Skin Lesion Classification on the MILK10K Dataset

Kinley Gyeltshen
U3264610
Bacholor of Information Technology
Univesity of Canberra
Canberra, Australia
u3264610@uni.canberra.edu.au

***Abstract*— Skin cancer diagnosis relies heavily on both dermoscopic imaging and patient-specific clinical information. However, the inherent class imbalance and heterogeneity in lesion appearance often limit the performance of traditional image-based models. This paper presents, a custom multimodal deep learning architecture designed to jointly learn from clinical images, dermoscopic images, and structured metadata to classify skin lesions across 11 diagnostic categories from the MILK10K dataset. The proposed framework employs a dual-branch convolutional neural network (CNN) for image processing and parallel multi-layer perceptrons (MLPs) for metadata encoding, followed by feature fusion at both modality and global levels. To address severe class imbalance, two strategies class-weighted loss and weighted random sampling were evaluated. Experimental results show that the weighted random sampler configuration achieved the best performance, with a test accuracy of 59.7% and a weighted F1-score of 0.4930, outperforming both baseline and class-weighted models. These findings demonstrate that probabilistic oversampling yields better balance between sensitivity and precision across lesion types. The study underscores the effectiveness of multimodal fusion in improving diagnostic accuracy for imbalanced dermatological datasets and provides a scalable foundation for future clinical AI systems.**

Keywords—Convolutional Neural Network(CNN), Multi-layer Perceptron Layer(MLP,)), dermoscopic, clinical, MultiModalLesionNet

I. Introduction

Skin cancer, particularly melanoma, poses a significant global health concern due to its high mortality rate when diagnosed at advanced stages. Early and accurate classification of skin lesions is critical for improving patient outcomes. Recent advances in deep learning have enabled automated diagnostic systems capable of integrating both visual and clinical information. The MILK10 dataset, a multimodal dataset comprising clinical images, dermoscopic images, and associated patient metadata, provides a robust foundation for training such systems.

This study proposes a multimodal deep learning architecture that fuses image features with structured clinical metadata to classify skin lesions into multiple diagnostic categories. Unlike traditional CNN-based image classifiers, my approach introduces metadata boost mechanisms and dual branch feature fusion to enhance diagnostic robustness in class-imbalanced settings.
The objective of this work is to evaluate the performance of a custom multimodal architecture on the MILK10K dataset using advanced training techniques such as class weighting, balanced sampling, and learning rate scheduling.

II. Related Work

Several studies have leveraged CNNs for medical image classification. For instance, [2] demonstrated that CNNs can effectively extract hierarchical features from dermoscopic images for melanoma detection. Data augmentation strategies have been employed to improve generalization, particularly in imbalanced datasets [3]. In particular, the work presented in [4] utilized the MILK10 dataset and demonstrated that combining dermoscopic and clinical modalities improved diagnostic precision. Hybrid models incorporating metadata, such as patient age and lesion location, have shown superior performance over image-only systems.

During my initial attempt on this dataset, [8] I have explored classical machine learning models, including Support Vector Classifier (SVC), Logistic Regression, and Naïve Bayes, evaluated via five-fold cross-validation. Logistic Regression achieved the best macro-F1 score (≈0.32), outperforming SVC (≈0.25) and Naïve Bayes (≈0.05). This prior investigation highlighted the challenges of class imbalance, feature correlation, and the limitations of purely linear models in handling multimodal (image + metadata) data. The findings serve as a foundational baseline for the present work, which extends this approach to deep neural architectures for improved feature representation and classification accuracy.

This work builds upon these efforts by designing a fully bespoke CNN+MLP hybrid architecture tailored to MILK10K's multimodal structure.

