CS682 Project Report Format

Course Title: CS682 - Software Development Laboratory I, Fall 2024
Project Name: Explainable AI Medical Imaging
Team Members: SangHyuk Kim (Project Manager, Developer), Yomil Shah (Developer), Kunal Sangurmath (Developer)

GEIR/Department Information	
Company Name	**Machine Psychology group**
Industry	**Medical Imaging Research Facility**
Company Contact (Name, Phone, Email)	**SangHyuk Kim, 5405585645, sanghyuk.kim001@umb.edu**
Brief Overview of Venture	**The research group, directed by Daniel Haehn, researches machine learning and medical imaging technologies.**
Scope of Work	**Machine learning, biomedical imaging research, machine perception in multi-modal.**

1. Executive Summary

The project aims to develop a robust melanoma detection framework by incorporating a duplicate image detection algorithm and a melanoma segmentation algorithm to enhance machine learning training. Duplicate images and random noise in raw data are significant challenges when fitting advanced machine learning models, as they can introduce bias and hinder accurate learning. This issue is particularly prevalent in large datasets, such as the one used in this project, which contains over 20,000 images—making manual analysis infeasible.

To address this, our solution automates the process of duplicate image detection and segmentation, ensuring the dataset is clean and optimized for training. The duplicate detection algorithm systematically identifies and removes redundant images, minimizing error-prone training caused by data duplication. Simultaneously, the segmentation algorithm focuses on isolating melanoma regions, enhancing the quality of data input for machine learning models.

The framework statistically analyzes the dataset, providing researchers with detailed findings on image duplicates and segmentation results. By automating these processes, our project improves the reliability of melanoma detection models and offers a streamlined approach for future researchers to clean and preprocess large-scale datasets effectively. This solution paves the way for more accurate and efficient melanoma detection, contributing to advancements in medical imaging and machine learning research.

2. Introduction

Project Background: Melanoma is one of the most aggressive forms of skin cancer, making early detection crucial for significantly improving survival rates. This project aims to develop an automated framework for melanoma detection that addresses key challenges in medical imaging, particularly the presence of duplicate images and the need for precise lesion segmentation. These challenges are critical because high-risk applications like melanoma detection rely heavily on machine learning (ML) models. Any malfunction caused by poor data quality can lead to severe consequences.

Project Goals: A significant focus of this project is data integrity. Duplicate images, often with different filenames, are common in large medical datasets and can introduce bias or degrade model performance. To address this issue, we have developed an automated system to:

- Detect and manage duplicate images within individual datasets and across multiple datasets.
- Segment melanoma lesions precisely, concentrating only on the required regions of interest.
- Present analysis findings with numbers.

The framework provides tools to identify, remove, and export results of duplicate image detection in a user-friendly manner. By automating these processes, the project ensures the quality and statistical reliability of the dataset, which is essential for building robust ML models.

3. Team Contributions

Team Member	Role	Contribution Summary
Kunal Prabhakar Sangurmath	Developer	Developed and tested ConvLSTM and U-Net models for melanoma segmentation, leveraging superpixels and ISIC datasets (2016–2020).
Yomil Shah	Developer	Developed a duplicate detection system that uses a hashing algorithm and stores outputs in JSON format, an efficient file format for scientific research.
SangHyuk Kim	Project Manager, Developer	Set the group objectives and distribute the work to the team members. Regular check of updates and alignment of objectives. Provide a code base for the team members to perform melanoma imaging work and guide the team members for any help throughout the semester.

4. Project Timeline

- **Milestones**:

Week 1–2

- Model Selection:
  - Select models such as U-Net and ConvLSTM models for segmentation performance.
  - U-Net model training began with 10 epochs.

Week 3

- U-Net Model Progress:

    - U-Net achieved test accuracy of 86.43%, IoU 55.17%, and loss of 37.51%.
    - Parameter tuning and additional training initiated.
- A dataset dictionary was prepared for organized dataset management.

Week 4

- Extended U-Net Training:
    - U-Net trained for 22 epochs. Results: Test accuracy 89%, IoU 60.7%, loss 32%.
    - Results were uploaded for segmentation and bounding box evaluation.

