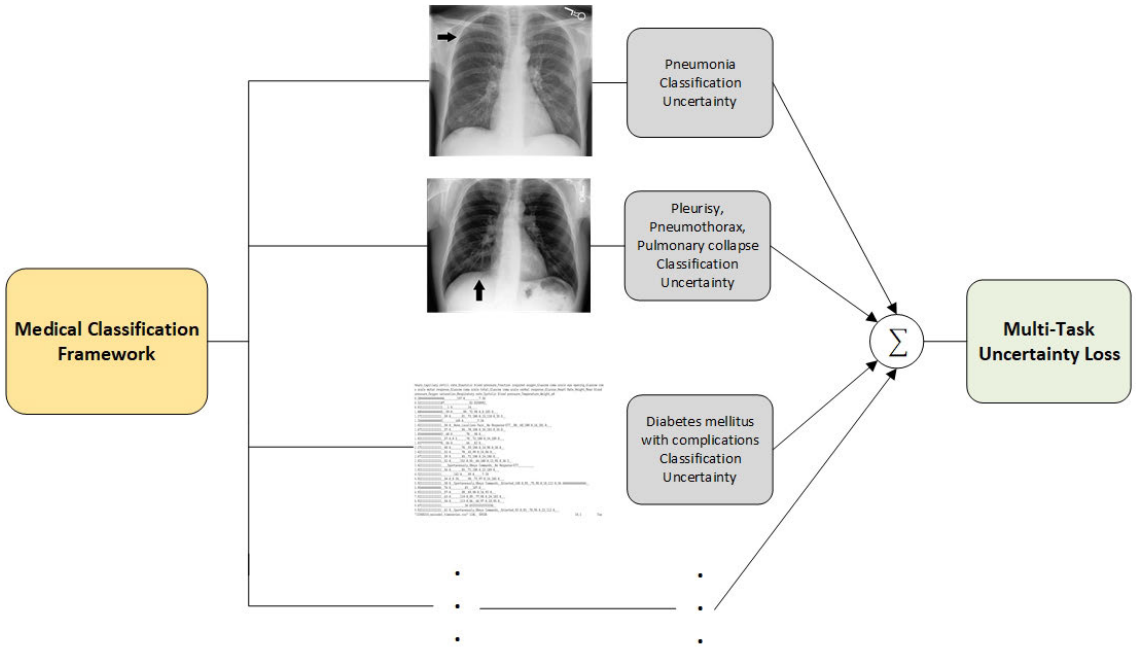


FIGURE 3. We combine and weigh multiple losses according to the uncertainty of each task to compute the multi-task uncertainty loss.

series data for in-hospital mortality prediction or phenotype classification. It is worth noting that, despite the use of unpaired data during pre-training, MedFuse requires labels for a distinct task in the image modality. For experimental evaluation, the publicly available MedFuse model is applied to the multimodal MIMIC dataset. The dataset is partitioned into similar splits as those employed in our work and other baseline models, ensuring a fair comparison.

Contrastive-based [38] Contrastive learning is a machine learning paradigm that aims to teach a model the differences and similarities between different data modalities. The fundamental idea behind contrastive learning is to embed similar samples closer to each other in a latent space while pushing dissimilar samples apart. In establishing this baseline, we have adopted an approach similar to that in [39], wherein we construct a model with 2 heads, using a ViT-Base image encoder [40] and an LSTM time series encoder as the backbone of the model. One header is designated for inter-modality optimization, while the other optimizes the intra-modality loss. For every pair of image and time series data, we apply random augmentations and try to maximize the similarity of the data and the augmented version. We also employ the other header to align the representations of the image and time series more closely.

Diffusion-based classifier In the domain of classifiers utilizing diffusion mechanisms, our literature review revealed an absence of pre-existing multimodal diffusion-based classifiers. We extended the recent unimodal diffusion-based classifier CARD [41] to a multimodal version. The original architecture employs an encoder(ResNet-34) to convert image data samples into prior vectors. Subsequently, the diffusion backward process, facilitated by a denoising deep neural network, aims to denoise these prior vectors, refining them into more accurate feature vectors for classification. The final denoised prior vectors are then used for classification. To create a multimodal diffusion-based classifier, we replaced the CARD encoder with an encoder concatenating the separate embeddings of image and time series data.

C. IN-HOSPITAL MORTALITY PREDICTION

For the critical task of in-hospital mortality prediction, we evaluate the performance of our proposed model against established baselines, including MedFuse, CARD, and Contrastive Learning. Table 2 presents a comparative analysis of the macro average F1-score, binary F1-score, AUROC, and AUPRC metrics as well as the parameter count of each model. Our model consistently outperforms state-of-the-art approaches, demonstrating its efficacy in predicting mortality across diverse modalities.

For this experiment we use time series data and chest X-ray images, to predict the in-hospital mortality of the patients according to the first 48 hours of ICU stay in a binary classification task. There are 18845 samples in our training set, 2138 samples in our validation set and 5243 samples in our test set. All samples include time series data, but 4885, 540 and 1373 samples have both modalities in the training, validation and test sets, respectively. The rest of the samples in each set have missing chest X-ray images.

