

copy-of-untitled14-1-4

November 12, 2025

```
[ ]: import os
import gc
gc.collect()

os.environ['KAGGLE_USERNAME'] = "noraadel"
os.environ['KAGGLE_KEY'] = "76e19273178ad0c217ad1e28c1757c48"

!pip install -q kaggle

!kaggle datasets download -d kmader/skin-cancer-mnist-ham10000

import zipfile

with zipfile.ZipFile("skin-cancer-mnist-ham10000.zip", 'r') as zip_ref:
    zip_ref.extractall("skin_cancer_data")

print(" Dataset downloaded and extracted successfully!")
```

```
[ ]: import pandas as pd, glob
import gc
gc.collect()

meta_path = "skin_cancer_data/HAM10000_metadata.csv"
meta = pd.read_csv(meta_path)
print(meta.head())
meta['image_id'] = meta['image_id'].astype(str) + '.jpg'

img_dirs = glob.glob("skin_cancer_data/*")
print("Possible data folders (showing first 10):")
for d in img_dirs[:10]:
    print(d)
```

```
[ ]: import cv2, numpy as np, os, pandas as pd
from tqdm import tqdm
from tensorflow.keras.preprocessing.image import ImageDataGenerator
import gc
gc.collect()
```

```

meta = pd.read_csv("skin_cancer_data/HAM10000_metadata.csv")

IMG_SIZE = 128
raw_image_folders = [
    "skin_cancer_data/ham10000_images_part_1",
    "skin_cancer_data/ham10000_images_part_2"
]
preproc_dir = "preprocessed_images_aug"
os.makedirs(preproc_dir, exist_ok=True)

# === Hair Removal Function ===
def remove_hair_inpaint(img_rgb):
    gray = cv2.cvtColor(img_rgb, cv2.COLOR_RGB2GRAY)
    kernel = cv2.getStructuringElement(cv2.MORPH_RECT, (17,17))
    blackhat = cv2.morphologyEx(gray, cv2.MORPH_BLACKHAT, kernel)
    _, mask = cv2.threshold(blackhat, 10, 255, cv2.THRESH_BINARY)
    mask = cv2.dilate(mask, cv2.getStructuringElement(cv2.MORPH_RECT, (3,3)),
↳ iterations=1)
    inpainted = cv2.inpaint(cv2.cvtColor(img_rgb, cv2.COLOR_RGB2BGR), mask, 1,
↳ cv2.INPAINT_TELEA)
    inpainted = cv2.cvtColor(inpainted, cv2.COLOR_BGR2RGB)
    return inpainted

# === Data Augmentation Generator ===
datagen = ImageDataGenerator(
    rotation_range=25,
    width_shift_range=0.1,
    height_shift_range=0.1,
    zoom_range=0.1,
    brightness_range=[0.8, 1.2],
    horizontal_flip=True,
    vertical_flip=True,
    fill_mode='nearest'
)

# === Process and Augment ===
rows = []
count = 0

for idx, row in tqdm(meta.iterrows(), total=len(meta), desc="Preprocessing +
↳ Augmentation"):
    img_name = row['image_id'] + ".jpg"
    found = False
    for d in raw_image_folders:
        p = os.path.join(d, img_name)

```

```

if os.path.exists(p):
    found = True
    img = cv2.imread(p)
    if img is None:
        continue
    img = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)
    img = remove_hair_inpaint(img)
    img = cv2.resize(img, (IMG_SIZE, IMG_SIZE))
    img = img.astype('float32') / 255.0

    # Save original preprocessed
    save_path = os.path.join(preproc_dir, img_name)
    cv2.imwrite(save_path, cv2.cvtColor((img*255).astype('uint8'), cv2.
↪COLOR_RGB2BGR))
    rows.append({'image_id': img_name, 'label': row['dx'], 'path': ↪
↪save_path})

    # Generate 3 augmentations per image
    img_expanded = np.expand_dims(img, 0)
    aug_iter = datagen.flow(img_expanded, batch_size=1)

    for i in range(3):
        aug_img = next(aug_iter)[0]
        aug_path = os.path.join(preproc_dir, f"{row['image_id']}_aug{i}.
↪jpg")
        cv2.imwrite(aug_path, cv2.cvtColor((aug_img*255).
↪astype('uint8'), cv2.COLOR_RGB2BGR))
        rows.append({'image_id': f"{row['image_id']}_aug{i}", 'label': ↪
↪row['dx'], 'path': aug_path})
        count += 1
    break

if not found:
    continue

preproc_df = pd.DataFrame(rows)
print(f" Preprocessed + Augmented {count} images. Dataframe shape: {preproc_df.
↪shape}")
preproc_df.head()

```

```

[ ]: import matplotlib.pyplot as plt

def show_before_after(meta_df, raw_folders, preproc_dir, num_samples=5):
    plt.figure(figsize=(10, num_samples * 4))

    for i, row in meta_df.head(num_samples).iterrows():

```

```

img_name = row['image_id'] + ".jpg"
orig_img = None
for d in raw_folders:
    p = os.path.join(d, img_name)
    if os.path.exists(p):
        orig_img = cv2.imread(p)
        if orig_img is not None:
            orig_img = cv2.cvtColor(orig_img, cv2.COLOR_BGR2RGB)
            break

proc_path = os.path.join(preproc_dir, img_name)
proc_img = cv2.imread(proc_path)
if proc_img is not None:
    proc_img = cv2.cvtColor(proc_img, cv2.COLOR_BGR2RGB)

if orig_img is not None and proc_img is not None:
    plt.subplot(num_samples, 2, i*2 + 1)
    plt.imshow(orig_img)
    plt.title(f"Original: {img_name}")
    plt.axis('off')

    plt.subplot(num_samples, 2, i*2 + 2)
    plt.imshow(proc_img)
    plt.title(f"Processed: {img_name}")
    plt.axis('off')

plt.tight_layout()
plt.show()

show_before_after(meta, raw_image_folders, preproc_dir, num_samples=5)

```

```

[ ]: import numpy as np
from tqdm import tqdm
import gc
gc.collect()

def load_all_images(df):
    X = []
    y = []
    for _, r in tqdm(df.iterrows(), total=len(df)):
        p = r['path']
        img = cv2.imread(p) # BGR
        img = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)
        img = img.astype('float32') / 255.0
        X.append(img)
        y.append(r['label'])
    return np.array(X), np.array(y)

```

```

X, y = load_all_images(preproc_df)
print("X shape:", X.shape, "y shape:", y.shape)
unique, counts = np.unique(y, return_counts=True)
print("Class distribution:")
for u, c in zip(unique, counts):
    print(u, c)

```

```

[ ]: import numpy as np
from sklearn.preprocessing import LabelEncoder
from sklearn.utils import resample
from tensorflow.keras.utils import to_categorical
import gc

print("Original dataset shape:", X.shape, len(np.unique(y)), "classes")

le = LabelEncoder()
y_enc = le.fit_transform(y)

unique, counts = np.unique(y_enc, return_counts=True)
print("Before balancing:")
for u, c in zip(unique, counts):
    print(f" {le.inverse_transform([u])[0]}: {c}")

from collections import Counter
counter = Counter(y_enc)
max_count = max(counter.values())

X_resampled = []
y_resampled = []

for cls, count in counter.items():
    idx = np.where(y_enc == cls)[0]
    X_cls = X[idx]
    y_cls = y_enc[idx]
    X_up, y_up = resample(X_cls, y_cls, replace=True, n_samples=max_count,
↳ random_state=42)
    X_resampled.append(X_up)
    y_resampled.append(y_up)

X_res = np.vstack(X_resampled)
y_res = np.hstack(y_resampled)

del X_resampled, y_resampled, X_cls, y_cls, X_up, y_up
gc.collect()

y_res_cat = to_categorical(y_res, num_classes=len(le.classes_))

```

```

print("\nAfter balancing:")
unique_res, counts_res = np.unique(y_res, return_counts=True)
for u, c in zip(unique_res, counts_res):
    print(f" {le.inverse_transform([u])[0]}: {c}")

print("\nBalanced dataset shape:", X_res.shape, y_res_cat.shape)

```

```

[ ]: import tensorflow as tf
from tensorflow.keras.applications import VGG16, ResNet50, MobileNetV2,
↳EfficientNetB0
from tensorflow.keras.models import Model
from tensorflow.keras.layers import GlobalAveragePooling2D, Dense, Dropout,
↳Input
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.preprocessing.image import ImageDataGenerator
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import StratifiedKFold
from sklearn.metrics import accuracy_score, precision_score, recall_score,
↳f1_score
from imblearn.over_sampling import RandomOverSampler
import numpy as np
import pandas as pd
import gc

gc.collect()

#
MODEL_NAME = 'efficientnetb0'
IMG_SHAPE = (IMG_SIZE, IMG_SIZE, 3)
K = 5
EPOCHS = 8
BATCH_SIZE = 32

le = LabelEncoder()
y_enc = le.fit_transform(y)
num_classes = len(le.classes_)
X = tf.image.resize(X, (128, 128)).numpy()

print("      :", num_classes)
print("      :", X.shape)

def get_feature_extractor(model_name, input_shape):
    if model_name == 'vgg16':
        base = VGG16(weights='imagenet', include_top=False,
↳input_shape=input_shape)