III. Dataset and Preprocessing

[5] The MILK10 dataset consists of 5240 lesions across 11 diagnostic classes which includes the following:

Diagnostic Category	Abbreviation
Actinic keratosis/intraepidermal carcinoma	AKIEC
Basal cell carcinoma	BCC
Other benign proliferations including collisions	BEN_OTH
Benign keratinocytic lesion	BKL
Dermatofibroma	DF
Inflammatory and infectious	INF
Other malignant proliferations including collisions	MAL_OTH
Melanoma	MEL

Melanocytic Nevus, any type	NV
Squamous cell carcinoma/keratoacanthoma	SCCKA
Vascular lesions and hemorrhage	VASC

. Each lesion is accompanied by:

- **Clinical Image:** Macroscopic view of the skin lesion.
- **Dermoscopic Image:** High-magnification view obtained via dermatoscope.

Metadata: Patient information (age, sex, lesion site), and image-based MONET annotations for the two images. The meta includes the following features:

- age_approx
- sex
- skin_tone_class
- site
- MONET_ulceration_crust
- MONET_hair
- MONET_vasculature_vessels
- MONET_erythema
- MONET_pigmented
- MONET_gel_water_drop_fluid_dermoscopy_liquid
- MONET_skin_markings_pen_ink_purple_pen

A. Data Cleaning and Filtering

The data was cleaned by removing incomplete or corrupted lesion samples. The following criteria were enforced:

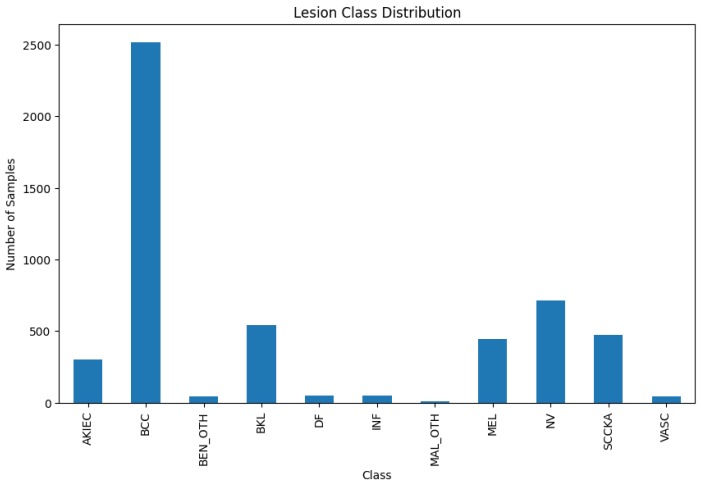
- Lesions missing either clinical or dermoscopic image were removed.
- Metadata entries containing `NaN` or invalid values were excluded.
- Only lesions with exactly two valid image modalities were retained.

A summary of dataset cleaning:
Total lesions before filtering: 5240
Lesions removed due to missing data: 51
Final usable lesions: `5189`

B. Data Label

Original labels were provided in one-hot format. Labels were converted to integer class indices (`0-10`).The indices of the class were saved for converting the raw index into labeled class. During further exploration on the class distribution of the data set, I found that the dataset was very unbalanced with class count as the following:

AKIEC 303
BCC 2518
BEN_OTH 43
BKL 541
DF 52
INF 49
MAL_OTH 9
MEL 444
NV 712
SCCKA 473
VASC 45



C. Data Augmentation

Due to imbalance of the class distribution, it was imminent that data augmentation might be required. However, due to the nature and structure of the dataset, I have decided not to do so. Since [5] each lesion image's meta data was annotated using the MONET framework, with probabilities for the following concept term groups included in the metadata:

- Ulceration, crust
- Hair
- Vasculature, vessels
- Erythema
- Pigmentation
- Gel, water drop, fluid, dermoscopy liquid
- Skin markings, pen ink, purple pen

Augmenting the images in any way such as rotation, flipping, scaling, and colour jitter can make the meta data obsolete.