Upon reviewing the performance of the models, it's clear that the CARD model falls short, demonstrating the lowest performance among all our baselines. The contrastive model, which employs the ViT architecture, surpasses MedFuse across all metrics, with the exception of AUROC. Our multimodal attention-based model outperforms MedFuse and CARD on all metrics and surpasses the contrastive model on macro average F1-score and AUROC. By training our proposed model with the CLAHE augmentation on images, we achieve superior results in all metrics excluding the macro average F1-score, where the filter slightly decreases the performance of our attention-based model.

D. PHENOTYPE CLASSIFICATION

In the domain of phenotyping, our model is evaluated against baselines, including MedFuse, CARD, and Contrastive learning. Table 3 illustrates the proficiency of our models in addressing the complex task of multi-label phenotyping and the number of parameters per each model. Metrics such as macro average F1-score, binary F1-score, AUROC, and AUPRC are employed to evaluate the model's accuracy in capturing diverse phenotypic characteristics. The results showcase the model's ability to handle the intricacies of multi-label phenotyping, outperforming the baselines.

The objective of this experiment is to predict 25 different conditions given to patients during the length of their ICU stay. We utilize time series data and chest X-ray images for this multi-label classification task. There are 42628 samples in our training set, 4802 samples in our validation set and 11914 samples in our test set. All samples include time series data, but 7756, 882 and 2166 samples have both modalities in the training, validation and test sets. The rest of the samples in each set have missing chest X-ray images.

When we compare the models, we find that the MedFuse model performs better than the CARD model in all metrics, with the exception of the binary F1-score. Our attention-based model trained on images augmented by the CLAHE filter surpasses CARD and MedFuse in macro average F1-score and AUROC but it has a lower binary F1-score compared to CARD. When we optimize our model with the uncertainty loss it outperforms all of our baselines in all metrics except for AUROC, where it appears that combining uncertainty loss with the CLAHE augmentation slightly decreases the performance of our model. It's important to note that all the best results across various metrics were achieved by our models.

E. ABLATION STUDY

The ablation study examines the model's architecture, exploring distinctive configurations to understand their impact. These configurations include variations such as using time series only, images only and fusing image and time series data through LSTM or attention mechanisms. The results are shown in Table 4, providing insights into the contributions of the model parts.

In our primary observation, the LSTM-fused multimodal model demonstrates superior performance in phenotype classification compared to uni-modal models across all metrics. This outcome underscores the effectiveness of the multimodal approach in fusing the information derived from both modalities for classification purposes. Subsequently, the attention-based fusion model surpasses the LSTM-fused model, emphasizing the contribution of the transformer layers in enhancing the multimodal model's classification performance.

Incorporating uncertainty loss into the model yields improved performance, particularly evident in AUPRC and AUROC metrics, without significantly impacting the macro F1-score. A minor decrease (approximately 0.003) is observed in the binary F1-score. In the end, we analyzed the impact of the CLAHE filter on CXR images. Results indicate that using CLAHE improves the performance of our attention-based model across all metrics. Also, adding this filter increases our attention-based uncertainty model's performance across all metrics except for AUROC.

In summary, the comprehensive analysis presented in this table highlights the positive impact of the modules and methods employed, collectively contributing to improved model performance, particularly evident in terms of AUPRC and AUROC metrics.

F. TASK-WISE UNCERTAINTIES

To understand the importance of utilizing homoscedastic uncertainty in our multi-label phenotyping task, we have compared the performance of our attention-based model fine-tuned using the uncertainty loss with the same model fine-tuned using the Cross-Entropy loss.

In table 7, The uncertainty loss not only enhances the average performance of our proposed framework but more importantly, this loss function increases the individual

TABLE 2. In-hospital mortality prediction performance.

Model	Macro Average F1-score	Binary F1-score	AUROC	AUPRC	Number of Parameters
MedFuse	0.677 (0.662, 0.693)	0.412 (0.374, 0.450)	0.857 (0.842, 0.871)	0.507 (0.470, 0.543)	23869264
Contrastive + ViT	0.690 (0.660, 0.716)	0.441 (0.383, 0.490)	0.852 (0.840, 0.870)	0.520 (0.470, 0.564)	87877496
CARD	0.660 (0.645, 0.675)	0.400 (0.385, 0.415)	0.690 (0.681, 0.699)	0.341 (0.329, 0.353)	52180932
Attention	**0.691** (0.669, 0.712)	0.438 (0.399, 0.476)	0.857 (0.842, 0.872)	0.514 (0.476, 0.556)	23871824
Attention + CLAHE	0.685 (0.653, 0.710)	**0.453** (0.432, 0.479)	**0.858** (0.831, 0.886)	**0.524** (0.457, 0.594)	23871824

TABLE 3. Phenotype classification performance.

Model	Macro Average F1-score	Binary F1-score	AUROC	AUPRC	Number of Parameters
MedFuse	0.589 (0.567, 0.607)	0.282 (0.264, 0.301)	0.763 (0.752, 0.775)	0.422 (0.401, 0.445)	23893387
CARD	0.585 (0.565, 0.605)	0.344 (0.324, 0.364)	0.600 (0.589, 0.611)	0.310 (0.298, 0.322)	52180932
Attention + CLAHE	0.611 (0.600, 0.622)	0.327 (0.307, 0.347)	**0.770** (0.758, 0.781)	0.431 (0.409, 0.454)	23895947
Attention + Uncertainty + CLAHE	**0.614** (0.589, 0.639)	**0.362** (0.330, 0.396)	0.759 (0.742, 0.776)	**0.466** (0.433, 0.502)	23895972

TABLE 4. Ablation study results.

