SKINSCAN AI: SKIN LESION CLASSIFICATION TOOL

Keshav Singal
Student# 1006277903
keshav.jindal@mail.utoronto.ca

Pasut Aranchaiya (Sean)
Student# 1011835935
pasut.aranchaiya@mail.utoronto.ca

ABSTRACT

Skin cancer is the most commonly diagnosed cancer in the United States, with over 5.4 million cases diagnosed annually in US alone. Melanoma, the deadliest form of skin cancer, accounted for 287,723 new cases and 60,712 deaths worldwide in 2018. Early detection significantly increases survival rates, with a 5-year survival rate of over 98.4%. This project aims to develop a hybrid deep learning model that leverages Convolutional neural networks (CNN's) via transfer learning and Artifical neural networks (ANN's). The model will classify dermoscopic images from the HAM10000 and BCN20000 dataset into seven distinct categories of skin lesions. The combined HAM and BCN dataset has 22,428 images of skin lesions that range from benign (safe) conditions like melanocytic nevi to malignant ones like melanoma and basal cell carcinoma. Not all skin lesions are cancerous; for example, Melanocytic Nevi are commonly known as "moles". Patient metadata such as age and sex will be taken into consideration. Prior research demonstrates that CNNs outperform classical models in image classification, achieving 89% accuracy compared to the 86% accuracy of Decision Trees and other traditional approaches. Our model architecture integrates CNNs for visual analysis and ANN's for processing patient metadata to enhance classification accuracy. Ethical concerns, such as bias in datasets and data privacy, are also addressed to ensure fairness and transparency in predictions. This project has the potential to assist dermatologists in accurate skin lesion classification, reducing miss-classification rates and false negatives in malignant case detection. A dangerous false negative is when a Melanoma skin lesion is classified as benign by the dermatologist. Our final model uses ensembling to average results from 4 different models developed, to achieve a balanced accuracy exceeding 74%.

—-Total Pages: 10

1 INTRODUCTION

Melanoma is a subset of skin cancer that occurs when pigment-producing cells (melanocytes) undergo adverse genetic mutation from UV radiation exposure (Bhattacharya et al., 2017). Over the last two decades, the number of patients diagnosed with melanoma has risen steadily, with 73,000 new estimated cases and over 9,000 deaths reported in 2015 (Bhattacharya et al., 2017). Early detection yields high survival rates (Stage 1A at 97%). However, without early diagnosis and preventive care, the cancer can quickly spread and become fatal (Stage 4 survival of 10-20%) (Bhattacharya et al., 2017). Poor diagnostic precision adds around $673 million in overall cost for managing the disease (Bhattacharya et al., 2017).

Current clinical diagnosis relies on the ABCDE method, which evaluates asymmetry, border irregularity, color variegation, lesion diameter, and evolution (Bhattacharya et al., 2017). However, diagnostic accuracy varies significantly among healthcare providers, with dermoscopy experts achieving 90% sensitivity, while general practitioners only reach 62% sensitivity (Menzies, 2005). This disparity is particularly concerning given Ontario's limited dermatologist availability of only 1.62 per

100,000 people as of 2014 (R., n.d.), highlighting the urgent need for accessible and accurate diagnostic tools.

Our project aims to develop a deep learning system that can classify dermatological images into seven distinct categories of skin lesions: melanocytic nevi (common benign moles), melanoma (aggressive skin cancer), basal cell carcinoma (slow-growing malignant cancer), actinic keratoses (precancerous lesions), benign keratosis-like lesions, vascular lesions, and dermatofibroma (Tschandl et al., 2018b). Deep learning presents an ideal solution for this multi-class classification challenge, as demonstrated by modern models achieving lower error rates (3.57%) than human baseline performance (5.1%) in complex image recognition tasks (APS360: Lecture 5, 2024). By developing an automated diagnostic tool capable of distinguishing between these specific lesion types, we can provide a scalable, cost-effective solution to improve early detection rates and address the critical shortage of dermatology specialists.

