Multimodal Models for Skin Cancer Classification using Clinical Freetext 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

Abstract

Background

Skin cancer is one of the most prevalent cancers globally, with early detection critical to ensure reduced mortality risk. To aid early detection, machine learning (ML) skin cancer detection models have been proposed, currently with a focus on dermatoscopic imaging only. However, freetext may provide extra diagnostic information that is not present in images alone.

Methods

We constructed a multimodal dataset comprising 5,481 dermatoscopic images from 4,538 patients, including patient metadata and clinical notes, with binary labels (benign vs. malignant, 7% malignant). To assess and mitigate bias from leading language, we developed a clinical text preprocessing pipeline combining regular expressions and large language models, enabling multiple levels of filtering. We train multimodal ML models on this dataset to explore the effect of freetext on model performance.

Results

Our results show that incorporating unfiltered text significantly improves classification performance (0.970 AUROC) compared to visual data alone (0.909 AUROC); even with leading language removed, performance gains persist (0.948 AUROC).

Conclusions

This work benchmarks clinical freetext inclusion in skin lesion classification, demonstrating that clinical text contributes predictive value beyond that available in images alone. The model's high performance on unfiltered clinical text highlights the high levels of bias, and possible shortcutting, present in this text which may make it unsuitable for inclusion in some ML models. By systematically filtering clinical notes via our proposed technique, we show that multimodal models retain improved accuracy while reducing bias. These results provide practical guidance for integrating clinical text into real-world skin cancer detection systems and establish a foundation for future multimodal research in dermatology.

Plain Language Summary

Prompt detection of skin cancer improves survival, but diagnosis must be made by clinicians. Image-based machine learning models for skin cancer classification have shown promise. However, key information is often only recorded in clinical notes, such as whether a lesion has changed, itches, or bleeds. By creating a dataset that contains images, patient data, and freetext descriptions of the problem, we train a series of machine learning models on both images and freetext to predict skin cancer. We show that the inclusion of freetext significantly enhances model performance, but that care must be taken to ensure the freetext does not unintentionally bias the model. These models could be used in multiple points in a skin cancer clinical workflow to either support more accurate referrals to dermatology, or direct patient access to dermatology services, potentially reducing wait times and improving patient outcomes.

Introduction

Skin cancer is one of the most prevalent cancers globally and, despite efforts towards improving prevention, its incidence continues to increase [1]. However, the 5-year net survival rate for early stage detection of skin cancer, particularly melanoma, is extremely high relative to other cancers, reaching up to 99.6% for stage 1 melanoma [2] (although known health inequalities in dermatology exist [3]). Effective screening via self-examination helps the early detection of skin cancer that might otherwise result in mortality [4]. Skin self-examination is technically straightforward: an unaided visual inspection of the lesion. If the patient identifies any concerning features they are encouraged to present to a healthcare professional (HCP) for a clinical assessment which includes a detailed medical history and examination of the lesion, often with magnification (a dermatoscope). If the HCP suspects skin cancer they will refer the patient to a specialist service using a suspected cancer pathway [5].

The UK primary care system is under unprecedented pressure with an increasing demand put on referrals into specialist services. To address these pressures, work is needed to improve both referral efficiency and diagnostic accuracy within primary care [6]. The digitisation of primary care records, technical ease of skin inspection, and high early survival rates together mean that a machine learning (ML) powered detection framework for skin cancer has the potential to deliver these efficiencies and improve skin cancer related mortality.

Encouragingly, the past few years have seen a plethora of work bringing the power of ML to the problem of skin lesion classification [7, 8, 9], yielding models with diagnostic accuracies that outperform some primary care HCPs. Though earlier databases of labelled skin lesion images are relatively small by modern standards, they were instrumental in evaluating the relatively impressive classification systems of their era [10, 11]. Over time, increasingly comprehensive dermatology datasets have been published [12, 13] with datasets such as HAM10000 [14, 15] and BCN20000 [16, 17]. Most recently, the ISIC 2020 image challenge [18, 19] contains images that distinguish between over a dozen classes of malignant lesion types. Cassidy et al. [20] provide a thorough analysis of the ISIC database.

While these datasets primarily use of images for skin lesion classification, recent approaches have sought to improve model performance by introducing additional modalities of input data. Metadata of the *patient* such as age and biological sex is a common source of additional information that many approaches add as features into models [21, 22]. Other research uses metadata about the *lesion* itself as additional inputs [23, 24, 25, 26, 27]. Using such patient and lesion metadata modalities has become standard practice for state-of-the-art lesion classification [23]. However, to the best of our knowledge, while multimodal ML models have gained traction within other medical specialities [28, 29], there has been no research into the use of clinical freetext notes alongside categorical metadata and skin lesion images for lesion classification.

In this paper, we take the next step in multimodal skin lesion classification by introducing natural language text information as an input modality. Such clinical freetext provides extensive natural language descriptions of lesion appearance and any associated symptoms over time. It contains detailed information beyond that conveyed in the categorical metadata (*e.g.,* "patient has an uncle who had melanoma in their 30s"), and additional input features such as working diagnosis and management plans (*e.g.,* "the mole contains some worrying features, refer for biopsy to exclude melanoma").

However, this type of freetext often contains 'leading language' - that is, text written in a way which may state or imply a particular diagnosis by its sentiment or management plan. Using this text during model training risks learning 'shortcuts' or spurious correlations - that is, features of the text which are correlated with the target variable but not causally related to it [30, 31].

To the best of our knowledge, our work is the first to identify and remove diagnoses and other such forms of leading language from clinical freetext for the purposes of an ablation study in skin lesion classification. Previous work has proposed a system for detecting uncertainty in clinical text [32, 33, 34, 35]. While clinical named entity recognition [36] techniques are well established for detecting diagnoses and other important clinical units of text, these are designed to detect specific words and phrases (*e.g.,* conditions as present in UMLS [37]) whereas our proposed methods are designed to detect and remove more broad forms of leading language. Although methods exist that aim to debias clinical text (or, often, text embeddings from a language model), these are aimed at reducing gender, age, ethnicity bias rather than label leakage or clinical language [38, 39]. Similarly, Ji et al. [40] and Wiest et al. [41] use large language models (LLM) for clinical named entity recognition. Where these works focus on the task of detecting these components in and of itself, we extend the usage of LLMs to detecting leading language as a preprocessing step to skin lesion classification. This approach aims to mitigate the risk of bias and label leakage that has shown to be present in clinical freetext, and can be extended to medical machine learning applications outside of dermatology [42, 43].

To this end, we develop an approach to carefully remove increasing levels of leading language, letting us shed light on the performance that certain forms of allowable freetext can unlock for a wide variety skin lesion classification systems. We combine this leading language preprocessing approach with an ablation study on the four distinct subcategories of clinical freetext in our dataset, providing a thorough exploration of performance improvements that a diverse array of text can bring to skin lesion classification. Our results show that unfiltered text yields a large performance increase to the classification task (0.970 AUROC). Further, our extensive component-wise analysis of text demonstrates high levels of bias and label leakage and thus that great care should be taken to preprocess the text in scenarios where leading language is undesirable; this component-wise analysis also shows that our proposed LLM-based pipeline is successful in removing this leading language. Understanding the potential for unintended bias in medical ML [44], we evaluate our best performing skin lesion classification models in terms of their bias. In summary, this paper proposes exploiting LLMs to implement a preprocessing step to adjust levels of leading language in clinical freetext - an essential measure to mitigate bias in multimodal classification - and demonstrates state of the art classification performance using these techniques for skin lesion classification.

Methods

To analyse the effect of the different levels of freetext on ML models, we train multimodal (see Supplementary Methods for a discussion of multimodality) ML classifiers for skin lesion classification using what is, to the best of our knowledge, the first large-scale skin lesion dataset that contains dermatoscopic images, patient metadata, and clinical freetext.

Multimodal Dataset

Our dataset is derived from anonymised secondary care electronic health records (EHR) of individual patients with skin conditions, obtained from Community Dermatology Services. A redacted and fully anonymised data set was provided by a healthcare service provider in order to develop algorithmic (AI) tools to support and improve the diagnosis of skin conditions within their service provision. This process was reviewed and approved by the healthcare provider's data protection officer, is in line with UK GDPR requirements and is part of the healthcare provider's usual business practice and services (see the Ethics statement for more details).

The dataset contains demographic data, photographs of skin conditions taken by HCPs and/or patients, data on medical conditions and medications, clinical opinions and examinations, and where appropriate histology findings. As highlighted in Table 1, our dataset contains additional linked information and the non-image modalities it provides. Natural language freetext clinical notes contain longitudinal information about the history and changes of the skin lesion, the patients family history of skin disease, and historical exposure to sunlight *i.e.,* information that is not available from the lesion image alone.

Results reports dataset statistics. All patients had a single skin lesion condition diagnosed - the final dataset does not include rashes, nor patients with multiple diagnoses. Patients had at least one dermatoscopic image taken; all other fields (both freetext and patient metadata) were optional, with missing data being represented by an empty string when passed to the text classifier. Figure 1 shows exact inclusion/exclusion criteria for our dataset.

Dataset Curation Details

Patient data is encapsulated in FHIR format (NHS Digital - FHIR), and extracted as 'clinical impression' objects.

As this dataset comes directly from dermatology departments it naturally filters out almost all irrelevant data points *i.e. non-skin related diseases*. However, we further ensure the relevance of each data point by requiring that the SNOMED-CT (Systematized Nomenclature of Medicine Clinical Terms) diagnosis code (SNOMED International) is a child of the SNOMED-CT code for 'Skin Lesion' (*i.e. that the type of disease is a taxonomical child of 'Skin Lesion'*). In this way, SNOMED-CT codes in our dataset can be thought of as existing on a taxonomical tree with the parent SNOMED-CT of 'Skin Condition' as its root. Many of the SNOMED-CT codes in our dataset exist on the *leaves* of the taxonomical tree and are thus very granular and specific *e.g.* 'Malignant melanoma of skin of chin', and others are more general *e.g.,* 'Actinic keratosis'.

