Enhancing Cancer Detection Through Natural Language Processing

Rishabh Medhi, Dravin Raj, Joshua Valiaveedu, Sachi Hansalia, Ela Guo
Lead: Philip Lee   Supervisor: Dr. Berrak Sisman

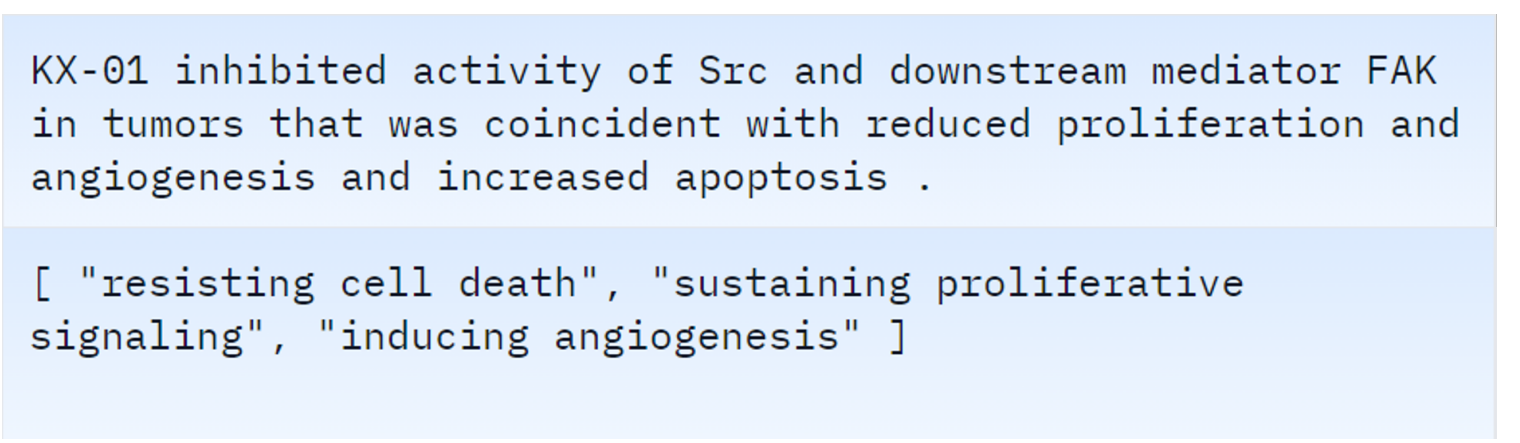
Introduction

Cancer, one of the world's leading causes of death, has been the subject of study in many different fields. If detected early, before progression into its fatal stages, cancer is treatable. The current issue is detecting cancer in its early stages. We aim to detect these early signs of cancer by implementing computational tools and semantics.
The past decade has seen leaps in machine learning research, especially with the design of the Transformer Large Language Model (LLM), developed by Google Brain researchers. This model can be utilized for Natural Language Processing (NLP), which has a number of applications such as translation and text generation.[2] Using the RoBERTa model, we created a model to detect cancer in text data based on the ten major hallmarks of cancer.

Purpose

We sought to evaluate the efficiency of implementing an LLM, represented by RoBERTa, in identifying indicators for cancer.
Our goal is to build a model that can accurately classify semantic data according to the ten hallmarks of cancer, thereby improving cancer detection while it is in a preventable stage. This has the potential to provide patients more time for treatment. The model could be beneficial to parts of the world with a lack of adequate medical testing equipment, by functioning as a patient screening and diagnostic tool.

Dataset



```
KX-01 inhibited activity of Src and downstream mediator FAK
in tumors that was coincident with reduced proliferation and
angiogenesis and increased apoptosis .

[ "resisting cell death", "sustaining proliferative
signaling", "inducing angiogenesis" ]
```