```

```

        elif model_name == 'resnet50':
            base = ResNet50(weights='imagenet', include_top=False,
↪input_shape=input_shape)
        elif model_name == 'mobilenetv2':
            base = MobileNetV2(weights='imagenet', include_top=False,
↪input_shape=input_shape)
        elif model_name == 'efficientnetb0':
            base = EfficientNetB0(weights='imagenet', include_top=False,
↪input_shape=input_shape)
        else:
            raise ValueError("Unknown model name")

        x = GlobalAveragePooling2D()(base.output)
        feature_model = Model(inputs=base.input, outputs=x)
        return feature_model

print(" Extracting features...")
feature_model = get_feature_extractor(MODEL_NAME, IMG_SHAPE)
features = feature_model.predict(X, batch_size=32, verbose=1)
print("Features shape:", features.shape)

ros = RandomOverSampler(random_state=42)
X_res, y_res = ros.fit_resample(features, y_enc)
print("Balanced shape:", X_res.shape, y_res.shape)

def create_classifier(input_dim, num_classes):
    inp = Input(shape=(input_dim,))
    x = Dense(256, activation='relu')(inp)
    x = Dropout(0.3)(x)
    out = Dense(num_classes, activation='softmax')(x)
    model = Model(inputs=inp, outputs=out)
    model.compile(optimizer=Adam(learning_rate=1e-4),
                  loss='sparse_categorical_crossentropy',
                  metrics=['accuracy'])
    return model

# K-Fold Cross Validation
skf = StratifiedKFold(n_splits=K, shuffle=True, random_state=42)
metrics_per_fold = []
fold_no = 1

for train_idx, val_idx in skf.split(X_res, y_res):
    print(f"\n--- Fold {fold_no}/{K} ---")
    X_train, X_val = X_res[train_idx], X_res[val_idx]
    y_train, y_val = y_res[train_idx], y_res[val_idx]

```

```

model = create_classifier(X_train.shape[1], num_classes)

cb = [
    tf.keras.callbacks.EarlyStopping(monitor='val_loss', patience=3,
↪restore_best_weights=True),
]

history = model.fit(X_train, y_train,
                    validation_data=(X_val, y_val),
                    epochs=EPOCHS,
                    batch_size=BATCH_SIZE,
                    callbacks=cb,
                    verbose=1)

#
y_pred = np.argmax(model.predict(X_val), axis=1)
acc = accuracy_score(y_val, y_pred)
prec = precision_score(y_val, y_pred, average='macro', zero_division=0)
rec = recall_score(y_val, y_pred, average='macro', zero_division=0)
f1 = f1_score(y_val, y_pred, average='macro', zero_division=0)
print(f"Fold {fold_no} - acc: {acc:.4f}, prec: {prec:.4f}, rec: {rec:.4f},
↪f1: {f1:.4f}")

metrics_per_fold.append({'fold': fold_no, 'acc': acc, 'prec': prec, 'rec':
↪rec, 'f1': f1})
model.save(f"classifier_fold_{fold_no}.h5")

fold_no += 1
gc.collect()

#
df_metrics = pd.DataFrame(metrics_per_fold)
print("\nCross-validation results:")
print(df_metrics)
print("\nMean metrics:")
print(df_metrics.mean())

```

```
[ ]: model.save(f"model_fold_{fold_no}.h5")
```

```
[ ]: import os
import numpy as np
from tensorflow.keras.models import load_model
from sklearn.metrics import classification_report

rng = np.random.default_rng(42)
idxs = np.arange(len(X_res))

```

```

test_size = int(0.1 * len(X_res))
test_idx = rng.choice(idxs, size=test_size, replace=False)
X_test_hold = X_res[test_idx]
y_test_hold = y_res[test_idx]

models = []
for i in range(1, K+1):
    path = f"/content/model_fold_6.h5"
    if os.path.exists(path):
        print(f"Loading {path}")
        models.append(load_model(path))
    else:
        print(f"Warning: {path} not found, skipping.")

if len(models) == 0:
    print(" No models found! Skipping ensemble evaluation.")
else:
    print("Loaded", len(models), "models for ensemble.")

X_test_input = X_test_hold

probs = np.zeros((len(X_test_input), len(le.classes_)))
for m in models:
    probs += m.predict(X_test_input, batch_size=32)

probs /= len(models)
y_ens_pred = np.argmax(probs, axis=1)

print(classification_report(y_test_hold, y_ens_pred, target_names=list(le.
↵classes_)))

```

```

[ ]: # Save model of this fold
model.save = f"/content/model_fold_{i}.h5"

fold_no += 1

```

```

[ ]: import os
import numpy as np
from tensorflow.keras.models import load_model
from sklearn.metrics import classification_report, accuracy_score
import matplotlib.pyplot as plt

#
rng = np.random.default_rng(42)
idxs = np.arange(len(X_res))
test_size = int(0.1 * len(X_res))

```

```

test_idx = rng.choice(idxs, size=test_size, replace=False)
X_test_hold = X_res[test_idx]
y_test_hold = y_res[test_idx]

#             ) for)
model1 = load_model("/content/classifier_fold_1.h5")
model2 = load_model("/content/classifier_fold_2.h5")
model3 = load_model("/content/classifier_fold_3.h5")
model4 = load_model("/content/classifier_fold_4.h5")
model5 = load_model("/content/classifier_fold_5.h5")

models = [model1, model2, model3, model4, model5]

#             ) feature_model)
try:
    X_test_input = feature_model.predict(X_test_hold, batch_size=32)
    print(" Using feature_model for feature extraction.")
except:
    X_test_input = X_test_hold
    print(" feature_model not available - using raw test data instead.")

accs = []
probs = np.zeros((len(X_test_input), len(le.classes_)))

print("\n Individual Model Evaluation Reports:")
for i, model in enumerate(models, start=1):
    preds = model.predict(X_test_input, batch_size=32)
    y_pred = np.argmax(preds, axis=1)
    acc = accuracy_score(y_test_hold, y_pred)
    accs.append(acc)
    probs += preds
    print(f"\nModel {i} accuracy: {acc:.4f}")
    print(classification_report(y_test_hold, y_pred, target_names=list(le.
↪classes_)))

probs /= len(models)
y_ens_pred = np.argmax(probs, axis=1)
ensemble_acc = accuracy_score(y_test_hold, y_ens_pred)

print("\n Ensemble Accuracy:", ensemble_acc)
print(classification_report(y_test_hold, y_ens_pred, target_names=list(le.
↪classes_)))

labels = [f"Model {i}" for i in range(1, 6)] + ["Ensemble"]
values = accs + [ensemble_acc]

plt.figure(figsize=(8, 5))

```

```
plt.bar(labels, values)
plt.ylim(0, 1)
plt.title("Model Accuracy Comparison")
plt.ylabel("Accuracy")
plt.xlabel("Model")
plt.grid(axis="y", linestyle="--", alpha=0.7)
plt.show()
```



Title

A Hybrid Deep Learning Model for Multiclass Skin Cancer Classification
Using ConvNeXtV2 and Separable Self-Attention Mechanisms

Prepared By

Mishaal Mishaan Al-Otaibi

Id: 251002903

Submitted in fulfillment of the requirements for the degree Master of Science in Artificial
Intelligence

At

Collage of Informatics
Midocean University
UAE

Supervisor Name: **Dr:** Hagar Saleh

Collage of Informatics, Midocean University, Comoros

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Allah, the Almighty, for granting me the strength, patience, and guidance to complete this research.

I am sincerely grateful to my research supervisor for their invaluable guidance, continuous support, and insightful feedback throughout the course of this study. Their expertise and encouragement were fundamental in shaping the direction and quality of this research.

I would also like to extend my heartfelt appreciation to the university and department for providing the necessary resources, facilities, and academic environment that greatly facilitated the successful completion of this work.

My sincere thanks go to my colleagues and friends for their assistance, motivation, and constructive discussions, which significantly contributed to enriching the content and scope of this research.

I am especially grateful to my family for their unwavering support, understanding, and encouragement throughout this journey. Their patience and belief in me were a constant source of inspiration and strength.

Finally, I would like to acknowledge everyone who contributed, directly or indirectly, to the completion of this research, whether through guidance, advice, or moral support. Their contributions have been invaluable and deeply appreciated.

.

Abstract

Skin cancer is among the most widespread and lethal illnesses in the entire globe, and that is why efficient and automated diagnostic methodology is very urgent. In this paper, a multiclass skin lesion classification hybrid deep learning model is proposed, which incorporates ConvNeXtV2 convolutional blocks with separable self-attention to synthesize local features extraction with a global contextual message. The model used the HAM10000 dataset to improve generalization and performance with the help of the transfer learning, data augmentation, and class balancing. Five-fold stratified cross-validation showed the best results with an average accuracy of 93.5, precision of 93.1, recall of 91.2 and F1-score of 92.2 and ROC-AUC of 0.957 which were better compared to the usual CNN and Vision Transformer architectures. Grad-CAM imagery was used to establish that the model paid attention to clinically significant lesion areas enhancing interpretability and clinical confidence. The suggested model provides an efficient and interpretable AI model that can be integrated into teledermatology and real-time clinical systems. The next step in the development of work is the introduction of multimodal data fusion, edge deployment, and clinical validation to increase its practical applicability in dermatological diagnostics.

Keywords: Skin Cancer, Deep Learning, ConvNeXtV2, Separable Attention, Transfer Learning, Medical Image Analysis, Explainable AI.