2 ILLUSTRATION

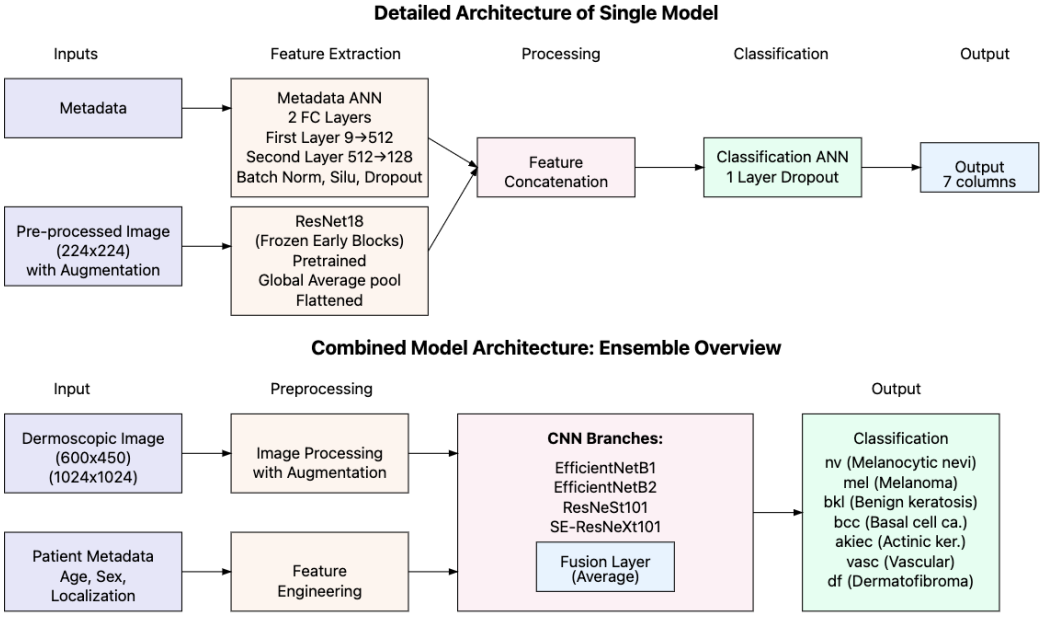


Figure 1: Project Architecture

The final model uses ensembling to average the results from 4 individual models. These models use EfficientNet B1 & B2, ResNeSt101, SE-ResNeXt101 for CNN transfer learning, and a common ANN architecture to process metadata.

3 BACKGROUND & RELATED WORK

Skin cancer classification research has evolved through numerous significant advances. Khan et al. (2021) established baseline comparisons using ISIC 2018 (Codella et al., 2018), demonstrating CNN's initial superiority (89% accuracy) over classical methods. (Dildar et al., 2021) furthered this by achieving 98% accuracy with CNNs versus lower-performing ANNs (86.66%), GANs (86.1%), and KNNs (93.15%). (Wu et al., 2022) revealed key challenges in CNN implementation, particularly regarding dataset diversity and diagnosis consistency. (Mahbod et al., 2020) introduced successful ensemble methods combining ResNet-50 and DenseNet-121, while ISIC 2019's winning solution achieved 92.5% accuracy through ensembling mainly on EfficientNets and advanced preprocessing.

These developments are particularly significant when considering (Menzies, 2005) findings that dermoscopy experts achieve 90% sensitivity in melanoma detection, while general practitioners

only reach 62% sensitivity. The relatively high accuracy and recall rates of modern ML approaches suggest promising potential for aiding doctors in diagnosing skin lesions.

Table 1: Performance Comparison of Skin Cancer Classification Models

Model Type	Accuracy	Dataset	Recall	Key Characteristics
Classical Models (Khan et al., 2021)				
Decision Trees	86%	ISIC 2018	0.70	Our baseline
KNN(K-Nearest Neighbor)	81%	ISIC 2018	0.68	Traditional model
Logistic Regression	88%	ISIC 2018	0.70	Our baseline
Neural Network	89%	ISIC 2018	0.70	Best Model
Neural Networks (Dildar et al., 2021)				
CNN	98%	ISIC	0.95	Highest accuracy
ANN	86.66%	90 dermoscopic images	N/A	Limited dataset
GAN	86.1%	ISIC	N/A	Experimental
KNN(Kohonen Self-Organizing Neural Network)	93.15%	DermQuest and Dermnet datasets	N/A	Second highest accuracy
Advanced Methods (2019-2020)				
EfficientNet Ensemble	92.5%	ISIC 2019	0.453	An effective approach

4 Data processing

4.1 Data Sourcing

Our data will be sourced from "Skin Cancer MNIST: HAM10000" on Kaggle (Tschandl et al., 2018a). The HAM10000 dataset is a large collection of labeled preprocessed dermatoscopic images with metadata columns aimed at supporting skin lesion classification research. The preprocessed images are centered on the skin lesions without any black border and the metadata include diagnosis type (dx type), age, sex, and localization. It contains 10,015 images of important pigmented skin lesions from 7 classes (Mader, 2018), including Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (vasc)(Mader, 2018). The HAM10000 dataset is imbalanced, leading to unoptimized performance; therefore, we have included the BCN20000 dataset to increase the amount of data and boost our model's performance.

