Toward Thyroid Cancer BRAF V600E Mutation Prediction via Multimodal Large Language Model: Dataset and Model Development

Pengfei Hao[1], Shuaibo Li[1], Hongqiu Wang[1], Sixu He[1], Chunwang Huang[2(✉)], and Lei Zhu[1,3(✉)]

[1] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China leizhu@hkust-gz.edu.cn
[2] Guangdong Provincial People's Hospital
[3] The Hong Kong University of Science and Technology

Abstract. Thyroid cancer is the ninth most common malignancy globally, with Papillary Thyroid Carcinoma (PTC) representing the most prevalent subtype. The BRAF V600E mutation has emerged as a decisive diagnostic and prognostic biomarker of PTC, essential for risk stratification and personalized therapy. However, its gold-standard detection relies on complex and expensive sequencing technology. Therefore, developing advanced deep learning methods to predict BRAF mutation by integrating multimodal clinical data offers a cost-effective solution to facilitate accurate diagnosis and individualized risk stratification. In this paper, we present the first comprehensive multimodal dataset for thyroid cancer BRAF V600E mutation prediction, **MM-BRAF dataset**, integrating 3120 high-quality ultrasound images, annotated tumor regions, ultrasound reports, clinicopathological records, and BRAF mutation status. Furthermore, we propose **BRAF-MLLM**, a novel and first Multimodal Large Language Model (MLLM) framework for thyroid cancer BRAF V600E mutation prediction. Given that mutation-relevant signatures in ultrasound imagery are often subtle and localized within minuscule tumor regions, we design the Tumor-centric Multimodal Dual-scale Alignment (TMDA) module. During the training stage, TMDA explicitly decouples visual features into intra-tumoral and extra-tumoral scales, aligning them with fine-grained clinical semantics to suppress background noise and enhance the perception of tumor-specific indicators. To validate the effectiveness of our approach, we establish a comprehensive benchmark for this novel task using traditional multimodal fusion models and mainstream MLLMs. Extensive experiments subsequently demonstrate that BRAF-MLLM consistently outperforms all baseline methods, showcasing its significant potential for precise clinical decision-making. The code and dataset are available at: https://github.com/FiFi-HAO467/BRAF-MLLM.

Keywords: Thyroid Cancer · BRAF Mutation · Multimodal Large Language Model

✉ Chunwang Huang and Lei Zhu are the corresponding authors of this work.

1 Introduction

Thyroid cancer ranks as the ninth most common malignancy globally, with Papillary Thyroid Carcinoma (PTC) representing the predominant subtype, accounting for approximately 84% of all cases [7]. Although PTC is generally considered highly curable, its clinical management is complicated by a significant recurrence rate and potential mortality in advanced cases [31]. Crucially, the B-type raf kinase (BRAF) V600E mutation has been identified as the most prevalent genetic alteration in PTC, serving as a decisive driver of tumor aggressiveness [26]. Given its strong correlation with poor prognosis, increased risk of recurrence, and diminished radioiodine uptake sensitivity [27,28], the BRAF V600E status has become a cornerstone for precise risk stratification. Consequently, precise assessment of this mutation is of paramount importance for prognostic evaluation and the formulation of adjuvant therapeutic strategies.

Currently, BRAF mutation detection relies on invasive surgical or fine-needle aspiration specimens analyzed via Sanger sequencing or other sequencing technology [24]. Many clinical studies have confirmed that the BRAF V600E mutation is significantly associated with aggressive clinicopathological features, such as extrathyroidal extension and lymph node metastasis [13]. Furthermore, despite minor discrepancies in specific indicators, existing studies consistently demonstrate the correlation between BRAF V600E mutation and the tumor's morphological phenotypes in ultrasound (US) images [16,17]. However, translating these clinical insights into accurate mutation prediction is highly challenging. Manual statistical evaluations often fail to integrate diverse data sources effectively, while early AI explorations remain restricted to single-modal US analysis [30] . Neither approach fully exploits the complex synergy between visual phenotypes and clinicopathological narratives, which often manifests as latent, high-dimensional correlations. Consequently, a multimodal framework capable of comprehensively understanding US images and clinical information is essential to transcend the limitations of single-source analysis for precise BRAF prediction.

