Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study

Ahed J Alkhatib[1*]

[1]*Department of Legal Medicine, Toxicology and Forensic Medicine, Jordan University of Science & Technology, Jordan*

***Correspondence:** Alkhatib A.J, Department of Legal Medicine, Toxicology and Forensic Medicine, Jordan University of Science & Technology, Jordan; E-mail: ajalkhatib@just.edu.jo

Received date: 20 Oct 2025: **Accepted date:** 10 Nov 2025: **Published date:** 17 Nov 2025

Copyright: © 2025 Alkhatib A.J. This article is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, provided the original author(s) and source are properly credited.

Abstract

Background: There is an increasing research on automated recognition and assessment of vitiligo using AI. Nevertheless, the high classification performance may be partly due to technical differences between classes rather than disease-related morphology.

Objective: The aim of this study was to evaluate the integrity as well as the technical characteristics of a publicly available vitiligo image dataset. Besides that, it aimed to determine whether non-clinical image metadata alone can discriminate vitiligo from healthy-image labels.

Methods: We performed a systematic data-quality audit of a publicly available dataset consisting of 3,628 images classified as vitiligo or healthy. We checked the font quality, total number of classes in the dataset, file format, width and height dimensions, file size, aspect ratio, pixel count, and duplication. MD5 hashing was used to identify duplicate files and remove them prior to the modeling stage. Mann-Whitney U test and rank-biserial correlations were used to compare variables and groups on technical characteristics. A logistic regression was calculated on image width, height, file size, aspect ratio, megapixel count (mp), and bytes per pixel, as a negative-control experiment. Pixels in the image as well as the dermatological features are not included. We evaluated performance by stratified five-fold cross-validation, and estimated 95% confidence intervals (CIs) using 2,000 bootstraps.

Results: All images (3,628) were legible including the (2,100) vitiligo and (1,528) healthy images. A total of forty redundant exact copies were identified, all belonging to the class vitiligo, leaving only 3588 unique images. Technical characteristics showed major differences between classes. In the healthy category, images had an average resolution of 1600 × 1200 pixels, whereas in the vitiligo category, the average resolution was 416 × 416 pixels. The median file size was 66,374 and 17,323 bytes respectively. The classes significantly differed with regard to all the technical variables examined ($p < .001$). Interestingly, the six-variable metadata-only model achieved an area under the curve (AUC) of 0.942, with an accuracy of 0.875, sensitivity of 0.917, and specificity of 0.817 (95% CI: 0.933-0.950). A model solely based on image width and height achieved an AUC of 0.929

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences*, 1(2):1-15.

Conclusion: Technical attributes of the images in the public vitiligo dataset are not clinically-related. As a result, high internal classification correctness achieved by the data set may partly relate to source or acquisition shortcuts rather than vitiligo-specific morphology. Before determining clinical claims based on models developed using such data, we should audit the dataset for issues like duplication or deviation from established sources and standards.

Keywords: Vitiligo, Artificial intelligence, Dermatology, Dataset bias, Image classification, Shortcut learning, Machine learning, Data leakage

Introduction

Vitiligo is acquired depigmenting disorder characterized by selective loss or dysfunction of epidermal melanocytes. The causes of the illness are not solitary but are numerous. The causatives may be genetic, or it may be a malfunctioning immunity mechanism. Further, it could be productive stress, or it could be abnormal melanocyte survival [1-3]. This disorder can affect all ethnic groups and can have a significant psychosocial effect especially if lesions are on visible areas of the body [4]. Accurate diagnosis, assessment of disease activity and extent, and individualized treatment are emphasized in modern clinical guidelines [5].

Use of standardized terminology has enabled classification of vitiligo into segmental, non-segmental and other clinical types [6]. The standardized instruments like the Vitiligo Extent Score can also help in measuring extent of disease, which allows reproducible measurement of body surface area affected [7]. However, the photographic documentation of skin diseases is becoming increasingly important for disease monitoring.

Dermatological image can be analyzed efficiently using AI because of the latest development in computer vision. Deep neural networks have performed well in a doctor's level (dermatologist level) for certain classification of skin-lesions tasks in controlled conditions [8]. The mass image collection resources have further enabled the implementation of algorithms and benchmarking. However, studies looking at demographic diversity and external testing have shown that high internal performance does not guarantee similar performance across populations or acquisition settings [9,10].

Artificial intelligence has been recently used for vitiligo. Guo et al. [11] created a hybrid AI approach able to detect vitiligo lesions and assess their severity. The authors Zhong et al. [12] later compared the ResNet and Swin Transformer architectures where these obtained a strong discrimination for vitiligo and non-vitiligo. A systematic review finds convolutional neural networks to dominate the literature on automated vitiligo recognition although considerable methodological heterogeneity persists among studies [13].

