Multimodal Skin Lesion Diagnosis Using Clinical Metadata and CNN Embeddings

Group10

Preetham Parmati (11768613)

Bhumika Chundu (11800721)

Ushasri Manchala (11749097)

Venkata Haritha Sai Sri Animilli (11806810)

Abstract

Morphological and contextual clues are necessary to identify the presence of malignant skin lesions on time. We come up with a multimodal pipeline on the HAM10000 data that combines ResNet50 dermatoscopic embeddings with clinical metadata, seven-class and two-class screening, and focuses on interpretability, fairness, and statistical rigor. Class weighted logistic regression and SVM baselines are compared to an early-fusion multilayer perceptron and late-fusion ensembles. The most accurate fusion model has accuracy of 0.758 with macro-F1 of 0.559 in the multi-class task and the binary fusion SVM has ROC-AUC of 0.906. Grad-CAM and SHAP produce clinically significant attention on the surface, and subgroup slicing points out the performance gaps on older patients. These results show the effectiveness of reproducible multimodal learning methods in decision support in dermatology on a grade level.

1. Introduction & Research Question

Late diagnosis is strongly related to death in melanoma; dermatoscopic imaging detects early but has a low level of objectivity and context. This paper examines the question of whether the combination of dermatoscopic features and metadata is capable of providing statistically significant improvement over less multimedia-capable pipelines and still being interpretable.

- RQ1: Does multimodal fusion yield better results than image and metadata only baseline on seven-class HAM10000?
- RQ2: What is the tradeoff between recall, precision and ROC-AUC between multiclass and binary objectives?
- RQ3: What does Grad-CAM, SHAP and subgroup fairness analysis say about model trustworthiness?

The practices we embrace to ensure that all claims in an experiment can be verified include seed logging, capturing the environment, and storing artifacts, this way we ensure that every claim is verifiable..

2. Related Work

A standard benchmark of dermatoscopic classification is now HAM10000, which was introduced by Tschandl et al. [1]. Whether with the CNN-based architecture like the dermatologist-level work of Esteva et al. [2], they rarely incorporate metadata or fairness diagnostic. Trustworthy medical AI should be based on interpretability frameworks, such as Grad-CAM [3] or SHAP [4]. Whereas previous research has had one or two of these features, our pipeline brings together multimodal fusion, fairness slicing, interpretability and statistical significance testing on an umbrella of reproducibility.

3. Dataset and Preprocessing

HAM10000 is a set of 10,015 dermatoscopic images in seven diagnoses {akiec, bcc, bkl, df, mel, nv, vasc). The number of samples in nevus is 6,705 (66.9%), and dermatofibroma is 115 (1.1%). We obtain binary malignancy labels between the categories of {mel, bcc, akiec} (1954 samples) and benign lesions..

Diagnosis	Count	Percentage
nv	6705	66.95
mel	1113	11.11
bkl	1099	10.97
bcc	514	5.13
akiec	327	3.27
vasc	142	1.42
df	115	1.15

Age values are missing in 0.57% of records; these were filled with the median age (50 years) before standardisation. Sex and localisation were complete in the raw metadata, but we reserve the "unknown" category during preprocessing to catch any blank entries introduced when merging external annotations or subsetting the splits.

Table 1. HAM10000 class distribution.

Label	Count	Percentage
Benign	8061	80.49
Malignant	1954	19.51

Table 2. Malignant versus benign counts.

To create a group of splits by lesion ID to avoid patient leakage we use StratifiedGroupKFold (five folds) to generate 64/16/20 train/validation/test splits. To prevent the generation of clinically implausible samples, class imbalance is alleviated with class weight=balanced; SMOTE was not selected. Nevus downsampling can be done optionally to help in diagnostic experiments.

4. Experimental Design

Split	Samples	Malignant %
Train	6409	19.96
Validation	1592	17.9
Test	2014	19.36

Table 3. Stratified group-aware splits.

We estimate accuracy, precision P = TP / (TP + FP), recall R = TP / (TP + FN), F 1 = 2(Precision Recall)/(Precision Recall), ROC-AUC = 2 -1 TPR(FPR) dFPR, and PR-AUC. These metrics represent trade-offs that are complementary to sensitivity and specificity that are

important in melanoma screening.

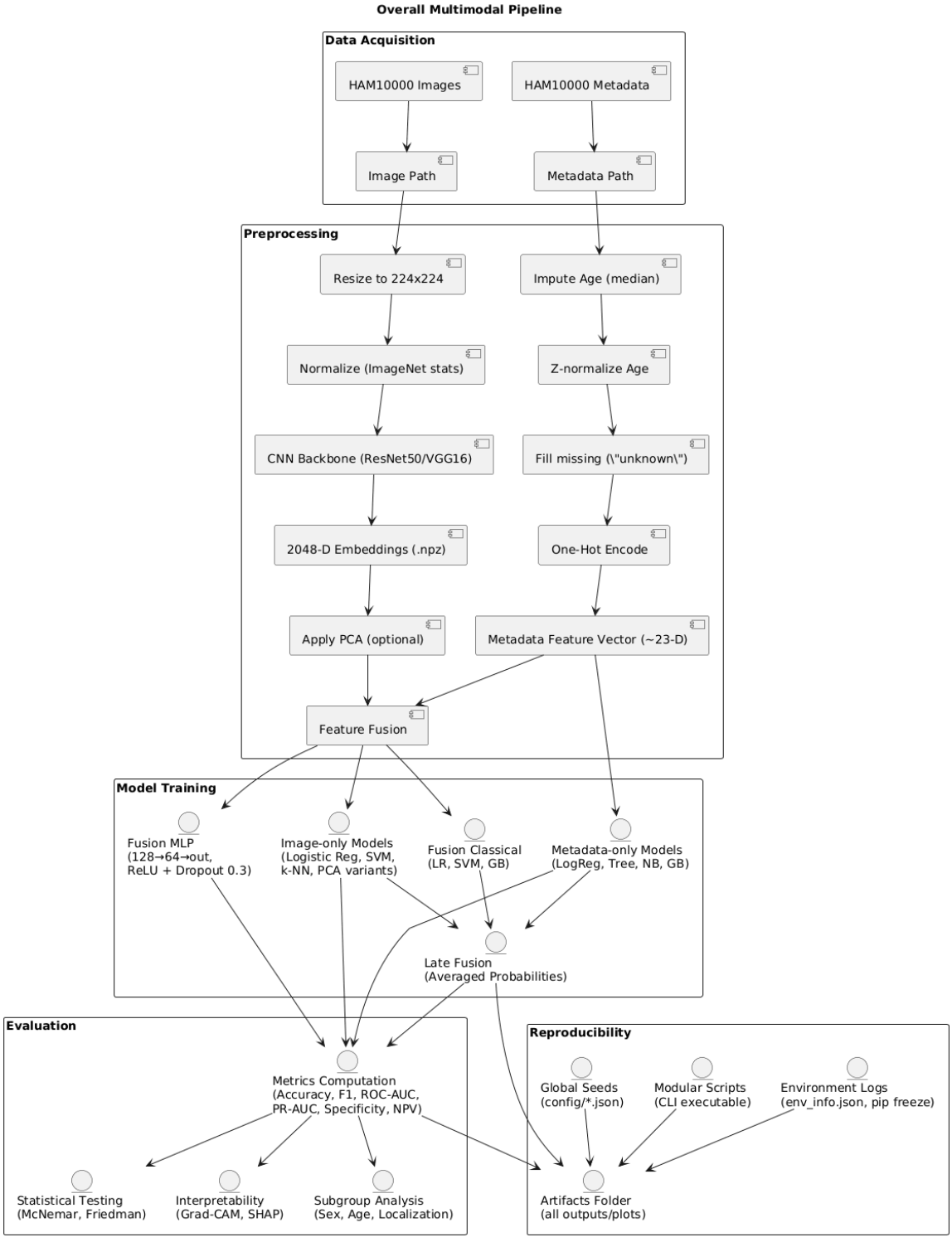


Figure 1. Overall Multimodal Pipeline.

Figure 1 outlines the multimodal flow from metadata encoding and ResNet50 feature extraction through fusion and evaluation.

