Multimodal Models for Skin Cancer Classification using Clinical Free Text and Dermatoscopic Images

Matthew Watson[1,2], Thomas Winterbottom[1,2], Thomas Hudson[1], Benedict Jones[1], Hubert P. H. Shum[2], Amir Atapour-Abarghouei[2], Toby Breckon[2], James Harmsworth King[1,2], and Noura Al Moubayed[1,2]

[1]Evergreen Life, UK
{*george.hudson, benedict.jones*}*@evergreen-life.co.uk*
[2]Department of Computer Science, Durham University, UK
{*matthew.s.watson, hubert.shum, amir.atapour-abarghouei, toby.breckon, james.h.king, noura.al-moubayed*}*@durham.ac.uk*

February 4, 2026

Supplementary Methods

Machine Learning Multimodality: A Careful Definition

It is important that —given our work is multidisciplinary— we are clear here on terms that have many different connotations in different fields.

- **Modality:** A mode or form in which information is experienced or expressed: *e.g.,* the English language.
- **Medium:** The means by which information is stored and delivered. *e.g.,* text is the **medium** through which the English Language **modality** is expressed.
- **Multimodal:** An adjective for a task, dataset, model, or other methodology that uses or requires information from more than one modality. Note that machine learning research generally does not yet distinguish between 'multi-medium' and 'multi-modality' scenarios.

In machine learning, the terms 'multimodality' and 'multimodal processing' are sometimes used to distinguish information from two fundamentally different sources *e.g.,* text and images. It is also sometimes (perhaps more precisely) used to distinguish between different modalities of the same medium *e.g.,* RGB and infrared representations modalities of the image medium.

Llama Prompts

Box 1: Prompt used with LLaMa 3.1 for the STag freetext filtering level. Examples have been removed for patient privacy.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a helpful AI assistant for processing clinical notes about skin lesions. Your goal is to replace
    language that implies diagnosis or treatment with some tags.

I will provide you with some clinical text. Please complete the following for me:
Return a 'Lvl 4' version of the text which replaces the following kinds of information in the text with
    corresponding tags:
    - @NT@ for any words to the effect of 'no treatment required'
    - @RTH@ for any referal to hospital or for biopsy.
    - @T@ for any treatment for skin conditions
    - @D@ for any skin condition diagnosis

Rules:
    - Only use the 4 tags defined above
    - If i do not supply you with any text, return 'nan' instead.
    - Begin your response with only 'Lvl 4 :'
    - Remove all details about how to use treatment and how often
    - Remove all details of creams, gels, and instructions on using them

[20 examples, removed for anonymisation purposes]

<|eot_id|>
<|start_header_id|>user<|end_header_id|>

Do NOT include details about application or frequency of treatments. These should be included in the @T@
    tag

Remember that discharged does not necessarily always mean @NT@

<|start_header_id|>assistant<|end_header_id|>
```

Box 2: Prompt used with LLaMa 3.1 for the FFilt freetext filtering level. Examples have been removed for patient privacy.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a helpful AI assistant for processing clinical notes about skin lesions. Your goal is to return a
     version of the text with the following things removed: names of skin conditions, treatments, and
    details of hospital removals or discharges.

I will provide you with some clinical text. Complete the following task:
- Return a trimmed version of the text that keeps only facts that don't imply a diagnosis. e.g. do not
    use the word benign, acne, psoriasis, alopecia, and so on...

Rules:
    - If i do not supply you with any text, return 'nan' instead.
    - Begin your response with only 'Lvl 5:', even if the task has an empty answer.
    - Do not include descriptions of creams or gel treatments

[14 examples, removed for anonymisation purposes]

<|eot_id|>
<|start_header_id|>user<|end_header_id|>

- Do not include details on creams or gels or ointment
- Do not include 'discharged', 'no treatment required', 'refer to hospital', or any other treatment for
    the patient.
- Do not include any name of any skin or hair condition at all.

<|start_header_id|>assistant<|end_header_id|>
```