The ability to predict some features of a disease with high accuracy does not establish that the model has learnt their biologically meaningful connection. Unforeseen associations between labels and other variables such as the resolution, device, compression, or institutional source, and nonclinical factors may be exploited by the machine-learning system. The shortcut learning phenomenon was reported in publication [14]. Medical imaging has had similar issues with algorithms that utilize hospital- or acquisition-specific features rather than the pathology of interest [15-17]. In dermatology, skin markings may also serve as unexpected predictive signals [18].

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

As a result, creating a dataset is a major source of bias. Varoquaux and Cheplygina [19] stressed that methodological flaws in medical-imaging machine learning can inflate estimates of generalizability. Castro et al. [20] showed that biases in medical image datasets may propel forward by deep-learning systems. Abhishek et al. [21] identified duplication and other data quality issues in widely used dermatology image datasets. This emphasizes the need for systematic auditing of datasets before the development of models.

So the current analysis evaluated an image dataset of vitiligo-vs-healthy images which is available in public domain Instead of only inquiring whether the model can make the distinction between the two classes, we posed the negative-control question whether vitiligo labels can be predicted without looking at the dermatological content of the image. The dataset was characterized for non-clinical images in the study provided. Exact duplicates were identified by authors. Properties were compared between classes. Whether technical metadata could predict supplied disease labels was assessed.

Materials and Methods

Study design

The current study examined the public domain dermatological picture dataset used in excessive number of picture-matching studies. The analytical structure was conceived to evaluate possible technical confounding before developing a classical image-based diagnostic framework. The approach is in line with broader recommendations that studies of medical AI assess the integrity of datasets involved, possible sources of bias, validation of the model and reproducibility before making claims about clinical performance [22,23] (Figure 1).

Figure 1. Study workflow

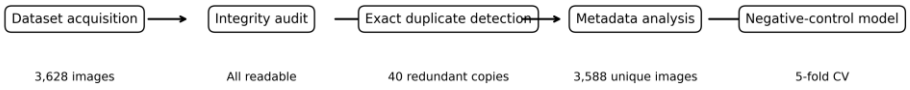


Figure 1. Study workflow from dataset acquisition through negative-control classification.

Dataset

The photos were taken from the Vitiligo dataset, a dataset hosted in Kaggle. The archive contained two image folders labeled healthy and vitiligo. No more pictures were received from other sources.

The dataset did not include demographic or clinical information at the participant level. In detail, there was no information available concerning age, sex, ethnicity, Fitzpatrick skin type, anatomical locality, vitiligo type, intra-anatomical time course, activity, Vitiligo Extent Score, history of treatment, laboratory measurements or identifiers of patients. Thus, the present study evaluates label-class technical specifications, using images rather than patient diagnostics.

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

By showing representative examples from the healthy and vitiligo classes, respectively, (Figure 2) demonstrates the visual nature and heterogeneity of the public dataset. These images are meant to serve illustrative descriptions. They were not individually interpreted clinically in this manuscript separately.



Figure 2. Representative healthy and vitiligo images from the public dataset

Image integrity and technical metadata

We programmatically opened every image file in the archive to check for corrupted or unreadable files. The data collected included width, height, file size, format, aspect ratio, count of pixels, megapixels, bytes per pixels, etc. There is no actual dermatological examination but it can show systematic differences in image acquisition, processing or source.

Duplicate detection

Exact duplication comprising MD5 cryptographic hashes calculated from the binary content of each file was investigated. Documents with the same MD5 hashes were deemed identical. The duplicated sets were analyzed to establish whether the duplicates occurred in the healthy class, in the vitiligo class, or between the two classes. Only one copy from each cluster of identical byte patterns was retained for further statistical modeling. Exact hashing can't find the same source image if it isn't exactly the same due to altering (resizing, cropping, re-compressing, rotating).

Statistical analysis

Because technical image characteristics were markedly non-normal, continuous variables were summarized using medians and interquartile range (IQR). The Mann-Whitney U test was used to analyze the healthy and vitiligo images. Effect size was measured using rank-biserial correlation. Positive and negative coefficients mean, respectively, larger values among healthy images and larger values among vitiligo images. Statistically significant was a two-sided p value of <0.05.

Metadata-only negative-control classification

A logistic regression model with negative control was created to see if class labels can be predicted without examining skin phenotype. The only six technical predictors the model used were width, height, file size, aspect ratio, megapixel

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

count and bytes per pixel. Most importantly, the model did not receive pixel intensities, color information, lesion borders, pigmentation characteristics, textures, or dermatological input.