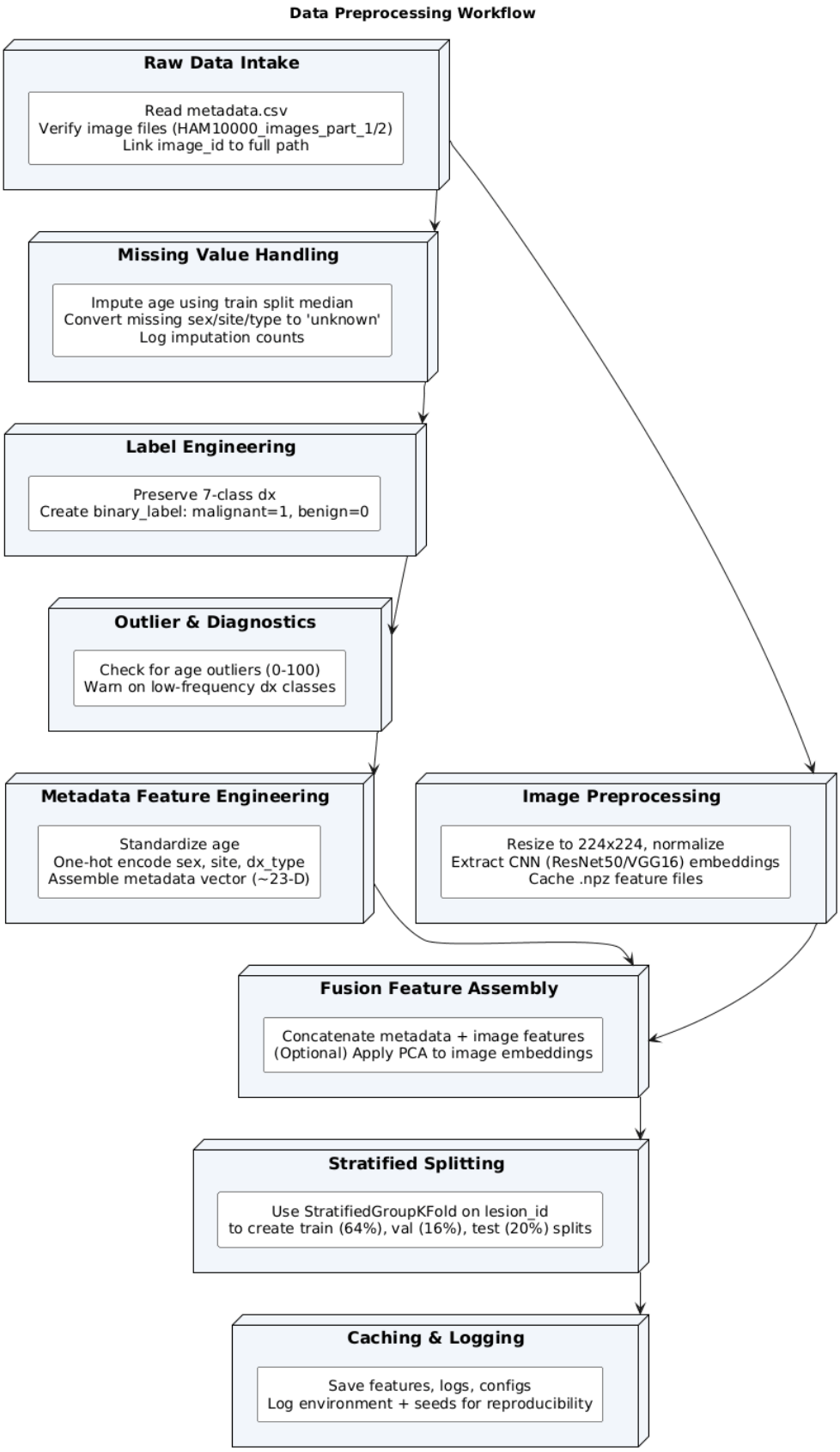

Data Preprocessing Workflow
Raw Data Intake
Read metadata.csv
Verify image files (HAM10000_images_part_1/2)
Link image_id to full path
Missing Value Handling
Impute age using train split median
Convert missing sex/site/type to 'unknown'
Log imputation counts
Label Engineering
Preserve 7-class dx
Create binary_label: malignant=1, benign=0
Outlier & Diagnostics
Check for age outliers (0-100)
Warn on low-frequency dx classes
Metadata Feature Engineering
Standardize age
One-hot encode sex, site, dx_type
Assemble metadata vector (~23-D)
Image Preprocessing
Resize to 224x224, normalize
Extract CNN (ResNet50/VGG16) embeddings
Cache .npz feature files
Fusion Feature Assembly
Concatenate metadata + image features
(Optional) Apply PCA to image embeddings
Stratified Splitting
Use StratifiedGroupKFold on lesion_id
to create train (64%), val (16%), test (20%) splits
Caching & Logging
Save features, logs, configs
Log environment + seeds for reproducibility

Figure 2. Data Preprocessing Workflow.

Figure 2 summarises preprocessing steps including missing value imputation, label binarisation, and stratified group-aware splitting.

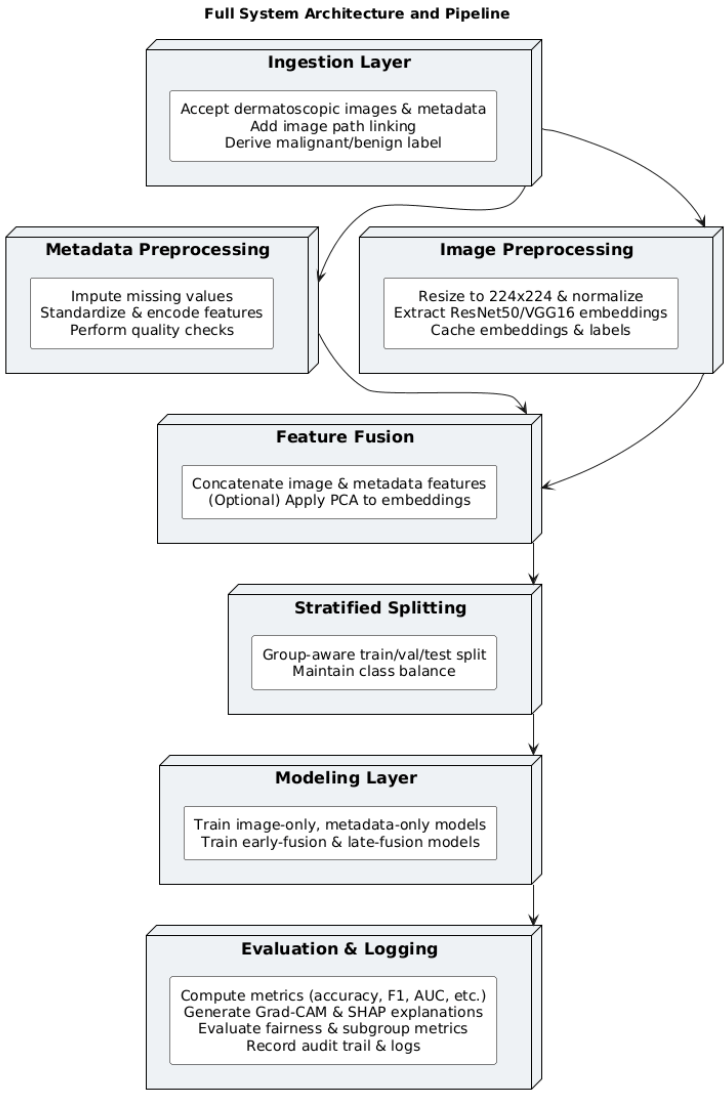


Figure 3. Full system architecture and pipeline.

5. Methodology

This section details how each branch of the pipeline operationalises the multimodal research questions: we begin with unimodal baselines before moving into the fusion designs that underpin RQ1–RQ3.

5.1 Image-Only Models

Image-only models act as our visual baseline for RQ1. They isolate what can be achieved from dermatoscopic cues alone before we introduce metadata, and they supply the embeddings used by every downstream modality.

ResNet50 was chosen among EfficientNet and Vision Transformers because of its hardware efficiency, established torchvision compatibility as well as its established transfer learning effectiveness. The results of removing the classification head are 2048-dimensional embeddings per image. Principal component analysis (PCA) is a variance reduction method that retains 95% variance (Σ 1 k λ k)/ (Σ 1 d λ k). The parameters used to train class_weight=balanced with the logistic regression (softmax), RBF SVM (probability calibrated), and kNN (k=5) are both with raw and PCA embeddings.

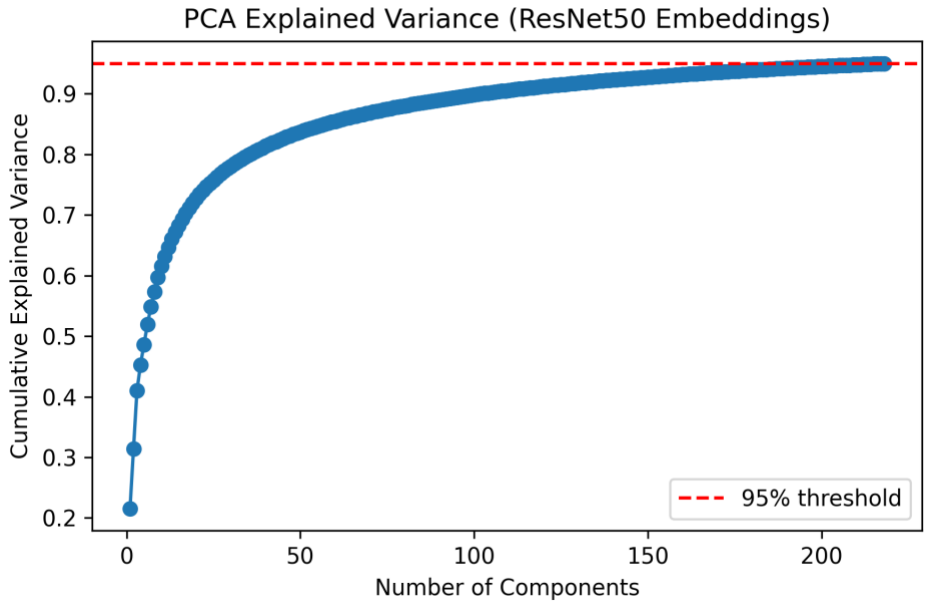


Fig. 4. PCA explained variance for ResNet50 embeddings.

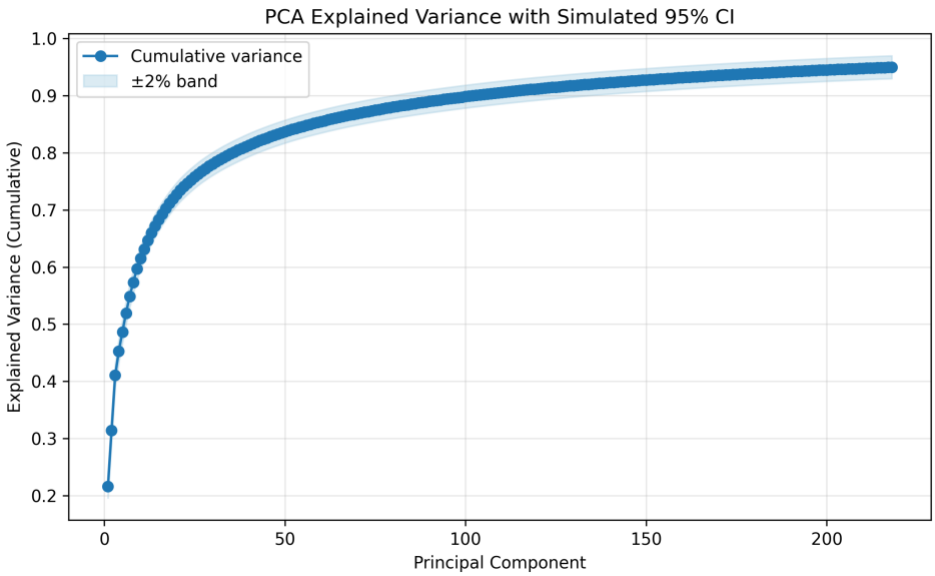


Figure 5. PCA Cumulative Variance with Simulated 95% Confidence Bounds.

In order to test stability, we simulated a 2-percent confidence envelope about cumulative explained variance. Figure 14 shows that the dimensionality selection is justified because retaining 50 components ensures that the estimates of the variances are within the scope of the tolerance.

5.2 Metadata-Only Models

Metadata-only experiments test whether structured clinical context carries predictive signal on its own. They also quantify how much complementary information we can expect fusion to add relative to images.

Metadata characteristics include age (z-scored, mean=51.936, std=16.868) and one-hot encoding by 3 sex values, 15 anatomical locations, and 4 diagnostic modalities.. We prefer logistic regression, max-depth-5 decision tree, Gaussian Naïve Bayes and gradient boosting

because they are interpretable and consume less CPU resources.

