# LEARNING FROM UGLY DUCKLINGS: HIERARCHICAL MODELING OF INTRA-PATIENT VARIANCE FOR MELANOMA PREDICTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Medical AI shows great potential to improve the efficiency and accuracy of diagnosis from medical images. However, disease manifestations vary across individuals, such that a similar observation may be typical for one patient but concerning for another. Most medical image classifiers assume independent and identically distributed (IID) observations and therefore do not account for intrapatient variation. In melanoma detection, dermatologists account for this variation by assessing both a lesion's appearance and whether it stands out from the patient's other lesions, a pattern known as the ugly-duckling principle. We model this setting with a hierarchical model that estimates a patient reference from unlabeled same-patient lesions and derives how deviations from this reference modify lesion-level evidence. We introduce EIDER (Evidence In Deviations from an Estimated Reference), which estimates this reference and learns how lesion-to-reference deviations adjust lesion-level predictions. We evaluate EIDER across three dermatology settings. On total-body photography, EIDER achieves an AUROC of 0.959, compared with 0.921 for our strongest image-only baseline. Controlled analyses show that replacing a patient's reference with that of another patient decreases performance, larger reference sets are associated with greater gains over lesion-only prediction, and combining signed and squared deviations performs better than either restricted deviation form; we further evaluate EIDER in dermoscopic and longitudinal settings, and assess its applicability to EEG seizure detection.

## 1 INTRODUCTION

Standard supervised learning typically treats observations as independent and identically distributed (IID) samples from a common population distribution (Wu et al., 2022). Medical imaging data often have a hierarchical structure: patients differ from one another, while multiple observations from the same patient are correlated (Thrall et al., 2021; Moskowitz et al., 2022; Lee et al., 2022). These inter-patient differences and within-patient correlations create information that is not represented when each observation is treated separately.

Melanoma detection provides a clear example. Although nevi vary substantially across individuals, each patient tends to exhibit a limited number of characteristic nevus patterns, described as an "individual reference system" (Wazaefi et al., 2013; Gaudy-Marqueste et al., 2017). Dermatologists use this reference through the well-established "ugly duckling" principle: lesions that depart from the patient's usual patterns are considered more suspicious and warrant closer inspection (Grob & Bonerandi, 1998). Gaudy-Marqueste et al. (2017) distinguish this *intrapatient comparative analysis* from *lesion-focused analysis*, in which a lesion is assessed from its own morphology (Fig. 1a). Clinical assessment can combine both forms of analysis.

Total-body photography (TBP) captures large numbers of lesions from the same patient, making the patient's other lesions available when classifying a target lesion (Rayner et al., 2018; Ji-Xu et al., 2021; Kurtansky et al., 2025). Most automated melanoma classifiers nevertheless predict from the target lesion alone, optionally together with metadata associated with that lesion (Esteva et al., 2017; Haenssle et al., 2018; Brinker et al., 2019; Tschandl et al., 2020). Methods that use the patient's other lesions have compared them with the target lesion through distance or outlier measures (Soenksen



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

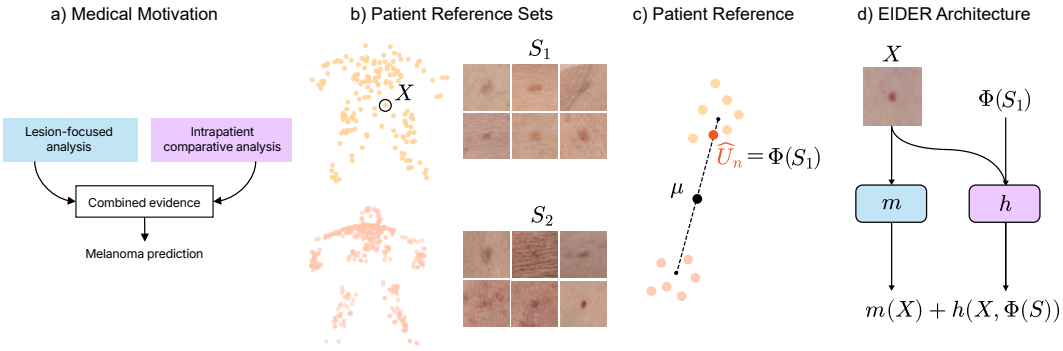


Figure 1: **Overview of prediction with patient reference sets and EIDER.** (a) Clinical motivation for combining lesion-focused and intrapatient evidence. Adapted from Gaudy-Marqueste et al. (2017), who distinguish the two when selecting lesions for biopsy. (b) We study prediction for a target lesion $X$ given an unlabeled reference set $S$ of other lesions from the same patient. (c) The reference set is summarized by an estimated patient reference $\widehat{U}_n = \Phi(S)$, partially pooled toward the population reference $\mu$. (d) EIDER combines a target-only term $m(X)$ with a reference term $h(X, \Phi(S))$ to produce the final prediction.

et al., 2021; Mohseni et al., 2021; Mendes & Barata, 2023), or jointly represented multiple lesions before prediction (Yu et al., 2021). These methods differ in two basic choices: how to summarize the patient's other lesions, and how to use that summary when predicting the target lesion. We study these two choices directly.

Building on intrapatient comparative analysis, we formalize the prediction problem of classifying a target observation using other unlabeled observations from the same patient. In melanoma, these other lesions form a reference set from which a patient reference can be estimated. We use a hierarchical model to derive how the patient reference depends on the number of available reference lesions and how the target lesion's deviation from this reference changes its predicted malignancy. The derivation further shows that this comparison cannot in general be reduced to the magnitude of the deviation: deviations in different directions can change the prediction differently.

Guided by this result, we introduce EIDER (**E**vidence **I**n **D**eviations from an **E**stimated **R**eference). EIDER estimates a patient reference from the reference lesions and learns how the target lesion's deviation from that reference changes its malignancy prediction. The final prediction combines the target lesion's lesion-focused prediction with this learned adjustment.

We evaluate EIDER across three dermatology settings and use controlled interventions to test its use of the patient reference. On total-body photography, EIDER achieves an AUROC of $0.959$, compared with $0.921$ for our strongest image-only baseline. Replacing a patient's reference with one estimated from another patient decreases performance as expected. Increasing the number of reference lesions increases the adjustment to the target-lesion prediction, and preserving the direction of the deviation is more informative than using its magnitude alone. We also evaluate transfer to dermoscopic and longitudinal settings. Finally, we evaluate the same prediction setting for EEG seizure detection, where repeated observations from the same patient are also available.

Our contributions are:

1. We formalize prediction of a target observation using an unlabeled set of other observations from the same patient, and derive a hierarchical model in which these observations estimate a patient reference.
2. We introduce EIDER, which estimates a patient reference from other lesions and learns how the target lesion's deviation from this reference changes its malignancy prediction.
3. We evaluate EIDER across total-body, dermoscopic, and longitudinal melanoma settings, use controlled reference interventions to test its dependence on the reference lesions, and evaluate the same prediction setting for EEG seizure detection.

# 2 Related Work

## 2.1 Melanoma Prediction

Melanoma detection has traditionally relied on lesion-focused criteria. Clinical and dermoscopic rules such as pattern analysis (Argenziano et al., 2003), the ABCD rule (Stolz et al., 1994), the Menzies method (Menzies, 2002), and the 7-point checklist (Argenziano et al., 1998) assess properties such as asymmetry, color, border, and dermoscopic structures. Automated methods replace these criteria with learned lesion representations (Esteva et al., 2017; Haenssle et al., 2018; Brinker et al., 2019; Yan et al., 2025b). These representations provide the basis for comparing a query with the patient's other lesions.

Some methods reduce this comparison to predefined intrapatient quantities. Soenksen et al. (2021) derive an intrapatient saliency score, related approaches identify ugly ducklings through distances in learned embedding spaces (Mohseni et al., 2021; Useini et al., 2024), and Mendes & Barata (2023) use distance and reconstruction-error statistics. Total Body Photography (TBP) systems similarly use patient-normalized measurements, lesion summaries, and clustering (Kurtansky et al., 2025). These approaches therefore assume in advance which aspects of the patient reference are useful for prediction.

Other work learns patient relationships from data rather than fixing the comparison statistics. Yu et al. (2021) jointly model same-patient lesions with a Transformer, allowing the prediction of each lesion to depend on the others. Patient-specific metric learning instead uses same-patient relationships during training to learn a representation in which ugly ducklings are easier to isolate (Naranpanawa et al., 2024). These approaches allow richer intrapatient representations, but do not establish what information from the reference must be preserved for malignancy prediction.

## 2.2 Set-Conditioned Prediction

Unlabeled observations at prediction time also appear in related settings. Domain and test-time adaptation use them to estimate distribution shift (Sun et al., 2016; Chang et al., 2019; Roy et al., 2019; Courty et al., 2017; Wang et al., 2021), while contextual anomaly detection uses them to define what is unusual (Breunig et al., 2000; Carmona et al., 2021; Li & van Leeuwen, 2023). Adaptive Risk Minimization similarly conditions predictions on unlabeled test batches (Zhang et al., 2021). Our setting instead uses other observations from the same group as the query to predict an externally defined label for that query. We compare these formulations in Appendix A.2.

Mixed-effects models address group variation by estimating a low-dimensional group effect with partial pooling (Laird & Ware, 1982; Robinson, 1991), with recent extensions to nonlinear predictors (Simchoni & Rosset, 2023; Mandel et al., 2023). Set models make fewer assumptions about the form of this group information and learn permutation-invariant representations through aggregation (Zaheer et al., 2017) or attention (Lee et al., 2019). Both provide mechanisms for incorporating group information, but neither specifies what the resulting representation must preserve for the query prediction.

Latent-variable methods make the shared group state explicit. Neural Statistician infers latent properties from an unlabeled collection (Edwards & Storkey, 2017), Bayesian Context Aggregation combines observations through inference over a shared latent variable (Volpp et al., 2021), and Bayesian meta-learning infers task-specific states from support examples (Ravi & Beatson, 2019; Gordon et al., 2019). Distribution regression similarly represents a finite sample through properties of its underlying distribution (Szabó et al., 2016). These works motivate estimating shared group information from a reference set; our focus is which part of that information must be retained to predict an individual query.

# 3 Prediction with Patient References

We study how an unlabeled reference set from the same patient can inform prediction for a target observation, and what information from this set should be retained. We first define predictive sufficiency for a reference representation, then use Bayes' rule to decompose prediction into a target-only term and a reference term. We then introduce a hierarchical model that derives a predictively suffi-

cient patient reference representation from the reference set. Finally, we characterize how deviations of the target observation from this reference contribute to prediction.

**Definition 3.1** (Prediction setting)**.** Let $X$ denote a target observation with binary label $Y$, and let $S$ be the unlabeled reference set containing the other observations from the same patient (see Figure 1b). At test time, $X$ and $S$ are observed, and the prediction target is $P(Y = 1 \mid X, S)$.

### 3.1 Reference Representation

A reference representation should retain all information in $S$ relevant to predicting $Y$ once $X$ is known. We formalize this requirement through predictive sufficiency.

**Definition 3.2** (Predictively sufficient statistic)**.** A reference representation $\Phi(S)$ is predictively sufficient for $S$ given $X$ if

$$Y \perp S \mid (X, \Phi(S)).$$

Equivalently, $P(Y = 1 \mid X, S) = P(Y = 1 \mid X, \Phi(S))$ almost surely.

Thus, a predictively sufficient representation retains exactly the information from the reference set needed for prediction.

### 3.2 Prediction Decomposition

Predictive sufficiency specifies what the reference representation should retain. We next characterize how this information contributes to the prediction.

**Lemma 3.3** (Bayes decomposition)**.** *For binary $Y$, Bayes' rule gives*

$$\operatorname{logit} P(Y = 1 \mid X, S) = \operatorname{logit} P(Y = 1 \mid X) + \log \frac{p(S \mid X, Y = 1)}{p(S \mid X, Y = 0)}. \tag{1}$$

The first term depends only on the target observation $X$, while the second captures the additional contribution of the reference set $S$. If $\Phi(S)$ is predictively sufficient, then $S$ can be replaced by $\Phi(S)$ without changing the prediction. This motivates predictors of the form

$$\widehat{\ell}(X, S) = m(X) + h(X, \Phi(S)), \tag{2}$$

where $m(X)$ is the target-only term and $h(X, \Phi(S))$ is the reference term.

