Dean&Francis
ISSN 2959-409X

Comparative Methods Review: Predicting Tumor Mutational Burden from Routine Pathology Slides

Ruilin Wen [1, *]

[1]Wooster School, Danbury, United States of America
*Corresponding author: rw3772@nyu.edu

Abstract:

Sequencing remains the reference standard for tumor mutational burden (TMB) but is costly, slow, and tissue intensive. Routine hematoxylin-and-eosin whole-slide images (WSIs) are inexpensive to digitize, motivating interest in whether AI can estimate TMB to support triage and prioritize sequencing. This review synthesizes more than thirty studies of TMB-from-WSI, standardizing task framing (primarily binary TMB-high versus TMB-low, with occasional regression) and evaluation practice (AUROC/AUPRC, external validation, calibration, and decision-curve analysis). Reported internal performance frequently falls around AUROC 0.70–0.82; independent-site external results are lower, approximately 0.65–0.73, yet directionally supportive. Multimodal fusion of H&E with basic clinical variables and the use of stronger representation self-supervised encoders and pathology foundation models-improve robustness, but performance remains sensitive to label definitions, class prevalence, tumor purity, and site/scanner domain shift. Reporting calibration quality, clinical net benefit, and subgroup analyses is inconsistent across studies. Overall, the current evidence supports TMB-from-WSI as a tool for triage and sequencing prioritization rather than a replacement for sequencing. This review recommends multicenter external validation, a minimal reporting set with a practical "benchmark card," and post-deployment monitoring of discrimination, calibration, and drift. Foundation and vision–language models with few-shot adapters are promising for cross-site transfer; prospective multicenter evaluations will be pivotal for clinical credibility.

Keywords: tumor mutational burden; whole-slide imaging; multiple-instance learning; self-supervised learning; foundation models

ISSN 2959-409X

1. Introduction

Molecular biomarkers such as tumor mutational burden (TMB) and microsatellite instability (MSI) increasingly guide immunotherapy selection and prognosis in solid tumors. However, comprehensive sequencing is often required to ascertain these biomarkers, and it carries practical constraints-cost, turnaround time, and tissue requirements-that limit access and delay decisions in routine care [1–3]. In contrast, hematoxylin-and-eosin (H&E) whole-slide images (WSIs) are produced for almost every patient at negligible incremental cost, motivating a complementary strategy: inferring molecular surrogates directly from routinely available morphology.

Over the past five years, deep learning on WSIs has progressed from weakly supervised multiple-instance learning (MIL) to hybrid supervision and, more recently, self-supervised and large-scale foundation models. Across lung, colorectal, and other cancers, internal test performance around AUROC 0.70–0.80 has been reported for predicting TMB or MSI from H&E slides [4–7]. Yet generalizability remains the central weakness: performance often degrades under domain shift introduced by staining protocols, scanners, or patient populations, and truly independent multicenter validation is still uncommon [8–10].

This review synthesizes evidence on predicting TMB directly from H&E WSIs and organizes the topic as follows. As shown in Figure 1, the end-to-end workflow comprises tiling and quality control, feature extraction (CNN/SSL/foundation), MIL aggregation, and a slide-level classifier/regressor that outputs predictions and attention/uncertainty maps. Section 2 formalizes the tasks and metrics (e.g., TMB-H vs TMB-L, AUROC/AUPRC, calibration, decision-curve analysis). Section 3 surveys methods by supervision and representation (MIL, hybrid/graph, SSL, foundation models). Section 4 consolidates per-cancer evidence with an emphasis on external, multicenter results. Sections 5–7 cover datasets/benchmarks, clinical translation, and open challenges.

