Missingness-aware Multimodal Learning for Four-class Adrenal Tumor Classification

Dehua Chen[1], Jing Fang[1], Qihong Wu[2], Pengfei Xia[1] *, and Huimin An[2] *

[1] School of Information and Intelligent Science, Donghua University, Shanghai, China
x6635570@163.com

[2] Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
huiminan12@fudan.edu.cn

Abstract. Accurate classification of adrenal tumors is essential for distinguishing hormonally active tumors from non-functioning adenomas and guiding appropriate management. Current diagnostic workflows typically rely on multi-step endocrine testing and contrast-enhanced CT, which may introduce additional radiation exposure and contrast-related risks. In this study, we investigate four-class adrenal tumor classification—Cushing's syndrome, primary aldosteronism, pheochromocytoma, and non-functioning adenoma—using non-contrast CT combined with routinely collected non-hormonal clinical variables. This task is challenged by substantial variable-level missingness in structured clinical data and pronounced class imbalance in real-world surgical cohorts. We propose a missingness-aware multimodal learning framework that explicitly encodes missing patterns and conditions cross-modal fusion on input reliability. Structured embeddings modulate imaging features via feature-wise affine transformation to enable interaction, while a reliability descriptor derived from missing patterns predicts adaptive fusion weights to dynamically integrate imaging, clinical, and interaction representations under incomplete inputs. To further address imbalance, we introduce a distribution-aware prototype regularization objective on the fused embedding space, encouraging compact intra-class representations and enlarged inter-class margins. Experiments on a retrospective cohort of 1,108 patients demonstrate that the proposed method achieves a macro-F1 of 0.822, balanced accuracy of 0.825, and AUC of 0.963 on a held-out test set, outperforming state-of-the-art unimodal and multimodal baselines. Notably, the method improves recall for minority tumor categories while maintaining stable performance on the majority class. Code is available at https://github.com/fj-sunny/AMAF-Net.

Keywords: Multimodal Fusion · Missingness-aware Learning · Prototype Regularization · Class Imbalance · Adrenal Tumor Classification.

* Co-Corresponding Author

1 Introduction

Adrenal incidentalomas are detected in approximately 2–7% of abdominal CT examinations, with prevalence increasing with age [3,24]. Although most lesions are benign adenomas, distinguishing hormonally active tumors—including primary aldosteronism, pheochromocytoma, and cortisol excess—from non-functioning adenomas is essential for appropriate management [9,24]. Current evaluation integrates imaging and endocrine assessment: non-contrast CT is recommended as the initial modality, whereas further characterization often requires contrast-enhanced multiphase imaging [9], which increases radiation exposure and introduces contrast-related risks [5]. Endocrine work-up (aldosterone–renin ratio, plasma/urinary metanephrines, and dexamethasone suppression testing) is multi-step and can be affected by medication use and patient compliance [1,14,17]. Therefore, reliable classification based on non-contrast CT and routinely available non-hormonal clinical variables, while reducing reliance on complex hormonal testing and contrast-enhanced imaging, remains a clinically important goal.

Recent studies on adrenal tumor classification mainly rely on imaging-based models. Moawad et al. [16] and O'Shea et al. [18] combined radiomics with conventional classifiers to differentiate indeterminate lesions and lipid-poor adenomas from metastases, while Singh et al. [20] and Zhao et al. [26] adopted 3D CNNs and transformer-based architectures for benign–malignant discrimination on contrast-enhanced CT. However, these efforts predominantly focus on binary tasks under enhanced imaging, and a multimodal four-class framework integrating non-contrast CT and clinical data remains lacking. To exploit heterogeneous information, multimodal fusion has evolved from feature concatenation [21] to interaction-based designs: Perez et al. introduced FiLM for affine feature modulation [19], and Zhang et al. incorporated cross-attention for bidirectional interaction [25]. In medical imaging, Wolf et al. proposed DAFT to inject tabular information into CNN backbones [22], and Duenias et al. developed HyperFusion via hypernetworks for dynamic adaptation [8]. Yao et al. proposed DrFuse to improve robustness under missing modalities [23], while Han et al. introduced FuseMoE for mixture-of-experts routing [11]. Despite these advances, missing clinical variables are often handled by imputation or dropout, rather than explicitly guiding adaptive fusion using missingness patterns. Moreover, clinical cohorts frequently exhibit class imbalance: Lin et al. proposed focal loss [15], Cui et al. introduced class-balanced reweighting [4], and Du et al. improved minority discrimination with probabilistic prototypical contrastive learning [7]. Yet these strategies mainly operate at the loss level and are largely developed for unimodal image learning, leaving a gap for multimodal frameworks that jointly address missing clinical data and class imbalance.