Week 5

- Duplicate Image Filtering:
    - Unique image datasets were created by removing duplicates from ISIC datasets (2016–2020).
    - U-Net segmentation results tested across ISIC datasets (2016–2019).

Week 6

- Model Testing Challenges:
    - ConvLSTM predictions tested; bounding box cropping issues identified.
    - Cropping of original test images based on bounding boxes initiated for further analysis.

Week 7

- Dataset Refinement and Testing:
    - Testing of U-Net and ConvLSTM models on finalized datasets completed.
    - Duplicate images between 2016–2020 datasets were identified for removal.
- Image quality and segmentation accuracy discussions continued.

Week 8

- Final Model Results:
    - ConvLSTM model results for ISIC datasets (2017–2020) finalized and shared.
    - Unique datasets with duplicates removed were confirmed for all years.

Week 9–10

- Final Report Writing:
    - Drafting of the final report began, covering:
    - Collaborative editing of the report in progress.

- **Gantt Chart** (optional):

Task	Week 1-2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9
Compare U-Net and ConvLSTM models	X							
Train U-Net (10 epochs)	X							
Tune and train U-Net (22 epochs)		X						
Prepare dataset dictionary		X						
Remove duplicate images			X	X				
Test U-Net on ISIC datasets			X	X				
Train and test ConvLSTM model				X	X	X		
Complete bounding box & cropping					X			
Finalize results for all datasets						X	X	
Write and finalize the report								X

5. Technical Implementation

- **Technology Stack**: Python, TensorFlow, Keras, PyTorch (for U-Net), OpenCV, PIL, Matplotlib..
- **Architecture Overview**:
    - The project incorporates duplicate image detection and segmentation algorithms for accurate and stable melanoma detection. In the segmentation phase, we use U-Net and ConvLSTM models for skin lesion segmentation on the ISIC dataset.

While detecting duplicate images, we use a hashing algorithm. After duplicate detection and image segmentation, we crop images based on bounding boxes to refine regions of interest. Finally, we evaluate segmentation model performance using accuracy and IoU.

- **Development Process**:

  Input: Gather skin cancer images and analyze their size, file format, and structure for segmentation and duplicate detection.
  Hash Computation: We concluded that the hash algorithm is appropriate for duplicate image detection as a hash can detect identical images.
  Segmentation: We use Google Colab to implement a segmentation algorithm and share it among the team members so that each member can contribute.
  Duplicate Detection: We designed a program to find duplicates across the datasets so the program won't miss any duplicate images for the combined dataset.
  Output: Unique datasets with removal of duplicate files and an option to generate JSON files having filenames and their respective duplicates across the datasets.

6. Key Features Developed

- ConvLSTM Model:
  - Leveraged sequence-based processing for accurate skin lesion segmentation.
  - Achieved high performance with superpixels as additional input features
- U-Net Model:
  - Encoder-decoder architecture effectively captured fine-grained lesion boundaries.
  - ResNet34 encoder improved feature extraction for complex patterns.
- Bounding Box and Cropping Pipeline:
  - Developed functionality to generate bounding boxes on predicted masks.
  - Cropped regions from original images for easier clinical validation.
- Testing Across ISIC Datasets:
  - Implemented extensive data augmentation and normalized inputs.
  - Evaluated the models on datasets from 2016 to 2020 for robustness.
  - Generated visualizations comparing original images, predicted masks, and bounding boxes.
- Duplicate Image Detection Within Datasets:
  - Identify images within a single dataset. Provide an option to list or delete duplicates.
- Duplicate Image Detection Across Datasets:
  - Comparing multiple datasets to find duplicate images.
- JSON-based Output
  - Production of duplicate images in a particular dictionary format. Hence, making it possible to analyze and debug the output when needed.
- Performance Optimization
  - Parallel hash computation ensures scalability for large datasets.
- User Control

    - Providing an option for the user to keep all the files intact or remove duplicates automatically

7. Testing and Quality Assurance

- **Testing Strategies:**
    - Model Validation:
        - Used ISIC 2017 validation dataset to tune hyperparameters.
        - Employed Dice and IoU metrics to evaluate segmentation accuracy.
    - Multi-dataset Testing:
        - Tested models on ISIC datasets from 2016, 2018, 2019, and 2020 to assess generalization.
    - Visualization Testing:
        - Verified the accuracy of bounding box placements and cropped outputs.
    - Bug Tracking:
        - Performed weekly peer-reviews and gathered the opinions of the team members.
        - Individuals checked skin cancer images manually and reported bugs, such as misaligned bounding boxes, scaling, duplicate images, or JSON file structure.