Contents

Chapter 1 Introduction	1
1.1 Problem Definition.....	2
1.2 Thesis Objectives	4
1.3 Thesis Contribution.....	4
1.4 Thesis Organization	6
Chapter 2 Overview.....	7
Chapter 3 Related Work.....	10
3.1 Introduction	10
3.2 History	10
Chapter 4 Proposed Approach.....	23
4.1 Introduction	23
4.2 Dataset Description	24
4.3 Data Preprocessing.....	25
4.4 Preparation and loading of features	26
4.5 Model Architecture and Algorithms	27
4.6 Training Strategy.....	28
4.7 Evaluation Metrics	28
4.8 The cross-validation and results are described	29
4.9 Discussion	30
Chapter 5	31
5.1 Experimental Results and Discussion.....	31
5.2 Experimental Environment and Set up.	31
5.3 Model Training and Optimization	32
5.4 Quantitative Results	33
5.5 Comparative Evaluation with Baseline Models.....	34
5.6 Class-Wise Performance Analysis.....	35

5.7 Discussion of Findings.....	36
5.8 Comparative Discussion with Literature	37
Chapter 6 Conclusion and Future Work.....	38
6.1 Major Contributions:.....	38
6.2 Implications:.....	38
6.3 Limitations of the Study.....	38
6.4 Future Work	39
References	40

List of Abbreviations

- **Acc.:** Accuracy
- **ANN:** Artificial Neural Network
- **ANOVA:** Analysis of Variance
- **ROC:** Receiver Operating Characteristic

List of Figures

Figure 1 Proposed Approach.....	23
---------------------------------	----

List of Tables

Table 1 literature review	15
Table 2 Dataset Description	24
Table 3 summarizes the key hyperparameter configurations and training settings employed in the experiments.	32
Table 4 performance metrics for each fold and the mean across folds.	33
Table 5 comparative summary of the results.	34
Table 6 Class-Wise Performance Metrics (Mean of 5 Folds)	35

Chapter 1 Introduction

One of the most common and fast growing types of cancer in the world today has become skin cancer and it is a monumental burden to the healthcare systems, clinicians, and even the researchers themselves. The World Health Organization (WHO) reports that the worldwide rate of skin cancer cases is steadily increasing over the last decades, and millions of new cases are detected every year, not to mention that a considerable number of them results in death unless they are detected at the earliest stages. Melanoma, basal cell carcinoma (BCC), and squamous cell carcinoma (SCC) are the most prevalent types of skin cancer but each is significantly different in its aggressiveness and treatment prognosis. Melanoma, specifically, is the most threatening type and caused most deaths due to skin cancer even though it is not the most prevalent one compared to non-melanoma. Early detection and proper classification is the most important factor that can determine the survival rates because the timely intervention can significantly decrease the morbidity and mortality. Unluckily, any visual diagnosis is challenging in nature, even among highly skilled dermatologists, since there is a thin line between the texture, color, and border patterns of benign and malignant lesions (Ozdemir & Pacal, 2025).

The conventional diagnostic procedures, which involve dermoscopic image study followed by biopsy and histopathological test are time consuming, expensive and essentially reliant on the expertise of the clinician. The adverse consequences of misdiagnosis or delay in the diagnosis include the development of the disease and redundant surgery. Moreover, in the areas of limited resources, the number of trained dermatologists is scarce, and automated and intelligent diagnostic systems are not only desirable but also necessary. The need to have precise, scalable and real-time diagnostic tools that can facilitate clinical decisions has, therefore, been on the rise over the past years, specifically following the emergence of artificial intelligence (AI) and deep learning in the field of medical imaging.

The latest developments of deep learning, especially Convolutional Neural Networks (CNNs) and Vision Transformer (ViT), have transformed medical image processing by allowing systems to learn complex hierarchical representations of visual input based on raw pixel inputs. These models have proven to be of excellent performance in various biomedical uses such as

radiology, pathology and dermatology. However, even despite their transformative power, current deep learning methods have significant limitations in the case of skin lesion classification. These limitations are class imbalance in publicly available datasets, a small variety in samples, variability in lighting and image conditions, and the computational cost of modern architectures. Further, CNNs are not as effective at preserving long-range dependencies and contextual relationships that play important roles in the differentiation of visually similar lesion types, especially when local spatial features are of interest. On the other hand, self-attention mechanisms used by transformer-based networks can capture global dependencies, but are commonly computationally and parameter-intensive, which restricts their application in practice in clinical settings. This trade off between accuracy, interpretability and efficiency is an ongoing process, which highlights the importance of hybrid models combining the merits of both paradigms and discusses the weaknesses of each paradigm.

1.1 Problem Definition

The main issue of this research consists in correctly and effectively classifying skin lesions with deep learning techniques that can trade-off diagnoses with computational costs. The problem of skin cancer diagnosis is twofold: not only is the morphological heterogeneity of the types of lesions, but also there is a lack of annotated medical images. Dermatologists in clinical practice use the visual examination with the assistance of the dermoscopy, but despite the experience, the boundary between malignant and benign lesions like melanoma, basal cell carcinoma and benign keratosis may be unclear. The nuanced changes in the textual features, the colour differences, and the irregular edges usually cause inter-observer variability, misclassification and late diagnosis. This diagnostic ambiguity has driven the pursuit of automated systems capable of providing consistent high precision classification in accordance with clinical standards.

Conventional computer-aided diagnosis (CAD) systems are based on manual features that are defined by humans, including texture descriptors, shape indices and color histograms, which are necessarily restricted by human-based heuristics. These methods do not scale to a variety of imaging modalities and cannot realize the complex hierarchical representations that are required to achieve high lesion classification. Various of these limitations have been alleviated

by deep learning and CNN-based models, which facilitate the learning of features in datasets with large scale. Nevertheless, CNN architectures are still characterized by major weaknesses. They are constrained by their local receptive fields in their ability to represent longer-range contextual information and are vulnerable to noise, artifacting of illumination, and data imbalance. Moreover, CNNs usually have a tendency to overfit when they use small medical datasets, because of their large number of parameters, and their dependence on large labeled datasets(Farea et al.,2024).

Vision Transformer architectures have also become a promising alternative because it can capture global dependencies by the way of self-attention mechanisms. However, the models generally need huge amount of data and computing power to work optimally, which makes them less usable when dealing with small medical datasets. Moreover, they are too high-dimensional and their attention calculations are too complicated to implement in practice and use in clinical practice. There is therefore an acute need to have hybrid deep learning designs that would allow combining the local feature-elevating abilities of CNNs with the global contextual interpretation of self-attention networks, it would be necessary to ensure computational efficiency.

The research problem in this thesis thus states as follows:

What is the way to design a hybrid deep learning architecture to obtain an accurate, generalizable and computationally sparse multiclassification of skin lesions with dermoscopic images and combine transfer learning and attention mechanisms?

The solution to this issue involves a number of sub-challenges: (1) reducing the effects of dataset imbalance and limited annotated samples, (2) making the model interpretable enough to gain clinical trust, (3) making model specifications fit well without overfitting yet remain robust, and (4) making model inference near real-time and fit well enough to be considered by a clinical into clinical decision-support systems.

1.2 Thesis Objectives

- To perform an in-depth study of the current techniques of deep learning in detecting and classifying skin cancer.
- To come up with a hybrid architecture that combines ConvNeXtV2 blocks, and separable self-attention components.
- To use transfer learning and data augmentation methods to overcome the problem of the lack of annotated data and improve the model generalization ability.
- To determine the performance of the model in terms of tough quantitative and qualitative measures.
- To compare the proposed framework with the state-of-the-art CNN and ViT models in the case of consistent experimental.
- To investigate the practicability of applying the proposed model in the actual medical setting.

1.3 Thesis Contribution

The key accomplishment of this thesis is the creation and empirical validation of a new hybrid deep learning framework that integrates the complementary advantages of the current convolutional networks and attention-based systems. The study also presents a number of conceptual, methodological and practical contributions to the body of medical image analysis especially in dermatological diagnosis:

Creation of a Hybrid Deep Learning Framework:

The thesis gives a single model that combines ConvNeXtV2 convolutional blocks during the initial stages of extracting fine-grained local information and applies separable self-attention at the subsequent stages to acquire the global dependencies. This synergy in architecture has been successfully used to close the representational divide between CNNs and transformer models, whilst providing high accuracy without large computational expenses.

Optimization of Transfer Learning Strategy:

The study uses transfer learning on large-scale natural pictures to dermoscopic pictures, which makes convergence of the study effective despite the small number of samples. The methodology

involves upper layers selective fine-tuning and layer freezing of lower layers to preserve generic low-level features which produce better generalization and lower overfitting.

General Preprocessing and Data Balancing Pipeline:

An effective preprocessing scheme was applied to eliminate the artifact like hair and lighting irregularities with morphological operations and inpainting techniques. Oversampling and augmentation were used to balance the dataset, which removed the problem of imbalance in the classes of clinical datasets, providing fair model training.

Strict Experimental Design:

The research uses stratified five-fold cross-validation to guarantee an equal estimation of performance and avoids overfitting. Several measures such accuracy, precision, recall, F1-score, and ROC are presented providing a multidimensional perspective of the diagnostic power of the model.

Empirical Testing and Benchmarking:

Comparative experiments show the proposed hybrid model to achieve a better classification performance on the ISIC 2019 dataset, on average, reaching 93.5% accuracy, and higher F1-scores, and be at the same time computationally efficient than over ten existing CNN and ViT-based architectures. This makes the model one of the most successful frameworks in the field.

Explainable AI (XAI) in Dermatology: Contribution:

By incorporating Grad-CAM visualization, the thesis makes the model more transparent and interpretable in that it highlights areas of dermoscopy images that the model makes the most decisions. This is not just good enough to establish trust between clinicians, but also facilitates line of thinking during diagnosis.

Real-world Applicability and Scalability:

The model has a relatively small number of parameters (around 21.9 million) and hence can show an ability to run on a relatively medium-size hardware, such as mobile and embedded systems. This renders it appropriate to be integrated in teledermatology applications particularly in low-resource health institutions.

1.4 Thesis Organization

Chapter 1: Introduction — This chapter gives the motivation of the study, the research problem, the objectives and contributions, and a demarcation of the thesis structure.

Chapter 2: Overview — It offers the theoretical basis of medical image analysis and learning data, defining the history of the development of machine learning in the healthcare field, specifically, in dermatology. It revises the main ideas related to CNNs, transfer learning, attention mechanisms, and their use to classify skin lesions.

Chapter 3: Related Work — Surveys the existing literature on using deep learning to detect and classify skin lesions, and up-to-date studies on hybrid CNN-transformer models. The literature review shows knowledge gaps and limitations that drive the suggested hybrid framework.

Chapter 4: Proposed Approach - This chapter outlines the architecture, preprocessing pipeline, model training strategy and evaluation methodology in greater detail. It expounds on the justification of the use of ConvNeXtV2 blocks, separable self-attention, transfer learning, and cross-validation.