In this paper, we propose a missingness-aware multimodal learning framework for four-class adrenal tumor classification, covering primary aldosteronism, Cushing's syndrome, pheochromocytoma, and non-functioning adenoma. The framework integrates non-contrast CT and non-hormonal clinical variables under incomplete and imbalanced clinical settings. Our contributions are threefold:

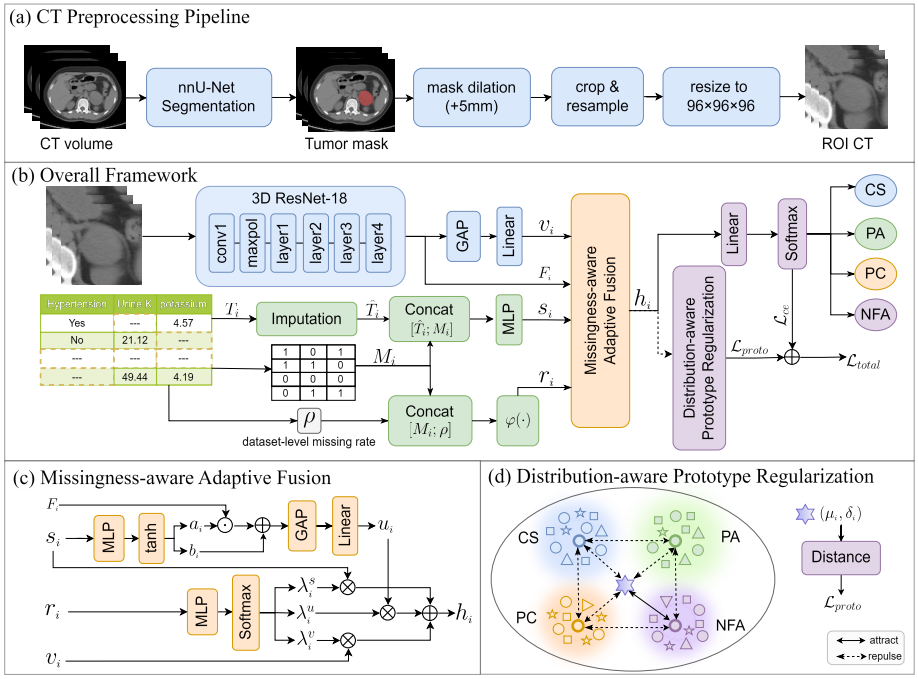

Fig. 1. Overview of the proposed framework. (a) CT preprocessing pipeline using nnU-Net segmentation and ROI normalization. (b) Overall multimodal architecture with explicit missing modeling. (c) Missingness-aware adaptive fusion with element-wise modulation and reliability-guided weighting. (d) Distribution-aware prototype regularization for class imbalance mitigation.

(1) a missingness-aware adaptive fusion mechanism that conditions cross-modal interaction on clinical missingness patterns for sample-specific modulation under incomplete inputs; (2) a distribution-aware prototype regularization objective that promotes compact intra-class representations and enlarged inter-class margins to improve minority-class separability under imbalance; and (3) extensive experiments on a retrospective cohort of 1,108 patients demonstrating consistent improvements over imaging-based and multimodal baselines, with notable gains in macro-F1 and minority-class sensitivity under realistic incomplete and imbalanced conditions.

2 Materials and Method

2.1 Dataset Description and Preprocessing

This retrospective study was approved by the Institutional Ethics Committee of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, and informed

consent was waived due to data de-identification. A total of 1,108 patients who underwent adrenal tumor resection between January 2020 and August 2025 were included, comprising Cushing's syndrome (CS, n = 141), primary aldosteronism (PA, n = 707), pheochromocytoma (PC, n = 143), and non-functioning adenoma (NFA, n = 117), with evident class imbalance.