8. Challenges Faced

- Computational Overhead:
    - Processing and integrating superpixels in real time increased computational complexity. To solve this, precomputed superpixel masks and stored them for efficient loading.
    - Large datasets were straining computational resources. Parallel processing was used to reduce runtime significantly.
- Model Performance:
    - Models trained on ISIC 2017 showed reduced performance on other datasets due to variations in image quality and lesion characteristics. So, we used extensive data augmentation and normalized inputs to remedy the performance issue.
- Integrating file extensions:
    - Some datasets also contained non-image files. Hence, careful filtering of files was required. So, we used ‘imutils.path.list_images()’ to ensure only valid image files are processed.

9. Outcomes and Learnings

- **Results**:
    - ConvLSTM Model:
        - Achieved an Accuracy of 85%, IoU of 48% and Loss of 33% on ISIC 2017 test data.
        - Demonstrated robustness across ISIC datasets with consistent bounding box outputs.

  - U-Net Model:
    - Achieved an Accuracy of 86.43 %, IOU of 55.17%, and Loss of 37.51% on ISIC 2017 test data.
    - Also, demonstrated robustness across ISIC datasets with consistent bounding box outputs.
  - Duplicate image detection:
    - Duplicates were successfully detected and removed from Melanoma datasets, enhancing their reliabilities.
- **Learnings**: We developed and tested ConvLSTM for temporal data and U-Net with a ResNet34 encoder for skin lesion segmentation. We utilized superpixels for feature engineering and applied visualization techniques like bounding boxes for better model interpretability. We also learned about duplicate image detection using hashing algorithms, optimized Python programs through parallel processing, and evaluated model performance using metrics such as IoU, accuracy, and loss.

10. Future Work and Recommendations

- **Improvements**: With additional time and resources, model performance can be enhanced by exploring more advanced architectures, such as attention-based models or transformer variants, for segmentation tasks. Incorporating larger and more diverse datasets would improve generalization while further optimizing hyperparameters and training strategies can lead to better accuracy and efficiency. Additionally, implementing more robust duplicate image detection algorithms and enhancing parallel processing for faster computations could streamline workflows.
- **Recommendations for Future Teams**: Future teams should maintain a well-documented pipeline to ensure smooth progress and collaboration. Emphasizing regular testing, validation, and visualization of results will help identify issues early. Leveraging advanced libraries and frameworks, such as PyTorch or TensorFlow, can simplify development, and integrating automated performance monitoring tools will optimize the process further.

11. Conclusion

The project involved developing and testing ConvLSTM and U-Net architectures for skin lesion segmentation, integrating advanced techniques such as ResNet34 encoders, superpixels, and visualization tools like bounding box generation. This process made significant progress in understanding temporal data handling, encoder-decoder models, and performance evaluation metrics such as IoU and accuracy. Additionally, key technical contributions were exploring image hashing for duplicate detection and optimizing Python programs with parallel processing. This journey improved model performance and provided valuable learning experiences in research, development, and collaboration.

12. Appendix (Optional)

- **Code Repositories**:
  - U-Net segmentation

    - ConvLSTM segmentation
    - Duplicate image repository
    - Image Dictionary Generator
    - Melanoma detection code base
- **References:**
    - U-Net: Convolutional Networks for Biomedical Image Segmentation
    - Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting
    - SLIC Superpixels Compared to State-of-the-Art Superpixel Methods
- **Documentation**:
    - Melanoma detection with uncertainty quantification
    - Web-based melanoma detection