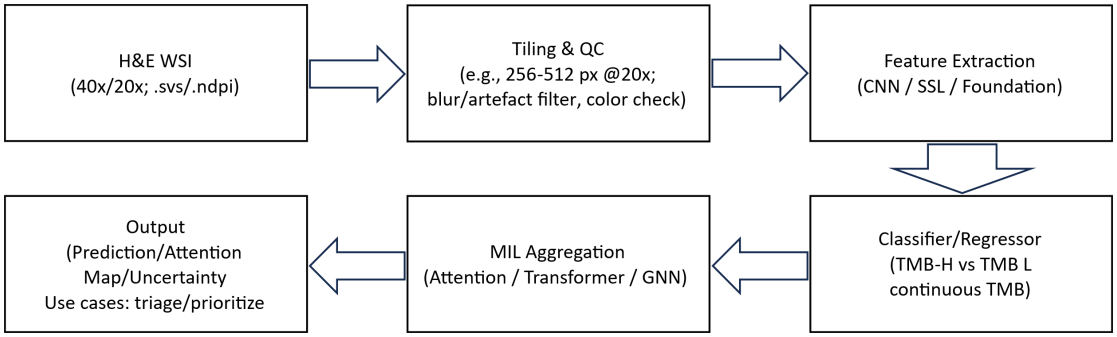

Fig 1. End-to-end workflow for WSI-based TMB prediction. WSIs are tiled and quality-checked; features are extracted by CNN/SSL/foundation encoders, tiles are aggregated via MIL (attention/transformer/GNN), and a slide-level head produces predictions and attention/uncertainty maps for triage/prioritization.

2. Tasks & Metrics

This section defines what is being predicted and how success is measured, so results in this section are comparable on a clinically meaningful scale. This review first specifies task scope and label rules, then the unit of analysis and split protocol, followed by primary metrics (AUROC/AUPRC), validation schemes (internal vs external), and clinical utility (calibration and decision-curve analysis).

2.1 Problem definitions and labels

This section defines the predictive tasks and the metrics by which model performance would be judged, so that results reported later can be interpreted on a common and clinically meaningful scale. In pathology-based biomarker inference, downstream tasks fall into four categories: (i) binary classification, (ii) multiclass classification, (iii) regression of continuous targets, and (iv) time-to-event prediction. Examples include classifying tumor mutational burden–high (TMB-H) versus tumor mutational burden–low (TMB-L), and microsatellite instability–high (MSI-H) versus microsatellite stable (MSS) in binary; mutations per mega base for regression; and overall/disease-free survival for time-to-event. For WSI models, labels are almost always defined at the patient or slide level, while predictions arise from tile-level features aggregated by MIL or related schemes. Ground truth for TMB typically comes from panel sequencing or whole-exome sequencing (WES) with heterogeneous counting rules and thresholds; for MSI, reference standards include PCR, IHC for MMR proteins, or NGS. These choices materially affect cross-study comparability and must be reported explicitly

[11–13].

2.2 Evaluation dimensions

Evaluation should answer three mutually independent questions: discrimination, calibration, and clinical utility. Discrimination assesses whether the model ranks cases correctly; calibration evaluates whether predicted probabilities are numerically meaningful; clinical utility examines whether using the model improves decisions for real patients [11–13].

2.3 Discrimination metrics

For binary or multiclass tasks, report AUROC with 95% confidence intervals (bootstrap or DeLong) [14-16]. Under class imbalance, include AUPRC, as it is more sensitive to prevalence than AUROC. Provide threshold-based summaries-sensitivity/recall, specificity, precision/PPV, NPV, F1-and confusion matrices at clinically relevant cutoffs (e.g., the literature threshold for TMB-high). For regression tasks, report R^2, MAE, and MSE, and correlate predictions with reference values; for survival, report Harrell's C-index and time-dependent AUROC [17].

2.4 Calibration and reliability

Because deployment decisions rely on probabilities, calibration should be evaluated using reliability plots, Brier score, and Expected Calibration Error (ECE). When miscalibration is detected, Platt scaling or isotonic regression on a held-out set should be considered, and calibration after threshold tuning should be re-checked. Poorly calibrated models can achieve high AUROC yet be unsafe for triage [18]. Figure 2 shows contents discussed in this section.