Model	Macro Average F1-score	Binary F1-score	AUROC	AUPRC
Time series only	0.581 (0.564, 0.599)	0.270 (0.239, 0.301)	0.759 (0.747, 0.772)	0.421 (0.400, 0.444)
Image only	0.535 (0.519, 0.550)	0.200 (0.176, 0.225)	0.670 (0.641, 0.700)	0.358 (0.323, 0.400)
Multimodal LSTM Fusion	0.589 (0.580, 0.599)	0.282 (0.266, 0.299)	0.763 (0.752, 0.775)	0.422 (0.401, 0.445)
Multimodal Attention	0.604 (0.594, 0.615)	0.314 (0.295, 0.332)	0.765 (0.753, 0.777)	0.424 (0.402, 0.447)
Attention + Uncertainty Loss	0.604 (0.593, 0.615)	0.311 (0.292, 0.331)	0.767 (0.755, 0.779)	0.427 (0.405, 0.450)
Attention + CLAHE	0.611 (0.600, 0.622)	0.327 (0.307, 0.347)	**0.770** (0.758, 0.781)	0.431 (0.409, 0.454)
Attention + Uncertainty + CLAHE	**0.614** (0.589, 0.639)	**0.362** (0.330, 0.396)	0.759 (0.742, 0.776)	**0.466** (0.433, 0.502)

TABLE 5. Robustness experiment results on models trained on noisy data.

Model	Percentage of Noise	AUROC	AUPRC	Macro Average F1-score	Binary F1-score
Attention	%10	**0.768** (0.750, 0.785)	**0.426** (0.394, 0.460)	**0.601** (0.585, 0.618)	**0.306** (0.278, 0.335)
MedFuse	%10	0.765 (0.747, 0.783)	0.422 (0.390, 0.456)	0.585 (0.570, 0.600)	0.274 (0.247, 0.300)
Attention	%20	**0.763** (0.745, 0.780)	**0.423** (0.390, 0.457)	**0.600** (0.588, 0.623)	**0.308** (0.284, 0.349)
MedFuse	%20	0.761 (0.743, 0.779)	0.420 (0.387, 0.455)	0.590 (0.579, 0.616)	0.281 (0.234, 0.334)
Attention	%30	0.751 (0.728, 0.769)	**0.403** (0.366, 0.436)	**0.610** (0.589, 0.628)	**0.322** (0.292, 0.353)
MedFuse	%30	**0.757** (0.739, 0.775)	0.410 (0.379, 0.445)	0.591 (0.573, 0.616)	0.288 (0.269, 0.316)
Attention	%40	**0.763** (0.744, 0.780)	**0.421** (0.388, 0.454)	**0.602** (0.587, 0.619)	**0.310** (0.281, 0.340)
MedFuse	%40	0.757 (0.747, 0.767)	0.414 (0.400, 0.430)	0.594 (0.586, 0.602)	0.295 (0.283, 0.308)
Attention	%50	**0.761** (0.743, 0.779)	**0.417** (0.385, 0.445)	**0.594** (0.578, 0.609)	**0.291** (0.263, 0.320)
MedFuse	%50	0.757 (0.738, 0.775)	0.410 (0.377, 0.444)	0.583 (0.568, 0.598)	0.270 (0.244, 0.297)
Attention	%60	0.755 (0.736, 0.773)	0.412 (0.380, 0.446)	**0.602** (0.586, 0.618)	**0.312** (0.284, 0.341)
MedFuse	%60	0.755 (0.737, 0.773)	0.412 (0.380, 0.445)	0.580 (0.569, 0.598)	0.271 (0.246, 0.297)

TABLE 6. Robustness experiment results on models trained on noise-free data.

Model	Percentage of Noise	AUROC	AUPRC	Macro Average F1-score	Binary F1-score
Attention	%10	**0.761** (0.743, 0.779)	**0.418** (0.385, 0.452)	**0.590** (0.581, 0.612)	**0.299** (0.271, 0.327)
MedFuse	%10	0.757 (0.739, 0.775)	0.412 (0.380, 0.445)	0.581 (0.561, 0.605)	0.284 (0.257, 0.311)
Attention	%20	0.737 (0.719, 0.755)	0.387 (0.355, 0.420)	0.581 (0.570, 0.598)	**0.279** (0.254, 0.304)
MedFuse	%20	**0.743** (0.710, 0.773)	**0.395** (0.346, 0.440)	**0.559** (0.536, 0.578)	0.239 (0.213, 0.260)
Attention	%30	**0.757** (0.739, 0.774)	**0.411** (0.379, 0.445)	**0.581** (0.566, 0.598)	**0.268** (0.240, 0.296)
MedFuse	%30	0.753 (0.734, 0.772)	0.404 (0.373, 0.434)	0.557 (0.543, 0.571)	0.220 (0.196, 0.245)
Attention	%40	**0.743** (0.724, 0.762)	0.396 (0.365, 0.430)	**0.570** (0.556, 0.586)	**0.246** (0.221, 0.272)
MedFuse	%40	0.724 (0.704, 0.743)	0.369 (0.339, 0.402)	0.557 (0.544, 0.571)	0.202 (0.179, 0.226)
Attention	%50	**0.732** (0.713, 0.750)	**0.379** (0.349, 0.411)	**0.561** (0.545, 0.575)	**0.245** (0.220, 0.269)
MedFuse	%50	0.717 (0.698, 0.736)	0.363 (0.334, 0.394)	0.546 (0.533, 0.560)	0.204 (0.181, 0.228)
Attention	%60	**0.689** (0.670, 0.709)	**0.334** (0.306, 0.364)	**0.540** (0.528, 0.552)	**0.195** (0.174, 0.217)
MedFuse	%60	0.674 (0.653, 0.694)	0.315 (0.290, 0.344)	0.510 (0.497, 0.522)	0.153 (0.135, 0.173)