All patients underwent non-contrast CT. The CT preprocessing pipeline is illustrated in Fig. 1(a). A nnU-Net [13] segmentation model was trained on 300 scans manually annotated by radiologists (Dice = 0.79 on a held-out set) and applied to the full cohort. Tumor masks were expanded by 5 mm, cropped, resampled to $1\times1\times1$ mm, resized to $96\times96\times96$, and intensity-normalized. Structured clinical variables included demographic information, laboratory measurements (biochemical and metabolic indices), 24-hour urinary tests, and tumor-related descriptors. Several variables exhibited substantial variable-level missingness. In particular, hypertension status showed over 50% missingness, while multiple metabolic and 24-hour urinary indicators (e.g., fasting glucose, urine electrolytes, and lipid profiles) had missing rates exceeding 35–45%. This heterogeneous and high missingness motivated the missingness-aware modeling strategy.

2.2 Method Overview

Given a CT ROI and structured clinical variables with potential missing entries, the proposed framework performs missing-aware multimodal adrenal tumor classification under incomplete and imbalanced conditions. The model consists of three key components. First, multimodal feature extraction encodes imaging features from the CT ROI, structured variables, and missing patterns to obtain modality-specific embeddings along with a reliability descriptor. Second, these representations are integrated through a missing-aware adaptive fusion mechanism, which enables cross-modal interaction and dynamically adjusts modality contributions based on input completeness. Third, the fused embedding is optimized with distribution-aware prototype regularization to enhance class separability under imbalance. The overall architecture is illustrated in Fig. 1(b).

Missing Pattern Encoding and Multimodal Feature Extraction Given multimodal samples $\{(X_i, T_i, y_i)\}_{i=1}^{N}$, X_i denotes the CT volume and $T_i \in \mathbb{R}^d$ denotes structured clinical variables with missing entries. A binary missing indicator $M_i \in \{0,1\}^d$ is introduced to encode variable-level incompleteness. The dataset-level missing ratio is defined as

$$\rho = \frac{1}{Nd}\sum_{i,j} M_{ij}. \tag{1}$$

Missing entries are imputed using training-set statistics and standardized. The concatenated vector $[\hat{T}_i; M_i]$ is fed into a multilayer perceptron (MLP) to produce a structured embedding $\mathbf{s}_i \in \mathbb{R}^D$. The imaging modality is encoded by a 3D ResNet-18 backbone followed by global average pooling and a linear projection layer, producing an imaging embedding $\mathbf{v}_i \in \mathbb{R}^D$. To separate reliability

information from discriminative features, the missing pattern is further encoded as

$$\mathbf{r}_i = \Psi([M_i; \rho]), \tag{2}$$

where $\Psi(\cdot)$ is a lightweight MLP. The descriptor $\mathbf{r}_i$ summarizes sample-level incompleteness and serves as a control signal in the subsequent fusion stage.

Missingness-aware Adaptive Fusion Inspired by feature-wise modulation methods such as FiLM [19] and DAFT [22], we propose a missingness-aware adaptive fusion mechanism that integrates structured clinical information into the imaging encoder while accounting for input reliability (Fig. 1(c)).

Specifically, let $\mathbf{F}_i$ denote the high-level feature map extracted from the imaging backbone. The structured embedding $\mathbf{s}_i$ is passed through a lightweight multilayer perceptron with non-linear activation, whose output is split channel-wise into two transformation parameters $\mathbf{a}_i$ and $\mathbf{b}_i$. These parameters share the same channel dimensionality as $\mathbf{F}_i$ and are used to perform feature-wise affine modulation:

$$\tilde{\mathbf{F}}_i = \mathbf{a}_i \odot \mathbf{F}_i + \mathbf{b}_i, \tag{3}$$

where $\odot$ denotes element-wise multiplication. After global average pooling and linear projection, the modulated feature map produces an interaction embedding $\mathbf{u}_i$. Following the same pooling-and-projection scheme, the original imaging embedding $\mathbf{v}_i$ is obtained from the unmodulated imaging pathway.