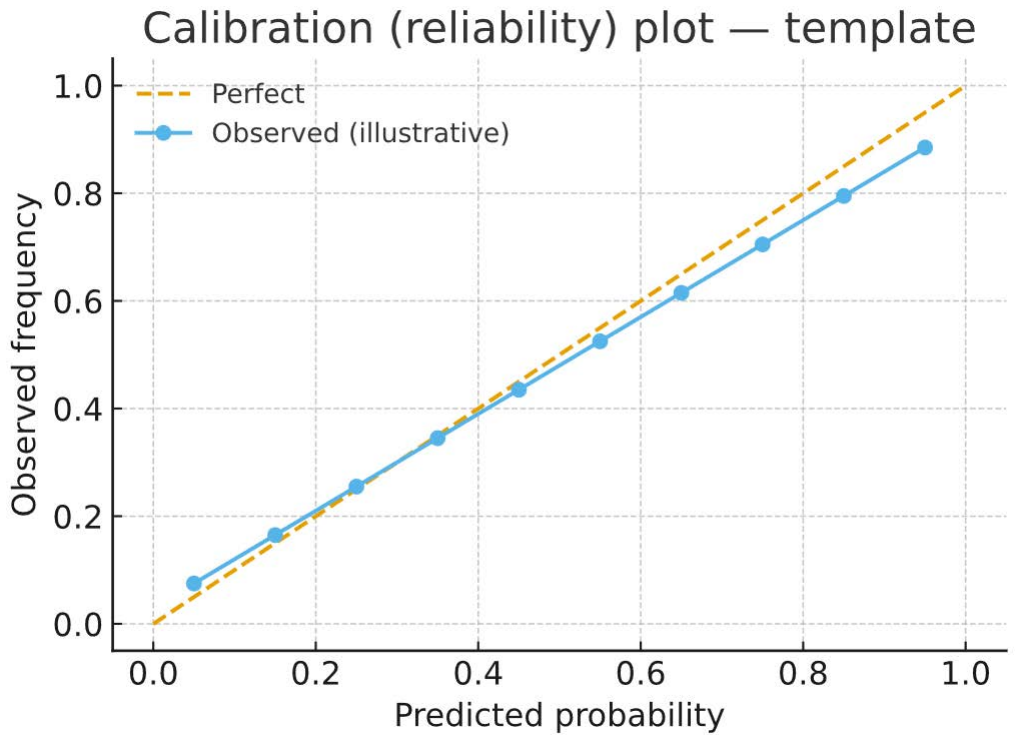

Figure 2. Reliability plot with perfect-calibration reference (dashed). Points are per-bin mean predictions vs observed event rates; binning and CIs are detailed in Section 2.4.

2.5 Clinical utility and decision analysis

Clinical utility is commonly summarized with decision-curve analysis (DCA), which plots net benefit across threshold probabilities and situates model use relative to "treat-all" and "treat-none" strategies. When applicable, studies also report reclassification metrics (e.g., Net Reclassification Improvement) and operational consequences, such as the proportion of sequencing that could be avoided at a prespecified miss rate for TMB-high. Publications usually specify the intended operating point and provide a brief clinical rationale for that choice [15]. Figure 3 shows contents discussed in this section.