performance on most labels in our multi-label setup. This loss function allows our model to focus more on tasks that are easier to predict and have less uncertainty without causing a significant decrease in performance in the less certain and more complicated tasks, thus achieving a higher average performance. Although the overall improvements might not be considered drastic, the constant improvement in different tasks, especially those that had high performances prior to using the uncertainty loss, further supports the claim that the loss function makes our model more flexible and allows it to

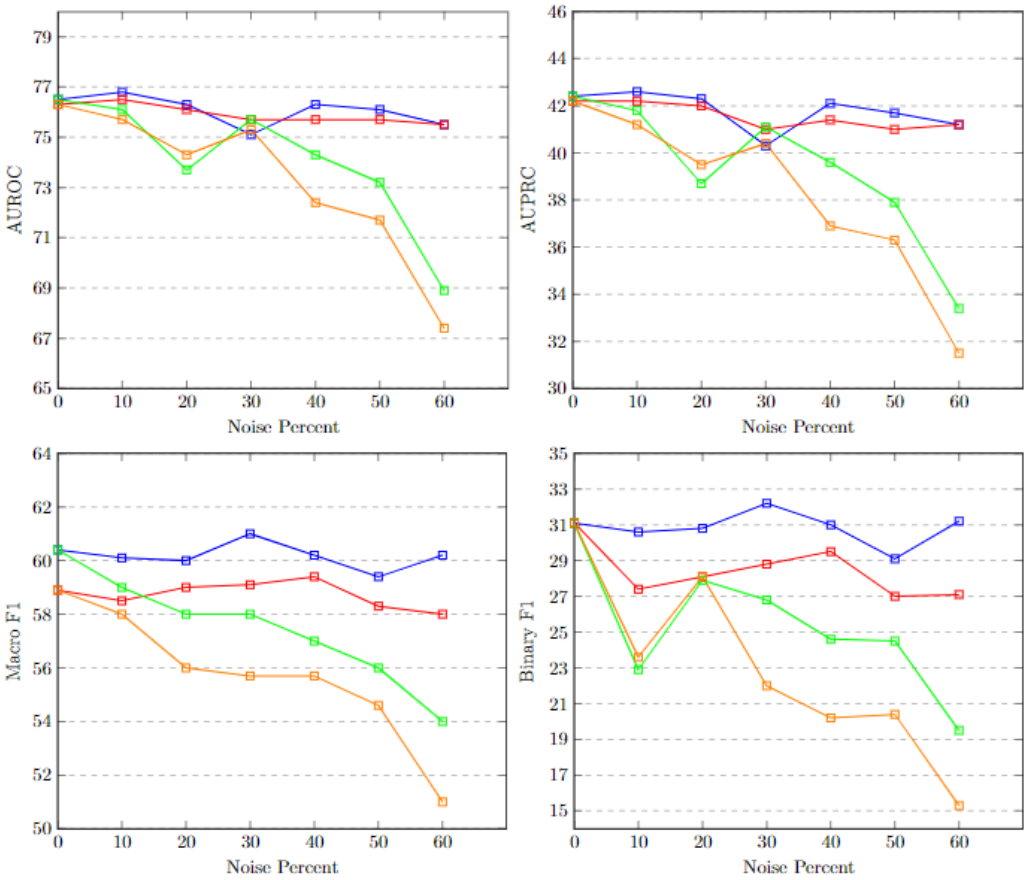


FIGURE 4. Performance comparison of models trained on noisy or noise-free datasets, and evaluated on noisy datasets. The plot employs different colors to represent specific configurations: blue indicates the attention-based fusion model trained on noisy data; red shows the MedFuse model trained on noisy data; green denotes the attention-based fusion trained on noise-free data; and orange represents the MedFuse model trained on noise-free data. As noise levels increase, a general decline in performance is observed for all models across various metrics. Notably, the use of attention mechanisms appears to mitigate performance degradation, showcasing enhanced robustness against noise.

focus on more certain tasks, and indicates the importance of using this loss function in our multi-label setup.

G. ROBUSTNESS

In order to explore the robustness of our attention-based model, we compared the performance of our model against MedFuse in noisy configurations. To do so, we prepared a noisy version of the multimodal MIMIC dataset. Table 5 presents the results across various noise levels, ranging from 10% to 60%, on both the training and testing sets. In this experiment, we subject all samples to varying levels of noise. For instance, in the case of images, we introduce noise by perturbing 10% of the pixels within the data. Similarly, for the time series data, we apply noise to 10% of the time steps in each sample. The introduced noise is Gaussian and its mean and standard deviation are estimated by measuring these parameters in 1000 random samples of the data. These parameters are calculated individually for each feature in the time series data and all pixels in the images.