To preserve complementary evidence, we explicitly maintain three representations: the imaging embedding $\mathbf{v}_i$, the structured embedding $\mathbf{s}_i$, and the interaction embedding $\mathbf{u}_i$. To adaptively adjust modality contributions under heterogeneous missingness, the reliability descriptor $\mathbf{r}_i$ (derived from clinical missing patterns) predicts fusion weights:

$$[\lambda_i^u, \lambda_i^s, \lambda_i^v] = \mathrm{Softmax}\big(\mathrm{MLP}(\mathbf{r}_i)\big), \tag{4}$$

where the weights are non-negative and sum to one.

The final fused embedding is computed as

$$\mathbf{h}_i = \lambda_i^u \mathbf{u}_i + \lambda_i^s \mathbf{s}_i + \lambda_i^v \mathbf{v}_i. \tag{5}$$

Distribution-aware Prototype Regularization Prototype-based contrastive learning, such as ProCo [7], models class-wise feature distributions to improve representation compactness and inter-class separability under imbalance. We apply this idea in the multimodal fusion setting by regularizing the fused representation space $\mathbf{h}_i$, as illustrated in Fig. 1(d).

Specifically, $\mathbf{h}_i$ is projected into a center descriptor $\boldsymbol{\mu}_i$ and a dispersion descriptor $\boldsymbol{\delta}_i$, where $\boldsymbol{\delta}_i$ is constrained to be non-negative to characterize intra-class uncertainty. For each class $k \in \{1, \ldots, K\}$, class-level prototypes $(\bar{\boldsymbol{\mu}}_k, \bar{\boldsymbol{\delta}}_k)$ are estimated from training samples and updated during optimization (in our setting, $K = 4$). The discrepancy between sample i and class k is defined as

$$D(i,k) = \|\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_k\|_2^2 + \beta \sum_j (\boldsymbol{\delta}_{i,j} + \bar{\boldsymbol{\delta}}_{k,j}), \tag{6}$$

where β balances center deviation and dispersion regularization. The prototype regularization loss is formulated as

$$\mathcal{L}_{proto} = -\log \frac{\exp(-D(i, y_i))}{\sum_{k=1}^{K} \exp(-D(i, k))}. \tag{7}$$

The final training objective combines classification and prototype regularization:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_{proto}\mathcal{L}_{proto}. \tag{8}$$

3 Experiment

3.1 Experimental Settings and Evaluation

We adopted patient-level five-fold stratified cross-validation with an independent 20% held-out test set for final reporting. Models were trained using AdamW (learning rate 1e-4, weight decay 0.1, batch size 8) for up to 100 epochs, with early stopping (patience = 15) based on validation macro-F1 and ReduceLROnPlateau scheduling. The training objective combined cross-entropy with the proposed prototype alignment loss. The prototype loss weight was set to $\lambda_{proto} = 0.1$, the uncertainty coefficient in Eq. (6) to $\beta = 0.1$, and the temperature to $\tau = 0.07$. Class prototypes were updated by exponential moving average with momentum 0.9. All models incorporating structured clinical inputs used the same preprocessing pipeline, including statistical imputation and concatenation of binary missing indicators. Test-set performance was evaluated using macro-F1 (primary metric), balanced accuracy (BAcc), overall accuracy (Acc), and macro-averaged AUC. We further report per-class recall and the Pearson correlation between clinical missing ratio and fusion weights to characterize behavior under high missingness.

3.2 Comparison with State-of-the-Art Methods

We compared the proposed method with representative unimodal and multimodal baselines, reporting performance on the held-out test set. Unimodal models included ResNet18 [12], 3D U-Net [2], Vision Transformer (ViT) [6], multilayer perceptron (MLP), and FT-Transformer [10], while multimodal baselines covered Concat, Late Fusion, Cross-attention [25], DAFT [22], HyperFusion [8], and DrFuse [23]. As summarized in Table 1, multimodal methods consistently outperformed unimodal models, and interaction-based fusion strategies yielded further gains over simple concatenation. The proposed method achieved the best overall performance (macro-F1 0.822, BAcc 0.825, AUC 0.963), exceeding all compared approaches and improving macro-F1 by 4.6% over the strongest baseline (DrFuse, 0.776). Table 2 further shows that improvements are most evident in minority categories. Recall for CS increased from 0.750 to 0.821 and for NFA from 0.667 to 0.708, while performance on the majority class PA remained stable (0.943). These gains directly contribute to higher macro-F1 and balanced accuracy, indicating enhanced discrimination of underrepresented classes rather than shifts in overall prediction bias.

