Attention-based Multimodal Bilinear Feature Fusion for Lung Cancer Survival Analysis

Hongbin Na[1*†], Lilin Wang[2*], Xinyao Zhuang[2], Jianfei He[2], Zhenyu Liu[2], Zimu Wang[3], Hong-Seng Gan[2]
[1]*School of Computer Science and Engineering, University of New South Wales, Sydney, Australia*
[2]*School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, China*
[3]*School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China*
h.na@student.unsw.edu.au, Lilin.Wang20@student.xjtlu.edu.cn

2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE) | 979-8-3503-9311-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBE60311.2023.00042

***Abstract*—Survival analysis (SA) is an essential task that aims to predict survival status and duration, determine individual and precise treatment strategies, and assess disease intensity and direction. However, the current research on multimodal SA has identified three unique challenges: inefficient cross-modal information integration, insufficient inter-modal key features, and noisy data. In this paper, we propose a novel SA framework, named Attention-based Multimodal Bilinear Feature Fusion (AMBF)-SA, to address the aforementioned challenges. Specifically, AMBF-SA first performs feature extraction with the off-the-shelf models on each modality separately, then fuses the features between multiple sources and modalities using our proposed AMBF method, and finally outputs the survival prediction by a multi-layer perception (MLP). Experimental results on the Non-small Cell Lung Cancer (NSCLC) Radiogenomics dataset demonstrate remark performance of AMBF-SA compared with the rest of the experimented models, including the models trained with single and combined modalities under the Mean Absolute Error (MAE) and the Concordance Index (C-index) evaluation metrics, indicating the usefulness of our proposed framework.**

***Keywords*—Attention mechanism, feature fusion, lung cancer, multimodal machine learning, survival analysis.**

I. Introduction

Survival analysis (SA) is an essential task that aims to predict survival status and duration, determine individual and precise treatment strategies, and assess disease intensity and direction, allowing physicians to select the most opportune moment for therapeutic intervention, thus avoiding over-treatment and optimizing the optimal use of medical resources [1], [2]. Simultaneously, it can also be utilized to assist patients in planning the remainder of their lives and achieving a more comprehensive life [3]. Such studies focus on predicting survival rates for complex ailments by incorporating multi-source and diverse features extracted from both clinical and medical image data to enhance the accuracy and personalization of the predictions. Combining multi-source information for SA not only assists doctors in better assessing the condition and treatment plan [2] but also enhances patients' understanding of their condition and improves the quality of survival.

Original research on SA mainly focused on the predictions over a single modality of data. As shown in Fig. 1, the modalities leveraged for previous SA research include clinical information [4], gene expressions [5], and medical images [6], such as computed tomography (CT) and positron emission tomography (PET) scans, and failed to fully exploit the heterogeneous information represented by different modalities of data. Additionally, some work also combined the features extracted from the combination of aforementioned sources to obtain better performances [7], [8]. Recently, some studies attempted to employ straightforward feature splicing techniques [9]–[11]; however, such approaches involved linear and uncomplicated combinations, failed to fully acknowledge the deep integration of semantic information from multi-source data. Considering that directly splicing heterogeneous features from diverse distributions may also encounter the issue of feature mismatch; therefore, constructing SA models capable of integrating heterogeneous features from multiple sources is a crucial scientific issue. Further experiments and research are required to enhance the expressiveness and robustness of SA models through cross-modal feature transformation and fusion.

Fig. 1. Modalities and sources commonly focused on in current SA research, including clinical information, gene expressions, and medical images, such as CT and PET scans.

In terms of multimodal SA, the current research in this field has identified the following unique challenges: (1) **Inefficient Cross-modal Information Integration.** In current research on multimodal SA, different modalities and sources, such as CT and PET scans, provide valuable information from multiple aspects. For instance, CT scans are able to scan detailed anatomical structures, while PET scans can showcase lesion activities. Another layer of complexity emerges with datasets providing clinical details, such as patient histories. Therefore, the integration of these diverse sources to maximize predictive value is critical [12]. (2) **Insufficient Intra-modal**

* Equal contribution.
† Partly done while H.Na was at Xi'an Jiaotong-Liverpool University.

2471-7819/23/$31.00 ©2023 IEEE
DOI 10.1109/BIBE60311.2023.00042

Key Features. Within modalities such as CT or PET, it's crucial to extract key features. Beyond addressing contrast and resolution differences, capturing core attributes within each modality is essential. Neglecting these can yield incomplete insights [10]. (3) **Noisy Data.** Noisy data inevitable exists in current datasets, which may hinder the models' training performances. With growing data volumes, it is paramount to maintain computational efficiencies and model scalabilities without succumbing to the noisy samples [13].