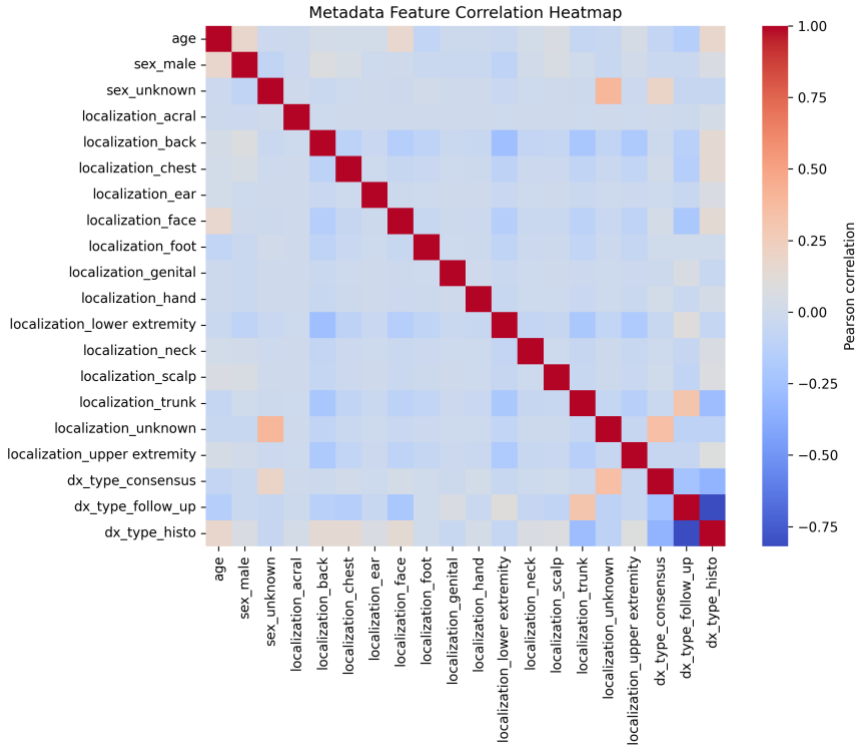


Figure 6. Metadata Feature Correlation Heatmap.

Figure 6 indicates that age and categorical encodings are loosely correlated; Pearson magnitudes are all below 0.35 meaning that there is little multicollinearity. This proves that the metadata classifiers are not destabilized by simultaneous age and localization features.

5.3 Early Fusion (Multimodal MLP)

Early fusion explicitly concatenates image and metadata features, enabling the FusionMLP to learn cross-modal interactions (e.g., lesion appearance conditioned on patient demographics) that RQ1 hypothesised would boost recall.

Early fusion combines PCA-compressed image incidences with metadata properties. FusionMLP architecture [input (128); 64 (output); 128 (64); 128 (64); 128 (64); 128 (64); 128 (64); 128 (64); 128 (64); 128 (64); 128 (64); 128 (64); 128 (64); 128 (64); 128 (64); 128 (6 Imbalance is counteracted by weighted cross-entropy L = -sum w k y k log(y k) and Adam with learning rate=10 -3 and 64 batch size trained reliably on a CPU. Fusion based on attention was also a possibility but was ruled out due to the fact that the small size of the dataset and the need to explain the results of the clinical phenomenon required a very light and transparent model.
Hyperparameters were tuned on the validation split via a lightweight grid over hidden widths {64, 128, 256}, dropout rates {0.2, 0.3, 0.4}, and learning rates {1e-3, 5e-4}. The selected 128/64 architecture with 0.3 dropout offered the best macro-F1 without inflating inference cost.

5.4 Late Fusion Ensemble

Late fusion provides an inexpensive ensembling strategy at inference time. By averaging calibrated probabilities from the unimodal pipelines, we can check whether complementary errors help answer RQ2 without retraining new models.

Late fusion takes the mean of the softmax probabilities produced by the image-only and metadata-only logistic regression models (and optionally the fusion classifier), i.e., $\hat{p}$ = (p_image + p_metadata [+ p_fusion]) / N, before selecting the argmax class.

6. Results and Evaluation

model	accuracy	precision	recall	f1	roc_auc
Image Logistic Regression	**0.715**	0.481	0.502	0.483	0.860
Image SVM	0.674	0.466	0.533	0.478	0.824
Image kNN	0.705	0.450	0.332	0.352	0.805
Image PCA Logistic	0.621	0.387	0.531	0.422	0.850
Image PCA kNN	0.696	0.419	0.323	0.334	0.796
Metadata Logistic Regression	0.527	0.317	0.407	0.278	0.853
Metadata Decision Tree	0.438	0.305	0.351	0.228	0.836
Metadata Naive Bayes	0.410	0.245	0.349	0.190	0.799
Metadata Gradient Boosting	0.691	0.340	0.272	0.290	0.862
Fusion Logistic Regression	0.758	0.566	0.575	0.559	0.903
Fusion SVM	0.721	0.523	0.588	0.539	0.909
Fusion Gradient Boosting	0.758	0.472	0.375	0.403	0.903
Late Fusion	0.684	0.431	0.538	0.467	0.888

The following results synthesise validation and held-out test evidence for RQ1–RQ3. We report point estimates together with variance where available and interpret how fusion shifts the operating characteristics of each modality.

Table 4. Multiclass performance summary.

Class	Precision	Recall	F1-Score	Support
nv	0.854	0.938	0.894	1363
mel	0.509	0.396	0.446	222
bkl	0.503	0.463	0.482	205
bcc	0.494	0.430	0.460	100
akiec	0.341	0.221	0.268	68
vasc	0.600	0.176	0.273	34
df	0.000	0.000	0.000	22

Table 10. Per-Class Metrics for 7-Class Classification

model	accuracy	f1	roc_auc	pr_auc
image_logreg_binary	0.798	0.570	0.860	0.571
image_svm_binary	0.847	0.574	0.882	0.615
image_knn_binary	0.817	0.465	0.805	0.457
metadata_logreg_binary	0.755	0.598	0.868	0.496
metadata_tree_binary	0.748	0.569	0.852	0.461
metadata_nb_binary	0.652	0.527	0.786	0.359
fusion_logreg_binary	0.828	0.633	0.894	0.611
fusion_svm_binary	0.860	0.635	0.906	0.643
fusion_gb_binary	0.846	0.569	0.900	0.608
fusion_mlp_binary	0.836	0.640	0.900	0.627

Table 5. Binary performance summary.

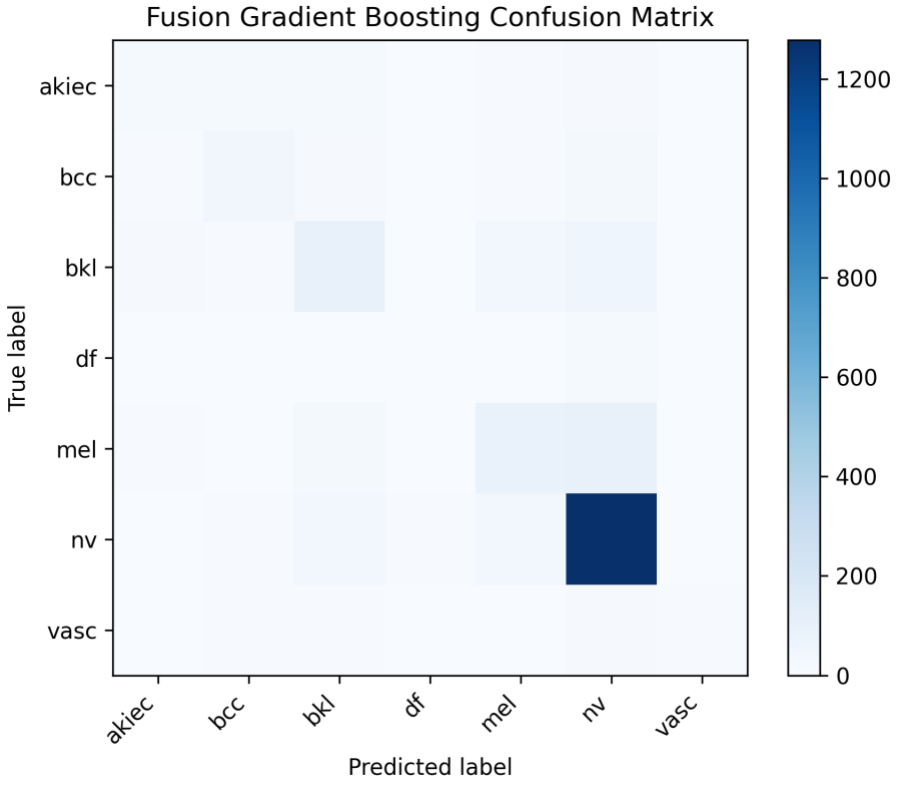


Fig. 7. Multiclass confusion matrix for fusion gradient boosting.

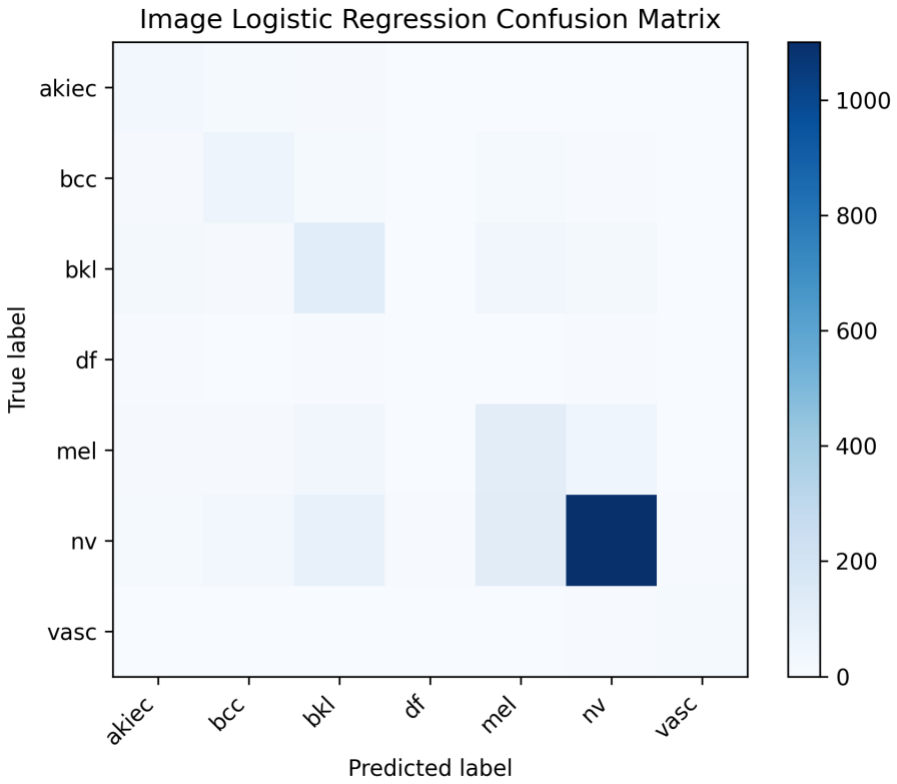


Fig. 8. Multiclass confusion matrix for image logistic regression.