Patient Metadata and Ground Truth

The dataset contains RGB colour dermatoscopic images of lesions from secondary care. The average image resolution for the dataset is 992x1333.

Ground Truth is determined by a SNOMED-CT code for the specific type of skin disease identified in the dermatoscopic image. SNOMED-CT is a structured clinical vocabulary used in EHRs to standardise the recording of medical information and, in this case, are recorded by a HCP upon diagnosis. These SNOMED-CT codes are the source of the benign/malignant labels used for classification. A list of *unique* skin condition SNOMED-CT codes present in our dataset were sorted by dermatologists into the classes: 'benign', 'malignant', 'pre-malignant' or 'ambiguous'. Patients with an ambiguous diagnosis were removed, and we then transform this into a binary classification task by treating 'pre-malignant' codes as 'malignant'. Although inter-coder agreement on SNOMED-CT coding has sometimes been shown to be low [45], this issue is largely mitigated in our dataset through the manual grouping of all SNOMED-CT codes present in our dataset into one of these 4 categories (hence two similar codes will be given the same ground truth label).

All malignancies were confirmed through histopathological results whereas, as most patients with suspected benign conditions do not undergo biopsy, clinically confirmed benign diagnoses (i.e., a diagnosis from a dermatologist visually inspecting the lesion, without a further biopsy) were included. While this may mean that a small number of benign samples in our dataset may be mislabelled (due to clinician misdiagnosis), clinician false negative rates in dermatology have been shown to be low [46, 47] and the size of the benign dataset is large enough that this error rate should be no larger than the label noise shown to be present in many other ML datasets. Therefore, we expect the modern ML models used in this study to overcome this issue [48, 49]. Conversely, clinician false positive rates are much higher [46, 47] due to the desire to avoid missing a skin cancer diagnosis. Due to the small proportion of malignant samples in our dataset, these errors may disproportionately affect model training and therefore we restrict our data to histopathology-confirmed malignant diagnoses, which have a much lower error rate.

Discrete Patient Metadata: Our dataset contains four discrete features about the patients themselves that can be considered relevant to skin lesion classification: *age*; *sex*; *Fitzpatrick skin score* (a discrete scale of the tone of healthy skin, 1 as lightest and 6 as darkest) [50]; and an *index of multiple deprivation decile*, a score of poverty with 1 as most deprived and 10 least deprived, derived using the patient's Lower Layer Super Output Area (LSOA) [51].

Clinical Freetext Components

The clinical freetext used in this experiment is formed from notes of various types contained in the EHR. Though such data is often noisy, this subsection describes the following four components that are abundantly and richly annotated, making them suitable for inclusion in multimodal predictive models (examples are in Table 2).

Family History of Skin Cancer: This text is short (2.82 words on average) but dense with information. Typically, this field is a single, short phrase regarding knowledge of any family members with skin cancer; examples include "no history" or "father had melanoma". The potential classifying power that this modality of clinical freetext has is in theory derived from genetic predisposition to skin cancer [52, 53]. Though the average sentence in this section is short, the total vocabulary for this component is over 1,500, reflecting its diversity.

Exposure to Sunlight: Similarly to family history of skin cancer, this is another short (4.4 words on average) but information dense modality of text. These phrases serve as an approximation of a patient's sun exposure, doing so in diverse but detailed language (reflected by its 2,400+ vocabulary). This often requires contextual knowledge: for instance, "patient lived in Brazil for 10 years" requires the knowledge and reasoning that the sunny weather in Brazil means most of the population have higher than average ultra violet radiation exposure. Such information is well suited for analysis by the language models we use in our experiments, with exposure to sun known to increase the risk of skin cancer [52, 53].

Surgical Consultation Notes: The surgical consultation notes are longer (23.86 words on average) and broadly comprised of a combination of one or more of following three sub-categories:

1. **Lesion description:** A factual description of the lesion with no direct implication to a suspected diagnosis, *e.g.,* "lesion on right shoulder is dark red and bleeds when scratched"

2. **Diagnosis attempt:** Attempts to directly diagnose the lesion by a HCP *e.g.,* "basal cell carcinoma, refer to hospital". Such text naturally contains significant leading language.

3. **Treatment prescription:** Notes of which treatment has been prescribed. This can also contain significant classifying power as prescriptions for typical skin medication are almost always for benign skin conditions, with suspected malignancies instead referred directly to hospital for biopsy.

Although surgery consultation notes contain targeted and specific language that may aid in cancer classification, they also include language elements that strongly indicate the ground truth label, potentially introducing significant bias in ML models. We outline our methods for controlling such leading language in order to ensure proper ML model behaviour in Controlling Leading Language in the Clinical Text.

History and Observations: The longest (47.84 words on average) and most diverse component with a total vocabulary of over 6,900, containing descriptive language of details pertinent to the skin lesion *without* making a direct diagnosis *i.e.,* its size and colour. Perhaps most crucially, these histories describe changes in the lesions over time. Such information can greatly aid diagnosis *e.g.,* it is important to know if a lesion has grown or changed colour [54]. As shown in Table 2, compared to the surgical consultation notes, such free-form descriptive language is longer, and has a larger total vocabulary with more lexically diverse sentences, but much less leading language.

Controlling Leading Language in the Clinical Text

To the best of our knowledge, this study is the first to consider clinical freetext for ML based skin lesion classification. Naively adding all components of freetext into a classification system is, however, not always desirable. Such real-world text data can often be unstructured and messy, and can contain domain-specific terminology, including but not limited to: working diagnoses of diseases, medication, and symptoms [40, 41]. These facets of the text may strongly indicate the clinician's differential - either implicitly or explicitly - which could unintentionally bias a classification model through 'data leakage' or 'short-cutting'. Leading language is a significant issue in clinical text; Table 2 demonstrates that 26% of the surgical consultation notes in our dataset contain medical terminology.

It is important to note that the leading language in physician-provided freetext may be a result of clinical acumen and intuition, rather than just a summary of the clinical examination. In cases like these, it could be argued such leading language is beneficial to an ML model as it can provide information that is impossible to otherwise learn. In this study we directly explore this effect, through varying the levels of leading language allowed during model training, to better understand how leading language effects classification model performance.

Indeed, the level of acceptable leading language in the training of ML models depends on the point in the care pathway the model will be deployed. For example, a skin lesion classification model could be deployed within primary care to assist referrals to secondary care - in this case it may be acceptable to allow some forms of leading language from the primary care physician. On the other hand, it would not make sense to train on the same set of freetext data when aiming to deploy a skin lesion classification model directly to patients - there would be significant data shift between the clinical freetext used for training, and the patient-provided text in deployment. However, careful removal of leading language in the clinical freetext could shift the data distribution of clinical freetext towards that seen from patients, enabling training of models that could be deployed in patient-facing contexts even when the training data contains only clinician-provided freetext.

We compare and contrast traditional regular expression text removal techniques with LLMs. Using Llama 3.1 [55], we build upon existing work that used Llama 2 [56] to detect assertions, diseases, exams and dosages in clinical text [40, 41].

Strategy for Removing Leading Language

For a thorough analysis of the effects of leading language in our text features on the classification task, we process the **surgical consultation notes** at one of four levels of increasingly strict removal of 'leading' clinical content (see Table 3).

Our designated **Original (Orig)** text features the full clinical freetext. **ConditionFiltering (CFilt)** uses a regular expression to remove the names of any skin conditions, and **DiagnosisFiltering (DFilt)** further removes the words 'benign' and 'malignant', in effect removing all diagnoses. Though regular expressions can remove the exact diagnoses themselves, they are much less effective at removing any phrasing or text structure that could imply a diagnosis (*e.g.,* "referred to hospital" typically implies concern around a possibly malignant skin lesion). They are also ineffective at removing language that contains typos and/or spelling errors, and are difficult to generalise to new concepts [57]. We aim to capture and control for this difference by introducing two even higher standards of text preprocessing that utilise LLMs.

Our strictest scenario **FullyFiltered (FFilt)** is designed to remove any clinical leading language, effectively allowing only statements of known fact. The FullyFiltered filtering level is designed to remove any leading language written by a clinician that could not be gained by asking the patient a question - it is, effectively, a combination of the CFilt and DFilt levels with the addition of also removing proposed treatments. This means simple statements of fact, such as sun exposure or skin type, are allowable at this level as this information can be easily gathered by asking the patient a series of questions. The aim of this filtering level is to shift the text data distribution to more closely match what a patient could provide, to explore whether skin lesion classification models could be patient-facing rather than clinician-facing.

SemanticTagging (STag) is designed purely for analysis, where any phrasing pertaining to five common diagnostic scenarios are replaced by tags (Table 3). Any combination of these tags can then be removed/allowed to assess their classifying power.

An example of a typical piece of clinical freetext under each level is as follows (with filtered text replaced with underscores for visualisation purposes only - in the actual data, the terms are fully removed):

- **Original:** "lesions looks like benign mole. no treatment required."
- **ConditionFiltering:** "lesions looks like benign ____. no treatment required."
- **DiagnosisFiltering:** "lesions looks like ___________. no treatment required."
- **FullyFiltered:** "lesions;"
- **SemanticTagging:** "lesion looks @DB@. @NT@."

As FullyFiltered and SemanticTagging are complicated to derive from the raw text, we process them using Llama 3.1. This model was not finetuned, but rather processed each of the training and validation examples one at a time using over 20 examples from the test set as guiding examples in the prompt (*i.e.,* few-shot prompting). The prompts used are available in the Supplementary Methods; when performing text filtering, only the freetext to be filtered is provided to the LLM (not any associated patient metadata, diagnoses, etc). Random samples of the resulting processed text were manually inspected to ensure data quality.

