2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230356

IMPROVED MULTIMODAL FUSION FOR SMALL DATASETS WITH AUXILIARY SUPERVISION

Gregory Holste[1,2], *Douwe van der Wal*[2], *Hans Pinckaers*[2], *Rikiya Yamashita*[2],
Akinori Mitani[2], *Andre Esteva*[2]

[1] The University of Texas at Austin, Austin, TX USA
[2] Artera, Inc, Mountain View, CA USA

ABSTRACT

Prostate cancer is one of the leading causes of cancer-related death in men worldwide. Like many cancers, diagnosis involves expert integration of heterogeneous patient information such as imaging, clinical risk factors, and more. For this reason, there have been many recent efforts toward deep multimodal fusion of image and non-image data for clinical decision tasks. Many of these studies propose methods to fuse learned features from each patient modality, providing significant downstream improvements with techniques like cross-modal attention gating, Kronecker product fusion, orthogonality regularization, and more. While these enhanced fusion operations can improve upon feature concatenation, they often come with an extremely high learning capacity, meaning they are likely to overfit when applied even to small or low-dimensional datasets. Rather than designing a highly expressive fusion operation, we propose three simple methods for improved multimodal fusion with small datasets that *aid optimization by generating auxiliary sources of supervision during training*: **extra supervision**, **clinical prediction**, and **dense fusion**. We validate the proposed approaches on prostate cancer diagnosis from paired histopathology imaging and tabular clinical features. The proposed methods are straightforward to implement and can be applied to any classification task with paired image and non-image data.

Index Terms— multimodal fusion, histopathology, prostate cancer, automated diagnosis

1. INTRODUCTION

Prostate cancer is the third most frequently diagnosed cancer worldwide and one of the leading causes of cancer-related death in men [1, 2]. Diagnosis and risk stratification is primarily performed by manual analysis of histopathology imaging from tissue biopsy in the context of patient history, and risk factors like age and prostate-specific antigen (PSA) level [3]. Like many cancers, diagnosis relies on the integration of multiple streams of heterogeneous information about the patient. Given this clinical reality and the ongoing success of deep learning techniques for automated diagnosis, there has been recent attention toward multimodal fusion for automated cancer risk stratification; this involves using deep learning techniques to jointly learn from patient imaging (radiological, histopathological, or other), patient risk factors, and other streams of data like lab results and genomic profiles [4, 5, 6]. In this line of research, a central open question remains: what is the best way to fuse information from each modality?

In their review of multimodal fusion approaches for imaging and electronic health record (EHR) data, Huang *et al.* [7] divide approaches into three categories: early fusion (joining inputs from each modality), joint fusion (joining learned intermediate features from each modality), and late fusion (joining predicted probabilities from each modality). Most studies find that some form of joint fusion (often called "**late joint fusion**"), combining learned modality-specific feature vectors to generate a single multimodal representation, to be most effective [7, 8]. Given, for example, an image representation and a clinical data representation, there are many ways to aggregate these two feature vectors – the most common being concatenation and elementwise averaging. Observing room for improvement, researchers have proposed more complex and expressive fusion operations [9, 10] and architectures [11, 12] to optimally join multimodal representations. For example, Chen *et al.* [9, 11] use cross-modal attention gating and Kronecker product fusion to aggregate genomic and histopathology imaging features for cancer outcome prediction. Braman *et al.* [10] build upon this by adding an orthogonality loss that encourages unimodal representations to provide complementary information during fusion.

While these fusion methods have shown significant improvements on diagnosis tasks, the resulting models often have a very high learning capacity (large number of learnable parameters). Thus when applied to small datasets or low-dimensional inputs, these methods are more likely to overfit to spurious patterns in the training data [13]. We hypothesize that another avenue toward improved generalization *without dramatically increasing model capacity* is to ease model optimization with multiple sources of supervision. In this work, we propose three straightforward methods that aid optimization by generating auxiliary sources of supervision: extra