BCN20000 was aimed to be used in classification of unconstrained dermoscopic images of pigmented lesions in hard-to-diagnose locations (Hernández-Pérez et al., 2024). Due to this and the fact that BCN20000 is not preprocessed, images from this dataset suffer from irregularities like black borders, uncentered images, and distinct patterns of images from various locations. The BCN20000 consists of 18,946 dermoscopic images, and it includes 3 features: age, sex, and localization (Hernández-Pérez et al., 2024). Each image and corresponding features belongs to one of eight classes, which are seven classes of HAM10000 and an additional class Squamous carcinoma (SCC).

4.2 Data Cleaning

4.2.1 HAM10000

The dataset contains certain rows with values of unknown for "sex", "localization" as well as "age". These values need to be removed by manner of dropping said rows from the dataset. Overall, there are 57 rows with null values for the age column, 234 rows with unknown values for the localization column, and 57 rows with unknown values for the sex column. The dataset also has multiple duplicates where values are the same in all columns; therefore, we removed those rows. Next, we remove the irrelevant columns, specifically "lesion id" and "dx type". Then, as values in "age" are represented in intervals of 5 (0 refers to age from 0 to 5), we add a constant value of 2.5 to this column to get the center of the interval. To prepare HAM10000 for concatenation, we used certain mapping to map values from "localization" to suitable values of "anatom_site_general" column of BCN20000 dataset, allowing it to be concatenated.

4.2.2 BCN20000

BCN20000 contains rows with N/A values for "age_approx" with 74 rows, "anatom_site_general" with 173 rows, and "sex" with another 74 rows. Unlike HAM10000, BCN20000 does not have "unknown" values and only has N/A values. We removed these rows by dropping them. From here the data must be purged of any duplicates. We conducted a test to find which combination of columns with the same values is considered duplicates. We found that duplicates in "lesion_id" are pictures of the same lesions, but they can either be taken on the same date or different date. So, we will consider pictures of the same lesions from different capture dates as different data points, and drop duplicates in "lesion_id" and "capture_date". From here we drop irrelevant columns namely "lesion_id", "capture_date", and "split", while removing rows with an irrelevant class named "SCC". Similar to HAM10000, we added a constant value of 2.5 to "age_approx". Next, we mapped the values in the "diagnosis" column to match the "dx" column in HAM10000 and renamed the columns in the dataset to match those of the HAM10000 dataset in preparation for data concatenation.

The two datasets can now be concatenated and shuffled. The "sex" and "localization" columns are then encoded using one-hot encoding, while the "age" column is retained as numerical values. Finally, a 70%-15%-15% train-validation-test split is done on the cleaned dataset, and the minority classes in the training set are duplicated to solve data imbalance which results in 31,012 data points.

4.3 Processing

4.3.1 HAM10000

As images from the HAM10000 dataset are fairly preprocessed, we do minor adjustment to the images with the purpose to include some randomness and eliminate unnecessary features by cropping. For each image, we applied the Shades of Gray method with Minkowski norm p=6 which is believed to be essential when working with multiple datasets. Then, we resize so the height is 1.25x the input height, while maintaining the aspect ratio. Then, a random square center crop with size between 0.8 and 1 of the resized image is implemented. However, the validation and test set take a fixed crop with size 0.9. After that, the image is resized again to our desired input size. Finally, the image is normalized. For training data, the processed images will be augmented; otherwise, they are ready to be input into the model.

4.3.2 BCN20000

For images from BCN20000, their borders are removed first before following the same processing procedure as the HAM10000 dataset. To remove the borders, we looked at the images and found that the borders are circle and centered on the images. We binarize each image and find the left and top edges of the circle by scanning along the center row to detect the left edge and scanning along the center column to detect the top edge where the pixel transitions from black to white. We then crop the image according to the edges and if the resulting area is more than 0.9 of the old image area, we consider the image as having no border and do not crop.