Classification Model Training

We use stratified sampling (on the ground truth label, *i.e.,* benign/malignant) to split our dataset into three subsets: 3,277 (60%) in the training set; 1,081 (20%) in the validation set; and 1,123 (20%) in the test set. We stratified on a per-patient basis, so a patient with multiple images will not be distributed across different splits.

As our dataset contains medical image and text features, we cannot rely on pretrained models and must train models from scratch; due to computational limitations, large-scale multimodal LLMs are beyond the resources of this study. To this end, we purposefully select vision and text component models that are reliable and have been widely used as benchmarks for different applications. To classify the dermatological images, we use ConvNext-Large-224 [58] as a convolution-based vision encoder. ConvNext was selected for its ability to match or outperform vision transformers on a range of medical imaging tasks [59, 60], and with the knowledge that convolution-based classifiers are generally more effective than transformer-based networks for dermatoscopic images due to their structural characteristics [61]. We transform tabular data to text, in the format `ColumnName1 is Value1, ..., ColumnNameN is ValueN`, as previous studies have shown this provides superior performance over other tabular data encoding techniques [62]. To encode this free-text and tabular data, we use BioClinicalBERT [63] as a transformer-based text encoder. BioClinicalBERT has been pre-trained on a large corpus of medical and clinical text, including clinician notes, and has been consistently shown to out-perform non-medical language models on clinical tasks [62, 64]. We extract vector representations from each used model for late fusion (concatenation) as input into a final non-linear classifier head (Supplementary Figure 1), thus using *both* visual and textual inputs. To address the significant class imbalance, all models were trained using weighted cross entropy loss with class weights computed as $\frac{N}{2 \cdot n_i}$, where N is the total number of samples and n_i denotes the number of samples in class i.

As dermatologist false negative rates have been shown to be around 1 in 20 [46, 65, 47, 66], we set the threshold for each classifier such that it achieves 95% sensitivity on the test set. We then report the resulting specificity at this threshold, and complement these results with the overall AUROC and average precision (AP) scores. AP was chosen over other metrics due to its suitability for heavily imbalanced classification problems [67]. AP is defined as the area under the precision-recall curve and thus an unskilled model has an AP value equal to the ratio of the positive class (0.07 in this study), rather than 0.5 like other metrics such as AUROC. For generalisability and improved external validation, Supplementary Table 1 presents metrics on the test set when using a decision threshold computed by requiring 95% sensitivity on the validation set. We use Integrated Gradients [68] to compute feature attributions for our best performing multimodal models, using the all-zero tensor as a baseline for the image input and the PAD token for the text input. Integrated Gradients was chosen over other explainability techniques as it can easily be extended to handle inputs of different modalities at the same time, crucial for our multimodal modelling.

Statistics and Reproducibility

All analyses were conducted using Python 3.10. Following best practice for both BioClinicalBERT and ConvNext [69, 58], models were trained with a learning rate of 1×10^{-5} for 5 epochs, with a batch size of 32. We repeat experiments with the same hyperparameters (but different random seeds) 5 times, and present both the mean and 95% confidence intervals of these 5 runs for each performance metric in Tables 4, 5, and 6.

Results

After filtering down to eligible patients (Figure 1) our retrospective dataset includes 5,481 images from 4,538 patients. Patients are aged 18 to 99 years (mean: 55; Figure 2(a)) and 62% female, with a modal Fitzpatrick Skin Score [50] of 2 (Figure 2(b)). Among these patients, 7% were diagnosed with malignant lesions, confirmed via biopsy (Supplementary Figure 2). The remaining 93% had either a clinically or histopathology confirmed benign diagnosis. Table 1 compares our dataset with other publicly available skin lesion datasets. Table 4 serves as the main results table for the paper, containing the majority of our multimodal experiments. Table 4 references Table 5 for the ablation on components of discrete metadata, and Table 6 for the ablation on leading language for surgical consultation notes. Results are visualised in Figure 3. The following subsections highlight the results of each of the tables.

Using the vision model only results in an AUROC of 0.909, a baseline result that all following experiments will be measured against.

Metadata Ablation

The results for this subsection are in Table 5. When disabling the input of the dermatoscopic image and relying only on the patient's age as input to the text model, our model achieves an AP and AUROC of 0.188 and 0.749 respectively. It is expected that age alone can yield moderate classifying power as skin cancer is more prevalent in older people [70]. Though of course one cannot judge a skin condition on the age of the patient alone, the above-random performance of this experiment is testament to age bias in such clinical scenarios. We find that the other 3 components of discrete patient metadata: Fitzpatrick Skin Score, Index of Multiple Deprivation Decile, and Sex each exhibit noticeably less classifying power than age, with models often unable to achieve the desired 95% sensitivity level without consistently predicting the positive class. The combination of all 4 discrete patient metadata yield a 0.793 AUROC (0.05 higher than with age alone). Given this knowledge of the bias contained within the metadata, it follows that when we include them alongside the dermatoscopic image, we find little change in performance compared to the vision model alone (0.923 vs 0.909).

STag Ablation

This subsection focuses on the component-wise ablation of leading language in the most powerful of our four clinical freetext groups - the **surgical consultation notes** (see Table 6).

Row 1 in Table 6 shows the result of using what remains of the surgical consultation notes after all components of leading language have been removed, finding an AUROC of 0.824 and an AP of 0.285. Rows 3 to 7 show the increase in performance of including each of the five components individually. We find the individual allowance of any benign diagnosis attempt (@DB@), or any malignant diagnosis attempt (@DM@) increases performance to the level of the standalone vision model (0.931, or 0.943 respectively). Such high performance from this text alone is unsurprising as it is expected that confident diagnosis opinions correlate heavily with the ground truth. Synonyms of ‘refer to hospital’ (@RTH@) yield a noticeably lower AUROC of 0.857. Any mention of non-hospital treatments such as prescription of medications or emollients (@T@) do not improve text model performance: 0.798 AUROC. The inclusion of all components of leading language (row 10) yields a 0.942 AUROC, with an AP of 0.531. Finally, including the dermatoscopic images with text gives an AUROC of 0.955 when all leading language is removed (0̃.05 increase over unimodal vision), and 0.975 when all components of leading language are allowed to remain, with an AP of 0.779.

Full Multimodal Ablation

Table 4 encapsulates both the majority and rest of our ablation experiments. Rows 2 through 9 show the rest of our analysis of the surgical consultation notes usage. The unfiltered surgical consultation notes *alone* (row 2) yield an AUROC of 0.963 and an AP of 0.587, an extremely high result with an increase in AUROC of over 0.05 above the vision-only baseline; this likely reflects the high levels of leading language in this text (i.e., the model may be learning textual shortcuts). Our highest level of text filtering (using Llama 3.1 to remove *all* leading language; FFilt) on rows 9 and 10 give a 0.824 AUROC on the text alone; when combined with the dermatoscopic image model, we achieve an AUROC of 0.955 and AP of 0.667. This result is higher than either modality alone (over 0.05 more than vision-only and more than 0.1 more than the FFilt surgical consultation notes only). We therefore argue that our efforts to remove leading language have yielded a model and data scenario that uses both modalities effectively, reducing the extent to which one modality inappropriately short-cuts across the other.

As detailed in Table 2, the remaining three components of clinical freetext contain significantly less leading language. When we use *only* family history of skin cancer (row 15) or exposure to sunlight (row 16), we find relatively low classification performance (APs of 0.211 and 0.192, respectively). This implies that such text alone does not offer much classifying power. The longer and highly-descriptive ‘histories and observations’ freetext (rows 11 through 14) contain very occasional leading language in the form of casual references to an opinion of the skin lesion *e.g.* “this **mole** is small and dark...”. As this long-form freetext often contains typos and acronyms, it is unsuitable to use a regular expression to filter it (levels CFilt and DFilt). Therefore, we only consider the unfiltered histories and observations (Orig; rows 11 and 12), and the direct removal of the occasional skin disease name via Llama 3.1 (FFilt; rows 13 and 14). We find that the unfiltered histories and observations yields an AUROC of 0.818, and the FFilt filtering an AUROC of 0.782. This small difference is testament to the very mild and occasional leading language contained in these freetext fields.

The final 4 rows of the table (18 through 21) combine all components of the clinical freetext and metadata with and without images. We see from rows 18 and 19 that the unfiltered text components yield results on par with the unfiltered surgical consultation notes, with or without the image as inputs. This highlights the significant power of uncontrolled leading language and the ‘shortcutting’ it can facilitate. Row 20 shows that the full multimodal combination of *filtered* text yields an AUROC of 0.870, and the inclusion of images in row 21 giving an AUROC of 0.948.

Supplementary Figures 5 and 6 show Integrated Gradient explanations for our best multimodal model on a subset of the ISIC 2020 [18, 19] dataset. Both figures reveal nuances in the freetext explainability, notably that some importance is attributed to tabular feature names (e.g., ‘fitzpatrick’). This partly reflects limitations of the ISIC dataset, and the strengths of the dataset in this paper: because ISIC lacks freetext, we generated synthetic descriptions, potentially introducing artefacts into both model outputs and model interpretability. This phenomenon also relates to the way the Transformer architecture handles attention, and suggests that the presence of a feature in text may carry some weight for the model (e.g., the fact the Fizpatrick score is not missing is information itself).

Despite these caveats, the freetext explanations exhibit trends consistent with clinical reasoning. For instance, in both figures, low Fitzpatrick skin types are highlighted in green, indicating an association with malignant diagnosis. This is consistent with the higher incidence of skin cancer in Fitzpatrick types I and II [71]. Similarly, in Supplementary Figure 6, terms such as 'asymmetrical' and the information that the lesion has grown are emphasized, again reflecting clinical practice.