Figure 1. Sample extracts of dataset

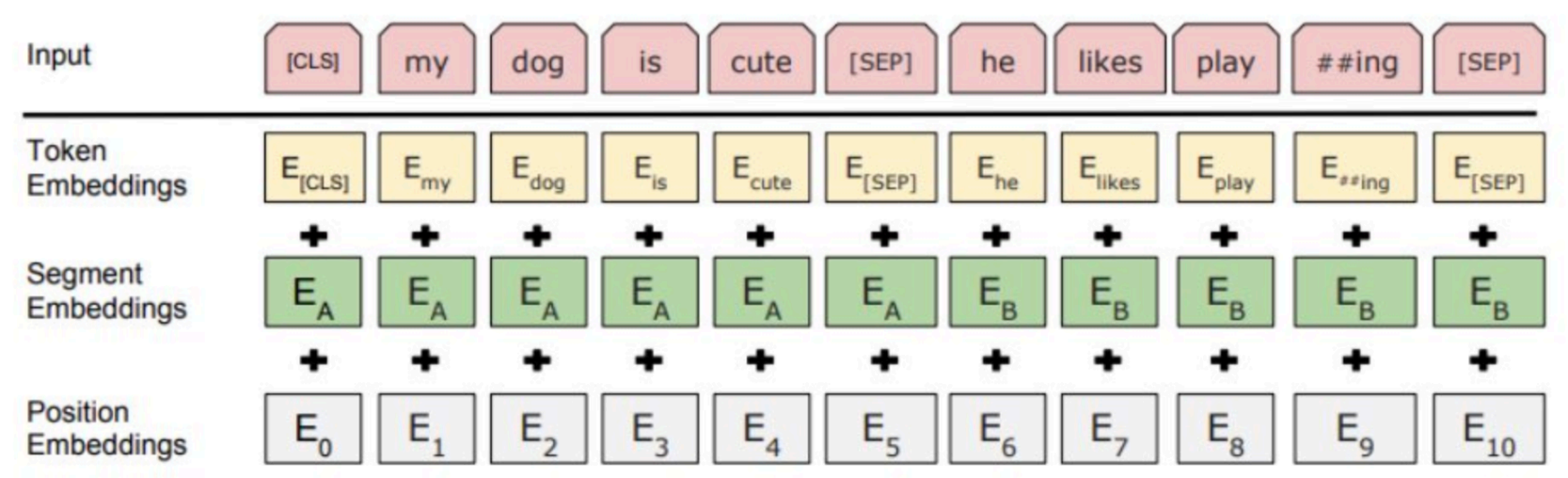


Figure 2. Text Preprocessing

The "Hallmarks of Cancer" is a dataset from *Automatic semantic classification of scientific literature according to the hallmarks of cancer*. It consists of extracts from approximately 1500 research paper abstracts from various fields such as molecular biology, public health, and clinical medicine. These extracts are labeled with numbers according to one of the hallmarks they present for the purpose of promoting research into ways to improve healthcare through AI/ML techniques.[3] To prepare the dataset for training, we cleaned and resampled from the set as follows:

- The training set for our model consisted of over 12k entries
- There was an overrepresentation of the label hallmark 7, "Genome instability mutation (GI)"
- To prevent bias towards the majority class, we randomly sampled 1000 of those entries to train the model
- Given the multi-label nature of the task, the outputs were one-hot-encoded to represent the presence or absence of labels across multiple classes

