AI for Skin Detection: Review and Example Implementation

David Franz
Victoria University of Wellington

***Abstract*—This project serves as a review and example implementation of the problem of using convolutional neural networks for skin cancer detection. We use the ISIC skin cancer dataset which consists of nine classes- three of which are cancerous conditions, one precancerous, and five benign conditions. We begin with a review of the state of the art in literature, then implement a CNN using PyTorch in a Google Colab notebook, then implement a second neural network based on experimentation which achieves better results than the previous one. We compare the results, and finish with a discussion of the implications (both positive and negative) and risks of this technology.**

I. Introduction

Skin cancer detection is a potential application of AI which carries a real risk of loss of human life- if a healthcare system is reliant on the accuracy of such a system, then a potential false negative result (someone has skin cancer but the system classifies it incorrectly as benign or even precancerous), this might significantly delay patient care which can lead to loss of life. Melanoma is the most deadly skin cancer, but early detection and treatment significantly improves the chance of survival and recovery. In this sense, this is really a safety critical system- software which if it failed has the potential for loss of human life. Unlike elevator software or rocket guiding software, CNN networks can't be evaluated to be as robust as those systems- which can be extensively tested with test cases and formally verified with LSAT equipped theorem prover languages. However, in an increasingly overworked healthcare system, a system which could automate the discovery of this has huge potential for beneficial outcomes for patients and improving of healthcare accessibility. We begin with a review of recent state of the art results applying CNNs to this problem, then implement an architecture recommended in a paper, then a second network based on our experimentation which gives a better result, and finish with a discussion of potential implications and risks. Beyond the lack of ability to formally verify deep learning models at the time of this paper, there is a particular risk for racial bias in the datasets- they are weighted heavily towards people with lighter skin leading to huge potential for racial inequality if patient outcomes rely on these models.

A. Colab links

1) Base CNN architecture based on paper: https://drive.google.com/file/d/1TtiIPg75nVJmL7R3knoLHSCyUZvTIDZ9/view?usp=sharing

2) Custom architecture: https://drive.google.com/file/d/12Ia2-hhVrSYvEVz6MjH_y5-N4tLVKxvx/view?usp=sharing

II. CNNs for Skin Cancer: literature review

We give a brief review of the history of CNNs for skin cancer detection as described section 'Convolutional Neural network in skin cancer detection' of the paper 'A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN)'.

The paper starts by noting the value CNNs posses for image processing over basic fully connected ANNs. Namely, by using convolutional filters of various sizes over our images, we are able to extract increasing more complex patterns with far fewer parameters. This efficiency can further be enhanced with techniques such as pooling layers and drop out.

A. Brinker et al., 2018

Provided the first systematic review of state of the art CNN models used for skin cancer detection. 13 papers were examined and showed high potential. However, the paper notes some of these used non publicly accessible datasets.

B. Hasan et al., 2019

Proposed a study using CNNs, doing simple binary class detection- benign or malignant. This achieved an accuracy of 89.5% and a training accuracy of 93.7% after using the public accessible data set. The paper suggests this was a state of the art performance at the time.

C. Helker et al, 2020

This paper proposed merging CNN datasets with patient information- giving the network access to data such as ethnicity, sex, age, location of lesion, etc lead to state of the art performance giving an accuracy of 97.49%. However, this approach ethical concerns for patient privacy. Also, due to doctor patient privacy laws, this information probably can't be assumed to be accessible in all regions.

D. Raja Subramanian et al., 2021

Proposed a paper using CNN which makes use of historical data of clinical images. The main aim of the research is to make a CNN model that has an accuracy of greater than 80%, false negative rate of less than 10% and a precision level of greater than 80% Several research papers and methods were surveyed and tried. An accuracy of 80% and greater was obtained with the HAM10000 dataset.

III. Training our models

A. Dataset

Kaggle Skin Cancer ISIC (9 classes of different skin conditions).[1] After basic cleaning we have 2,357 images: 1,697 train, 475 test, 185 validation. The images come from Kaggle pre split into train and test, but we manually combine them and split them randomly in the split above. Images are resized to a fixed square (28x28 for the first network as described in the paper; 54x54 in the second custom network), normalised, and augmented with flips, small rotations, and mild colour jitter.