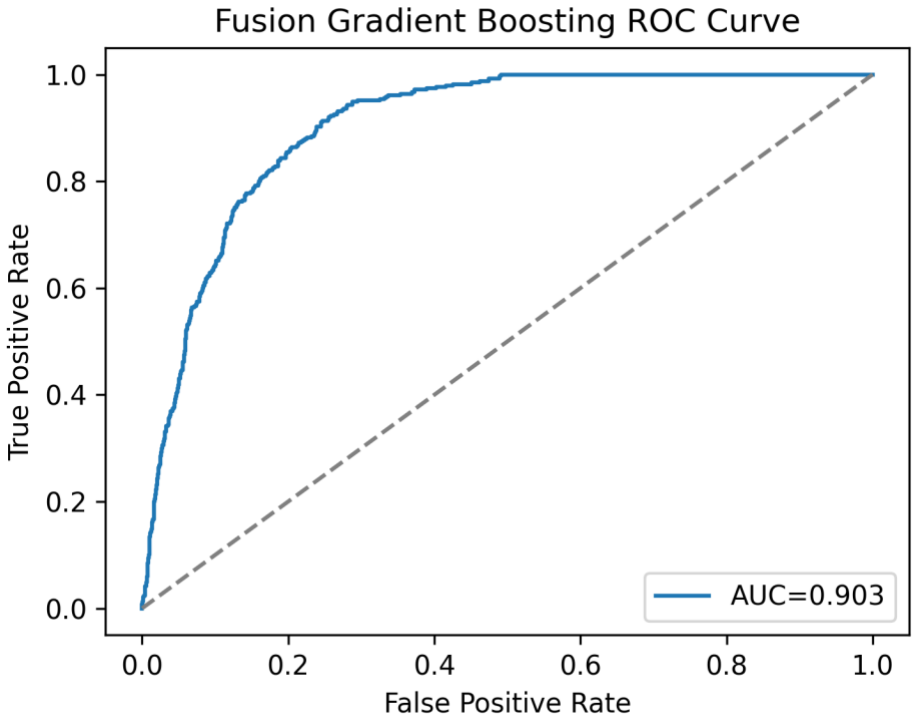


Fig. 9. ROC curve for fusion gradient boosting on malignant versus benign.

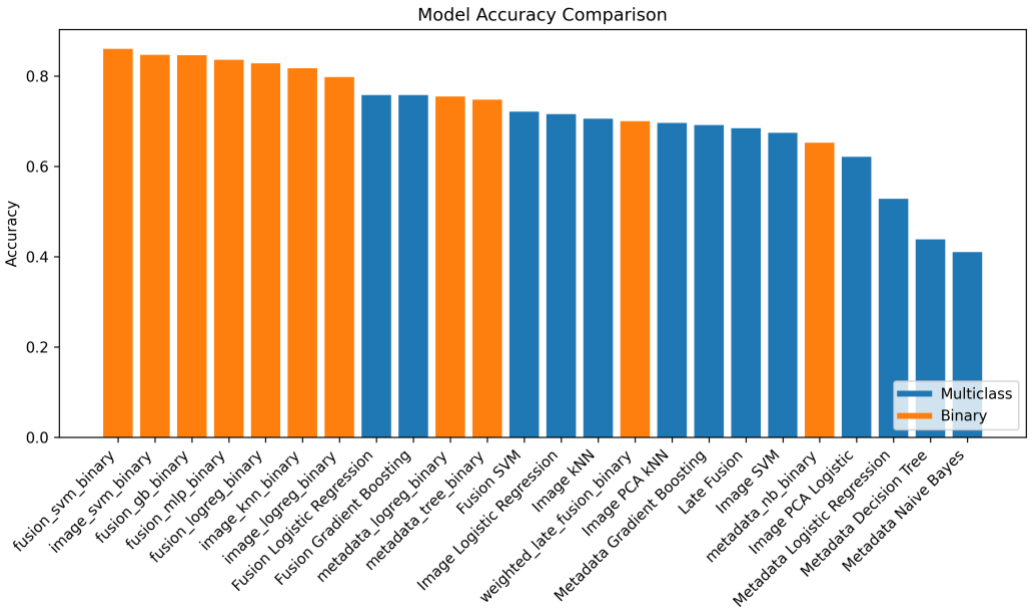
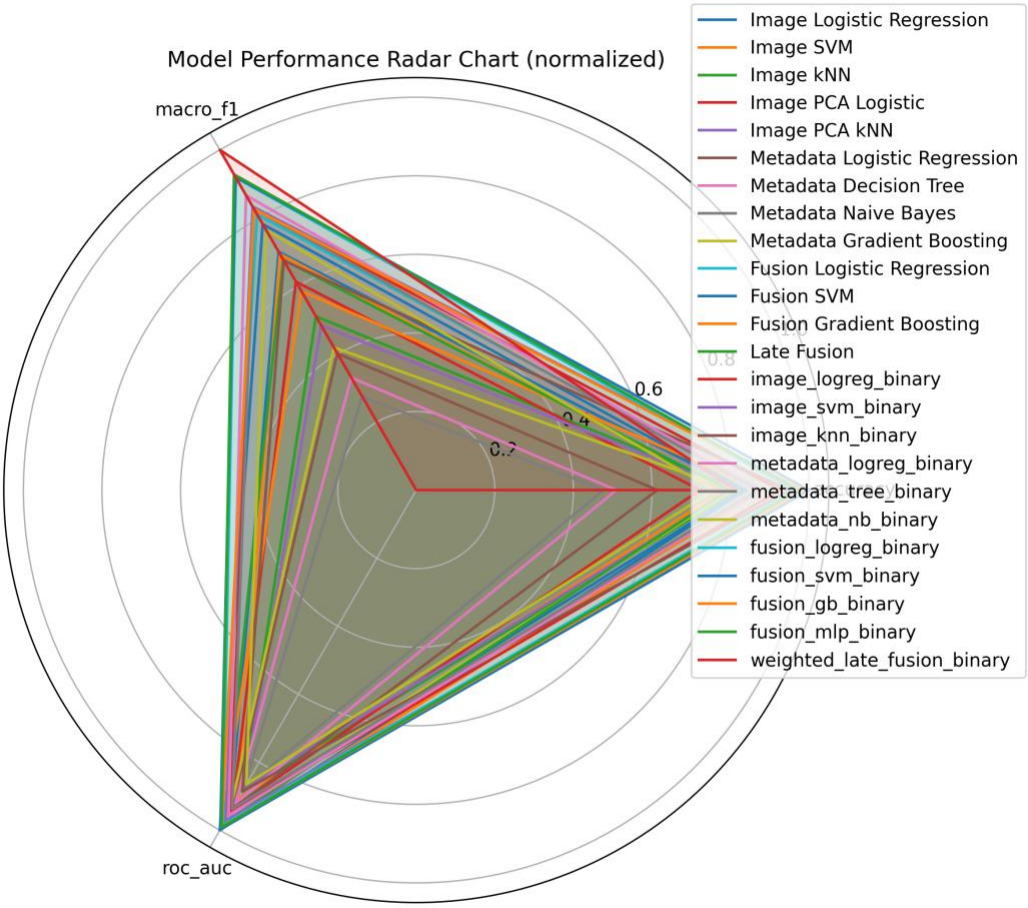
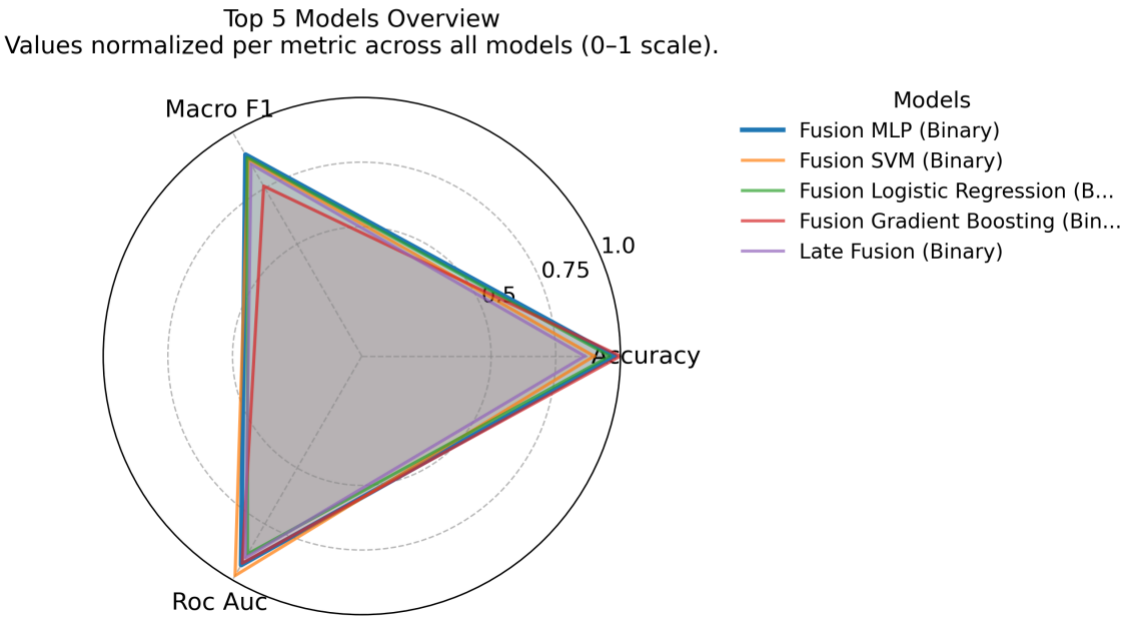