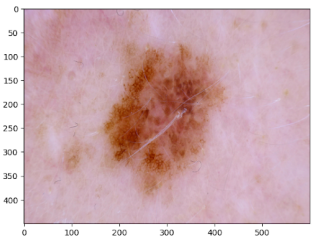


Figure 2: Unprocessed image

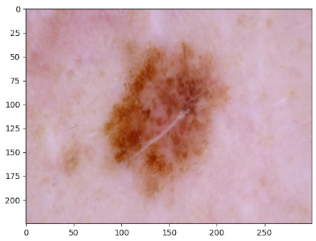


Figure 3: Processed image

4.4 Data Augmentation

We have a transformation pipeline that applies a series of data augmentation to enhance the diversity of input images. Considering the high number of duplicated data points, we followed fairly complex

data augmentation techniques to increase the difficulty of models' learning and reduce overfitting. The pipeline begins with spatial transformations like random transpositions, vertical and horizontal flips, each applied with a 50% probability, to introduce spatial variability. Brightness and contrast are adjusted randomly with up to a 20% limit, and a probability of 75%. Various blur and noise effects are randomly applied to simulate real-world imperfections in the images. These include motion blur, median blur, Gaussian blur, and Gaussian noise. Each effect has specific parameters to control its intensity: motion blur with a blur limit of 5, median blur with a blur limit of 5, Gaussian blur with a blur limit of 5, and Gaussian noise with a variance limit ranging from 5.0 to 30.0. One of these augmentations will be applied to the image with a probability of 70%. To introduce geometric distortions, one of the following augmentations is applied with a 70% probability: optical distortion with a distortion limit of 1.0, grid distortion with 5 steps and a distortion limit of 1.0, or elastic deformations with an alpha value of 3. Color adjustments, like histogram equalization (CLAHE) and hue-saturation-value shifts, further diversify color profiles. Next, random shifts, scaling, and rotations are applied to enhance spatial resilience, followed by resizing the image to a fixed size (e.g., 224x224 pixels). A cutout augmentation removes a rectangular region. Finally, the images are normalized to a standard scale and converted into tensors for use.

5 ARCHITECTURE

Our final model is an ensemble of four different models where each model uses pretrained CNN for the images, fully connected layers on the metadata, and fully connected layers for classification. Throughout the four models, we varied the CNNs, while keeping the metadata layers and classification layers the same. For the CNN part, we used EfficientNetB1, EfficientNetB2, SE-ResNeXt101, ResNeSt101, all of which are pre trained on ImageNet. For the metadata layers, we used a combination of two layers. The first layer receives a 9 dimensional vector of the one hot encoded metadata and transforms it to an intermediate representation of 512 dimensions before applying batch normalization, SiLU activation function, and dropout of probability 0.3. The intermediate representation is then passed to the second linear layer with batch normalization and SiLU activation function, transforming the 512 dimensional vector to a 128 dimensional vector. To obtain the prediction for each of the models, we pass the output from the CNN layers through a global average pooling layer and flatten it. Then, we concatenate the flattened vector with the output from metadata layers before dropping out the concatenated output with probability 0.5. Finally, we obtain the 7 classes prediction through the classification layer which consists of one linear layer with input size according to the distinct output size of each CNN. From this, the ensemble takes the predictions of the four models and averages them to get the final prediction.

To train each of the model, we processed all images to 224x224 pixels with augmentation and used a batch size of 64. Our training method alternates between training the few final CNN blocks and the metadata/classification layers. Specifically, we train the few last CNN blocks for 20 epochs, select the best model based on balanced accuracy, then switch to training the metadata and classification layers for another 20 epochs. This cycle continues until training is complete. In our approach, we trained the model for a total of 60 epochs. We found that this approach is able to achieve better results than training the whole model, while making it harder for the model to overfit.

To address data imbalance, we duplicated data points for minority classes and used weighted cross entropy loss. The weights are computed by $\sqrt{N/N_k}$, where N is the total number of data points before duplication and Nk is the number of datapoints for the k class before duplication. As we have duplicated a lot of data (to mitigate against class imbalance), our models can learn the pattern of the data very easily and tend to overfit. Therefore, we used complex data augmentation strategies and Adam optimizer with weight decay. We also added one cycle learning rate scheduling strategy to make the models converge faster and escape the local minima more efficiently. From all of these, we still saw the pattern of overfitting. Hence, we applied the MixUp strategy on every batch with a probability of 50%. We also added gradient clipping to solve exploding gradients from our large CNN architecture. Finally, we select the best model with the highest balanced accuracy across all epochs for ensembling.