D. Metadata Processing

[5] Meta data includes some categorical features like, sex and site. I used one hot encoding to convert them into one hot encoded features, resulting in new features like:

- sex_female
- sex_male
- site_foot
- site_genital
- site_hand
- site_head_neck_face
- site_lower_extremity
- site_trunk
- site_upper_extremity

Numerical metadata fields like Age Approximation, Skin tone class were normalized using z-score scaling via `StandardScaler`. Due to the nature of the other meta features they were not scaled

E. Image Preprocessing

All images were resized to `256 x 256`, were kept in RGB format, and then normalized using `ToTensor()` to convert the images pixel value from 0 to 1.

IV. Model Architecture

The proposed MultiModalLesionNet is a custom deep learning architecture designed to jointly learn from *dermoscopic* and *clinical* skin lesion images, along with corresponding patient metadata. This multi-branch framework enables the model to integrate two different information sources, improving diagnostic robustness and interpretability.

A. Dual-Stream Image Processing

The network incorporates two parallel convolutional branches, one for *clinical* images and one for *dermoscopic* images, each consisting of two sequential convolutional blocks. Each block comprises a pair of convolutional layers followed by ReLU activations and a 2×2 max pooling operation. After the second block, the feature maps are flattened into 1D representations to facilitate multimodal fusion.

B. Metadata Encoding

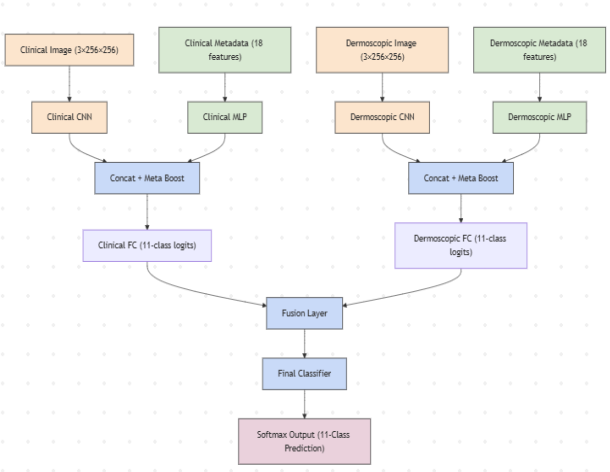
In parallel, metadata are processed through separate multi-layer perceptron (MLPs) for each modality. Each metadata MLP contains a configurable number of hidden, with ReLU activations after each linear transformation. To control the influence of the metadata features, a learnable scalar parameter is applied, allowing the model to dynamically adjust the contribution of metadata relative to the image-derived features during training.

C. Modality-Level Fusion

For each modality, the flattened CNN feature vector is concatenated with its corresponding boosted metadata embedding. This combined representation passes through a modality-specific fully connected (FC) fusion network that projects the multimodal embedding into an 11-dimensional logit vector, corresponding to the 11 lesion classes in the MILK-10 dataset.

D. Final Fusion and Classification

The outputs from the two modality-specific fusion networks are concatenated to form a joint multimodal representation. This representation is passed through a final classifier consisting of a sequence of linear layers and ReLU activations, culminating in a softmax output layer that produces the final class probabilities. This late-fusion strategy allows the model to integrate global contextual cues from both clinical and dermoscopic modalities for final decision-making.



V. Loss Function, Optimizer and Imbalance Handling

The Cross-Entropy Loss function was employed as the primary training objective. This loss is the standard choice for single-label multi-class classification problems, where each input belongs exclusively to one of *K* possible categories [6].

For optimization, the Adam (Adaptive Moment Estimation) optimizer [7] was utilized. Adam adaptively adjusts learning rates for each parameter based on estimates of both the first and second moments of the gradients, effectively combining the benefits of AdaGrad (robustness to sparse gradients) and RMSProp (stability under non-stationary objectives). This makes Adam particularly well-suited for multimodal networks with heterogeneous feature distributions, such as the proposed model.

