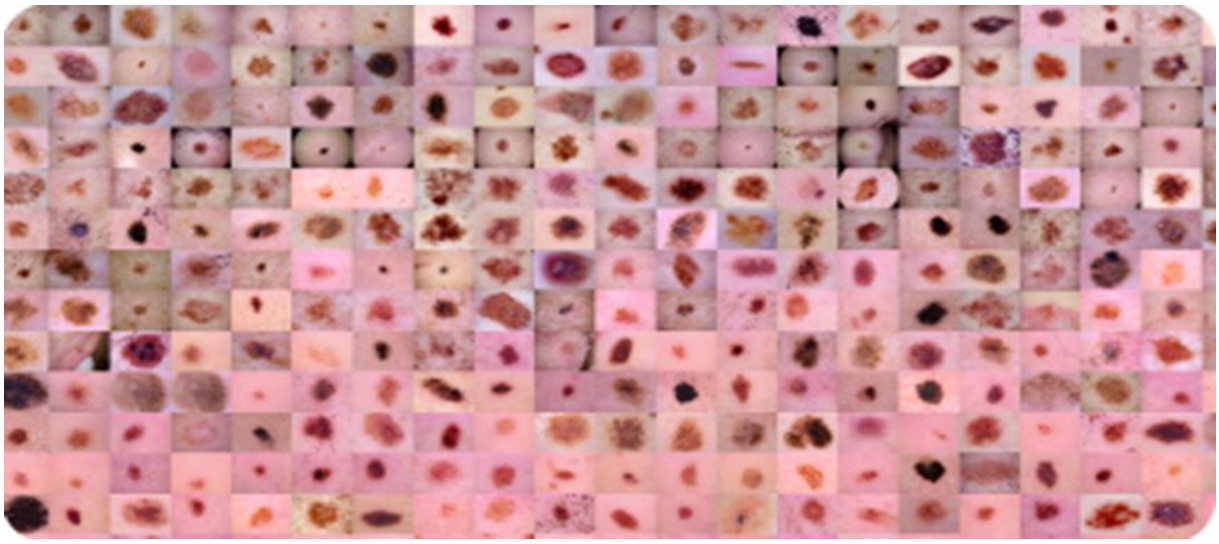
Classification of Skin Diseases with Ontology-Based Validation

DS50 Final Report



Professor:
Serge Iovleff
Tutors :
Fatima Ez Zahra BENKIRANE
Mohamed KAS

Team:
Tolgahan GÜRCÜOGLU
Diego CICERI
Caleb Mario BAVING
Wenxuan GUI
Haoqi TAN

Abstract

This project combines convolutional neural networks with ontology-based semantic reasoning to classify skin lesions from dermoscopic images. Using the HAM10000 dataset and OWL ontologies, we achieve both strong performance and interpretability, offering potential support for clinical decision-making.

1. Introduction

1.1 Context and Objectives

Skin disease diagnosis is a critical task in dermatology. Given the visual similarity between benign and malignant skin lesions, even experienced professionals can face difficulty during diagnosis. This project addresses the need for accurate and explainable automated classification systems for dermoscopic images.

The main objectives of the project were:

- To classify dermoscopic skin images using machine learning and deep learning models.
- To validate the predictions through ontology-based semantic reasoning, providing explainable and trustworthy outputs for medical applications.

1.2 Dataset

We used the HAM10000 dataset, which contains 10,015 dermoscopic images across seven diagnostic categories:

- akiec (Actinic keratoses)
- bcc (Basal cell carcinoma)
- bkl (Benign keratosis-like lesions)
- df (Dermatofibroma)
- mel (Melanoma)
- nv (Melanocytic nevi)
- vasc (Vascular lesions)

2. Methods and Implementation

2.1 Data Preprocessing

- Images were resized to **224×224** pixels to match CNN input requirements.
- **Normalization** was applied based on ImageNet mean and std.
- **Data augmentation** (random rotations, flips, color jitter) was used to reduce overfitting and improve generalization.

2.2 Modeling

- **Baseline (Traditional ML) :**
  - Principal Component Analysis (PCA) for dimensionality reduction.
  - Random Forest (RF) for classification.
- **Deep Learning Models :**
  - Pretrained **ResNet50** and **MobileNetV2** architectures.
  - Both ResNet50 and MobileNetV2 models were initialized with pretrained weights from ImageNet (pretrained=True) to leverage transfer learning and reduce training time on the medical dataset.
  - A classifier head was added, and the training was done in two phases:
    - Freeze base and train classifier.
    - Unfreeze layers and fine-tune with a lower learning rate.
  - Training used **AdamW** optimizer and **ReduceLROnPlateau** scheduler.
  - Models were evaluated using accuracy, F1 scores, and confusion matrix.