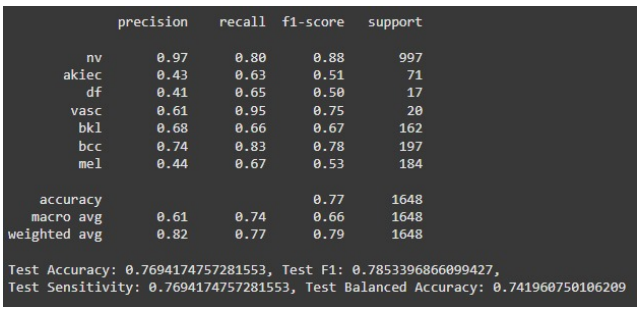
6 BASELINE MODEL

We implemented a simple baseline model which is Logistic Regression to evaluate our primary model's performance. As our data suffers from data imbalance, we have used balanced accuracy and sensitivity as our evaluation metrics.

6.1 LOGISTIC REGRESSION

Based on a research named "Towards Skin Cancer Classification Using Machine Learning and Deep Learning Algorithms: A Comparison", the performance of Logistic Regression on skin cancer classification task is slightly worse than CNN (Khan & Jr, 2021). Due to this and the fact that Logistic Regression is a simple machine learning model which requires minimal tuning, we have determined that it is a suitable baseline model. For multiclass classification, Logistic Regression is a model which uses the softmax function to map input features to the probabilities of being in a class. In our implementation, we used Scikit-learn's Logistic Regression model with "lbfgs" solver, 1000 maximum iteration, and regularization parameter C of 0.001. We adjusted C to prevent overfitting, and modified iteration limit to make sure the model is able to converge. To prepare inputs for the model, unprocessed images are resized to a width of 24 pixels and a height of 18 pixels without any augmentation. These resized images are then flattened and concatenated with the metadata to form a comprehensive input vector. The model takes the standardization of these input vectors as inputs. Due to its low complexity, the model demonstrated remarkable robustness against over fitting, enabling it to perform well even on a dataset with high number of duplications and features. This led to an impressive balanced accuracy of 0.473 and sensitivity of 0.588 on the test set.

7 QUANTITATIVE RESULTS



	precision	recall	f1-score	support
nv	0.97	0.80	0.88	997
akiec	0.43	0.63	0.51	71
df	0.41	0.65	0.50	17
vasc	0.61	0.95	0.75	20
bkl	0.68	0.66	0.67	162
bcc	0.74	0.83	0.78	197
mel	0.44	0.67	0.53	184
accuracy			0.77	1648
macro avg	0.61	0.74	0.66	1648
weighted avg	0.82	0.77	0.79	1648

Test Accuracy: 0.7694174757281553, Test F1: 0.7853396866099427,
Test Sensitivity: 0.7694174757281553, Test Balanced Accuracy: 0.741960750106209

Figure 4: Performance metrics in test set

Our multi-class skin lesion classification model achieved strong overall performance with a test accuracy and recall of roughly 76.9% and a weighted F1-score of 0.79 across seven distinct lesion classes. The model demonstrates particularly strong performance in identifying melanocytic nevi (nv) with precision of 0.97 and F1-score of 0.88. The model achieves clinically relevant sensitivity of 67% for melanoma detection. This exceeds typical physician performance rates (0.62 recall) for melanoma detection.

The class-wise performance metrics reveal robust capabilities across different lesion types, with notably high recall for vascular lesions (0.95) and strong balanced performance for basal cell carcinoma (precision: 0.74, recall: 0.83). Our model's balanced accuracy of 74.2% is within the range of the ISIC 2018 winner's balanced accuracy (88.5%) (Codella et al., 2018). In addition, our model is able to outperform Logistic Regression in both balanced accuracy and sensitivity. The model exceeds a recall of 60% across all lesion types despite extreme imbalances in dataset. The largest class, melanocytic nevi (nv), contains 6,649 cases, while rare classes like dermatofibroma (df) only have 113 cases, a nearly 60 fold drop.

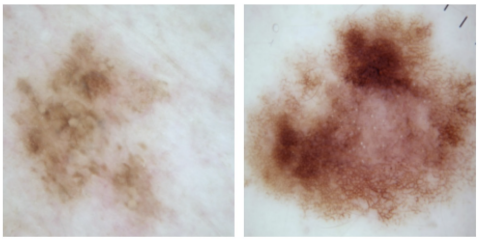


Figure 5: Mis-classfied images of NV - Melanocytic nevi (common moles) as Melanoma

8 Qualitative Results

Our analysis of melanocytic nevi (common moles) misclassifed as melanoma revealed that the deep learning model independently discovered and applied the ABCDE melanoma diagnostic criteria without explicit human training. While the model effectively detected **A**symmetry, **B**order irregularities, **C**olor variations, and **E**volving color patterns, it struggled with **D**iameter assessment. The absence of scale references and dimensional metadata in the preprocessing pipeline prevented the model from evaluating the critical diameter threshold of 6mm, leading to false positive classifications of melanoma. This limitation will become particularly evident when the model encounters small benign moles that display melanoma-like visual characteristics but fall below the size threshold for concern. This exposes a crucial gap in the model's diagnostic framework.