To mitigate the severe class imbalance inherent in the MILK-10 dataset, two model will be trained using each of the methods below:

- a class-weighted loss function was implemented. Specifically, the Cross Entropy Loss function was parameterized with class weights inversely proportional to the frequency of each label in the training set, as follows:

  loss_fn
  = nn.CrossEntropyLoss(weight=class_weights)

  This weighting scheme ensures that minority classes contribute a proportionally higher penalty to the loss, thereby compelling the model to pay greater attention to underrepresented lesion categories.

- a Weighted Random Sampler was employed within the data loader to further counteract the imbalance during batch formation. This sampler assigns each sample a probability proportional to the inverse of its class frequency and performs random sampling with replacement. Consequently, minority class samples are more frequently presented to the network throughout training, improving representation diversity and stabilizing convergence under highly skewed data distributions.

VI. HYPERPARAMETERS

Several hyperparameters were tuned to optimize the performance and stability of the proposed MultiModalLesionNet. The final configuration was determined through iterative experimentation, balancing model convergence, computational efficiency, and generalization to unseen data. The optimized parameters are summarized in below:

Parameter	Value
Optimizer	Adam
Loss Function	CrossEntropyLoss
Learning Rate	0.001
Batch Size	32
Epochs	50
Hidden Units	10
Hidden layer DNN	10

Hidden Units is the number convolutional layer inside the each convolutional block
Hidden layer DNN is number of hidden layer inside the MLP blocks

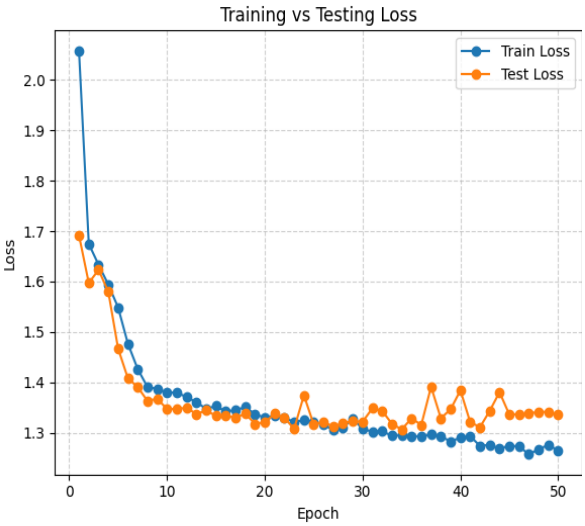
VII. EXPERIMENTAL RESULTS

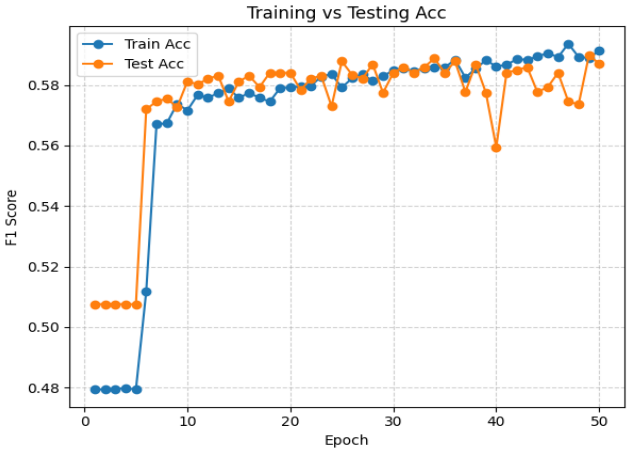
A. *Model 0 – Base Model*

Metric	Train Value	Test Value
Accuracy	0.59	0.58
Macro F1	0.1271	0.125
Weighted F1	0.4539	0.4529