B. Baseline CNN (PaperStyleCNN)

We implement the CNN architecture described in the paper 'A Skin Cancer Detection Interactive Application Based on CNN and NLP'. While the dataset used in this paper is different, we used this as a base example implementation. The architecture is as described in the paper:

Three 3×3 conv blocks (Conv$\rightarrow$BN$\rightarrow$ReLU), two 2×2 pools, then an MLP head with dropout (p=0.5). Widths: 32/64/128. FC head: 1152$\rightarrow$128$\rightarrow$64$\rightarrow$32$\rightarrow$9. Total params: 251,689.

C. Parallel-kernel CNN (CustomStyleCNN)

This was a network developed based on the PaperStyleCNN but with improvements made over time leading to improved accuracy, particular when examining the 'superclass' (cancerous, precancerous, benign). The archictecture is as follows:

Five parallel input kernels $(3, 5, 7, 9, 27)$, BN per branch, concat, a 3×3 mixing stack (64 channels), global average pooling, dropout (p=0.27), and a 9-way classifier. Total params: 290,025.

D. Training

Both trained with cross entropy loss (multiclass classification), Adam optimiser (lr 10^{-4}), batch 64, 100 epochs (but we use the result which performed best on the validation set during training for the evaluation of the models. This setup mirrors the regularisation/discipline recommended in [2], [3].

IV. Experiments and Results

A. Evaluation

Primary metric: top-1 accuracy on the 9-class validation set. We also collapse to superclasses (cancer / precancer / benign) to reflect clinical diagnosis. We report accuracy and class-wise recall (to expose false negatives). We look at both the overall accuracy of the model of the class predicition, the superclass prediction and the recall and the chance of the false negative result (model says benign when lesion is cancerous potentially leading to reduced patient care).

[1]https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic/data

TABLE I
Model comparison (validation/test).

Model	Params ($\sim$)	9-class Val Acc	9-class Test Acc	Superclass Val Acc
Baseline CNN	2.52×10^5	0.5892	–	0.6324
Parallel-kernel	2.90×10^5	0.6324	0.6253	0.7405

TABLE II
Superclass recall (rows=true).

Class	Baseline	Parallel
Cancer	$86/109 = 0.789$	$92/109 = 0.844$
Precancer	$25/33 = 0.758$	$28/33 = 0.848$
Benign	$6/43 = 0.140$	$17/43 = 0.395$

B. Main results and model comparison

Table I summarises the results.[2] The custom model learns faster and reaches a higher 9-class ceiling. Test accuracy (0.6253) tracks validation, suggesting limited extra overfit beyond late-epoch drift. Collapsing to superclasses lifts accuracy, with a larger gain for the parallel model and a clear improvement in malignant recall.

C. False negatives and operating point

A false negative on a malignant lesion is the failure mode that matters. Accuracy alone is not sufficient.

Observed recall (superclass). Using the confusion matrices, malignant recall improves from 0.789 (baseline: $86/109$) to 0.844 (parallel: $92/109$). This is fairly significant- in practise it means that 85% of the time that cancer is present in any form, the improved model will detect it. However, 15% non detection is probably still too high to be safe for real world use. Precancer recall rises from 0.758 to 0.848. Benign recall is low in both (0.140 vs. 0.395).

V. Bias and Transparency

Bias. Public skin cancer data sets are imbalanced and skew light-skin; performance can drop for darker skin tones and rare classes [1]. Our errors cluster among visually similar benign types, consistent with imbalance. Mitigations: reweighting or focal loss, targeted augmentation for minority classes and darker skin tones, domain balancing, and reporting per-group recall.

Transparency. CNNs remain hard to audit. A minimal model card should state: dataset composition, augmentation, thresholds, calibration, malignant recall, and known failure modes, with a misclassification gallery. The lack of proven safety in a safety critical system suggests that this should only be used for help, not a tool to be relied on yet.