ISSN 2959-409X

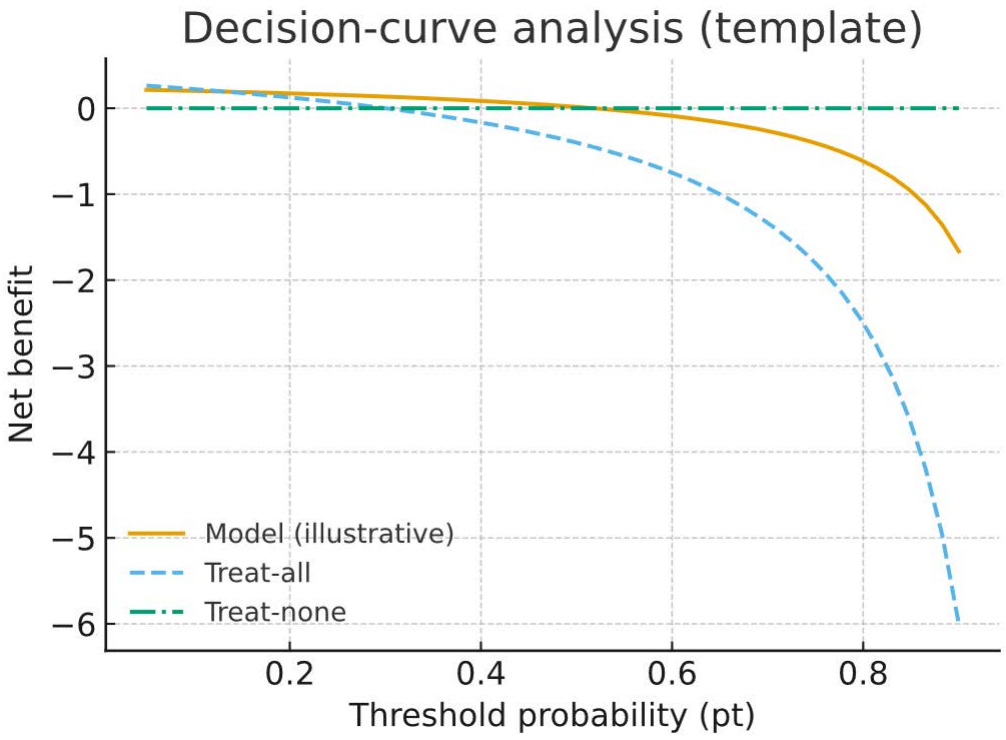

Figure 3. DCA comparing the model with treat-all and treat-none strategies across threshold probabilities; the vertical line marks the intended operating point (details in Section 2.5).

2.6 Validation design and leakage control

Validation designs in this area generally rely on patient-level splits, keeping all slides from the same patient within a single fold, and avoiding any mixing of tiles from one slide across training and test. Model development commonly uses stratified hold-out or nested cross-validation, while an untouched external validation cohort from an independent center/scanner/population is retained. Where feasible, multicenter evaluation is used to probe domain shift attributable to staining protocols, scanners, or case mix. In line with current reporting guidance, hyperparameters are not tuned on the external set, preprocessing is kept consistent across splits, and scanners, magnifications, staining protocols, and any color normalization procedures are documented [11–13].

2.7 Minimum reporting set (the checklist)

Major reporting frameworks (TRIPOD-AI, STARD-AI, SPIRIT-AI/CONSORT-AI, DECIDE-AI) converge on a core set of items for reproducible WSI-based biomarker studies. Table 1 provides a compact checklist tailored to TMB prediction from H&E WSIs. At minimum, studies describe the task and ground-truth rules, characterize the cohort and acquisition pipeline, specify preprocessing/tiling, document split strategy and leakage controls, report discrimination, calibration, and decision analysis with uncertainty, present external/multicenter results and subgroup analyses, and provide artifacts for reproducibility (code/weights/tiling manifest) [11–14,19].

Table 1. Minimum reporting set for WSI-based TMB prediction (aligned with TRIPOD-AI/STARD-AI/DECIDE-AI).

Field	What to report
Task type & label rules	Primary task (binary TMB-H vs TMB-L or regression) and label source/cutoffs; assay/panel noted
Class prevalence & dataset sizes	Prevalence of TMB-H (%); N per split at patient/slide/tile levels
Scanners, staining, magnifications, color normalization	Vendors/models, staining protocols, objective magnifications, color normalization method

Discrimination Calibration Clinical utility	AUROC (±95% CI), AUPRC; confusion matrices at chosen cutoffs Brier/ECE and a reliability plot (slope/intercept if available) Decision-curve analysis; chosen operating point with PPV/NPV/Sens/Spec and rationale
External/multicenter & subgroups	Independent site/multicenter results; subgroup analyses (site/scanner/stage/ancestry/sex)
Model availability & reproducibility	Code/weights availability; random seeds; versioned data/commit hashes

3. Methods

Building on the task framing in Section 2, this section organizes model design choices by supervision level and representation learning, highlighting how each family trades off label cost, robustness, and interpretability.

3.1 Weakly supervised multiple-instance learning (MIL) and attention

Whole-slide images are partitioned into tiles ("instances") and grouped into slide-level bags; models learn to map bag-level labels from instance-level features without region annotations. Attention-based pooling and transformer-style aggregators (e.g., CLAM-like attention pooling or TransMIL-like architectures) highlight informative tiles while suppressing noise [9,10]. Typical pipelines extract tile embeddings with a CNN/ViT backbone (either ImageNet-pretrained or WSI-pretrained), then train a slide-level aggregator with cross-entropy or focal losses. Multi-scale tiling (e.g., 5×/10×/20×) and hard-example mining are frequently used to capture morphology at different resolutions and reduce label noise. Taken together, these design choices yield a pragmatic trade-off: the model can be trained at scale using only slide-level labels, yet it still produces tile-level attention maps that pathologists can review. This makes MIL pipelines attractive for retrospective cohorts and workflows where region annotations are scarce. At the same time, the reliance on weak labels and patch sampling can make performance brittle under stain/scanner or population shifts, and attention may occasionally concentrate on non-causal correlates ("saliency illusions"), underscoring the need for external checks [9,10].