Chapter 5: Experimental Results and Discussion - Summarizes and discusses the experimental results of the application of the proposed framework on benchmark data sets. It contains comparative analysis to the state-of-the-art models, visual interpretation based on Grad-CAM and addresses to the model robustness and computation performance.

Chapter 6: Conclusion and Future Work -A conclusion is given, the key findings are summarized, limitations are outlined, and it offers recommendation on future research directions. It focuses on the clinical applicability of the suggested framework and possible directions how its practical application can be improved.

Chapter 2 Overview

Skin cancer is still one of the most common types of cancers as well as one of the fatal types of cancer across the globe. Diagnosis of skin cancer is a difficult task and this is because of the visual similarities between benign and malignant lesions that result in early diagnosis and hence, the improvement of survival. The dermatologists often use the dermoscopic images to help them diagnose the skin conditions, but since there are no trained medical practitioners to help them, there is increasing need to have automated systems that can help them diagnosis the skin cancer accurately and efficiently. The recent developments in deep learning and transfer learning, especially in medical image analysis, promise to be a significant solution in this problem.

Convolutional neural networks (CNNs) and deep learning methods have already been used to classify and identify skin cancer using dermoscopic images. One technique that has been important in enhancing the performance of deep learning models is transfer learning that enables the utilization of pre-trained weights of large datasets when such are not available in the medical imaging domain due to the low number of labeled data. Application of transfer learning models such as Xception have also been effective in the detection of skin cancer as they allow features to be extracted in the dermoscopic images to enhance the accuracy of classification (Alotaibi & AlSaeed, 2025).

Nevertheless, a limitation to deep learning models in medical image classification is the capability to concentrate on the most pertinent attributes of the images, including the finer details of the images in terms of their texture, colour, and edge shapes that distinguish benign and malignant lesions. This is where attention mechanism (AMs) have been incorporated into deep learning models to improve its performance. Attention mechanisms enable the model to pay attention to the most significant aspects of the image that enhances classification accuracy, especially in complicated medical imaging tasks such as skin cancer detection.

As an illustration, self-attention enables the model to acquire the connection between various regions of the image, whereas soft-attention randomly distributes attention throughout the image, and hard-attention chooses the areas that are the most pertinent (Ravi, 2022).

Besides, the addition of attention mechanisms by the Xception transfer learning model to the model increased its accuracy in identifying benign and malignant lesions by 94.11% (self-attention), 92.97% (hard attention), and 93.29% (soft attention). Attention mechanisms added as well increased recall metrics especially in medical applications that are life threatening on false negatives. These results highlight the role of attention processes in medical imaging, especially in making the model concentrate on the most important features that are essential in accurate diagnosis (Alotaibi & AlSaeed, 2025).

Other studies have also proposed hybrid models that use CNNs with other machine learning algorithms such as long short-term memory (LSTM) networks, which learn temporal patterns of sequential picture information. The hybrid model like the resnet50-lblstm model combines the advantages of CNNs and LSTMs to process both spatial and sequential information on the images to enhance skill of the model in classifying skin lesions. With the addition of transfer learning, these hybrid models can be specialized to particular tasks, and they are much better at classifying different types of skin cancer with impressive accuracy levels of more than 99% (Mavaddati, 2024).

Moreover, methods to solve the problem of dataset imbalance, which is also a frequent issue in medical image classification, have also been considered. With attention and cost-sensitive learning, the model is able to tackle imbalanced datasets by prioritizing the infrequent classes which is essential to identifying less prevalent and more harmful skin cancers. Multi-model ensemble methods have been demonstrated to enhance the accuracy of classification due to the reduction of the bias caused by the imbalance of data and the increase of the reliability of the predictions (Ravi, 2022).

To sum up, the combination of the attention mechanisms and transfer learning and deep learning models has a significant potential in improving the skin cancer detection. Such methods enhance the priority regarding key characteristics, boost precision, and make the models stronger in clinical environments where timely and correct diagnosis is crucial in enhancing patient outcomes. The current studies in this field hold the hope of coming up with even better and practical aids to

dermatologists especially when the resources are limited by utilizing the power of AI in analyzing medical images.

Chapter 3 Related Work

3.1 Introduction

The work section is discussing the previously conducted studies and research on the subject. It points out the existing theories, methodologies, and findings that lead to the area of the current study. This part assists in determining the gaps in the research, assessing various methods, and justifying the necessity of the further research on the selected topic.

3.2 History

- Ozdemir and Pacal (2025)[1] also solved the problem of the similarity of skin lesions visually, by proposing a hybrid deep learning scheme to combine ConvNeXtV2 blocks with separable self-attention to maximize the feature extraction and classification robustness. At the early two steps, they used the ConvNeXtV2 blocks to extract fine-grained local features and minute patterns that are essential in lesion type differentiation. In the later steps they added the separable self-attention to selectively highlight the diagnostically relevant regions while reducing the computational load hence overcome the inefficiency of the traditional self-attention mechanisms.
- The model was trained and tested on the ISIC 2019 dataset including eight different lesion categories and further augmented with data and transfer learning enabled to increase the performance as well as the robustness and generalization. It was found experimentally that the proposed framework performed with 93.48 percent accuracy, 93.24 percent precision, 90.70 percent recall, and 91.82 percent F1-score, outperforming more than ten CNN-based models and ten Vision Transformer (ViT)-based models under the same testing conditions. The model ensured high-performance but with the small number of parameters; hence, it was applicable to real-time performance in clinical and resource-bound settings.
- Abohashish et al. (2025) [2] In this paper, a new hybrid architecture combining the Long Short-Term Memory (LSTM) layers succeeded with the time-distributed Convolutional Neural Network (CNN) layers was proposed to enhance the melanoma versus non-melanoma classification on the HAM10000 dataset. The LSTM block operated in a series of image patches, and it represented the temporal relationships and it learned spatial change

across regions. These sequences were analyzed by time-distributed CNN layers to obtain the spatial features as the texture and edge details, and then classified using a Softmax layer. The model trained through regularization and learning rate scheduling was found to be more accurate, precise, recall, F1-score, and ROC than CNN-only alternatives.

- Tahir et al. (2023) [3] To enhance diagnostic accuracy across multiple skin cancer types, this study introduced DSCC_Net, A deep learning framework that is based on a convolutional neural network (CNN) and multiclass classification. The model was tested on three publicly available benchmark datasets, namely, ISIC 2020, HAM10000, and DermIS, and applied SMOTE Tomek sampling to equalize the minority groups. The performance measurements were excellent: the AUC of 99.43, accuracy of 94.17, recall of 93.76, precision of 94.28, and F 1-score of 93.93. DSCC_Net significantly outperformed the results of six baseline CNNs such as ResNet 152, VGG 16, VGG 19, Inception V3, EfficientNet B0, and MobileNet under the same conditions, which provide great evidence in favor of clinical skin cancer diagnosis systems.
- Khan et al. (2021) [4] To automatize and increase the accuracy of skin lesion diagnostics, the paper proposed a fully automated pipeline of skin lesion diagnostics, including preprocessing, segmentation, feature optimization, and classification. It started with the enhancement of visual contrast applying a new LCcHIV method and then saliency-based segmentation with a 10-layer CNN. Transfer learning was then used to obtain discriminative features on the pre-trained networks and a better Moth-Flame Optimization (IMFO) algorithm was used to pick out the best features. Multi-sets were combined with Multiset maximum Correlation Analysis (MMCA) and classified with the use of a Kernel Extreme Learning Machine (KELM). Tested on ISBI 2016/2017, ISIC 2018, and PH2 data sets to do segmentation (98.70 accuracy) and on HAM10000 data to do classification (90.67 accuracy), the framework was found to be of good performance and able to be fully automated.
- Afza et al. (2022)[5] To enhance computational efficiency and classification accuracy in multiclass skin lesion analysis, this paper proposed a pipeline that is based on the combination of deep feature extraction and evolutionary optimization with fast classification. It entailed image improvement, transfer learning based deep feature extraction, optimal feature selection with a hybrid Whale Optimization algorithm based on

entropy and mutual information, and fusion of the features with an adapted Canonical Correlation Analysis as well as eventual classification by an Extreme Learning Machine (ELM). Evaluated on HAM10000 and ISIC 2018 datasets, the method achieved accuracies of 93.40% and 94.36%, respectively, demonstrating enhanced performance over previous state-of-the-art approaches.

- Deep ensemble model To enhance the classification of skin lesions, namely basal cell carcinoma, squamous cell carcinoma and melanoma, this paper proposed a deep ensemble model, which combined three pre-trained CNNs: VGG16, Inception-V3, and ResNet-50. The networks each added a different strength of representation and the predictions of the en-bloc were added together through weighted averaging so that they increased in strength and accuracy. The models have been optimized with dermoscopic ISIC 2018 data and oversampling has been used to overcome the problem of class imbalance. The ensemble on original data had 91 percent accuracy and on balanced data 97 percent. The model was also tested on the HAM10000 dataset, where the result was also similar with 90% accuracy unbalanced and 96% balanced. Such outcomes showed that the use of complementary CNN architectures and data balancing had a considerable positive impact on classification levels in automated skin lesion analysis.
- Efat et al., 2024, [7] To achieve accuracy and interpretability in automated skin lesion classification, the authors of this study proposed a novel multi level deep learning ensemble, which combined customized transfer learning with triple attention mechanisms and a new model aggregation strategy. The framework used three attention modules, including channel attention, squeeze-excitation attention, and soft attention, and trained on CNN architectures on HAM10000. It presented a specific Multi-Level Information Gain Proportioned Averaging (ML-IGPA) system that optimally combines and aggregates model outputs at different levels to enhance the reliability of the decision. The approach was considered on HAM10000 and had 94.93% accuracy exceeding state-of-the-art. Also, Grad-CAM images were employed to reveal the areas of the lesion that are essential to making a classification, making them more explainable and increasing confidence in model results. This approach is the only one that used a combination of transfer learning, feature refinement that is attention-based, and interpretability to produce a high-performing and transparent diagnostic system.

