Data Leakage Report for HAM10000 Dataset

Hassan Tamer 7405 - Sohail Waleed 7372 - Hana Waleed 7599

December 9, 2024

Introduction

Data leakage poses a critical issue in machine learning pipelines, especially in datasets intended for medical imaging tasks such as the HAM10000 skin lesion dataset. Data leakage occurs when information from the testing set influences the training process, leading to overly optimistic evaluation metrics. This report identifies potential sources of data leakage in the HAM10000 dataset, and offers visualization of what was discovered during the task.

Identified Sources of Data Leakage

1. Duplicate Lesion Images Across Splits

The HAM10000 dataset includes multiple images of the same lesion.The image may be exactly the same or under different conditions such as zoomed out or from a different angle. If these duplicates are distributed across training, validation, and testing sets, the model can "memorize" specific features rather than generalizing to unseen data.

Impact:

- Inaccurate evaluation metrics that appear big in development and a lot smaller in production.
- Reduced generalization to unseen data in real time.

Mitigation:

- Use feature-based similarity measures to detect duplicate images and keep only unique ones.

This was done using mobileNet architecture and removing the classification head to get representation. The representations are then calculated pairwise and compared against a threshold.

Visualization: Figure 1 shows an image that is duplicated twice in the dataset as a sample. Further examples of these are shown in data leakage notebook.

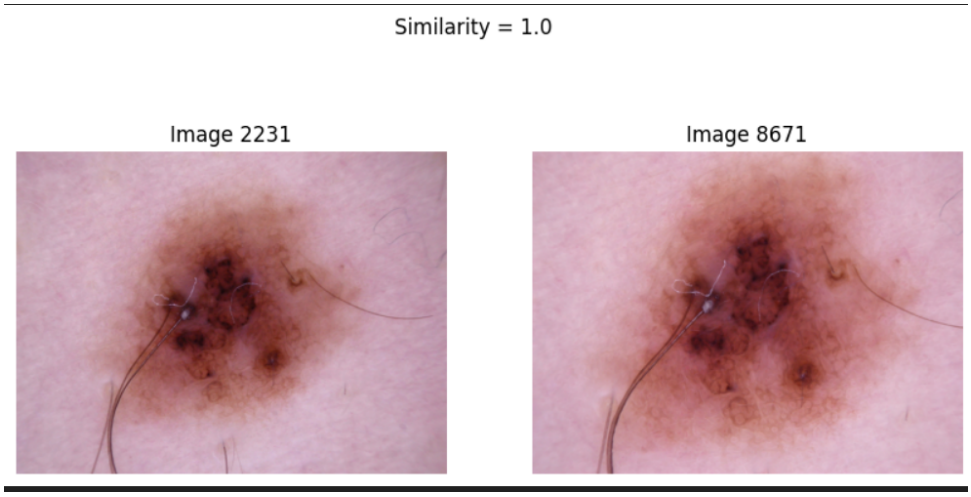

Figure 1: Duplicate Image

2. Leakage Through Masks

Similar to images, masks associated with the same lesion may contain duplicate patterns. Overlapping masks across dataset splits can lead to leakage similar to that discussed above.

Impact:

- Misleading evaluation of model performance.

Suggested Mitigation:

- Compute similarity scores between masks using metrics such as Intersection over Union (IoU) or Dice coefficient.
- Group masks by lesion and assign them consistently to a single split.

In our code we tried to calculate the pairwise similarity between masks, but that required a lot of resources as computing IOU score for all pairwise images is computationally expensive.

3. Class Imbalance

The dataset exhibits significant class imbalance, with some classes (e.g., nevus) dominating the dataset and others (e.g., vascular lesions) being underrepresented.

Impact:

- Models trained on imbalanced data may perform poorly on minority classes.
- Overall performance metrics may be skewed in favor of majority classes.

Mitigation:

- Use oversampling or undersampling techniques to balance the dataset.
- Apply weighted loss functions during training to penalize errors in minority classes more heavily.

This was also not implemented in the actual process due to limited resources. **Visualization:** Figure 2 illustrates the class distribution..

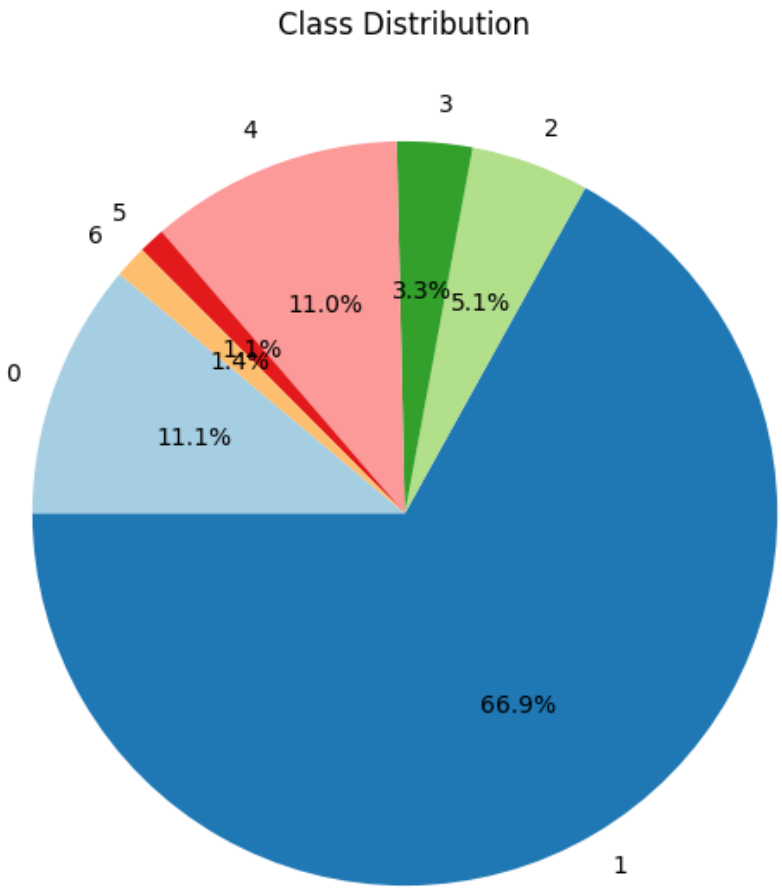

Figure 2: Class Distribution in the Dataset

Conclusion

The HAM10000 dataset requires careful handling to avoid data leakage. By addressing duplicate images and masks, ensuring consistent dataset splits, and managing class imbalance, we can build models that generalize effectively on it.

References

- HAM10000 Dataset: `https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000`
- GitHub Repository on HAM10000 Analysis: `https://github.com/sofglide/lesion-diagnosis`