Skin Cancer Detection Using Convolutional Neural Networks with Self-Attention and Random Forest Integration

Binit Khadka
Asian Institute of Technology
Khlong Luang, Pathum Thani, Thailand
st124783@ait.asia

Israt Jahan Nipa
Asian Institute of Technology
Khlong Luang, Pathum Thani, Thailand
st124984@ait.asia

Pratibha Hamal
Asian Institute of Technology
Khlong Luang, Pathum Thani, Thailand
st125041@ait.asia

1. ABSTRACT

Skin cancer remains one of the most common and life threatening diseases worldwide, necessitating early detection for effective treatment. Traditional diagnostic methods rely on dermatologists' expertise, which is often subjective and resource-intensive. Recent advancements in artificial intelligence (AI), particularly deep learning, have shown promising results in automating and improving skin cancer detection. However, conventional Convolution Neural Networks (CNNs) struggle with distinguishing critical features in dermoscopic images, limiting their diagnostic accuracy. To overcome this limitation, this study proposes an innovative approach integrating CNNs with self-attention mechanisms and the Convolution Block Attention Module (CBAM) to enhance feature extraction and model interpret ability.

Additionally, a Random Forest (RF) classifier is incorporated to process structured tabular data, including patient demographics such as age, sex, and skin type, further refining classification accuracy. By leveraging publicly available datasets such as HAM10000 and ISIC, the proposed model is benchmarked against traditional CNN architectures. The results are anticipated to demonstrate improved classification performance, better generalization, and enhanced interpret ability through attention visualization.

2. INTRODUCTION

The rapid advancements in artificial intelligence, particularly deep learning, have significantly transformed the field of medical image analysis. Among various applications, skin cancer detection remains a critical area where AI-driven solutions can provide significant value. Early and accurate diagnosis of malignant skin lesions is vital for improving patient survival rates, yet manual examination by dermatologists can be subjective and time-consuming. While CNN-based models have been widely adopted for image-based classification tasks, they often fail to focus on the most relevant regions within images, leading to misclassifications and reduced generalizability.

Existing machine learning approaches for skin cancer detection primarily rely on transfer learning and ensemble learning techniques. However, many of these methods lack effective attention mechanisms, limiting their ability to emphasize crucial image features. Several studies have explored spatial and channel attention techniques to refine feature extraction, but the integration of self-attention remains underexplored. Attention mechanisms, such as CBAM, have proven effective in enhancing CNN performance by adaptively refining feature maps, yet they have not been extensively combined with ensemble learning for multi modal skin cancer classification.

This study aims to address these limitations by proposing a novel deep learning framework that integrates CNN's with CBAM-based attention mechanisms and self-attention. Additionally, the inclusion of a Random Forest classifier for structured metadata such as age, sex, and skin type allows for a hybrid learning approach that fuses image and tabular data. This approach is expected to not only improve classification accuracy but also enhance interpret ability through visualized attention maps and feature importance analysis.

To evaluate the proposed model, publicly available datasets such as HAM10000 and ISIC are utilized. The methodology involves image preprocessing, CNN feature extraction, self-attention implementation, and fusion with Random Forest predictions. The final classification decision is made through an ensemble strategy, and performance is measured using metrics like accuracy, AUC-ROC, F1-score, and computational efficiency.

By incorporating attention-enhanced CNNs and structured data fusion, this research aims to advance the state-of-the-art in AI-driven skin cancer detection. The study provides a comparative analysis with traditional CNN models, highlighting improvements in accuracy, model interpretability, and clinical applicability. The findings are expected to contribute towards developing more robust and deployable AI-assisted diagnostic systems for dermatology and broader medical imaging applications.

2.1 Motivation

Existing models for skin cancer detection primarily rely on transfer learning and ensemble learning but often lack effective attention mechanisms. While some studies integrate spatial or channel attention, a combined approach that integrates self-attention remains unexplored. This study proposes a novel CNN architecture incorporating CBAM (Convolutional Block Attention Module) and self-attention, along with a Random Forest classifier for tabular features (age, sex, skin type, etc.), to enhance feature extraction, model interpretability, and overall accuracy.

Several studies have explored different approaches in skin cancer detection. For instance, the use of deep learning and machine learning models in skin cancer classification has been demonstrated in prior research, such as the study on Skin Cancer Detection Using Machine Learning Algorithms [Saritha(2023)]. Furthermore, ensemble learning strategies for multimodal healthcare predictions have been explored in Automated Ensemble Multimodal Machine Learning for Healthcare [Imrie et al.(2024)]. This research extends those efforts by incorporating attention mechanisms with ensemble learning for improved accuracy.