The image-based explanations also demonstrate alignment with clinically relevant patterns. For example, much of the highlighted image regions in Supplementary Figure 6 are around lesion borders, mimicking areas clinicians are trained to examine [72]. While it is difficult to make further conclusions from these explainability images (it is difficult to know, for example, what area of a lesion one would look at to decide if it is asymmetric, or if you are assessing its colour), it is encouraging that some logical conclusions can be drawn. Although image attributions in Supplementary Figure 5 appear noisier, similar border-focused patterns emerge in the second and third samples, while the fourth sample shows concentrated attention on the lesion centre. The noisier nature of benign diagnosis explanations may also be due to there not (necessarily) being a defined set of features that can diagnose one of the many possible benign diagnoses - unlike malignant cases, there is a much wider variety of benign diagnoses, each of which will have its own diagnostic criteria, necessitating a wider range of attribution maps. It is also important to note that explainability values can be high due to an image feature being *missing* - e.g., that the lesion is small, or is not shaped in a certain way - and that this could also explain the noisier nature of the benign samples.

Discussion

Our ultimate result is that a full multimodal combination of all relevant data modalities present in our skin lesion data set, under the most thorough debiasing and filtering techniques, yields a model with an AUROC of 0.948 and a specificity of 71.55% for the desired fixed sensitivity threshold of 94.81%. We note that the sometimes large confidence intervals for specificity are due to differences in model calibration between runs that occasionally affect the performance of finding a suitable decision threshold for a 94.81% sensitivity. The smaller confidence intervals for AP and AUROC indicate the models achieve similar levels of performance across all runs.

By analysing model error rates across patient age and Fitzpatrick Skin Score (Figure 4), we have shown that our Vision only models exhibit low levels of bias across the protected characteristics we can measure in our dataset, despite some bias being present in the training data. The low levels of bias shown on Fitzpatrick Skin Scores is of particular importance, as it suggests the model has successfully learned clinically relevant features that are relevant to specific skin types [73, 74]. Supplementary Figure 3 shows similar levels of bias across our fully multimodal model, highlighting that the addition of text and patient metadata to the model does not increase model bias.

Our CFilt and DFilt filtering experiments used regular expressions to filter out all names of skin diseases and then further the words 'benign' and 'malignant' respectively (rows 4 through 7, Table 4). Though one might assume that removing such forms of leading language would result in a large performance drop, we found only a small impact *i.e.,* none of these experiments drop below 0.939 AUROC. This motivates our much more thorough and nuanced approach to control for components of leading language. Our Semantic Tagging experiments in Table 6 are crucial in empirically proving that synonyms of language such as 'refer to hospital' (@RTH@) and 'no treatment recommendations' (@NT@) retain nearly as much classifying power as an outright diagnosis attempt.

These results underline the importance of thoroughly analysing all training data for ML models, particularly multimodal models which we have shown to be especially susceptible to (modality) bias. The higher performance of the models using unfiltered freetext underscores the susceptibility of freetext modelling to exhibit data bias and label leakage found in previous studies [42, 43]. Nevertheless, through the application of our proposed freetext pre-processing techniques we have also shown that this data can greatly, and safely, improve the performance of ML models for skin lesion classification. We hope that the relative ease-of-use of our LLM-based text pre-processing pipeline encourages its use across other medical ML applications where the inclusion of freetext could be beneficial to model performance.

Our work in this paper introduces the modality of clinical freetext to the task of skin lesion classification. We have demonstrated that certain pieces of clinical text with leading language are a very powerful addition to a multimodal classifier. We highlight subtle difficulties if one wishes to control for such language to varying degrees, and have proposed an LLM based approach that we show can successfully control for leading language. We show that such efforts yield a significant performance boost over using only visual inputs, demonstrating that clinical freetext of many different types are powerful additions to a skin lesion classification system. We have also discussed how models trained with different levels of text pre-processing could be deployed in different scenarios. For example, our models trained at the FFilt level of pre-processing may be suitable for deployment direct to patients, whereas those trained at the CFilt/DFilt level may be more appropriate within primary care, after an initial clinical assessment. The high sensitivity achieved by our models (Table 4 and Supplementary Figure 4) ensures the models are suitable for clinical workflows, where the cost of missing a skin cancer is much higher than that of missing a true negative diagnosis.

As these models use dermatoscopic images as part of their diagnosis, they would be best placed within primary care to support improved referral efficiency to dermatologists. However, prior to their use in clinical practice, more model evaluation is needed. Prospective analysis of the use of the proposed models in clinical practice are required to fully understand their real-world impact and performance. While our models achieve both high sensitivity and specificity levels, our techniques are tuned to a high sensitivity (i.e., reduce false negatives and the number of missed cancers). Supplementary Figure 4 highlights that, in term of raw numbers, this comes at the cost of many false positives as the benign case-load overwhelms the malignant case-load. However, previous retrospective studies show that referrals of high-risk lesions from primary care to dermatology services have even higher false positive rates; in the UK, studies have shown that up to 87% of patients referred by GPs with suspected skin cancer will receive a benign diagnosis at histopathology [75, 76, 6]. In comparison, at a fixed sensitivity of 95% (compared to the 70-88% sensitivity reported for primary care clinicians [6]), our proposed techniques only have a false positive rate of 29%. This further supports using this model in primary care to reduce false positives and improve the referral accuracy beyond the current standard of care.

Nevertheless, direct comparisons of model and dermatologist/primary care physician performance (depending on the type of model being used) are necessary to uncover our proposed techniques' impact on clinical workflows. Additionally, although our dataset was collected from a range of clinics across the UK, our model's generalisability to other datasets (with different demographics and case mix) should also be investigated in future work. The application of model explainability techniques have shown that these attribution methods can be used to help explain model decisions, possibly aiding clinical decision making further. However, these techniques alone are not sophisticated enough to ensure models are *consistently* applying clinically relevant decisions, and future work should develop techniques to explore this. Furthermore, patients and the public should be involved in these prospective evaluations to ensure the model's impact on the patient experience is explored.

Conclusion

Thorough and component-wise analysis throughout our benchmark aims to inform future research using clinical freetext with a variety of different data scenarios. We wish to encourage future research integrating clinical freetext into medical visual classification scenarios to carefully consider components of text they wish to integrate, and to be vigilant and resourceful in any pre-processing they consider. By incorporating appropriately filtered clinical freetext from electronic health records we have demonstrated a boost in performance on top of vision only models that in the future could be used improve the referral efficiency and diagnostic accuracy of skin lesions within primary care.

References

[1] Katelyn Urban, Sino Mehrmal, Prabhdeep Uppal, Rachel L Giesey, and Gregory R Delost. The global burden of skin cancer: A longitudinal analysis from the global burden of disease study, 1990–2017. *JAAD international*, 2:98–108, 2021.

[2] Office for National Statistics. Cancer survival in england: Latest bulletin, 2024. URL https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancersurvivalinengland/latest. Accessed: 2025-01-14.

[3] Kesha J Buster, Erica I Stevens, and Craig A Elmets. Dermatologic health disparities. *Dermatologic clinics*, 30(1):53–59, 2012.

[4] David Rowell, Kim Huong Nguyen, Peter Baade, and Monika Janda. Evaluation of a skin self-examination programme: a four-stage recursive model. *Asian Pacific Journal of Cancer Prevention: APJCP*, 18(4):1063, 2017.

[5] National Institute for Health and Care Excellence. CKS: How should i assess a lesion? https://cks.nice.org.uk/topics/melanoma/diagnosis/assessment/, 2022. [Accessed 20-03-2025].

[6] Harish Shivakumar, Upamanyu Leo Chanda, and Ogba Onwuchekwa. Evaluating the diagnostic accuracy and challenges of the two-week wait referral pathway for skin cancers in primary care. *Cureus*, 17(1), 2025.

[7] Mohamed A. Kassem, Khalid M. Hosny, Robertas Damaševičius, and Mohamed Meselhy Eltoukhy. Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review. *Diagnostics*, 11, 2021.

[8] Daniel Sauter, Georg Lodde, Felix Nensa, Dirk Schadendorf, Elisabeth Livingstone, and Markus Kukuk. Deep learning in computational dermatopathology of melanoma: A technical systematic literature review. *Computers in Biol-*

ogy and Medicine, 163:107083, 2023. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2023.107083. URL https://www.sciencedirect.com/science/article/pii/S0010482523005486.

[9] V. Vipin, Malaya Kumar Nath, V. Sreejith, Nikhil Francis Giji, Adithya Ramesh, and M. Meera. Detection of melanoma using deep learning techniques: A review. In *2021 International Conference on Communication, Control and Information Sciences (ICCISc)*, volume 1, pages 1–8, 2021. doi: 10.1109/ICCISc52257.2021.9484861.

[10] Stefania Seidenari, Giovanni Pellacani, and Alberto Giannetti. Digital videomicroscopy and image analysis with automatic classification for detection of thin melanomas. *Melanoma Research*, 9(2):163–172, 1999.

[11] Peter A Lio and Paul Nghiem. Interactive atlas of dermoscopy: Giuseppe argenziano, md, h. peter soyer, md, vincenzo de giorgio, md, domenico piccolo, md, paolo carli, md, mario delfino, md, angela ferrari, md, rainer hofmann-wellenhof, md, daniela massi, md, giampiero mazzocchetti, md, massimiliano scalvenzi, md, and ingrid h. wolf, md, milan, italy, 2000, edra medical publishing and new media. *Journal of the American Academy of Dermatology*, 50(5):807–808, 2004.

[12] Teresa Mendonça, Pedro M. Ferreira, Jorge S. Marques, André R. S. Marcal, and Jorge Rozeira. Ph2 - a dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5437–5440, 2013. doi: 10.1109/EMBC.2013.6610779.

