Multimodal Learning for Predicting Mortality in Patients with Pulmonary Arterial Hypertension

Mohammod N. I. Suvon[1], Prasun C. Tripathi[1], Samer Alabed[2,3,4], Andrew J. Swift[2,3,4], Haiping Lu[1,3,*]
[1]*Department of Computer Science,* [2]*Department of Infection, Immunity and Cardiovascular Disease,*
[3]*INSIGNEO, Institute for in silico medicine, The University of Sheffield, Sheffield, United Kingdom,*
[4]*Department of Clinical Radiology, Sheffield Teaching Hospitals, Sheffield, United Kingdom*
{mnisuvon1, p.c.tripathi, s.alabed, a.j.swift, *h.lu}@sheffield.ac.uk
*Corresponding author.

***Abstract*—Pulmonary Arterial Hypertension (PAH) is a life-threatening disorder. The prediction of mortality in PAH patients can play a crucial role in the clinical management of this disease. The prediction of mortality from one modality is a difficult task that may only provide limited performance. Therefore, we propose a multimodal learning approach in this work to predict one-year mortality in PAH patients. We have utilised three modalities, which include extracted numerical imaging features, echo report categorical features, and echo report text features from Electronic Health Records (EHRs) of patients. We have proposed a feature integration module to combine features from multiple modalities. The text features have been extracted from the echo reports using the Bidirectional Encoder Representations from Transformers (BERT). An attention mechanism and a weighted summation method are also adopted during the process of feature integration. We have performed different experiments to evaluate the performance of the proposed framework for mortality prediction. The experimental results indicate that we can achieve the best AUC score of 0.89 for predicting one-year mortality by combining all three modalities. The source code of this paper is available at https://github.com/Mdnaimulislam/MultimodalTab.**

***Index Terms*—Data integration, Mortality prediction, Multimodal learning, Pulmonary Arterial Hypertension**

I. Introduction

Pulmonary Arterial Hypertension (PAH) is a disease that shortens life and eventually results in right heart failure and death if left untreated [1]. PAH is frequently diagnosed at the advanced stage because it does not show early symptoms. Technological advancements have improved healthcare over the years. However, PAH is still considered as one of the deadliest diseases [2]–[4]. Therefore, the effective clinical management of PAH patients is crucial during the treatment. Mortality prediction can help physicians to find out high-risk patients in a large cohort. Prognostic risk variables are often analysed to determine the likelihood of complications in a large population of patients.

In the literature, several studies [5]–[16] have been reported for diagnosis and prognosis of PAH patients. Alonzo et al. [5] have developed a method for survival prediction from hemodynamic data. Benza et al. [6], [7] have developed the Registry to Evaluate Early and Long-Term PAH Disease Management (REVEAL) risk calculator. This risk calculator utilizes 12 clinically relevant features to predict one-year mortality for PAH patients. Furthermore, the REVEAL score has been enhanced and named REVEAL 2.0 [17] by adding one more variable and tweaking another to improve risk prediction.

The morphological features of Cardiac Magnetic Resonance Imaging (CMRI) have been utilized to predict mortality in [10]. In this work, the authors monitored pulmonary artery stiffness to find out the changes in the area and shape for estimating mortality. Sachdev et al. [11] utilized various clinical features to identify the risk of heart failure in PAH patients. Furthermore, electrocardiogram-based features have also been used for the diagnosis and mortality prediction in [12]. A machine learning-based pipeline has been developed in [13] to diagnose PAH patients using tensor-based features learned from CMRI scans. Uthoff et al. [14] predicted the mortality of PAH patients based on geodesically smoothed tensor features. Recently, Alabed et al. [15] utilized tensor-based features to perform one-year mortality prediction for PAH patients. A deep learning-based framework has been developed in [16] for the prognosis of PAH patients using electrocardiogram features. These existing works for mortality prediction utilize a specific modality of data such as electrocardiograph features, clinical features, imaging features, etc. This limits the performance of mortality prediction.