Driven by these insights, we present the first large-scale and comprehensive dataset, **MM-BRAF dataset**, specifically curated for BRAF V600E mutation prediction, integrating 3120 high-quality US images with their corresponding US reports, expert-annotated tumor regions, clinicopathological records, and ground-truth BRAF mutation status. Building upon this foundation, we propose **BRAF-MLLM**, a novel and first Multimodal Large Language Model (MLLM) [29] framework meticulously designed for thyroid cancer BRAF V600E mutation prediction. Precisely predicting BRAF mutation is a cognitively demanding task that requires not only a thorough understanding of clinical semantics and visual characteristics but also the logical reasoning capability to decipher the subtle, deep-seated correlations between them. Recognizing this inherent complexity, we actively employ MLLM as our foundational architecture, leveraging its robust cross-modal reasoning [9] and inference capabilities [10,11] to bridge the gap between heterogeneous clinical data and molecular genotypes.

However, general MLLMs often lack expertise in understanding US images and specialized clinical terminology. Moreover, clinical experience indicates that

mutation-relevant visual signatures are typically confined to minuscule tumor regions, which are frequently obscured by the inherent speckle noise and blurred margins of US images. Consequently, these subtle diagnostic signals are often overwhelmed within global image representations, making it difficult for MLLMs to accurately align subtle visual cues with their corresponding clinical descriptions, thereby limiting the precision of BRAF prediction [22]. To overcome these limitations, we design the **Tumor-centric Multimodal Dual-scale Alignment (TMDA)** module. During the training stage, by explicitly decoupling visual features into intra-tumoral and extra-tumoral scales and precisely aligning them with clinical semantic descriptions, the TMDA module effectively suppresses background noise and focuses on localized pathological signatures. This fine-grained perception improves the MLLM's capability to interpret nuanced US manifestations, ultimately enabling a more robust mapping from imaging phenotype to BRAF genotype and markedly improving prediction performance.

The main contributions of this work are summarized as follows: (1) we present the first comprehensive multimodal dataset, MM-BRAF dataset, for thyroid cancer BRAF V600E mutation prediction, integrating ultrasound images, annotated tumor regions, ultrasound reports, BRAF mutation status, and clinicopathological records; (2) we propose BRAF-MLLM, a novel and first MLLM framework for thyroid cancer BRAF V600E mutation prediction; (3) we design a Tumor-centric Multimodal Dual-scale Alignment (TMDA) module that enhances the capture of subtle signatures through fine-grained dual-scale cross-modal alignment; (4) we establish a comprehensive benchmark for this novel task. Extensive experiments demonstrate that our proposed BRAF-MLLM achieves state-of-the-art performance, showcasing its significant potential for precise clinical decision-making.

2 Methodology

2.1 Overview of BRAF-MLLM Framework

The BRAF-MLLM framework is designed to comprehend multimodal clinical data for the precise prediction of BRAF V600E mutations. As illustrated in Fig. 1, the overall architecture follows a three-stage pipeline: multimodal feature extraction, tumor-centric dual-scale alignment, and MLLM-based reasoning. Specifically, for each case, the input consists of a US image along with its corresponding US report and clinicopathological records. We employ the pretrained vision encoder and tokenizer of the MLLM to extract the visual features and text embeddings, respectively. Subsequently, these features are processed by the TMDA module, where intra-tumoral and extra-tumoral signatures are explicitly decoupled and aligned with fine-grained clinical semantics. The aligned multimodal representations are then fused and fed into the MLLM. To adapt the model to the specific domain of thyroid oncology, we utilize Low-Rank Adaptation (LoRA) [14] for efficient fine-tuning of the MLLM. Leveraging its enhanced cross-modal reasoning and logical inference capabilities, the model processes the integrated clinical context to output the final BRAF mutation prediction.