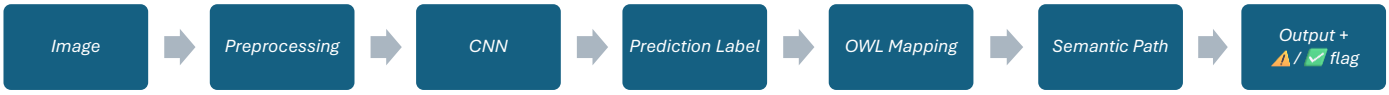
2.3 Ontology Integration

- **OWLready2** was used to load and query a custom dermatology ontology.
- Each predicted label (e.g., mel) was mapped to its corresponding OWL class (e.g., DERMO_0000970).
- The .ancestors() function was used to extract hierarchical semantic paths.
- Malignancy was flagged using pre-defined malignant paths (e.g., Melanoma and BCC branches).
- An explanation string was generated (e.g., “Melanoma → Non-Melanoma Skin Cancer → Skin Cancer”).

2.4 Development Tools

Tool	Purpose
PyTorch	Deep learning framework
scikit-learn	Traditional ML + evaluation
OWLready2	Ontology management and queries
matplotlib	Visualization of metrics

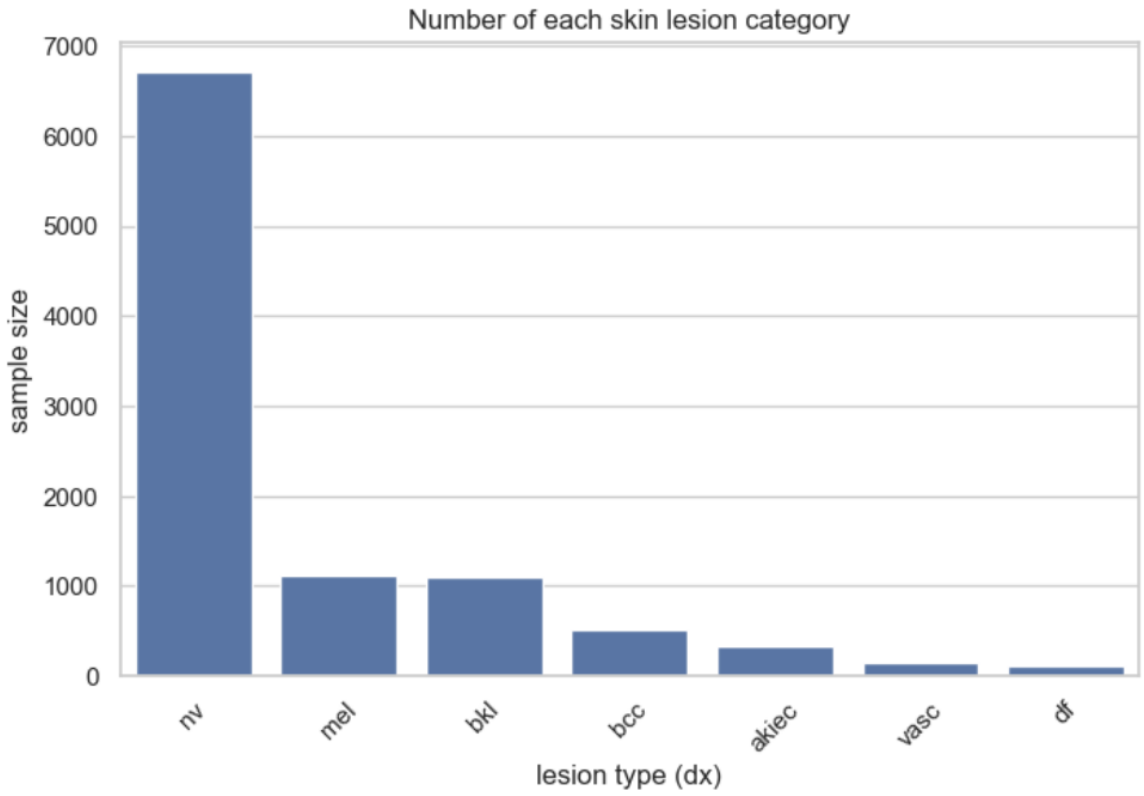
Figure 1: End-to-end pipeline for skin lesion classification and semantic reasoning



3. Difficulties encountered and solutions

3.1 Class Imbalance

- **Problem**: The dataset was imbalanced, with classes like “df” and “vasc” severely underrepresented.