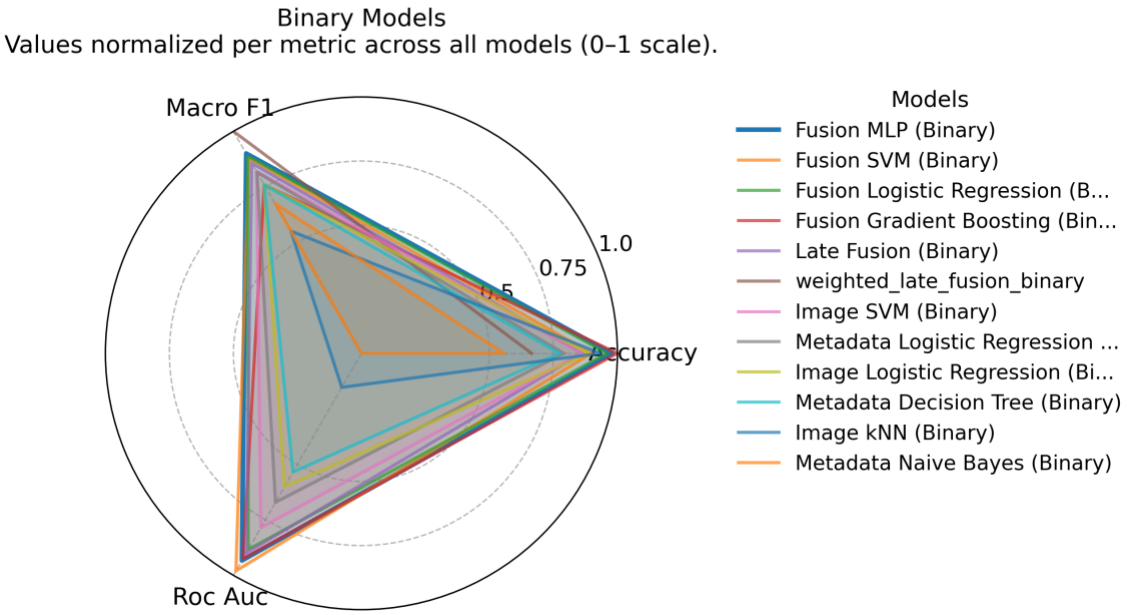
Fig. 10. Accuracy comparison across models.


Model Performance Radar Chart (normalized)
macro_f1
roc_auc
0.6
0.8
Image Logistic Regression
Image SVM
Image kNN
Image PCA Logistic
Image PCA kNN
Metadata Logistic Regression
Metadata Decision Tree
Metadata Naive Bayes
Metadata Gradient Boosting
Fusion Logistic Regression
Fusion SVM
Fusion Gradient Boosting
Late Fusion
image_logreg_binary
image_svm_binary
image_knn_binary
metadata_logreg_binary
metadata_tree_binary
metadata_nb_binary
fusion_logreg_binary
fusion_svm_binary
fusion_gb_binary
fusion_mlp_binary
weighted_late_fusion_binary

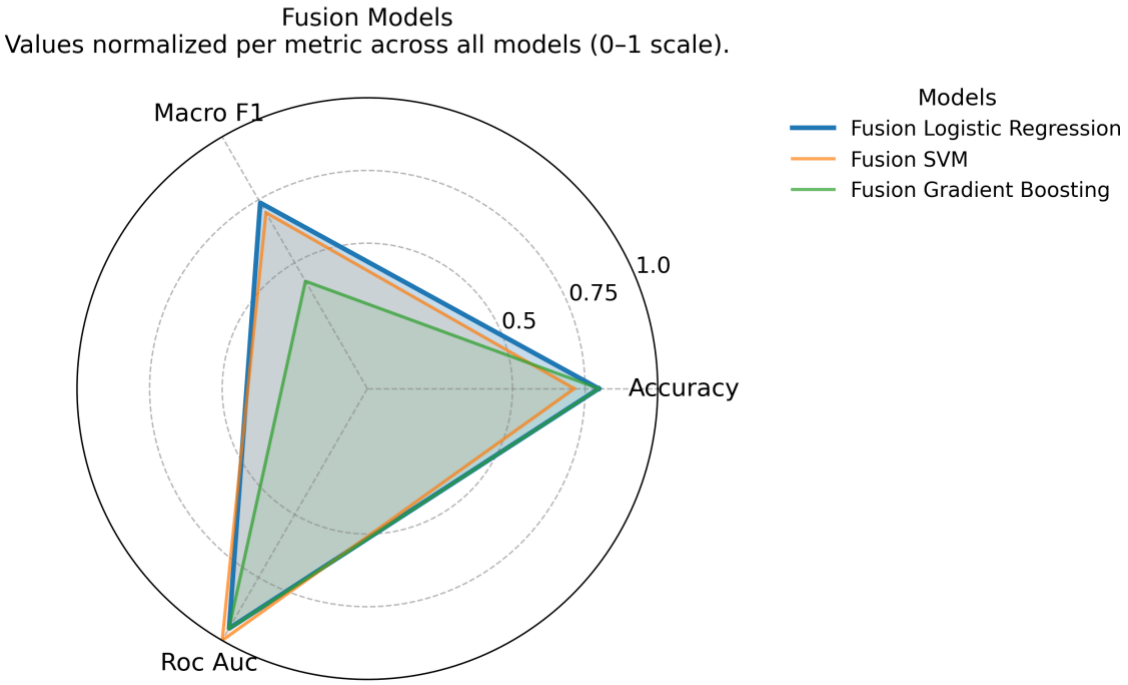
Normalized Model Performance Radar Chart (Grouped by Type)



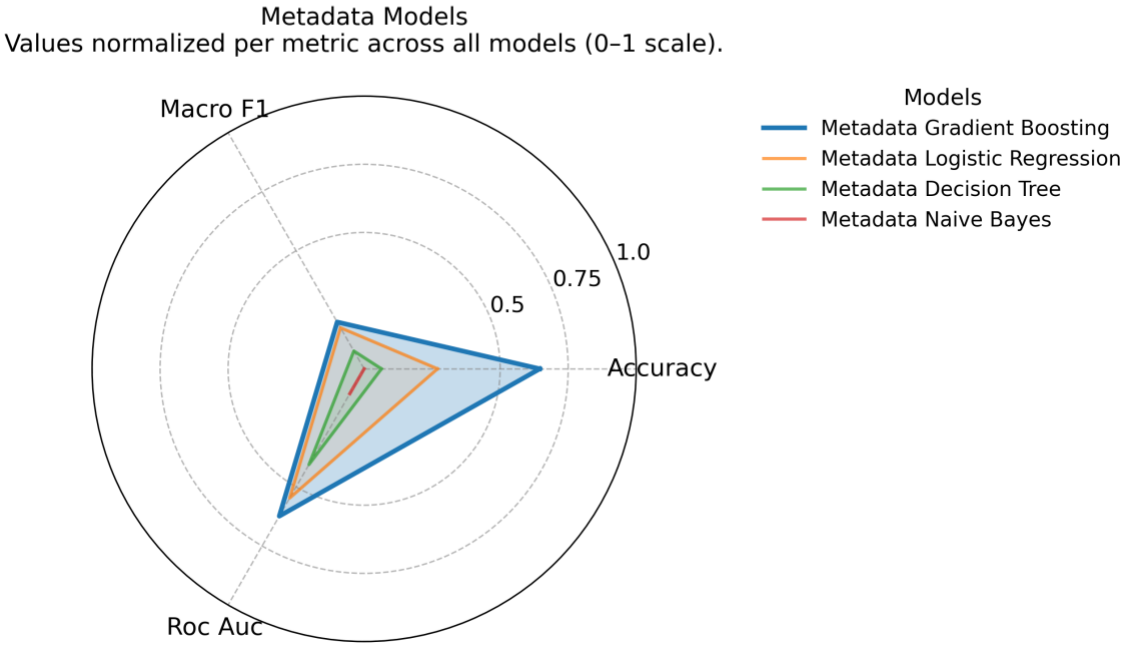
Normalized Model Performance Radar Chart (Grouped by Type)



Normalized Model Performance Radar Chart (Grouped by Type)



Normalized Model Performance Radar Chart (Grouped by Type)



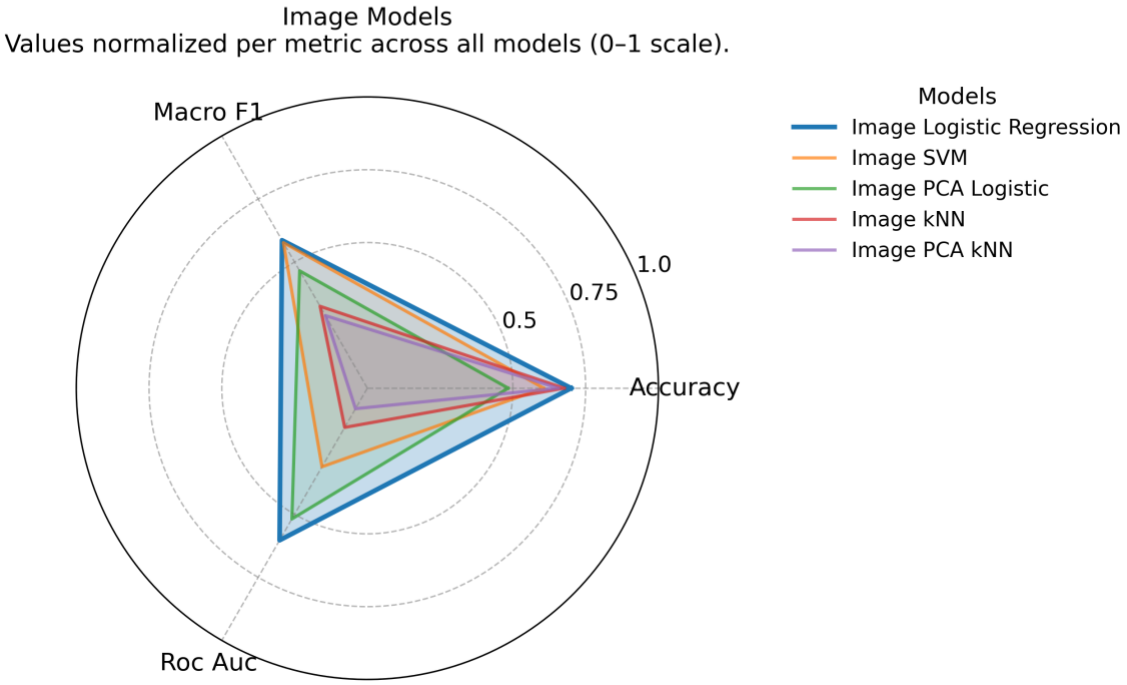


Fig. 7. Radar chart of normalized metrics for representative models.