Two different modes are considered for the evaluation process. In one, models are trained and tested on noisy datasets. In the other, testing is conducted on noisy datasets without any prior training on noisy datasets. The latter scenario is very common in real-world applications, where unexpected noise is often encountered in the data. The results for each setting are presented in Table 5 and Table 6, while the corresponding Figure 4 mirrors the tabulated results. It is evident that as noise levels increase, the model accuracy decreases. However, the attention-based fusion results in superior overall performance compared to the MedFuse model.

Unlike the MedFuse model, which exhibits significant performance degradation in the presence of noise, our attention-based model demonstrates high levels of robustness. For instance, even at 60% noise levels, the decrease in performance is minimal, with results showcasing as little as a 2% reduction in performance. These results underscore the efficacy of our attention mechanism in mitigating the effects of noise, ensuring consistent

TABLE 7. Task-wise uncertainty impacts.

	Attention with Uncertainty		Attention without Uncertainty	
Phenotype	AUROC	AUPRC	AUROC	AUPRC
Acute and unspecified renal failure	0.793 (0.784, 0.801)	**0.592** (0.575, 0.610)	0.793 (0.784, 0.801)	0.590 (0.573, 0.608)
Acute cerebrovascular disease	0.908 (0.895, 0.919)	**0.468** (0.429, 0.510)	0.908 (0.895, 0.918)	0.462 (0.425, 0.501)
Acute myocardial infarction	0.760 (0.745, 0.774)	0.216 (0.192, 0.242)	0.762 (0.746, 0.776)	0.221 (0.199, 0.248)
Cardiac dysrhythmias	0.681 (0.671, 0.692)	**0.484** (0.469, 0.502)	0.681 (0.671, 0.691)	0.482 (0.466, 0.499)
Chronic kidney disease	**0.746** (0.735, 0.757)	**0.439** (0.418, 0.460)	0.736 (0.725, 0.747)	0.429 (0.412, 0.450)
Chronic obstructive pulmonary disease and bronchiectasis	**0.704** (0.690, 0.715)	**0.292** (0.271, 0.314)	0.701 (0.688, 0.715)	0.289 (0.270, 0.311)
Complications of surgical procedures or medical care	**0.731** (0.719, 0.742)	**0.406** (0.384, 0.428)	0.730 (0.719, 0.742)	0.405 (0.384, 0.425)
Conduction disorders	**0.718** (0.702, 0.733)	0.246 (0.227, 0.269)	0.717 (0.702, 0.732)	0.248 (0.226, 0.272)
Congestive heart failure; nonhypertensive	**0.765** (0.755, 0.775)	**0.522** (0.502, 0.541)	0.760 (0.751, 0.770)	0.516 (0.497, 0.534)
Coronary atherosclerosis and other heart disease	**0.761** (0.753, 0.771)	**0.591** (0.573, 0.609)	0.760 (0.751, 0.769)	0.588 (0.571, 0.606)
Diabetes mellitus with complications	**0.900** (0.892, 0.908)	**0.595** (0.569, 0.622)	0.897 (0.889, 0.906)	0.592 (0.565, 0.620)
Diabetes mellitus without complication	**0.789** (0.778, 0.799)	**0.413** (0.393, 0.436)	0.787 (0.777, 0.797)	0.406 (0.387, 0.428)
Disorders of lipid metabolism	0.704 (0.694, 0.714)	**0.619** (0.604, 0.635)	0.706 (0.696, 0.714)	0.617 (0.602, 0.632)
Essential hypertension	**0.668** (0.658, 0.678)	**0.579** (0.563, 0.595)	0.660 (0.651, 0.670)	0.570 (0.555, 0.585)
Fluid and electrolyte disorders	0.759 (0.750, 0.768)	0.644 (0.628, 0.659)	0.760 (0.751, 0.769)	0.644 (0.629, 0.659)
Gastrointestinal hemorrhage	0.772 (0.757, 0.786)	0.213 (0.192, 0.242)	0.772 (0.758, 0.786)	0.215 (0.191, 0.242)
Hypertension with complications and secondary hypertension	**0.738** (0.726, 0.748)	**0.432** (0.411, 0.452)	0.728 (0.717, 0.739)	0.421 (0.404, 0.441)
Other liver diseases	**0.741** (0.727, 0.752)	0.297 (0.276, 0.319)	0.737 (0.724, 0.750)	0.307 (0.284, 0.331)
Other lower respiratory disease	0.655 (0.639, 0.672)	0.165 (0.151, 0.182)	0.657 (0.641, 0.673)	0.169 (0.154, 0.188)
Other upper respiratory disease	**0.752** (0.727, 0.775)	**0.261** (0.224, 0.303)	0.747 (0.724, 0.769)	0.249 (0.211, 0.289)
Pleurisy; pneumothorax; pulmonary collapse	0.708 (0.692, 0.727)	0.155 (0.140, 0.177)	0.714 (0.696, 0.732)	0.158 (0.143, 0.180)
Pneumonia	0.**813** (0.802, 0.824)	**0.384** (0.360, 0.410)	0.811 (0.799, 0.821)	0.382 (0.358, 0.408)
Respiratory failure; insufficiency; arrest (adult)	**0.872** (0.865, 0.880)	0.565 (0.542, 0.590)	0.871 (0.863, 0.879)	0.565 (0.540, 0.589)
Septicemia (except in labor)	**0.846** (0.838, 0.855)	**0.522** (0.499, 0.547)	0.842 (0.833, 0.852)	0.512 (0.489, 0.538)
Shock	**0.890** (0.882, 0.898)	**0.569** (0.542, 0.595)	0.888 (0.880, 0.895)	0.552 (0.526, 0.579)
Average	**0.767** (0.755, 0.779)	**0.427** (0.405, 0.450)	0.765 (0.753, 0.776)	0.424 (0.402, 0.447)

and reliable performance across diverse environmental conditions.