Equation 2 identifies three choices for models using a reference set: the target-only term $m$, the reference representation $\Phi$, and the reference term $h$. Appendix A.3 uses these choices to characterize prior melanoma models. The decomposition leaves the choice of $\Phi(S)$ unspecified; we next derive one from a hierarchical model of the patient reference.

### 3.3 Hierarchical Model of Patient Reference

We model a patient's lesions as repeated observations around a latent patient reference. A Gaussian random-effects model (Laird & Ware, 1982; Robinson, 1991) combines population and patient information to estimate this reference while accounting for the number of observed reference lesions. The notation, assumptions, formal results, and proofs are given in Appendix B.3.

**Theorem 3.4** (Patient reference and target deviation (informal))**.** *For each feature $k$, the reference set yields the partially pooled patient reference estimate*

$$\widehat{U}_{n,k} = \frac{n}{n + \kappa_k} \overline{Z}_k + \frac{\kappa_k}{n + \kappa_k} \mu_k, \qquad \kappa_k = \frac{\sigma^2_{c,k}}{\tau^2_k}, \tag{3}$$

*where $\overline{Z}_k$ is the mean of feature $k$ across the $n$ reference lesions, $\mu_k$ is its population mean, and $\sigma^2_{c,k}$ and $\tau^2_k$ describe within- and between-patient variation. Figure 1c shows a toy example with two patients.*

*Conditional on the latent patient reference $U$, let $r = Z_X - U$ denote the target deviation. Then*

$$\operatorname{logit} P(Y = 1 \mid Z_X, U) = b(U) + c + r^\top \alpha r + \beta^\top r \tag{4}$$

Equation 3 shows that as the reference size $n$ increases, $\widehat{U}_n$ places more weight on the patient's reference lesions, while its posterior uncertainty decreases. Proposition B.4 further shows that $\Phi(S) = (\widehat{U}_n, n)$ is predictively sufficient. Thus, under the model, the reference representation consists of the estimated patient reference and the reference size.

Equation 4 shows that prediction can depend on both signed and squared deviations from the patient reference. EIDER therefore uses $\widehat{U}_n$, $n$, and the signed and squared estimated deviations $\widehat{r} = Z_X - \widehat{U}_n$ to learn the reference term. Appendix C evaluates the assumptions and predictions of the hierarchical model.

### 3.4 EIDER

EIDER (Evidence In Deviations from an Estimated Reference) follows the decomposition in Section 3.2 by combining a target-only prediction with an additive reference term. Figure 1d shows the target and reference branches and their additive fusion.

Given a target observation $X$ and reference set $S$, EIDER first estimates the patient reference $\widehat{U}_n$ from $S$ using the partially pooled estimator derived in Section 3. It then represents the target through its signed and squared deviations from this reference, together with the reference estimate and reference size. A reference branch maps these quantities to the reference term $h(X, \Phi(S))$. A separate target branch predicts from $X$ alone, and the reference term is added to its logit as defined by the preceding decomposition. The target observation is excluded from its reference set.

We train the target branch first and then freeze it while training the reference branch on the combined prediction. This keeps the target-only predictor fixed and learns only the adjustment contributed by the reference set. Architecture and dataset-specific implementation details are given in Section 4.1 and Appendix D.

## 4 Melanoma Detection

Melanoma prediction provides a natural setting for prediction with a reference set because patients may have multiple lesions, and clinical assessment considers whether a target lesion differs from the patient's other lesions (Grob & Bonerandi, 1998). We first evaluate EIDER against target-only models and prior methods using reference lesions. We then examine the mechanisms predicted by the theory through patient reference estimation and signed and squared deviations. Finally, we evaluate the same formulation in dermoscopic and longitudinal settings with different reference-set structures.

### 4.1 Experimental Setup

**Datasets.** We evaluate EIDER on three melanoma datasets using the patient's other lesions as its reference set. ISIC 2024 (SLICE-3D) (Kurtansky et al., 2025) contains 401,059 TBP lesion crops from 1,042 patients, including 393 malignant lesions, with accompanying lesion and patient metadata. SIIM–ISIC 2020 (Rotemberg et al., 2021) contains 33,126 dermoscopic images from 2,056 patients, including 584 melanomas. The longitudinal cohort (Ghahari et al., 2025) contains 35,355 observations of 9,311 lesions from 458 patients. For this cohort, the reference set contains only lesions observed during that visit, so the evaluation follows the same formulation as the other two datasets rather than modeling longitudinal change. Appendix D gives dataset, reference-set construction, and preprocessing details.

**Metrics.** Following Kurtansky et al. (2025), we report two primary metrics. Because lesion triage prioritizes high sensitivity, our main metric is the partial ROC area above 80% sensitivity ($\mathrm{pAUC}_{80}$; maximum $0.2$). We also report patient recall among the top 15 ranked lesions ($\mathrm{PR}_{15}$), defined as the fraction of patients with a positive lesion for whom at least one positive appears among that patient's 15 highest-ranked lesions. Appendix D.3 defines metrics, and E.2 reports AUROC and threshold-based metrics including specificity and sensitivity.

Table 1: Performance across three dermatology datasets. All values are scaled by $10^2$; higher is better. Bold indicates the best result in each column. Uncertainty denotes a conservative symmetric radius derived from 95% patient-bootstrap intervals with 2,000 resamples. Longitudinal results use observation-level evaluation.

| | ISIC 2024 | | SIIM–ISIC 2020 | | Longitudinal | |
|---|---|---|---|---|---|---|
| Model | $\mathrm{pAUC}_{80}$ | $\mathrm{PR}_{15}$ | $\mathrm{pAUC}_{80}$ | $\mathrm{PR}_{15}$ | $\mathrm{pAUC}_{80}$ | $\mathrm{PR}_{15}$ |
| *Lesion-only baselines* | | | | | | |
| ResNet-18 | 14.3 ±1.2 | 64.5 ±5.8 | 8.7 ±1.0 | 92.8 ±2.6 | 10.7 ±2.2 | 83.3 ±8.7 |
| ConvNeXtV2 | 13.0 ±1.3 | 62.9 ±5.9 | 13.1 ±0.8 | 95.2 ±2.5 | 9.2 ±1.8 | 86.9 ±7.6 |
| LightGBM | 12.6 ±1.6 | 52.5 ±5.9 | 3.1 ±0.9 | 87.2 ±3.3 | 6.9 ±1.4 | 70.2 ±10.3 |
| PanDerm MLP | 10.4 ±1.4 | 46.7 ±5.9 | 10.6 ±1.1 | 95.8 ±2.0 | **11.7 ±2.2** | **89.3 ±7.1** |
| MONET MLP | 11.1 ± 1.3 | 51.7 ± 6.2 | 11.2 ± 0.8 | 94.4 ± 2.3 | 9.1 ± 2.6 | 82.1 ± 7.9 |
| *Intrapatient baselines* | | | | | | |
| Image–LightGBM | 13.9 ±1.2 | 56.8 ±6.1 | 12.6 ±0.8 | 96.3 ±2.0 | 8.5 ±3.6 | 83.3 ±8.4 |
| Deep Sets | 10.7 ± 1.4 | 51.7 ± 6.1 | 10.3 ± 1.1 | 95.3 ± 2.1 | 8.1 ± 3.7 | 86.9 ± 7.6 |
| Set Transformer | 11.6 ± 1.5 | 56.4 ± 6.0 | 9.8 ± 1.1 | 94.4 ± 2.3 | 9.3 ± 1.9 | 86.9 ± 7.8 |
| DMT-Quadruplet | 11.4 ±1.2 | 50.2 ±6.2 | 12.0 ±0.9 | 96.3 ±2.0 | 9.5 ±3.0 | 79.8 ±9.0 |
| *Ablations* | | | | | | |
| EIDER (lesion branch) | 12.6 ±1.5 | 64.1 ±5.9 | 13.0 ±0.8 | 95.6 ±2.0 | 9.1 ±3.9 | 84.5 ±8.0 |
| **EIDER** | **16.5 ±1.2** | **75.7 ±5.4** | **13.4 ±0.8** | **97.0 ±1.7** | 9.3 ±3.3 | 84.5 ±8.2 |

### 4.2 Models

We compare EIDER with two groups of baselines: target-only models that predict each lesion using only the target lesion, and intrapatient models that use relationships among lesions from the same patient during prediction or representation learning. Implementation details are given in Appendix D.2.

**Target-only baselines.** We evaluate ResNet-18 (He et al., 2016) and ConvNeXtV2 (Woo et al., 2023) on lesion images, LightGBM (Ke et al., 2017) on lesion metadata. To test two dermatology foundation encoders, we also train the same MLP head on frozen PanDerm (Yan et al., 2025a) and MONET (Kim et al., 2024) embeddings. All models predict from the target lesion alone.

**Intrapatient baselines.** We compare against several approaches that use relationships among lesions from the same patient. An image–LightGBM stacker combines image predictions with absolute and within-patient normalized metadata, following the normalization strategy used in strong ISIC systems (Kurtansky et al., 2025). Deep Sets (Zaheer et al., 2017) and Set Transformer (Lee et al., 2019) encode the reference lesions and condition target-lesion prediction on the resulting set representation. DMT-Quadruplet (Naranpanawa et al., 2024) uses same-patient relationships during representation learning and predicts from the target lesion alone at inference.

**EIDER.** For the dermatology experiments, the lesion branch is a ConvNeXtV2 model operating on the target lesion image. The patient branch uses the lesion representation defined for each dataset in Section D: WB360 morphology features for ISIC 2024 and frozen MONET embeddings for SIIM–ISIC 2020 and the longitudinal cohort. The patient reference is estimated with $\kappa = 5$, and a two-hidden-layer MLP receives $\widehat{U}_n$, the signed and squared target deviations, and the reference size $n$ to predict the additive adjustment. The lesion branch is trained first and then frozen while the patient branch is trained on the combined prediction. The lesion-branch ablation sets the patient-branch adjustment to zero. Full preprocessing, architecture, and optimization details are given in Appendix D.

### 4.3 Comparison with Lesion-Only and Intrapatient Baselines

Table 1 compares EIDER with target-only and intrapatient baselines across the three dermatology settings. EIDER performs best on both cross-sectional datasets. On ISIC 2024, it reaches 16.5

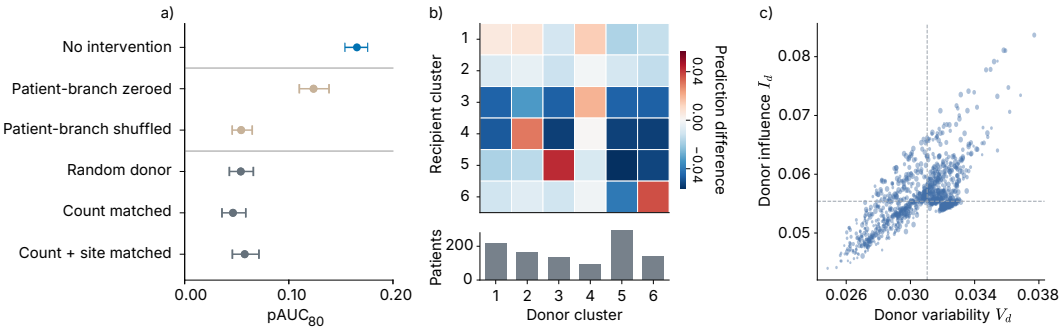


Figure 2: **Manipulating the patient reference changes predictive performance and individual predictions.** (a) $\mathrm{pAUC}_{80}$ after removing or shuffling the reference adjustment and under reference substitution. For reference substitution, the target lesion and lesion-only prediction remain fixed while its original reference set is replaced by another patient's reference set, chosen at random, matched by lesion count, or matched by both lesion count and acquisition site. (b) Mean prediction change under reference substitution, $\Delta(X, S') = \hat{p}(X, S) - \hat{p}(X, S')$, grouped by the clusters of the original and substituted reference sets. Positive values indicate that substitution lowers the predicted malignancy probability, while negative values indicate that it raises it. (c) Each substituted reference set is summarized by the mean absolute prediction change it induces across target lesions and the standard deviation of these changes.

