Bioinformatics Summer School Report

Name: Yahya Alhabboub
Internship Organization: University of Toronto
Supervisor's Name: Jüri Reimand
Project title: From CNA Profiles to Prognostic Predictions: Utilizing Hilbert Curve Images for Cancer Subtype Classification and Survival Analysis

1. Abstract

Cancer results from DNA mutations, including Copy Number Alterations (CNAs) where large DNA segments are duplicated or deleted. This study examines the potential of CNAs as biomarkers for cancer subtype classification, tumor stage prediction, and survival analysis, using nearly 10,000 samples across 25 cancer subtypes. We employed Hilbert Curve visualizations of CNA profiles and applied computer vision models, including YOLOv8 and VGG16, for classification tasks. A logistic regression model using patient cytoband data served as a control. YOLOv8 performed best in cancer subtype classification, surpassing VGG16 and the control model, although tumor stage classification yielded poor results across all models. Survival analysis with DeepSurv showed fluctuating C-index results, indicating issues with overfitting.

2. Objectives

- **Exploring CNAs:** Investigate their role in early cancer diagnosis, treatment, and tumor characterization, including subtype identification, staging, and resistance.
- **AI Model Development:** Develop AI models to distinguish between cancerous and normal mutations.
- **1D to 2D Conversion:** Transform CNA sequences into two-dimensional formats for better analysis, applying computer vision techniques like transfer learning for enhanced interpretation.
- **Model Comparison:** Compare single-task vs. multi-task models in cancer classification tasks.
- **CV vs. ML Models:** Evaluate complex CV models against simpler ML models like logistic regression using patient cytoband data.
- **Survival Analysis:** Use CNAs in survival analysis through the DeepSurv framework.

3. Materials and Methodology

3.1 Dataset

We utilized the Copy Number Alterations (CNA) dataset from the TCGA - PanCanAtlas Publications collection [1]. This dataset was employed to generate CNA images and calculate cytoband values for each patient.

3.2 Preprocessing workflow:

Figure 1 is a workflow that summarizes the key details of each Preprocessing steps:

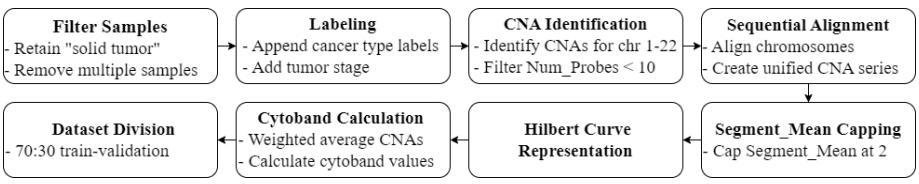


Figure 1: Preprocessing workflow

3.3 Modeling:

- **YOLOv8 Classification:** YOLOv8n-cls was applied to classify cancer subtypes using CNA images, with the following hyperparameters: epochs=150, imgsz=480, lr=0.01, batch=16, optimizer=SGD.
- **Multi-task VGG16 Model:** A multi-task model based on VGG16 was employed for classifying both cancer subtypes and tumor stages using CNA images. The hyperparameters were: lr=0.0001, optimizer=Adam, epochs=300, imgsz=224, batch=32, with "LeakyReLU" as the activation function between layers and "Softmax" in the final layer, using "CrossEntropyLoss" as the loss function.
- **Control Model:** Logistic Regression was used to classify cancer subtypes using cytobands, with hyperparameters: max_iter=800, solver='liblinear', multi_class='ovr', class_weight='balanced'.
- **Survival Analysis Model:** Cytobands were used with DeepSurv, implemented using pre-existing code with modifications[2]. The model used "SELU" activation between layers, a "Linear" function in the final layer, and "NegativeLogLikelihood" as the loss function, with hyperparameters: optimizer=Adam, learning_rate=0.000001, epochs=200, lr_decay_rate=0.001.

3.4 Post-processing:

Classification tasks: Precision - Recall. -1-score.

Survival analysis tasks: C-index.

4. Results and Discussion

4.1 Cancer Subtypes Task:

Figure 2 displays the results of all metrics for all subtypes of the different models:

Performance Variation: F1-scores varied, with TGCT showing the highest and ESCA the lowest.

Model Comparison: YOLOv8n-cls performed best, followed by the control model, then VGG16.

Metric Consistency: Precision, recall, and F1-scores were generally aligned.