Before fitting the model, continuous variables were standardized and balanced class weighting was used. The evaluation of the performance was done using stratified five-fold cross-validation with shuffling and a fixed random seed. Each image was evaluated by a model that had not been trained on that image, which utilized out-of-fold predictions. The metrics used include accuracy, balanced accuracy, sensitivity, specificity, F1 score, ROC AUC. The threshold for classification was 0.50 Using 2,000 bootstrap resamples, 95% CIs were calculated for the principal six-variable model. The evaluation included two additional negative-control secondary model, which were image width plus height only, and file size only.

Interpretation as a negative-control experiment

The metadata-only model was never intended to be used for diagnosis. It was used for a negative controlนั้น The size of the image or features related to its file storage don't have any biological ability for diagnosing vitiligo. Thus, strong class discrimination by these variables would indicate that the assigned class labels are correlated with the technical characteristics of the dataset. This method has a conceptually relation with studies showing shortcut learning and hidden confounding in medical image classification [14-17].

Reporting and methodological considerations

The design of the study was based on the already known and new recommendations for transparent reporting of diagnostic and AI studies including STARD principles, CLAIM, updated CLAIM guidance, and STARD-AI [24-27]. Frameworks designate explicit description for aspects of the data set selection, preprocessing, validation, reference standards, and bias.

Ethical considerations

The investigation utilized secondary computational analysis of open-access image files and did not involve human subject participation. The downloaded dataset did not have any patient identifiers. The fact that a database is publicly accessible does not solve questions of original consent, licensing, privacy, or image provenance. Check these issues against source dataset requirements before submission to the journal. The authors' institution and target journal's policies should determine the need for institutional ethical review.

Results

Dataset Composition and Image Integrity

We identified a total of 3,628 image files and all successfully opened them and decoded. The original project collected 2,100 vitiligo images (57.9%) and 1,528 healthy images (42.1%). The category comprising healthy photos only had JPEG images. The category vitiligo contained 1,984 JPEG and 116 PNG images (Table 1).

Characteristic	Healthy	Vitiligo	Total
Original images	1,528	2,100	3,628
Percentage of original dataset	42.1%	57.9%	100%

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

Readable images	1,528	2,100	3,628
Unreadable/corrupted images	0	0	0
JPEG images	1,528	1,984	3,512
PNG images	0	116	116
Redundant exact copies	0	40	40
Unique images after exact deduplication	1,528	2,060	3,588

Table 1. Composition and integrity of the vitiligo image dataset

Note: Exact duplication was determined from identical MD5 file hashes.

Exact duplicate images

There were found forty unnecessary exact copies. Everything happened in the class of vitiligo. The healthy and vitiligo classes did not share any exact duplicate file. Following the retention of a solitary copy of each byte-identical image, the analysis dataset comprised 3588 unique images of 2060 vitiligo and 1528 healthy. As a result, about 1.1% of the original archive was made up of redundant exact copies (Figure 3).

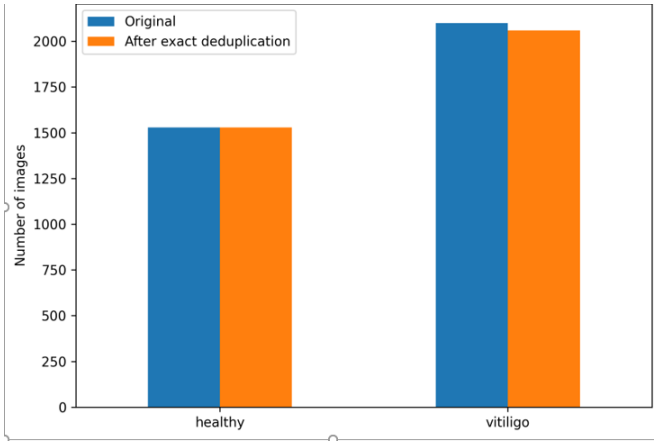


Figure 3. Class composition before and after exact deduplication

Technical differences between healthy and vitiligo images

After exact deduplication, the two classes exhibited marked technical differences. According to the findings, the image widths were around 416 pixels for vitiligo images and 1,600 pixels for healthy images, making healthy images about four times wider than vitiligo images. The average height among healthy images was 1,200 pixels while it was 416 pixels among vitiligo images. The median image areas were 1.920 MP and 0.173 MP, respectively. Images which were classified as healthy also had much larger median file sizes. The median file size was 66,374 bytes for healthy images. This is as compared to 17,323 bytes for vitiligo images (Table 2).