[13] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. *Color medical image analysis*, pages 63–86, 2013.

[14] Philipp Tschandl. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, 2018. URL https://doi.org/10.7910/DVN/DBW86T.

[15] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[16] Marc Combalia, Noel C. F. Codella, Veronica M Rotemberg, Brian Helba, Verónica Vilaplana, Ofer Reiter, Allan C. Halpern, Susana Puig, and Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild. *ArXiv*, abs/1908.02288, 2019.

[17] Marc Combalia, Noel C. F. Codella, Veronica M Rotemberg, Brian Helba, Verónica Vilaplana, Ofer Reiter, Allan C. Halpern, Susana Puig, and Josep Malvehy. Bcn20000 collection (isic archive, collection 249). https://api.isic-archive.com/collections/249/. Accessed: 2025-09-25.

[18] 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.

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

[20] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. Analysis of the isic image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis*, 75:102305, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.102305. URL https://www.sciencedirect.com/science/article/pii/S1361841521003509.

[21] Bless Lord Y. Agbley, Jianping Li, Amin Ul Haq, Edem Kwedzo Bankas, Sultan Ahmad, Isaac Osei Agyemang, Delanyo Kulevome, Waldiodio David Ndiaye, Bernard Cobbinah, and Shoistamo Latipova. Multimodal melanoma detection with federated learning. In *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 238–244, 2021. doi: 10.1109/ICCWAMTIP53232.2021.9674116.

[22] Sutong Wang, Yunqiang Yin, Dujuan Wang, Yanzhang Wang, and Yaochu Jin. Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis. *IEEE Transactions on Cybernetics*, 52(12):12623–12637, 2022. doi: 10.1109/TCYB.2021.3069920.

[23] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019. doi: 10.1109/JBHI.2018.2824327.

[24] Yuheng Wang, Jiayue Cai, Daniel C Louie, Z Jane Wang, and Tim K Lee. Incorporating clinical knowledge with constrained classifier chain into a multimodal deep network for melanoma detection. *Computers in Biology and Medicine*, 137:104812, 2021.

[25] Lucas Schneider, Christoph Wies, Eva I. Krieghoff-Henning, Tabea-Clara Bucher, Jochen S. Utikal, Dirk Schadendorf, and Titus J. Brinker. Multimodal integration of image, epigenetic and clinical data to predict braf mutation status in melanoma. *European Journal of Cancer*, 183:131–138, 2023. ISSN 0959-8049. doi: https://doi.org/10.1016/j.ejca.2023.01.021. URL https://www.sciencedirect.com/science/article/pii/S095980492300045X.

[26] Wanqiu Zhang, Nathan Heath Patterson, Nico Verbeeck, Jessica L Moore, Alice Ly, Richard M Caprioli, Bart De Moor, Jeremy L Norris, and Marc Claesen. Multimodal maldi imaging mass spectrometry for improved diagnosis of melanoma. *medRxiv*, pages 2022–11, 2022.

[27] Xin Lai, Jinfei Zhou, Anja Wessely, Markus Heppt, Andreas Maier, Carola Berking, Julio Vera, and Le Zhang. A disease network-based deep learning approach for characterizing melanoma. *International Journal of Cancer*, 150(6):1029–1044, 2022.

[28] Xiaogen Zhou, Yiyou Sun, Min Deng, Winnie Chiu Wing Chu, and Qi Dou. Robust semi-supervised multimodal medical image segmentation via cross modality collaboration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 57–67. Springer, 2024.

[29] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

[30] Rhys Compton, Lily Zhang, Aahlad Puli, and Rajesh Ranganath. When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations. In *Machine learning for healthcare conference*, pages 110–127. PMLR, 2023.

[31] Susu Sun, Lisa M Koch, and Christian F Baumgartner. Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 425–434. Springer, 2023.

[32] Noa P. Cruz Díaz, Manuel J. Maña López, Jacinto Mata Vázquez, and Victoria Pachón. A machine-learning approach to negation and speculation detection in clinical texts. *J. Assoc. Inf. Sci. Technol.*, 63:1398–1410, 2012. URL https://api.semanticscholar.org/CorpusID:17246092.

[33] Richárd Farkas, Veronika Vincze, György Móra, János A. Csirik, and György Szarvas. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. In *CoNLL Shared Task*, 2010. URL https://api.semanticscholar.org/CorpusID:549335.

[34] Kazuki Fujikawa, Kazuhiro Seki, and Kuniaki Uehara. A hybrid approach to finding negated and uncertain expressions in biomedical documents. In *MIXHS '12*, 2012. URL https://api.semanticscholar.org/CorpusID:15585121.

[35] Shaodian Zhang, Tian Kang, Xingting Zhang, Dong Wen, Noémie Elhadad, and Jianbo Lei. Speculation detection for chinese clinical notes: Impacts of word segmentation and embedding models. *Journal of biomedical informatics*, 60, 02 2016. doi: 10.1016/j.jbi.2016.02.011.

[36] David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, 177:105122, 2023.

[37] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[38] Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120, 2020.

[39] Ankita Agarwal, Tanvi Banerjee, William L Romine, and Mia Cajita. Debias-clr: A contrastive learning based debiasing method for algorithmic fairness in healthcare applications. In *2024 IEEE International Conference on Big Data (BigData)*, pages 6411–6419. IEEE, 2024.

[40] Yuelyu Ji, Zeshui Yu, and Yanshan Wang. Assertion detection in clinical natural language processing using large language models. *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 242–247, 2024. URL https://api.semanticscholar.org/CorpusID:267334862.

[41] Isabella Catharina Wiest, Dyke Ferber, J. Zhu, M. van Treeck, Sonja K Meyer, Radhika Juglan, Z. I. Carrero, Daniel Paech, Jens Kleesiek, M. P. Ebert, Daniel Truhn, and J. N. Kather. From text to tables: A local privacy preserving large language model for structured information retrieval from medical documents. In *medRxiv*, 2023. URL https://api.semanticscholar.org/CorpusID:266082605.

[42] Sharon E Davis, Michael E Matheny, Suresh Balu, and Mark P Sendak. A framework for understanding label leakage in machine learning for health care. *Journal of the American Medical Informatics Association*, 31(1):274–280, 2023.

[43] Anshul Thakur, Tingting Zhu, Vinayak Abrol, Jacob Armstrong, Yujiang Wang, and David A Clifton. Data encoding for healthcare data democratization and information leakage prevention. *Nature Communications*, 15(1):1582, 2024.

[44] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[45] Michael F Chiang, John C Hwang, C Yu Alexander, Daniel S Casper, James J Cimino, and Justin Starren. Reliability of snomed-ct coding by three physicians using two terminology browsers. In *AMIA Annual Symposium Proceedings*, volume 2006, page 131, 2006.

[46] Hafsa Faarax Shirwac, Hannah Morgan, Huw Greenish, Sarah Carswell, Harry Heath, James Koutsis, Pascale Guitera, and Alexander DG Anderson. P023 real-world melanoma miss rate in an nhs hospital dermatology service. *British Journal of Dermatology*, 191(Supplement_1):i24–i25, 2024.

[47] Frans HJ Rampen, Idy JAMG Casparie-van Velsen, Barbara EWL van Huystee, Lambertus ALM Kiemeney, and Leo J Schouten. False-negative findings in skin cancer and melanoma screening. *Journal of the American Academy of Dermatology*, 33 (1):59–63, 1995.

[48] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[49] Rajmadhan Ekambaram, Dmitry B Goldgof, and Lawrence O Hall. Finding label noise examples in large scale datasets. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2420–2424. IEEE, 2017.

[50] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

[51] UK Government Office for National Statistics. English indices of deprivation 2015 — gov.uk. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015, 2015. [Accessed 19-02-2025].

[52] Jason E Hawkes, Amanda Truong, and Laurence J Meyer. Genetic predisposition to melanoma. In *Seminars in oncology*, volume 43, pages 591–597. Elsevier, 2016.

[53] Vishal Madan, John T Lear, and Rolf-Markus Szeimies. Non-melanoma skin cancer. *The lancet*, 375(9715):673–685, 2010.

[54] H Kittler, M Seltenheim, M Dawid, H Pehamberger, K Wolff, and M Binder. Morphologic changes of pigmented skin lesions: a useful extension of the abcd rule for dermatoscopy. *Journal of the American Academy of Dermatology*, 40(4):558–562, 1999.

[55] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[56] Hugo Touvron, Louis Martin, Kevin Stone, and et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

[57] Duy Duc An Bui and Qing Zeng-Treitler. Learning regular expressions for clinical text classification. *Journal of the American Medical Informatics Association*, 21(5):850–857, 2014.

[58] Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022.

[59] Zhimeng Han, Muwei Jian, and Gai-Ge Wang. Convunext: An efficient convolution neural network for medical image segmentation. *Knowledge-based systems*, 253:109512, 2022.

[60] Tolgahan Gulsoy and Elif Baykal Kablan. Focalnext: A convnext augmented focalnet architecture for lung cancer classification from ct-scan images. *Expert Systems with Applications*, 261:125553, 2025.

[61] Md Saiful Islam Sajol, Syada Tasmia Alvi, and Chowdhury Abida Anjum Era. Performance assessment of advanced cnn and transformer architectures in skin cancer detection. In *2024 11th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pages 1–8. IEEE, 2024.

[62] Matthew Watson, Stelios Boulitsakis Logothetis, Darren Green, Mark Holland, Pinkie Chambers, and Noura Al Moubayed. Performance of machine learning versus the national early warning score for predicting patient deterioration risk: a single-site study of emergency admissions. *BMJ Health & Care Informatics*, 31(1):e101088, 2024.

[63] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[64] Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. Evaluating pretraining strategies for clinical BERT models. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 410–416, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.43/.