3.2 Hybrid supervision, label-noise mitigation, and graph-structured MIL

Hybrid supervision combines slide-level labels with a small number of region-level annotations or high-confidence pseudo-labels to refine localization and reduce label noise. Typical designs include partial-label learning, consistency regularization, and curriculum schedules that progressively trust finer-grained supervision while retaining slide-level constraints. Graph-structured MIL treats tiles as nodes connected by spatial proximity or tissue-type similarity, enabling message passing that captures glandular architecture, stromal context, and tumor–immune interfaces [20]. In practice, hybrid supervision often improves robustness and interpretability at the cost of additional annotation effort and engineering complexity; pseudo-label pipelines can also propagate early mistakes if not carefully curated. Graph-based approaches frequently pair a pretrained tile encoder with a graph neural network (GNN) or transformer-style graph aggregator, yielding competitive slide-level performance and saliency maps that better align with tissue structures [20].

3.3 Self-supervised pretraining and retrieval/representation learning

Self-supervised learning (SSL) pretrains encoders on unlabeled WSIs using objectives such as contrastive alignment or masked-region reconstruction, then fine-tunes on downstream tasks [21]. Compared with training from scratch, SSL backbones generally yield more stable features under small data regimes and improved cross-site generalization [21]. Instance retrieval systems (e.g., SISH-style nearest-neighbor search) leverage SSL embeddings to find morphologically similar regions, supporting both case review and weakly supervised labeling [21]. In practice, however, realizing these benefits introduces operational trade-offs: the same design choices that enable powerful representations-strong augmentations, large pretraining corpora, and domain-specific pipelines-also determine stability, cost, and bias characteristics. Common pitfalls include collapse without sufficiently strong augmentations, heavy computation during pretraining, and augmentation choices that inadvertently encode dataset or site-specific bias [21].

3.4 Foundation and vision–language models

Foundation models are trained at scale (often millions of tiles across diverse centers) to produce reusable pathology encoders that transfer across tasks and institutions [22–24].

ISSN 2959-409X

For WSI-level inference, a common recipe freezes the pretrained encoder and learns a lightweight adapter or MIL aggregator; few-shot or parameter-efficient fine-tuning is also used [22–24]. Vision–language variants pair images with pathology reports to align visual features with clinical semantics, enabling zero-/few-shot transfer and text-conditioned outputs [22–24]. Evidence to date suggests stronger cross-hospital robustness than task-specific models, although gains depend on pretraining diversity and governance of data overlap, privacy, and model updates [22–24].

4. Applications

With the methodological landscape in view, this section consolidates evidence on predicting tumor mutational burden (TMB), and, secondarily, microsatellite instability (MSI), from routine H&E whole-slide images across major cancer types, emphasizing external and multicenter validity; it highlights task setups, representative performance ranges, sources of heterogeneity, and practical considerations for clinical translation.

4.1 Tumor Mutational Burden (TMB)

Early studies explored direct TMB prediction from lung adenocarcinoma H&E slides using CNN-style pipelines, followed by MIL-based approaches and multi-scale designs [5,6,8]. Across internal test sets, reported AUROC typically falls in the 0.70–0.80 range, with performance depending on cohort composition, tumor purity, and patch sampling strategies [5–8]. External validation remains limited but is gradually increasing, including multicenter analyses in squamous cell carcinoma settings [7]. Comparability across studies is affected by heterogeneous ground-truth definitions (e.g., panel vs. WES, counting rules) and threshold choices for defining "TMB-high"[5–8]. Practical takeaways include reporting prevalence, clarifying label sources, and probing robustness under site/scanner changes before considering triage use [5–8].

4.2 Microsatellite Instability (MSI)

Seminal work demonstrated the feasibility of predicting MSI directly from routine H&E slides, and subsequent studies expanded to multicenter validations across gastrointestinal cancers [1–3]. Weakly supervised frameworks further explored molecular pathways and mutation patterns in colorectal cancer, reinforcing the link between morphology and genotype [4]. Performance varies with cancer subtype and population; prevalence differences and site-specific workflows can introduce domain shift, underscoring the need for external and cross-population testing [1–3]. Attention maps in several studies align with histopathologic hallmarks implicated in mismatch-repair deficiency, supporting face validity while not proving causality [1–4].