Characteristic	Healthy, median (IQR)	Vitiligo, median (IQR)	p value	Rank-biserial r
Width, pixels	1,600 (612-1,600)	416 (300-416)	< .001	0.754
Height, pixels	1,200 (408-1,200)	416 (226-416)	< .001	0.488
Megapixels	1.920 (0.250-1.920)	0.173 (0.067-0.173)	< .001	0.725
File size, bytes	66,374 (23,225-79,957)	17,323 (12,134-23,419)	< .001	0.622
Aspect ratio	1.333 (1.333-1.483)	1.000 (1.000-1.500)	< .001	0.334
Bytes per pixel	0.050 (0.040-0.110)	0.122 (0.092-0.303)	< .001	-0.609

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

Table 2. Technical image characteristics after exact deduplication

Note: IQR=interquartile range. Positive rank-biserial correlations indicate higher values among healthy images; negative values indicate higher values among vitiligo images.

The strongest class separation was associated with image width (r=0.754) and megapixel count (r=0.725). These findings demonstrate that the two image classes were generated, stored, selected, or processed under markedly different technical conditions (Figure 4).

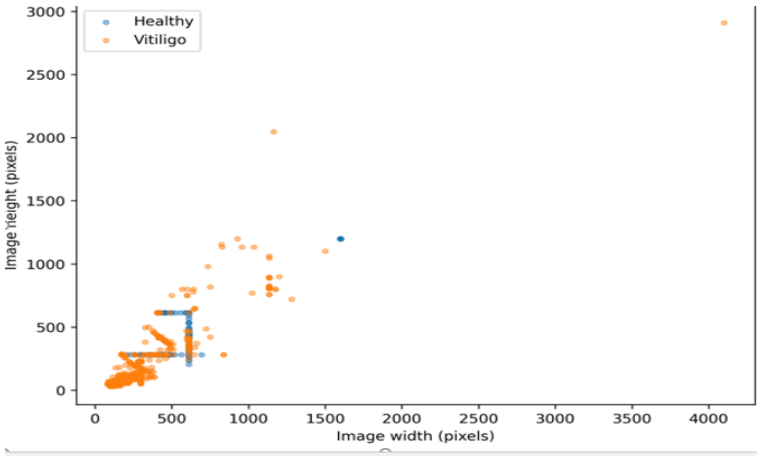


Figure 4: Image dimensions by class after exact deduplication.

Metadata-only classification

The main negative-control analysis revealed the supplied disease labels can be predicted with good discrimination, and without analysing the dermatological content of the photos. The model using six metadata variables gave accuracy=0.875, balanced accuracy=0.867, sensitivity=0.917, specificity=0.817, F1 score=0.894 and ROC AUC=0.942. The bootstrap 95% confidence interval for the AUC was 0.933-0.950. The confusion matrix out-of-fold comprised of 1,249 true negatives, 279 false positive, 171 false negatives and 1899 true positives (Table 3).

Model	Accuracy	Balanced accuracy	Sensitivity	Specificity	F1 score	ROC AUC
Six technical variables	0.875	0.867	0.917	0.817	0.894	0.942
Width + height only	0.863	0.856	0.905	0.807	0.884	0.929
File size only	0.759	0.735	0.900	0.571	0.811	0.800

Table 3. Performance of metadata-only negative-control classifiers

Note: For the principal six-variable model, bootstrap 95% CIs were: accuracy, 0.863-0.885; balanced accuracy, 0.856-0.879; sensitivity, 0.905-0.929; specificity, 0.798-0.837; F1 score, 0.884-0.903; and ROC AUC, 0.933-0.950. Six-variable predictors were width, height, file size, aspect ratio, megapixel count, and bytes per pixel.

Notably, limiting the model to width and height only resulted in an AUC of 0.929. Just the file size alone gave good discriminatory information with a value of 0.800 (Figure 5).