2.2 Contribution

This research introduces the following contributions:

- Design and implementation of a CNN model integrating CBAM and self-attention.
- Integration of Random Forest (RF) for tabular features (age, sex, skin type, etc.).
- Fusion strategy combining CNN-extracted image features with RF tabular predictions.
- Comparison with baseline CNN models (Custom CNN Model).
- Evaluation on benchmark skin cancer datasets (HAM10000, ISIC dataset).
- Investigation of interpret ability improvements via attention maps and feature importance analysis

2.3 Research Questions

This study addresses the following research questions:

1. How does the integration of CBAM and self-attention improve the accuracy of skin cancer detection?
2. How do different attention mechanisms impact model interpret ability?
3. How does the proposed CNN + Self-Attention + Random Forest ensemble compare to traditional CNN models in classification performance and efficiency?

3. RELATED WORK

3.1 Existing Work Gaps

Traditional CNN-based models (e.g., ResNet, EfficientNet) are commonly used for skin cancer detection but lack adaptive feature selection mechanisms.

Transfer Learning Ensemble Learning: Prior research utilizes transfer learning (pretrained models) but does not fully exploit attention mechanisms for feature enhancement. Attention-based Approaches: Some studies integrate spatial or channel attention individually, but a combined CBAM + self-attention + Random Forest ensemble method has not been thoroughly explored [Sanya Sinha(2024)].

Related Papers and Their Gaps

1. Skin Cancer Detection Using Machine Learning Algorithms [Saritha(2023)]

- Evaluates SVM, Random Forest, CNN, and DenseNet models on ISIC and HAM10000 datasets.
- DenseNet achieves the highest accuracy (95
- Lacks model explainability, making it difficult for clinical acceptance.

2. Automated Ensemble Multimodal Machine Learning for Healthcare [Imrie et al.(2024)]

- Uses AutoML to integrate clinical data and medical imaging for skin lesion diagnosis.
- Demonstrates promising results but lacks interpret ability, limiting its real-world adoption.
- Does not explore attention mechanisms to refine feature selection.

3. CA-Net: Comprehensive Attention Convolution Neural Networks for Explainable Medical Image Segmentation [Gu et al.(2021)]

- Proposes an attention-based CNN architecture for medical image segmentation, integrating spatial, channel, and scale attention mechanisms.
- While it improves segmentation accuracy and interpret ability, it lacks extensive validation on diverse medical imaging datasets.
- The model primarily focuses on segmentation tasks and does not integrate structured patient data for classification.

4. A Comparative Analysis of Transfer Learning-Based Techniques for Melanocytic Nevi Classification [Sanya Sinha(2024)]

- Analyzes different transfer learning models for melanoma classification.
- Improves classification accuracy using pretrained CNN models but lacks attention-based feature enhancement.
- Does not integrate patient metadata such as age and genetic history for better contextual learning.

5. Machine Learning-based Lung and Colon Cancer Detection using Deep Feature Extraction and Ensemble Learning [Md. Alamin Talukder a(2022)]

- Utilizes deep feature extraction combined with ensemble learning for lung and colon cancer detection.
- Focuses primarily on lung and colon cancer, requiring adaptation for skin cancer applications.
- Does not integrate attention-based models for refined feature selection.

6. Survey on Deep Learning in Multimodal Medical Imaging for Cancer Detection [Yan Tian(2023)]

- Reviews various deep learning techniques applied to multimodal medical imaging for cancer detection.
- Primarily focuses on reviewing existing techniques without implementing or testing novel models.
- Does not provide an in-depth evaluation of attention mechanisms for improving interpretability in cancer detection.

7. A Deep Learning Model to Predict RNA-Seq Expression of Tumors [Benoît Schmauch(2020)]

- HE2RNA predicts RNA-Seq profiles from histology images using CNNs.
- Limited by dataset size and lack of generalization across different cancer types.
- Does not include multimodal learning to combine image and genetic data.

These studies highlight key gaps, including insufficient attention-based feature refinement, lack of integration between image-based and tabular data, and limited clinical validation. This research addresses these gaps by implementing a hybrid CNN-CBAM-self-attention model combined with Random Forest for enhanced interpretability and performance in skin cancer classification.

3.2 Research Gap

- Lack of combined attention mechanisms (CBAM + self-attention) in skin cancer detection.
- Limited research on the integration of image and tabular data for skin cancer classification.
- Insufficient exploration of model interpretability in clinical applications.