In the past few years, multimodal learning has been applied in different application domains to enhance prediction performance. These studies [18]–[20] combine different types of modalities, such as electrocardiography, text, and image to improve the performance. Motivated by the success of multimodal learning, we propose a multimodal learning-based method for mortality prediction in PAH patients, as illustrated in Fig. 1. The three main contributions of this work are summarised as follows:

1) Firstly, we propose a novel method for predicting mortality in PAH patients utilising multimodal learning to achieve improved performance. Specifically, we combine the features from three modalities using different feature integration mechanisms and compare them. Bidirectional Encoder Representations from Transformers (BERT) are also used for the feature extraction from the echo report text data.
2) Secondly, we perform various experiments to assess the performance of the proposed method on real data from

978-1-6654-6819-0/22/$31.00 ©2022 IEEE

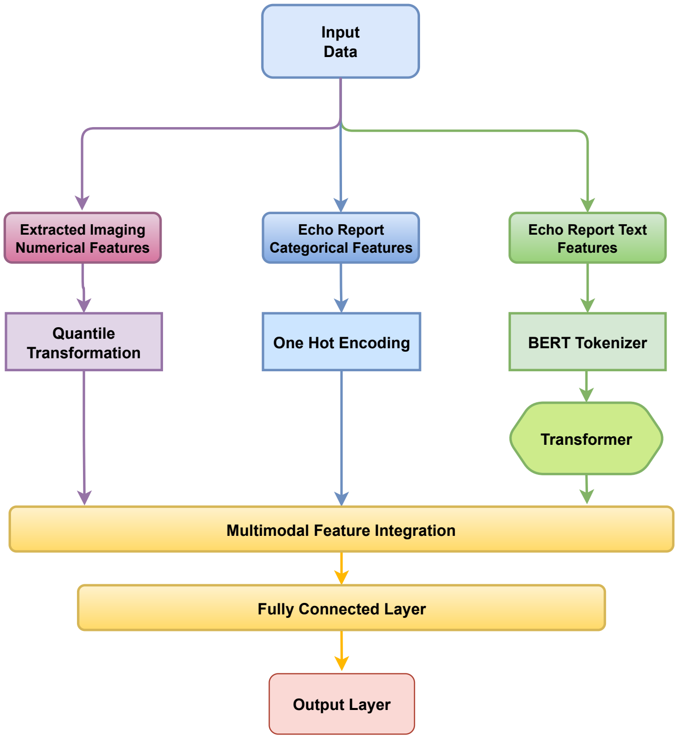
Input
Data
Extracted Imaging Numerical Features
Echo Report Categorical Features
Echo Report Text Features
Quantile Transformation
One Hot Encoding
BERT Tokenizer
Transformer
Multimodal Feature Integration
Fully Connected Layer
Output Layer

Fig. 1: The proposed multimodal learning framework for the prediction of mortality in PAH patients. Three modalities are integrated in multimodal feature integration.

2, 563 PAH patients (which includes 233 cases in the positive class and 2, 330 cases in the negative class).

3) Lastly, we also compare the results with respect to the REVEAL score to determine the clinical effectiveness of the proposed work.

The rest of the paper is structured as follows. Section II explains the materials and the proposed method of our study. In Section III, the performance of our proposed method has been demonstrated. Finally, concluding observations are described in Section IV.

II. Materials and Methods

In this section, we firstly describe the data acquisition process and then provide the detailed description of the proposed method.

A. Data Acquisition

a) Study Population: The ASPIRE registry [21] was used to locate all consecutive PAH patients who had not received any treatment and had been referred for a baseline CMRI between 2008 and 2019. Eligibility requirements include: (i) a baseline CMRI test completed within 14 days of the RHC-confirmed diagnosis of PAH and before the start of PAH medication. (ii) Death within 12 months of the CMRI test or a minimum 12-month follow-up. A sum of 2, 567 successive incident patients with PAH was recognised, and 233 patients died. The number of the negative class samples (patients who did not die) was 2, 330, and the number of the positive class samples (patients who died) was 233.