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

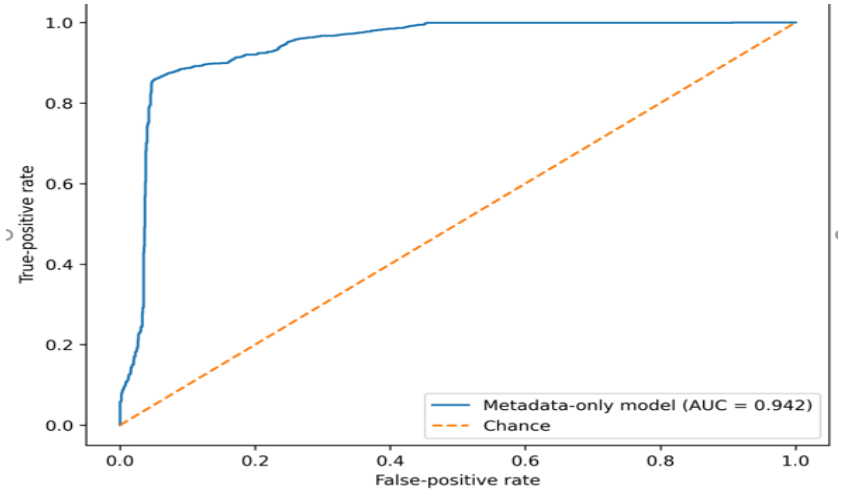


Figure 5. Receiver operating characteristic curve for the six-variable metadata-only logistic regression model.

File-size distribution

There were significant file size differences between classes. The distribution of healthy images moved towards bigger files while vitiligo images were primarily smaller files. This variable may serve as a proxy for systematic differences between image sources because file size can reflect image dimensions, compression, image complexity, file format, and acquisition or preprocessing practices (Figure 6).

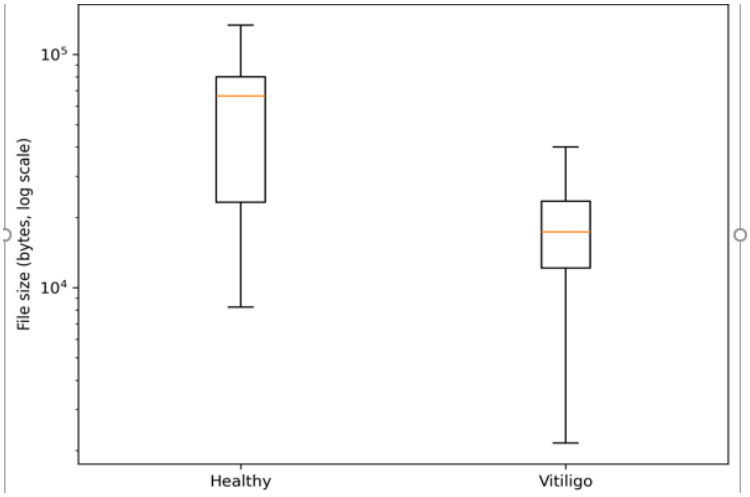


Figure 6. File-size distribution according to image class after exact deduplication.

Discussion

Principal findings

The main finding of the study is that publicly available vitiligo image dataset has sufficiently strong technical confounding for its disease labels to be predicted with AUC of .942, without looking at the skin itself. A classifier that used only the width, height, file size, aspect ratio, number of megapixels and bytes/pixel showed discrimination equivalent to

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

performance sometimes reported for image-based disease classifiers more complex. Only the width and height of the image gave an AUC of 0.929.

According to these variables, melanocyte loss cannot be identified biologically nor vitiligo diagnosed. The performance data suggests the input class labels exhibit a strong correlation with technical characteristics of the images. The image sizes were more different than anything. The vitiligo image was 416 × 416 pixels and healthy image was 1600 × 1200 pixels which were median values. There were very substantial differences in file size and megapixel count. This study shows evidence of class associated technical confounding from this dataset.

Shortcut learning as an explanation

When a model finds predictive associations that work well in the data but do not correspond to the biological or clinical phenomenon of interest, we have a shortcut learner [14]. This issue is very much pertinent to medical imaging. The pneumonia classifiers trained on chest radiographs do not perform the same across the institutions, as Zech et al [15] have shown. One reason for this is the fact that the models could learn to use certain characteristics that were specific to the institution. According to Badgeley et al [16], the use of confounding variables can not only induce success in a model but can also prevent doctors from properly analyzing their results. Research by DeGrave et al. [17] showed that COVID-19 imaging models rely on irrelevant shortcuts.

Our results show that the same phenomenon is a prerequisite within the current vitiligo dataset; the class label itself is strongly connected with technical image properties. A future deep-image learning system trained on these images may therefore also have access to resolution-dependent interpolation patterns, JPEG or PNG encoding characteristics, image sharpness, compression artifacts, back-ground composition, photographic framing, source-dependent coloring characteristics, cropping practices, lighting, anatomical framing, and potential source-specific visual artifacts. Subsequent research involving dermatology shows that “innocuous”, which includes surgical skin markings, features can change neural-network behavior [18].

Changing the size of each image to the same neural-network input dimension does not eliminate these differences. Information related to the source can remain encoded in texture, compression, background, sharpness and other pixel-level characteristics after resizing.