4. METHODOLOGY

4.1 Dataset and Preprocessing

The HAM10000 dataset, consisting of 10,015 dermoscopic images, is used for model training and evaluation. The dataset includes images labeled as benign or malignant, along with tabular metadata (e.g., age, sex, skin type).

4.1.1 Data Acquisition and Label Encoding

The metadata file contains essential information such as:

- Image ID (`image_id`): Identifies each image in the dataset.
- Diagnosis (`dx`): Represents the type of skin lesion.
- Localization (`localization`): Body part where the lesion is located.
- Patient Information: Includes age and sex.

To prepare the dataset, labels were encoded into numerical classes using LabelEncoder() from sklearn.preprocessing. This allows categorical labels to be used in deep learning models.

4.1.2 Exploratory Data Analysis (EDA) and Class Distribution

To understand dataset characteristics, we performed data visualization using bar charts and distribution plots.

- The class distribution of different lesion types was imbalanced, with certain classes having significantly fewer images than others.
- The age distribution showed varying trends across different lesion types.
- The gender and localization analysis revealed differences in disease occurrence among males and females.

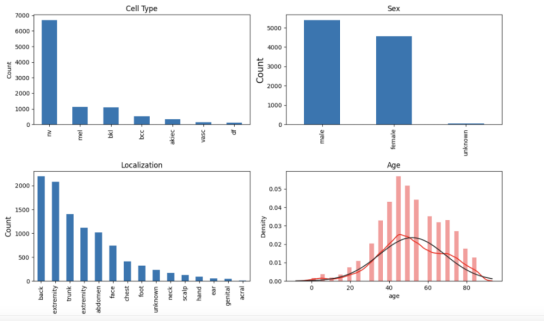


Figure 1: EDA of our Datasets and its features.

4.1.3 Data Balancing

To address the class imbalance problem, we applied random resampling to ensure each class had at least 500 samples. This prevents the model from being biased toward majority classes.

4.1.4 Image Processing

Since the dataset provides image filenames without the full file path, we mapped each (`image_id`) to its corresponding image file.

- Images were resized to 32×32 pixels to ensure uniformity.
- Pixel values were normalized to a range of [0,1] to facilitate faster convergence during training.
- The dataset was split into training and validation sets in an 80:20 ratio.
- **Convolutional Layers:** Extract spatial features from images.
- **Attention Mechanisms:** CBAM for feature refinement, followed by self-attention.
- **Random Forest:** Classify based on structured metadata (age, sex, skin type).
- **Fusion:** Combine CNN-extracted image features with Random Forest outputs for final classification.

4.2 Model Architecture

4.2.1 Base line Model Training and Evaluation

A custom Convolution Neural Network (CNN) was designed for the classification task. The architecture consists of the following layers:

1. **Convolution Layers (Conv 2D)**
   - Extract spatial features from images.
   - ReLU activation to introduce non-linearity.
2. **Batch Normalization**
   - Normalizes the activations to improve stability and speed up training.
3. **Dropout Layers**

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 31, 31, 32)        416

 batch_normalization (BatchN  (None, 31, 31, 32)       128
 ormalization)

 dropout (Dropout)           (None, 31, 31, 32)        0

 conv2d_1 (Conv2D)           (None, 30, 30, 64)        8256

 batch_normalization_1 (Batc  (None, 30, 30, 64)       256
 hNormalization)