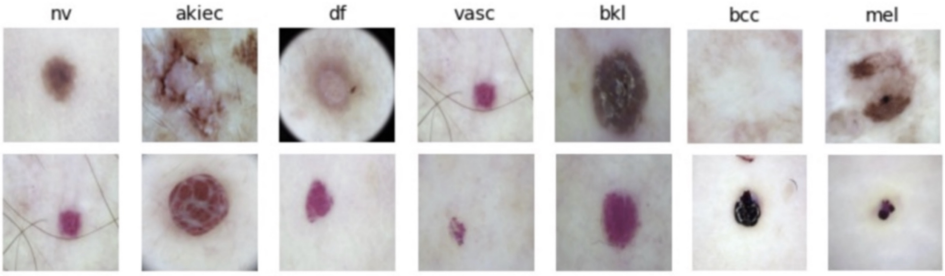


Figure 6: First row: 7 different skin lesions, Second row: Vascular cell lesions images

It's evident from figure 6 and the confusion matrix that our model can easily classify vascular lesions with a 95% recall. We have plotted one image from each skin lesion class (figure 7) to understand the model's excellent performance in classifying vascular lesions. Upon examination, we can see that vascular lesions have a distinct characteristic of being an extremely pink dot. We see 10 more images to validate our theory. Each vascular lesion picture is predominantly either a pink or black dot within a white surrounding. This suggests that classifying vascular lesions is easier than classifying other pigmented lesions due to their distinct characteristics.

9 Evaluate model on new data

Our model was able to get a balanced accuracy exceeding 74%, and was trained on HAM10000 and BCN20000 data. This is close to ISIC 2018 winner's 88.5% (Mahbod et al., 2020). It is important to note that we used both HAM10000 and BCN20000 for training our model, whilst ISIC 2018 only used mainly HAM10000 data. Also, our balanced accuracy exceeds the leaders of ISIC2019 (74% VS 63.6%). The ISIC2019 dataset consisted of images from both HAM10000, BCN20000 and MSK datasets. We theorize that the ISIC winners have a lower performance than our model due to having additional training images from the MSK dataset which is not used by us. This would make their training process more difficult as they have more variation in image types. Furthermore, ISIC 2019 competitors had their models tested on different data compared to ours. Hence, we cannot be conclusively sure that our solution would perform better than them. However, our model has a similar/higher accuracy than ISIC winners from 2018 and 2019. This can be interpreted as a positive sign for our model performance on real world data (relative to other deep learning approaches).

The model is robust for real world data as it is trained on multiple datasets (HAM10000 & BCN20000). This ensures that the model learns from multiple data sources and develops robust feature recognition that can generalize better on new, unseen data from different clinical environments. Training on mutliple datasets also ensures that the model does not get a high accuracy by fixating on dataset specific characteristics.

To evaluate our dataset, we set aside 15% of the data that was unseen by the model during training. This data was completely untouched during hyperparameter tuning. We had a test accuracy and recall of roughly 77%, with a recall score of 79%. Furthermore, the model showed robust performance by achieving a recall exceeding 60% on all classes.

During development, we implemented complex data augmentations like random brightness adjustments (±20%), contrast variations (±20%), spatial transformations (50% probability of flips/transpositions), multiple blur types (motion/median/Gaussian with blur limit of 5), and Gaussian noise (variance 5.0-30.0). These augmentations prepare the model for variable image quality, lighting conditions, different camera angles etc that are common in real world scenarios. Additionally, the model has strong performance across images in both HAM10000 and BCN20000 datasets. These images were captured in different clinical conditions, providing evidence for the model's ability to generalize to different imaging methods.

Furthermore, the model had strong performance across skin lesion classes despite extreme class imbalances (6,649 melanocytic nevi cases vs. 113 dermatofibroma cases). This is a 59:1 ratio. This demonstrates robust generalization even with limited training examples for rare conditions.