For this retrospective research, written consent was waived, and ethical approval was received from the local ethics council (ref c06/Q2308/8).

b) CMR Imaging Protocol: A 1.5 Tesla GE HDx (GE Healthcare, Milwaukee, USA) system consisting of an eight-channel cardiac coil was used to carry out cardiac magnetic resonance. A cardiac-gated multislice balanced steady-state free precession sequence was used to obtain short-axis (SA) and four-chamber (4Ch) cine images. With both ventricles completely covered from base to apex, a stack of radiographs in the SA region was taken. The cavity zone at the end-systole was believed to be the smallest. End-diastole was recognized as the biggest volume, the first cine phase of the R-wave triggered capture. The patients were lying on their backs with an ECG gated retrospectively and a surface coil. At end-diastole and end-systole on the SA images, volumetric and ventricular function analysis was executed by contouring the ventricular endocardial borders using MASS software. Papillary muscles and trabecula were incorporated into the blood volume. Later on, different imaging feature measurements were extracted from this.

c) Clinical and Mortality Data: Before starting treatment, clinical information comprising the lung function test, intermittent shuttle walking test, serum NT-proBNP level, etc. were gathered. The electronic medical system was used to gather demographic information, PAH subgroup diagnosis, WHO functional status, and prognosis information. Data on mortality was gathered from the National Health Service (NHS) and Private Demographics Service's electronic files. Once a death was reported in the UK, the NHS updated the mortality records automatically. All patients were observed as part of the national service definition for patients with PAH for at least 12 months, and no patients have been lost to follow-up.

d) Multimodal Feature Preprocessing and Selection: The initial dataset contains 299 features which have been divided into 2 categories: 1) extracted numerical imaging feature, and 2) patient echo report (text and categorical). In this dataset, we had many missing values for different features. The features that contain more than 40% missing data have been discarded from this study, and the rest of the features with missing values have been imputed using the modified Nonlinear Iterative Partial Least Squares (NIPALS) method [22].

After eliminating these features, we got **111 features**. These features have been divided into numerical imaging features, echo report categorical features, and echo report text features. The different combinations of the data-set and its modalities for this study have been summarised in Table I. In Table I all the unimodals have one modality, such as extracted numerical imaging features. An echo report normally consists of 2 modalities, which are categorical (e.g., left ventricle dilated?, right ventricle dilated?, mitral regurgitation, etc.) and textual (e.g., machine report-1, doctor report-1, machine report-2, doctor report-2, summary, etc.). We have separated the categorical and textual features from the echo report and

TABLE I: Combinations of different features categorized into different modalities.

Modalities	Number of categorical features	Number of numerical features	Number of text features	Total number of features
Numerical imaging features (unimodal)	0	93	0	93
Textual echo report (unimodal)	9	0	0	9
Categorical echo report (unimodal)	0	0	9	9
Textual and categorical echo report (bi-modal)	9	0	9	18
Numerical imaging features + textual echo report (bimodal)	0	93	9	102
Numerical imaging features + categorical echo report (bimodal)	9	93	0	102
Numerical imaging features + textual and categorical echo report (trimodal)	9	93	9	111

made two separate unimodal. Finally, all the unimodal features have been merged in various combinations to make several multimodal feature datasets.

B. An End-to-End Multimodal Framework

The architecture of the proposed multimodal learning framework is depicted in Fig. 1. We discuss the components of the proposed framework in the following.

1) Multimodal Input and Preprocessing: The input to the proposed method contains multimodal features. This feature consists of three modality which include imaging numerical, echo report categorical, and echo report text features. In the preprocessing stage, each type of data is handled using a specific method.

First, the imaging numerical features have been preprocessed using the **quantile transform** method. By using this method, the features are transformed to have a uniform or normal distribution. As a result, this transformation generally spreads around the most frequent values for a given feature. This makes it a robust preprocessing approach because it also minimizes the impact of outliers. Each feature has been transformed individually. The original values are converted to a uniform distribution using an estimate of a feature's cumulative distribution function. Using the associated quantile function, the values are then transferred to the appropriate output distribution. Feature values from unseen or new data that lie below or beyond the fitted range are transferred to the output distribution boundaries.