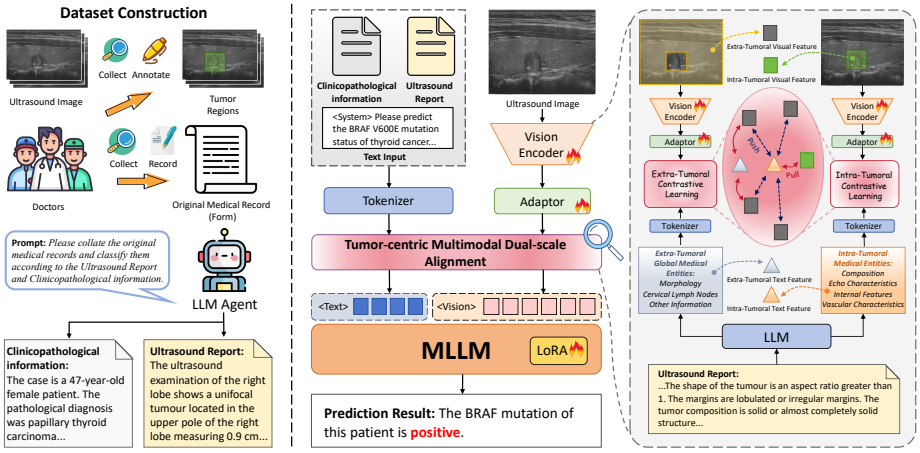

Fig. 1. Overview of our proposed method. The left part shows the construction pipeline of our MM-BRAF dataset. The middle part illustrates the BRAF-MLLM framework, where US images are aligned with clinical text and subsequently input into the MLLM for BRAF V600E prediction. The right part details the Tumor-centric Multimodal Dual-scale Alignment (TMDA) module, designed to explicitly decouple visual features into intra-tumoral and extra-tumoral scales and align them with fine-grained clinical semantics to enhance tumor-specific perception.

2.2 Dataset Construction

The construction of our MM-BRAF dataset follows a structured pipeline comprising data acquisition, expert annotation, and automated curation. As illustrated in the left part of Fig. 1, we first collect a comprehensive set of raw data for each patient, including US images, original medical records, and gold-standard BRAF gene mutation status. Subsequently, experienced clinicians perform professional annotations on the US images to generate precise tumor regions. To transform the unstructured medical forms into standardized clinical insights, we employ GPT-4o as a Large Language Model (LLM) [32] agent to synthesize and structure the raw data. Guided by task-specific prompts, the LLM Agent automatically collates and categorizes these records into two distinct components: clinicopathological information and US reports. Clinicopathological information integrates clinical demographics (e.g., age and gender) with pathological findings (e.g., histopathological diagnosis and lymph node metastasis). US reports include tumor shape, sonographic features, and so on. This hybrid curation approach ensures the generation of high-quality, fine-grained multimodal data, providing a robust foundation for the subsequent training of BRAF-MLLM.

2.3 Tumor-centric Multimodal Dual-scale Alignment

Clinical experience suggests that BRAF V600E mutations are intrinsically linked to specific tumor phenotypic characteristics [16,17]. While general MLLMs ex-

hibit robust reasoning in natural domains, they often encounter "domain blindness" when interpreting thyroid US images. This is primarily attributed to two factors: the inherent complexity of sonographic patterns coupled with specialized clinical terminology, and the fact that the lesion typically occupies only a small fraction of the image, where mutation-relevant signatures are frequently obscured by speckle noise. Traditional global alignment methods tend to bridge textual cues with the entire visual field [21]; consequently, irrelevant background features interfere with the modeling of subtle clinical indicators, causing the sparse but critical diagnostic signals to be obscured or lost in the global feature space.

To surmount these challenges, we design the TMDA module, which implements the decouple-then-align strategy. From a clinical perspective, indicators of BRAF mutations are distributed across two scales: intra-tumoral features (e.g., microcalcifications and solid composition) and extra-tumoral signatures (e.g., cervical lymph node and peritumoral morphology). By leveraging clinician-annotated tumor regions, TMDA explicitly decouples the US images into intra-tumoral and extra-tumoral visual features. Simultaneously, we utilize an LLM agent to categorize text from US reports into intra-tumoral medical entities and extra-tumoral medical entities based on predefined clinical rules. Finally, through contrastive learning, these dual-scale visual features are aligned with their corresponding semantic entities. This mechanism compels the model to perform focal precision sensing on the lesion itself while maintaining contextual awareness of the surrounding tissues [20]. Such fine-grained alignment ensures that the MLLM can bypass extraneous noise and focus on the genotype-phenotype associations within the tumor region, thereby enhancing the accuracy of BRAF prediction.

Specifically, the visual representations are extracted via a pre-trained vision encoder and a learnable adaptor, which are explicitly decoupled into the intra-tumoral visual feature f_{intra} and the extra-tumoral visual feature f_{extra}. Concurrently, the US report T_{us} is partitioned by an LLM agent into two sets of semantic entities, T_{intra} and T_{extra}, representing intra-tumoral and extra-tumoral clinical descriptions. These entities are processed by the tokenizer to generate the corresponding semantic embeddings t_{intra} and t_{extra}. The mathematical formulations are as follows:

$$T_s = \Psi_{\text{LLM}}(T_{\text{us}}, s), \tag{1}$$

$$f_s = \phi(\Phi_v(I, M_s)), \quad t_s = \text{Tokenizer}(T_s), \quad s \in \{\text{intra, extra}\}, \tag{2}$$

where Ψ_{LLM} represents the LLM agent, Φ_v denotes the vision encoder and ϕ represents the learnable adaptor.