In this paper, we propose a novel SA framework, named **A**ttention-based **M**ultimodal **B**ilinear Feature **F**usion (AMBF)-SA, to address the aforementioned challenges. Specifically, as shown in Fig. 2, AMBF-SA first performs feature extraction with the off-the-shelf models on each modality separately, including the medical images, clinical information, and tumor segmentations. Consequently, feature fusion between multiple sources and modalities is performed by utilizing our proposed AMBF method, which enables handling heterogeneous multimodal data to enhance the SA process. Finally, the survival prediction is output by a multi-layer perception (MLP). Notably, AMBF not only considers the uniqueness of each pattern but also fully exploits the potential correlations across them, leading to more accurate survival predictions. Experimental results on the Non-small Cell Lung Cancer (NSCLC) Radiogenomics dataset [14] demonstrate remark performance of AMBF-SA compared with the rest of the experimented models, including the models trained with single and combined modalities under the Mean Absolute Error (MAE) and the Concordance Index (C-index) evaluation metrics, indicating the usefulness of our proposed framework.

In summary, our main contributions are as follows:

- We introduce a novel SA framework, named AMBF-SA, to adeptly manage heterogeneous features present in multimodal medical data, subsequently improving the performance of SA.
- We design a two-stage feature fusion strategy, named AMBF, that fully considers the uniqueness of each modality and the associations between them. It first processes the features of each modality with multiple heads of attention and then cross-fertilizes them to generate the final output.
- Experimental results demonstrate that AMBF-SA performs optimally among the experimented models, better than those trained with single and subset modalities.

II. Related Work

In the field of medical image analysis, it is common to perform feature extractions and survival analysis (SA) over single modalities, such as clinical information [4], gene expressions [5], and medical images [6], such as computed tomography (CT) and positron emission tomography (PET) scans. Some work also combined the features extracted from the combination of aforementioned sources to obtain better performances [7], [8]. While the integration of machine learning and deep learning models with medical images offers promising potential [15], [16], they have achieved performance that competes with and even exceeds doctors in some cases [17]; however, it is rarely employed for mortality prediction. As a result, this represents a unique and challenging domain, and the overall performance of available survival analysis techniques is generally inadequate. Wu *et al.* [11] proposed the first multimodal deep learning method for Non-small Cell Lung Cancer (NSCLC) survival analysis, known as DeepMMSA, to address the aforementioned problem. Unlike conventional methods that relied on clinical data for lung cancer survival analysis and provided statistical probabilities, DeepMMSA extracted features from multiple modalities, including CT images and clinical data, and fused them for survival prediction. Extensive experimental results demonstrated the underlying relationship between prognostic information and radiomic images, together with the superiority of DeepMMSA over traditional unimodal approaches, leading to increased accuracy for survival prediction. To improve the extraction of latent features from medical images, Wang *et al.* [18] retrieved radiomic features from regions of interest (ROI) and combined these features for survival prediction outcomes. Additionally, they employed multidimensional intra- and peritumoral features for patients with clinical stage and pathologic stage IA pure-solid NSCLC so as to provide personalized survival risk stratification. This method demonstrated the efficacy of stratifying survival risks for patients with clinical and pathologic stage IA pure-solid NSCLC with the utilization of multiregional radiomics signature, improving the discriminative ability beyond conventional clinical predictors.

Previous research has demonstrated the effectiveness of machine learning and radiomics analysis methods in the overall survival (OS) of NSCLC predictions. Sun *et al.* [19] extracted tumor features from pre-processed CT images, quantifying tumor phenotypic characteristics based on shape, size, intensity statistics, and texture. With the utilization of 5 feature selection methods and 8 machine learning models, they concluded that the gradient boosting linear models based on Cox's partial likelihood (GB-Cox) with the concordance index (CI) feature selection method achieved the overall optimal performance. For enhanced application of radiomic features in survival analysis, Müller-Franzes *et al.* [20] conducted reliability analysis on CT and MRI radiomic features, improving the predictive capability of the underlying model for clinical imaging modalities and tumor entity patient survival prior to reliability analysis and selecting the most reliable radiomic features. Blanc-Durand *et al.* [21] adopted LASSO Cox regression to obtain progression-free survival (PFS) and OS-pPET-RadScores when predicting survival of hepatocellular carcinoma (HCC). Kaplan-Meier method was used to estimate the survival curve to explore the potential association of the PET radiomics signature with the PFS and overall survival OS.