- **Solution** :
  - Used **class_weight='balanced'** in classical models.

```
classes = np.unique(targets)
class_weights_np = compute_class_weight(
    class_weight='balanced',
    classes=classes,
    y=targets
)
```

  - Used **f1 score** as an evaluation method.

```
# Upload F1 score for each class to wandb
f1_logs = {f"f1_score/{label}": report[label]["f1-score"] for label in class_names}

# Also log macro and weighted averages
f1_logs["f1_score/macro_avg"] = report["macro avg"]["f1-score"]
f1_logs["f1_score/weighted_avg"] = report["weighted avg"]["f1-score"]
```

3.2 CNN Training Instability

- **Problem**: ResNet50 would not converge when training all layers at once.

- **Solution** :
    - First trained only the classifier head.
    - Later unfroze and fine-tuned the backbone.

```
for p in model.features.parameters():
    p.requires_grad = False
for p in model.classifier.parameters():
    p.requires_grad = True
```

    - Used **ReduceLROnPlateau** to handle learning rate decay based on validation loss.

```
scheduler = ReduceLROnPlateau(optimizer, mode='min', factor=0.5, patience=3)
```

3.3 Label-Ontology Mismatch

- **Problem**: Predicted labels (e.g., 'bkl') didn't match OWL class names.
- **Solution** :
    - Constructed a custom **label-to-IRI mapping**.

```
label_to_class_name = {
    'nv':    "DERMO_0000337",
    'bkl':   "DERMO_0000108",
    'df':    "DERMO_0000771",
    'mel':   "DERMO_0000970",
    'vasc':  "DERMO_0001950",
    'bcc':   "DERMO_0000395",
    'akiec':"DERMO_0000395"
}
```

    - Used search_one(iri=...) to retrieve the correct OWL class.

```
for label, cls_name in label_to_class_name.items():
    owl_class = onto.search_one(iri=f"*{cls_name}")
    label_to_owl_class[label] = owl_class
```

    - Mappings were verified manually to ensure semantic accuracy.

```
for label, owl_class in label_to_owl_class.items():
    print(f"{label}: {owl_class}")
```

3.4 Semantic Ambiguity in OWL Paths

- **Problem**: Some ontology paths were long or difficult to interpret.

- **Solution** :
  - Defined **malignant label sets** and searched for them in ancestor chains.

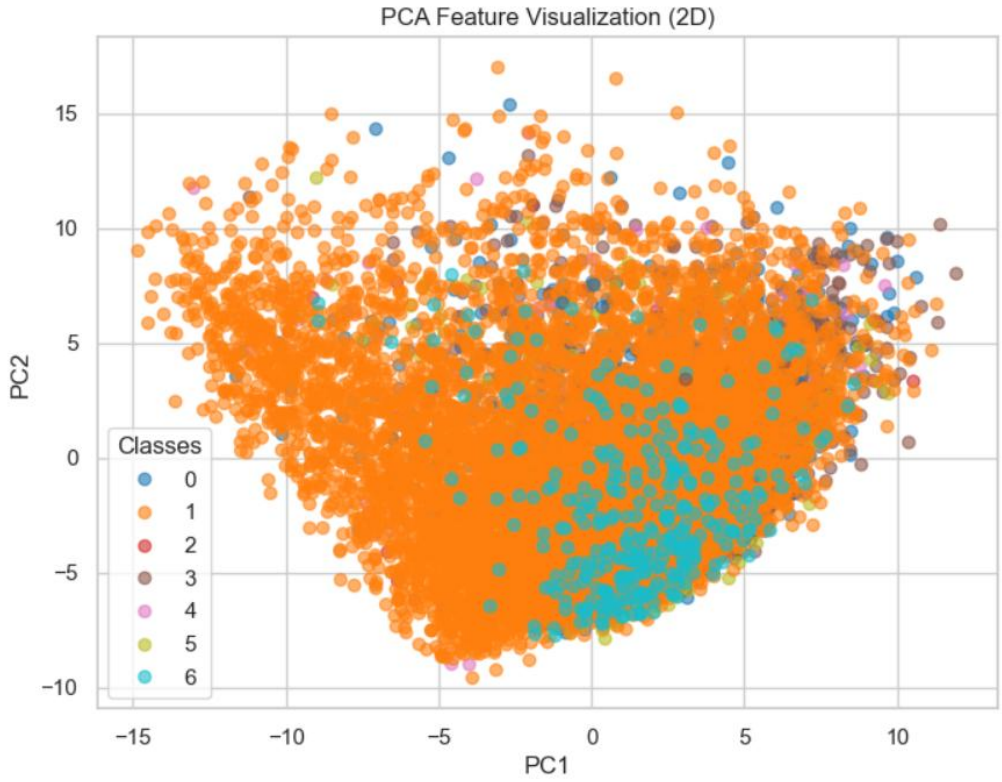
```
mel_class = label_to_owl_class['mel']
nmsc_class = label_to_owl_class['bcc']

def is_malignant_owl(label):
    cls = label_to_owl_class[label]
    return (mel_class in cls.ancestors()) or (nmsc_class in cls.ancestors())
```