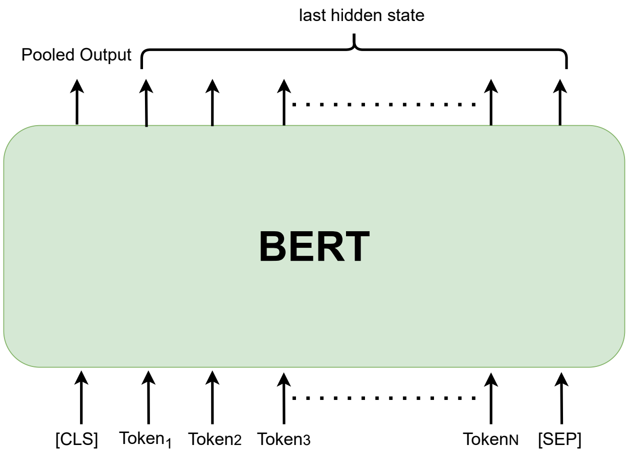

Fig. 2: The extraction of text features as pooled outputs using BERT architecture.

Second, while processing, echo report categorical data must be transformed into numerical form. We perform a **one-hot encoding** to convert categorical features as one-hot numeric arrays. The length of the numeric array is equal to the number of categories in the categorical features. During the processing, it takes a column with categorical data that has been label-encoded and then separates the subsequent column into numerous columns. Based on the column values, the numbers are replaced by either 0 or 1.

Third, we perform tokenization on the echo report text features. We have used the **BERT Tokenizer [23]** to determine tokens from the text data of Echo records. This tokenizer uses a word-piece tokenizer to split the text into words and convert words into either their full forms (e.g., one word converted into one token) or into word pieces, where one word can be split into multiple tokens (e.g. character N-gram).

2) Transformer: We have utilised the popular BERT transformer [23] to process text data. The pre-trained transformer has been incorporated into our method. This transformer has been trained with a huge corpus of text data. The transformer exploits positional encoding and attention mechanisms in its architecture. The BERT-base has 12 BERT layers, and each BERT layer produces token embeddings. We obtain a total of 13 layers since the model adds one extra embedding layer at the beginning. Generally, the BERT model returns 2 outputs that are **pooled output** and **sequence to sequence output**. The first output from the BERT model is sequence to sequence, and it is the output from the final layer. The output size of this layer is ([number of batches, number of tokens in each batch, and the size of the hidden layer]). **In this study**, we have extracted the pooled output from the last layer hidden state of the first token of the sequence (classification token [CLS]) of the pre-trained BERT architecture as shown in Fig. 2. The pooled output that corresponds to the last hidden state of a [CLS] token has no interpretability, but it is the best option as an input for a classifier that can be fine-tuned on a separate dataset. The main reason for this is the manner in which it is pre-trained.

It is important to remember that we are utilising the BERT model as transfer learning, which has already been pre-trained on massive quantities of data. The [CLS] token is always used as a starting token when performing prediction, which is why a pooled output has no interpretability but captures all the information for a specific input. The extracted embedded output has been used as the text features for the Multimodal feature combiner.

3) Multimodal Feature Combiner: In this module, we integrate different types of feature representations. We denote numerical features as **n**, categorical features as **c**, and the output of the transformer for text features as **t**. These features are combined and represented as **x**. Despite the fact that cross-modal attention is already incorporated into the middle layers of multimodal transformers, we opted for a design where the transformer comes earlier than the modality integration because this module may simply be expanded to accommodate more transformers in the future.

We have implemented four techniques for merging the multiple representations into a unique feature space for classification. These techniques are concatenation, MLP+concatenation, attention, and weighted summation. In the concatenation, we combine the three modalities of features as given in Eq. (1):

$$\mathbf{x} = \mathbf{t} \oplus \mathbf{c} \oplus \mathbf{n}, \tag{1}$$

where $\oplus$ denotes a simple concatenation operator that connects multiple tensor inputs.

In the second type of integration, we use Multi-layer Perceptron (MLP) to combine the categorical echo report and numerical imaging features. In the MLP+contenation scheme, categorical and numerical imaging features are passed through MLP. Eq. (2) represents the process of feature integration:

$$\mathbf{x} = \mathbf{t} \oplus \mathrm{MLP}(\mathbf{c}) \oplus \mathrm{MLP}(\mathbf{n}), \tag{2}$$

In the third scheme, we have utilized an attention mechanism [24] to integrate numerical imaging, textual, and categorical echo report features.

The fourth method utilises weighted summation method [25] to perform feature fusion on numerical imaging, textual, and categorical echo report features. The feature integration techniques used in this work are based on related research in multimodal transformers [26]–[29] and other baselines like concatenation and MLP.