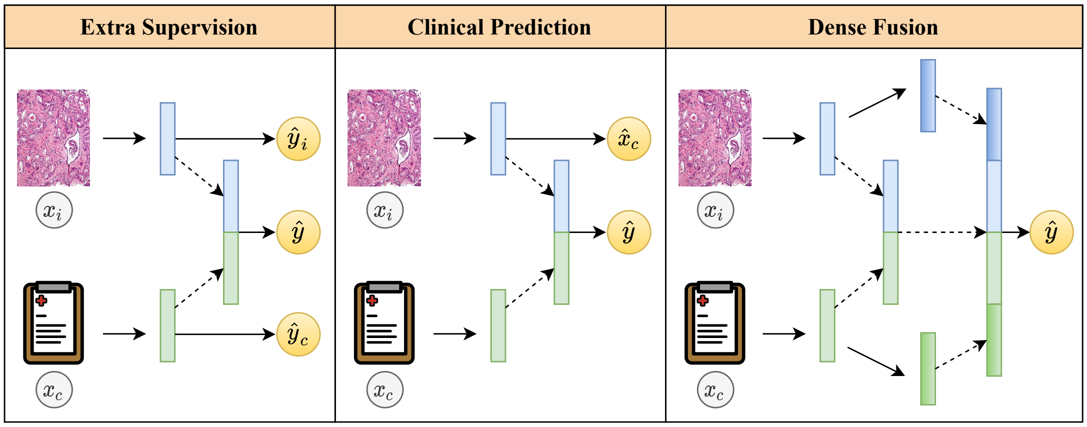

Fig. 1. Diagram of proposed multimodal fusion approaches for small or low-dimensional datasets. **Extra supervision** (left) generates additional modality-specific predictions, **clinical prediction** (middle) uses the learned image representation to directly predict the associated non-image inputs, and **dense fusion** (right) encourages dense interaction of features by joining information from different stages of the network.

supervision, clinical prediction, and dense fusion. We validate these approaches on prostate cancer classification using paired histopathology image features and tabular clinical risk factors. Our experiments show that a combination of these three methods improves upon late joint fusion (with concatenation *or* Kronecker product fusion) and can be readily applied to any task with paired image and non-image data.

2. MATERIALS & METHODS

2.1. Dataset Description

This study uses data collected from five prospective, randomized phase III clinical trials of men with localized prostate cancer conducted by NRG Oncology: NRG/RTOG-9202, 9408, 9413, 9910, and 0126 [14, 15, 16, 17, 18]; a subset of $4,581$ patients was made available for machine learning research in this work. For each patient, we use paired histopathology imaging and clinical risk factor data to predict distant metastasis (DM), the binary event that cancer spreads from the original tumor site. Image features were pre-extracted by a self-supervised learning model [19] so that each slide is represented by a lower-dimensional "bag" of image features $\boldsymbol{x_i} \in \mathbb{R}^{K \times 128}$, where K is the number of fixed-size (256 × 256) patches. Each associated clinical input vector $\boldsymbol{x_c} \in \mathbb{R}^6$ includes six tabular descriptors of the patient: age, PSA, T-stage, and pathologist-determined Gleason scores and patterns [20]. In total, DM occurs in 12.2% of the 4,581 patients. For full details on data acquisition and image feature extraction, please refer to Esteva *et al.* [19].

2.2. Fusion Methods for Small Datasets

The standard "late joint fusion" method of multimodal classification involves learning modality-specific representations, fusing those feature vectors into a single multimodal representation, then using this feature vector to perform classification. Concretely, let $\boldsymbol{x_i}$ represent a patient's histopathology imaging and $\boldsymbol{x_c}$ represent the patient's associated clinical information. Given image encoder $f_i(\cdot)$ and clinical encoder $f_c(\cdot)$, we compute representations $\boldsymbol{h_i} = f_i(\boldsymbol{x_i})$ and $\boldsymbol{h_c} = f_c(\boldsymbol{x_c})$, where i refers to the imaging modality and c refers to the clinical (or non-imaging) modality. Then image and non-image representations are fused via concatenation: $\boldsymbol{h} = \text{concat}([\boldsymbol{h_i}, \boldsymbol{h_c}])$. Finally, a classifier $g(\cdot)$ is used to generate a final prediction of the outcome via $\hat{y} = g(\boldsymbol{h})$. This model is optimized by minimizing the loss $\mathcal{L} = \ell(y, \hat{y})$, where $\ell(\cdot)$ is the binary cross-entropy loss function.