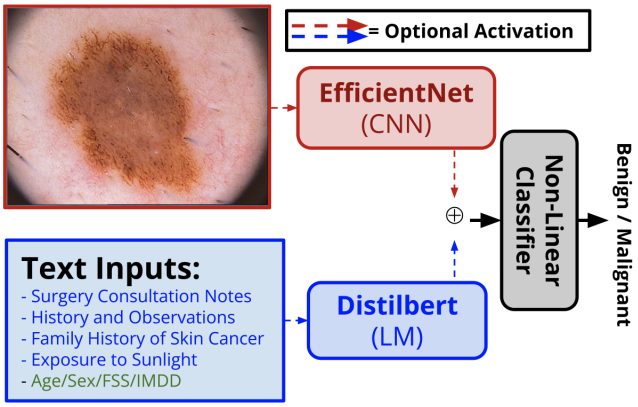
Supplementary Results

Performance with Validation Derived Threshold

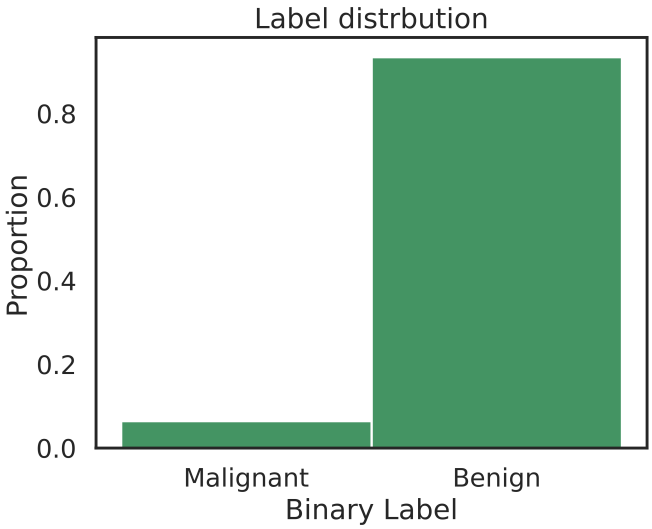
	Surgery Consultation Notes	History & Observations	Family History of Skin Cancer	Exposure to Sunlight	Sex	Age	FSS	IMDD	Averaged Metrics (5 Runs)				Label Key
Vision	Text Components								AP	Spec	Sens	AUROC	
✓									0.547 ±0.017	57.09% ±7.15%	96.10% ±1.97%	0.909 ±0.006	Vision Only
	Orig								0.587 ±0.039	86.50% ±4.09%	98.44% ±0.72%	0.963 ±0.006	Surg Cons (Orig)
✓	Orig								0.802 ±0.047	84.42% ±7.85%	97.14% ±4.47%	0.980 ±0.004	V+Surg Cons (Orig)
	CFilt								0.424 ±0.021	77.25% ±8.18%	98.44% ±1.77%	0.944 ±0.005	Surg Cons (CFilt)
✓	CFilt								0.658 ±0.096	83.89% ±1.81%	99.03% ±1.54%	0.967 ±0.007	V+Surg Cons (CFilt)
	DFilt								0.405 ±0.021	84.53% ±6.40%	94.81% ±2.28%	0.939 ±0.008	Surg Cons (DFilt)
✓	DFilt								0.726 ±0.068	78.70% ±6.95%	99.22% ±1.44%	0.970 ±0.006	V+Surg Cons (DFilt)
	FFilt								0.285 ±0.005	15.79% ±8.93%	98.44% ±3.50%	0.824 ±0.013	Surg Cons (Lvl4)
✓	FFilt								0.667 ±0.026	63.15% ±3.99%	98.96% ±1.35%	0.955 ±0.007	V+Surg Cons (Lvl4)
		Orig							0.267 ±0.012	41.61% ±16.2%	93.51% ±6.94%	0.818 ±0.035	Hist & Obs (Orig)
✓		Orig							0.616 ±0.043	66.35% ±8.29%	89.61% ±1.14%	0.898 ±0.020	V+Hist & Obs (Orig)
		FFilt							0.257 ±0.016	22.49% ±9.69%	96.88% ±3.71%	0.782 ±0.013	Hist & Obs (Lvl4)
✓		FFilt							0.560 ±0.02	64.21% ±3.93%	90.91% ±1.97%	0.908 ±0.011	V+Hist & Obs (Lvl4)
			✓						0.211 ±0.003	—%	—%	0.705 ±0.007	Fam Hist of Skin Cancer
				✓					0.192 ±0.018	18.83% ±9.70%	94.55% ±4.62%	0.699 ±0.017	Exposure to Sunlight
	Orig	Orig	✓	✓	✓	✓	✓	✓	0.734 ±0.081	89.27% ±2.09%	97.40% ±2.55%	0.976 ±0.005	All Text (Leading Language)
✓	Orig	Orig	✓	✓	✓	✓	✓	✓	0.747 ±0.094	87.04% ±3.55%	95.32% ±3.14%	0.970 ±0.009	V+All Text (Leading Language)
	FFilt	FFilt	✓	✓	✓	✓	✓	✓	0.444 ±0.033	40.42% ±24.12%	94.45% ±3.30%	0.870 ±0.028	All Text (No Leading Language)
✓	FFilt	FFilt	✓	✓	✓	✓	✓	✓	0.679 ±0.038	69.37% ±8.30%	96.62% ±2.93%	0.948 ±0.013	V+All Text (No Leading Language)

Supplementary Table 1: The main results table for the paper when the decision threshold is calculated using the validation set. Each result is comprised from 5 runs with exactly the same experimental setup and hyperparameters. Mean and the 95% confidence interval of each metric is reported. Where sensitivity and specificity are unavailable, the model was consistently unable to reach a 95% sensitivity level without always predicting the positive class. Red and green cells indicate dermatoscopic images and patient metadata, respectively, were included. Blue cells indicate freetext was included, with darker blues designating higher levels of freetext filtering. AP: Average Precision (baseline AP: 0.07); AUROC: Area Under the Receiver Operating Characteristic Curve.