Under the multimodal setting, Chen *et al.* [22] assessed the association of radiological imaging and gene expression with patient outcomes in NSCLC and constructed a nomogram by combining selected radiomic, genomic, and clinical risk factors with the extraction of handcrafted radiomic features and deep learning genomic features. To address the costly and

time-consuming shortcomings in tumor localization, Dao *et al.* [23] proposed a tumor segmentation model, named Multi-scale Aggregation-based Parallel Transformer Network (MAP-TransNet), by employing parallel transformer mechanism to capture the global context of multi-scale encoder feature maps and then concatenated them to obtain global context at multi-scale maps. Furthermore, they concluded the utilization of multimodal features offered abundant information pertaining to survival analysis task in NSCLC.

III. Methodology

A. Problem Definition

Given a multimodal medical dataset with CT images, PET images, clinical data, and a region of interest (ROI), our objective is to predict the corresponding survival time. For a CT image I_{CT}, lung features are extracted using the combination of `lungmask` and ResNet-18, represented by $F_{\text{CT}} = E_{\text{lung}}(I_{\text{CT}})$. Simultaneously, for a PET image I_{PET}, features are delineated using ResNet-18, denoted as $F_{\text{PET}} = E_{\text{res}}(I_{\text{PET}})$. Clinical data, referred to as D_{clinic}, undergoes pre-processing techniques, such as one-hot encoding, and scaling to produce the feature F_{clinic}. Radiomic features from the ROI, R, in the CT image I_{CT}, are extracted using PyRadiomics, resulting in $F_{\text{rad}} = E_{\text{rad}}(R, I_{\text{CT}})$. In the feature fusion phase, we utilize our proposed AMBF method, formulated as $F = \text{AMBF}(F_{\text{clinic}} + F_{\text{rad}}, \text{AMBF}(F_{\text{CT}}, F_{\text{PET}}))$. This fused feature F is then input into a designed multi-layer perceptron (MLP) for predicting the survival time, expressed as $y = \text{MLP}(F)$. The overarching aim is to leverage features from various medical modalities through an efficient fusion strategy to predict survival times accurately.

B. Feature Extraction

1) CT/PET Image Feature Extraction: We employ a U-Net model [24] pre-trained on a on a subset of the LTRC dataset to extract features from CT images, with the utilization of a renowned `lungmask` toolkit [25], providing a high capability to efficiently reduce false positives, enhancing prediction accuracy, and also facilitates direct execution of the R231 and LTRCLobes models via fusion results. Having resized the CT images to a uniform 224×224 pixels, they are fed into the U-Net model. The model subsequently outputs a lung region mask with a resolution of 512×512 pixels. To maintain data integrity, we institute a filtering criterion: images are excluded if the non-zero pixel count in the mask falls below 5% of the total possible pixel count for a 512×512 dimension.

Following the lung field pre-processing, CT and PET images, though processed separately, are both channeled into ResNet-18 [26] for individual feature extraction. ResNet-18, a sophisticated structure of convolutional layers, pooling segments, and a diverse range of residual units, operates based on the foundational equation, as shown in Eq. (1) as follows:

$$y = F(x, W) + x, \tag{1}$$

where x and y represent the input and output, F is the neural operations within the residual block, and W is their associated weights. At the end of the model, we add a linear layer to adjust the feature dimensions for each patient, and both PET and CT features are saved with a feature dimension of $1,500$ to facilitate the use of image features in the next step. With ResNet-18, we achieve a comprehensive feature representation for both CT and PET images, each set capturing distinct features inherent to their respective modalities.

2) Radiomic Feature Extraction: In this research, we perform an in-depth quantitative analysis of medical images so as to utilize radiomics features. Using the PyRadiomics software package [27], 18 categories of radiomics features, amounting to a total of $1,682$ individual features, are extracted from CT images and their corresponding ROI segmentation. These features encompass the original images, as well as their various transformations, including exponential, gradient, square, and square root images. Additionally, features derived from images transformed by specific filters such as wavelets, Gaussian Laplace, and various local binary patterns are also considered. Such transformations provides a multi-dimensional viewpoint for the extraction of features. For instance, edges and structures are accentuated in gradient images [28], while specific intensity ranges are highlighted in exponential and square function images, facilitating enhanced visualization and feature extraction from diverse image regions [29].