  - Rendered the semantic path into user-friendly format.

	Label	Ontology Path	Malignancy	Recommendation	Sample Index	Validation	Age	Sex	Location
0	nv	Thing → disease → cutaneous disease → pigmenta...	✅ Benign	No immediate concern	0	✅ OK	80.0	male	scalp
1	nv	Thing → disease → cutaneous disease → pigmenta...	✅ Benign	No immediate concern	1	✅ OK	80.0	male	scalp
2	bcc	Thing → disease → cutaneous disease → disorder...	⚠️ Malignant	Seek medical attention	2	❌ Location: scalp	80.0	male	scalp
3	vasc	Thing → disease → cutaneous disease → vascular...	✅ Benign	No immediate concern	3	❌ Not for elderly / ❌ Location: scalp	80.0	male	scalp
4	mel	Thing → disease → cutaneous disease → disorder...	⚠️ Malignant	Seek medical attention	4	✅ OK	75.0	male	ear
5	nv	Thing → disease → cutaneous disease → pigmenta...	✅ Benign	No immediate concern	5	✅ OK	75.0	male	ear

4. Results and Analysis

4.1 Feature Space Visualization



To better understand the model's input space, we applied **PCA** to project the features into 2D.

As shown in the figure, class **1 (nv)** dominates the feature space and heavily overlaps with other classes, especially class 0 (bkl) and class 3 (mel).

Minority classes such as *df* (2), *vasc* (4), and *akiec* (6) are sparsely scattered and show no clear separation.

4.2 Model Performance Metrics

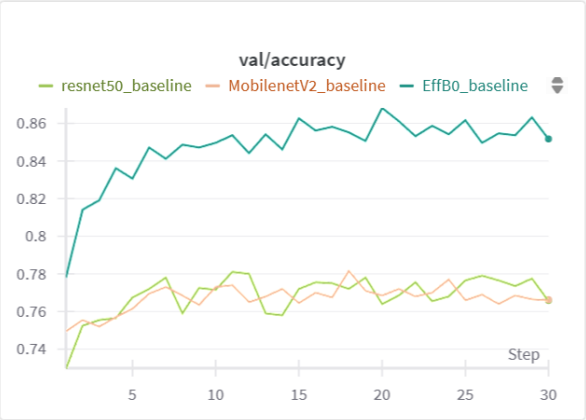
1. **Accuracy of ML model – Random Forest**

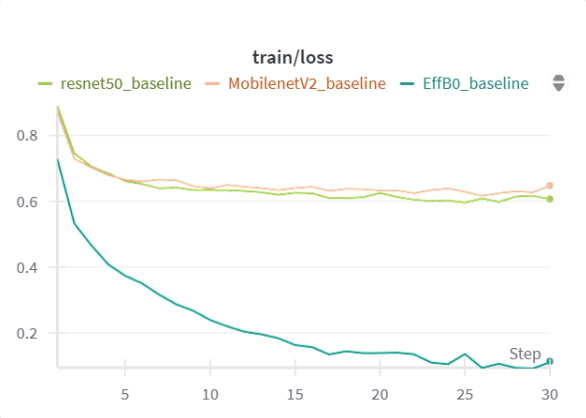
	precision	recall	f1-score	support
0	0.64	0.16	0.26	220
1	0.70	1.00	0.82	1341
2	0.00	0.00	0.00	23
3	0.67	0.08	0.14	223
4	0.00	0.00	0.00	28
5	0.88	0.07	0.13	103
6	0.50	0.02	0.03	65
accuracy			0.70	2003
macro avg	0.48	0.19	0.20	2003
weighted avg	0.67	0.70	0.60	2003