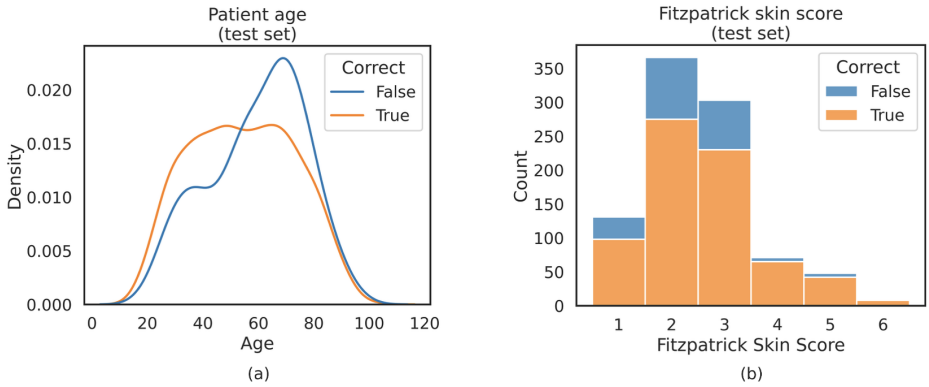
Supplementary Figures



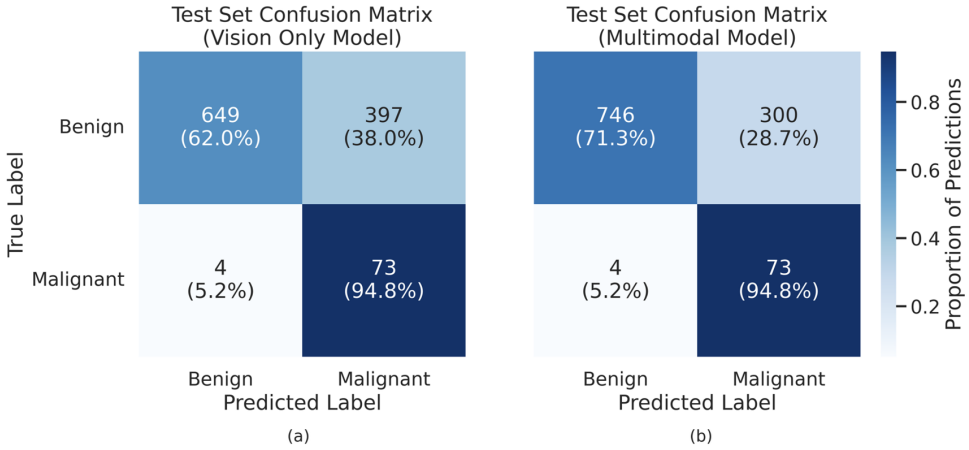
Supplementary Figure 1: The multimodal classification framework. Inputs from any of the data sources can optionally be included in feature concatenation for classification as represented by dashed arrows. The image of the skin lesion is taken from the ISIC dataset [1].



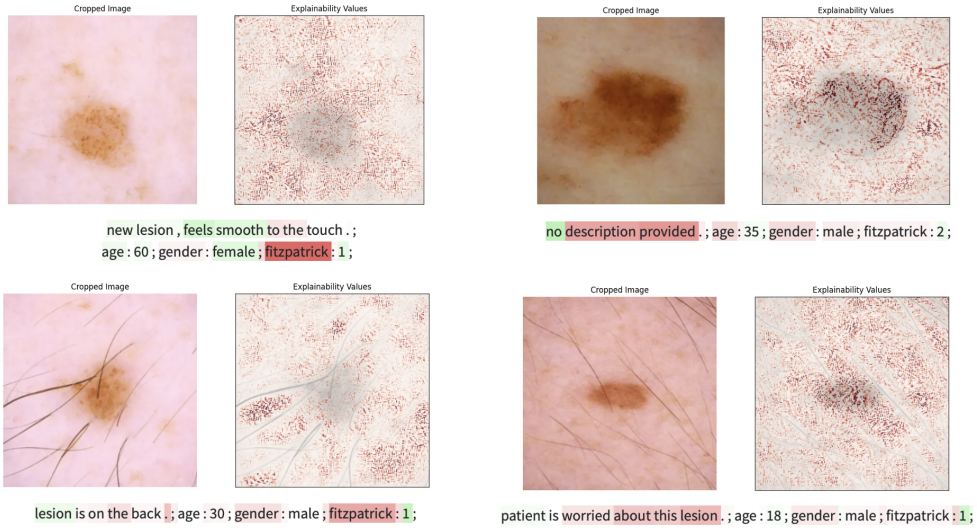
Supplementary Figure 2: Proportion of images with each ground truth label (*i.e.*, diagnosis)