RoBERTa Model

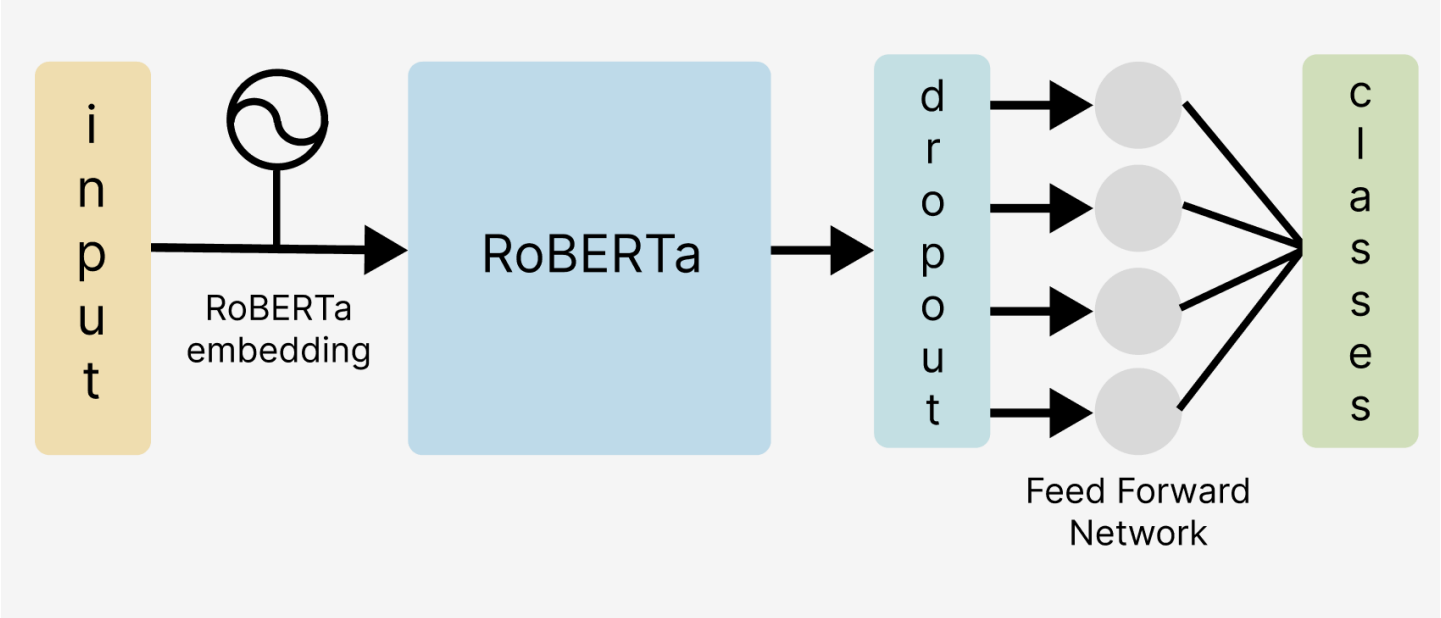


Figure 3. RoBERTa Classification Architecture

We used the RoBERTa **(Robustly optimized Bidirectional Encoder Representations from Transformers approach)** model, to identify cancer hallmarks.
RoBERTa utilizes the **Transformer architecture**, introduced by the Google Brain team, which consists of encoder and decoder components utilizing self-attention mechanisms, substituting traditional recurrent and convolutional neural networks. RoBERTa incorporates dynamic masking, a stabilized training process, and eliminates the prediction of additional sentences[4]
The model processes inputs by tokenizing the text into vectors encoded with the word's position and relationship to other words in the sentence, creating a matrix of token vectors. RoBERTa then applies dynamic masking and multiple attention heads to the matrix of vectors, allowing it to capture complex relationships in the text as well as generalize its training.