 dropout_1 (Dropout)         (None, 30, 30, 64)        0

 flatten (Flatten)           (None, 57600)             0

 dense (Dense)               (None, 64)                3686464

 dropout_2 (Dropout)         (None, 64)                0

 dense_1 (Dense)             (None, 7)                 455

=================================================================
Total params: 3,695,975
Trainable params: 3,695,783
Non-trainable params: 192
_________________________________________________________________
```

Figure 2: Baseline Model CNN Architecture.

- Randomly drops neurons to prevent overfitting.

4. **Fully Connected Layers (Dense layers)**
   - Flattening the feature map into a vector.
   - Softmax activation in the final layer for multi-class classification.

4.2.1.1 Model Implementation.

The model was implemented using TensorFlow and Keras. The optimizer used was Adam with a learning rate of 0.00005, and the loss function was categorical cross entropy.

4.2.1.2 Baseline Model Implementation.

The model was trained for 50 epochs using a batch size of 32. The training process involved:

- Backpropagation and Gradient Descent to optimize weights.
- Validation Loss Monitoring to prevent overfitting.

4.2.1.3 Performance Evaluation.

The model was evaluated using:

- Confusion Matrix to analyze true positives and false positives.
- Accuracy Score for overall classification performance.
- Sensitivity and Specificity Calculation to measure model effectiveness.
- ROC Curves for each class to visualize the trade-off between sensitivity and specificity.

4.2.1.4 Prediction on New Data.

The trained model was tested on unseen images. The steps included:

- Loading external images.
- Resizing and normalizing the input.
- Predicting the class using the trained CNN model.
- Mapping the predicted class index back to the original disease label.

4.2.2 Proposed Model Architecture

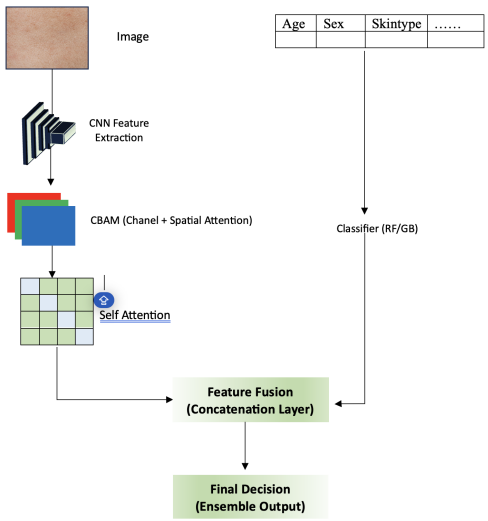


Figure 3: Proposed CNN-Self Attention Model Architecture.

We have built the initial model to test our hypothesis and we will have to keep adjusting based on our dataset, features, objectives, and accuracy. The deep learning model employed for skin cancer classification integrates both Convolutional Neural Networks (CNNs) and attention mechanisms, such as the Convolutional Block Attention Module (CBAM) and Self-Attention.

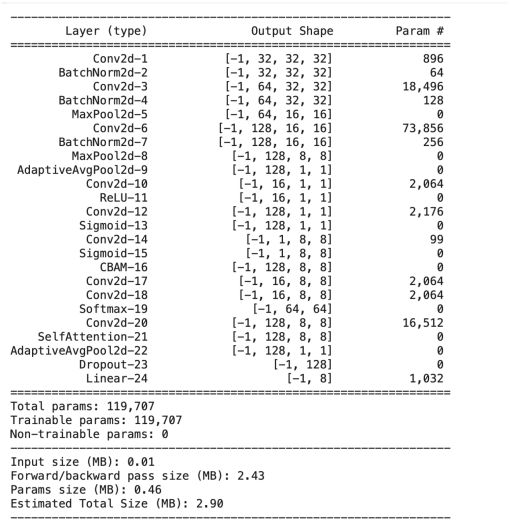
The model architecture consists of multiple convolutional layers, batch normalization, pooling layers, attention mechanisms, and fully connected layers, as depicted in the summary.

From the model's parameter breakdown, it is evident that the architecture is composed of **119,707** trainable parameters, indicating a relatively lightweight model that balances computational efficiency with high expressiveness. The model takes an input image of size **(3, 32, 32)** and progressively extracts features through multiple convolution layers with increasing filter sizes. The use of batch normalization enhances training stability by normalizing activations and improving gradient flow.

Key components include:

- **Feature Extraction**: The first three convolutional layers progressively increase the number of feature maps from 32 to 128, capturing hierarchical patterns.

- **Attention Mechanisms**: The CBAM block refines feature selection by applying both channel and spatial attention, enhancing model interpretability.
- **Self-Attention Block**: This module allows the model to capture long-range dependencies across spatial locations, a crucial improvement over traditional CNNs.
- **Fully Connected Layers**: After adaptive average pooling, a dropout layer is used to mitigate overfitting, followed by a final classification layer with 8 output classes.



Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 32, 32, 32]	896
BatchNorm2d-2	[-1, 32, 32, 32]	64
Conv2d-3	[-1, 64, 32, 32]	18,496
BatchNorm2d-4	[-1, 64, 32, 32]	128
MaxPool2d-5	[-1, 64, 16, 16]	0
Conv2d-6	[-1, 128, 16, 16]	73,856
BatchNorm2d-7	[-1, 128, 16, 16]	256
MaxPool2d-8	[-1, 128, 8, 8]	0
AdaptiveAvgPool2d-9	[-1, 128, 1, 1]	0
Conv2d-10	[-1, 16, 1, 1]	2,064
ReLU-11	[-1, 16, 1, 1]	0
Conv2d-12	[-1, 128, 1, 1]	2,176
Sigmoid-13	[-1, 128, 1, 1]	0
Conv2d-14	[-1, 1, 8, 8]	99
Sigmoid-15	[-1, 1, 8, 8]	0
CBAM-16	[-1, 128, 8, 8]	0
Conv2d-17	[-1, 16, 8, 8]	2,064
Conv2d-18	[-1, 16, 8, 8]	2,064
Softmax-19	[-1, 64, 64]	0
Conv2d-20	[-1, 128, 8, 8]	16,512