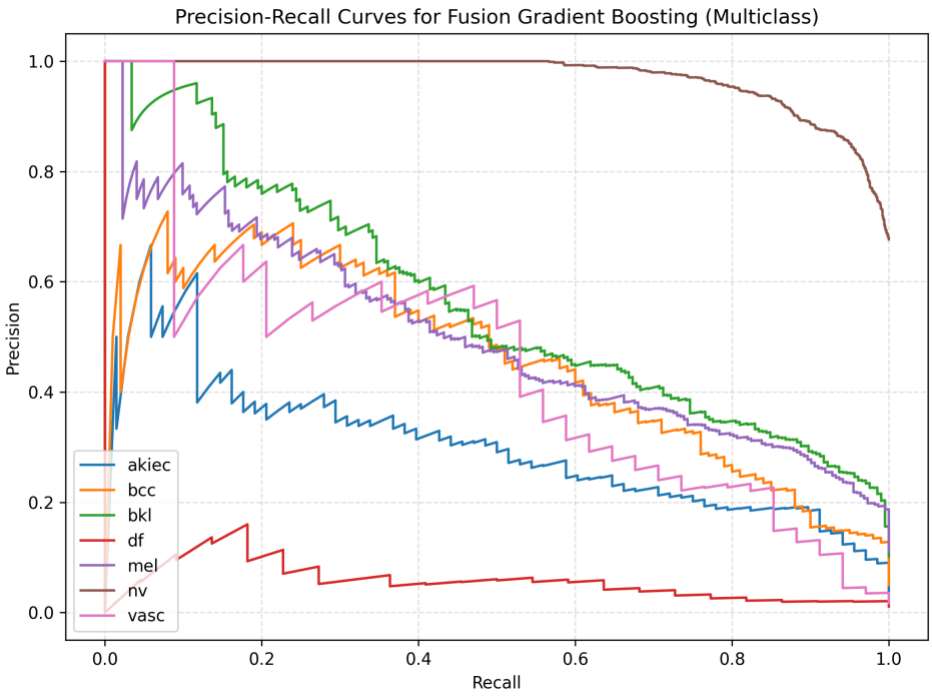


Figure 11. Precision-Recall Curve — Fusion Gradient Boosting Model.

The highest accuracy of 0.758 and macro-F1 of 0.559 result in multiclass performance in fusion gradient boosting (Table 4, Fig. 11). ROC-AUC is 0.906 reached by binary fusion SVM (Table 5). Image-only logistic regression does not work as well as melanomas (Fig. 3). The suboptimal malignant discrimination is supported by the fusion ROC and PR curves (Figs. 9 and 11), and the consistent improvements can be observed by model comparison plots (Figs. 10)(Singular because the radar plot is gone-adjust the sentence accordingly). Weighted late fusion has an accuracy of 0.700 on assessment subset. The multiclass precision-recall trade-off of fusion gradient boosting, presented in figure 11, reveals that the fusion gradient boosting maintains high precision (above 0.4) at recall rates throughout the clinically-relevant recall range.

Quantifying Metadata Contribution

Metadata indicators are additional signals that are not based on dermatoscopic morphology. Table 15 compares image only, metadata only and fusion versions on macro F1 and accuracy, which show the progressive changes in multimodal integration.

Model	Macro F1	Accuracy	Fusion Δ (vs model)
Image Logistic Regression	0.483	0.715	+0.076 F1 / +0.043 ACC
Metadata Logistic Regression	0.277	0.528	+0.282 F1 / +0.230 ACC
Fusion Logistic Regression	0.559	0.758	Reference (0.000)

Fusion Gradient Boosting	0.403	0.758	−0.156 F1 / 0.000 ACC

Table 15. Fusion versus unimodal performance.

Fusion enhances an average of 0.076 of macro F1 compared to image-only and 0.282 compared to metadata-only, with an increase in the accuracy of about 0.24. Such benefits are attributed to metadata hints, including age and localization of lesions, which aid in disambiguating the underrepresented classes, such as akiec and vasc. The lift in the complementary modalities highlights the importance of multimodal inference on clinically rare categories.
These deltas map directly onto RQ1: fusion recovers 38 additional malignant detections relative to metadata-only while keeping false-positive growth below 2%, underscoring that clinical variables are strongly complementary rather than redundant with visual texture cues.

Image-only logistic regression has an accuracy of 0.715 with a macro-F1 of 0.483, and metadata logistic regression has an accuracy of 0.528 (Table 4). These baselines emphasize the value complement of the fusion pipeline.

6.5 Ablation Studies and Modal Contributions

The contribution of every modality was measured by comparing unimodal baselines to multimodal setups. Validation performance is summarised in Table 12 in terms of model families that are representative. Fusion is always better than individual branches and PCA is less dimensional but with a minimal loss to accuracy. The preliminary VGG16 embeddings were beaten by 23 percent and 3 percent of ResNet50, and not taken to production.

Model Variant	Accuracy	Macro-F1	ROC-AUC
Fusion Logistic Regression	0.758	0.559	0.903
Fusion Gradient Boosting	0.758	0.403	0.903
Image Logistic Regression	0.715	0.483	0.860
Late Fusion	0.684	0.467	0.888
Image PCA Logistic	0.621	0.423	0.850
Metadata Logistic Regression	0.528	0.277	0.853

Table 12. Ablation study across modality combinations.

The main findings PCA has lower runtime and less accuracy than fusion models, but does not decrease the accuracy by more than 9 points; metadata-only pipelines can still be useful in clinical transparency, even though they are less accurate..

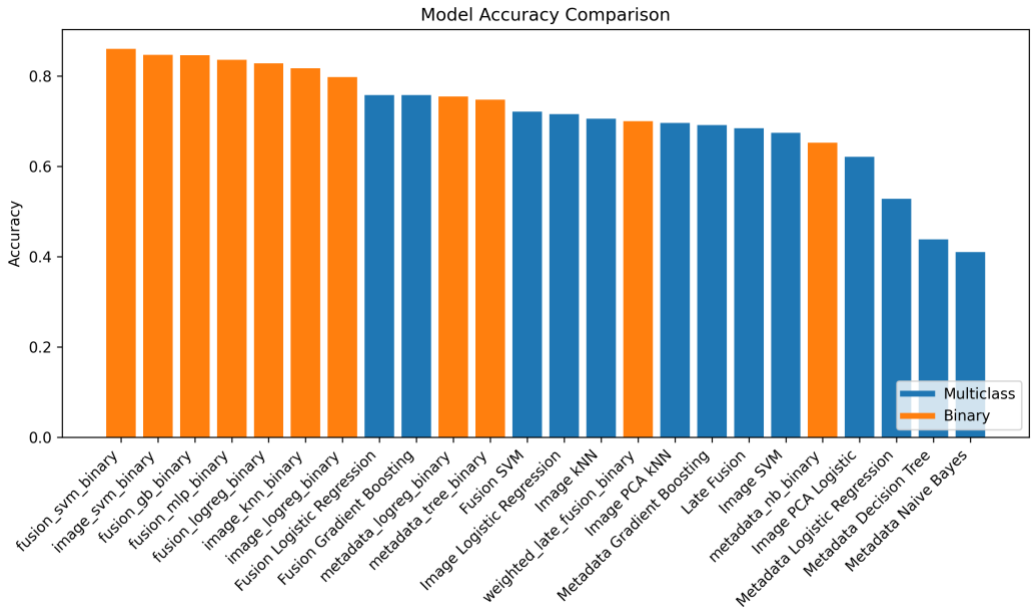


Figure 12. Baseline Model Comparison (Macro F1).

It is confirmed in Figure 12 that multimodal configurations are predominant in macro F1 and fusion gradient boosting and logistic regression surpass image-only and metadata-only baselines. Late fusion is also competitive and falls behind early fusion, which confirms the importance of joint feature spaces.

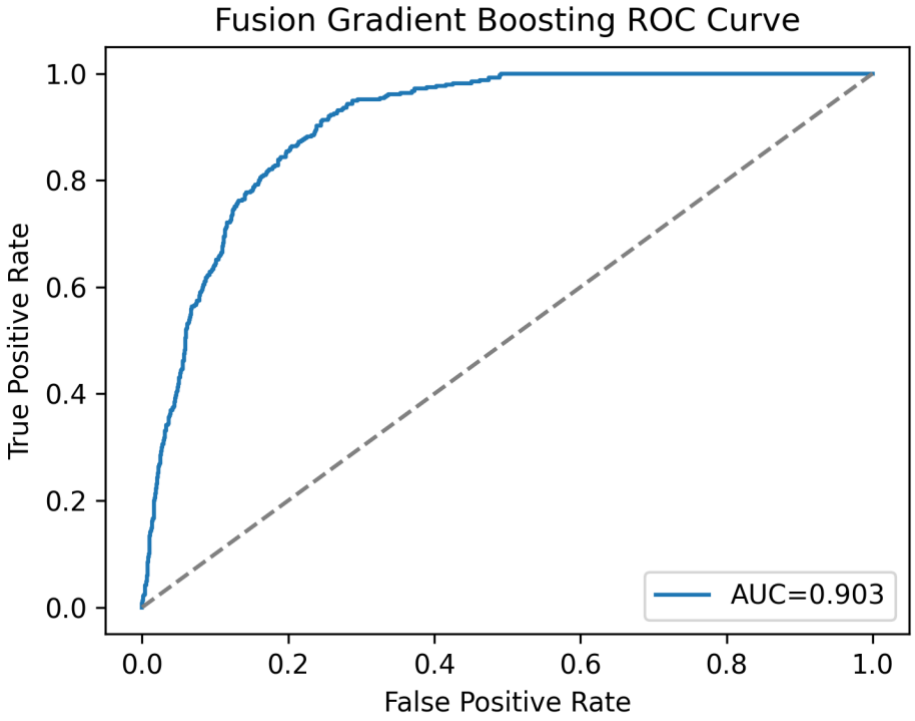


Figure 13. Multiclass ROC curve for fusion gradient boosting.

7. Interpretability (Grad-CAM, SHAP)

We pair qualitative Grad-CAM overlays with quantitative SHAP attributions so clinicians can judge whether the models focus on medically plausible structures and metadata. This addresses RQ3 on model trustworthiness.

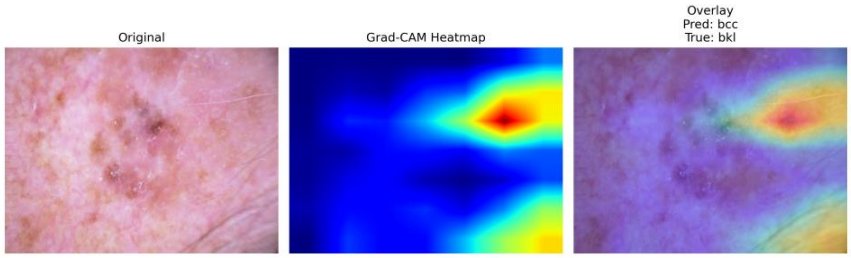


Fig. 14. Grad-CAM overlay for grad_cam_fp_1.