10 Discussion

Our final model ensembles four different neural network architectures which utilize both dermoscopic images and metadata. For each model in the ensemble, we used a fairly large architecture since we found that simple architectures cannot capture the complex pattern in classifying pigmented lesion types, resulting in underfitting. However, larger models like EfficientNet tend to overfit considering the amount of duplications in HAM10000 and BCN20000 that are needed to solve data imbalance. Therefore, we leverage complex data augmentation pipeline, weight decay, and CNN layers freezing. We found a suitable architecture which is an ensemble of EfficientNetB1, EfficientNetB2, SE-ResNeXt101, and ResNeSt101, and suitable training techniques for the architecture. From this, the model is able to achieve an impressive test set balanced accuracy of 0.742 and a test set sensitivity of 0.769. The result exceeded our expectations as it achieved a comparable score to the ISIC 2018 which had a balanced accuracy of 0.885 (Codella et al., 2018). We outperformed all of our baseline models.

From evaluating qualitative results, we found that the model is able to classify melanoma based on ABCDE approach used by dermatology experts. We didn't provide the model with any instruction to use ABCDE. The model is unable to make out lesion diameters from the image itself. Lesion diameters plays a crucial role in diagnosing melanoma. This could suggest why the confusion matrix had significant mis-classifications of melanocytic nevi as melanoma. Melanoma's are on average 6 mm in diameter, while nv's are between 1-3 mm. Furthermore, The model handled extreme class imbalances in the dataset very well. The dataset had over 6,649 cases of melanocytic nevi (nv) and just 113 cases of dermatofibroma (df) (a 59:1 ratio). Despite this, the model maintained strong recall across all classes, with a minimum recall of 63%. The model achieved a sensitivity of 67% for melanoma detection, which outperformed general doctors (62% recall) in classify melanoma. Our model used complex ensemble approaches, multiple CNN architectures with sophisticated training techniques. This successfully addressed the data imbalance problems and enabled the model to effectively understand the complex characteristics of skin lesions.

11 Ethical Considerations

The deployment of AI in skin cancer diagnosis requires careful consideration of patient safety and diagnostic integrity. The most critical concern is the potential impact of false negatives, where melanomas are incorrectly classified as benign lesions. With Stage 1 melanoma having a 98% five-year survival rate compared to just 10 - 20 % for Stage 4, diagnostic error delays can have severe

consequences for patient outcomes Bray et al. (2018). To address this risk, our system implements mandatory dermatologist verification for all AI predictions, with automated escalation protocols for cases showing low confidence scores or concerning features. Additionally, we maintain comprehensive data security protocols in compliance with HIPAA U.S. Department of Health Human Services (2020) regulations, implementing end-to-end encryption and strict access controls to protect sensitive patient information.

System errors also raise important questions about medical liability and accountability. When an AI system contributes to diagnostic decisions, determining responsibility for adverse outcomes becomes more complex Price (2017). We address this through a clear framework where AI serves as a supportive tool rather than an autonomous decision-maker, with final diagnostic authority remaining with board-certified dermatologists. Any predictions falling below defined confidence thresholds automatically trigger senior dermatologist review, ensuring appropriate clinical oversight and maintaining physicians' essential role in patient care. This structured approach helps mitigate risks while leveraging AI's capabilities to enhance rather than replace medical expertise Topol (2019).

12 Project Difficulty/Quality

The fundamental challenge of skin cancer classification is evidenced by (Menzies, 2005) findings that dermoscopy experts achieve 90% sensitivity in melanoma detection while general practitioners only reach a small 62% recall. Our dataset compounds this challenge with extreme imbalances across seven classes - from melanocytic nevi (n=6,649) to dermatofibroma (n=113), creating a 59:1 ratio. Additional complexity comes from duplicate lesion IDs between BCN and HAM datasets (8,479 and 2,543 respectively). We employed oversampling with multipliers scaled by class size (50x for dermatofibroma, 42x for vascular lesions, 5x for melanoma) to make the dataset balanced. We also used anti-overfitting techniques including high dropout (0.3), extensive augmentation with 12 transformations, and mixup training with α =0.4. Furthermore, our model classifies skin lesion's in 7-8 distinct classes. This problem is much more challenging that a binary skin cancer/ no skin cancer classification problem.

The medical nature of the dataset leads to strong data imbalances, more ethical regulations, and much higher adverse consequences of false negatives. We trained 8 different CNN's (EfficientNet variants, ResNet18, ResNeSt101, SE-ResNeXt) totaling over 100 million parameters. From these 8 trained CNN models, we only picked 4 (EfficinetNet B1,B2, ResNest101, SE-ResNeXt101) which did well on validation set for ensembling. Each network processes 224x224 pixel dermoscopic images through CNN branches while parallel ANNs handle 9 metadata features through batch-normalized layers with SiLU activation.