Results

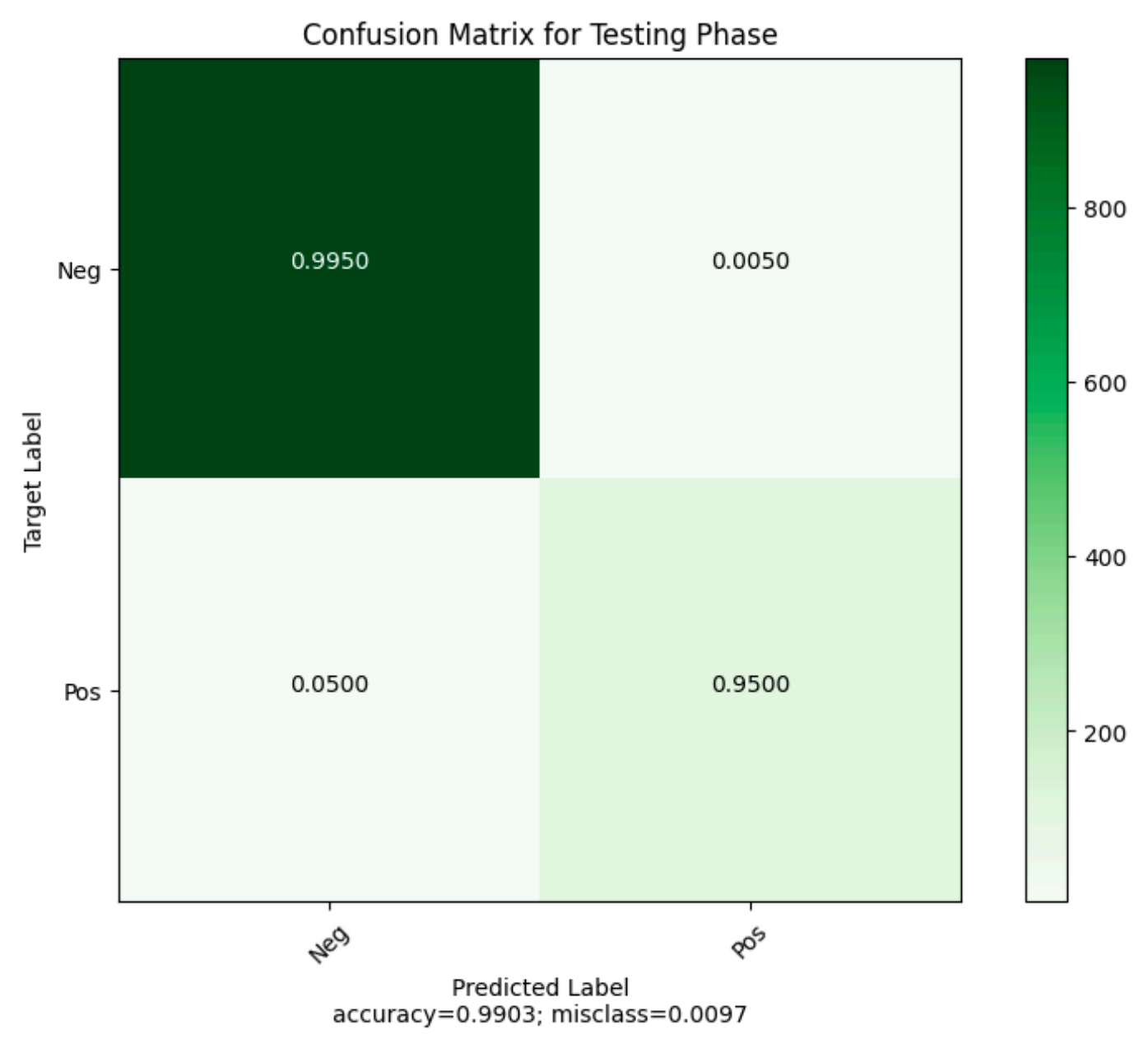


Figure 4. Heatmap denoting Confusion Matrix

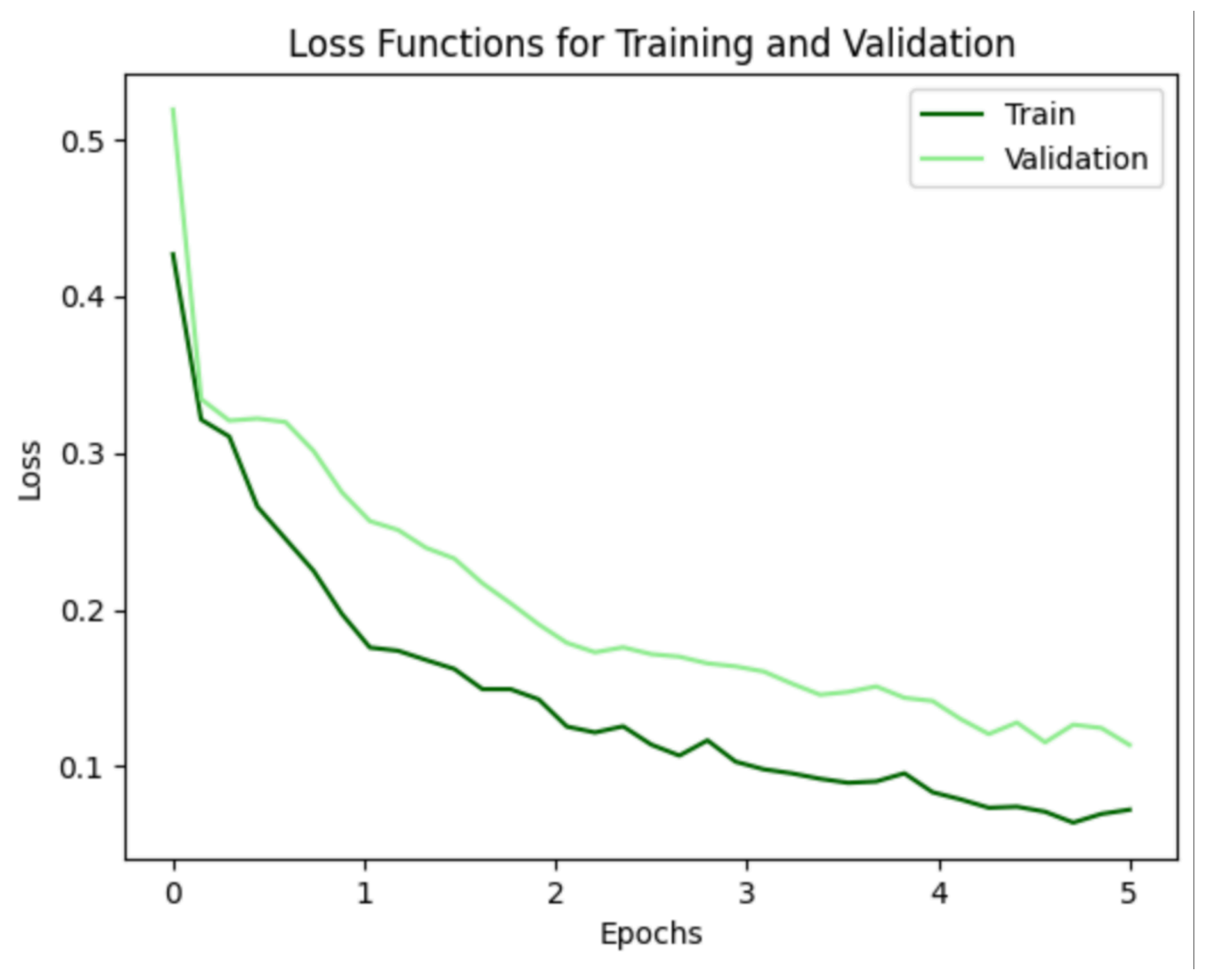


Figure 5. Loss Function

We evaluated the model's performance using metrics such as accuracy, F1, precision, and recall. The values for each metric are as follows:

- **Accuracy Score**: 0.907
- **F1 Score (Micro)**: 0.832
- **F1 Score (Macro)**: 0.841
- **F1 Score (Weighted)**: 0.826
- **Precision (Weighted)**: 0.849
- **Recall (Weighted)**: 0.820

Analysis

The text was tokenized through the input embedding layer and the specifications of the training process were set. The data was then split into the training set (**80%** of data) and test set (**20%** of data). **DataLoader** from PyTorch was used to parallelize data loading as multiple "heads." Finally, the data was passed through a pretrained RoBERTa model, and those "heads" were given label classifications. Our best-performing model had the following parameters.

- **Max Embedding Length**: 256
- **Batch Size**: 8
- **Epochs**: 5
- **Learning Rate**: 1e-05

We used the following metrics to measure our model's performance.

Precision is the percentage of positive samples that are correctly identified as positive out of all samples identified as positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

Accuracy is measured as the percentage of true positive and true negative predictions out of all predictions that are made.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

F1 Score Curve: The F1 score is fundamental in obtaining the model's confidence score that best balances the precision and recall values.

$$\text{F1 Score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \quad (3)$$

Where TP is proportion of true positive, TN is true negative, FP is false positive, and FN is false negative.