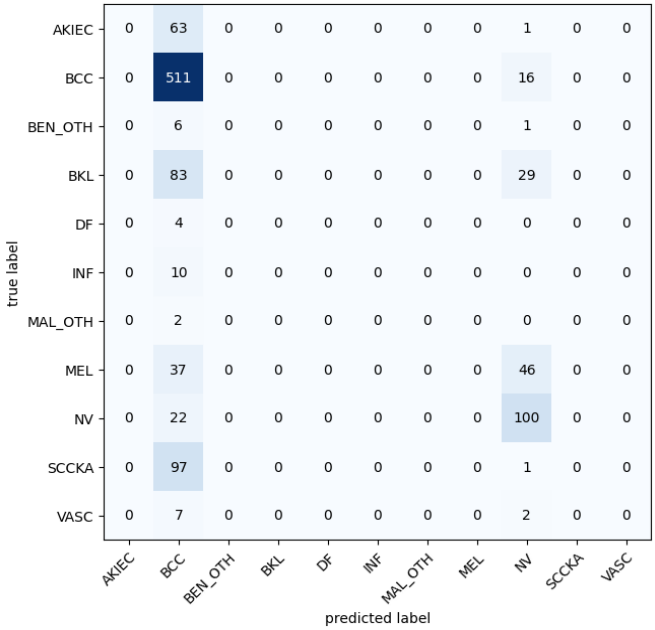
For more detail classification report refer [9]

1) *Model learning trend*





2) *Confusion Matrix for its prediction on the test data*



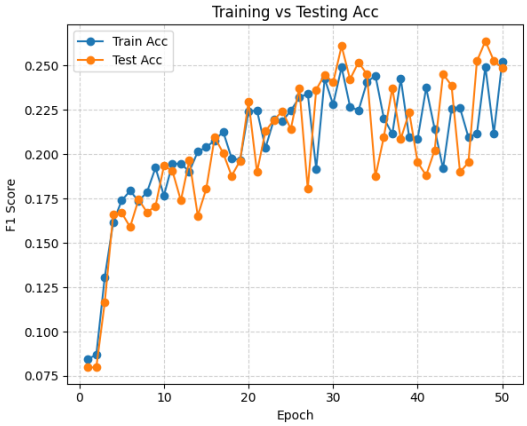
Although the overall accuracy reached approximately 58 %, the macro-F1 score remained low (≈ 0.125), indicating that performance was dominated by the majority class. This aligns with the previously observed imbalance in the dataset, where one class represented nearly 50 % of all samples. Nevertheless, compared to the earlier machine-learning baselines [8] (Logistic Regression ≈ 0.32 macro-F1, SVC ≈ 0.25 macro-F1), the current deep multimodal approach demonstrates improved feature representation and higher overall accuracy on unseen data.

B. *Model 1 - Model with class-weighted loss function*

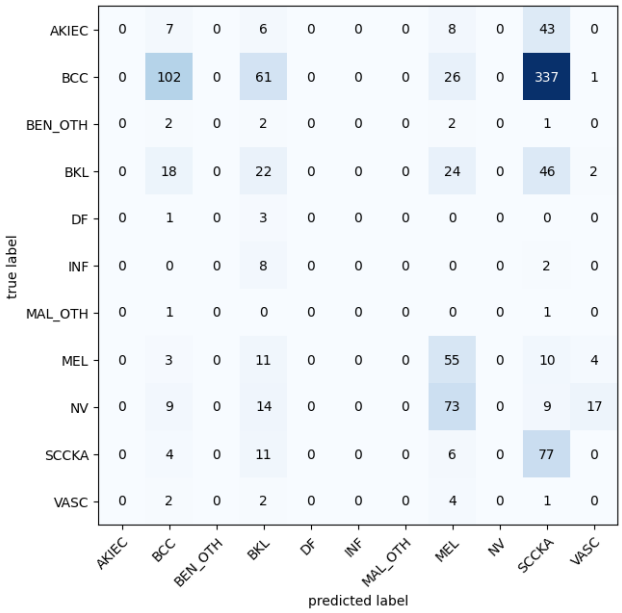
Metric	Train Value	Test Value
Accuracy	0.2524	0.2485
Macro F1	0.1325	0.1013
Weighted F1	0.2469	0.2266

For more detail classification report refer [9]

1) *Model learning trend*



2) Confusion Matrix for its prediction on the test data



Introducing class weighting in **Model 1** led to a reduction in overall accuracy but a slight increase in **macro-F1**, indicating improved sensitivity to minority classes at the cost of misclassifying dominant classes. Even from the confusion matrix we could see that the model classifying some other class rather.