Table 1. Comparison with Uni- and Multimodal Baselines for Adrenal Tumor Classification.

Category	Method	AUC↑	Macro-F1↑	Acc↑	BAcc↑
Image-only	ViT [6]	0.738	0.427	0.622	0.462
	3D U-Net [2]	0.871	0.606	0.772	0.584
	ResNet18 [12]	**0.857**	**0.635**	**0.783**	**0.628**
Clinical-only	FT-Transformer [10]	0.935	0.651	0.800	0.655
	MLP	0.938	0.657	0.800	0.663
	MLP+Missingness	**0.943**	**0.702**	**0.822**	**0.703**
Fusion baselines	Concat	0.924	0.677	0.806	0.677
	Cross-attention [25]	0.943	0.698	0.827	0.703
	Late Fusion	0.941	0.734	0.842	0.725
	DAFT [22]	0.928	0.725	0.832	0.720
	HyperFusion [8]	**0.954**	0.759	0.849	0.764
	DrFuse [23]	0.949	**0.776**	**0.855**	**0.788**
Proposed	Ours	**0.963**	**0.822**	**0.887**	**0.825**

Table 2. Per-class recall (CS=28, PA=141, PC=29, NFA=24).

Method	CS	PA	PC	NFA
Cross-attention [25]	0.714	0.929	0.793	0.375
HyperFusion [8]	0.786	0.922	0.724	0.625
DrFuse [23]	0.750	0.908	0.828	0.667
Ours	**0.821**	**0.943**	**0.828**	**0.708**

3.3 Mechanism Analysis

Missing-aware Behavior and Representation Analysis As the missing ratio increases, the clinical weight (λ_s) decreases significantly (Pearson $r = -0.94$, $p < 0.001$), while imaging (λ_v) and interaction weights (λ_u) increase ($r = +0.62$ and $+0.71$), indicating that the model adaptively shifts reliance from incomplete clinical inputs to complementary modalities, as shown in Table 3. Embedding visualization further reveals that Cross-Entropy leads to evident class overlap, and Focal Loss only partially improves cluster compactness. In contrast, prototype regularization produces tighter intra-class clusters and clearer inter-class

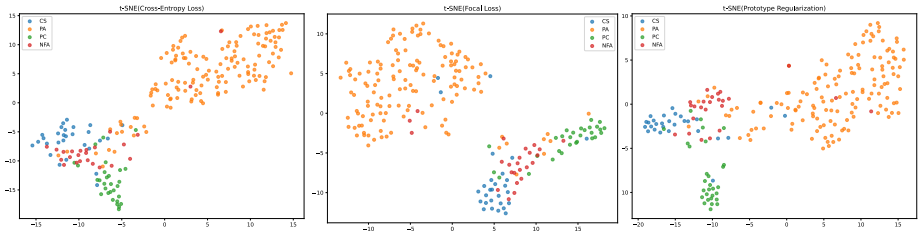

Fig. 2. t-SNE Visualization of Embedding Distributions under Different Objectives.

Table 3. Missingness-conditioned fusion weights on the test set. Pearson correlation coefficients between missing ratio and fusion weights are all statistically significant ($p < 0.001$). Arrows indicate the direction of change as missing ratio increases.

Missing Ratio	$\lambda_v \uparrow$	$\lambda_s \downarrow$	$\lambda_u \uparrow$
0–10%	0.330	0.291	0.379
10–30%	0.342	0.274	0.384
30–50%	0.355	0.258	0.387
>50%	0.370	0.235	0.395
Pearson r	**+0.62**	**-0.94**	**+0.71**

Table 4. Ablation study on missing-aware fusion and imbalance objective.