[2]Custom architecture model best validation at epoch 72 (0.6324). Baseline peaked at 0.5892 around epoch 48.

```
Superclass Names: ['cancer', 'precancer', 'benign']
Superclass Accuracy: 0.6919
Mean predicted probability per superclass:
  - cancer    : 0.5921
  - precancer: 0.2029
  - benign    : 0.2050

Superclass Confusion Matrix (rows=true, cols=pred):
             cancer  precance    benign
   cancer        86         7        16
 precance         7        25         1
   benign        17         9        17
```

Fig. 1. Base model superclass results

VI. Conclusion

Two compact CNNs on ISIC reach 0.589 (baseline) and 0.632 (parallel) validation accuracy on 9 classes; the parallel model is 0.625 on test. Superclass collapse yields 0.740 vs. 0.632, with higher malignant recall for the parallel model.

Appendix: Tables and Figures

A.1 Dataset split

TABLE III
Dataset split (stratified).

Split	Images	Share
Train	1697	72.0%
Validation	475	20.2%
Test	185	7.9%

A.2 Hyperparameters

TABLE IV
Training hyperparameters.

Loss	Cross-entropy
Optimizer	Adam (lr $= 10^{-4}$, β_1=0.9, β_2=0.999)
Batch size	64
Epochs (max)	100, early stopping on val acc
Regularisation	Dropout (p=0.5/0.27), BatchNorm
Augment	Flips, small rotations, mild colour jitter
Checkpoint	Save best on val acc

A.3 Architectures (concise)

TABLE V
Model specs.

Baseline (251,689)	Conv3×3-32 → BN → Conv3×3-64 → BN → Pool → Conv3×3-128 → BN → Pool → MLP (dropout 0.5) → 9-way FC
Parallel (290,025)	Parallel stem: $\{3, 5, 7, 9, 27\}$ + BN → Concat → Conv3×3 mix (64) → BN → GAP → Dropout 0.27 → 9-way FC