VI. CONCLUSION

A. SOCIAL IMPLEMENTATION AND ETHICAL CONSIDERATIONS

As our work involves patient data, the ethical and social implications of deploying multimodal deep neural networks in clinical settings must be carefully considered. In real-world applications, the use of sensitive patient data, such as medical images and time series, necessitates strict adherence to privacy regulations, including data anonymization and compliance with healthcare data standards. Furthermore, ensuring that the model's predictions do not inadvertently perpetuate biases present in the data is crucial to avoid ethical pitfalls. By incorporating uncertainty estimation in our model, we aim to provide more transparent decision-making, giving clinicians the ability to understand and appropriately weigh the model's confidence in its predictions, thereby enhancing trust and accountability in clinical environments.

From a broader societal perspective, the deployment of such models introduces both opportunities and challenges. While improving clinical decision support systems has the potential to significantly enhance patient outcomes, the scalability and integration of these models into existing healthcare infrastructures present technical and ethical challenges. Practical issues such as the need for continuous model monitoring, maintaining data security in deployment, and ensuring equitable access to the benefits of advanced AI systems across diverse populations are central to the responsible application of our approach. Addressing these concerns is essential for ensuring that the benefits of AI-driven healthcare can be realized without compromising ethical standards or patient trust.

B. SUMMARY AND FUTURE DIRECTIONS

Our research introduces an innovative approach to multimodal deep neural networks, specifically designed for integrating heterogeneous data modalities such as images and time series data in mortality prediction and phenotyping label assignments. By employing dedicated encoders for each modality, our model effectively captures nuanced patterns inherent in both visual and temporal information, thus enhancing predictive capabilities. Our experiments conducted under noisy settings demonstrate the robustness of our model, surpassing state-of-the-art methods and showcasing its efficacy in handling real-world challenges with noisy clinical data. Additionally, our innovative use of an uncertainty loss function addresses the complexity of multi-label classification, contributing to improved model performance. Furthermore, the integration of attention mechanisms for modality fusion enhances adaptability by dynamically allocating attention based on task relevance. In summary, our research advances robust multimodal deep learning for clinical applications, offering a flexible and robust framework capable of addressing challenges in real-world clinical data, with promising outcomes for clinical decision support systems.

Looking ahead, several promising directions for future research emerge from the outcomes of this study. First and foremost, enhancing the interpretability of the multimodal deep neural network remains a critical area of investigation. Developing methodologies that shed light on the decision-making process of the model will be essential for building trust in clinical applications. This may involve exploring novel visualization techniques or model-agnostic interpretability tools to dissect how the network integrates information from diverse modalities, providing clinicians with valuable insights into its decision rationale.

Our research sets the stage for exploring the integration of additional data modalities into our multimodal deep neural network framework. While our current focus has been on combining images and time series data for mortality prediction and phenotyping, there's immense potential in extending our approach to include other modalities such as textual data or genomic information. By incorporating these additional modalities, we can investigate how our model performs across a broader spectrum of clinical data and further enhance its predictive capabilities. This exploration opens up avenues for understanding how different types of data interact and contribute to the overall predictive power of our model, helping to have more comprehensive and robust clinical decision support systems. Additionally, addressing why certain tasks introduce more uncertainty for the network, particularly focusing on whether the uncertainty and also the difficulty of each task is primarily driven by image modalities or time series data, will be important. Further research could also explore how different data types contribute to overall model performance, and examine more deeply the sources of uncertainty in various modalities.

REFERENCES

[1] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019, *arXiv:1901.07042*.

[2] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-W. H. Lehman, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," *Sci. Data*, vol. 10, no. 1, Jan. 2023, Art. no. 1.

[3] N. Hayat, K. J. Geras, and F. E. Shamout, "MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images," in *Proc. Mach. Learn. Healthcare Conf. (MLHC)*, vol. 182, Durham, NC, USA, Z. C. Lipton, R. Ranganath, M. P. Sendak, M. W. Sjoding, and S. Yeung, Eds., Aug. 2022, pp. 479–503.

[4] S. Qiu et al., "Multimodal deep learning for Alzheimer's disease dementia assessment," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 3404.

[5] N. Rahim, S. El-Sappagh, S. Ali, K. Muhammad, J. D. Ser, and T. Abuhmed, "Prediction of Alzheimer's progression based on multimodal deep-learning-based fusion and visual explainability of time-series data," *Inf. Fusion*, vol. 92, pp. 363–388, Apr. 2023.

[6] K. Niu, K. Zhang, X. Peng, Y. Pan, and N. Xiao, "Deep multi-modal intermediate fusion of clinical record and time series data in mortality prediction," *Frontiers Mol. Biosci.*, vol. 10, Mar. 2023, Art. no. 1136071.

[7] L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. V. Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas, "Integrated multimodal artificial intelligence framework for healthcare applications," *NPJ Digit. Med.*, vol. 5, no. 1, Sep. 2022, Art. no. 149.