To bridge the genotype-phenotype gap, we formulate a dual-scale contrastive learning objective. For a batch of N samples, the intra-tumoral alignment loss $\mathcal{L}_{\text{intra}}$ and the extra-tumoral alignment loss $\mathcal{L}_{\text{extra}}$ are independently defined as:

$$\mathcal{L}_{\text{intra}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(f_{i,\text{intra}}, t_{i,\text{intra}})/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(f_{i,\text{intra}}, t_{j,\text{intra}})/\tau\right)}, \tag{3}$$

$$\mathcal{L}_{\text{extra}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(f_{i,\text{extra}}, t_{i,\text{extra}})/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(f_{i,\text{extra}}, t_{j,\text{extra}})/\tau\right)}. \tag{4}$$

The total multimodal alignment loss $\mathcal{L}_{\text{align}}$ is then computed as a weighted combination of the two scale-specific alignment objectives:

$$\mathcal{L}_{\text{align}} = \alpha\mathcal{L}_{\text{intra}} + \beta\mathcal{L}_{\text{extra}}, \tag{5}$$

where α and β are weighting coefficients that control the relative contributions of the intra-tumoral and extra-tumoral alignment losses, respectively. $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is the temperature hyper-parameter.

3 Experiments and Results

3.1 Dataset and Implementation Details

Dataset. The MM-BRAF dataset consists of 3,120 multimodal samples from a clinical cohort of 1,258 patients, each of whom has a single confirmed lesion with several multi-view US images. To capture complementary spatial information, these distinct US views are treated as individual samples. Specifically, each sample integrates a US image, its associated clinical report, annotated tumor region, clinicopathological record, and the BRAF mutation status. To prevent data leakage, all data from the same patient were assigned to the same subset, and the dataset was split into training and test sets at a 7:3 ratio. After patient-level splitting, the training set was kept approximately balanced, with 56.36% BRAF V600E-positive cases and 43.64% negative cases, while the test set retained its natural distribution, with 66.99% positive cases.

Implementation Details. For performance evaluation, we employ standard classification metrics for MLLM, including Accuracy (Acc) [12], F1-Score [12], and Recall [12]. Our framework is implemented in PyTorch and trained for 2 epochs on a cluster of 8 NVIDIA RTX 4090 GPUs. Utilizing Qwen2.5-VL-7B as the base model, we apply LoRA for efficient MLLM fine-tuning. The learning rate is set to 1×10^{-5} with a per-device batch size of 1 and gradient accumulation steps of 16.

3.2 Experiment Results

As illustrated in Table 1, we evaluate the performance of our BRAF-MLLM against several baseline models on the MM-BRAF dataset. The comparison covers two main categories: traditional multimodal fusion methods and MLLMs.

Comparisons with Traditional Multimodal Fusion Methods. As shown in Table 1, our proposed BRAF-MLLM consistently outperforms all traditional multimodal fusion baselines across all evaluation metrics. Specifically, compared to the strongest traditional baseline, GPT2, our model achieves an accuracy of 0.7735, representing a 2.78% absolute improvement over GPT2's 0.7457. Regarding F-Score and Recall, BRAF-MLLM yields 0.6979 and 0.7129, surpassing

Table 1. Performance comparison of traditional multimodal fusion methods, Multimodal Large Language Models, and our proposed BRAF-MLLM on the MM-BRAF dataset. SFT denotes Supervised Fine-Tuning.

Method	Acc	F-Score	Recall
Traditional Multimodal Fusion Methods			
VisualBERT [19]	0.7382	0.6632	0.6701
BlockTucker [6]	0.7040	0.6462	0.6510
MUTAN [5]	0.7115	0.6588	0.6534
GPT2 [25]	0.7457	0.6644	0.6529
Multimodal Large Language Models			
GPT-4o [15]	0.6049	0.6346	0.6185
Qwen3-VL [3]	0.5695	0.6443	0.6521
Gemini 2.5 [8]	0.6128	0.5884	0.6005
Claude 3.7 Sonnet [1]	0.5591	0.6640	0.6418
Qwen-VL-Max [2]	0.6004	0.6400	0.6127
LLaVA-Med [18]	0.4280	0.3971	0.4795
LLaVA-1.5-7B [23] (SFT)	0.7211	0.6371	0.6215
Qwen2.5-VL-7B [4] (SFT)	0.7443	0.6598	0.6383
BRAF-MLLM (Ours)	**0.7735**	**0.6979**	**0.7129**

the category leaders (GPT2 at 0.6644 and VisualBERT at 0.6701) by 3.35% and 4.28%, respectively. These results demonstrate that our MLLM-based architecture possesses a superior capability for multimodal reasoning and feature integration compared to traditional fusion methods.