Building on this baseline, we present three methods that aid optimization by adding auxiliary sources of supervision during training: **extra supervision**, **clinical prediction**, and **dense fusion** (Figure 1).

2.2.1. Extra Supervision

While the baseline late joint fusion approach generates a single prediction based on fused image and non-image features, it does not directly encourage the modality-specific representations to be predictive of the outcome. We can remedy this by adding classification heads on top of each modality-specific representation. Given image-only classifier $g_i(\cdot)$ and clinical-only classifier $g_c(\cdot)$, we additionally compute $\hat{y}_i = g_i(\boldsymbol{h_i})$ from image features alone and $\hat{y}_c = g_c(\boldsymbol{h_c})$ from clinical features alone. This approach, **extra supervision**, now generates three predictions of the outcome, which can together be used to optimize the entire network via the loss $\mathcal{L} = \ell(y, \hat{y}) + \ell(y, \hat{y}_i) + \ell(y, \hat{y}_c)$. This way, additional supervisory signal from the modality-specific features can flow back to optimize all parameters in the network. Similar approaches have appeared in Kawahara *et al.* [21] and Holste *et al.* [6].

2.2.2. *Clinical Prediction*

In **clinical prediction**, we build upon the standard late joint-fusion approach by using the learned image representation to directly predict (regress) the associated non-image input features. Specifically, we compute $\hat{\boldsymbol{x}}_{\boldsymbol{c}} = g_{i\to c}(\boldsymbol{h_i})$ with an additional classification head $g_{i\to c}(\cdot)$. In addition to the predicted outcome, this new predicted clinical feature vector $\hat{\boldsymbol{x}}_{\boldsymbol{c}}$ serves as an auxiliary source of supervision during training via the loss $\mathcal{L} = \ell(y, \hat{y}) + \text{sim}(\boldsymbol{x_c}, \hat{\boldsymbol{x}}_{\boldsymbol{c}})$, where $\ell(\cdot)$ is the binary cross-entropy loss and $\text{sim}(\cdot)$ is the mean squared error (MSE) loss for continuous inputs. This approach can be interpreted as an auxiliary prediction task that aligns the learned image representations with the clinical non-image features. Since the image and non-image "views" of the patient are expected to be correlated, this alignment of modalities may enable more optimal fusion of image and non-image information.

2.2.3. *Dense Fusion*

Inspired by Hu *et al.* [22], **dense fusion** improves upon standard late joint fusion by allowing for denser interaction between modalities during training. As in late joint fusion, we learn image-only and non-image-only representations via $\boldsymbol{h_i^{(1)}} = f_i^{(1)}(\boldsymbol{x_i})$ and $\boldsymbol{h_c^{(1)}} = f_c^{(1)}(\boldsymbol{x_c})$, and then those representations are fused to form $\boldsymbol{h^{(1)}} = \text{concat}([\boldsymbol{h_i^{(1)}}, \boldsymbol{h_c^{(1)}}])$; the superscript (1) is denotes that this is the first intermediate layer of the network. However, instead of placing a classifier on top of the fused representation $\boldsymbol{h^{(1)}}$, we instead learn a deeper representation of the image and non-image features with $\boldsymbol{h_i^{(2)}} = f_i^{(2)}(\boldsymbol{x_i})$ and $\boldsymbol{h_c^{(2)}} = f_c^{(2)}(\boldsymbol{x_c})$, where $f_i^{(2)}$ and $f_c^{(2)}$ are fully-connected layers. We then form a final fused representation that not only aggregates the image-only and clinical-only features (as in late joint fusion), but also incorporates the fused representation from the *previous* stage of the network: $\boldsymbol{h^{(2)}} = \text{concat}([\boldsymbol{h_i^{(2)}}, \boldsymbol{h_c^{(2)}}, \boldsymbol{h^{(1)}}])$. Finally, a classifier $g(\cdot)$ is used to generate $\hat{y} = g(\boldsymbol{h^{(2)}})$, and the model is trained by optimizing the same binary cross-entropy loss as in late joint fusion. This allows for dense interaction of features from each modality, aggregating information across different stages of the network.