[65] Edge Health. Evaluating pathways for ai dermatology in skin cancer detection: A white paper. Technical report, NHS England Outpatient Recovery and Transformation Programme, July 2024. URL https://www.edgehealth.co.uk/wp-content/uploads/2024/08/Evaluating-Pathways-for-AI-Dermatology-in-Skin-Cancer-Detection.pdf.

[66] Suephy C Chen, Dena M Bravata, Evette Weil, and Ingram Olkin. A comparison of dermatologists' and primary care physicians' accuracy in diagnosing melanoma: a systematic review. *Archives of dermatology*, 137(12):1627–1634, 2001.

[67] Jean-Gabriel Gaudreault, Paula Branco, and João Gama. An analysis of performance metrics for imbalanced classification. In *International Conference on Discovery Science*, pages 67–77. Springer, 2021.

[68] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[69] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[70] Elliot J Coups, Alan C Geller, Martin A Weinstock, Carolyn J Heckman, and Sharon L Manne. Prevalence and correlates of skin cancer screening among middle-aged and older white adults in the united states. *The American journal of medicine*, 123(5): 439–445, 2010.

[71] Adèle Green and Diana Battistutta. Incidence and determinants of skin cancer in a high-risk australian population. *International journal of cancer*, 46(3): 356–361, 1990.

[72] Ana F Duarte, Bernardo Sousa-Pinto, Luís F Azevedo, Ana M Barros, Susana Puig, Josep Malvehy, Eckart Haneke, and Osvaldo Correia. Clinical abcde rule for early melanoma detection. *European journal of dermatology*, 31(6):771–778, 2021.

[73] Hugh M Gloster Jr and Kenneth Neal. Skin cancer in skin of color. *Journal of the American Academy of Dermatology*, 55(5):741–760, 2006.

[74] Katina M Byrd, Dawn C Wilson, Suzanna S Hoyler, and Gary L Peck. Advanced presentation of melanoma in african americans. *Journal of the American Academy of Dermatology*, 50(1):21–24, 2004.

[75] NH Cox. Evaluation of the uk 2-week referral rule for skin cancer. *British Journal of Dermatology*, 150 (2):291–298, 2004.

[76] A Haile-Redai and J O'Connor. Diagnostic accuracy amongst two week wait referrals for skin malignancy. *J Dermatol Res Ther*, 7:111, 2021.

[77] Teresa Mendonça, Pedro M. Ferreira, Jorge S. Marques, André R. S. Marcal, and Jorge Rozeira. Ph^2 database. https://www.fc.up.pt/addi/ph2%20database.html, 2013. Accessed: 2025-09-25.

[78] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. Dermofit image library. https://licensing.edinburgh-innovations.ed.ac.uk/product/dermofit-image-library. Accessed: 2025-09-25.

[79] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. 7-point criteria evaluation database (derm7pt). https://derm.cs.sfu.ca/Welcome.html, 2025. Accessed: 2025-09-25.

[80] David A. Gutman, Noel C. F. Codella, M. E. Celebi, Brian Helba, Michael Armando Marchetti, Nabin K. Mishra, and Allan C. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172, 2016.

[81] 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.

[82] National Library of Medicine. Medical Subject Headings (MeSH). https://www.ncbi.nlm.nih.gov/mesh, 2024. Accessed: 2024-11-27.

Acknowledgements

This work was supported by Innovate UK grant number 10027358.

Author Contributions

All authors contributed to the design of the study. MW and TW conducted the experiments and analysis, which were designed and conceptualised by all authors. All authors interpreted results. MW and TW drafted the manuscript, and all authors critically revised the manuscript for important content.

Ethics Statement

This was a retrospective analysis of data that is collected as part of routine clinical practice for the purpose of improving provision of therapeutic medicine and related research. A redacted and fully anonymised data set was then provided by Healthcare service provider, Omnes Healthcare, to the University of Durham and Evergreen Life Ltd in order to develop algorithmic (AI) tools to support and improve the diagnosis of skin conditions within their service provision. The related research described in this paper was undertaken as part of the development of this algorithmic tool. The above process has been reviewed by the companies DPOs and is in line with UK GDPR requirements. It is part of Omnes Healthcare's usual business practice and services as a healthcare provider. Omnes Healthcare's Institutional Review Board approved this study, and waived the need for informed consent for secondary use of patient data due to all data being fully anonymised.

Data Availability

The dataset may be made available upon reasonable request the corresponding author. Individual performance metrics for each of the 5 runs of all experiments are available in the Supplementary Material as `raw-experiment-results.csv`; this is source data for Figure 5, Tables 4-6.

Code Availability

Supporting code may be made available upon reasonable request to the corresponding author.

Competing Interests

All authors declare no competing interests.

Figure Information

For each figure, the title is in **bold**, followed by the figure description and any acronyms used.

- Figure 1 - **Patient inclusion/exclusion criteria**. Filtering performed on our anonymised dataset.
- Figure 2 - **Dataset patient demographics**. Figures reporting (a) patient sex, age, and (b) Fitzpatrick Skin Score.
- Figure 3 - **Model performance across all experiments**. Horizontal line plot of AUROC scores. The bars, which are coloured by experiment group, report the mean AUROC across 5 runs for that experiment. Error bars indicate the 95% confidence interval for the mean, and the AUROC for each individual run is plot as a grey dot. AUROC: Accuracy under the Receiver Operating Characteristic curve; FSS: Fitzpatrick Skin Score; IMDD: Index of Multiple Deprivation; MD: Metadata; V: Vision; Surg Cons: Surgical Consultation Notes; Orig: Original text; CFilt: ConditionFiltered text; DFilt: DiagnosisFiltered text; FFilt: FullyFiltered text; Hist & Obs: History and Observations; Fam Hist of Skin Cancer: Family History of Skin Cancer.
- Figure 4 - **Model bias across patient subgroups**. Vision only model error rates across (a) patient age and (b) patient Fitzpatrick Skin Score (on test set, 1,123 images).

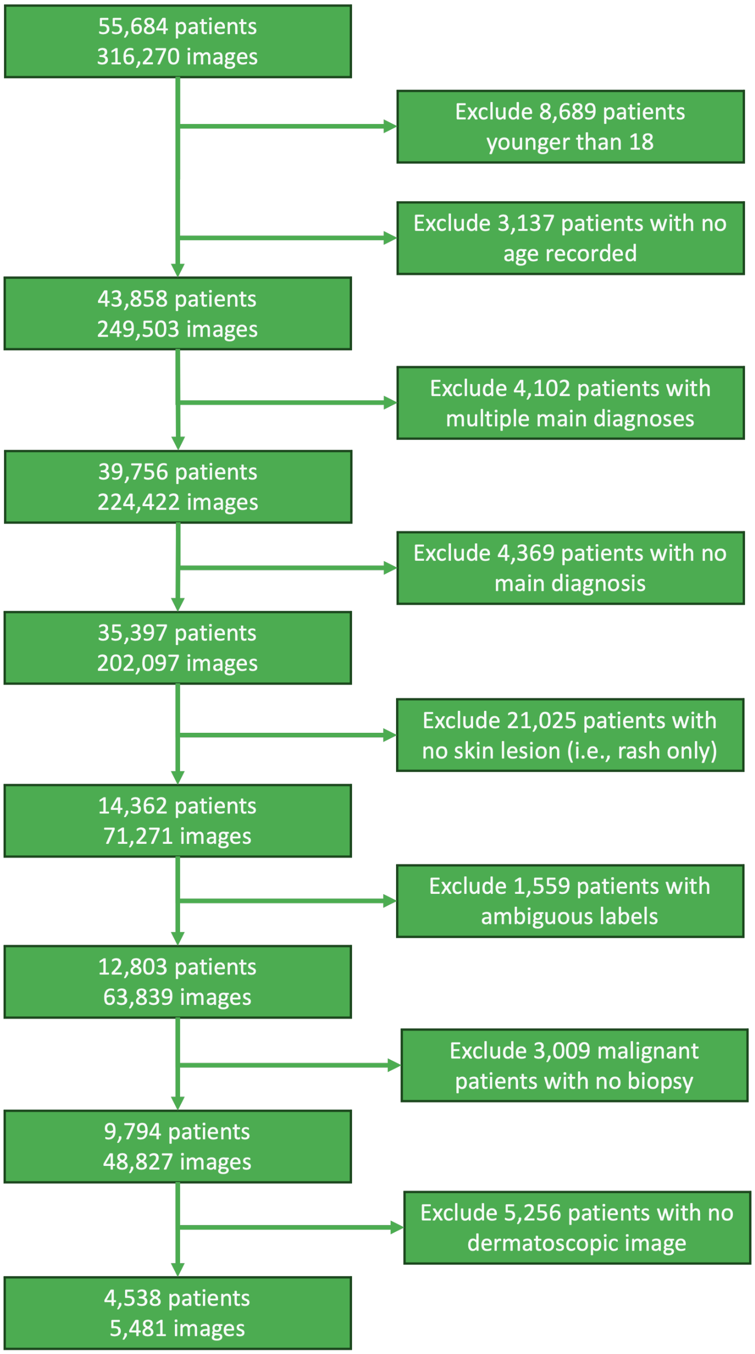
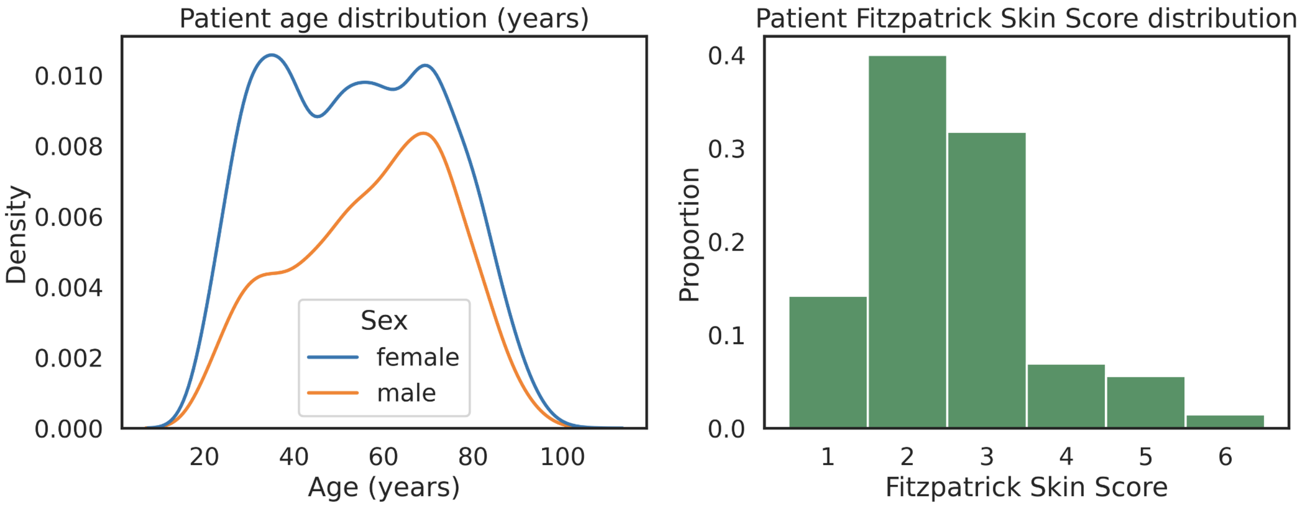