Why high internal accuracy could be misleading

Accuracy/sensitivity/specificity/AUC is often the deep-learning study end-point. Performance validation is meaningful only when the design tests for intended clinical signal rather than unintended dataset characteristics. A number of methodological reviews show how inappropriate splitting, heterogeneity of data and confounding as well as limited external validation result into optimistic estimates of medical AI performance [19,20].

As a result, a model may perform well on a subset of the present dataset chosen at random, yet still fail on images acquired using another smartphone, camera, clinic, country, compression method, or image processing pipeline. This difference is important for the clinical translation. As Kelly et al. [22] suggested, for an algorithm to be classified as effective healthcare AI, it needs to prove generalizability among other things. According to Wiens et al [23], it is not enough for companies to only assess predictive performance.

Citation: Alkhatib A.J (2025). “Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study” *Journal of Integrated Human Health Sciences,* 1(2):1-15.

Implications for vitiligo AI research

AI holds great potential for vitiligo detection and disease quantification. An artificial intelligence system for the detection and evaluation of vitiligo lesions was proposed by Guo et al. [11]. Moreover, this was evaluated employing recent ResNet and Swin Transformer architectures by Zhong et al. [12]. According to the systematic review conducted by Tanvir et al. [13], research activity is increasing in machine-learning approaches to vitiligo.

It is not the case that AI cannot recognize vitiligo morphology. Quite the contrary. They show that performance achieved from a certain dataset cannot be ascribed to disease-specific learning automatically. It is particularly relevant in cases where a publicly built dataset pools together heterogeneous online or photographic images. The question is therefore not just how well a model can classify images, but what information it uses to do that.

Duplicate Images and Data Leakage

The original archive contained forty identical redundant copies from the vitiligo category alone. Even though this only comprises around 1.1% of the whole dataset, duplication can become significant when images are randomly split up for training and testing. When the same images are present in both sets, it can enhance the performance of the model because the testing set is not entirely unseen.

According to Abhishek et al. [21], the current resource is not the only one in which duplication and other data-quality problems occur. Recently, audits of dermatological image datasets have shown that this resource also shows such problems. As a result, it is necessary to audit public datasets before model training, rather than taking every downloaded file as an independent observation. MD5 hashing primarily detects exact binary copies only. If the original photograph has been resized, cropped, rotated, recompressed, converted between formats, or edited slightly, it will not be detected as such. Therefore, the current duplication count should be considered as the minimum redundancy of images. A following study need to do near-duplicates of images families by perceptual hashing and image-embedding similarity.

Importance of patient-level independence

The absence of patient identifiers is a more serious limitation. The same person can be a subject of more than one photograph, different lesions of the same person or repeats of the same lesion. These potentials cannot be determined without identifiers of the participant. This has big changes for machine-learning validating. When a patient has multiple images and this image pool is divided randomly between training and test sets, the algorithm can infrequently learn patient-specific characteristics. It may escape learning disease characteristics that are generalizable in nature. The statistically correct unit of partitioning in clinical image research is generally the patient and not the separate photograph wherever multiple images per participant are possible. The current dataset does not allow for patient-level partitioning and this should be communicated clearly in any future study via these images.

Dermatological dataset diversity

Skin-image AI also varies with demographics and phenotypes. The Fitzpatrick17k resource was developed by Groh et al [9] partly to analyze algorithmic performance across skin types. Subsequent research by Daneshjou et al [10] demonstrated that different dermatology AI systems performed variably when tested on a heterogeneous clinical dataset.

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

The visible symptoms of vitiligo make it easy to diagnose in any skin type. Nevertheless, lesion contrast with background pigmentation, light and camera exposure can all influence diagnosis using photographic evidence. Due to the lack of Fitzpatrick skin type, race, ethnicity, or geographic origin in the current dataset, we cannot evaluate subgroup fairness and generalizability. It is particularly relevant while progressing from the classification of datasets to claims of clinical deployment.

Explainability is necessary but not sufficient

Techniques that point out areas in an image contributing to a model's prediction like Grad-CAM [28] fall in this category. An effective vitiligo classifier would ideally focus on the morphology and borders of the depigmented lesions and not on the clothes, borders, backgrounds, text, etc.

Nevertheless, one can't consider that explanation methods do provide conclusive proof for the biological validity of a model. As demonstrated by Adebayo et al [29], several saliency based methods appears visually plausible, with model parameters and/or labels randomly permuted. Explainability should supplement dataset auditing, controlled experiments, bias mitigation, external validation, and clinically independent testing, not supplant them.

Recommended framework for future modeling

The results here suggest that future machine-learning experiments based on this dataset should include a staged bias-control framework.