While dense fusion does not explicitly generate additional sources of supervision like the other two approaches, it can be combined with extra supervision and clinical prediction to obtain even more sources of supervision than either approach alone. Dense fusion provides multiple intermediate feature vectors; when combined with extra supervision, these feature vectors are used to generate *additional* outcome predictions. Further, **all three methods are complementary and can be applied in any combination**. For example, Figure 2 depicts an architecture combining all three proposed approaches; this model minimizes the loss $\mathcal{L} = \ell(y, \hat{y}) + \ell(y, \hat{y}_f) + \ell(y, \hat{y}_i^{(1)}) + \ell(y, \hat{y}_i^{(2)}) + \ell(y, \hat{y}_c^{(1)}) + \ell(y, \hat{y}_c^{(2)}) + \text{sim}(\boldsymbol{x_c}, \hat{\boldsymbol{x}}_{\boldsymbol{c}})$. For all models, the final prediction at inference time is taken to be $\hat{y}$, the output learned from the final fused multimodal representation.

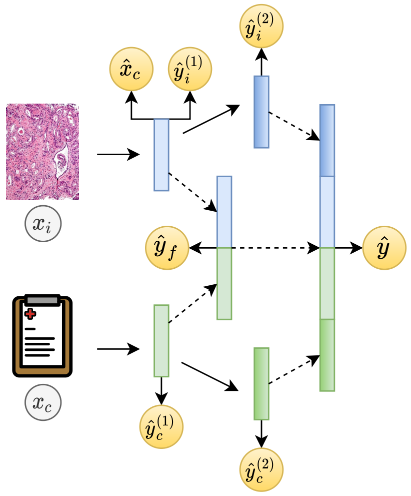

Fig. 2. Multimodal fusion architecture with all three approaches combined: extra supervision, clinical prediction, and dense fusion. The model is trained by optimizing the sum of six cross-entropy losses (one for each predicted $\hat{y}$) and one mean squared error loss (for regressing clinical features $\hat{\boldsymbol{x}}_{\boldsymbol{c}}$).

2.3. Implementation Details

As in Esteva *et al.* [19], all six clinical features were treated as numeric and standardized. To generate a 128-dimensional representation for each bag of imaging features $\boldsymbol{x_i} \in \mathbb{R}^{K\times 128}$, attention weights were learned to modulate certain features via elementwise multiplication, then sum pooling was utilized across the patch dimension. All models were trained with the AdamW optimizer [23] with learning rate 0.0075, batch size 128, and dropout 0.9 on classification heads. Each model was trained for 100 epochs with class-balanced resampling per fold to combat imbalance. For the first 20 epochs, only the clinical encoder $f_c(\cdot)$ and classifier $g_c(\cdot)$ were optimized; then, the parameters of $f_c(\cdot)$ were frozen while all other parameters were optimized for the remaining 80 epochs.

2.4. Experimental Setup

In our first experiment, we train a late joint fusion model (Section 2.2) with (i) concatentation, (ii) Kronecker product fusion, (iii) concatenation + all three proposed auxiliary supervision methods, and (iv) Kronecker fusion + all three auxiliary supervision methods. A comparison of these four models is performed to reveal the tradeoff between model complexity and generalization when applied to our low-dimensional dataset. To understand the contribution of each proposed technique, we also perform an ablation-style exper-

Fusion Operation	Auxiliary Supervision	# Params	AUC
Concatenation		17.1K	0.781 ± 0.024
Kronecker		66.3K	0.770 ± 0.018
Concatenation	✓	43.1K	**0.792 ± 0.014**
Kronecker	✓	207.1K	0.781 ± 0.013

Table 1. Distant metastasis (DM) classification results with different fusion techniques and our proposed auxiliary supervision approaches. "Auxiliary Supervision" = trained with extra supervision, clinical prediction, and dense fusion. The baseline late joint fusion approach is highlighted in gray.

iment that progressively adds each technique to the baseline of late joint fusion with concatenation. With the same hyperparameters and an otherwise identical architecture, we train the baseline model plus all possible combinations of the three proposed approaches: extra supervision (Section 2.2.1), clinical prediction (Section 2.2.2), and dense fusion (Section 2.2.3). We adopt five-fold cross-validation for model training and use area under the receiver operating characteristic curve (AUC) for evaluation. Performance is summarized by the mean and standard deviation of AUC across all five folds.