[8] D. Nie, J. Lu, H. Zhang, E. Adeli, J. Wang, Z. Yu, L. Liu, Q. Wang, J. Wu, and D. Shen, "Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages," *Sci. Rep.*, vol. 9, no. 1, p. 1103, Jan. 2019.

[9] B. Srinivas and G. S. Rao, "Segmentation of multi-modal MRI brain tumor sub-regions using deep learning," *J. Electr. Eng. Technol.*, vol. 15, no. 4, pp. 1899–1909, Jul. 2020.

[10] D. Muduli, R. Dash, and B. Majhi, "Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 102825.

[11] R. A. Khan, M. Fu, B. Burbridge, Y. Luo, and F.-X. Wu, "A multi-modal deep neural network for multi-class liver cancer diagnosis," *Neural Netw.*, vol. 165, pp. 553–561, Aug. 2023.

[12] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 841–850, May 2019.

[13] S. Joo, E. S. Ko, S. Kwon, E. Jeon, H. Jung, J.-Y. Kim, M. J. Chung, and Y.-H. Im, "Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer," *Sci. Rep.*, vol. 11, no. 1, Sep. 2021, Art. no. 18800.

[14] Z. Zeng, L. Yao, A. Roy, X. Li, S. Espino, S. E. Clare, S. A. Khan, and Y. Luo, "Identifying breast cancer distant recurrences from electronic health records using machine learning," *J. Healthcare Informat. Res.*, vol. 3, no. 3, pp. 283–299, Sep. 2019.

[15] G. Harerimana, J. W. Kim, H. Yoo, and B. Jang, "Deep learning for electronic health records analytics," *IEEE Access*, vol. 7, pp. 101245–101259, 2019.

[16] S. Daberdaku, E. Tavazzi, and B. Di Camillo, "A combined interpolation and weighted K-nearest neighbours approach for the imputation of longitudinal ICU laboratory data," *J. Healthcare Informat. Res.*, vol. 4, no. 2, pp. 174–188, Jun. 2020.

[17] D. Zikos, A. Shrestha, and L. Fegaras, "A cross-sectional study to predict mortality for medicare patients based on the combined use of HCUP tools," *J. Healthcare Informat. Res.*, vol. 5, no. 3, pp. 300–318, Sep. 2021.

[18] A. Guo, N. R. Mazumder, D. P. Ladner, and R. E. Foraker, "Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0256428.

[19] J. Jeon, P. J. Leimbigler, G. Baruah, M. H. Li, Y. Fossat, and A. J. Whitehead, "Predicting Glycaemia in type 1 diabetes patients: Experiments in feature engineering and data imputation," *J. Healthcare Informat. Res.*, vol. 4, no. 1, pp. 71–90, Mar. 2020.

[20] A. Mir and S. N. Dhage, "Diabetes disease prediction using machine learning on big data of healthcare," in *Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2018, pp. 1–6.

[21] V. J. Gogi and M. N. Vijayalakshmi, "Prognosis of liver disease: Using machine learning algorithms," in *Proc. Int. Conf. Recent Innov. Electr., Electron. Commun. Eng. (ICRIEECE)*, Jul. 2018, pp. 875–879.

[22] N. G. Maity and S. Das, "Machine learning for improved diagnosis and prognosis in healthcare," in *Proc. IEEE Aerosp. Conf.*, Mar. 2017, pp. 1–9.

[23] S. Niu, Q. Yin, Y. Song, Y. Guo, and X. Yang, "Label dependent attention model for disease risk prediction using multimodal electronic health records," 2022, *arXiv:2201.06779*.

[24] M. N. I. Suvon, P. C. Tripathi, S. Alabed, A. J. Swift, and H. Lu, "Multimodal learning for predicting mortality in patients with pulmonary arterial hypertension," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Las Vegas, NV, USA, D. A. Adjeroh, Q. Long, X. M. Shi, F. Guo, X. Hu, S. Aluru, G. Narasimhan, J. Wang, M. Kang, A. Mondal, and J. Liu, Eds., Dec. 2022, pp. 2704–2710.

[25] T. Zhu, K. Li, J. Chen, P. Herrero, and P. Georgiou, "Dilated recurrent neural networks for glucose forecasting in type 1 diabetes," *J. Healthcare Informat. Res.*, vol. 4, no. 3, pp. 308–324, Sep. 2020.

[26] M. Balbi, A. Caroli, A. Corsi, G. Milanese, A. Surace, F. Di Marco, L. Novelli, M. Silva, F. L. Lorini, A. Duca, R. Cosentini, N. Sverzellati, P. A. Bonaffini, and S. Sironi, "Chest X-ray for predicting mortality and the need for ventilatory support in COVID-19 patients presenting to the emergency department," *Eur. Radiol.*, vol. 31, no. 4, pp. 1999–2012, Apr. 2021.