Conclusion

Our model excels in effectively pinpointing the distinctive hallmarks associated with cancer within the diverse textual dataset that was utilized. This functionality holds immense promise, not only in providing valuable support to physicians during diagnostic processes but also in significantly streamlining the classification of cancer research materials, thereby catalyzing advancements in the broader field of medicine.
Future improvements include expanding the model to differentiate between various types of cancer and exploring multi-modal capabilities with text and images to further enhance its detection capability. In our forthcoming approach, we advocate for an advanced integration of the ResNet-50 architecture and the RoBERTa Large Language Model. This combination aims to significantly elevate early cancer detection by concatenating their distinctive features. A Feed Forward Network will be also connected to optimize any multi-labeling task.

References

1. Farida B Ahmad and Robert N Anderson. *The leading causes of death in the us for 2020*. Jama, 325(18):1829–1830, 2021
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762
3. Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, Anna Korhonen. *Automatic semantic classification of scientific literature according to the hallmarks of cancer*, Bioinformatics, Volume 32, Issue 3, February 2016, Pages 432–440, https://doi.org/10.1093/bioinformatics/btv585
4. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arxiv:1907.11692). arXiv.
5. Iannantuono, G. M., Bracken-Clarke, D., Floudas, C. S., Roselli, M., Gulley, J. L., Karzai, F. (2023). *Applications of large language models in cancer care: current evidence and future perspectives*. Frontiers in oncology, 13, 1268915. https://doi.org/10.3389/fonc.2023.1268915