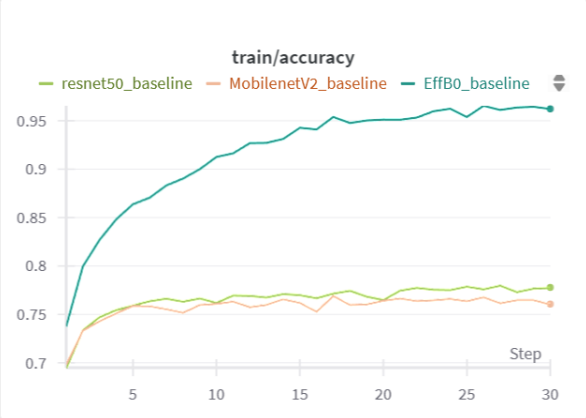
As a classical machine learning baseline, we evaluated a **Random Forest classifier** on the same dataset.

The model achieved an **overall accuracy of 70%**, with a **weighted average F1-score of 0.60**. However, the **macro average F1-score was only 0.20**, indicating that the model struggled with minority classes.

2. Classification accuracy of baseline CNN models





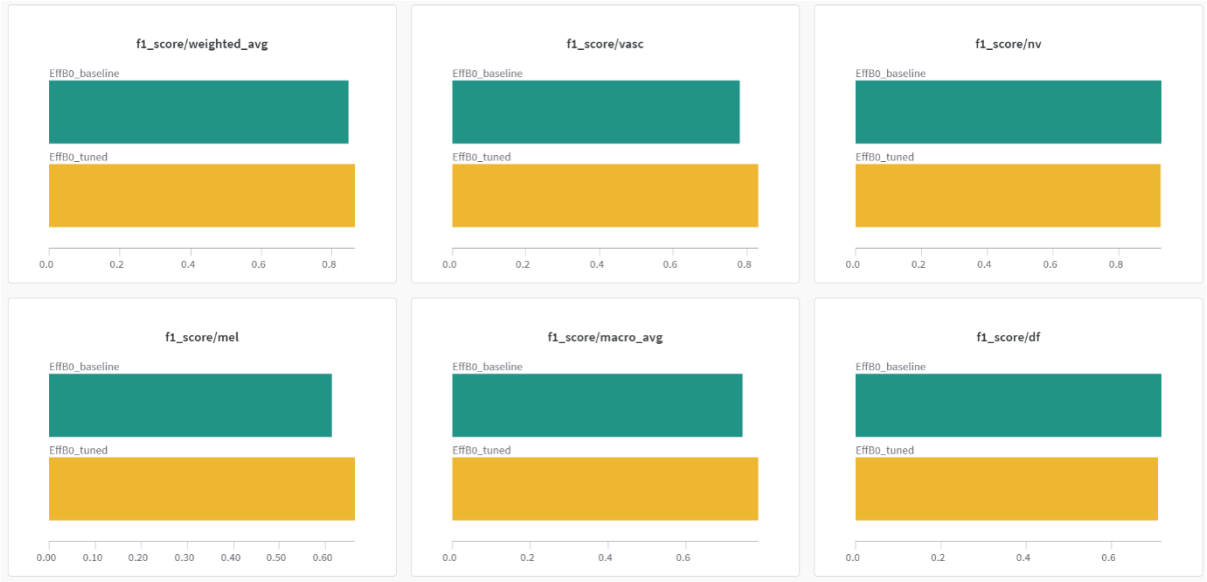


We evaluated 3 baseline CNN models, **ResNet50**、**EfficientNet** and **MobileNet.**

Among the 3 models, **EfficientNet-B0** achieved the highest performance, reaching over **86% validation accuracy** and nearly **96% training accuracy**.

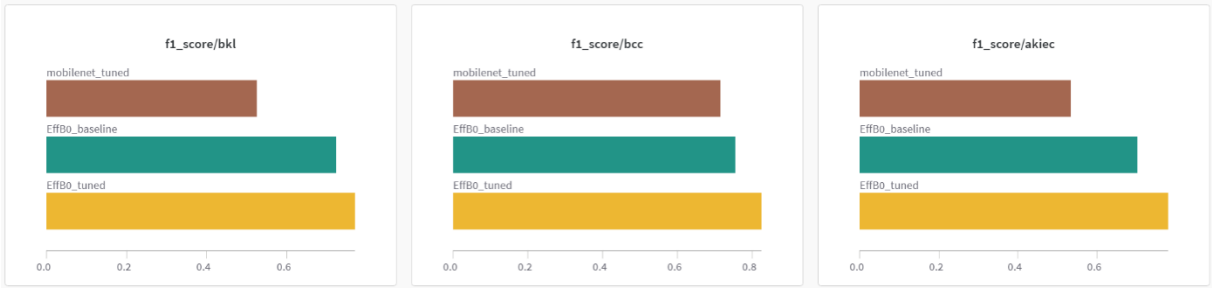
In contrast, **ResNet50** and **MobileNetV2** showed lower and more stable performance, with validation accuracy plateauing around **76–78%**.