4) The Output Layer: In the output layer, we calculate the final prediction score. The formulation of this score is given using Eq. (3):

$$p = \sigma_s(\boldsymbol{\omega}^T \mathbf{x} + \mathbf{b}), \tag{3}$$

where $\boldsymbol{\omega}$, $\mathbf{b}$, and p stand for the model weight, bias vector, and prediction score, respectively. Additionally, the probabilities are produced using a non-linear activation function called $\sigma_s(\cdot)$, which can be altered for different tasks. For instance, in this case, the classification task is performed using a sigmoid activation function. The cross-entropy of prediction over the labels, which is expressed in Eq. (4), is the loss function given the label $y \in \{0, 1\}$:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log p_i + (1 - y_i) \log(1 - p_i), \tag{4}$$

where N represents the total number of training samples.

III. Experimental results and analysis

We have performed different experiments to study the performance of the proposed methods. As our dataset is highly imbalanced (233 positive and $2,330$ negative class samples), we have utilized a custom cross-validation scheme to test the performance. In this scheme, we made 10 folds where for every fold, the positive class has the same 233 samples, and the negative class has 233 unique negative samples from the $2,330$ negative samples. So every fold contains total 466 samples, which have been split into training and validation with an 80:20 ratio. The model's performance is decided based on the average testing performance of all folds. To perform the training, we optimise the loss function in Eq. (4) by using the stochastic gradient descent approach. The learning rate has been set as 0.0001.

Table II reports the results for different types of integration schemes. In this table, we have compared the results for four types of integration schemes which include concatenation, MLP+concatenation, attention, and weighted summation. The performance of these schemes is compared based on mean Area Under Curve (AUC) scores obtained. It is evident from this table that weighted summation produces superior results for the combined modalities model rather than unimodal models.

The proposed method produces the best result when we use the weighted summation method and the tri-modal model with all three modalities. In this case, the proposed method achieves a mean AUC of 0.89. The bi-modal model with numerical imaging and textual echo report features also provides the same mean AUC score for the weighted summation method. However, we select the tri-modal model as the best because it contains smaller variations in results as compared to the bi-modal model (see Table II). The REVEAL model allows the evaluation of one-year mortality using some clinical features. These features are taken from a single modality. The REVEAL model provides an AUC score of 0.70 for the mortality prediction. This score is considered as a gold standard for mortality prediction in PAH patients. A score above 0.70 indicates the clinical applicability of the proposed method. This shows that utilizing multimodal data helps to enhance the performance of one-year mortality prediction. The features of a single modality are insufficient for mortality prediction (all the unimodal models have lower performance results than bi-modal models and tri-modal model). Multimodal learning helps enhance the performance of a model. The integration methods, such as concatenation and weighted summation, can be trained with a single modality. However, the other two

TABLE II: The performance on different modalities in terms of mean AUC for different types of integration methods. The integration methods, such as concatenation and weighted summation, can be trained with a single modality. However, other two integration methods require both imaging numerical and categorical modalities for the training. Due to this, results for some experiments are not available in the table. (**Best**, Second best).

Modalities	**Concatenation**	**MLP+Concatenation**	**Attention**	**Weighted Summation**
Numerical imaging features (unimodal)	0.69±0.03	–	–	0.83±0.02
Textual echo report (unimodal)	0.63±0.05	–	–	0.72±0.04
Categorical echo report (unimodal)	0.58±0.04	–	–	0.67±0.01
Textual and categorical echo report (bi-modal)	0.64±0.02	–	–	0.73±0.03
Numerical imaging features + textual echo report (bi-modal)	0.70±0.03	–	–	0.89±0.02
Numerical imaging features + categorical echo report (bi-modal)	0.69±0.02	0.72±0.03	0.78±0.02	0.86±0.01
Numerical imaging features + textual and categorical echo report (tri-modal)	**0.71±0.03**	**0.74±0.02**	**0.81±0.01**	**0.89±0.01**

TABLE III: The performance of the best integration method weighted summation in other metrics for different modalities (**Best**, Second best).