Variant	AUC↑	Macro-F1↑	Acc↑	BAcc↑
w/o Missing Indicator	0.949	0.744	0.809	0.777
w/o Reliability Descriptor	0.962	0.787	0.868	0.802
w/o Adaptive Weighting	0.958	0.778	0.865	0.783
w/o Interaction Branch	0.947	0.739	0.822	0.762
w/o Imaging Branch	0.949	0.747	0.833	0.771
w/o Clinical Branch	0.957	0.765	0.851	0.778
w/o $\mathcal{L}_{proto}$	0.960	0.779	0.868	0.788
Replace with Focal Loss	0.955	0.793	0.865	0.809
Ours	**0.963**	**0.822**	**0.887**	**0.825**

margins, resulting in a more discriminative representation space, as illustrated in Fig. 2. This structural improvement aligns with the observed gain in Macro-F1.

Ablation Study Performance consistently drops when removing missing-aware components (indicator, reliability descriptor, or adaptive weighting), demonstrating the necessity of dynamic fusion. Eliminating modality or interaction branches further degrades results, indicating that both unimodal representations and cross-modal interactions are essential, as summarized in Table 4. Replacing prototype alignment with cross-entropy or Focal Loss reduces Macro-F1, showing that prototype regularization enforces stronger class-level structure than reweighting strategies. The complete model achieves the best overall performance.

4 Conclusion

In summary, this study advances four-class adrenal tumor classification by moving beyond conventional single-modality approaches and integrating non-contrast CT with routinely collected non-hormonal clinical data. In contrast to prior imaging-dominated studies, our work demonstrates the feasibility and benefit of multimodal modeling under realistic clinical conditions characterized by substantial missingness and class imbalance. Experimental results on a retrospective cohort of 1,108 patients show that our framework achieves superior macro-F1, balanced accuracy, and AUC compared with state-of-the-art baselines, with notably improved recall for minority tumor categories. Future work will focus on

external multi-center validation to further evaluate generalizability and support clinical translation.

Acknowledgments. This work was supported in part by the National Key Research and Development Program of China (Grant No. 2024YFE0102100), the Fundamental Research Funds for the Central Universities, the DHU Distinguished Young Professor Program, Noncommunicable Chronic Diseases-National Science and Technology Major Project (Grant No. 2025ZD0552200).

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Adler, G.K., Stowasser, M., Correa, R.R., et al.: Primary aldosteronism: an endocrine society clinical practice guideline. Journal of Clinical Endocrinology & Metabolism **110**(9), 2453–2495 (2025). https://doi.org/10.1210/clinem/dgaf284
2. Alimu, P., Fang, C., Han, Y., et al.: Artificial intelligence with a deep learning network for the quantification and distinction of functional adrenal tumors based on contrast-enhanced CT images. Quantitative Imaging in Medicine and Surgery **13**(4), 2675 (2023). https://doi.org/10.21037/qims-22-539
3. Bovio, S., Cataldi, A., Reimondo, G., et al.: Prevalence of adrenal incidentaloma in a contemporary computerized tomography series.Journal of endocrinological investigation **29**(4), 298–302 (2006). https://doi.org/10.1007/BF03344099
4. Cui, Y., Jia, M., Lin, T.-Y., et al.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR). pp. 9268–9277. IEEE, Long Beach (2019). https://doi.org/10.1109/CVPR.2019.00949
5. Davenport, M.S., Perazella, M.A., Yee, J., et al.: Use of intravenous iodinated contrast media in patients with kidney disease: consensus statements from the American College of Radiology and the National Kidney Foundation. Radiology **294**(3), 660–668 (2020). https://doi.org/10.1148/radiol.2019192094
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021). https://doi.org/10.48550/arXiv.2010.11929
7. Du, C., Wang, Y., Song, S., et al.: Probabilistic contrastive learning for long-tailed visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). **46**(9), 5890–5904 (2024). https://doi.org/10.1109/TPAMI.2024.3369102
8. Duenias, D., Nichyporuk, B., Arbel, T., et al.: HyperFusion: a hypernetwork approach to multimodal integration of tabular and medical imaging data for predictive modeling. Medical Image Analysis. **102**, 103503 (2025). https://doi.org/10.1016/j.media.2025.103503
9. Fassnacht, M., Tsagarakis, S., Terzolo, M., et al.: European Society of Endocrinology clinical practice guidelines on the management of adrenal incidentalomas. European Journal of Endocrinology. **189**(1), G1–G42 (2023). https://doi.org/10.1093/ejendo/lvad066