Subtle textures and patterns, which are often challenging to discern in the original images, are effectively captured through specific filter transformations, such as the Local Binary Pattern (LBP) method [30]. Wavelet filters, on the other hand, allowed for the decomposition of images into distinct frequency components, emphasizing finer details and rougher structures within the image [31]. Specifically, first-order statistics are employed to describe the voxel intensity distribution within regions of the image defined by masks, thereby furnishing fundamental image information. Shape-based descriptors provided insights into the geometric shapes and sizes of structures within 2D and 3D images. Furthermore, texture matrix features, encompassing the Gray-Level Co-occurrence Matrix (GLCM), Gray-Level Dependence Matrix (GLDM), Gray-Level Run Length Matrix (GLRLM), Gray-Level Size Zone Matrix (GLSZM), and Neighboring Gray Tone Difference Matrix (NGTDM), present comprehensive quantitative depictions of intricate patterns or textures, granting deeper understandings of the inherent image characteristics.

C. Feature Fusion

We introduce a novel feature fusion methodology, named Attention-based Multimodal Bilinear Feature Fusion (AMBF), inspired by both the Transformer architecture [32] and multimodal compact bilinear pooling [33], that leverages the multi-head attention mechanism, designed to address the intricate challenge of modeling sophisticated cross-modal interactions, aiming to capture nuanced relationships within and across modalities. As shown in Fig. 3, AMBF consists of two primary stages: intra-modal feature attention, which focuses on indi-

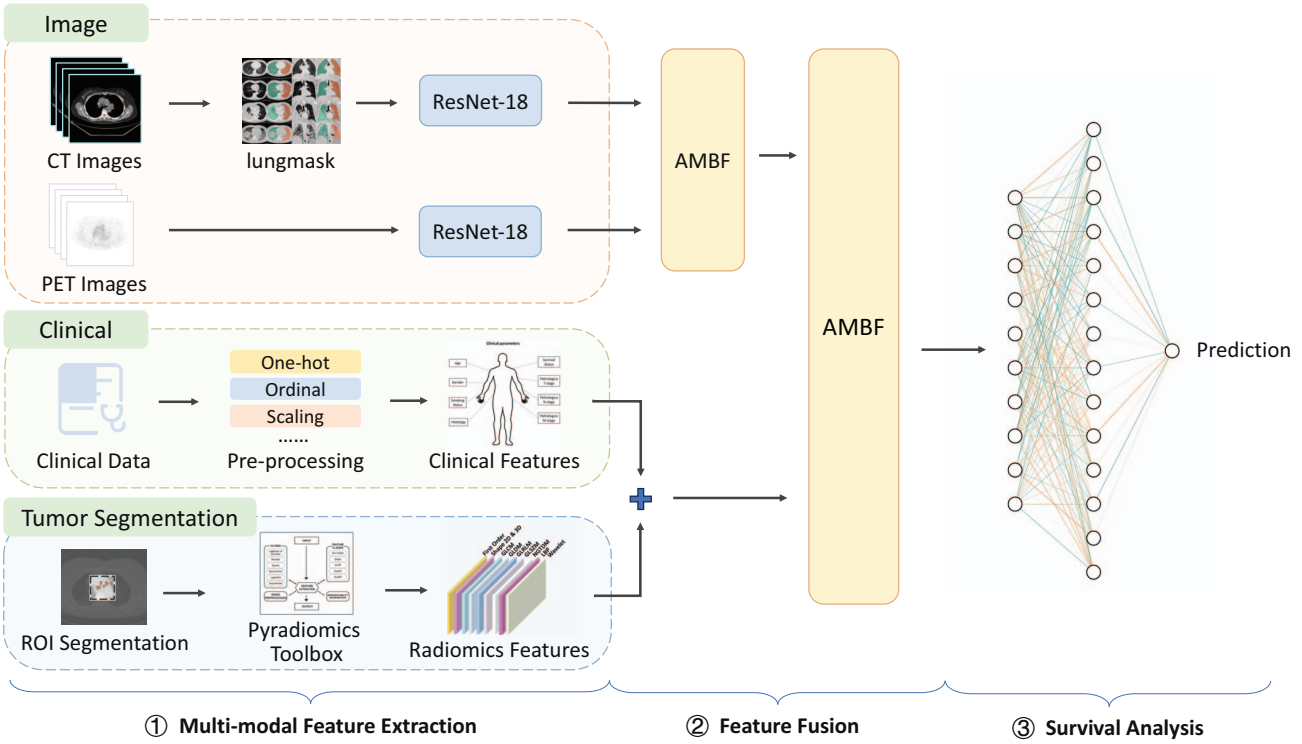

Fig. 2. The overall framework of our proposed AMBF-SA framework, in which the features are firstly extracted with the off-the-shelf models on each modality separately, and then the feature fusion between multiple sources and modalities is performed by utilizing our proposed AMBF method. Finally, the survival prediction is output by a MLP model.

vidual modality characteristics, and inter-modal feature fusion, where cross-modal interactions are harmoniously combined.