$\mathrm{pAUC}_{80}$, compared with 14.3 for the strongest target-only baseline and 13.9 for the strongest intrapatient baseline, while improving $\mathrm{PR}_{15}$ from 64.5 to 75.7 over the strongest baseline. On SIIM–ISIC 2020, EIDER achieves 13.4 $\mathrm{pAUC}_{80}$ and 97.0 $\mathrm{PR}_{15}$, compared with 13.1 and 96.3 for the strongest baselines. The larger improvement on ISIC 2024 is consistent with total-body photography providing a richer set of other lesions from which to estimate the patient reference. The matched ablation shows a similar pattern. On ISIC 2024, adding the patient-reference adjustment increases $\mathrm{pAUC}_{80}$ from 12.6 for the lesion branch to 16.5 for EIDER. The improvement is smaller on SIIM–ISIC 2020, from 13.0 to 13.4. Thus, the contribution of the patient reference is strongest in the setting with the richest cross-lesion information. The longitudinal cohort shows a weaker intrapatient signal. EIDER reaches 9.3 $\mathrm{pAUC}_{80}$, close to its lesion branch at 9.1, while the strongest model is the lesion-only PanDerm MLP at 11.7. The same pattern holds across intrapatient baselines, suggesting that same-examination reference lesions add little beyond the target lesion in this cohort. Because the dataset also contains repeated observations over time, a longitudinal extension could separately model cross-lesion differences and temporal change.

### 4.4 Dependence on the Patient Reference

We test whether EIDER depends on the patient reference by perturbing or replacing the reference set while keeping the target lesion and lesion-only prediction fixed, and measuring the resulting changes in ranking and individual predictions.

**Reference perturbations.** We first test whether EIDER depends on the patient reference. Figure 2a removes the reference adjustment, shuffles reference-adjustment scores across target lesions, or substitutes the reference set. For substitution, the target lesion and lesion-only prediction remain fixed while the original reference set is replaced by another patient's reference set, selected at random or matched by lesion count, with or without matching acquisition site. All perturbations reduce $\mathrm{pAUC}_{80}$, and matching lesion count and acquisition site does not recover the original performance. Thus, EIDER depends on the specific patient reference rather than only on reference size or acquisition site. **Prediction differences.** The decrease in $\mathrm{pAUC}_{80}$ shows that reference substitution matters overall; we next test whether its effect depends on the receiver and donor. We call the target lesion's patient the *receiver*, with reference set $S$, and the patient supplying the replacement reference set the *donor*, with reference set $S'$. We compute $\Delta(X, S') = \hat{p}(X, S) - \hat{p}(X, S')$; positive values indicate that the donor reference lowers the prediction, while negative values indicate that it raises it. We cluster reference sets into six groups based on their prediction-difference profiles. Figure 2b reports the mean prediction difference for each receiver–donor cluster pair. Effects are larger

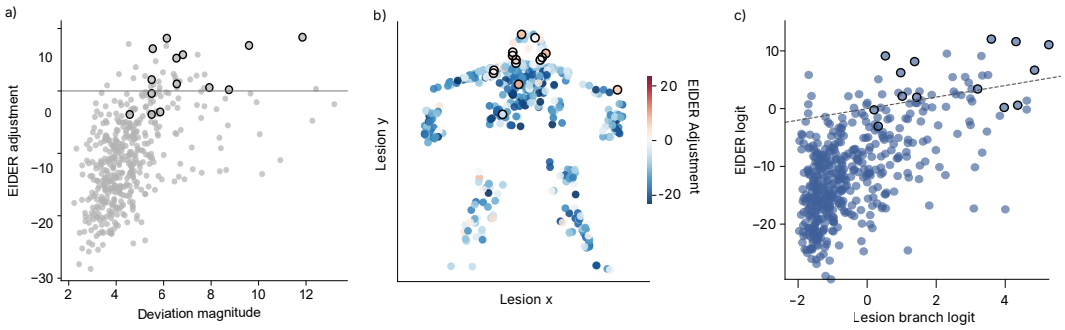


Figure 3: **Patient-level interpretation of EIDER.** For one representative patient, we visualize how the patient branch modifies lesion-level predictions. **(a)** EIDER adjustment $h(X, \Phi(S))$ as a function of deviation magnitude $\|Z_X - \widehat{U}_n\|_2$. Lesions with similar deviation magnitudes can receive substantially different adjustments. **(b)** The same lesions at their recorded body locations, colored by EIDER adjustment $h(X, \Phi(S))$. **(c)** Final EIDER logit against the lesion-branch logit $m(X)$; the dashed line denotes $y = x$, so deviations from the line correspond to the patient-branch adjustment. Malignant lesions are outlined in black throughout.

for receiver clusters 3–5 than for clusters 1–2, and their direction depends on the donor cluster. For receiver cluster 4, donors from cluster 2 lower predictions, whereas donors from clusters 1, 3, 5, and 6 raise them. Thus, the effect depends jointly on the receiver and donor references.

**Variation across donors.** We next test whether some donor reference sets have stronger effects across target lesions. Figure 2c summarizes each donor by its mean absolute prediction difference and the standard deviation of these differences across target lesions. Most donors have similar effects, while a smaller group induces larger and more variable changes. Mean absolute prediction difference correlates with its standard deviation ($\rho = 0.62$), but only weakly with donor lesion count ($\rho = 0.34$), indicating that donor effects are not explained by reference size alone.

### 4.5 Patient-Level Interpretation

Aggregate metrics do not show how EIDER uses the patient reference to modify individual predictions. We therefore inspect one patient's lesions to test whether the learned adjustment is driven only by deviation magnitude, whether it forms coherent patterns across the patient, and how it changes the lesion-branch prediction.

Figure 3a relates the EIDER adjustment to deviation magnitude $\|Z_X - \widehat{U}_n\|_2$. Lesions at similar distances from the patient reference can receive markedly different, and sometimes oppositely signed, adjustments. Thus, EIDER does not reduce the deviation to a scalar atypicality score; which morphological features differ from the reference also matters. Figure 3b maps the same adjustments to body location. Several localized groups receive similar negative adjustments, consistent with nearby lesions sharing morphology and appearing typical for that patient. The positive adjustments are more sparse and correlate with malignant labels, consistent with ugly duckling principle. Figure 3c shows the resulting change to the frozen lesion-branch logit: lesions above the identity line are promoted, while those below are suppressed. Together, the panels show that EIDER learns structured, lesion-specific corrections rather than simply increasing risk for lesions far from the patient reference. Appendix F shows the same analysis for additional patients.

## 5 Generalization to EEG Seizure Detection

To test whether the same formulation extends beyond dermatology, we apply EIDER to seizure detection on CHB–MIT (Guttag, 2010). EEG varies substantially across patients, while recordings from the same patient exhibit recurring seizure and non-seizure patterns that have motivated patient-specific seizure detection (Shoeb et al., 2004). This makes EEG a natural setting for comparing a target observation against a reference estimated from the same patient.

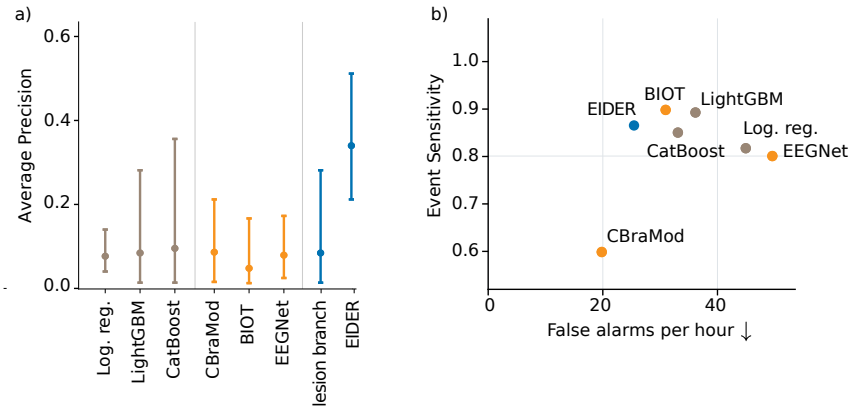


Figure 4: **Generalization to EEG seizure detection.** (a) Average precision on CHB–MIT. Band-power baselines use spectral features computed from each EEG window, whereas waveform baselines operate directly on the EEG signal. Points show means and error bars show 95% patient-bootstrap confidence intervals. (b) Event sensitivity versus false alarm events per hour at a validation-selected 1% window budget. Contiguous flagged windows are merged into alarm events; event sensitivity is the fraction of seizure events detected by at least one alarm, and false alarm events per hour counts alarms outside seizures per hour of non-seizure EEG. Higher sensitivity and fewer false alarms are preferred.

**Setup.** We split the EEG recordings into eight-second windows. Each window is a target observation, and the patient reference is estimated from earlier unlabeled windows from the same patient. The dataset contains 16,199 windows from 24 patients, including 220 seizure windows from 28 seizure events. We use patient-disjoint training, validation, and test splits, with preprocessing and model selection restricted to the training patients.

We compare against band-power classifiers using spectral features—logistic regression, LightGBM, and CatBoost—and waveform models operating directly on the EEG signal—EEGNet (Lawhern et al., 2018), BIOT (Yang et al., 2023), and CBraMod (Wang et al., 2025). Because seizure windows are rare, we use average precision (AP) as the primary window-level metric, which summarizes precision and recall without depending on a decision threshold. Additional window-level metrics are reported in Appendix G. For event-level evaluation, contiguous flagged windows form one alarm event. Event sensitivity is the fraction of seizure events detected by at least one alarm, while false alarm events per hour counts alarms outside seizures per hour of non-seizure EEG. We select on the validation patients the score threshold that flags 1% of windows and apply it unchanged to the test patients.

EIDER achieves the highest AP among the evaluated models at 0.362 (Fig. 4a), substantially above the target-only model. At the event level, EIDER lies on the Pareto frontier between sensitivity and false alarms (Fig. 4b), having the least false alarm per hour among the high sensitivity baselines.

## 6 Conclusion

We study prediction of a target observation using other unlabeled observations from the same patient to estimate an individual reference. Across dermatology datasets, EIDER improves prediction by combining target-specific information with deviations from this reference, and controlled interventions show that the identity and quality of the reference matter.

These results suggest one way to account for inter-patient heterogeneity when repeated observations from the same patient are available. The same formulation may be relevant to other tasks in which an observation is interpreted relative to an individual's usual measurements, as explored here with EEG seizure detection. More broadly, the work illustrates how clinical reasoning principles such as the ugly-duckling principle can motivate prediction structures that are not captured by treating observations independently.

## 7 AI DISCLOSURE STATEMENT

In this work, we used generative AI tools to assist in writing proofs, provide feedback on research methodology and experiments, and implement methods. We did not use generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, generate synthetic datasets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, assist with translation, clean or reformat datasets, or interpret results. Qualitative and thematic data analysis was not applicable to this work.

Additionally, we used generative AI tools to create or edit software code, summarize and analyze existing literature, brainstorm ideas, search for and source information, edit the manuscript for readability, and propose titles or keywords. We did not use generative AI tools to create or modify scientific figures or images, suggest experimental parameters, identify research topics or gaps, or determine the structure of the paper.

All AI-assisted mathematical arguments, proofs, code, literature summaries, and manuscript edits were reviewed by the authors. Mathematical claims and proofs were checked against the stated assumptions and derivations, generated or modified code was inspected and tested against the intended experimental procedures, and literature-related outputs were verified against the cited sources. The authors made the final decisions regarding the methodology, experiments, analysis, and presentation of the work. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

## REFERENCES

Giuseppe Argenziano, Gabriella Fabbrocini, Paolo Carli, Vincenzo De Giorgi, Elisa Sammarco, and Mario Delfino. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Archives of Dermatology*, 134(12):1563–1570, 1998. doi: 10.1001/archderm.134.12.1563.