3. RESULTS

Compared to the standard late joint fusion approach, our proposed auxiliary supervision methods considerably improve DM classification performance (Table 1). Our proposed model (row 3) reaches 0.792 ± 0.014 AUC, while late joint fusion achieves 0.781 ± 0.024 AUC with concatenation and 0.770 ± 0.018 AUC with Kronecker product fusion. We find that Kronecker fusion increases model capacity by 4-5× the number of parameters with a consistent adverse affect on generalization in our setting. While our proposed methods *also* increase model capacity, the benefits of auxiliary supervision are demonstrated by the fact that our proposed model (row 3) outperforms late joint Kronecker fusion (row 2).

In our ablation study, we find that best results are achieved with all three proposed auxiliary supervision approaches (Table 2). While the baseline reaches 0.781 ± 0.024 AUC, a model that additionally uses extra supervision, clinical prediction, and dense fusion reaches 0.792 ± 0.014 AUC. Interestingly, while the addition of *only* extra supervision or *only* dense fusion does not improve performance, any combination of the two approaches improves upon the baseline. We also find that clinical prediction is the single most impactful of the three techniques; the three models trained with clinical prediction achieved the three largest mean AUCs and three smallest standard deviations in AUC across folds.

Extra Supervision	Clinical Prediction	Dense Fusion	AUC
			0.781 ± 0.024
✓			0.778 ± 0.021
	✓		0.789 ± 0.013
		✓	0.776 ± 0.025
✓	✓		0.787 ± 0.015
✓		✓	0.785 ± 0.016
	✓	✓	0.790 ± 0.012
✓	✓	✓	**0.792 ± 0.014**

Table 2. Distant metastasis (DM) classification results with all combinations of the proposed approaches. Best results appear in bold, and second-best results are underlined. The baseline late joint fusion approach is highlighted in gray.

4. DISCUSSION & CONCLUSION

In summary, we proposed three simple approaches for improved deep multimodal fusion on low-dimensional datasets. Rather than designing highly expressive fusion or attention operations, the proposed techniques are less likely to overfit because they aid optimization with auxiliary sources of supervision rather than adding significant extra learning capacity. We validate these three approaches – extra supervision, clinical prediction, and dense fusion – on the task of learning jointly from digital histopathology imaging and tabular clinical data to predict prostate cancer metastasis, observing that the three proposed approaches improve upon the standard approach of late joint fusion. Further, the proposed approaches outperform the recently popular Kronecker product fusion while being more parameter-efficient. Though these methods were validated on one specific application and dataset, they can be readily applied to any multimodal classification task with paired image and non-image data.

This study could be expanded to perform more extensive validation of our fusion approaches on other classification tasks and even in other domains beyond hisopathology. While the combination of all three proposed methods achieved best performance, it remains to be seen whether this is the optimal combination of methods across different tasks and datasets. Lastly, our study could potentially benefit from more thorough evaluation using metrics that are resistant to class imbalance, such as average precision or balanced accuracy. We chose AUC as a metric to match the evaluation of Esteva *et al.* [19] and avoid the need to choose a decision threshold, though AUC can become inflated with class imbalance [24]. It would also be valuable to compare our approaches with other baselines such as human expert performance. Further, a feature importance analysis to uncover which clinical features are most predictive of DM in this multimodal setting could enrich the clinical value of our findings.

5. COMPLIANCE WITH ETHICAL STANDARDS

IRB approval was obtained my NRG Oncology through IRB00000781. Informed consent was waived because all data was anonymized.

6. ACKNOWLEDGMENTS

This project was supported by U10CA180822 (NRG Oncology SDMC), U10CA180868 (NRG Oncology Operations), and U24CA196067 (NRG Specimen Bank) from the National Cancer Institute.

7. REFERENCES

[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, et al., "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018.