1) Intra-modal Feature Attention: Given an input feature set X from a specific modality, we deploy a multi-head attention mechanism. Formally, for each input feature $x_i \in X$, the attention mechanism computes a weighted sum of all features in X, weighted by the attention scores between x_i and every other feature in X. The attention scores are computed as Eq. (2) as follows:

$$A(x_i, x_j) = \frac{\exp(x_i \cdot x_j)}{\sum_{k \in X} \exp(x_i \cdot x_k)}, \tag{2}$$

where $x_i \cdot x_j$ is the dot product between the query representation of x_i and the key representation of x_j. The output of this stage is a set of attention-enhanced features, which capture the most crucial features within each modality.

2) Inter-modal Feature Fusion: After capturing the essential features within each modality, we focus on fusing features across modalities. We use a compact bilinear pooling method known as the count sketch (CS) transformation. For two modalities A and B with feature vector v_A and v_B respectively, the sketch count transformation is given by Eq. (3) as follows:

$$\mathrm{CS}(v) = \sum_{i=1}^{d} s_i \cdot v_i \cdot \delta(h_i), \tag{3}$$

where h_i and s_i are randomly generated parameters, δ is the Kronecker delta function, and d is the feature dimension. Subsequently, an element-wise multiplication is performed between the transformed feature vectors of the two modalities to yield the final fused feature representation, thereby enhancing the cross-modal interaction representation, as shown in Eq. (4) as follows:

$$v_{\text{fused}} = \mathrm{CS}(v_A) \odot \mathrm{CS}(v_B), \tag{4}$$

in which the symbol $\odot$ represents element-wise multiplication between two vectors. The combined effect of the attention mechanism and the fusion technique allows our approach to selectively focus on essential features within each modality while also effectively capturing interactions between them.

D. Survival Prediction

Our approach to survival prediction leans on a neural architecture, marking a departure from classic Kaplan-Meier or Cox regression frameworks. We employ a feedforward neural network tailored to predict the survival times. The network consists of a single hidden layer of 64 neurons, integrated with a Sigmoid activation function for non-linearity. To enhance generalization and prevent overfitting, a dropout mechanism with a rate of 0.3 is integrated. Central to our training process is the Mean Absolute Error (MAE) loss, which quantifies the discrepancy between the predicted and actual survival times.

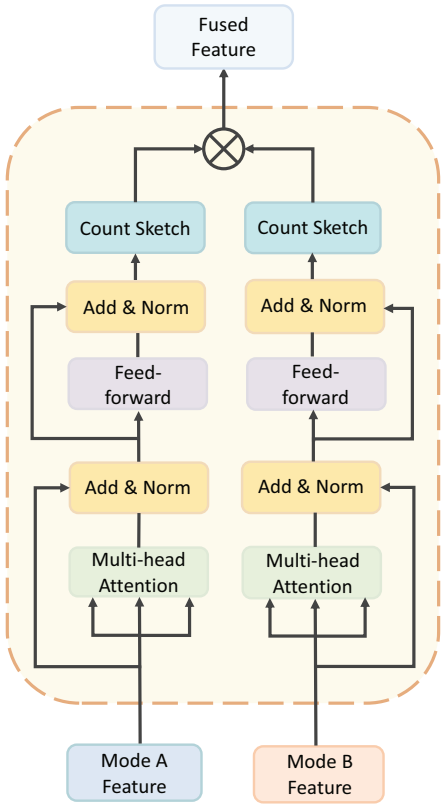

Fig. 3. Procedures of our proposed AMBF multimodal feature fusion method.

Formally, for N samples, the MAE loss is given by Eq. (5) as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} \left| y_i - \hat{y}_i \right|, \tag{5}$$

where y_i represents the ground truth and $\hat{y}_i$ is the predicted survival time. Through this loss function, the model is refined to offer nuanced predictions, bridging the gap between traditional and contemporary predictive paradigms.