Modalities	**Sensitivity**	**Specificity**	**Positive Predictive Value (PPV)**	**Negative Predictive Value (NPV)**	**Accuracy**
REVEAL score (unimodal)	0.66±0.3	0.70±0.02	0.67±0.02	0.69±0.02	0.68±0.02
Numerical imaging features (unimodal)	0.69±0.06	0.86±0.06	0.87±0.05	0.69±0.04	0.76±0.03
Textual echo report (unimodal)	0.59±0.12	0.71±0.05	0.72±0.02	0.59±0.07	0.64±0.05
Categorical echo report (unimodal)	0.54±0.05	0.64±0.05	0.68±0.02	0.53±0.01	0.59±0.01
Textual and categorical echo report (bi-modal)	0.52±0.04	0.82±0.02	0.78±0.02	0.57±0.02	0.65±0.02
Numerical imaging features + textual echo report (bi-modal)	0.69±0.03	**0.90±0.03**	0.90±0.03	0.70±0.01	0.78±0.01
Numerical imaging features + categorical echo report (bi-modal)	67±0.03	0.85±0.02	0.85±0.02	0.67±0.02	0.75±0.02
Numerical imaging features + textual and categorical echo report (tri-modal)	**0.73±0.05**	0.89±0.03	**0.90±0.02**	**0.72±0.03**	**0.80±0.01**

integration methods require numerical imaging and categorical modalities for training.

We have also analyzed the performance of the best model with the weighted summation integration method in Table III. In this table, the results are compared based on five performance metrics. These performance metrics include sensitivity, specificity, PPV (Positive Predictive Value), NPV (Negative Predictive Value), and accuracy. It can be observed from this table that the proposed method produces promising results for different metrics. The proposed method shows the best performance for each metric (except specificity) when we utilize all modalities. Finally, we have reported Receiver Operating Characteristics (ROC) curve in Fig. 3 for the best models. This curve depicts the results for the best integration method, which is the weighted summation method. It can be noticed from this figure that the mean AUC score of the proposed methods is the highest for the tri-modal model with all three modalities. The bi-modal model with numerical imaging and categorical echo report features produces lower performance than the tri-modal model and bi-modal model with numerical imaging and

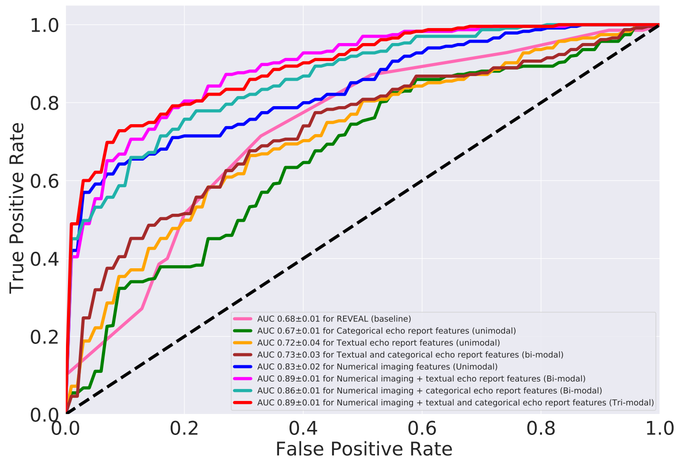

Fig. 3: Mean receiver operation characteristic curve analysis showing predictive accuracy of the best integration method (weighted summation) with different modalities. This figure is best viewed in color print or on screen.

textual echo report features. Therefore, the utilisation of the textual echo report feature helps to enhance the performance of the model.

IV. Conclusion and future work

Automatic diagnosis and prognosis of PAH patients play a crucial role in clinical practice. It is usually required to identify high-risk patients from a large set in order to provide targeted course of the treatment. In this work, we have developed a method to predict one-year mortality in PAH patients. The proposed method exploits multimodal data to enhance the performance. We have utilised numerical imaging data extracted from CMRI and echo report categorical and text data extracted from electronic health records. We have incorporated a transformer in the proposed method to process text features. Several combinations of modalities and integration methods have been tested to observe the performance. We have achieved the best performance when we combined all modalities of data using the weighted summation method. The proposed method has achieved an AUC score of 0.89, which is better than the REVEAL score (AUC = 0.70) used in the clinical practice. Some of the features used in the dataset contains a huge number of missing values. These features have been excluded in the current study. In the future, we will utilise a better missing data handling scheme and use all features present in the dataset to enhance the performance.