55,684 patients
316,270 images
Exclude 8,689 patients younger than 18
Exclude 3,137 patients with no age recorded
43,858 patients
249,503 images
Exclude 4,102 patients with multiple main diagnoses
39,756 patients
224,422 images
Exclude 4,369 patients with no main diagnosis
35,397 patients
202,097 images
Exclude 21,025 patients with no skin lesion (i.e., rash only)
14,362 patients
71,271 images
Exclude 1,559 patients with ambiguous labels
12,803 patients
63,839 images
Exclude 3,009 malignant patients with no biopsy
9,794 patients
48,827 images
Exclude 5,256 patients with no dermatoscopic image
4,538 patients
5,481 images

Figure 1: Filtering performed on our anonymised dataset.



(a)

(b)

Figure 2: Figures reporting (a) patient sex, age, and (b) Fitzpatrick Skin Score.

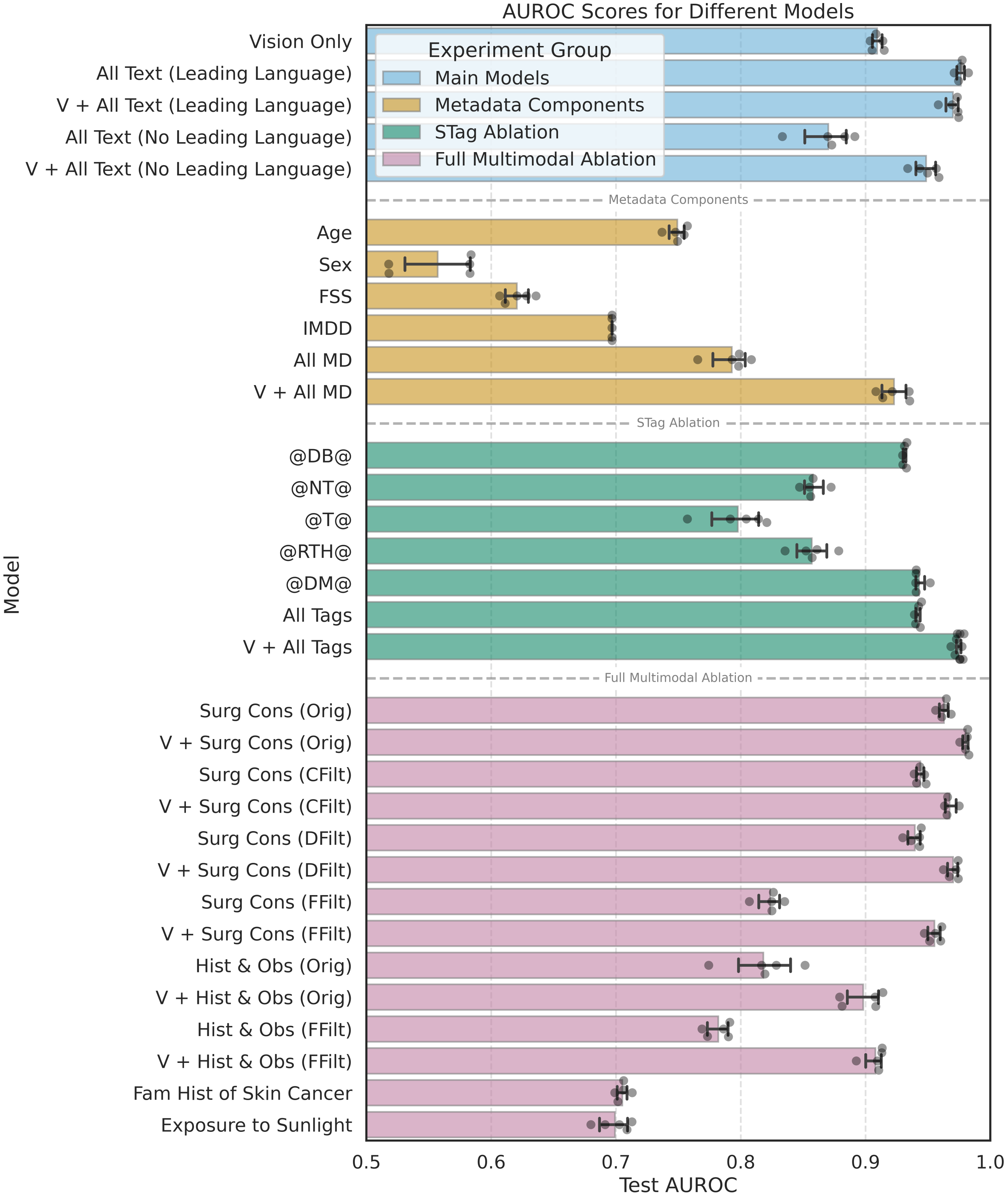
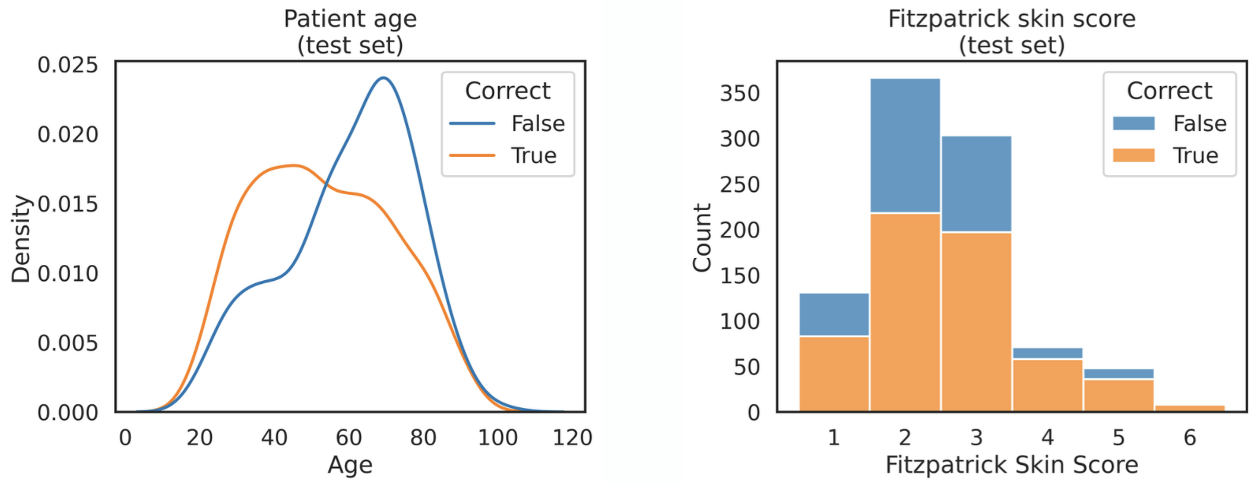


Figure 3: Horizontal line plot of AUROC scores. The bars, which are coloured by experiment group, report the mean AUROC across 5 runs for that experiment. Error bars indicate the 95% confidence interval for the mean, and the AUROC for each individual run is plot as a grey dot.



(a)

(b)

Figure 4: Vision only model error rates across (a) patient age and (b) patient Fitzpatrick Skin Score (on test set, 1,123 images).

			Dermatoscopic Image	Sex	Age	Fitzpatrick Skin Score	Index Multiple Deprivation Decile	Surgical Consultation Notes	History and Observations	Family History of Skin Disease	Exposure to Sun
Dataset	**Year**	**Size**	**Data Sources**								
			Image	**Metadata**				**Text**			
DBDermo-MIPS [10]	1999	424	✓								
Interactive Atlas of Dermoscopy [11]	2004	~2000	✓								
PH2 [12, 77]	2013	200	✓								
Dermofit Image Library [13, 78]	2013	1300	✓								
7-Point Criteria Database [23, 79]	2017	2045	✓								
HAM10000 [14, 15]	2018	10,015	✓	✓	✓						
BCN20000 [16, 17]	2019	20,000	✓	✓	✓						
ISIC 2017 [80, 19] (Tasks 1,2,3)	2017	2600	✓								
ISIC 2018 [14, 81, 19] (Tasks 1,2)	2018	3695	✓								
ISIC 2018 [14, 81, 19] (Task 3)	2018	11,720	✓								
ISIC 2019 [14, 80, 16, 19] (Tasks 1,2)	2019	33,569	✓	✓	✓						
ISIC 2020 [18, 19]	2020	43,683	✓	✓	✓						
Ours	2025	5481	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: A comparison of skin lesion classification datasets. Our dataset includes additional patient metadata and, most notably, freetext clinician notes.