Giuseppe Argenziano, H. Peter Soyer, Sergio Chimenti, Renato Talamini, Roberto Corona, Franco Sera, et al. Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the internet. *Journal of the American Academy of Dermatology*, 48(5):679–693, 2003. doi: 10.1067/mjd.2003.281.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000. doi: 10.1145/342009.335388.

Titus J. Brinker, Achim Hekler, Alexander H. Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Tim Holland-Letz, Jochen S. Utikal, Christof von Kalle, Wiebke Ludwig-Peitsch, Judith Sirokay, Lucie Heinzerling, Magarete Albrecht, Katharina Baratella, Lena Bischof, Eleftheria Chorti, Anna Dith, Christina Drusio, Nina Giese, Emmanouil Gratsias, Klaus Griewank, Sandra Hallasch, Zdenka Hanhart, Saskia Herz, Katja Hohaus, Philipp Jansen, Finja Jockenhöfer, Theodora Kanaki, Sarah Knispel, Katja Leonhard, Anna Martaki, Liliana Matei, Johanna Matull, Alexandra Olischewski, Maximilian Petri, Jan-Malte Placke, Simon Raub, Katrin Salva, Swantje Schlott, Elsa Sody, Nadine Steingrube, Ingo Stoffels, Selma Ugurel, Anne Zaremba, Christoffer Gebhardt, Nina Booken, Maria Christolouka, Kristina Buder-Bakhaya, Therezia Bokor-Billmann, Alexander Enk, Patrick Gholam, Holger Hänßle, Martin Salzmann, Sarah Schäfer, Knut Schäkel, Timo Schank, Ann-Sophie Bohne, Sophia Deffaa, Katharina Drerup, Friederike Egberts, Anna-Sophie Erkens, Benjamin Ewald, Sandra Falkvoll, Sascha Gerdes, Viola Harde, Axel Hauschild, Marion Jost, Katja Kosova, Laetitia Messinger, Malte Metzner, Kirsten Morrison, Rogina Motamedi, Anja Pinczker, Anne Rosenthal, Natalie Scheller, Thomas Schwarz, Dora Stölzl, Federieke Thielking, Elena Tomaschewski, Ulrike Wehkamp, Michael Weichenthal, Oliver Wiedow, Claudia Maria Bär, Sophia Bender-Säbelkampf, Marc Horbrügger, Ante Karoglan, Luise Kraas, Jörg Faulhaber, Cyrill Geraud, Ze Guo, Philipp Koch, Miriam Linke, Nolwenn Maurier, Verena Müller, Benjamin Thomas, Jochen Sven Utikal, Ali Saeed M. Alamri, Andrea Baczako,

Carola Berking, Matthias Betke, Carolin Haas, Daniela Hartmann, Markus V. Heppt, Katharina Kilian, Sebastian Krammer, Natalie Lidia Lapczynski, Sebastian Mastnik, Suzan Nasifoglu, Cristel Ruini, Elke Sattler, Max Schlaak, Hans Wolff, Birgit Achatz, Astrid Bergbreiter, Konstantin Drexler, Monika Ettinger, Sebastian Haferkamp, Anna Halupczok, Marie Hegemann, Verena Dinauer, Maria Maagk, Marion Mickler, Biance Philipp, Anna Wilm, Constanze Wittmann, Anja Gesierich, Valerie Glutsch, Katrin Kahlert, Andreas Kerstan, Bastian Schilling, and Philipp Schrüfer. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113:47–54, 2019. ISSN 0959-8049. doi: https://doi.org/10.1016/j.ejca.2019.04.001. URL https://www.sciencedirect.com/science/article/pii/S0959804919302217.

Chris U. Carmona, François-Xavier Aubet, Valentin Flunkert, and Jan Gasthaus. Neural contextual anomaly detection for time series. *arXiv preprint arXiv:2107.07702*, 2021. URL https://arxiv.org/abs/2107.07702.

Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7354–7362, 2019. doi: 10.1109/CVPR.2019.00753.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017. doi: 10.1109/TPAMI.2016.2615921.

Harrison Edwards and Amos Storkey. Towards a neural statistician. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HJDBUF5le.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017. doi: 10.1038/nature21056.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1704–1713, 2018.

Caroline Gaudy-Marqueste, Yanal Wazaefi, Yvane Bruneu, Raoul Triller, Luc Thomas, Giovanni Pellacani, Josep Malvehy, Marie-Françoise Avril, Sandrine Monestier, Marie-Aleth Richard, Bernard Fertil, and Jean-Jacques Grob. Ugly duckling sign as a major factor of efficiency in melanoma detection. *JAMA Dermatology*, 153(4):279–284, 2017. doi: 10.1001/jamadermatol.2016.5500.

Nima Ghahari, Liam Caffery, Brigid Betz-Stablein, Adam Mothershaw, Dilki Jayasinghe, Clare Primiero, Shekhar S. Chandra, Joachim Torrano, H. Peter Soyer, and Monika Janda. A longitudinal dataset of tile and corresponding dermoscopic images with metadata for identifying skin cancers. *Scientific Data*, 12:5880, 2025. doi: 10.1038/s41597-025-05880-2. URL https://doi.org.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E. Turner. Meta-learning probabilistic inference for prediction. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkxStoC5F7.

Jean-Jacques Grob and Jean-Jacques Bonerandi. The "ugly duckling" sign: Identification of the common characteristics of nevi in an individual as a basis for melanoma screening. *Archives of Dermatology*, 134(1):103–104, 1998. doi: 10.1001/archderm.134.1.103-a.

John Guttag. CHB-MIT Scalp EEG Database. *PhysioNet*, 1.0.0, 2010. doi: 10.13026/C2K01R. URL https://doi.org/10.13026/C2K01R.

Holger A Haenssle, Christine Fink, Tomek Jakob Schneider, Ferdinand Toberer, Timo Buhl, Andreas Blum, A Aicollis, et al. Man against machine: diagnostic performance of a deep learning

convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018. doi: 10.1093/annonc/mdy166. URL https://doi.org.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

A. Ji-Xu, J. Dinnes, and R.N. Matin. Total body photography for the diagnosis of cutaneous melanoma in adults: a systematic review and meta-analysis. *British Journal of Dermatology*, 185(2):302–312, 2021. doi: https://doi.org/10.1111/bjd.19759. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/bjd.19759.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Chanwoo Kim, Soham U. Gadgil, Alex J. DeGrave, Jesutofunmi A. Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nature Medicine*, 30:1154–1165, 2024. doi: 10.1038/s41591-024-02887-x.

Nicholas R. Kurtansky, Maura C. Gillis, Noel C. F. Codella, Brian M. D'Alessandro, Zongyuan Ge, Pascale Guitera, Allan C. Halpern, Harald Kittler, Josep Malvehy, et al. Automated triage of cancer-suspicious skin lesions with 3d total-body photography. *npj Digital Medicine*, 8:708, 2025. doi: 10.1038/s41746-025-02070-7.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4): 963–974, 1982. doi: 10.2307/2529876.

Vernon J Lawhern, Amelia J Solon, Waywel D Waytowich, Stephen M Gordon, Chou P Hung, and Lance E Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, jul 2018. doi: 10.1088/1741-2552/aace8c. URL https://doi.org.

Hyo Min Lee, Fatemeh Fadaie, Ravnoor Singh Gill, Benoît Caldairou, Veronique Sziklas, Jeremy Crane, Andrea Bernasconi, and Neda Bernasconi. Decomposing mri phenotypic heterogeneity in epilepsy: a step towards personalized classification. *Brain*, 145(3):897–908, 2022. doi: 10.1093/brain/awab425.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, 2019.

Zhong Li and Matthijs van Leeuwen. Explainable contextual anomaly detection using quantile regression forests. *Data Mining and Knowledge Discovery*, 37(6):2517–2563, 2023. doi: 10.1007/s10618-023-00967-z.

Francesca Mandel, Riddhi Pratim Ghosh, and Ian Barnett. Neural networks for clustered and longitudinal data using mixed effects models. *Biometrics*, 79(2):711–721, 2023. doi: 10.1111/biom.13615.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

António Mendes and Catarina Barata. Intrapatient diagnosis for melanoma skin cancer. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023. doi: 10.1109/ISBI53787.2023.10230696.

Scott W Menzies. Dermoscopy: alternative melanocytic algorithms—the abcd rule of dermatoscopy, menzies scoring method, and 7-point checklist. *Clinics in Dermatology*, 20(3):240–247, 2002. doi: 10.1016/S0738-081X(02)00236-5.

Mohammadreza Mohseni, Jordan Yap, William Yolland, Arash Koochek, and Stella Atkins. Can self-training identify suspicious ugly duckling lesions? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1829–1836, 2021.

Chaya S. Moskowitz, Mattea L. Welch, Michael A. Jacobs, Brenda F. Kurland, and Amber L. Simpson. Radiomic analysis: Study design, statistical analysis, and other bias mitigation strategies. *Radiology*, 304(2):259–270, may 2022. doi: 10.1148/radiol.211597. URL https://pubs.rsna.org/doi/abs/10.1148/radiol.211597.

Nathasha Naranpanawa, H. Peter Soyer, Adam Mothershaw, Gayan K. Kulatilleke, Zongyuan Ge, Brigid Betz-Stablein, and Shekhar S. Chandra. A tiered quadruplet network with patient-specific mining and dynamic margin for improved ugly duckling lesion classification. In *2024 IEEE International Symposium on Biomedical Imaging*, pp. 1–5, 2024. doi: 10.1109/ISBI56570.2024.10635546.

Sachin Ravi and Alex Beatson. Amortized bayesian meta-learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rkgpy3C5tX.

Jenna E. Rayner, Antonia M. Laino, Kaitlin L. Nufer, Laura Adams, Anthony P Raphael, Scott W Menzies, and H. Peter Soyer. Clinical perspective of 3d total body photography for early detection and screening of melanoma. *Frontiers in Medicine*, Volume 5 - 2018, 2018. ISSN 2296-858X. doi: 10.3389/fmed.2018.00152. URL https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2018.00152.

George K. Robinson. That blup is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32, 1991. doi: 10.1214/ss/1177011926.

Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8(1):34, 2021. doi: 10.1038/s41597-021-00815-z.

Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulò, Nicu Sebe, and Elisa Ricci. Unsupervised domain adaptation using feature-whitening and consensus loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9463–9472, 2019. doi: 10.1109/CVPR.2019.00970.

Ali Shoeb, Herman Edwards, Jack Connolly, Blaise Bourgeois, S. Ted Treves, and John Guttag. Patient-specific seizure onset detection. *Epilepsy & Behavior*, 5(4):483–498, 2004. ISSN 1525-5050. doi: https://doi.org/10.1016/j.yebeh.2004.05.005. URL https://www.sciencedirect.com/science/article/pii/S1525505004001593.

Giora Simchoni and Saharon Rosset. Integrating random effects in deep neural networks. *Journal of Machine Learning Research*, 24(156):1–57, 2023. URL https://jmlr.org/papers/v24/22-0501.html.

Luis R. Soenksen, Timothy Kassis, Susan T. Conover, Berta Marti-Fuster, Judith S. Birkenfeld, Jason Tucker-Schwartz, Asif Naseem, Robert R. Stavert, Caroline C. Kim, Maryanne M. Senna, José Avilés-Izquierdo, James J. Collins, Regina Barzilay, and Martha L. Gray. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Science Translational Medicine*, 13(581):eabb3652, 2021. doi: 10.1126/scitranslmed.abb3652. URL https://www.science.org/doi/abs/10.1126/scitranslmed.abb3652.

Wilhelm Stolz, Andreas Riemann, Armand B. Cognetta, Laurent Pillet, Wolfgang Abmayr, Dieter H"olzel, Peter Bilek, Frank Nachbar, Michael Landthaler, and Otto Braun-Falco. ABCD rule of dermatoscopy: A new practical method for early recognition of malignant melanoma. *European Journal of Dermatology*, 4:521–527, 1994.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. doi: 10.1609/aaai.v30i1.10306.

Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016. URL https://www.jmlr.org/papers/v17/14-510.html.

James H. Thrall, David Fessell, and Pari Pandharipande. Rethinking the approach to artificial intelligence for medical image analysis: The case for precision diagnosis. *Journal of the American College of Radiology*, 18(2):174–179, 2021. ISSN 1546-1440. doi: 10.1016/j.jacr.2020.07.010.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020. doi: 10.1038/s41591-020-0942-0.

Vullnet Useini, Stephanie Tanadini-Lang, Quentin Lohmeyer, Mirko Meboldt, Nicolaus Andratschke, Ralph P. Braun, and Javier Barranco García. Automatized self-supervised learning for skin lesion screening. *Scientific Reports*, 14(1):12697, 2024. doi: 10.1038/s41598-024-61681-4.

Michael Volpp, Fabian Flürenbrock, Lukas Grossberger, Christian Daniel, and Gerhard Neumann. Bayesian context aggregation for neural processes. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=ufZN2-aehFa.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. TENT: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025.

Yanal Wazaefi, Caroline Gaudy-Marqueste, Marie-Françoise Avril, Joseph Malvehy, Giovanni Pellacani, Luc Thomas, Raoul Triller, Yvane Bruneu, Sandrine Monestier, Marie-Aleth Richard, Bernard Fertil, and Jean-Jacques Grob. Evidence of a limited intra-individual diversity of nevi: Intuitive perception of dominant clusters is a crucial step in the analysis of nevi by dermatologists. *Journal of Investigative Dermatology*, 133(10):2355–2361, 2013. ISSN 0022-202X. doi: https://doi.org/10.1038/jid.2013.183. URL https://www.sciencedirect.com/science/article/pii/S0022202X15359911.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142, June 2023.

Jiqing Wu, Inti Zlobec, Maxime Lafarge, Yukun He, and Viktor Koelzer. Towards iid representation learning and its application on biomedical data. In Ender Konukoglu, Bjoern Menze, Archana Venkataraman, Christian Baumgartner, Qi Dou, and Shadi Albarqouni (eds.), *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pp. 1375–1402. PMLR, 06–08 Jul 2022. URL https://proceedings.mlr.press/v172/wu22a.html.

Siyuan Yan, Ming Hu, Yiwen Jiang, Xieji Li, Hao Fei, Philipp Tschandl, Harald Kittler, and Zongyuan Ge. Derm1M: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12681–12690, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yan_Derm1M_A_Million-scale_Vision-Language_Dataset_Aligned_with_Clinical_Ontology_Knowledge_ICCV_2025_paper.html.

Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 31:2691–2702, 2025b. doi: 10.1038/s41591-025-03747-y.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. Biot: biosignal transformer for cross-data learning in the wild. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Zhen Yu, Victoria Mar, Anders Eriksson, Shakes Chandra, Paul Bonnington, Lei Zhang, and Zongyuan Ge. End-to-end ugly duckling sign detection for melanoma identification with transformers. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12907 of *Lecture Notes in Computer Science*, pp. 176–184. Springer, 2021. doi: 10.1007/978-3-030-87234-2_17.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017.

Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

# A Extended Literature Review

This appendix expands three aspects of our positioning: the clinical basis for combining target-only evidence with evidence from the reference set, the corresponding design choices in prior melanoma models, and the distinction between prediction from a target observation and reference set and adjacent set-learning settings.

## A.1 Clinical Basis of Patient Heterogeneity

Nevi within an individual tend to share recurring morphological characteristics, providing a patient reference for identifying lesions that break that pattern (Grob & Bonerandi, 1998). At the same time, melanoma assessment depends on which characteristics differ—including pattern, color, symmetry, and dermoscopic structures—and their interpretation can vary by anatomical site (Argenziano et al., 2003; Menzies, 2002). A lesion's deviation from the patient reference therefore provides different information from target-only assessment, motivating the two sources of evidence illustrated in Fig. 1a and represented by the target-only term $m(X)$ and reference term $h(X, \Phi(S))$.

## A.2 Relationship to Set Learning Settings

Several learning settings condition prediction on additional observations, but differ in what those observations represent and how they affect the prediction target. Table 2 summarizes the objective of each setting.

Table 2: Problem formulations related to prediction from a target observation and an additional set. Here $x \in \mathcal{X}$ is the target observation, $S_X = \{x_i\}_{i=1}^n$ is an unlabeled set, and $S_{XY} = \{(x_i, y_i)\}_{i=1}^n$ is a labeled set. $\widehat{P}_S$ denotes the distribution estimated from $S$, and $D$ denotes a measure of how atypical $x$ is relative to that distribution. Subscripts $g$, $\tau$, and $d$ denote a group, task, and domain, respectively.

| Setting | Problem formulation |
|---|---|
| Latent set inference (Edwards & Storkey, 2017) | $S_X \longmapsto p(U \mid S_X)$ |
| Distribution regression (Szabó et al., 2016) | $S_X \sim P^n \longmapsto y(P)$ |
| Meta-learning / Neural Processes (Gordon et al., 2019; Ravi & Beatson, 2019; Garnelo et al., 2018; Volpp et al., 2021) | $(x, S^\tau_{XY}) \longmapsto p(y \mid x, S^\tau_{XY})$ |
| Test-time adaptation (Wang et al., 2021; Zhang et al., 2021) | $S^d_X \longmapsto \{p(y_i \mid x_i, S^d_X)\}_{i=1}^n$ |
| Contextual anomaly detection (Breunig et al., 2000; Carmona et al., 2021) | $(x, S_X) \longmapsto D\Big(x, \widehat{P}_{S_X}\Big)$ |
| Mixed-effects prediction (Laird & Ware, 1982; Robinson, 1991) | $(x, S^g_{XY}) \longmapsto p(y \mid x, S^g_{XY})$ |
| Ours | $(x, S^g_X) \longmapsto p(y \mid x, S^g_X), \quad x \notin S^g_X$ |

## A.3 Melanoma Models Using Reference Lesions

Motivated by Lemma 3.3, we compare melanoma methods that use reference lesions by their target-only term $m(X)$, reference representation $\Phi(S)$, and reference term $h(X, \Phi(S))$. Table 3 summarizes these choices.

Table 3: Melanoma methods characterized by their target-only term, reference representation, and reference term.

| Method | $m(X)$ | $\Phi(S)$ | $h(X, \Phi(S))$ |
|---|---|---|---|
| Mohseni et al. (2021) | Target embedding | Mean reference-lesion embedding | Distance from target to reference mean |
| Yu et al. (2021) | Target representation | Reference-lesion representations | Transformer interaction |
| Naranpanawa et al. (2024) | Metric-learned target representation | — | — |
| Kurtansky et al. (2025) | Target-lesion features | Reference-set summaries | Joint prediction |
| EIDER | Target-only prediction | Patient reference estimate and reference size | Prediction from signed and squared deviations |

# B Theory Details

We derive EIDER from a Bayes decomposition of prediction from a target observation and a reference set. We first show how the reference set contributes an additive reference term and how a predictively sufficient statistic $\Phi(S)$ can represent the information it contributes. We then instantiate $\Phi(S)$ with a hierarchical patient-reference model, derive a partially pooled patient reference estimate, and characterize how signed and squared deviations from this reference contribute to prediction. These results motivate the lesion-only and reference branches used by EIDER.

## B.1 Notation

For $x \in \mathbb{R}^d$, $\|x\|_2$ denotes the Euclidean norm and $I_d$ the $d$-dimensional identity matrix. For a diagonal matrix $A = \mathrm{diag}(a_1, \ldots, a_d)$ we write $a_k$ for its $k$-th diagonal entry, so every product of diagonal matrices below acts feature by feature. For a nonempty reference set $S = \{Z_1, \ldots, Z_n\}$, let $\overline{Z}$ be the empirical mean. We write $\pi_y = P(Y = y)$ and use $p(\cdot \mid \cdot)$ for conditional densities.

## B.2 Prediction Decomposition

*Proof of Lemma 3.3.* Bayes' rule gives, for $y \in \{0, 1\}$,

$$P(Y = y \mid X, S) = \frac{p(S \mid X, Y = y) P(Y = y \mid X)}{p(S \mid X)}.$$

Taking the ratio between $y = 1$ and $y = 0$,

$$\frac{P(Y = 1 \mid X, S)}{P(Y = 0 \mid X, S)} = \frac{P(Y = 1 \mid X)}{P(Y = 0 \mid X)} \frac{p(S \mid X, Y = 1)}{p(S \mid X, Y = 0)}.$$

Taking logarithms yields Eq. 1. □

## B.3 Hierarchical Model of Patient Reference

We use the prediction setting from Definition 3.1 and model the reference set as repeated observations around a latent patient reference, combining population and patient information through partial pooling (Laird & Ware, 1982; Robinson, 1991). Let $Z_X \in \mathbb{R}^d$ denote the representation of the target observation and $S = \{Z_1, \ldots, Z_n\}$ the reference representations in the same feature space.

**Assumption B.1** (Patient reference)**.** *Each patient has a latent patient reference*

$$U \sim \mathcal{N}(\mu, T), \qquad T = \mathrm{diag}(\tau_1^2, \ldots, \tau_d^2),$$

*where $\mu$ is the population reference and $\tau_k^2$ is the between-patient variance of feature $k$.*

This assumption models patient references as varying around a population reference. The feature-specific variances allow some features to vary more across patients than others.

**Assumption B.2** (Reference lesions)**.** *Conditional on the patient reference, the reference-lesion representations are independent draws,*

$$Z_j \mid U \overset{\text{iid}}{\sim} \mathcal{N}(U, \Sigma_c), \qquad j \in \{1, \dots, n\}, \qquad \Sigma_c = \operatorname{diag}(\sigma_{c,1}^2, \dots, \sigma_{c,d}^2).$$

This assumption models variation among lesions from the same patient around the patient reference. Together, $T$ and $\Sigma_c$ determine how strongly the estimate of each reference feature is pooled toward the population reference.

**Proposition B.3** (Patient-reference estimate)**.** *Under Assumptions B.1 and B.2,*

$$U \mid S \sim \mathcal{N}(\widehat{U}_n, V_n), \qquad V_n = \operatorname{diag}(v_{n,1}, \dots, v_{n,d}),$$

*where, for each feature $k$,*

$$\widehat{U}_{n,k} = \frac{\sum_{j=1}^n Z_{j,k} + \kappa_k \mu_k}{n + \kappa_k}, \qquad \kappa_k = \frac{\sigma_{c,k}^2}{\tau_k^2}, \qquad v_{n,k} = \frac{\tau_k^2 \sigma_{c,k}^2}{\sigma_{c,k}^2 + n\tau_k^2}. \tag{5}$$

*Larger $\kappa_k$ gives more weight to the population reference, while increasing $n$ reduces the posterior variance of the patient reference.*

*Proof.* Combining the Gaussian prior and reference likelihood gives

$$p(u \mid S) \propto \exp\left( -\tfrac{1}{2}(u-\mu)^\top T^{-1}(u-\mu) - \tfrac{1}{2}\sum_{j=1}^n (Z_j - u)^\top \Sigma_c^{-1}(Z_j - u) \right).$$

Completing the square gives

$$V_n = \left(T^{-1} + n\Sigma_c^{-1}\right)^{-1}, \qquad \widehat{U}_n = V_n \left(T^{-1}\mu + n\Sigma_c^{-1}\overline{Z}\right).$$

Because $T$ and $\Sigma_c$ are diagonal, these expressions reduce feature-wise to Eq. 5. □

**Proposition B.4** (Predictive sufficiency of the patient reference)**.** *Under Assumptions B.1 and B.2, and $S \perp (X, Y) \mid U$,*

$$p(U \mid S) = \mathcal{N}(\widehat{U}_n, V_n)$$

*depends on $S$ only through $(\widehat{U}_n, V_n)$. Since $V_n$ is determined by $n$, $\Phi(S) = (\widehat{U}_n, n)$ is predictively sufficient: $Y \perp S \mid (X, \Phi(S))$.*

*Proof.* By conditional independence,

$$p(X, Y \mid S) = \int p(X, Y \mid U)\, p(U \mid S)\, dU.$$

Hence $p(Y \mid X, S)$ depends on $S$ only through $p(U \mid S)$, which Proposition B.4 determines from $(\widehat{U}_n, n)$. □

### B.4 Prediction under the Hierarchical Model

The reference model thus reduces the patient's other lesions to a patient reference and its uncertainty. We next model how the target lesion differs from this reference. It then models the target lesion relative to the same patient reference, yielding separate terms for squared and signed lesion-to-reference deviations.