4.3 Radiomics as comparator

Radiomics seeks to predict immuno-oncology–relevant signals (including TMB or response surrogates) from CT/PET images, offering whole-organ coverage without tissue sampling. Compared with pathology-based models that capture micro-architectural cues at cellular scales, radiomics emphasizes macroscopic heterogeneity; the two modalities are therefore complementary rather than interchangeable. In workflows, one pragmatic view is to use radiology to flag candidates for biopsy and WSI acquisition, then use pathology-AI to prioritize sequencing when tissue is limited.

5. Datasets & Benchmarks

To interpret metrics consistently, this section details datasets and evaluation setups and introduces a practical benchmark card for evidence extraction. Typical WSIs are FFPE H&E slides scanned at 20× or 40×, with heterogeneous formats and scanners; slide-level labels are patient- or block-derived and may not localize tumor regions.

5.1 Landscape and dataset types

Public research datasets and institution-specific clinical repositories serve different purposes and exhibit different biases. Public sets enable comparability and ablation studies, while clinical repositories better reflect workflow constraints, slide variability, and case mix seen in practice. Common public sources include multi-cancer archives and challenge datasets; this review uses them primarily for methods development and sanity checks, not as surrogates for deployment evidence [11–13].

5.2 Preprocessing and tiling conventions

Pipelines commonly apply tissue detection, color normalization, and tiling into fixed-size patches (e.g., 256–512 px at a defined magnification). Magnification must be reported alongside the physical resolution (μm/px), since "20×" is not standardized across scanners. When multiple magnifications are used, document how tiles are sampled and fused (multi-scale MIL, pyramids) to ensure reproducibility.

5.3 Splits, external validation, and leakage con-

trol

Across published benchmarks, splits are typically defined at the patient level, thereby avoiding cross-contamination of slides or tiles; preprocessing pipelines are kept identical across splits to prevent drift. External validation is usually performed on a sealed cohort from an independent site/scanner/population, and multicenter testing is often included to probe domain shift [11–13,19].

Consistent with current reporting guidance, thresholds and hyperparameters are not tuned on the external set; instead, studies document scanners, staining protocols, magnifications (with μm/px), and color normalization as part of a concise "benchmark card" to support reproducibility [11–13,19].

5.4 Current benchmarks' limitations

Most public splits lack sealed, no-peek external sites and rarely stress-test scanner/stain shifts or prevalence changes. Few benchmarks report calibration, decision-curve analysis, or cost–benefit under realistic operating points. Subgroup performance by site, scanner, stage, and demographics is inconsistently reported, complicating fairness and generalizability claims.

5.5 A practical "benchmark card" (minimal fields)

In line with current reporting guidance, this review summarizes a compact benchmark card in Table 2. The card captures dataset scope, ground-truth rules and thresholds, cohort composition and acquisition, preprocessing/tiling, split strategy and leakage controls, core metrics (discrimination with uncertainty, calibration, and decision analysis), external/multicenter results with subgroup analyses, and basic reproducibility artifacts. Publishing code/weights together with a versioned slide manifest enables exact reproduction of tiling and sampling [11–13,19].

Table 2. Benchmark card for WSI-based TMB prediction (minimal fields aligned with TRIPOD-AI/STARD-AI/DECIDE-AI) [11,14,19].

Study	Cancer	Task	Data	External?	AUROC (int/ext)	Method
Sadhwani (2021) [5]	LUAD	TMB-H/L	TCGA LUAD	Partial	0.71	Histological-subtype features + clinical
Jain (2020) [6]	LUAD	TMB-H/L	TCGA LUAD	No	/	CNN (Inception v3) + Random Forest
Dammak (2023) [7]	LUSC	TMB-H/L	multicenter	Yes	0.65	Transfer learning CNN (VGG16)
Chen (2022) [8]	Multiple	TMB-H/L	TCGA, CPTAC	Yes	0.818 (CV) / 0.732	Multi-scale weak supervision + graph aggregation (MC-TMB)
Chen (2022) [21]	Multiple	WSI retrieval	TCGA, CPTAC, BWH	Yes	/	Self-supervised retrievals (SISH; VQ-VAE + DenseNet)