Comparisons with MLLMs. We further evaluate BRAF-MLLM against a range of outstanding MLLMs. First, compared to mainstream general-purpose MLLMs such as GPT-4o, Qwen3-VL, Gemini 2.5, Claude 3.7 Sonnet, and Qwen-VL-Max, our model demonstrates a substantial superiority across all metrics. This indicates that despite their vast pre-training knowledge, general MLLMs struggle with the specialized diagnostic reasoning required for BRAF mutation prediction. Second, although LLaVA-Med is specifically tailored for the medical domain, it underperforms on the BRAF mutation prediction task. This suggests that broad medical knowledge alone is insufficient to capture the unique multimodal patterns inherent in BRAF prediction, thereby underscoring the formidable challenge of this specialized diagnostic objective. Finally, while supervised fine-tuning (SFT) significantly boosts the performance of LLaVA-1.5-7B and Qwen2.5-VL-7B, our BRAF-MLLM still consistently outperforms these fine-tuned MLLMs. In summary, these results demonstrate that our proposed BRAF-MLLM achieves significant advantages over both zero-shot and fine-tuned MLLMs in the precise prediction of BRAF mutations.

Table 2. Ablation studies of different components in our BRAF-MLLM. "US-Img" and "Text" denote the inclusion of US images and clinical text. "TMDA" refers to our proposed Tumor-centric Multimodal Dual-scale Alignment module.

Method	US-Img	Text	Alignment	Acc	F-Score	Recall
M1	✓	–	–	0.7254	0.6432	0.6107
M2	–	✓	–	0.7115	0.6260	0.5924
M3	✓	✓	✗	0.7443	0.6598	0.6383
M4	✓	✓	CL	0.7610	0.6825	0.6779
Ours	✓	✓	TMDA	**0.7735**	**0.6979**	**0.7129**

3.3 Ablation Studies

To validate the effectiveness of the components in BRAF-MLLM, we conduct a series of ablation experiments as presented in Table 2. We first evaluate the necessity of utilizing multimodal data by comparing unimodal baselines, M1 (utilizing only US images) and M2 (utilizing only clinicopathological text), with the multimodal fusion approach M3. The superior performance of M3 over both M1 and M2 demonstrates that neither modality in isolation provides sufficient diagnostic information, confirming that multimodal integration is essential for capturing the complex genotype-phenotype associations required for accurate BRAF mutation prediction.

Furthermore, we evaluate our alignment strategy by comparing BRAF-MLLM against M3, which lacks explicit cross-modal alignment, and M4, which employs traditional global contrastive learning (CL) between the US images and the entire US reports. BRAF-MLLM significantly outperforms both baselines, demonstrating the superiority of the TMDA module. The results indicate that by aligning specific semantic entities with local regions, our approach learns a more discriminative representation of subtle diagnostic features that global contrastive alignment tends to obscure. This leads to more robust and accurate BRAF mutation predictions.

4 Conclusion

In this work, we introduce the MM-BRAF dataset, the first comprehensive multimodal dataset specifically designed for thyroid cancer BRAF V600E mutation prediction. Building upon this, we present BRAF-MLLM, a pioneering framework that leverages a TMDA module to align dual-scale visual features with fine-grained clinical semantics. By explicitly decoupling intra-tumoral and extra-tumoral signatures, our method effectively suppresses background noise and enhances the perception of subtle, mutation-relevant morphological indicators. Extensive experiments demonstrate that BRAF-MLLM consistently outperforms all baseline methods across all metrics. Future work will focus on expanding the dataset to include multi-center cohorts, aiming to further validate the model's robustness and generalizability across diverse clinical settings.