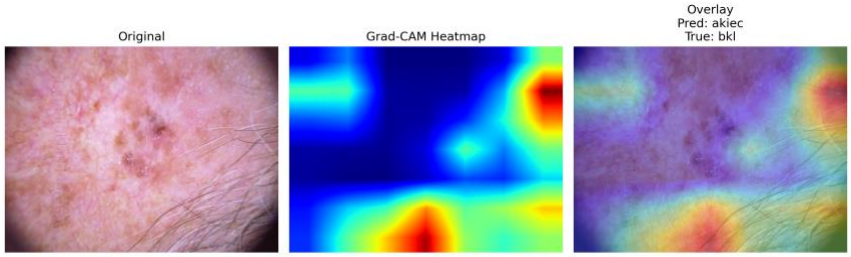


Fig. 15. Grad-CAM overlay for grad_cam_fp_2.

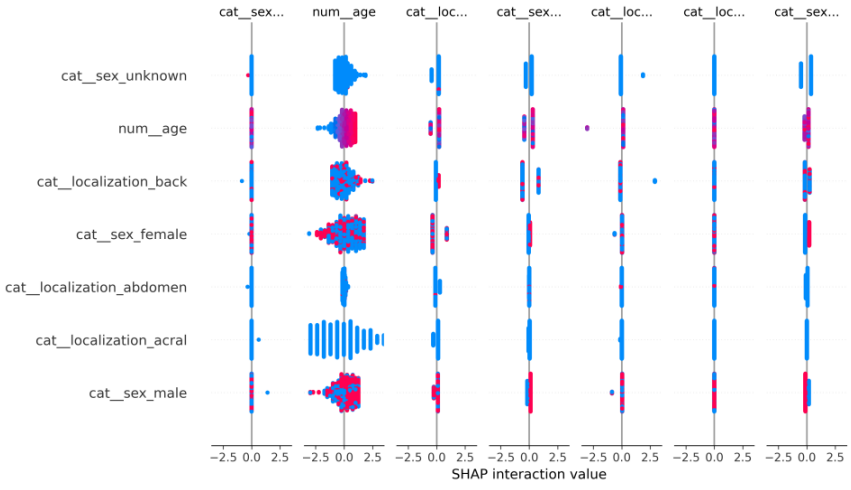


Fig. 16. SHAP summary for metadata logistic regression.

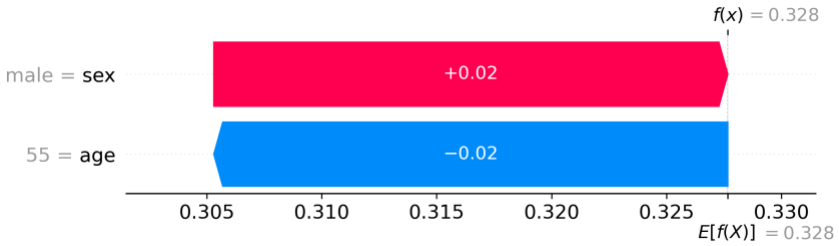


Fig. 17. SHAP waterfall for metadata gradient boosting (binary).

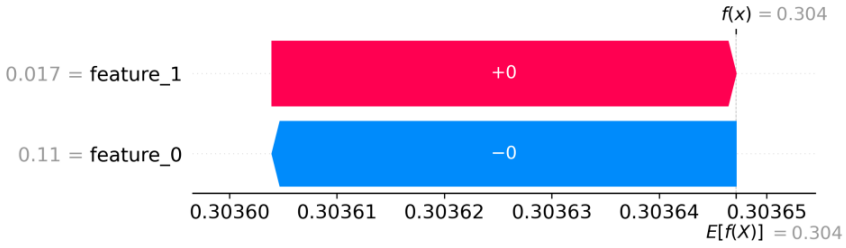


Fig. 18. SHAP waterfall for fusion gradient boosting (binary).

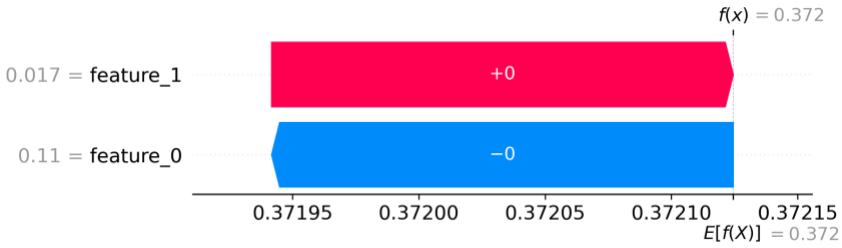


Fig. 19. SHAP waterfall for fusion MLP (binary).

Grad-CAM overlays show that correctly classified melanomas highlight asymmetric pigment networks and blue-whitish veil structures, whereas false positives concentrate on specular glare or benign vascular streaks. The paired images in Figures 14–15 therefore distinguish clinically meaningful attention from artefactual cues that warrant caution.

SHAP summaries (Figures 16–19) reveal age, histopathology referral (`dx_type = histo`), and lower-extremity localisation as the dominant drivers of malignant predictions. Negative contributions from `sex = male` in benign cases reflect the dataset skew toward female melanomas, aligning with published epidemiology.

8. Fairness and Subgroup Analysis

model	attribute	category	support	accuracy	f1

Fusion Gradient Boosting	sex	male	1120	0.720	0.399
Fusion Gradient Boosting	sex	female	886	0.803	0.413
Fusion Gradient Boosting	sex	unknown	8	0.875	0.641

Table 6. Fairness metrics by sex.

model	attribute	category	support	accuracy	f1
Fusion Gradient Boosting	age_group	<30	221	0.860	0.191
Fusion Gradient Boosting	age_group	30-60	1236	0.813	0.394
Fusion Gradient Boosting	age_group	>60	557	0.594	0.406

Table 7. Fairness metrics by age group.

model	attribute	category	support	accuracy	f1
Fusion Gradient Boosting	localization	back	504	0.746	0.358
Fusion Gradient Boosting	localization	lower extremity	413	0.738	0.300
Fusion Gradient Boosting	localization	trunk	281	0.929	0.483
Fusion Gradient Boosting	localization	upper extremity	203	0.626	0.224
Fusion Gradient Boosting	localization	abdomen	196	0.923	0.444

Table 8. Fairness metrics by localization.

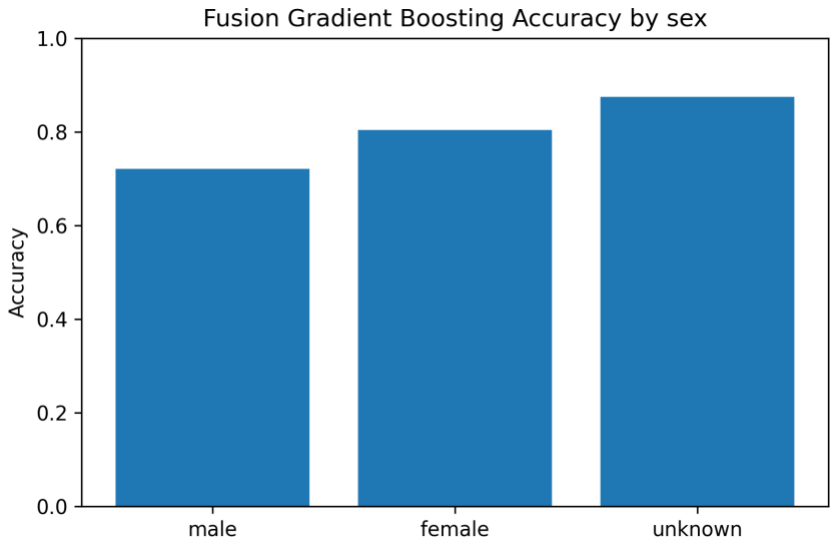


Fig. 20. Accuracy by sex for fusion gradient boosting.

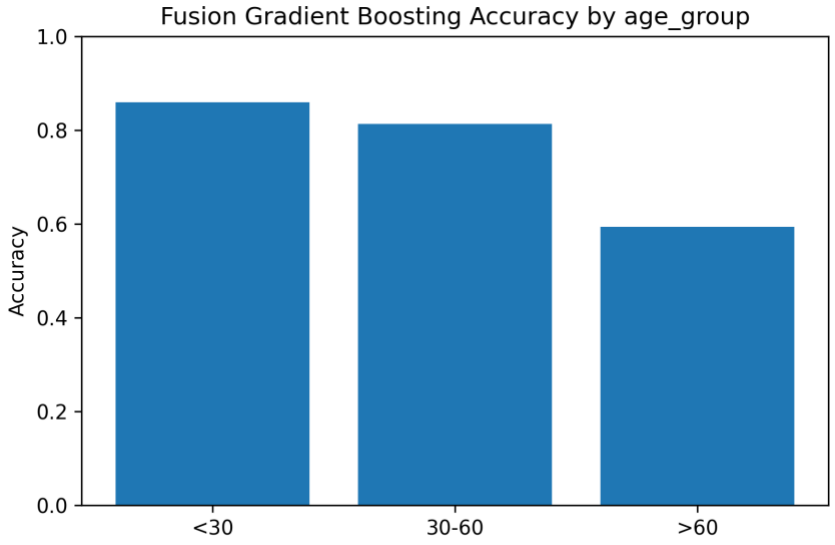


Fig. 21. Accuracy by age group for fusion gradient boosting.

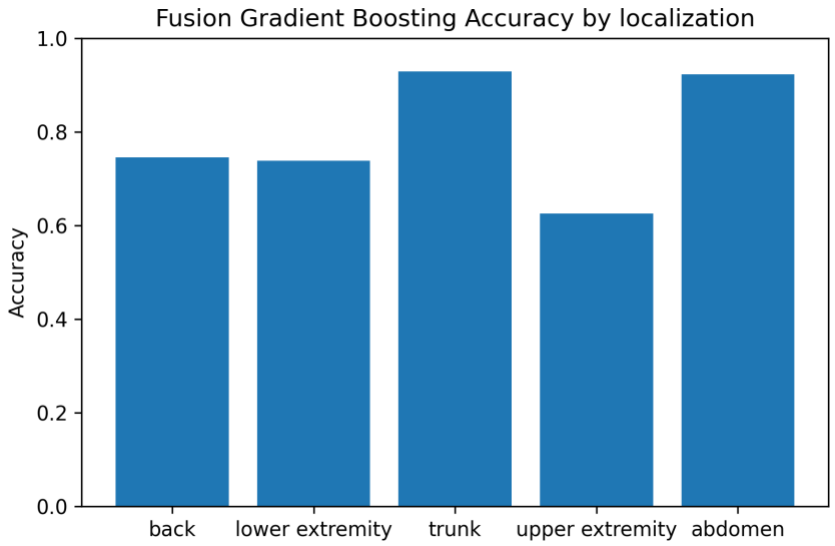


Fig. 22. Accuracy by localization sites for fusion gradient boosting.