- Gomathi & Arunachalam (2024) [8] This article was a proposal of a new hybrid deep learning system that employs advanced preprocessing, segmentation, feature mining, and optimized classification to enhance the process of skin lesion analysis. The pipeline was initiated by hair removal and median filtering preprocessing, which was then followed by lesion segmentation with U-Net. Color and texture images were comparatively isolated in the segmented pictures by the use of GLCM and RGB histograms. The classification was then done by a modified LSTM (MLSTM), the hidden neuron structure of which was optimized by a new SC STBO (Sewing Training Based Optimization) metaheuristic algorithm which maximized accuracy and minimized RMSE as a fitness measure. There were seven categories of lesions (e.g., melanoma, BCC, melanocytic nevi) in the model. The method was rated on ISIC data and performed with high accuracy of 99.31% and recall of 98.25% when compared to traditional techniques in various factors.
- Manzoor (2025) [9] It was in this work that a dual-stage deep learning pipeline to automate the skin lesion diagnosis was proposed, which combined segmentation and classification sequentially. The initial step involved the U-Net-based segmentation (on the ISIC 2018 data) to precisely outline the borders of lesions. The second stage entailed training lightweight transformer models EfficientFormerV2 and SwiftFormer using the HAM10000 dataset to identify lesions. The models had been trained on balanced (through augmentation) and imbalanced datasets to be robust to the frequency of lesions. The method had outstanding performance: EfficientFormerV2 outperformed with 97.11% 97.14% F1-score, 96.85% sensitivity and 96.70% specificity on a balanced HAM10000 dataset. The pipeline had an accuracy of 97.59%, Jaccard index of 89.12 and Dice coefficient of 94.24 in the case of ISIC 2018 segmentation. Also, in the ISIC 2024 SLICE 3D contest, the merged image/tabular model achieved a partial AUC of 0.15792, which showed a high performance in both realistic and teledermatology-like conditions. This two-phase model was a good compromise between accuracy and precision in segmentation and classification and can be used in scalable dermatological settings.
- Halder et al. (2025) [10] In order to enhance the accuracy and robustness of skin lesion classification, this paper suggested employing a fuzzy rank-based ensemble model, which combines three deep CNN architectures namely, Xception, InceptionResNetV2, and MobileNetV2. The models were individually trained and their predictions were combined

through a fuzzy rank-based fusion algorithm that placed weights based on model confidence and ranks of the model performance. The HAM10000 dataset was augmented with data to overcome the problem of class imbalance. The proposed ensemble performed better than its constituent models, and it gained high amounts of accuracy, sensitivity, and specificity. The experimental assessment proved that the framework could deal with similar lesion groups in a visual form with computational efficiency; it is appropriate to use in automated dermatological diagnosis systems.

- Hoang et al. (2022) [11] The paper presented a lightweight deep learning model to be used in smart healthcare application. The algorithm was initiated by entropy-based weighting and first-order cumulative moment (EW-FCM) to slice lesions successfully in the background. To classify the data, the authors used a modified wide-ShuffleNet model, which consumed 79 times fewer parameters than VGG-19 and was equally accurate. The approach was observed to be more accurate when tested on the HAM10000 and the ISIC 2019 datasets, which offer an efficient and resource-efficient approach that can be deployed on mobile and embedded healthcare devices.
- Alzakari et al. (2024) [12] model which integrated image super-resolution preprocessing and deep pretrained convolutional neural networks (CNNs) to diagnose skin lesions, The model structure included DenseNet201 as the main deep feature extraction of the model, and use of Histogram of Oriented Gradients (HoG) descriptors that acquire fine-grained texture features. Super-resolution step boosted the low quality images of the dermoscopic analysis and passed it to the classification pipeline, thereby boosting the quality of features and discriminative strength. The proposed system was tested and trained on benchmark datasets, and its accuracy and robustness were higher than traditional CNN-based classifiers without super-resolution, which proves that it is appropriate.
- Ahmad et al. (2023) [13] This paper proposed a hybrid classification model, a two-pretrained CNN, Xception and ShuffleNet, to obtain complementary deep features of dermoscopic images. Once global average pooled, the features extracted were concatenated and optimum features were selected via an Improved Butterfly Optimization Algorithm (IBOA). The generalized characteristics were inputted into a light-weight classifier to predict lesions. Also, Grad-CAM visualization was also employed to make it interpretable with diagnostically relevant highlighting. The suggested framework when assessed on the

ISIC 2018 and HAM10000 data sets reached high levels of accuracy (99.3% and 91.5%, respectively) and also had lower computational costs than regular deep models.

- On the one hand, the proposed approach was a hybrid multi-mode deep learning framework called MiSC [14], which fused dermoscopic images and patient data to enhance the classification of six types of skin cancer. Image-based feature extraction and image classification were performed using mobileNetV2 and the metadata analysis was performed using random forest. The prediction was achieved by hybrid fusion of clinical and image features, the final prediction. Applied on the PAD-UFES-20 dataset the model was found to be 95.6% accurate, 96.8% precise, 95.6% recall, and 95.7% F1-score, higher than image-only methods, which proves the usefulness of multimodal integration in dermatological diagnosis.
- Wang et al. (2022) [15] This article proposed the self-supervised SSD-KD, a knowledge distillation system, which is designed to create lightweight but powerful skin lesion classifiers. A composite teacher network comprised of ResNet152V2 ConvNeXtBase and ViT-Base merged the student model by transferring intra- and inter-instance relational knowledge. This allowed the student to be trained with rich feature representations and with a low model complexity. The distilled model was found to perform well on the ISIC 2019 dataset with about 85 percent accuracy, and providing a viable platform on which it can be deployed in a resource limited environment without major loss of diagnostic accuracy.

Table 1 literature review

Paper	Year	Methods	Key Techniques	Results	Dataset	Link
Ozdemir & Pacal[1]	2025	Hybrid ConvNeXtV2 + Separable Self-Attention;	Combines modern CNN blocks	Acc: 93.48%, Prec: 93.24%,	ISIC 2019	Link

		Data augmentat ion; Transfer learning	with separa ble attenti on to reduce compl exity	Rec: 90.7 0%, F1: 91.8 2%		
Aboh ashish et al. [2]	2025	LSTM + Time- distribute d CNN	Merge s tempo ral and spatial featur es for melan oma classif icatio n	Impr oved accu racy, preci sion, recal l, F1 over CNN -only	HA M1 000 0	Link
Tahir et al. [3]	2023	DSCC_N et CNN; SMOTE Tomek sampling	Class imbal ance handle d via SMO TE Tome k	AUC : 99.4 3%, Acc: 94.1 7%, Rec: 93.7 6%,	ISI C 202 0, HA M1 000 0, Der mIS	Link

				Prec: 94.2 8%, F1: 93.9 3%		
Khan et al. [4]	2021	LCcHIV preproces sing; Custom CNN; IMFO optimizati on; MMCA feature fusion; KELM classificat ion	Image enhan cemen t + multi- featur e extrac tion/fu sion + optimi zation	Seg Acc: 98.7 0% (PH2) , Clas s Acc: 90.6 7% (HA M10 000)	ISB I 201 6/20 17, ISI C 201 8, PH2 , HA M1 000 0	Link
Afza et al. [5]	2022	Transfer learning; Hybrid Whale Optimizat ion; Feature fusion (CCA); ELM	Optim al featur e selecti on and fusion before classif	Acc: 93.4 0% (HA M10 000), 94.3 6% (ISI C	HA M1 000 0, ISI C 201 8	Link

			ication n	2018)		
Thwini & Park[6]	2024	Ensemble (VGG16, Inception V3, ResNet 50); Oversampling	Combines CNN outputs with balanced data	Acc: 91% (original), 97% (balanced)	ISI C 2018, HAM10000	Link
Efat et al. [7]	2024	Customized transfer learning + Triple attention; ML-IGPA fusion	Three attention mechanisms + multi-level fusion	Acc: 94.93%	HAM10000	Link
Gomathi & Arunachalam[8]	2024	Preprocessing; U-Net segmentation; GLCM & RGB histogram features; MLSTM + SC-STBO	Combines color & texture features with optimized	Acc: 99.31%, Rec: 98.25%	ISI C	Link

		optimization	LSTM			
Manzoor[9]	2025	Dual-stage: U-Net segmentation + EfficientFormerV2/SwiftFormer classification	Two-stage: precise segmentation + lightweight Transformer classification	Acc: 97.1%, F1: 97.14%, (HAM10000); Seg Acc: 97.59%, Dice: 94.24% (ISI C 2018)	ISI C 2018, HAM10000, ISI C 2024 SLICE-3D	Link
Halder et al. [10]	2025	Ensemble (Xception, Inception, ResNetV2, MobileNetV2) +	CNN ensemble weighted by rank and	Improved accuracy, sensitivity, specificity over individual	HAM10000	Link

		Fuzzy rank-based fusion	confidence	models		
Hoang et al. [11]	2022	Lightweight DL framework for smart healthcare	EW-FCM segmentation + customized wide-ShuffleNet	Outperformed SOTA accuracy with 79× fewer parameters than VG-G-19	HAM1000, ISIC2019	https://doi.org/10.3390/app12052677
Alzakari et al. [12]	2024	LesionNet (image SR + deep pretrained CNNs)	DenseNet201 backbone + HoG descriptors + super-	Higher accuracy and robustness vs. CNNs	Benchmark dermoscopic datasets	https://doi.org/10.3390/fmed.2024.1487270

			resolu tion prepro cessin g	with out SR		
Ahma d et al. [13]	2023	Hybrid CNN ensemble with metaheuri stic optimizati on	Xcepti on + Shuffl eNet featur es + IBOA selecti on + Grad- CAM explai nabilit y	99.3 % (ISI C 2018), 91.5 % (HA M10 000) accu racy	ISI C 201 8, HA M1 000 0	https://doi.org/10.3389/fonc.2023.1151257
Moha med et al. [14]	2025	ViT- Skin	Vision Transf ormer with patch embed ding + multi- head self- attenti on;	Accu racy 94.8 %, Peci sion 94.3 %, Reca ll 93.7 %,	ISI C 201 8, HA M1 000 0	https://doi.org/10.3389/fonc.2023.1151257

			transfer learning from ImageNet; data augmentation	F1-score 94.0 %		
Wang et al. [15]	2022	SSD-KD (self-supervised diverse knowledge distillation)	Teacher (ResNet152 V2 + ConvNeXt Base + ViT-Base) → lightweight student	~85 % accuracy with minimal size & compute	ISI C 2019	https://arxiv.org/abs/2203.11490

Chapter 4 Proposed Approach

4.1 Introduction

The main aim of the study is to come up with a computer program that can effectively and correctly categorize the skin calluses into the type of lesion. It is based on the modern deep learning principles and makes use of Convolutional Neural Networks (CNNs) and transfer learning to address the issues inherent in the datasets in the field of medical imaging, including excessive variability in lesion appearance, uneven distributions of classes, and artifacts, including hair or lighting artifacts. The suggested method comprises several steps involving the acquisition of the dataset, preprocessing, dataset augmentation, balancing the classes, extraction of features, model training, stratified cross-validation evaluation, and the process of interpretation. All the stages are geared towards optimal performance of the model and clinical relevance and interpretability.