3. F1-scores



The **fine-tuned EfficientNet-B0** model consistently outperforms the baseline across all F1-score metrics.

The improvement is especially clear in the *mel* class, which is very important in real medical situations and usually appears less often in the data.

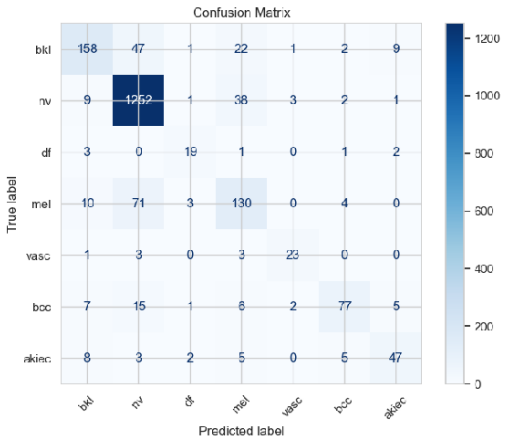


Meanwhile, using F1-score as the evaluation metric, we can see that the **fine-tuned EfficientNet-B0** achieved significantly better results on **less common classes** like *akiec* and *bcc*.

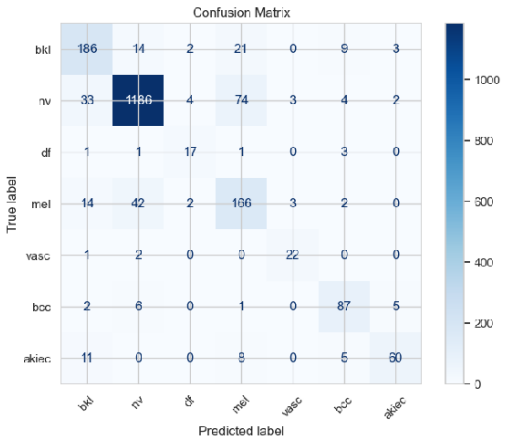
These classes are underrepresented in the dataset, so the improvement suggests the model has learned to recognize them more accurately after tuning.

4. Confusion matrix

EffB0_baseline



EffB0_tuned



The confusion matrices clearly show that after tuning, EfficientNet-B0 made more correct predictions for many classes, especially rare ones like *akiec* and *bcc*.

We also observe fewer misclassifications between *mel* and *nv*, which are commonly confused due to visual similarity.

5. Perspectives

- Incorporate multi-label classification (e.g., symptoms, anatomical location)
- Expand the ontology with richer properties (color, texture, location)
- Experiment with advanced architectures (e.g., Vision Transformers)
- Build a visual interface for clinicians to interact with predictions and semantic explanations
- Explore SWRL rule integration for stronger semantic validation

6. Conclusion

In this project, we evaluated three baseline CNN models — ResNet50, MobileNetV2, and EfficientNet-B0 — on the HAM10000 skin lesion dataset. Among them, **EfficientNet-B0 consistently outperformed the others**, achieving the highest F1-scores and accuracy, especially on rare and clinically significant classes such as *melanoma* and *akiec*.

By applying **fine-tuning** to the backbone, we further improved the performance of EfficientNet-B0. This adjustment significantly boosted its ability to classify underrepresented classes, as confirmed by confusion matrices and per-class F1-score analysis.

To ensure semantic correctness beyond raw predictions, we integrated an **ontology-based validation layer**. This approach checked whether predictions were consistent with domain knowledge (e.g., typical age, sex, and body location for each lesion class), and flagged potential mismatches. Additionally, **malignancy detection** via OWL reasoning added interpretability and clinical relevance to the output.

Overall, our best-performing and most reliable pipeline combines:

- **EfficientNet-B0 with fine-tuning** (for classification),
- **Ontology-based validation and explanation** (for semantic correctness and trust).

This integrated approach ensures both high prediction accuracy and medically meaningful output, supporting safer and more interpretable AI in dermatology.