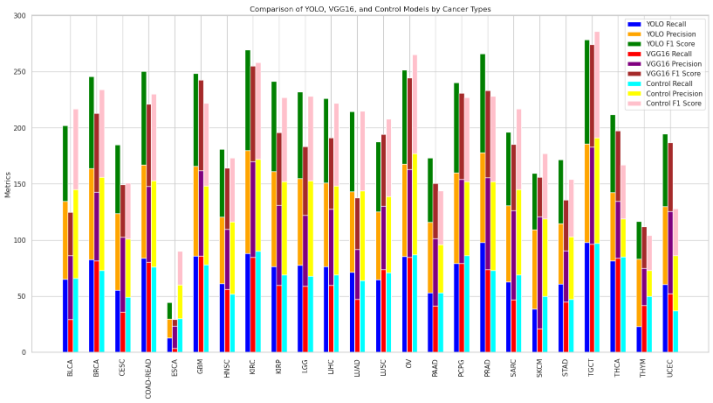


Figure 2: Cancer Subtypes Results

4.2 Tumor Stage Task:

Overall Performance:

Figure 3 shows uniformly poor classification results for tumor stages across all models.

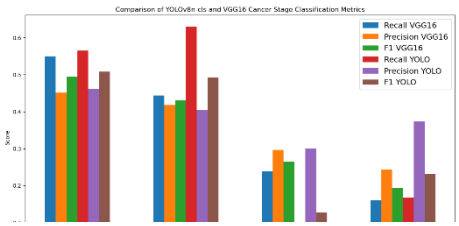


Figure 3: Tumor Stage Results

4.3 Survival Task:

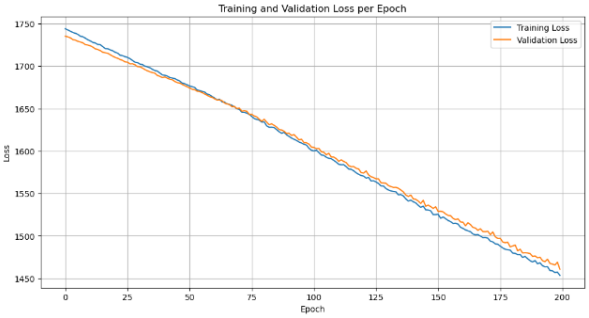


Figure 5:Survival Model Loss

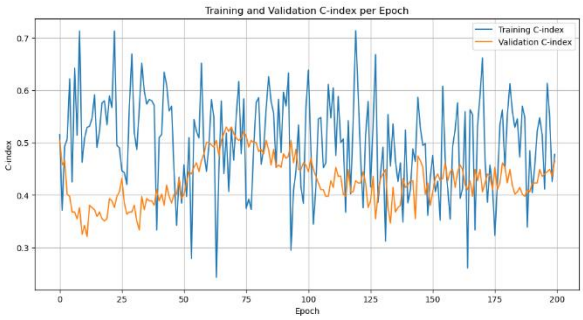


Figure 4: Survival Model Results

Model Stability: Although the loss function is ideal in Figure 4 convergence across epochs, Figure 5 shows the fluctuating C-index in training and validation sets suggesting issues with model overfitting or underfitting.

5. Skills and Knowledge Gained

-Gaining Insights into Mutations.
-Data Transformation.
-Complex Model Design

-Performance Evaluation
-Survival Analysis
-Scientific Processes

6. Challenges and Solutions

-Email Communication Ineffective → Increased frequency
-Overwhelming Ideas → Prioritized, canceled some
-Unavailable data → Used what was available

7. Conclusion

Primary Objective Explored the utility of CNAs in studying tumor characteristics, with some subtypes showing excellent performance, highlighting their potential in cancer research.

Challenges:

- Limited dataset (10,000 samples) made addressing multiple tasks challenging.
- Data augmentation was not feasible due to the reliance on specific image features.
- Pre-trained models may have been too complex for simple CNA images.

Discussion and Future Work

Research Foundation: The study provides a basis for further CNA exploration.

Next Steps:

- Expand the dataset and develop CNA-specific models.
- Include predictions for additional patient characteristics (age, gender).

8. Acknowledgments

The From CNA Profiles to Prognostic Predictions: Utilizing Hilbert Curve Images for Cancer Subtype Classification and Survival Analysis is supported by the Data Sciences Institute, University of Toronto. KAUST Academy is financially supporting Yahya Alhabboub.

9. References

[1] Zack, T., Schumacher, S., Carter, S. et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet 45, 1134–1140 (2013). https://doi.org/10.1038/ng.2760.

[2] C. Zifan, "DeepSurv.pytorch," GitHub, 10-Apr-2020. [Online]. Available: https://github.com/czifan/DeepSurv.pytorch/tree/master. [Accessed: 10-Aug-2024].