**Assumption B.5** (Class-dependent target lesion)**.** *For $Y \in \{0, 1\}$,*

$$Z_X \mid U, Y = y \sim \mathcal{N}(U + \delta_y, \Sigma_y), \qquad \Sigma_y = \operatorname{diag}(\sigma_{y,1}^2, \dots, \sigma_{y,d}^2).$$

This assumption formalizes the ugly-duckling principle by making malignancy evidence depend on how the target lesion differs from the patient's reference. The class-dependent mean $\delta_y$ captures which directions of deviation are associated with each class, while $\Sigma_y$ captures how much these deviations vary across features.

**Proposition B.6** (Patient-Reference Logit Decomposition)**.** *Under Assumption B.5 and with* $r = Z_X - U$*,*

$$\operatorname{logit} P(Y = 1 \mid Z_X, U) = b(U) + c + r^\top \alpha r + \beta^\top r, \tag{6}$$

*where* $b(U) = \operatorname{logit} P(Y = 1 \mid U)$ *is the patient risk,* $c$ *does not depend on* $r$*,* $\alpha = \frac{1}{2}\left(\Sigma_0^{-1} - \Sigma_1^{-1}\right)$*, and* $\beta = \Sigma_1^{-1}\delta_1 - \Sigma_0^{-1}\delta_0$*.*

*Because* $\alpha$ *is diagonal,*

$$r^\top \alpha r = \sum_k \alpha_k r_k^2, \qquad \alpha_k = \tfrac{1}{2}\left(\sigma_{0,k}^{-2} - \sigma_{1,k}^{-2}\right).$$

*Thus* $r^\top \alpha r$ *captures class evidence from squared feature-wise deviations, while* $\beta^\top r$ *captures class evidence from their direction.*

*Proof.* Bayes' rule gives

$$\operatorname{logit} P(Y = 1 \mid Z_X, U) = b(U) + \log \frac{p(Z_X \mid U, Y = 1)}{p(Z_X \mid U, Y = 0)}.$$

Substituting the Gaussian target-lesion distributions and expanding $(r - \delta_y)^\top \Sigma_y^{-1}(r - \delta_y)$ gives Eq. 6, with

$$c = -\tfrac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_0|} - \tfrac{1}{2}\delta_1^\top \Sigma_1^{-1}\delta_1 + \tfrac{1}{2}\delta_0^\top \Sigma_0^{-1}\delta_0.$$

□

*Remark* B.7 (EIDER Architecture). EIDER follows Proposition B.6 by estimating the patient reference with $\widehat{U}_n$ and representing the target lesion through signed and squared deviations from this estimate. Proposition B.4 shows that $(\widehat{U}_n, n)$ is predictively sufficient, with $n$ determining the uncertainty in $\widehat{U}_n$ and this uncertainty decreasing as more reference lesions are observed. The patient branch learns the resulting prediction adjustment without assuming that substituting $\widehat{U}_n$ for $U$ yields the exact marginalized likelihood ratio.

*Remark* B.8 (Squared and signed deviations). Proposition B.6 separates patient-level risk $b(U)$ from two forms of intrapatient evidence. Equal class covariances, $\Sigma_0 = \Sigma_1$, remove the squared-deviation term. When $\beta = 0$, the signed-deviation term vanishes and intrapatient evidence depends only on squared deviations. Any representation that retains only unsigned deviations therefore discards the directional evidence represented by $\beta^\top r$.

# C Theory Validation

The following tests examine four predictions of the hierarchical model: the shrinkage scale, reference uncertainty as the reference set grows, the single-reference assumption, and the form of the deviation contribution to prediction.

## C.1 Estimating Shrinkage

For each training fold and feature $k$, we estimate the between-patient variance $\tau_k^2$ and within-patient variance $\sigma_{c,k}^2$, and compute $\kappa_k = \sigma_{c,k}^2/\tau_k^2$. The median feature-wise estimate is $\kappa_k = 4.06$, close to the fixed $\kappa = 5$ used by EIDER; the estimated shared value is 3.87. Using the same prediction head, fixed, shared, and feature-wise shrinkage reach $\mathrm{pAUC}_{80}$ values of .1534, .1545, and .1544, respectively, with no clear paired difference. Thus, $\kappa = 5$ is consistent with the scale estimated from the training data, while feature-wise shrinkage provides no detectable ranking improvement.

### C.2 Reference Error as Reference Size Grows

Proposition B.3 predicts that the posterior variance of reference feature $k$ decreases with the number of reference lesions as $v_{n,k} = \sigma^2_{c,k}/(n + \kappa_k)$. For a held-out lesion from the same patient, the corresponding expected squared error is $e_{n,k} = \sigma^2_{c,k} + v_{n,k}$; we average this quantity across features to obtain the predicted error.

We estimate the variance components on training patients and evaluate the fixed prediction on held-out patients using the same frozen MONET representation used elsewhere in the experiments. The predicted error closely follows the measured error ($R^2 = .989$; Figure 5a). Partial pooling provides the largest reduction relative to the unpooled patient mean for small reference sets, while the two converge as the reference size increases.

### C.3 Testing the Single-Reference Assumption

Assumption B.1 represents a patient's reference lesions with a single patient reference. We test whether prediction benefits from representing several recurring lesion types instead.

For $K > 1$, we fit a $K$-component mixture to the training-fold lesion representations and assign each patient's reference lesions to these components. A separate partially pooled patient reference is estimated for each component, and each target lesion is compared with its closest patient component. Thus, $K = 1$ recovers the single-reference model, while $K > 1$ provides multiple candidate references.

The single-reference model achieves the highest $\mathrm{pAUC}_{80}$, with lower performance for every $K > 1$ in the sweep (Figure 5b). These results support the single-reference representation used by EIDER in this setting; they do not imply that the lesion distribution itself is unimodal.

### C.4 Patient Reference Logit Decomposition

Proposition B.6 predicts signed and squared terms in the target lesion's deviation from the patient reference. We test this functional form using a directional Gaussian head containing only the signed term, a magnitude Gaussian head containing only the squared term, and the full diagonal Gaussian rule containing both. We compare these constrained forms with EIDER's flexible residual MLP while keeping the patient reference estimate and lesion branch fixed.

The full diagonal Gaussian rule outperforms either restricted Gaussian form, while the residual MLP achieves the highest $\mathrm{pAUC}_{80}$ (Figure 5c). The comparison supports retaining both signed and squared deviations, while the MLP result indicates that a flexible mapping can capture structure beyond the Gaussian approximation.

## D Melanoma Experimental Details

**Datasets and Evaluation Protocol** We evaluate on three dermatology datasets with different amounts and types of same-patient information. ISIC 2024 contains 401,059 lesions from 1,042 patients, with a median of 241 lesions per patient and 393 malignant lesions. Following the ISIC 2024 setup, we use the 33 numerical measurements provided by the Vectra WB360 system to describe lesion morphology, including size, shape, color, symmetry, and 3D location. We refer to these measurements as the *WB360 features*; Table 4 lists the complete set.

SIIM–ISIC 2020 contains 33,126 dermoscopic images from 2,056 patients, including 584 melanomas, with a median of 12 images per patient. The longitudinal dataset contains 35,355 dermoscopic observations from 458 patients, including 427 positive observations from 84 patients. For the main observation-level evaluation, the reference set for each target contains the patient's other lesions observed at the same visit, excluding the target lesion. For comparison with prior lesion-level methods, we also collapse repeated observations to the latest visit of each lesion, yielding 9,311 lesions including 147 malignant lesions.

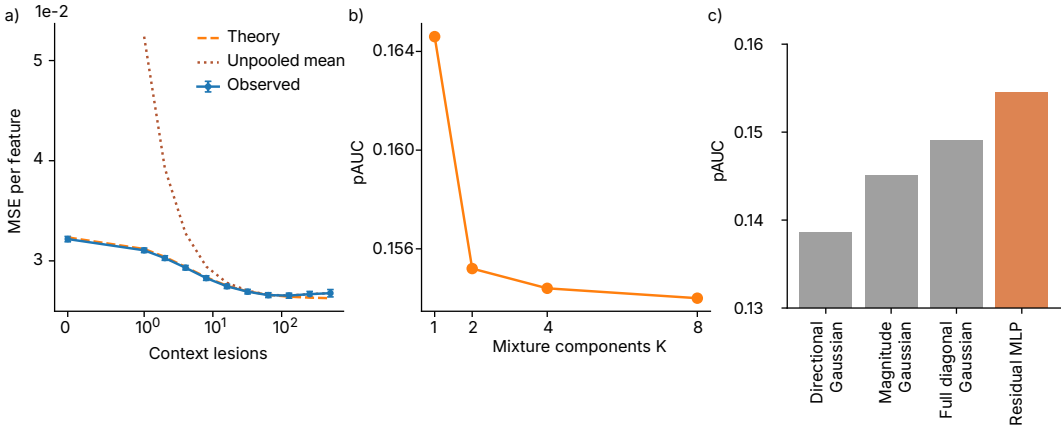


Figure 5: **Validation of the hierarchical patient-reference model.** **(a)** Predicted and observed held-out feature error as the number of reference lesions increases. The theoretical curve is fixed from training-fold variance estimates; partial pooling improves over the unpooled mean most for small reference sets. **(b)** High-sensitivity ranking for $K$-component patient references. The single-reference model ($K = 1$) gives the highest $\mathrm{pAUC}_{80}$ in the sweep. **(c)** Comparison of the deviation terms predicted by the logit decomposition. The full diagonal Gaussian rule combines signed and squared deviations; the residual MLP uses the same reference estimate with a flexible mapping.

Table 4: **Structured lesion measurements used from ISIC 2024.** These 33 numerical measurements form the WB360 representation used by EIDER's patient branch.

| Group | Features |
|---|---|
| Size and shape | `clin_size_long_diam_mm`, `tbp_lv_areaMM2`, `tbp_lv_area_perim_ratio`, `tbp_lv_eccentricity`, `tbp_lv_minorAxisMM`, `tbp_lv_norm_border`, `tbp_lv_perimeterMM`, `tbp_lv_symm_2axis`, `tbp_lv_symm_2axis_angle` |
| Color and appearance | `tbp_lv_A`, `tbp_lv_Aext`, `tbp_lv_B`, `tbp_lv_Bext`, `tbp_lv_C`, `tbp_lv_Cext`, `tbp_lv_H`, `tbp_lv_Hext`, `tbp_lv_L`, `tbp_lv_Lext`, `tbp_lv_color_std_mean`, `tbp_lv_deltaA`, `tbp_lv_deltaB`, `tbp_lv_deltaL`, `tbp_lv_deltaLB`, `tbp_lv_deltaLBnorm`, `tbp_lv_norm_color`, `tbp_lv_radial_color_std_max`, `tbp_lv_stdL`, `tbp_lv_stdLExt` |
| Other lesion measurements | `tbp_lv_nevi_confidence`, `tbp_lv_x`, `tbp_lv_y`, `tbp_lv_z` |

All datasets use five patient-disjoint folds. Confidence intervals and paired comparisons use 2,000 patient-level bootstrap resamples. All patient-aware models exclude the target lesion from its reference set during training and evaluation.

**Lesion representations.** Intrapatient models require a common feature space in which to represent the target lesion and the patient's other lesions. We choose representations that encode lesion morphology, so that comparisons between lesions reflect differences in clinically relevant appearance rather than raw image variation. On ISIC 2024, we use the structured WB360 morphology measurements listed in Table 4. On SIIM–ISIC 2020 and the longitudinal cohort, where comparable structured measurements are unavailable, we use frozen MONET (Kim et al., 2024) embeddings, which encode clinically relevant visual concepts from dermatology images.