IV. Experiments

A. Dataset

The Non-small Cell Lung Cancer (NSCLC) Radiogenomics dataset [14] from the Cancer Imaging Archive (TCIA)[1] is leveraged as the benchmark dataset in our research, which is a dataset established to elucidate potential correlations between the molecular attributes of tumors and the features of medical imaging, and subsequently to facilitate the development and evaluation of prognostic medical imaging biomarkers. For each subject, the available data encompasses CT images, ROI segmentation of tumors evident in the CT scans, and pertinent clinical information. The clinical parameters encompass age, gender, smoking status, TNM staging, overall staging (derived from TNM), and survival rates.

[1]https://www.cancerimagingarchive.net/

TABLE I
Demographic Characteristics of our Experimental Data

Feature	Categories	Description
Age	< 69 years	64 (45.39%)
	≥ 69 years	77 (54.61%)
Sex	Male	105 (74.47%)
	Female	36 (25.53%)
Histology	Adenocarcinoma	111 (78.72%)
	Squamous	27 (19.15%)
	Other	3 (2.13%)
Pathologica T Stage	T1 & Tis	72 (51.06%)
	T2	48 (34.04%)
	T3	16 (11.35%)
	T4	5 (3.55%)
Pathologica N Stage	N0	114 (80.85%)
	N1	10 (7.09%)
	N2	17 (12.06%)
Pathologica M Stage	M0	137 (97.16%)
	M1a	1 (0.71%)
	M1b	3 (2.13%)
Smoking Status	Non-smoking	22 (15.60%)
	Smoking	28 (19.86%)
	Former Smoking	91 (64.54%)
Survival Status	Alive	91 (64.54%)
	Dead	50 (35.46%)

Of the initial 211 subjects, data from 141 patients, encompassing clinical information, CT and PET images, and associated tumor segmentation labels, are incorporated into our study. The remaining 70 subjects were precluded due to either a lack of segmentation labels or failure to satisfy modality data prerequisites. For the selected data, the population of each clinical attribute are detailed in Table I.

B. Data Pre-processing

A comprehensive pre-processing strategy is implemented in this research in response to the phenomenon observed in the data, such as missing values and nominal attributes, including data selection, missing values fulling, and one-hot encoding. Additionally, the recurrence dates are simplified by categorizing them based on their respective years, a measure taken to bolster the reliability of our predictions given our dataset's size. To refine our imputation approach, the interval between `CT Date` and `Date of Last Known Alive` is calculated, which then informs the imputation of `Time to Death (days)`. This feature underwent min-max normalization to align its scale with other continuous variables, constraining its values between 0 and 1, as shown in Eq. (6) as follows:

$$X_{\mathrm{scaled}} = \frac{X - X_{\mathrm{min}}}{X_{\mathrm{max}} - X_{\mathrm{min}}}. \tag{6}$$

From an imaging perspective, the raw DICOM image data undergoes conversion to the NIfTI format, a preparatory step for downstream radiomics analysis and to ensure versatility in tool compatibility. Image standardization is achieved by

mapping pixel values to a mean of 0 and a standard deviation of 1, with set parameters of 0.5 for both mean and standard deviation.

C. Experiment Setup

To achieve a fair evaluation, we partition the NSCLC Radiogenomics dataset into training and test sets at a ratio of $8:2$. Additionaly, to ensure efficient and stable training, we utilize an NVIDIA GeForce RTX 3090 24GB graphics card for all computational tasks. Regarding our training strategy for the model, we opted for SGD as the optimizer, and we set an initial learning rate of $1e-3$ and employ the cosine annealing approach for dynamic learning rate adjustment. Furthermore, the number of epochs for training is set as 400.

D. Evaluation Metrics

We adopt the Mean Absolute Error (MAE) and the Concordance Index (C-index) to evaluate the survival prediction performance our experimented models. The MAE quantifies the average absolute disparity between the predicted and actual survival times, acting as an intuitive measure of the model's predictive accuracy, where a reduced MAE indicates superior prediction accuracy [1]. Simultaneously, the C-index gauges the alignment between the model's predicted survival time orderings and the actual observed order for patients categorized as high or low risk, which can be calculated as Eq. (7) as follows:

$$C-index=\frac{concordant\ pairs}{comparable\ pairs}. \tag{7}$$

In calculating the C-index, we rigorously adhered to the guidelines delineated by Harrell *et al.* [34], especially designed for datasets encompassing right-censored data. Within this framework, only pairs of samples with discernible event times, inclusive of right-censored data, are taken into account, and a pair is considered comparable if the observed event is not earlier than the event time of the associated sample.