It seems that the model penalizes more for misclassifying rare classes, causing the model to deviate from the majority distribution and reduce total accuracy. Since, misclassifying rare classes punishes the model severely than misclassifying dominant class, the model seems get the predominant class right than dominant class leaving the model to not care about the predominant class. The trade-off suggests that although **Model 1** better accounts for underrepresented lesions, further balancing methods and fine tuning are required to stabilize learning.

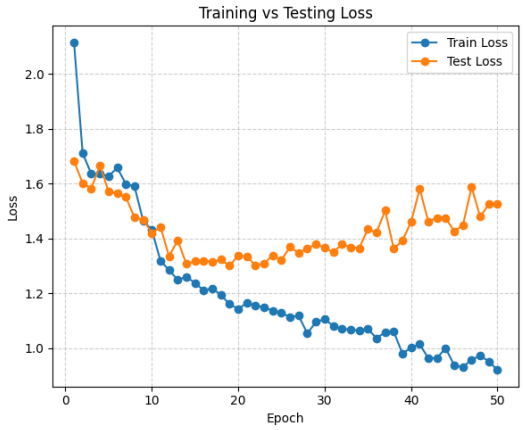
C. Model 2 - Model with Weighted Random Sampler

The third model configuration introduced a **Weighted Random Sampler** into the data loading process to address class imbalance more dynamically. Unlike the static class-weighted loss used in Model 1, this approach adjusts the **probability of sampling** each training example such that underrepresented lesion types appear more frequently during training. The sampler assigns higher selection probability to minority class samples, ensuring that the model is exposed to all lesion types with roughly equal frequency per epoch.

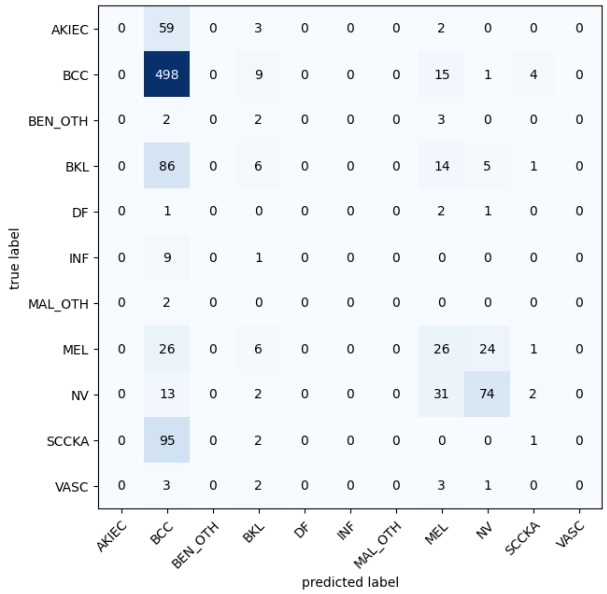
Metric	Train Value	Test Value
Accuracy	0.6740	0.5969
Macro F1	0.2339	0.1632
Weighted F1	0.5996	0.4930

For more detail classification report refer [9]

1) Model learning trend



2) Confusion Matrix for its prediction on the test data



Model 3 demonstrated a clear improvement in both training and test performance compared to the previous configurations. The incorporation of the Weighted Random Sampler allowed for more balanced exposure to underrepresented lesion classes during training, resulting in a test accuracy of 59.7% and a weighted F1 score of 0.4930 the highest among all models tested. This indicates that the model achieved a better balance between sensitivity and precision across classes, reducing the dominance of the majority *BCC* class. While the macro F1 score (0.1632) suggests that performance on the rarest categories remains limited, the overall results confirm that probabilistic oversampling is a more effective strategy for mitigating class imbalance than static class-weighted loss.