	History and Observations	**Family History of Skin Cancer**	**Exposure to Sunlight**	**Surgigcal Consultation Notes**
Average Length	47.84 words	2.82 words	4.40 words	23.86 words
Leading Language	Occasional (Table 3)	None	None	Substantial (Table 3)
Lexical Diversity	94.5%	78.2%	91.4%	84.0%
Medical Density	3.3%	2.2%	3.0%	26.0%
Total Vocab	6909 words	1527 words	2410 word	1828 words
Examples	"Patient says lesion is getting darker, wider, raised, flaky, smooth, not itchy" "It is raised and dark in colour. It has grown in size."	"None"; "No history" "Patient's 3rd cousin has skin cancer" "father had cancerous mole removed from leg"	"Works outside" "Used to live in Brazil" "travels abroad on holidays" "doesn't wear sun cream"	"lesion right side of nose looks like benign seborrhoeic keratosis. no treatment is required." "Basal cell carcinoma, refer to hospital"

Table 2: Statistics and examples for the 4 components of clinical freetext used in our experiments. Lexical diversity is the average proportion of unique words in a each component. Medical density is the average proportion of words in each sentence that are 'medical' terms from the medical subject headings (MeSH) [82].

Preprocess Level	Relevant Text Components	Description

Orig	Surgery Consult Notes;
	History & Observations;
	Exposure to Sunlight;
	Family History of Cancer;
		No changes or filtering.

Basic Preprocessing (RegEx)		
CFilt	Surgery Consult Notes;	Remove the names of any skin condition
DFilt	Surgery Consult Notes;	**Lvl 2** + remove the words 'benign' and 'malignant'
Advanced Preprocessing (Manual/LLM)		
STag	Surgery Consult Notes;	Replace the following 6 components of text with tags. Any combination of these tags can then be removed or included at training time. • **@DB@**: Benign diagnosis. • **@DM@**: Malignant diagnosis. • **@RTH@**: Any explicit 'refer to hospital' or 'refer for biopsy'. • **@NT@**: Any words to the effect of 'no treatment required'. • **@T@**: Any recommendation of non-biopsy treatment *e.g.* creams, avoid sunlight.

FFilt	Surgery Consult Notes;
	History & Observations;
		Remove any language implying, treatment, or any of the text above are removed. Only statements of known facts are allowed.

Table 3: Levels of preprocessing of the clinical freetext used for thorough analysis of performance. Translation from raw text to STag and FFilt were done via a Llama 3.1 8B LLM model with manual supervision.

	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	62.60% ±2.85%	94.81% ±0.00%	0.909 ±0.006	Vision Only
	Orig								0.587 ±0.039	90.80% ±2.07%	94.81% ±0.00%	0.963 ±0.006	Surg Cons (Orig)
✓	Orig								0.802 ±0.047	92.03% ±1.45%	94.81% ±0.00%	0.980 ±0.004	V+Surg Cons (Orig)
	CFilt								0.424 ±0.021	86.08% ±1.64%	94.81% ±0.00%	0.944 ±0.005	Surg Cons (CFilt)
✓	CFilt								0.658 ±0.096	89.22% ±1.16%	94.81% ±0.00%	0.967 ±0.007	V+Surg Cons (CFilt)
	DFilt								0.405 ±0.021	83.59% ±4.38%	94.81% ±0.00%	0.939 ±0.008	Surg Cons (DFilt)
✓	DFilt								0.726 ±0.068	89.67% ±3.21%	94.81% ±0.00%	0.970 ±0.006	V+Surg Cons (DFilt)
	STag								**Ablation of Leading Language (Lvl4T) Explored in Table 6**				
	FFilt								0.285 ±0.005	25.18% ±4.95%	94.81% ±0.00%	0.824 ±0.013	Surg Cons (Lvl4)
✓	FFilt								0.667 ±0.026	79.45% ±7.04%	94.81% ±0.00%	0.955 ±0.007	V+Surg Cons (Lvl4)
		Orig							0.267 ±0.012	38.01% ±12.3%	94.81% ±0.00%	0.818 ±0.035	Hist & Obs (Orig)
✓		Orig							0.616 ±0.043	50.75% ±4.58%	94.81% ±0.00%	0.898 ±0.020	V+Hist & Obs (Orig)
		FFilt							0.257 ±0.016	25.76% ±7.74%	94.81% ±0.00%	0.782 ±0.013	Hist & Obs (Lvl4)
✓		FFilt							0.560 ±0.02	62.18% ±8.12%	94.81% ±0.00%	0.908 ±0.011	V+Hist & Obs (Lvl4)
			✓						0.211 ±0.003	2.39% ±4.10%	94.81% ±0.00%	0.705 ±0.007	Fam Hist of Skin Cancer
				✓					0.192 ±0.018	15.70% ±4.25%	94.81% ±0.00%	0.699 ±0.017	Exposure to Sunlight
					✓				**Ablation of Metadata Explored in Table 5**				
	Orig	Orig	✓	✓	✓	✓	✓	✓	0.734 ±0.081	91.72% ±1.59%	94.81% ±0.00%	0.976 ±0.005	All Text (Leading Language)
✓	Orig	Orig	✓	✓	✓	✓	✓	✓	0.747 ±0.094	87.06% ±6.32%	94.81% ±0.00%	0.970 ±0.009	V+All Text (Leading Language)
	FFilt	FFilt	✓	✓	✓	✓	✓	✓	0.444 ±0.033	33.69% ±10.9%	94.81% ±0.00%	0.870 ±0.028	All Text (No Leading Language)
✓	FFilt	FFilt	✓	✓	✓	✓	✓	✓	0.679 ±0.038	71.55% ±15.6%	94.81% ±0.00%	0.948 ±0.013	V+All Text (No Leading Language)

Table 4: The main results table for the paper. 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. The label key entries correspond to plots in Figure 3. 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.

	Age	**Sex**	**FSS**	**IMDD**	**Averaged Metrics (5 Runs)**				**Label Key**
Vision	**Text Components**				**AP**	**Spec**	**Sens**	**AUROC**	
	✓				0.188 ±0.009	23.02% ±5.44%	94.81% ±0.00%	0.749 ±0.010	Age
		✓			0.113 ±0.046	—-% ± – %	—-% ± – %	0.557 ±0.04	Sex
			✓		0.178 ±0.002%	6.31% ±7.29%	94.81% ±0.00%	0.621 ±0.015	FSS
				✓	0.135 ±0.00%	—-% ± – %	—-% ± – %	0.697 ±0.00	IMDD
	✓	✓	✓	✓	0.318 ±0.035	26.12% ±3.36%	94.81% ±0.00%	0.793 ±0.020	All MD
✓	✓	✓	✓	✓	0.636 ±0.034	65.58% ±5.45%	94.81% ±0.00%	0.923 ±0.015	V+All MD

Table 5: Metadata inclusion table. The label key entries correspond to plots in Figure 3. Each result is comprised from 5 runs with exactly the same experimental setup and hyperparameters. Mean and 95% confidence interval width of each metric is reported. Where sensitivity and specificity values are unavailable, the models were unable to achieve 95% sensitivity without consistently predicting the positive class. AP: Average Precision (baseline AP: 0.07); AUROC: Area Under the Receiver Operating Characteristic Curve.

	Heavily Implies Benign	Heavily Implies Benign	Implies Benign	Implies Malignant	Heavily Implies Malignant					
	Benign Diagnosis (@DB@)	No Treatment Required (@NT@)	Treatments Recommended (@T@)	Refer to Hospital (@RTH@)	Malignant Diagnosis (@DM@)	Averaged Metrics (5 Runs)				Label Key
Vision	Surgical Consultation Notes: STag					AP	Spec	Sens	AUROC	
						0.285 ±0.005	25.18% ±4.95%	94.81% ±0.00%	0.824 ±0.013	No Tags
✓						0.667 ±0.026	79.45% ±7.04%	94.81% ±0.00%	0.955 ±0.007	
	✓					0.476 ±0.008	74.76% ±0.796%	94.81% ±0.00%	0.931 ±0.002	@DB@
		✓				0.304 ±0.008	44.99% ±14.45%	94.81% ±0.00%	0.858 ±0.011	@NT@
			✓			0.280 ±0.013	16.42% ±9.89%	94.81% ±0.00%	0.798 ±0.031	@T@
				✓		0.307 ±0.016	40.08% ±3.32%	94.81% ±0.00%	0.857 ±0.019	@RTH@
					✓	0.533 ±0.010	75.72% ±0.982%	94.81% ±0.00%	0.943 ±0.006	@DM@
	✓				✓	0.531 ±0.004	75.47% ±0.758%	94.81% ±0.00%	0.942 ±0.003	
		✓	✓	✓		0.305 ±0.0059	39.24% ±10.09%	94.81% ±0.00%	0.853 ±0.014	
	✓	✓	✓	✓	✓	0.531 ±0.004	75.66% ±0.479%	94.81% ±0.00%	0.942 ±0.003	All Tags
✓	✓	✓	✓	✓	✓	0.779 ±0.017	89.76% ±3.82%	94.81% ±0.00%	0.975 ±0.004	V+All Tags

Table 6: Results (mean ± 95% confidence interval width) for STag text and their accompanying tags. The label key entries correspond to plots in Figure 3. Rows with no STag show results when all 5 components of leading text have been removed. Red cells indicate experiments with dermatoscopic images included and blue cells those with the corresponding tag included. Text colour indicates the type of semantic tagging used. AP: Average Precision (baseline AP: 0.07); AUROC: Area Under the Receiver Operating Characteristic Curve.