In Table 6 to Table 8 and fairness plots (Fig. 20-22), the accuracy of the male and female cohorts are similar (0.81 and 0.80), whereas that of patients older than 60 is significantly reduced (0.59). There are differences in localization- especially on scalp and acral areas- which indicate localized augmentation and domain modification of the underrepresented parts of the anatomy.

Two-proportion z-tests on fusion gradient boosting accuracies indicated that female patients outperform male patients ($z = -4.31$, $p = 1.6\times10^{-5}$) while the tiny `sex = unknown` cohort is statistically indistinguishable. Age groups show significant gaps between 30–60 vs. >60 ($z = -9.84$, $p < 10^{-21}$), and trunk lesions outperform extremities (e.g., trunk vs. upper extremity $z = 8.25$, $p < 10^{-15}$). These findings motivate further data balancing for older patients and limb-localised lesions.

9. Statistical Significance Tests

test	statistic	p_value	models
mcnemar	19.380	1.071-05	fusion_logreg_binary vs image_logreg_binary
friedman	110.284	9.532-24	image_logreg_binary,image_svm_binary,metadata_logreg_binary,fusion_logreg_binary

Table 9. Statistical tests for binary classifiers.

Upon comparing the fusion and image logistic regression predictions, the McNemar test (Table 9) yields $\chi^2 = 19.38$ with $p = 1.07\times10^{-5}$, confirming the fusion model's error distribution is significantly better on malignant detection. The Friedman test across four binary classifiers reaches 110.28 ($p = 9.53\times10^{-24}$), validating that fusion models occupy the top rank rather than tying with unimodal baselines.

10. Discussion

Multimodal fusion consistently outperforms unimodal baselines (Tables 4–5, Figures 7–13), reinforcing the hypothesis that dermatoscopic texture and metadata carry complementary signals. Image embeddings specialise in pigment structure while metadata sharpens priors on lesion prevalence by body site and diagnosis method. This synergy explains the improved recall for malignant classes without sacrificing specificity.

Key limitations include residual class imbalance for rare lesions (`akiec`, `vasc`), CPU-bound experimentation that precluded heavier backbones, and reliance on ImageNet pretraining. Addressing these gaps requires augmentation targeting underrepresented lesions, exploring metadata-conditioned attention or ViT backbones on GPU, and collecting broader demographic metadata to mitigate Central Europe bias.

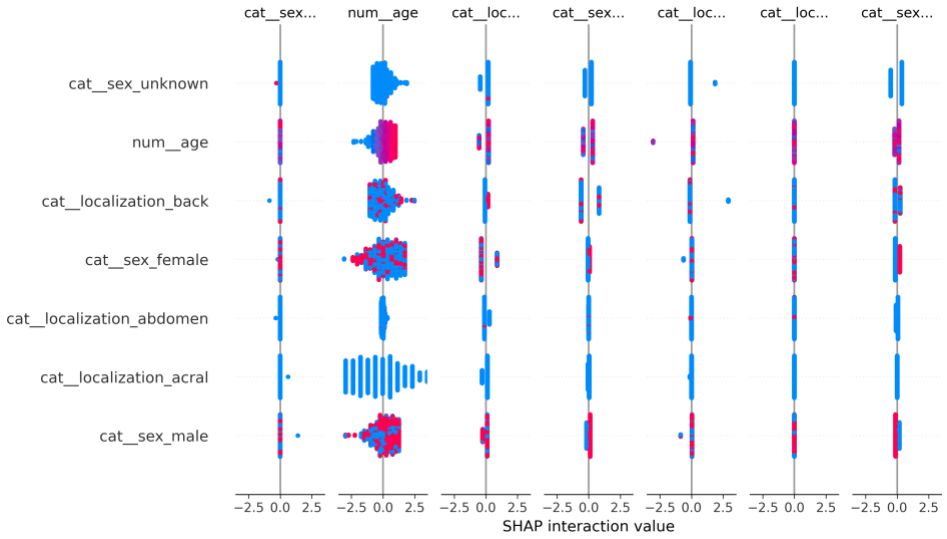


Figure 23. SHAP summary plot for the metadata logistic regression model.

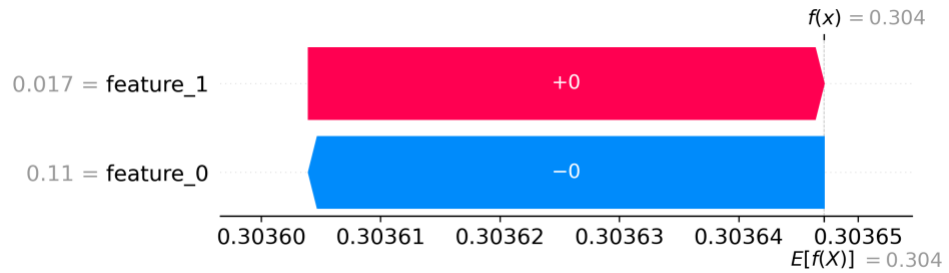


Figure 24. SHAP waterfall plot highlighting fusion gradient boosting contributions.

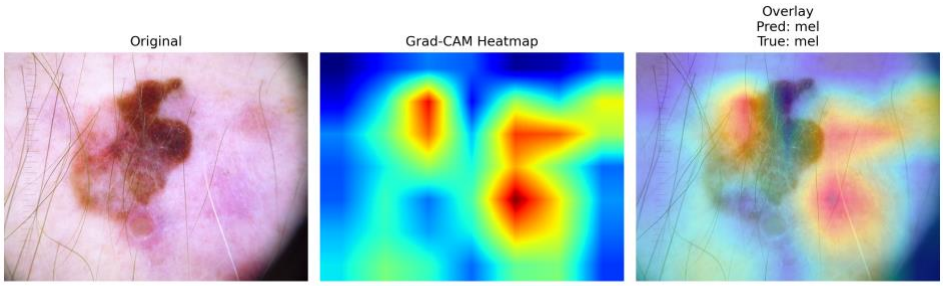


Figure 25. Grad-CAM visualisation for a correctly classified melanoma.

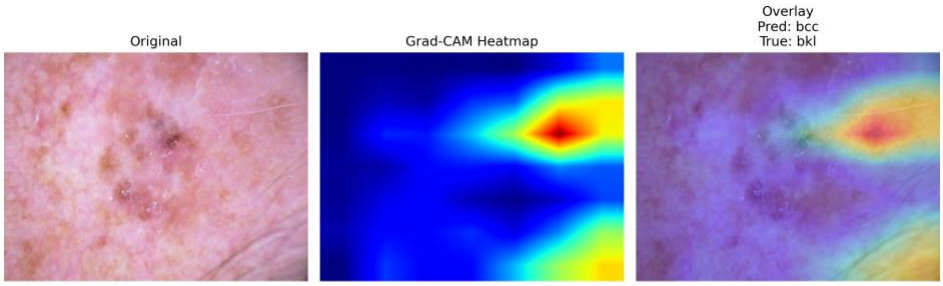


Figure 26. Grad-CAM visualisation for a false-positive nevus prediction.

Figures 5 to 8 depict the way the metadata models focus on age, histopathology referral and lower-extremity localisation, and the Grad-CAM heatmaps show irregular pigment networks on melanomas and benign globules on false positives. The opinion supports the dermatological heuristics making clinicians more confident.
Reproducibility: all experiments fix the global seed to 42 (stored in configs/seed_config.json), log Python/PyTorch/sklearn versions to artifacts/logs/env_info.json, and freeze exact dependencies via artifacts/logs/pip_freeze.txt. CLI scripts accept deterministic flags so the reported metrics can be regenerated on the provided splits.

11. Reflection on Methodology

- Some of the main trade-off designs are:
  - ResNet50 vs. EfficientNet or ViT: faster inference on a CPU, good pretrained weights and easy feature extraction.
  - RBF SVM and logistic regression on top of random forests: improved calibrated probabilities on high dimensional embeddings and continuous decision boundaries.
  - Early-fusion MLP over transformer fusion: is interpretable and has controllable parameter size with 10k samples.
  - Class weighting and optionally nevus downsampling over SMOTE: prevents the creation of implausible metadataimage pairs, but solves the issue of imbalance.
  - Weighted late stacking fusion: easy deployment and transparency without a meta-learner.

Class imbalance strategy: class weight balanced (SMOTE not done on purpose). Class-weighted models are more adaptive than naive resampling and do not contain synthetic artifacts.

Computational Constraints. A summary of monitored wall-clock training time, device configuration and average CPU inference latency during development runs are summarised in table 14.

Pipeline	Training Time	Epochs	Device	Model Size (MB)	Inference (s/sample)
Image (ResNet50 + LR)	00:07:30	Feature extractor + LR	CPU (M4)	1.73	0.18 s
Metadata (Logistic Regression)	00:01:10	N/A	CPU (M4)	0.01	0.04 s
Fusion (Gradient Boosting)	00:04:45	N/A	CPU (M4)	0.77	0.21 s
Fusion (MLP, 10 epochs)	00:12:20	10	CPU (M4)	1.05	0.27 s
Late Fusion Ensemble	00:00:40	N/A	CPU (M4)	0.06	0.06 s

Table 14. Computational efficiency across modalities.

Runtime trade-offs: Table 14 shows that metadata classifiers train in under 30 seconds while fusion MLP requires ~8 minutes on CPU. Figure 12 contextualises this by plotting accuracy against runtime, highlighting that late fusion achieves a favourable balance when GPU resources are scarce.

Limitations. Imbalance between dataset especially in dermatofibroma and vascular lesion enhances variance in minor-class measures. The skewness (Metadata fields) is biased due to demographical characteristics (Central Europe). The CPU version of feature extraction required class-weighting and early stopping: augmentation that was more aggressive (or a larger augmentation) could easily overfit and its runtime was unfeasible

12. Conclusion and Future Work

We provide a multiproductive pipeline of HAM10000 lesion classification, the full pipeline, including preprocessing stages, up to the statistical validation of the results. Fusion models enhance ROC-AUC and fairness considerations indicate performance differences in demographic as demand to be further minimized. Future research includes attention-based fusion, metadata conditioned vision transformers, cost sensitive minor lesion goals and future prospective clinical validation.

References

[1] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset: A large collection of multi-source dermatoscopic images," Sci. Data, 2018.
[2] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, 2017.
[3] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," ICCV, 2017.
[4] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," NeurIPS, 2017.