Acknowledgement. This work was supported by the Guangdong Science and Technology Department (2024ZDZX2004) and the Program of Guangdong Education Department.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anthropic, C.: 3.7 sonnet and claude code (2025)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
5. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2612–2620 (2017)
6. Ben-Younes, H., Cadene, R., Thome, N., Cord, M.: Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8102–8109 (2019)
7. Boucai, L., Zafereo, M., Cabanillas, M.E.: Thyroid cancer: a review. Jama **331**(5), 425–435 (2024)
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. Hao, P., Li, S., Wang, H., Kou, Z., Zhang, J., Yang, G., Zhu, L.: Surgery-r1: Advancing surgical-vqla with reasoning multimodal large language model via reinforcement learning. arXiv preprint arXiv:2506.19469 (2025)
10. Hao, P., Wang, H., Li, S., Xing, Z., Yang, G., Wu, K., Zhu, L.: Surgical-mamballm: Mamba2-enhanced multimodal large language model for vqla in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 573–583. Springer (2025)
11. Hao, P., Wang, H., Yang, G., Zhu, L.: Enhancing visual reasoning with llm-powered knowledge graphs for visual question localized-answering in robotic surgery. IEEE Journal of Biomedical and Health Informatics (2025)
12. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process **5**(2), 1 (2015)
13. Howell, G.M., Nikiforova, M.N., Carty, S.E., Armstrong, M.J., Hodak, S.P., Stang, M.T., McCoy, K.L., Nikiforov, Y.E., Yip, L.: Braf v600e mutation independently predicts central compartment lymph node metastasis in patients with papillary thyroid cancer. Annals of surgical oncology **20**, 47–52 (2013)

14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
15. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
16. Hwang, J., Hee Shin, J., Han, B.K., Ko, E.Y., Kang, S.S., Kim, J.W., Chung, J.H.: Papillary thyroid carcinoma with braf v600e mutation: Sonographic prediction. American Journal of Roentgenology **194**(5), W425–W430 (2010)
17. Kwak, J.Y., Kim, E.K., Chung, W.Y., Moon, H.J., Kim, M.J., Choi, J.R.: Association of brafv600e mutation with poor clinical prognostic factors and us features in korean patients with papillary thyroid microcarcinoma. Radiology **253**(3), 854–860 (2009)
18. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
19. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
20. Li, S., Ma, W., Guo, J., Xu, S., Li, B., Zhang, X.: Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12523–12533 (2024)
21. Li, S., Xu, S., Ma, W., Zong, Q.: Image manipulation localization using attentional cross-domain cnn features. IEEE Transactions on Neural Networks and Learning Systems **34**(9), 5614–5628 (2021)
22. Li, S., Zhang, L., Ma, W., Guo, J., Xu, S., Qiu, Z., Zha, H.: Realnet: Efficient and unsupervised detection of ai-generated images via real-only representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 38889–38898 (2026)
23. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
24. Nikiforova, M.N., Wald, A.I., Roy, S., Durso, M.B., Nikiforov, Y.E.: Targeted next-generation sequencing panel (thyroseq) for detection of mutations in thyroid cancer. The Journal of Clinical Endocrinology & Metabolism **98**(11), E1852–E1860 (2013)
25. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
26. Xing, M.: Prognostic utility of braf mutation in papillary thyroid cancer. Molecular and cellular endocrinology **321**(1), 86–93 (2010)
27. Xing, M., Alzahrani, A.S., Carson, K.A., Shong, Y.K., Kim, T.Y., Viola, D., Elisei, R., Bendlová, B., Yip, L., Mian, C., et al.: Association between braf v600e mutation and recurrence of papillary thyroid cancer. Journal of clinical oncology **33**(1), 42–50 (2015)
28. Xing, M., Alzahrani, A.S., Carson, K.A., Viola, D., Elisei, R., Bendlova, B., Yip, L., Mian, C., Vianello, F., Tuttle, R.M., et al.: Association between braf v600e mutation and mortality in patients with papillary thyroid cancer. Jama **309**(14), 1493–1501 (2013)
29. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)

30. Yu, Y., Zhao, C., Guo, R., Zhang, Y., Li, X., Liu, N., Lu, Y., Han, X., Tang, X., Mao, R., et al.: Deep learning model based on ultrasound images predicts braf v600e mutation in papillary thyroid carcinoma. IScience **28**(5) (2025)
31. Zhao, J., Liu, P., Yu, Y., Zhi, J., Zheng, X., Yu, J., Gao, M.: Comparison of diagnostic methods for the detection of a braf mutation in papillary thyroid cancer. Oncology Letters **17**(5), 4661–4666 (2019)
32. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 **1**(2), 1–124 (2023)