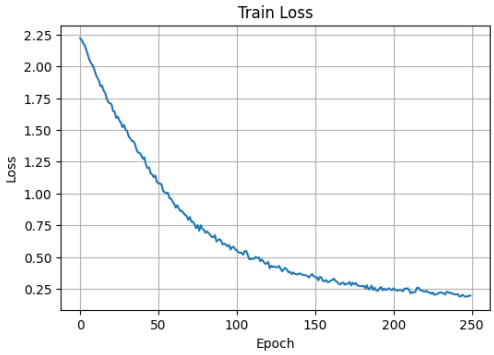

Fig. 2. Base model trained on 250 epochs

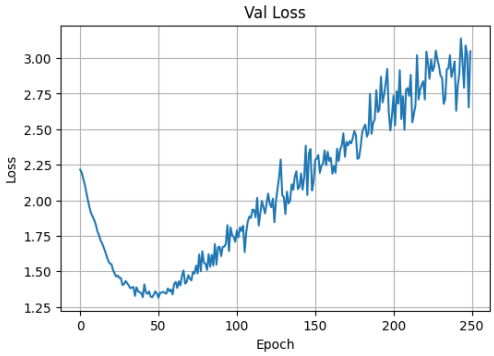

Fig. 3. Base model validation loss- starts increasing from 50th epoch

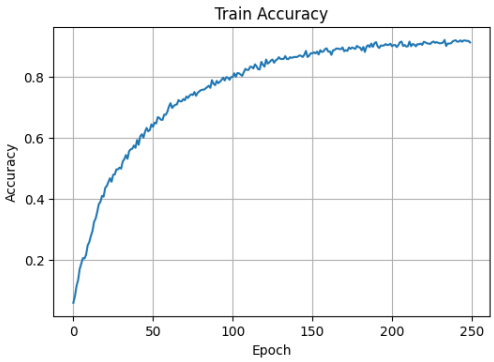

Fig. 4. Training data accuracy

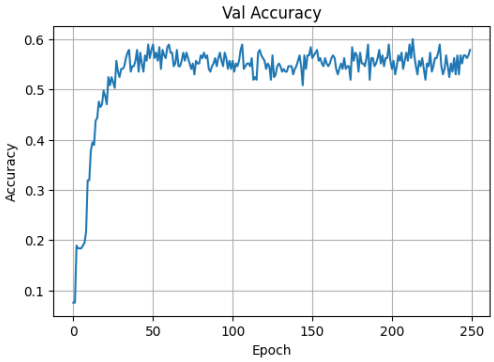

Fig. 5. Validation accuracy

```
Superclass Names: ['cancer', 'precancer', 'benign']
Superclass Accuracy: 0.7405
Mean predicted probability per superclass:
  - cancer    : 0.5739
  - precancer: 0.2032
  - benign    : 0.2228

Superclass Confusion Matrix (rows=true, cols=pred):
             cancer  precance    benign
   cancer        92         5        12
 precance         4        28         1
   benign        18         8        17
```

Fig. 6. Results from improved model architecture

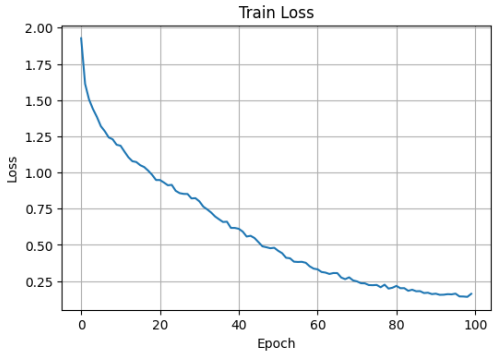

Fig. 7. Improved mode training loss

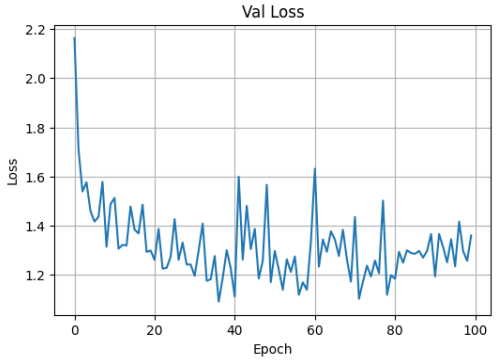

Fig. 8. Improved model validation loss- does not start to rise suggesting overfitting like the base model at epoch 50

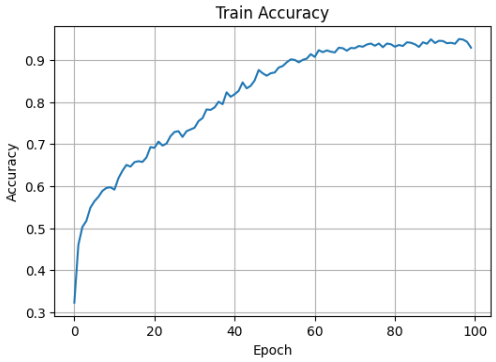

Fig. 9. Improved model training accuracy

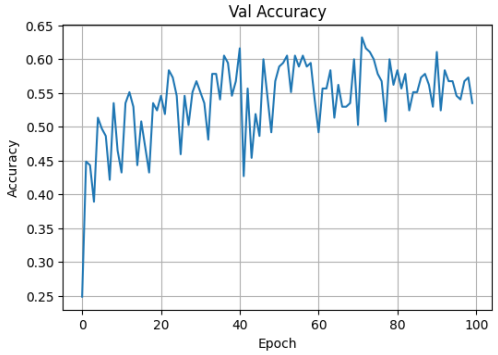

Fig. 10. Imrpoved model validation accuracy- more consistent

A.4 Figures

REFERENCES

[1] A. Shah *et al.*, "A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN)," *Clinical eHealth*, 6:76–84, 2023.

[2] X. Gong and Y. Xiao, "A Skin Cancer Detection Interactive Application Based on CNN and NLP," *Journal of Physics: Conference Series*, 2078:012036, 2021.

[3] M. M. Musthafa *et al.*, "Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification," *BMC Medical Imaging*, 24:201, 2024.

[4] "Skin Cancer ISIC (9 classes)," Kaggle. Available: https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic/data.