10. Gorishniy, Y., Rubachev, I., Khrulkov, V., et al.: Revisiting deep learning models for tabular data. In: Advances in Neural Information Processing Systems 34, pp. 18932–18943 (2021).
11. Han, X., Nguyen, H., Harris, C., et al.: FuseMoE: mixture-of-experts transformers for fleximodal fusion. In: Advances in Neural Information Processing Systems 37 (2024). https://doi.org/10.52202/079017-2167
12. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 770–778. IEEE, Las Vegas (2016). https://doi.org/10.1109/CVPR.2016.90
13. Isensee, F., Jaeger, P.F., Kohl, S.A.A., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
14. Lenders, J.W.M., Duh, Q.Y., Eisenhofer, G., et al.: Pheochromocytoma and paraganglioma: an endocrine society clinical practice guideline. The Journal of Clinical Endocrinology & Metabolism **99**(6), 1915–1942 (2014). https://doi.org/10.1210/jc.2014-1498
15. Lin, T.-Y., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision(ICCV), pp. 2980–2988. IEEE, Venice (2017). https://doi.org/10.1109/ICCV.2017.324
16. Moawad, A.W., Ahmed, A., Fuentes, D.T., et al.: Machine learning-based texture analysis for differentiation of radiologically indeterminate small adrenal tumors. Abdominal Radiology **46**(10), 4853–4863 (2021). https://doi.org/10.1007/s00261-021-03136-2
17. Nieman, L.K., Biller, B.M.K., Findling, J.W., et al.: The diagnosis of Cushing's syndrome: an endocrine society clinical practice guideline. The Journal of Clinical Endocrinology & Metabolism **93**(5), 1526–1540 (2008). https://doi.org/10.1210/jc.2008-0125
18. O'Shea, A., Kilcoyne, A., McDermott, E., et al.: Can radiomic feature analysis differentiate adrenal metastases from lipid-poor adenomas on single-phase contrast-enhanced CT abdomen? Clinical Radiology **77**(10), e711–e718 (2022). https://doi.org/10.1016/j.crad.2022.05.014
19. Perez, E., Strub, F., De Vries, H., et al.: FiLM: visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). https://doi.org/10.1609/aaai.v32i1.11671
20. Singh, Y., Kelm, Z.S., Faghani, S., et al.: Deep learning approach for differentiating indeterminate adrenal masses using CT imaging. Abdominal Radiology **48**(10), 3189–3194 (2023). https://doi.org/10.1007/s00261-023-03988-w
21. Spasov, S., Passamonti, L., Duggento, A., et al.: A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. NeuroImage **189**, 276–287 (2019). https://doi.org/10.1016/j.neuroimage.2019.01.031
22. Wolf, T.N., Pölsterl, S., Wachinger, C., et al.: DAFT: a universal module to interweave tabular data and 3D images in CNNs. NeuroImage **260**, 119505 (2022). https://doi.org/10.1016/j.neuroimage.2022.119505
23. Yao, W., Yin, K., Cheung, W.K., et al.: DrFuse: learning disentangled representation for clinical multi-modal fusion with missing modality. In: Proceedings of the AAAI Conference on Artificial Intelligence 38, pp. 16416–16424 (2024). https://doi.org/10.1609/aaai.v38i15.29578
24. Young, W.F.: The incidentally discovered adrenal mass. The New England Journal of Medicine **356**(6), 601–610 (2007). https://doi.org/10.1056/NEJMcp065470

25. Zhang, J., He, X., Liu, Y., et al.: Multi-modal cross-attention network for Alzheimer's disease diagnosis. Computers in biology and medicine **162**, 107050 (2023). https://doi.org/10.1016/j.compbiomed.2023.107050
26. Zhao, K., Sun, Z., Jiang, H., et al.: Transformer-based integration of radiomics and deep learning for differentiating lipid-poor adrenal adenomas from malignant tumors. Meta-Radiology 100183 (2025). https://doi.org/10.1016/j.metrad.2025.100183