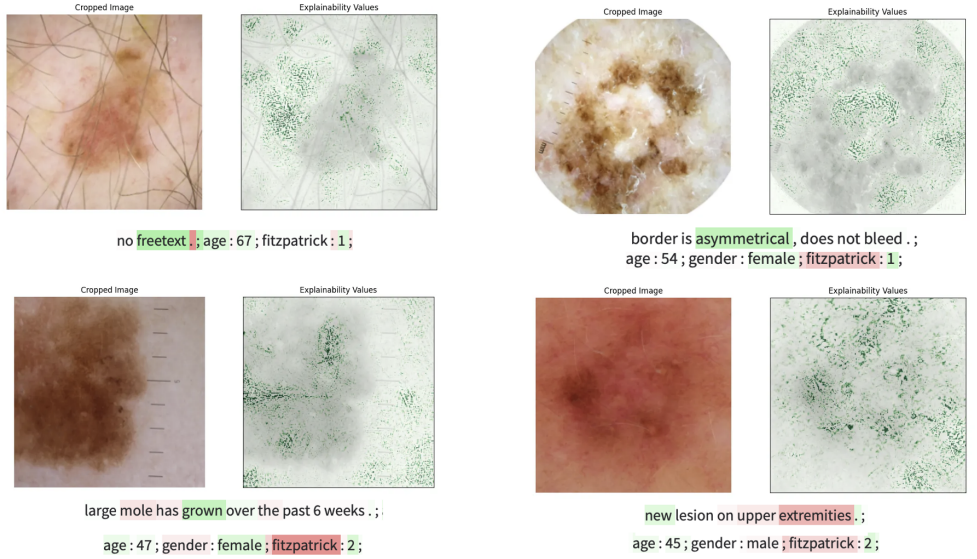
Supplementary Figure 3: Fully multimodal (using filtered text, age, sex, Fitzpatrick Skin Score, and dermatoscopic images) model error rates across (a) patient age and (b) patient Fitzpatrick Skin Score.



Supplementary Figure 4: Confusion matrices across the test set of (a) a vision only model, and (b) fully multimodal using filtered text, age, sex, Fitzpatrick Skin Score, and dermatoscopic images. A 95% sensitivity has been targeted. Darker blue areas indicate higher normalised values.



Supplementary Figure 5: 4 samples from ISIC 2020 [2, 3] which our multimodal classifier classify as benign. For each sample, the original image is shown beside an Integrated Gradients explanation from the model. Freetext descriptions passed into the model are also shown, again overlaid with the computed Integrated Gradients explanation (though note that, as ISIC 2020 does not contain freetext, these descriptions are entirely fictional). Green highlights features pushing the model towards a malignant diagnosis and red towards benign, with darker areas indicating more importance.



Supplementary Figure 6: 4 samples from ISIC 2020 [2, 3] which our multimodal classifier classify as malignant. For each sample, the original image is shown beside an Integrated Gradients explanation from the model. Freetext descriptions passed into the model are also shown, again overlaid with the computed Integrated Gradients explanation (though note that, as ISIC 2020 does not contain freetext, these descriptions are entirely fictional). Green highlights features pushing the model towards a malignant diagnosis and red towards benign, with darker areas indicating more importance.

References

[1] Noel C. F. Codella, Veronica M Rotemberg, Philipp Tschandl, M. E. Celebi, Stephen W. Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Armando Marchetti, Harald Kittler, and Allan C. Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *ArXiv*, abs/1902.03368, 2019.

[2] Veronica M Rotemberg, Nicholas R. Kurtansky, Brigid Betz-Stablein, Liam J. Caffery, Emmanouil Chousakos, Noel C. F. Codella, Marc Combalia, Stephen W. Dusza, Pascale Guitera, David Gutman, Allan C. Halpern, Harald Kittler, Kivanç Köse, Steve G. Langer, Konstantinos Liopryis, Josep Malvehy, Shenara Musthaq, Jabpani Nanda, Ofer Reiter, George Shih, Alexander J. Stratigos, Philipp Tschandl, Jochen Weber, and Hans Peter Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8, 2020.

[3] International Skin Imaging Collaboration (ISIC) Archive. Isic archive: Collections api. https://api.isic-archive.com/collections/. Accessed: 2025-09-25.