[27] W. Rahane, H. Dalvi, Y. Magar, A. Kalane, and S. Jondhale, "Lung cancer detection using image processing and machine learning HealthCare," in *Proc. Int. Conf. Current Trends Towards Converging Technol. (ICCTCT)*, Mar. 2018, pp. 1–5.
[28] G. Jacenków, A. Q. O'Neil, and S. A. Tsaftaris, "Indication as prior knowledge for multimodal disease classification in chest radiographs with transformers," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Kolkata, India, Mar. 2022, pp. 1–5.
[29] G. Bezirganyan, S. Sellami, L. Berti-ÉQuille, and S. Fournier, "M2-Mixer: A multimodal mixer with multi-head loss for classification from multimodal data," in *Proc. IEEE Int. Conf. Big Data (BigData)*, Sorrento, Italy, J. He, T. Palpanas, X. Hu, A. Cuzzocrea, D. Dou, D. Slezak, W. Wang, A. Gruca, J. C.-W. Lin, and R. Agrawal, Eds., Dec. 2023, pp. 1052–1058.
[30] H. Türkmen, O. Dikenelli, C. Eraslan, M. C. Çallı, and S. S. Özbek, "BioBERTurk: Exploring Turkish biomedical language model development strategies in low-resource setting," *J. Healthcare Informat. Res.*, vol. 7, no. 4, pp. 433–446, Dec. 2023.
[31] G. Yadav, S. Maheshwari, and A. Agarwal, "Contrast limited adaptive histogram equalization based enhancement for real time video system," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Delhi, India, Sep. 2014, pp. 2392–2397.
[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., Dec. 2017, pp. 5998–6008.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Y. Bengio and Y. LeCun, Eds., May 2015.
[36] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7482–7491.
[37] R. Beran, "An introduction to the bootstrap," in *The Science of Bradley Efron*. New York, NY, USA: Springer, 2008, pp. 288–294.
[38] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, vol. 119, pp. 1597–1607.
[39] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6991–7000.
[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, Austria, May 2021.
[41] X. Han, H. Zheng, and M. Zhou, "CARD: Classification and regression diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, New Orleans, LA, USA, Dec. 2022.

ALI RASEKH received the Bachelor of Science degree in computer engineering and the master's degree in artificial intelligence from the Sharif University of Technology, Tehran, Iran, in 2019 and 2021, respectively. He is currently pursuing the Ph.D. degree in artificial intelligence with the L3S Research Center, Hannover, Germany. Prior to the Ph.D. studies, he had research collaborations with the National University of Singapore and École Polytechnique Fédérale de Lausanne (ÉPFL). His research interests include multimodal learning, deep learning, and machine learning.

REZA HEIDARI is currently pursuing the bachelor's degree in computer engineering from the Sharif University of Technology. His research interests include multimodal vision-language models, visual question answering, and 3-D computer vision.

AMIR HOSEIN HAJI MOHAMMAD REZAIE was born in Tehran, Iran, in 2002. He is currently pursuing the bachelor's degree in computer engineering with the Sharif University of Technology, Tehran. His research interests include machine learning and deep learning. His awards and honors include the Gold Medal in Iranian National Astronomy and Astrophysics Olympiad, in 2019, and the Bronze Medal in the International Olympiad on Astronomy and Astrophysics, in 2020.

PARSA SHARIFI SEDEH is currently pursuing the bachelor's degree in computer engineering with the Sharif University of Technology, Tehran, Iran. His research interests include machine learning and deep learning, particularly their applications in medical imaging and vision-language models (VLMs) and the challenges of domain shifts in visual data.

ZAHRA AHMADI is a junior group leader at Peter L. Reichertz Institute for Medical Informatics of Hannover Medical School. She received her B.Sc. and M.Sc. in Computer Engineering from Sharif University of Technology, Iran, in 2007 and 2009, respectively. Later, she obtained her Ph.D. in Computer Science from Johannes Gutenberg University Mainz, Germany, in 2019. Her research interests are Machine learning and data mining, especially data stream and time series analysis, graph data, multi-label classification, deep neural networks, and explainable AI with applications in multimodal data (sensor, image, and text).

PRASENJIT MITRA received the B.Tech. degree (Hons.) in computer science and engineering from Indian Institute of Technology Kharagpur, Kharagpur, in 1993, the M.S. degree in computer science from The University of Texas at Austin, in 1994, and the Ph.D. degree in electrical engineering from Stanford University, in 2004. He is currently a Professor of Data Sciences and Artificial Intelligence in the College of Information sciences and Technology with The Pennsylvania State University. His research interests include social media analytics, artificial intelligence, big data analytics, information extraction and integration, applied machine learning, natural language processing, and visual analytics. His research has been funded by the NSF CAREER Award and by several grants from DoD, DoE, DHS, NGA, DTRA, Microsoft Corporation, Dow, Lockheed Martin, and Raytheon.

WOLFGANG NEJDL was born in 1960. He received the Dipl.Ing. and Dr.Techn. degrees in computer science from Vienna University of Technology, in 1984 and 1988, respectively. He has been a Full Professor of computer science with Leibniz University Hannover, since 1995. He was an Assistant Professor in Vienna, from 1988 to 1992, and an Associate Professor at RWTH Aachen University, from 1992 to 1995. He was a Visiting Researcher/a Professor with Xerox PARC, Stanford University, the University of Illinois at Urbana-Champaign, EPFL Lausanne, PUC Rio, Trento, and Politecnico di Milano. He heads the L3S Research Center, www.L3S.de, and the Data Science Institute/Knowledge-Based Systems at Leibniz University. From 2014 to 2019, he was a Principal Investigator of the ERC Advanced Grant ALEXANDRIA, where he worked on foundations for temporal search, exploration, and analytics in web archives. He has published more than 440 scientific articles listed at DBLP, with an H-index (based on Google Scholar) of 78. His research interests include information retrieval/search, web science, artificial intelligence, social and semantic web, digital libraries, and technology enhanced learning.