All intrapatient baselines use the same lesion representation within each dataset, so their differences reflect how they summarize and use the patient's other lesions rather than differences in the underlying representation. For EIDER, $Z_X$ denotes the target-lesion representation and $Z_1, \ldots, Z_n$ the reference-lesion representations in this space. These are used to estimate $\widehat{U}_n$ with Eq. 5 and compute $r = Z_X - \widehat{U}_n$.

Table 5: **Lesion representations used by EIDER's patient branch.**

| Dataset | Patient-branch representation $Z$ | Source |
|---|---|---|
| ISIC 2024 | Structured WB360 numerical measurements | Dataset-provided lesion measurements |
| SIIM–ISIC 2020 | Frozen MONET embedding | Dermoscopic image |
| Longitudinal cohort | Frozen MONET embedding | Dermoscopic image |

### D.1 EIDER Implementation

EIDER combines a lesion branch that predicts from the target lesion with a patient branch that adjusts this prediction using the patient's reference lesions. The lesion branch is trained first and then frozen while training the patient branch. Table 6 summarizes the main hyperparameters.

**Lesion branch.** The lesion branch uses a pretrained ConvNeXtV2-Tiny on RGB lesion crops. Training uses standard image augmentation and ImageNet normalization. Each fold retains all positive lesions and subsamples negatives, with inverse-frequency sampling. The model is trained with binary cross-entropy and AdamW, with model selection based on validation $\mathrm{pAUC}_{80}$. After training, the branch is frozen and provides the lesion-only logit.

**Patient reference.** Numerical features are median-imputed and standardized using statistics from the training fold. The training-fold population statistics are then used to estimate the patient reference from the reference lesions according to Eq. 5.

**Patient branch.** Let $\widehat{U}_n$ denote the patient reference and $r = Z_X - \widehat{U}_n$ the target lesion's deviation from it. Following Section 3, the patient branch receives $\widehat{U}_n$, $r$, $r^2$, and the reference size $n$. A two-hidden-layer MLP maps these inputs to an additive adjustment to the frozen lesion logit. The branch is trained on the resulting combined prediction using positive-weighted binary cross-entropy and AdamW, with early stopping on validation $\mathrm{pAUC}_{80}$.

**Prediction and ablation.** EIDER adds the patient-branch adjustment to the frozen lesion-branch logit and applies a sigmoid. The lesion-branch ablation sets the patient-branch adjustment to zero.

Table 6: EIDER hyperparameters.

| Component | Hyperparameter | Value |
|---|---|---|
| Lesion branch | Input resolution | $224 \times 224$ |
| | Negative-to-positive cap | 50:1 |
| | Learning rate | $10^{-4}$ |
| | Weight decay | $10^{-4}$ |
| | Batch size | 192 |
| | Maximum epochs | 20 |
| | Gradient clipping | 1 |
| Patient reference | Shrinkage $\kappa$ | 5 |
| Patient branch | Hidden widths | 256, 128 |
| | Dropout | 0.1 |
| | Learning rate | $10^{-3}$ |
| | Weight decay | $10^{-4}$ |
| | Batch size | 4096 |
| | Maximum epochs | 40 |
| | Gradient clipping | 1 |

### D.2 Baselines

Here, we describe the baselines used for the dermatology experiments. They include target-only models and intrapatient models that use same-patient lesions during prediction or representation learning.

#### D.2.1 TARGET-ONLY MODELS

**ResNet-18** (He et al., 2016) is a residual convolutional network trained directly on the target-lesion image.

**ConvNeXtV2** (Woo et al., 2023) is a convolutional image model trained directly on the target-lesion image.

**LightGBM** (Ke et al., 2017) is a gradient-boosted decision-tree model trained on target-lesion metadata without information from the reference lesions.

**PanDerm MLP** uses frozen 512-dimensional PanDerm (Yan et al., 2025b) embeddings of the target lesion. The classifier has one hidden layer of width 128, GELU activation, dropout 0.1, and is trained with AdamW for up to 12 epochs.

**MONET MLP** uses frozen 768-dimensional MONET (Kim et al., 2024) embeddings of the target lesion. It uses the same classifier architecture and training procedure as the PanDerm MLP.

#### D.2.2 INTRAPATIENT MODELS

**Image–LightGBM** combines image predictions with absolute and within-patient normalized lesion metadata, following the normalization strategy used in strong ISIC systems (Kurtansky et al., 2025).

**Deep Sets** (Zaheer et al., 2017) is a permutation-invariant set model that predicts the target-lesion label from the target and reference lesions represented in the same feature space as EIDER. We use hidden width 128 and dropout 0.1.

**Set Transformer** (Lee et al., 2019) uses self-attention over the target and reference lesions in the same feature space as EIDER. We use hidden width 128, dropout 0.1, four attention heads, and one set-attention layer.

**DMT-Quadruplet** (Naranpanawa et al., 2024) uses same-patient relationships during metric learning and predicts from the target lesion alone at inference. Using the ResNet backbone described in the original work did not reproduce its reported performance in our setting. We therefore use frozen MONET (Kim et al., 2024) features, which gave stronger performance, followed by layers of widths 256 and 128 that produce a normalized embedding and a supervised MLP classifier.

### D.3 METRICS

Here, we define the metrics used in the dermatology experiments.

**Partial area under the ROC curve ($\mathrm{pAUC}_{80}$).** Our primary metric measures performance in the high-sensitivity region of the ROC curve. Let $r$ denote sensitivity and $\mathrm{FPR}(r)$ the false-positive rate at the operating point with sensitivity $r$. We compute

$$\mathrm{pAUC}_{80} = \int_{0.8}^{1} [1 - \mathrm{FPR}(r)]\ dr,$$

which integrates specificity over sensitivities from $0.8$ to $1$ and has a maximum value of $0.2$.

**Patient recall at 15 ($\mathrm{PR}_{15}$).** For each patient, we rank lesions by predicted malignancy score and retain the 15 highest-ranked lesions, or all lesions when the patient has fewer than 15. $\mathrm{PR}_{15}$ is the fraction of patients with at least one positive lesion for whom at least one positive lesion appears in this retained set.

**AUROC.** Area under the receiver operating characteristic curve measures discrimination across all classification thresholds.

**Average precision (AP).** AP summarizes the precision–recall curve as the step-wise weighted mean of precision obtained as recall increases. We use average precision rather than trapezoidal interpolation of the precision–recall curve.

**Sensitivity.** Sensitivity is the fraction of positive lesions classified as positive.

**Specificity.** Specificity is the fraction of negative lesions classified as negative.

**F1 score.** F1 is the harmonic mean of precision and recall at the retained classification threshold.

Table 7: **Classical performance metrics across the three dermatology datasets.** All values are scaled by $10^2$. We report AUROC, average precision (AP), sensitivity, specificity, and F1. Higher is better for all metrics.

| | *ISIC 2024* | | | | | *SIIM–ISIC 2020* | | | | | *Longitudinal* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | AUROC | AP | Sens. | Spec. | F1 | AUROC | AP | Sens. | Spec. | F1 | AUROC | AP | Sens. | Spec. | F1 |
| ResNet-18 | 92.1 | 4.0 | 83.7 | 84.3 | 1.0 | 77.3 | 6.2 | 78.8 | 60.7 | 6.6 | 84.4 | 6.7 | **81.7** | 74.6 | 7.2 |
| ConvNeXtV2 | 90.5 | 2.5 | 78.9 | 85.2 | 1.0 | 74.6 | 5.1 | 75.2 | 60.7 | 6.4 | 82.5 | 8.1 | 75.2 | 74.7 | 6.7 |
| LightGBM | 89.2 | 1.7 | 82.2 | 83.3 | 1.0 | 64.3 | 4.4 | 49.7 | 73.9 | 6.2 | 73.5 | 3.4 | 62.5 | 74.7 | 5.6 |
| PanDerm MLP | 86.0 | 1.6 | 75.3 | 79.9 | 0.7 | 84.9 | **14.7** | 77.9 | 74.6 | 9.8 | **87.8** | 12.5 | 78.5 | 80.9 | 9.0 |
| MONET MLP | 86.7 | 1.7 | 67.4 | 88.0 | 1.1 | 84.8 | 11.9 | 80.3 | 73.0 | 9.5 | 83.4 | 9.2 | 72.4 | 82.4 | 9.0 |
| Image–LightGBM | 91.6 | 2.7 | 88.3 | 85.7 | 1.2 | 87.3 | **14.7** | 80.5 | 74.4 | 10.0 | 85.0 | 10.6 | 73.8 | **85.5** | **10.8** |
| Deep Sets | 86.9 | 2.9 | 75.6 | 81.6 | 0.8 | 83.7 | 12.4 | 75.7 | 76.1 | 10.1 | 83.1 | 9.2 | 76.4 | 80.0 | 8.4 |
| Set Transformer | 88.8 | 3.2 | 79.1 | 83.0 | 0.9 | 83.2 | 12.9 | 70.7 | **78.6** | **10.4** | 82.5 | 6.7 | 71.7 | 76.8 | 6.9 |
| DMT-Quadruplet | 87.1 | 1.6 | 76.1 | 80.4 | 0.8 | 84.8 | 8.6 | 83.7 | 71.3 | 9.4 | 83.8 | 8.5 | 79.2 | 76.7 | 7.6 |
| EIDER (lesion branch) | 89.7 | 2.4 | 81.7 | 81.6 | 0.9 | 87.7 | 12.7 | 80.8 | 76.0 | 10.6 | 83.9 | 12.5 | 74.0 | 81.0 | 8.6 |
| EIDER | **95.9** | **7.9** | **88.3** | **88.5** | **1.5** | **87.9** | 12.7 | **84.1** | 73.2 | 10.0 | 83.8 | **12.6** | 75.9 | 80.2 | 8.5 |

Classification thresholds are selected by validation Youden's $J$ statistic (i.e., the threshold maximizing sensitivity + specificity $-1$). The ranking metrics $\mathrm{pAUC}_{80}$, $\mathrm{PR}_{15}$, AUROC, and AP do not require a fixed classification threshold.

# E Additional Melanoma Results

## E.1 Additional Performance Metrics

To complement the high-sensitivity ranking metrics used in the main evaluation, we also report in Table 7 standard discrimination and classification metrics across all three dermatology datasets.

## E.2 Sensitivity to Reference Set Characteristics

The hierarchical model suggests that the value of the reference set depends on how precisely it estimates the patient reference and how much that reference differs from the population reference. We therefore examine performance as a function of reference size, heterogeneity, and distinctiveness.

**Reference set size.** Larger reference sets reduce uncertainty in the estimated patient reference. Figure 6a groups patients by the number of reference lesions. EIDER's advantage over the lesion branch increases with reference size, while the lesion branch and ResNet-18 show no corresponding trend. This is consistent with more reference lesions providing a more precise patient reference.

**Reference set heterogeneity.** Greater variation among a patient's reference lesions makes them less well summarized by a single patient reference. Figure 6b groups patients into quintiles of reference-set heterogeneity, defined as the mean within-patient variance across features ($\overline{\sigma_{0,k}}$). Performance decreases in the highest-heterogeneity group for both EIDER and the target-only models, suggesting that these patients are generally more difficult to predict rather than exposing a failure specific to the reference branch.

**Reference set distinctiveness.** A patient reference farther from the population reference provides a stronger test of whether patient reference is beneficial. Figure 6c groups patients into quintiles of patient-reference distinctiveness, measured by $\|\widehat{U}_n - \mu\|_2$. Target-only performance is comparatively stable across these groups, whereas EIDER decreases in the most distinctive group. Thus, greater departure from the population reference does not translate into a larger predictive benefit and may instead indicate patients whose reference sets are harder to model.