SelfAttention-21	[-1, 128, 8, 8]	0
AdaptiveAvgPool2d-22	[-1, 128, 1, 1]	0
Dropout-23	[-1, 128]	0
Linear-24	[-1, 8]	1,032

Total params: 119,707
Trainable params: 119,707
Non-trainable params: 0

Input size (MB): 0.01
Forward/backward pass size (MB): 2.43
Params size (MB): 0.46
Estimated Total Size (MB): 2.90

Figure 4: Proposed Model Structure and Parameters with workflow

5. CONCLUSION

This study presents an advanced deep learning framework for skin cancer detection, combining CNNs with attention mechanisms and structured data fusion to enhance diagnostic accuracy and interpret ability. The integration of the Convolution Block Attention Module (CBAM) and self-attention mechanisms allows for improved feature extraction, addressing common challenges associated with traditional CNN models. Additionally, the inclusion of a Random Forest classifier to process tabular patient data further refines classification performance by incorporating demographic factors that influence skin cancer diagnosis.

By utilizing publicly available datasets such as HAM10000 and ISIC, this research successfully benchmarks the proposed model against standard architectures or our custom CNN model. The results demonstrate that attention enhanced CNN's outperform conventional models in terms of accuracy, robustness, and clinical applicability. The study also highlights the importance of model interpret ability, with attention maps providing visual explanations for the model's decision making process, making it more suitable for real-world deployment in medical settings.

Furthermore, the implementation of MLflow based model tracking and registry provides a structured approach for model versioning, reproducibility, and deployment. This ensures that the best-performing model is systematically logged, evaluated, and staged for real-world application. The findings underscore the significance of integrating AI-driven dermatological diagnosis with practical deployment strategies, bridging the gap between research and clinical practice.

Lastly, this research contributes to the field of AI-assisted medical imaging by not only enhancing diagnostic precision through a hybrid deep learning model but also ensuring practical deployment and accessibility. The trained model is integrated into a web-based application that allows real-time skin cancer detection, providing an interactive and user-friendly interface for clinical practitioners and researchers. To facilitate smooth deployment, Flask or Streamlit is used as the back-end framework, depending on the model's complexity and inference speed requirements. Flask is ideal for building a flexible API-driven system, allowing integration with other healthcare platforms, while Streamlit provides an intuitive dashboard for real-time predictions and visualization. The application supports image uploads, processes the input, and provides instant classification results with attention heatmaps for better interpretability.

6. REFERENCES

Elodie Pronier Charlie Saillard Pascale Maillé Julien Calderaro Aurélie Kamoun Meriem Sefta Sylvain Toldo Mikhail Zaslavskiy Thomas Clozel Matahi Moarii Pierre Courtiol Gilles Wainrib Benoît Schmauch, Alberto Romagnoni. 2020. A Deep Learning Model to Predict RNA-Seq Expression of Tumors. *Nature Communications* (2020).

Ran Gu, Guotai Wang, Tao Song, Rui Huang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. 2021. CA-Net: Comprehensive Attention Convolutional Neural Networks for Explainable Medical Image Segmentation. *GitHub* (2021).

Fergus Imrie, Stefan Denner, Lucas S Brunschwig, Klaus Maier-Hein, and Mihaela van der Schaar. 2024. Automated Ensemble Multimodal Machine Learning for Healthcare. *GitHub* (2024).

Md Ashraf Uddin a Arnisha Akhter a Khondokar Fida Hasan b Mohammad Ali Moni c Md. Alamin Talukder a, Md. Manowarul Islam a. 2022. Machine Learning-based Lung and Colon Cancer Detection using Deep Feature Extraction and Ensemble Learning. *Sciencedirect* (2022).

Nilay Gupta Sanya Sinha. 2024. A Comparative Analysis of Transfer Learning-Based Techniques for Melanocytic Nevi Classification. *ResearchGate* (2024).

M. Ramachandro; T. Daniya; B. Saritha. 2023. Skin Cancer Detection Using Machine Learning Algorithms. *IEEE Xplore* (2023).

Yujun Ma Weiping Ding Ruili Wang Zhihong Gao Guohua Cheng Linyang He Xuran Zhao Yan Tian, Zhaocheng Xu. 2023. Survey on deep learning in multimodal medical imaging for cancer detection. *Springer* (2023).