VIII. Best Model

Among the three experimental configurations, Model 2 (Weighted Random Sampler) emerged as the best performing model. It achieved the highest test accuracy (59.7%) and weighted F1-score (0.4930), outperforming both the baseline model and the class weighted loss variant. The weighted F1-score, which accounts for label imbalance, provides a more reliable indication of real-world performance than accuracy alone. Model 2's improvements highlight the effectiveness of the Weighted Random Sampler in ensuring that minority lesion classes were adequately represented during training. While the macro F1-score (0.1632) indicates that prediction performance for rare classes still lags behind, the combination of higher accuracy and stronger weighted F1 performance suggests that the model achieved a better trade-off between sensitivity and precision across all 11 lesion categories. Thus, Model 2 was identified as the most balanced and generalizable configuration, providing the most robust classification results for this highly imbalanced dermatological dataset

A. Saving and Loading Model

After training the final version of the MultiModalLesionNet, the model was saved using **pickle.** This stores the model's learned parameters (weights and biases) in a .pkl file, ensuring the model can be reloaded later without retraining from scratch using the same pickle library. Refer [9] for full code implementation.

B. Fine Tuning with new data.

The saved model can be further fine-tuned as new labeled data becomes available. a crucial step for maintaining model performance across evolving clinical imaging standards or demographics. Fine-tuning typically involves:

1. Loading the pre-trained weights from the saved checkpoint.
2. Freezing earlier convolutional layers (to preserve learned low-level features).
3. Unfreezing and retraining higher-level or fully connected layers using a smaller learning rate
4. Updating class weights to reflect the new dataset's label distribution.

Over time, with continuous retraining and validation on updated datasets, the model's robustness and diagnostic accuracy can be progressively enhanced, making it more generalizable and clinically reliable

C. Ethical and Privacy Considerations

Developing artificial intelligence models for medical image analysis introduces a number of ethical, privacy, and fairness challenges that must be addressed to ensure the responsible use of AI in healthcare contexts. The dataset used in this study, MIlk10, contains clinical and dermoscopic skin lesion images, as well as associated patient metadata (e.g., age, sex, and body site). Although the dataset is anonymized and publicly released for research purposes [5], several ethical issues still arise concerning data handling, model bias, and interpretability.

All patient-identifiable information was removed before data preprocessing and model training. Only non-identifiable metadata fields (e.g., approximate age, lesion location, and binary gender) were used. The dataset complies with ethical data-sharing standards and institutional review protocols[5]. Additionally, during local experimentation, no patient data was uploaded to cloud storage; all computations were performed in a secure, offline environment to minimize potential data leakage.

Acknowledgement

Portions of the abstract and text editing were assisted using OpenAI's ChatGPT (GPT-5) to improve grammar, clarity, and conciseness. All technical content, data analysis, and interpretations were independently conducted and verified by the author.

REFERENCES

[1] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, pp. 115–118, 2017.

[2] Y. Yuan, “Automatic skin lesion segmentation with fully convolutional-deconvolutional networks,” *arXiv preprint arXiv:1703.05165*, 2017.

[3] K. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, 2019.

[4] Philipp, Tschandl, et al. "MILK10k: A hierarchical multimodal imaging learning toolkit for diagnosing pigmented and non-pigmented skin cancer and its simulators." *Journal of Investigative Dermatology,* 2025.

[5] MILK study team, “MILK10k Benchmark.” ISIC Archive, 2025, doi: 10.34970/262082.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[7] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.

[8] K. Gyeltshen, “Assignment 2 Part A”, *Unpublished coursework report*, PRML, University of Canberra, 2025.

[9] K.Gyeltshen, “Assignment 3 Code” *Unpublished coursework report*, PRML, University of Canberra, 2025.