6. Clinical Translation & Reporting

Translating reported performance into practice requires governance beyond AUC. In line with reporting and evaluation guidance, clinical claims rest on transparent methods and fit-for-purpose designs: TRIPOD-AI for prediction models, STARD-AI for diagnostic accuracy, SPIRIT-AI/CONSORT-AI for interventional trials and DECIDE-AI for early clinical evaluation [11–14,19]. Typical elements include prespecified endpoints, model lock before any external testing, documented data provenance, and enough procedural detail for replication. Evidence usually follows a validation ladder-internal validation first, then sealed external-site testing, and, where feasible, multicenter prospective studies [11–14,19]. Practical deployment tends to keep a pathologist in the loop, define operating thresholds and reflex-to-sequencing rules, and integrate with LIS/EHR; calibration maintenance and drift monitoring are scheduled, with audit logs and rollback procedures when thresholds or models change [11,18]. Decision-curve analysis (DCA) complements discrimination metrics by summarizing net benefit across thresholds; reporting operational consequences (e.g., the proportion of sequencing avoided at a fixed miss rate) clarifies value [15]. Governance spans data-use agreements, privacy

ISSN 2959-409X

protection, change control for model updates, and fairness auditing with subgroup reporting by site, scanner, stage, and demographics [11–14,19].

7. Discussion

This section synthesizes the main failure modes that separate promising discrimination from deployable clinical utility-domain shift, data/label quality, interpretability, reporting gaps, and post-deployment maintenance-and argues for a pragmatic path forward grounded in externally validated evidence and standardized reporting.

Performance drops under domain shift-differences in staining, scanners, and case mix-remain the primary risk; larger foundation models coupled with diverse pretraining, and few-shot adapters improve cross-site transfer but do not eliminate the problem [22–24]. Ceiling performance is further constrained by tumor purity, slide quality, and label noise; hybrids that add limited region cues or graph structure can help localize signal and reduce propagation of noise [4, 20]. Interpretability via attention or saliency maps enables slide review yet can surface non-causal correlates, reinforcing the need for human-in-the-loop verification [9,10]. Beyond discrimination, reproducibility and reporting are uneven-calibration, decision-curve analysis, and subgroup reporting are inconsistently presented-hence alignment with AI reporting guidance (e.g., TRIPOD-AI, SPIRIT-AI/CONSORT-AI, DECIDE-AI) is essential for credible claims and fair comparison [11–14,19]. For maintenance, clinical deployment should specify auditable update policies and post-deployment monitoring of discrimination, calibration, and case-mix drift; regulatory-grade practice further requires multicenter external testing and laboratory validation frameworks [11–14,19]. Looking ahead, foundation/vision–language pipelines plus standardized external test sets and early-phase multicenter evaluations offer the most credible route to safe triage/priority sequencing use in the near term [22–24].

8. Conclusion

H&E WSI–based models can infer molecular surrogates such as TMB and MSI with moderate discriminatory performance while leveraging ubiquitous, low-cost pathology data. Evidence to date supports their role as triage or prioritization tools rather than full replacements for sequencing, given domain shift, label heterogeneity, and calibration concerns. Progress will hinge on diversified pretraining, sealed external and multicenter validation, rigorous reporting (TRIPOD-AI/STARD-AI), and governance for updates and fairness.

References

[1] Kather, J. N., Pearson, A. T., Halama, N., et al. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Medicine, 25(7), 1054–1056. https://doi.org/10.1038/s41591-019-0462-y

[2] Yamashita, R., Long, J., Longacre, T. A., et al. (2021). Deep learning model to predict microsatellite instability from routine histology of colorectal cancer: Multicentre validation. The Lancet Oncology, 22(1), 132–141. https://doi.org/10.1016/S1470-2045(20)30535-0

[3] Echle, A., Grabsch, H. I., Quirke, P., et al. (2020). Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning of histology images. Gastroenterology, 159(4), 1406–1416.e11. https://doi.org/10.1053/j.gastro.2020.06.021

[4] Bilal, M., Raza, S. E. A., Azam, A., et al. (2021). Weakly supervised prediction of molecular pathways and key mutations in colorectal cancer from routine histology images: A retrospective study. The Lancet Digital Health, 3(12), e902–e912. https://doi.org/10.1016/S2589-7500(21)00180-1