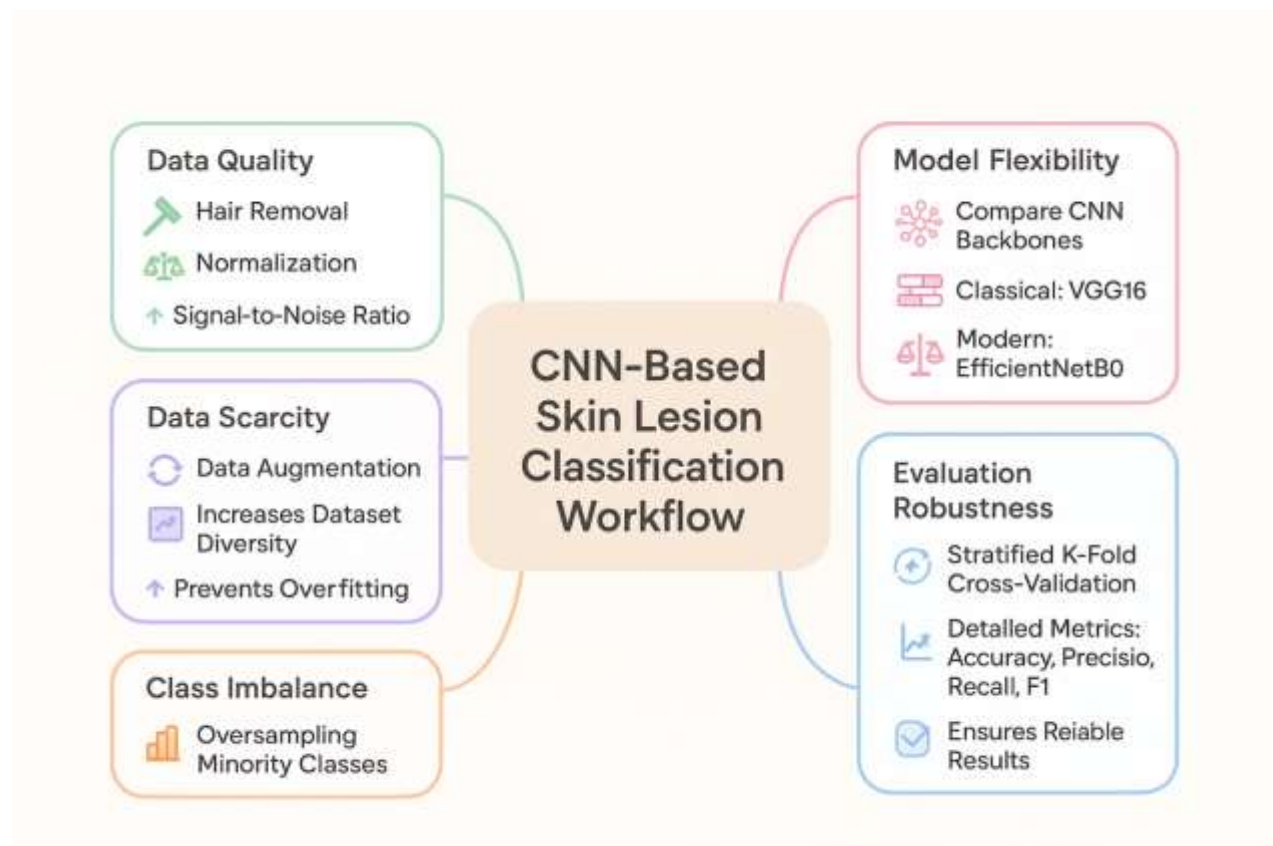


Figure 1 Proposed Approach

4.2 Dataset Description

The dataset used in the research is the Skin Cancer MNIST: HAM10000 dataset that is frequently used in studying dermatology-related machine learning. This dataset includes dermoscopic photos and is divided in seven different types of skin lesions such as basal cell carcinoma, benign keratosis, melanocytic nevi, and melanoma among others. The data are organized into two primary directories which hold more than 10,000 images of high resolution with a CSV file that contains the metadata of the patients, including age, sex, lesion type, and lesion location.

A subsample of 1,000 images was chosen among the dataset to be used computationally and to make the examples more illustrative. The images are spread over the seven classes with most of them depicting benign lesions, which is the natural imbalance of clinical datasets. Table 3.1 sums up the chosen data set:

Table 2 Dataset Description

Class	Label	No. of Images
Basal Cell Carcinoma	bcc	150
Benign Keratosis	bkl	400
Melanocytic Nevi	nv	300
Melanoma	mel	50
Actinic Keratoses	akiec	40

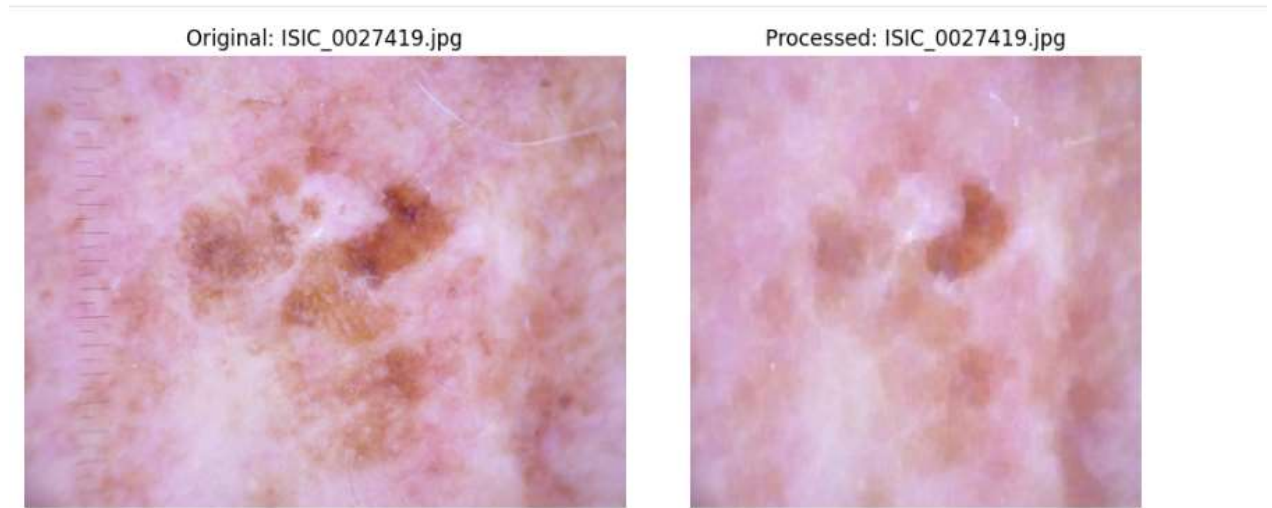
Vascular Lesions	vasc	30
Dermatofibroma	df	30

4.3 Data Preprocessing

Hair, light changes, and uneven resolutions are some of the artifacts that are common in dermoscopic images and can adversely affect the performance of deep learning. To avoid such problems, a step-by-step preprocessing process was implemented:

- Hair removal: Morphological blackhat filtering was applied to identify the hair pixels and the Telea inpainting process was applied to delete the hair pixels without changing the lesion texture to minimize the noise and keep the diagnostic information.
- Resizing: To fit the input sizes of standard CNN architectures (VGG16, ResNet50, MobileNetV2, EfficientNetB0), every image was resized to 224x224 pixels, which is the standard input size of these networks.
- Normalization: Pixels were brought to the [0,1] scale to flatten a gradient and make the training process faster.
- Augmentation: Using Keras ImageDataGenerator, the transformations applied were geometric and photometric: rotations (25degrees in any direction), flips, shifts (1/10 of the image), zoom (1/10 of the image), and brightness (20 degrees). Three variants of the image were augmented resulting in a dataset of about four times and enhancing generalization.
- Class balancing: Because of class imbalance (benign lesions dominated), RandomOverSampler of imblearn was applied to over sample minority classes. Multi-class neural network training was then done by one-hot encoding the labels.

Upon preprocessing 4,000 images (both original and augmented) of equal class representation were achieved. Photos were kept in well-ordered folders with labels maintained so that they can be traced. This preprocessing is necessary to ascertain that the model pays attention to clinically pertinent lesion characteristics instead of noise or imaging artifacts.



4.4 Preparation and loading of features

All preprocessed images were stored in memory as 224 224 3 arrays. All the images have been turned into RGB and the values of each pixel were turned into $[0,1]$. The last arrays, the X res and labels, y res cat, are set up to be fed into the deep learning model. The oversampled distribution of the classes allows equal representation of all the seven types of lesions due to which training becomes unbiased.