Acknowledgment

The study was supported by the Wellcome Trust grants 215799/Z/19/Z and 205188/Z/16/Z.

References

[1] V. V. McLaughlin, S. L. Archer, D. B. Badesch, R. J. Barst, H. W. Farber, J. R. Lindner, M. A. Mathier, M. D. McGoon, M. H. Park, R. S. Rosenson *et al.*, "Accf/aha 2009 expert consensus document on pulmonary hypertension: a report of the american college of cardiology foundation task force on expert consensus documents and the american heart association developed in collaboration with the american college of chest physicians; american thoracic society, inc.; and the pulmonary hypertension association," *Journal of the American college of cardiology*, vol. 53, no. 17, pp. 1573–1619, 2009.

[2] R. L. Benza, D. P. Miller, R. J. Barst, D. B. Badesch, A. E. Frost, and M. D. McGoon, "An evaluation of long-term survival from time of diagnosis in pulmonary arterial hypertension from the reveal registry," *Chest*, vol. 142, no. 2, pp. 448–456, 2012.

[3] H. W. Farber, D. P. Miller, A. D. Poms, D. B. Badesch, A. E. Frost, E. Muros-Le Rouzic, A. J. Romero, W. W. Benton, C. G. Elliott, M. D. McGoon *et al.*, "Five-year outcomes of patients enrolled in the reveal registry," *Chest*, vol. 148, no. 4, pp. 1043–1054, 2015.

[4] E. M. Lau, E. Giannoulatou, D. S. Celermajer, and M. Humbert, "Epidemiology and treatment of pulmonary arterial hypertension," *Nature Reviews Cardiology*, vol. 14, no. 10, pp. 603–614, 2017.

[5] G. E. D'Alonzo, R. J. Barst, S. M. Ayres, E. H. Bergofsky, B. H. Brundage, K. M. Detre, A. P. Fishman, R. M. Goldring, B. M. Groves, J. T. Kernis *et al.*, "Survival in patients with primary pulmonary hypertension: results from a national prospective registry," *Annals of internal medicine*, vol. 115, no. 5, pp. 343–349, 1991.

[6] R. L. Benza, D. P. Miller, M. Gomberg-Maitland, R. P. Frantz, A. J. Foreman, C. S. Coffey, A. Frost, R. J. Barst, D. B. Badesch, C. G. Elliott *et al.*, "Predicting survival in pulmonary arterial hypertension: insights from the registry to evaluate early and long-term pulmonary arterial hypertension disease management (reveal)," *Circulation*, vol. 122, no. 2, pp. 164–172, 2010.

[7] R. L. Benza, M. Gomberg-Maitland, D. P. Miller, A. Frost, R. P. Frantz, A. J. Foreman, D. B. Badesch, and M. D. McGoon, "The reveal registry risk score calculator in patients newly diagnosed with pulmonary arterial hypertension," *Chest*, vol. 141, no. 2, pp. 354–362, 2012.

[8] W.-T. N. Lee, Y. Ling, K. K. Sheares, J. Pepke-Zaba, A. J. Peacock, and M. K. Johnson, "Predicting survival in pulmonary arterial hypertension in the uk," *European Respiratory Journal*, vol. 40, no. 3, pp. 604–611, 2012.

[9] T. Thenappan, C. Glassner, and M. Gomberg-Maitland, "Validation of the pulmonary hypertension connection equation for survival prediction in pulmonary arterial hypertension," *Chest*, vol. 141, no. 3, pp. 642–650, 2012.

[10] C. T.-J. Gan, J.-W. Lankhaar, N. Westerhof, J. T. Marcus, A. Becker, J. W. Twisk, A. Boonstra, P. E. Postmus, and A. Vonk-Noordegraaf, "Noninvasively assessed pulmonary artery stiffness predicts mortality in pulmonary arterial hypertension," *Chest*, vol. 132, no. 6, pp. 1906–1912, 2007.

[11] A. Sachdev, H. R. Villarraga, R. P. Frantz, M. D. McGoon, J.-F. Hsiao, J. F. Maalouf, N. M. Ammash, R. B. McCully, F. A. Miller, P. A. Pellikka *et al.*, "Right ventricular strain for prediction of survival in patients with pulmonary arterial hypertension," *Chest*, vol. 139, no. 6, pp. 1299–1309, 2011.