[5] Sadhwani, A., Chang, H.-W., Behrooz, A., et al. (2021). Comparative analysis of machine learning approaches to classify tumor mutation burden in lung adenocarcinoma using histopathology images. Scientific Reports, 11, 16605. https://doi.org/10.1038/s41598-021-95747-4

[6] Jain, M. S., & Massoud, T. F. (2020). Predicting tumour mutational burden from histopathological images using multiscale deep learning. Nature Machine Intelligence, 2(6), 356–362. https://doi.org/10.1038/s42256-020-0190-5

[7] Dammak, S., Cecchini, M. J., Breadner, D., & Ward, A. D. (2023). Using deep learning to predict tumor mutational burden from multicenter H&E whole-slide images of lung squamous cell carcinoma. Journal of Medical Imaging, 10(1), 017502. https://doi.org/10.1117/1.JMI.10.1.017502

[8] Chen, S., Xiang, J., Wang, X., et al. (2022). Deep learning-based approach to reveal tumor mutational burden status from whole slide images across multiple cancer types. arXiv preprint arXiv:2204.03257. (No DOI)

[9] Lu, M. Y., Williamson, D. F. K., Chen, T. Y., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering, 5(6), 555–570. https://doi.org/10.1038/s41551-020-00682-w

[10] Shao, Z., Bian, H., Chen, Y., et al. (2021). TransMIL: Transformer-based multiple instance learning for whole slide image classification. In Advances in Neural Information Processing Systems (NeurIPS 34), 2136–2148. (No DOI) arXiv:2106.00908

[11] TRIPOD-AI Collaboration. (2024). Transparent reporting of a multivariable prediction model for AI/ML-based diagnosis and prognosis (TRIPOD-AI). BMJ, 386, e078378. https://doi.org/10.1136/bmj-2023-078378

[12] SPIRIT-AI Steering Group. (2020). SPIRIT-AI extension: Guidelines for clinical trial protocols involving artificial intelligence interventions. BMJ, 370, m3210. https://doi.org/10.1136/bmj.m3210

[13] CONSORT-AI Steering Group. (2020). CONSORT-AI extension: Reporting of clinical trials involving artificial intelligence interventions. BMJ, 370, m3164. https://doi.org/10.1136/bmj.m3164

[14] STARD-AI Steering Group. (2023). Reporting diagnostic accuracy studies that use AI: The STARD-AI protocol. BMJ Open, 13, e047709. https://doi.org/10.1136/bmjopen-2023-047709

[15] Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making, 26(6), 565–574. https://doi.org/10.1177/0272989X06295361

[16] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated ROC curves: A nonparametric approach. Biometrics, 44(3), 837–845. https://doi.org/10.2307/2531595

[17] Harrell, F. E. (2015). Regression Modeling Strategies (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-19425-7

[18] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017), PMLR 70, 1321–1330. (No DOI) http://proceedings.mlr.press/v70/guo17a.html

[19] DECIDE-AI Steering Group. (2022). DECIDE-AI: Reporting guidelines for early-stage clinical evaluation of AI-based decision support systems. Nature Medicine, 28, 924–933. https://doi.org/10.1038/s41591-022-01772-9

[20] Lu, W., Toss, M., Dawood, M., Rakha, E., Rajpoot, N., & Minhas, F. (2022). SlideGraph+: Whole-slide image-level graphs to predict HER2 status in breast cancer. Medical Image Analysis, 80, 102486. https://doi.org/10.1016/j.media.2022.102486

[21] Chen, C., Lu, M. Y., Williamson, D. F. K., et al. (2022). Self-supervised instance-level retrieval for computational pathology. Nature Biomedical Engineering, 6, 1420–1434. https://doi.org/10.1038/s41551-022-00929-8

[22] Xu, H., Usuyama, N., Bagga, J., et al. (2024). A whole-slide foundation model for digital pathology from real-world data. Nature, 629, 358–365. https://doi.org/10.1038/s41586-024-07441-w

[23] Wang, X., Yang, S., Zhang, J., et al. (2024). A pathology foundation model for cancer diagnosis and prognosis. Nature, 630, 131–138. https://doi.org/10.1038/s41586-024-07894-z

[24] Chen, R.-J., Ding, T., Williamson, D. F. K., et al. (2024). A generalist pathology foundation model. Nature Medicine, 30, 1703–1713. https://doi.org/10.1038/s41591-024-02857-3