E. Experimental Results

Experimental results of the models' performances over the test set of the NSCLC Radiogenomics dataset are presented in Table II, in which a reduced MAE and an elevated C-index indicate a model's superior performance. From the experimental results, we can make the following observations:

1) It's salient to observe that radiomic features attain a C-index of 0.6497, which notably exceeds the 0.5975 from clinical data. This observation accentuates the more comprehensive insights provided by radiomics. Combining both clinical and radiomics features escalates the C-index to 0.7488, distinctly outpacing individual modalities. This aligns with current literature emphasizing the superior efficacy of multimodal approaches over their unimodal counterparts.
2) In our image data fusion experiments, relying exclusively on CT+PET results in a C-index of 0.5432. Yet, integrating the proposed AMBF fusion strategy propels this value to 0.6768. This significant increment not only corroborates the robustness of our fusion strategy but also underlines its adeptness at mining more profound insights from image datasets.
3) A cornerstone of our contribution is the novel AMBF-SA algorithm. Seamlessly integrating data from Clinical + Radiomics and CT+PET+AMBF, it registers exceptional performance metrics. With an enviable MAE of 0.1341 and a stellar C-index of 0.8325, the AMBF-SA methodology eclipses the other techniques assessed. Such results underline the paramount efficacy of our proposed technique, hinting at its prospective utility in the broader research landscape.

TABLE II
PERFORMANCES OF THE MODELS ON THE NSCLC RADIOGENOMICS DATASET

Method	MAE↓	C-index↑
Clinical	0.2534	0.5975
Radiomics	0.2372	0.6497
Clinical+Radiomics	0.1728	0.7488
CT+PET	0.2773	0.5432
CT+PET+AMBF	0.2017	0.6768
AMBF-SA (Ours)	**0.1341**	**0.8325**

V. CONCLUSION AND FUTURE WORK

In this paper, we design a novel SA framework, named AMBF-SA, to address the three unique challenges identified in current SA research. We investigate the prognostic value of various data modalities, including clinical, radiomic, and fused image data, in predicting survival outcomes for patients within the NSCLC Radiogenomics dataset. Experimental results under the MAE and C-index metrics underscore the intrinsic merits of incorporating both clinical and radiomic features, in which the combined features' approach notably surpassed the individual performances of either modality, attesting to its efficacy. Furthermore, our proposed AMBF-SA methodology, which amalgamates the rest of the models, exhibited unprecedented accuracy, as evidenced by its superior MAE and C-index metrics, which not only solidifies the potential of multimodal strategies over their unimodal counterparts but also emphasizes the versatility and capability of our novel fusion mechanism.

For future endeavors, we aim to expand our dataset to enhance the generalizability of our model. Integrating more diverse imaging modalities may also provide richer information for predictions. We also foresee leveraging advanced deep learning techniques to further refine our fusion strategies, potentially driving even more robust and accurate prognostic models for NSCLC patients.

REFERENCES

[1] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Computing Surveys*, vol. 51, no. 6, feb 2019.

[2] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[3] C. K. Reddy, Y. Li, and C. Aggarwal, "A review of clinical prediction models," *Healthcare Data Analytics*, vol. 36, pp. 343–378, 2015.

[4] J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," *Journal of the American Medical Informatics Association*, vol. 20, no. 4, pp. 613–618, 03 2013.

[5] J. Hou, J. Aerts, B. den Hamer, W. van IJcken, M. den Bakker, P. Riegman, C. van der Leest, P. van der Spek, J. A. Foekens, H. C. Hoogsteden, F. Grosveld, and S. Philipsen, "Gene expression-based classification of non-small cell lung carcinomas and survival prediction," *PLOS ONE*, vol. 5, no. 4, pp. 1–12, 04 2010.

[6] J. E. van Timmeren, R. T. Leijenaar, W. van Elmpt, B. Reymen, C. Oberije, R. Monshouwer, J. Bussink, C. Brink, O. Hansen, and P. Lambin, "Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam ct images," *Radiotherapy and Oncology*, vol. 123, no. 3, pp. 363–369, 2017.

[7] N. G. Mikhaeel, D. Smith, J. T. Dunn, M. Phillips, H. Møller, P. A. Fields, D. Wrench, and S. F. Barrington, "Combination of baseline metabolic tumour volume and early response on pet/ct improves progression-free survival prediction in dlbcl," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 43, no. 7, p. 1209–1219, Feb 2016.