[2] Elizabeth M Ward, Recinda L Sherman, S Jane Henley, et al., "Annual report to the nation on the status of cancer, featuring cancer in men and women age 20–49 years," *J. Natl. Cancer Inst.*, vol. 111, no. 12, pp. 1279–1297, 2019.

[3] Edward Schaeffer, Sandy Srinivas, Emmanuel S Antonarakis, et al., "NCCN guidelines insights: Prostate cancer, version 1.2021: Featured updates to the NCCN guidelines," *J. Natl. Compr. Canc. Netw.*, vol. 19, no. 2, pp. 134–143, 2021.

[4] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, et al., "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Natl. Acad. Sci.*, vol. 115, no. 13, pp. E2970–E2979, 2018.

[5] Anika Cheerla and Olivier Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, 2019.

[6] Gregory Holste, Savannah C. Partridge, Habib Rahbar, et al., "End-to-end learning of fused image and non-image features for improved breast cancer classification from MRI," in *Proc. IEEE Int. Conf. Comput. Vis. Works.*, 2021, pp. 3294–3303.

[7] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, et al., "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–9, 2020.

[8] Can Cui, Haichun Yang, Yaohong Wang, et al., "Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: A review," *arXiv preprint 2203.15588*, 2022.

[9] Richard J Chen, Ming Y Lu, Jingwen Wang, et al., "Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Trans. Med. Imaging*, 2020.

[10] Nathaniel Braman, Jacob WH Gordon, Emery T Goossens, et al., "Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data," in *Med. Image Comput. Comput. Assist. Interv.* Springer, 2021, pp. 667–677.

[11] Richard J Chen, Ming Y Lu, Wei-Hung Weng, et al., "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 4015–4025.

[12] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, et al., "GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3942–3951.

[13] Trevor Hastie, Robert Tibshirani, and Jerome H Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2, Springer, 2009.

[14] Christopher U Jones, Daniel Hunt, David G McGowan, et al., "Radiotherapy and short-term androgen deprivation for localized prostate cancer," *N Engl. J. Med.*, vol. 365, no. 2, pp. 107–118, 2011.

[15] Jeff M Michalski, Jennifer Moughan, James Purdy, et al., "Effect of standard vs dose-escalated radiation therapy for patients with intermediate-risk prostate cancer: the NRG oncology RTOG 0126 randomized clinical trial," *JAMA Oncol.*, vol. 4, no. 6, pp. e180039–e180039, 2018.

[16] Thomas M Pisansky, Daniel Hunt, Leonard G Gomella, et al., "Duration of androgen suppression before radiotherapy for localized prostate cancer: radiation therapy oncology group randomized clinical trial 9910," *J. Clin. Oncol.*, vol. 33, no. 4, pp. 332, 2015.

[17] Eric M Horwitz, Kyounghwa Bae, Gerald E Hanks, et al., "Ten-year follow-up of radiation therapy oncology group protocol 92-02: a phase III trial of the duration of elective androgen deprivation in locally advanced prostate cancer," *J. Clin. Oncol.*, vol. 26, no. 15, pp. 2497–2504, 2008.

[18] Colleen A Lawton, Michelle DeSilvio, Mack Roach III, et al., "An update of the phase III trial comparing whole pelvic to prostate only radiotherapy and neoadjuvant to adjuvant total androgen suppression: updated analysis of RTOG 94-13, with emphasis on unexpected hormone/radiation interactions," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 69, no. 3, pp. 646–655, 2007.

[19] Andre Esteva, Jean Feng, Douwe van der Wal, et al., "Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials," *NPJ Digit. Med.*, vol. 5, no. 1, pp. 1–8, 2022.

[20] Donald F Gleason and George T Mellinger, "Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging," *J. Urol.*, vol. 111, no. 1, pp. 58–64, 1974.

[21] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, et al., "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 538–546, 2018.

[22] Di Hu, Chengze Wang, Feiping Nie, et al., "Dense multimodal fusion for hierarchically joint representation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2019, pp. 3941–3945.

[23] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Repr.*, 2019.

[24] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera, *Learning from imbalanced data sets*, vol. 10, Springer, 2018.