- Prior to partitioning the data, exact and near-duplicate removal, followed by perceptual-hash-based detection of transformed copies or embedding-based.
- The dimensions of images and their file encoding, cropping, and preprocessing should be technically standardized as far as possible.
- Images originating from the same source must remain in the same partition if the source information can be reconstructed.
- Using a cleaned dataset, train a CNN or vision-transformer model controlling for bias.
- The Grad-CAM or similar explainability investigation will analyze whether predictions depend on lesion regions or irrelevant structures. A similar approach was used [28]
- Various variability analysis that investigate the performance prior to and after duplicate removal and technical standardization; within strata of image resolution; and removal of images containing obvious source specific cue.
- Verification by an independent external agency using dataset created under a different acquisition process before clinical diagnostic claims are made.

This approach is in line with the overall aim of AI reporting standards to differentiate reproducible model development from clinically meaningful diagnostic validation [25-27].

Interpretation of the metadata-only AUC

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

It is not correct to consider the AUC of 0.942 as alternative diagnosis method for vitiligo. Rather, this quantifies how technical characteristics are found associated with class membership on this data. As a result, the experiment acts as a stress test for a dataset.

A homogeneous clinical dataset technically classified based on metadata alone will yield results close to chance unless there are bona fide differences in acquisition related to the workflow. In the current dataset, class construction and technical image properties are connected as deviation from chance is large. Because it is not a preliminary analysis, we consider the metadata-only classifier to be one of the most informative results.

Strengths

There are various strengths of this study. Before building a standard deep learning model, the dataset is investigated; every downloaded image is tested as readable; exact duplicates are sorted out systematically; multiple technical image properties are quantified rather than visually inspected; and the metadata-only negative-control experiment demonstrates directly that the class labels contains huge non-clinical predictive information. For the evaluation of the models, stratified five-fold cross-validation with out-of-fold predictions (different folds for train and test) rather than a single favorable train-test split was used and bootstrap confidence intervals for the main performance metrics were calculated. In conclusion, the analytical strategy is clear, straightforward and can readily reproduce infferspectral library other public medical image datasets.

Limitations

One should consider some limitations We cannot establish independence of individual photographs since no patient-level identifiers were available. Dermatologists did not independently confirm the validity of the original vitiligo and healthy labels. Exact MD5 hashing only identifies such files that are identical by bytes and can't help in case an image is modified or near identical. We do not have any clinical variables in the dataset such as disease subtype, duration, activity, severity, anatomical location, treatment, and diagnostic confirmation. We did not have data on any demographic variables (age, sex, ethnicity, Fitzpatrick skin type). The metadata model shows technical confounding but does not directly suggest which features a future CNN or transformer-based classifier uses. They did not include an independent external dataset and the provenance and acquisition conditions of individual images could not be reconstructed from the downloaded files. The current findings cannot be interpreted as estimates of clinical diagnostic performance because of these limitations.

Implications for Future Research

The next logical step is a bias-controlled image-modeling experiment, rather than a simple CNN classification exercise. An especially informative design would be to compare a conventional model trained on the original dataset; the same architecture after exact and near-duplicates were removed; the model after technical image standardization; performance after restricting images to overlapping ranges of resolutions; and performance on an independent external test set. If you paint leaves to look accurate but they don't look as accurate after bias control, that would somewhat be proof. On the flip

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

side, should authoritative performance persist in the context of stringent controls and external validation, confidence in the model's gain of true vitiligo morphology would sharpen. This experiment would provide more information than comparing various neural-network architectures on a random train-test split.

Conclusion

In a publicly available vitiligo image dataset, substantial class-associated technical heterogeneity was identified. We detected 40 redundant exact copies among 3628 images and left 3588 files. There was a significant difference between healthy and vitiligo images in resolution, megapixels, file size, aspect ratio, and bytes/pixel. The most significant finding is that a logistic regression model that only incorporates the non-clinical technical metadata is able to discriminate vitiligo from healthy labels with an AUC of 0.942 (95% CI: 0.933-0.950).

This finding illustrates that the class labels are strongly tied to technical features unrelated to vitiligo pathophysiology. This means that an algorithm that is performing well internally (which by all means will happen) with respect to the image-based models produces with this dataset should not be interpreted as the algorithm having learned the disease-specific morphology. Before the model development, auditing of datasets should hence be done. Before any clinical conclusions can be drawn, it is necessary to perform deduplication, assessment of independence of patients, technical standardization, shortcut analysis, explainability, and independent external validation.