[12] R. W. Scherptong, I. R. Henkens, G. F. Kapel, C. A. Swenne, K. W. van Kralingen, M. V. Huisman, A. J. Schuerwegh, J. J. Bax, E. E. van der Wall, M. J. Schalij *et al.*, "Diagnosis and mortality prediction in pulmonary hypertension: the value of the electrocardiogram-derived ventricular gradient," *Journal of Electrocardiology*, vol. 45, no. 3, pp. 312–318, 2012.

[13] A. J. Swift, H. Lu, J. Uthoff, P. Garg, M. Cogliano, J. Taylor, P. Metherall, S. Zhou, C. S. Johns, S. Alabed *et al.*, "A machine learning cardiac magnetic resonance approach to extract disease features and automate pulmonary arterial hypertension diagnosis," *European Heart Journal-Cardiovascular Imaging*, vol. 22, no. 2, pp. 236–245, 2021.

[14] J. Uthoff, S. Alabed, A. J. Swift, and H. Lu, "Geodesically smoothed tensor features for pulmonary hypertension prognosis using the heart and surrounding tissues," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 253–262.

[15] S. Alabed, J. Uthoff, S. Zhou, P. Garg, K. Dwivedi, F. Alandejani, R. Gosling, L. Schobs, M. Brook, Y. Shahin *et al.*, "Machine learning cardiac-mri features predict mortality in newly diagnosed pulmonary arterial hypertension," *European Heart Journal-Digital Health*, vol. 3, no. 2, pp. 265–275, 2022.

[16] G. P. Diller, M. L. Benesch Vidal, A. Kempny, K. Kubota, W. Li, K. Dimopoulos, A. Arvanitaki, A. E. Lammers, S. J. Wort, H. Baumgartner *et al.*, "A framework of deep learning networks provides expert-level

accuracy for the detection and prognostication of pulmonary arterial hypertension," *European Heart Journal-Cardiovascular Imaging*, 2022.

[17] R. L. Benza, M. Gomberg-Maitland, C. G. Elliott, H. W. Farber, A. J. Foreman, A. E. Frost, M. D. McGoon, D. J. Pasta, M. Selej, C. D. Burger *et al.*, "Predicting survival in patients with pulmonary arterial hypertension: the reveal risk score calculator 2.0 and comparison with esc/ers-based risk assessment strategies," *Chest*, vol. 156, no. 2, pp. 323–337, 2019.

[18] V. Radu, C. Tong, S. Bhattacharya, N. D. Lane, C. Mascolo, M. K. Marina, and F. Kawsar, "Multimodal deep learning for activity and context recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–27, 2018.

[19] C. Marechal, D. Mikolajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, and K. Wegrzyn-Wolska, "Survey on AI-based multimodal methods for emotion detection." *High-performance modelling and simulation for big data applications*, vol. 11400, pp. 307–324, 2019.

[20] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu *et al.*, "MultiBench: Multiscale benchmarks for multimodal representation learning," *arXiv preprint arXiv:2107.07502*, 2021.

[21] J. Hurdman, R. Condliffe, C. Elliot, C. Davies, C. Hill, J. Wild, D. Capener, P. Sephton, N. Hamilton, I. Armstrong *et al.*, "Aspire registry: assessing the spectrum of pulmonary hypertension identified at a referral centre," *European Respiratory Journal*, vol. 39, no. 4, pp. 945–955, 2012.

[22] C. Preda, G. Saporta, and M. H. Mbarek, "The nipals algorithm for missing functional data," *Revue roumaine de mathématiques pures et appliquées*, vol. 55, no. 4, pp. 315–326, 2010.

[23] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, pp. 4171–4186.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[25] P. C. Fishburn, "Additive utilities with incomplete product sets: Application to priorities and assignments," *Operations Research*, vol. 15, no. 3, pp. 537–542, 1967.

[26] M. Ostendorff, P. Bourgonje, M. Berger, J. Moreno-Schneider, G. Rehm, and B. Gipp, "Enriching bert with knowledge graph embeddings for document classification," *arXiv preprint arXiv:1909.08402*, 2019.

[27] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.

[28] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[29] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.