We implemented weighted cross-entropy loss (sqrt(N/Nk) class weights where N=10,986 total samples), one-cycle learning rate scheduling, and gradient clipping at 1.0. This stabilized training with duplicated samples that were made during data augmentation to mitigate dataset imbalances. The computational demands were substantial - the models were trained on A100 GPUs and batch size of 64. We paid for A100 GPU's, which are around 50-55 times faster than normal CPU's. Despite the added computational power, the entire team spent over 100 hours in training different models. Our final model achieved 75% balanced accuracy through ensembling, with particular emphasis on minimizing false negatives for critical cancerous lesions like melanoma and basal cell carcinoma. Early detection of melanoma improves 5-year survival rates from 15% to over 97% (Bhattacharya et al., 2017).

Our model uses both BCN20000 and HAM10000 datasets for training, enabling it to recognize skin lesions across different medical imaging systems and clinical environments. This approach ensures robust model performance in real-world settings, where image quality and capture conditions can vary substantially between healthcare facilities.

13 Google Colab Link

`https://colab.research.google.com/drive/1KCFxgy9dRUaq4mT5ytLKJ6Fz9WFWkvL8?usp=sharing`

REFERENCES

APS360: Lecture 5. Week 5: Convolutional neural networks - part ii, 2024. URL https://q.utoronto.ca/courses/363324/files/33584930?module_item_id=6191305. Lecture slides, University of Toronto.

Arjun Bhattacharya, Andrew Young, Alex Wong, Seth Stalling, Michelle Wei, and Drew Hadley. Precision diagnosis of melanoma and other skin lesions from digital images. In *AMIA Joint Summits on Translational Science Proceedings*. AMIA, July 26 2017. https://pmc.ncbi.nlm.nih.gov/articles/PMC5543387/#r1-2612795.

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 2018. doi: 10.3322/caac.21492.

Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Isic 2018: Skin lesion analysis towards melanoma detection, 2018. URL https://challenge.isic-archive.com/landing/2018/. International Skin Imaging Collaboration (ISIC) Challenge.

Muhammad Dildar, Muhammad Akram, Tahira Fatima, and Noorul Raza. Skin cancer detection: A review using deep learning techniques. *International Journal of Environmental Research and Public Health*, 18(10):5479–5490, may 2021. doi: 10.3390/ijerph18105479.

Carlos Hernández-Pérez, Marta Combalia, Sergi Podlipnik, Noel C. F. Codella, Veronika Rotemberg, Allan C. Halpern, Ondrej Reiter, Cristina Carrera, Antonio Barreiro, Brian Helba, Susana Puig, Veronica Vilaplana, and Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild. *Scientific Data*, June 17 2024. doi: 10.1038/s41597-024-03387-w. https://www.nature.com/articles/s41597-024-03387-w.

Iqra Khan and Ijist Jr. Towards skin cancer classification using machine learning and deep learning algorithms: A comparison. 3:110–118, 12 2021. doi: 10.33411/IJIST/2021030508.

Kai S. Mader. Skin cancer mnist: Ham10000. Kaggle, September 20 2018. URL https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000.

Amirreza Mahbod, Gerald Schaefer, Isabella Ellinger, Alain Pitiot, Chunliang Wang, and Rupert Ecker. Fusing fine-tuned deep features for skin lesion classification. *Computerized Medical Imaging and Graphics*, 84:101–111, 2020. doi: 10.1016/j.compmedimag.2020.101796.

Scott W. Menzies. The performance of solarscan. *Archives of Dermatology*, November 1 2005. URL https://jamanetwork.com/journals/jamadermatology/fullarticle/400665.

W Nicholson Price. Medical malpractice and black-box medicine. *Big Data, Health Law, and Bioethics*, pp. 295–306, 2017.

M. R. R. Changes in the practice patterns and demographics of ontario dermatologists. *Journal of Cutaneous Medicine and Surgery*, n.d. URL https://pubmed.ncbi.nlm.nih.gov/29519145/.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. Skin cancer mnist: Ham10000, 2018a. URL https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.

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, 2018b. doi: 10.1038/sdata.2018.161. https://www.nature.com/articles/sdata2018161.

U.S. Department of Health Human Services. Summary of the HIPAA security rule. 2020. URL https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html.

X. Wu, Y. Li, Q. Liu, and W. Jiang. A comparative study of convolutional neural network variants for skin cancer classification. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1402–1412, April 2022.