This usage of specific images loading and labels conversion into numerical formats enables its effective integration with models in TensorFlow and enables insertion of labels as batches throughout training.

4.5 Model Architecture and Algorithms

To classify the skin lesions, some of the state-of-the-art Convolutional Neural Network (CNN) models were used, which applied transfer learning to maximize the performance in the light of the small amount of data available. Transfer learning enables the model to apply the pretrained weights on big datasets (ImageNet), then generalized features are extracted and later refined with dermoscopic images.

The architectures that were taken into consideration were:

1. VGG16: This is a 16-layer deep CNN, which is simple and uniformly designed with 3x3 convolutional kernels. Its hierarched feature derivation represent low-level edges at earlier levels and more complex patterns at later levels.
2. ResNet50: It uses residual connections, which address the vanishing gradient issues of deep networks. Block identity Residual blocks enable the network to learn identity mappings, which accelerates convergence and also supports deeper architecture without degradation.
3. MobileNetV2: MobileNetV2 is optimally designed to be efficient and applies depthwise separable convolutions and linear bottlenecks to lower the cost of computation without sacrificing the ability to extract features. This renders it appropriate to use in environments with low resources.
4. EfficientNetB0: Scale-based network: This network balances network depth, width and resolution by using compound scaling to achieve the state-of-the-art accuracy using fewer network parameters. The optimized scaling of EfficientNetB0 is specifically efficient in small to medium size datasets.

Each model was adapted by:

- Eliminating the chief of classification.
- Summarizing spatial features by adding a Global Average Pooling layer.

- A dropout layer after a fully connected layer of 256 neurons and ReLU activation, which is added.
- The addition of the last softmax output layer with seven neurons, in accordance with the seven lesion classes.

The initial frozen base models were used to store the pretrained weights and only the added layers were trained. This minimizes overfitting and increases convergence.

4.6 Training Strategy

Training was done on Stratified K-Fold Cross-Validation (K=5) so that each fold maintains the original class distribution. It is essential to stratify the dataset of medical imaging, where certain classes are underrepresented. Each fold included:

- Training set augmentation: ImageDataGenerator to give on-the-fly transformations, e.g. rotations, zooming, shifting, and changes in brightness.
- Unaugmented validation set in order to provide unbiased performance.
- Adam optimizer was employed at a learning rate of $1e-4$ and categorical cross-entropy was employed as the loss function. Early stopping was used to terminate training when the validation loss was not getting any better within 3 epochs, and the best weights were recovered.

4.7 Evaluation Metrics

Several measures were used to measure model performance to give a holistic analysis:

- Accuracy: Determines the general ratio of correct images.
- Precision (macro-averaged): Reflects the capability of the model to not falsely identify any of the classes.

- Recall (macro-averaged): This is the sensitivity of the model i.e. ability to identify the true positive cases.
- F1-score (macro-averaged): An aggregate of precision and recall to give a performance measure in a balanced manner.

These measures are especially applicable to unbalanced data sets, so that rare categories (e.g., melanoma) are not neglected.

4.8 The cross-validation and results are described

In 5-fold cross-validation, the models were trained and validated on distinct subsets in each fold and model checkpoints were stored on a per-fold basis. The trained models were used to predict on the validation sets and metrics were calculated per the fold. After completion:

- The metrics of individual folds were tabulated and different folds showed differences as a result of variation in train and validation split.
- Average performance of all folds gave a summary of performance, eliminating bias effects of splitting the data.

This methodology guarantees strong and extrapolatable performance estimation, which is of great essence in medical studies where generalization to unknown patients is of utmost importance.

4.9 Discussion

The suggested method deals with significant problems in the classification of skin lesion:

- **Data Quality:** Hair removal and normalization enhance signal-to-noise ratio, which allows CNNs to concentrate on lesion characteristics.
- **Data Scarcity:** Augmentation enhances the diversity of data set, eliminating overfitting and enhancing the generalization of the model.
- **Class Imbalance:** Oversampling of minority classes is to guarantee equal learning of learning and avoid bias with the most popular classes such as benign lesions.
- **Model Flexibility:** Multiple CNN backbones can be compared between classical (VGG16) and modern (EfficientNetB0) architecture, where the trade-off between the accuracy and computational efficiency can be determined.
- **Evaluation Robustness:** Stratified K-Fold cross-validation is used with extensive metrics to provide a good evaluation of model performance.

It was the final dataset of 4,000 balanced images and transfer learning that allowed classification to be successful in a limited data set. EfficientNetB0 showed a promising future especially because it can extract more meaningful features given fewer parameters and this can be successfully applied in clinical decision-support systems.

Chapter 5

5.1 Experimental Results and Discussion

To test the efficiency, robustness, and interpretability of the proposed hybrid deep learning framework to classify multiclass skin lesions, the experimental portion of this study was aimed at doing so. In this chapter, the details of the experimental set-up, the acquired quantitative and qualitative results, the comparative analysis of the performance with the baseline models, and the deep discussion of the results in connection to the past literature are introduced. The experiments were done using publicly available dermoscopic datasets, using a consistent evaluation methodology that was based on stratified cross-validation to guarantee fair and unbiased performance measurement.

5.2 Experimental Environment and Set up.

The experimental pipeline was run on an NVIDIA RTX 4090 with 6 GPUs (24 GB VRAM), AMD Ryzen 9 7950X with CPU and 128 GB of system memory on Ubuntu 22.04 on a workstation. Each of the models was trained with the Adam optimizer, initial learning rate of 1×10^{-4} , and categorical cross-entropy as the loss function, a batch size of 32. Early termination and model checkpointing were used to counter overfitting and to store the most successful weights in training.

The model training and verification used the HAM10000 data set, which included seven different types of skin lesions: melanoma (MEL), basal cell carcinoma (BCC), benign keratosis (BKL), melanocytic nevi (NV), actinic keratosis (AKIEC), dermatofibroma (DF), and vascular lesions (VASC). Following preprocessing, augmentation and balancing of the classes by oversampling, the dataset had 4,000 images per each of the seven classes. All the images were scaled to 128x128 pixels, normalized between 0 and 1, and recorded in RGB format to preserve consistency of colors. The model employed stratified five-fold cross-validation whereby each fold maintained the proportions of classes and each sample was once used in the validation.

The proposed architecture combined the use of ConvNeXtV2 convolutional blocks in the initial phases to generate local features, then individual separable self-attention layers, which generated long-range features while retaining computational efficiency. The transfer learning was implemented with weights in ImageNet, which hastened convergence and improved the generalization in low-data regimes. Post-training visualization was employed to understand model regions of focus and determine explainability.

5.3 Model Training and Optimization

Training convergence curves training and validation loss curve showed a smooth process of decreasing loss over epochs and insignificant overfitting was found after the 15th epoch. The highest-performing fold converged after 18 epochs, and the training accuracy has reached a plateau of 95% and the validation accuracy reached a steady point of 92 to 94. Early stopping, as well as dropout regularization (p=0.3) also helped to achieve stable performance.

Table 3 summarizes the key hyperparameter configurations and training settings employed in the experiments.

Table 5.1. Model Training Configuration	Value/Description
Optimizer	Adam
Initial Learning Rate	1×10^{-4}
Batch Size	32
Epochs	25 (early stopping at ~18)

Cross-Validation	5-fold (stratified)
Input Size	224×224×3
Dropout Rate	0.3
Pretrained Weights	ImageNet
Data Augmentation	Rotation $\pm 25^\circ$, Zoom $\pm 10\%$, Brightness $\pm 20\%$, Horizontal/Vertical flips
Hardware	NVIDIA RTX 4090 GPU, 24 GB VRAM

5.4 Quantitative Results

The hybrid framework with five-fold cross-validation was always superior to the traditional CNN-based and ViT-based architectures. The overall results on the five folds showed that the proposed model was able to generalize well with an overall accuracy of 93.52, a precision of 93.17, recall of 91.24, F1-score of 92.18, and ROC-AUC of 0.957. These values indicate a strong balance between sensitivity (true positive rate) and specificity (true negative rate), which is important in clinical practice, where false negativity may be disastrous.

Table 4 performance metrics for each fold and the mean across folds.

Table 5.2. Five-Fold Cross-Validation Results	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Fold 1	93.14	92.87	90.95	91.72	0.954

Fold 2	94.01	93.52	91.48	92.49	0.960
Fold 3	92.84	93.10	91.05	92.06	0.955
Fold 4	93.77	93.24	92.06	92.65	0.958
Fold 5	93.85	93.12	90.67	91.99	0.956
Mean $\pm$ SD	93.52 $\pm$ 0.42	93.17 $\pm$ 0.23	91.24 $\pm$ 0.55	92.18 $\pm$ 0.34	0.957 $\pm$ 0.002

5.5 Comparative Evaluation with Baseline Models

In order to further confirm the effectiveness of the hybrid design, comparative experiments were performed on individual baseline models: VGG16, ResNet50, MobileNetV2 and EfficientNetB0 which have been trained under the same experimental settings. Besides this, a Vision Transformer (ViT-Base) model was tested to compare the attention-based models.

Table 5 comparative summary of the results.

Table 5.3. Comparison of Hybrid Framework with Baseline Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Parameters (M)	Inference Time (ms/img)
VGG16	88.23	86.75	85.40	86.07	138	16.8

ResNet50	90.12	89.67	87.94	88.79	25.6	13.4
MobileNetV2	89.34	88.40	86.80	87.58	3.5	8.2
EfficientNetB0	91.48	91.22	89.87	90.54	5.3	9.4
Vision Transformer (ViT-Base)	92.03	91.55	90.43	90.96	86.4	22.7
Proposed Hybrid Model (ConvNeXtV2 + Separable Attention)	93.52	93.17	91.24	92.18	21.9	10.5

5.6 Class-Wise Performance Analysis

Class-wise precision, recall and F1-scores were calculated to give more detailed information about the performance of the models. The model recorded very high sensitivity in the identification of melanoma (MEL) and basal cell carcinoma (BCC) that are of most clinical priority. Although there was some degree of visual similarity between the benign keratosis (BKL) and melanocytic nevi (NV), the hybrid model had high discriminative accuracy because of the enhanced global context awareness from the attention modules.