[8] L. Evangelista, L. Urso, M. Caracciolo, F. Stracuzzi, S. Panareo, A. Cistaro, and O. Catalano, "Fdg pet/ct volume-based quantitative data and survival analysis in breast cancer patients: A systematic review of the literature," *Current Medical Imaging*, vol. 19, no. 8, pp. 807–816, 2023.

[9] L. Sun, S. Zhang, H. Chen, and L. Luo, "Brain tumor segmentation and survival prediction using multimodal mri scans with deep learning," *Frontiers in Neuroscience*, vol. 13, 2019.

[10] Y. Li and L. Shen, "Deep learning based multimodal brain tumor diagnosis," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes, Eds. Cham: Springer International Publishing, 2018, pp. 149–158.

[11] Y. Wu, J. Ma, X. Huang, S. H. Ling, and S. Weidong Su, "Deepmmsa: A novel multimodal deep learning method for non-small cell lung cancer survival analysis," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2021, pp. 1468–1472.

[12] M. A. Saleh, A. A. Ali, K. Ahmed, and A. M. Sarhan, "A brief analysis of multimodal medical image fusion techniques," *Electronics*, vol. 12, no. 1, 2023.

[13] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.

[14] S. Bakr, O. Gevaert, S. Echegaray, K. Ayers, M. Zhou, M. Shafiq, H. Zheng, J. A. Benson, W. Zhang, A. N. Leung, and et al., "A radiogenomic dataset of non-small cell lung cancer," *Scientific Data*, vol. 5, no. 180202, Nov 2018.

[15] W. Zhu, C. Liu, W. Fan, and X. Xie, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 673–681.

[16] D. Riquelme and M. A. Akhloufi, "Deep learning for lung cancer nodules detection and classification in ct scans," *AI*, vol. 1, no. 1, pp. 28–67, 2020.

[17] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[18] T. Wang, Y. She, Y. Yang, X. Liu, S. Chen, Y. Zhong, J. Deng, M. Zhao, X. Sun, D. Xie, and C. Chen, "Radiomics for survival risk stratification of clinical and pathologic stage ia pure-solid non–small cell lung cancer," *Radiology*, vol. 302, no. 2, pp. 425–434, 2022, pMID: 34726531.

[19] W. Sun, M. Jiang, J. Dang, P. Chang, and F.-F. Yin, "Effect of machine learning methods on predicting nsclc overall survival time based on radiomics analysis," *Radiation Oncology*, vol. 13, no. 197, Oct 2018.

[20] G. Müller-Franzes, S. Nebelung, J. Schock, C. Haarburger, F. Khader, F. Pedersoli, M. Schulze-Hagen, C. Kuhl, and D. Truhn, "Reliability as a precondition for trust—-segmentation reliability analysis of radiomic features improves survival prediction," *Diagnostics*, vol. 12, no. 2, 2022.

[21] P. Blanc-Durand, A. Van Der Gucht, M. Jreige, M. Nicod-Lalonde, M. Silva-Monteiro, J. O. Prior, A. Denys, A. Depeursinge, and N. Schaefer, "Signature of survival: a 18f-fdg pet based whole-liver radiomic analysis predicts survival after 90y-tare for hepatocellular carcinoma," *Oncotarget*, vol. 9, no. 4, pp. 4549–4558, 2018.

[22] W. Chen, X. Qiao, S. Yin, X. Zhang, and X. Xu, "Integrating radiomics with genomics for non-small cell lung cancer survival analysis," *Journal of Oncology*, vol. 2022, no. 5131170, p. 1–8, Aug 2022.

[23] D.-P. Dao, H.-J. Yang, N.-H. Ho, S. Pant, S.-H. Kim, G.-S. Lee, I.-J. Oh, and S.-R. Kang, "Survival analysis based on lung tumor segmentation using global context-aware transformer in multimodality," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 5162–5169.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[25] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, vol. 4, no. 50, Aug 2020.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[27] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 10 2017.

[28] O. Vincent and O. Folorunso, "A descriptive algorithm for sobel image edge detection," *InSITE Conference*, 2009.

[29] X. Chen, J. Li, and Z. Hua, "Low-light image enhancement based on exponential retinex variational model," *IET Image Processing*, vol. 15, no. 12, pp. 3003–3019, 2021.

[30] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[31] D.-Z. Tian and M.-H. Ha, "Applications of wavelet transform in medical image processing," in *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)*, vol. 3, 2004, pp. 1816–1821 vol.3.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[33] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 457–468.

[34] J. Harrell, Frank E., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the Yield of Medical Tests," *JAMA*, vol. 247, no. 18, pp. 2543–2546, 05 1982.