# F Additional Interpretability

Figure 7 extends the patient-level analysis from Section 4.5 to patients with different numbers of observed lesions. Across these examples, lesions with similar deviation magnitudes can receive dif-

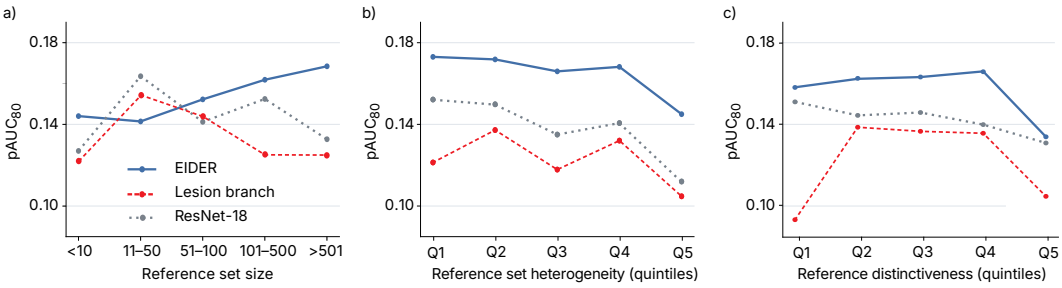


Figure 6: **Performance across reference-set characteristics.** We compare EIDER with the ablated lesion branch, and the best lesion-only model ResNet-18. (a) $\mathrm{pAUC}_{80}$ by reference size. EIDER's advantage over the lesion branch increases with the number of reference lesions. (b) $\mathrm{pAUC}_{80}$ by reference-set heterogeneity quintile $\bar{\sigma}_0$. Performance decreases for highly heterogeneous patients across both EIDER and target-only models. (c) $\mathrm{pAUC}_{80}$ by reference distinctiveness quintile $\|\widehat{U}_n - \mu\|_2$. Performance decreases for highly heterogeneous patients for EIDER, but not for target-only models.

ferent adjustments, while groups of nearby lesions can receive similar adjustments. These examples further illustrate that the patient branch uses the structure of the lesion-to-reference deviation rather than deviation magnitude alone.

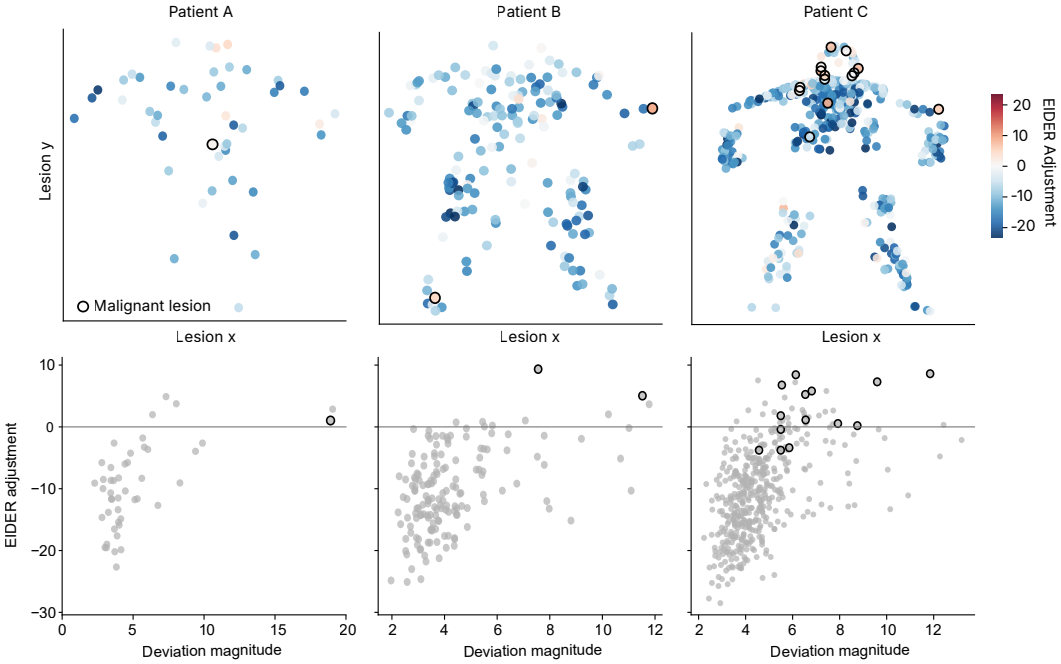


Figure 7: **Additional patient-level EIDER visualizations.** Three patients with different numbers of observed lesions are shown. **Top:** lesions at their recorded body locations, colored by the EIDER adjustment; malignant lesions are outlined. **Bottom:** EIDER adjustment against deviation magnitude $\|Z_X - \widehat{U}_n\|_2$ for the same lesions. The horizontal line indicates zero adjustment.

### F.1 Representation Analysis

We examine how subtracting the estimated patient reference changes the structure of the lesion representation. Using UMAP (McInnes et al., 2018), we project the same held-out lesions from the absolute representation $Z_X$ and from the deviation representation $Z_X - \widehat{U}_n$.

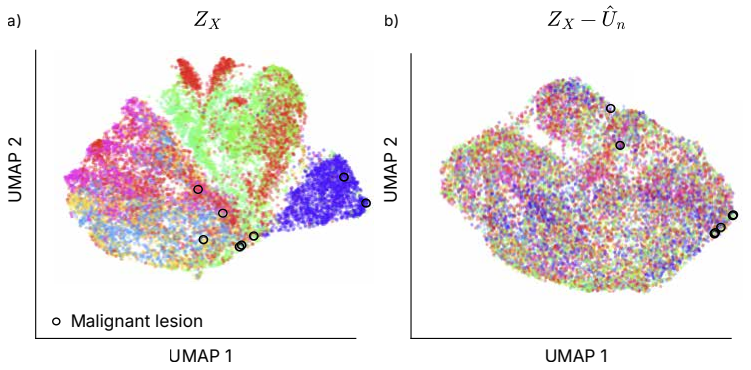


Figure 8: **Effect of the patient reference on lesion representations.** UMAP projections of the same held-out lesions from eight patients with the largest numbers of lesions. **(a)** Absolute lesion representations $Z_X$ form visible patient clusters. **(b)** After subtracting the estimated patient reference, $Z_X - \widehat{U}_n$, these clusters largely disappear and the malignant lesions shown are more peripheral in the shared representation. Points are colored by patient; malignant lesions are outlined.

Figure 8 shows clear patient clusters in the absolute representation. After subtracting the estimated patient reference, these clusters largely disappear, indicating that variation associated with differences between patient references has been reduced. The malignant lesions shown in the absolute space sometimes lie between patient clusters; in the deviation space, they appear predominantly toward the periphery of the shared representation. This is consistent with the intended role of the reference transformation: align variation that is typical for each patient while retaining deviations from that reference.

# G EEG Implementation Details

## G.1 Experiment Design

**Dataset and preprocessing.** We use CHB–MIT (Guttag, 2010), retaining one seizure-bearing recording from each of 24 subjects and the same 16 bipolar temporal and parasagittal channels for all recordings. We divide each recording into non-overlapping eight-second windows and label a window positive when it overlaps an annotated seizure interval. The resulting cohort contains 16,199 windows, including 220 seizure windows from 28 seizure events and 35.51 hours of non-seizure EEG.

Waveform models operate directly on the 16-channel signals. Feature-based models use 80-dimensional representations containing log power in the delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz) bands for each channel. As in the melanoma setting, preprocessing statistics and population-reference statistics are estimated from the training subjects only.

**Training.** We use five subject-disjoint folds, with separate training, validation, and test subjects in each fold. Hyperparameters are selected by validation AP.

**EIDER.** We apply the same patient-reference formulation to EEG, but adapt the prediction model to the band-power representation. For each target window, the reference set contains only earlier unlabeled windows from the same subject and recording; the target and all future windows are excluded. We estimate a partially pooled reference with $\kappa = 5$ from the 80-dimensional band-power representations using Eq. 5. The target is represented by its band-power features and signed deviation from this reference, and a LightGBM classifier predicts the seizure label from these inputs. The target-only ablation removes the reference-derived deviation features while keeping the same model and tuning procedure.

**Feature-based baselines.** Logistic regression, LightGBM, and CatBoost operate on the same 80-dimensional band-power representation. Logistic regression uses class-balanced loss with $C \in \{0.1, 1, 10\}$. For LightGBM, we compare 300 trees with learning rate 0.03 and 15 leaves against 500 trees with learning rate 0.02 and 31 leaves. For CatBoost, we compare 400 trees of depth 6 with learning rate 0.03 against 600 trees of depth 7 with learning rate 0.02. Missing values are median-imputed and features are standardized within each training fold where required.

**Waveform baselines.** EEGNet (Lawhern et al., 2018) uses the EEGNet-8,2 architecture and is trained from scratch on the 16-channel waveforms. BIOT (Yang et al., 2023) uses its released pretrained encoder after resampling each window to 1,600 samples and applying robust per-channel normalization. CBraMod (Wang et al., 2025) uses its released pretrained checkpoint with eight temporal patches of 200 samples. For BIOT and CBraMod, we evaluate both a frozen encoder with a trained classification head and full fine-tuning. All waveform models use class-weighted binary cross-entropy, AdamW with weight decay $10^{-4}$, batch size 32, gradient clipping at 1, and early stopping on validation AP over at most 20 epochs.

All baselines are retrained and evaluated under the same patient-disjoint protocol used for EIDER. Published CHB-MIT results for these models use different evaluation protocols and are therefore not directly comparable; we use the reproduced models only for within-protocol comparison.

**Metrics.** Similar to the melanoma setting, we report window-level AUROC and AP together with sensitivity, specificity, accuracy, and F1 at a threshold selected on the validation set by Youden's $J$ statistic. Uncertainty intervals use subject-level bootstrap resampling.

Following prior CHB–MIT seizure-detection evaluations that report seizure-event sensitivity together with false-alarm rate per hour (Wang et al., 2025), we additionally report event-level metrics. A seizure event is detected when at least one overlapping window is flagged. Consecutive flagged nonseizure windows within the same recording are merged into one false-alarm event, and false alarms are reported per hour of nonseizure EEG. For the fixed 1% alarm-budget analysis in the main text, we instead select on the validation subjects the threshold that flags 1% of windows and apply it unchanged to the test subjects.

### G.2 EEG Additional Results

We complement the main metrics with more classical classification metrics in Table 8.

Table 8: **EEG seizure-detection results.** Values are bootstrap mean ± conservative symmetric 95% confidence-interval radius over 2,000 patient-level bootstrap resamples. All values are reported in percentage points.

| | Method | AUROC | AP | Sens. | Spec. | F1 |
|---|---|---|---|---|---|---|
| Band-power inputs | Logistic regression | 69.6 ± 9.5 | 7.7 ± 6.3 | 66.6 ± 16.8 | 56.9 ± 17.1 | 4.3 ± 3.1 |
| | LightGBM | 68.0 ± 13.2 | 8.4 ± 19.7 | **67.1 ± 15.4** | 64.3 ± 17.4 | 5.3 ± 5.1 |
| | CatBoost | 70.8 ± 15.0 | 9.5 ± 26.1 | 66.2 ± 15.7 | 61.0 ± 18.5 | 4.8 ± 5.1 |
| Waveform inputs | EEGNet | 66.4 ± 7.5 | 7.9 ± 9.3 | 52.3 ± 14.0 | 78.3 ± 9.9 | 6.4 ± 4.5 |
| | BIOT | 61.5 ± 12.9 | 4.8 ± 11.9 | 60.9 ± 14.4 | 50.1 ± 18.9 | 3.4 ± 2.5 |
| | CBraMod | 59.6 ± 10.9 | 8.6 ± 12.6 | 41.7 ± 15.0 | 82.0 ± 14.1 | 7.1 ± 13.9 |
| EIDER | Target branch | 68.0 ± 13.2 | 8.4 ± 19.7 | 67.1 ± 15.4 | 64.3 ± 17.4 | 5.3 ± 5.1 |
| | Full | **84.0 ± 6.2** | **36.2 ± 14.7** | 67.0 ± 16.2 | **89.6 ± 8.6** | **16.4 ± 16.3** |