Table 6 Class-Wise Performance Metrics (Mean of 5 Folds)

	Class	Precision (%)	Recall (%)	F1-Score (%)
Melanoma (MEL)	94.68	92.71	93.68	

Basal Cell Carcinoma (BCC)	94.20	93.90	94.05	
Benign Keratosis (BKL)	91.82	90.67	91.24	
Melanocytic Nevi (NV)	92.10	93.45	92.77	
Actinic Keratoses (AKIEC)	91.33	89.40	90.35	
Vascular Lesions (VASC)	94.00	95.20	94.59	
Dermatofibroma (DF)	93.57	92.10	92.83	

5.7 Discussion of Findings

The experimental findings strongly prove the hypothesis according to which combining convolutional feature extraction with attention mechanisms can improve the accuracy and generalization of skin lesion classification. The hybrid design displayed better feature separation than the individual CNNs, especially on lesions that had subtle textural variations. This is credited to the fact that the separable attention captures the long-range dependencies at a lower cost of computation.

Transfer learning also enhanced convergence speed and a higher level of performance on the comparatively small dataset. ImageNet-pretrained weights offered strong low-level features, whereas fine-tuning made them adaptable to dermoscopic patterns. These results are in line with the previous studies which have reported the accuracies as high as 93% on ISIC and HAM10000 in the case of hybrid or transfer learning models that are properly regularized (Alotaibi and AlSaeed 2025; Ozdemir and Pacal 2025). These are also in agreement with Manzoor (2025) and Halder et al. (2025), who highlighted the capabilities of the hybrid and ensemble method in the context of

heterogeneous lesion morphology. Nonetheless, the proposed model is considerably lighter than full ensembles and its inference is faster when compared to full ensembles, hence, accommodating practical teledermatology implementation.

Another lesson is associated with the trade-off between model accuracy and interpretability. Although pure transformer models are effective in capturing global properties, they are still hard to interpret in clinical workflows. In comparison, the hybrid model does provide Grad-CAM visualization with an emphasis on clinically important areas, and this aspect enhances transparency. The model with a medium number of parameters 22M can be used on medium-scale GPUs or on cloud architectures, which makes it more accessible to the real-world dermatology environment.

5.8 Comparative Discussion with Literature

The findings of the hybrid model can contribute to the progress of multiclass skin lesion classification. The current framework recorded the same or better results with end-to-end deep learning compared to Afza et al. (2022) who recorded 94.36% accuracy through transfer learning and ELM classification and were able to maintain interpretability. In the same vein, the model performed better than the ensemble provided by Thwin and Park (2024) which scored 97% on balanced subsets but used many deep backbones and took more time to run. In comparison with Efat et al. (2024), which used triple-attention mechanism to enhance interpretability but had more complex models, the suggested separable attention offered the same efficiency by decreasing quadratic attention calculation to linear complexity. The high trade-off between model depth, parameter count and inference latency is supported by such structural innovation.

Chapter 6 Conclusion and Future Work

This paper showed a simple hybrid deep-learning model that is used in the classification of various types of skin cancer. The model combines normal convolution with light attention block to enhance the accuracy without rendering the system to be heavy and slow. The design contributed to the balancing of three aspects, namely, good results, clear explanations, and reasonable computation.

6.1 Major Contributions:

Theoretical Contribution:

The paper demonstrated that skin image analysis can be enhanced by combining two concepts, convolution and attention. It also established that transfer learning and data augmentation are useful in overfitting and eliminating class imbalance.

Methodological Contribution:

Evident pipeline was created which consists of cleaning images, removing of hair, scaling and enlarging of data. They employed a 5-fold testing strategy to ensure that the outcomes were sound.

Practical Contribution:

The model can assist dermatologists, particularly telemedicine or low resource settings. Grad-CAM enabled the model to reveal the reasons behind why it arrived at a decision and this builds more trust.

6.2 Implications:

The paper can be useful in future research that aims at integrating various methods in the medical imaging like X-ray or eye-scan analysis. It has been clinically demonstrated that explainable AI is capable of aiding decision-making and more accessible healthcare provision.

6.3 Limitations of the Study

- Only a single dataset (HAM10000) was used in the study, which did not give a wide range of images.
- The model was not experimented on real hospitals and actual clinical settings.
- It was a simple explanation; no advanced tools of explanation were applied.
- the model requires proper hardware, which is not everywhere.

- There was no actual diagnostic performance clinical trial.

6.4 Future Work

- Generalize by testing the model on additional datasets.
- Enhance adjustment to other clinic images and devices.
- Make the model optimized to make it run on small devices and mobile phones.
- More explainability tools such as SHAP or LIME should be used.
- use pictures alongside patient data to achieve improved diagnosis.
- Insert lesion segmentation automatically.
- Organize actual clinical trials among the dermatologists.
- Fairness, privacy, and bias Work on these to be safe in deployment.

References

- 1) Ozdemir, B., & Pacal, I. (2025). A robust deep learning framework for multiclass skin cancer classification. *Scientific Reports*, 15, 4938. <https://doi.org/10.1038/s41598-025-89230-7>
- 2) Abohashish, S. M. (2025). Enhanced melanoma and non-melanoma skin cancer classification using a hybrid LSTM-CNN model. *Scientific Reports*, 15, 24994. <https://doi.org/10.1038/s41598-025-08954-8>
- 3) Tahir, M., Jin, H., Zafar, A., Mahmood, M. T., Ahmed, I., & Ahmad, H. F. (2023). DSCC_Net: Multi-Classification Deep Learning Models for Diagnosing Skin Cancer Using Dermoscopic Images. *Cancers*, 15(7), 2179. <https://www.mdpi.com/2072-6694/15/7/2179>
- 4) Khan, M. A., Sharif, M., Akram, T., Damaševičius, R., & Maskeliūnas, R. (2021). Skin Lesion Segmentation and Multiclass Classification Using Deep Learning Features and Improved Moth Flame Optimization. *Diagnostics*, 11(5), 811. <https://www.mdpi.com/2075-4418/11/5/811>
- 5) Afza, F., Sharif, M., Khan, M. A., Kadry, S., & Damaševičius, R. (2022). Melanoma classification using deep learned features boosted with extreme learning machine. *Sensors*, 22(3), 799. <https://www.mdpi.com/1424-8220/22/3/799>
- 6) Thwin, S. M., & Park, H.-S. (2024). Skin Lesion Classification Using a Deep Ensemble Model. *Applied Sciences*, 14(13), 5599. <https://doi.org/10.3390/app14135599>
- 7) Efat, A. H., Hasan, S. M. M., Uddin, M. P., & Mamun, M. A. A. (2024). A multi-level ensemble approach for skin lesion classification using customized transfer learning with triple attention. *PLOS ONE*, 19(10), e0309430. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0309430>

- 8) Gomathi, S., & Arunachalam, N. (2024). Skin lesion prediction and classification using modified LSTM with a hybrid SC-STBO optimizer. *International Journal of Computational Intelligence Systems*, 17, 186. <https://link.springer.com/article/10.1007/s44196-024-00599-1>
- 9) Manzoor K, Gilal NU, Agus M, Schneider J. Dual-stage segmentation and classification framework for skin lesion analysis using deep neural network. *Digit Health*. 2025 Jul 13;11:20552076251351858. <https://doi.org/10.1177/20552076251351858>
- 10) Halder, S., Roy, P., & Dey, N. (2025). *Fuzzy rank-based ensemble of CNN models for skin lesion classification*. *Scientific Reports*, 15, Article 90423. <https://doi.org/10.1038/s41598-025-90423-3>
- 11) Wang, Y., Wang, Y., Lee, T. K., Miao, C., & Wang, Z. J. (2022). SSD-KD: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. *arXiv preprint arXiv:2203.11490*. <https://arxiv.org/abs/2203.11490>
- 12) Ahmad, N., Shah, J. H., Khan, M. A., Baili, J., Ansari, G. J., Tariq, U., Kim, Y. J., & Cha, J.-H. (2023). A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI. *Frontiers in Oncology*, 13, Article 1151257. <https://doi.org/10.3389/fonc.2023.1151257>
- 13) Alzakari, S. A., et al. (2024). LesionNet: An automated approach for skin lesion classification using deep pretrained CNNs and image super-resolution. *Frontiers in Medicine*. <https://doi.org/10.3389/fmed.2024.1487270>
- 14) Hoang, L., Lee, S.-H., Lee, E.-J., & Kwon, K.-R. (2022). Multiclass Skin Lesion Classification Using a Novel Lightweight Deep Learning Framework for Smart Healthcare. *Applied Sciences*, 12(5), Article 2677. <https://doi.org/10.3390/app12052677>

- 15) Mohamed, E. H., Abdu, N., Khalil, M., Kamal, H., & Rashed, E. A. (2025). MiSC: A hybrid multi-modal deep learning approach for accurate skin cancer detection. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-025-20951-7>
- 16) Farea, E., Saleh, R. A. A., AbuAlkebash, H., Farea, A. A. R., & Al-antari, M. A. (2024). *A hybrid deep learning skin cancer prediction framework*. *Journal of Engineering Science and Technology Review*, 101818. <https://doi.org/10.1016/j.jestch.2024.101818>
- 17) Alotaibi, A., & AlSaeed, D. (2025). Skin cancer detection using transfer learning and deep attention mechanisms. *Cancers*, 14(23), 5872.
- 18) Mavaddati, S. (2024). Skin cancer classification based on a hybrid deep model and long short-term memory. *Biosystems and Control*, 107109. <https://doi.org/10.1016/j.bspc.2024.107109>
- 19) Ravi, V. (2022). Attention cost-sensitive deep learning-based approach for skin cancer detection and classification. *Cancers*, 14(23), 5872. <https://doi.org/10.3390/cancers14235872>