References

1. Ezzedine, K., Eleftheriadou, V., Whitton, M., & van Geel, N. (2015). *Vitiligo Lancet 386 (9988): 74–84*.
2. Bergqvist, C., & Ezzedine, K. (2020). Vitiligo: a review. *Dermatology*, *236*(6), 571-592.
3. Seneschal, J., Bae, J. M., Ezzedine, K., Hamzavi, I., Harris, J. E., Bellei, B., ... & Picardo, M. (2025). The Vitiligo. *Nature Reviews Disease Primers*, *11*(1).
4. Bibeau, K., Ezzedine, K., Harris, J. E., van Geel, N., Grimes, P., Parsad, D., ... & Hamzavi, I. H. (2023). Mental health and psychosocial quality-of-life burden among patients with vitiligo: findings from the global VALIANT study. *JAMA dermatology*, *159*(10), 1124-1128.
5. Eleftheriadou, V., Atkar, R., Batchelor, J., McDonald, B., Novakovic, L., Patel, J. V., ... & British Association of Dermatologists' Clinical Standards Unit. (2022). British Association of Dermatologists guidelines for the management of people with vitiligo 2021. *British Journal of Dermatology*, *186*(1), 18-29.
6. Ezzedine, K., Lim, H. W., Suzuki, T. A., Katayama, I., Hamzavi, I., Lan, C. C. E., ... & Vitiligo Global Issue Consensus Conference Panelists. (2012). Revised classification/nomenclature of vitiligo and related issues: the Vitiligo Global Issues Consensus Conference. *Pigment cell & melanoma research*, *25*(3), E1-E13.
7. van Geel, N., Lommerts, J., Bekkenk, M., Wolkerstorfer, A., Prinsen, C. A., Eleftheriadou, V., ... & Speeckaert, R. (2016). Development and validation of the Vitiligo Extent Score (VES): an international collaborative initiative. *Journal of Investigative Dermatology*, *136*(5), 978-984.
8. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, *542*(7639), 115-118.
9. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., ... & Badri, O. (2021, June). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1820-1828). IEEE.

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

10. Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., ... & Chiou, A. S. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science advances*, *8*(31), eabq6147.
11. Guo, L., Yang, Y., Ding, H., Zheng, H., Yang, H., Xie, J., ... & Ge, Y. (2022). A deep learning-based hybrid artificial intelligence model for the detection and severity assessment of vitiligo lesions. *Annals of translational medicine*, *10*(10), 590.
12. Zhong, F., He, K., Ji, M., Chen, J., Gao, T., Li, S., ... & Li, C. (2024). Optimizing vitiligo diagnosis with ResNet and Swin transformer deep learning models: a study on performance and interpretability. *Scientific Reports*, *14*(1), 9127.
13. Tanvir, S., Syed, S. A., Hussain, S., Zia, R., Rashid, M., & Zahid, H. (2024). Detection of Vitiligo Through Machine Learning and Computer-Aided Techniques: A Systematic Review. *BioMed Research International*, *2024*(1), 3277546.
14. Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665-673.
15. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*.
16. Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., ... & Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ digital medicine*, *2*(1), 31.
17. DeGrave, A. J., Janizek, J. D., & Lee, S. I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, *3*(7), 610-619.
18. Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., ... & Haenssle, H. A. (2019). Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, *155*(10), 1135-1141.
19. Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, *5*(1), 48.
20. Castro, D. C., Walker, I., & Glocker, B. (2020). Causality matters in medical imaging. *Nature Communications*, *11*(1), 3673.
21. Abhishek, K., Jain, A., & Hamarneh, G. (2025). Investigating the quality of dermamnist and fitzpatrick17k dermatological image datasets. *Scientific Data*, *12*(1), 196.
22. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, *17*(1), 195.
23. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, *25*(9), 1337-1340.
24. Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., ... & STARD Group. (2015). STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*, *277*(3), 826-832.
25. Mongan, J., Moy, L., & Kahn Jr, C. E. (2020). Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiology: Artificial Intelligence*, *2*(2), e200029.
26. Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., ... & CLAIM 2024 Update Panel. (2024). Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiology: Artificial Intelligence*, *6*(4), e240300.
27. Sounderajah, V., Guni, A., Liu, X., Collins, G. S., Karthikesalingam, A., Markar, S. R., ... & Ashrafian, H. (2025). The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature medicine*, *31*(10), 3283-3289.
28. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.

29. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, *31*.

Citation: Alkhatib A.J (2025). "Hidden Technical Confounding in a Public Vitiligo Image Dataset: A Data-Quality Audit and Metadata-Only Classification Study" *Journal of Integrated Human Health Sciences,* 1(2):1-15.