Research Plan/Project Summary Instructions

A complete Research Plan/Project Summary is required for ALL projects and must accompany Student Checklist (1A).

1. All projects must have a Research Plan/Project Summary
   a. Written prior to experimentation following the instructions below to detail the rationale, research question(s), methodology, and risk assessment of the proposed research.
   b. If changes are made during the research, such changes can be added to the original research plan as an addendum, recognizing that some changes may require returning to the IRB or SRC for appropriate review and approvals. If no additional approvals are required, this addendum serves as a project summary to explain research that was conducted.
   c. If no changes are made from the original research plan, no project summary is required.

2. Some studies, such as an engineering design or mathematics projects, will be less detailed in the initial project plan and will change through the course of research. If such changes occur, a project summary that explains what was done is required and can be appended to the original research plan.

3. The Research Plan/Project Summary should include the following:
   a. **RATIONALE:** Include a brief synopsis of the background that supports your research problem and explain why this research is important and if applicable, explain any societal impact of your research.
   b. **RESEARCH QUESTION(S), HYPOTHESIS(ES), ENGINEERING GOAL(S), EXPECTED OUTCOMES:** How is this based on the rationale described above?
   c. Describe the following in detail:
      - **Procedures:** Detail all procedures and experimental design including methods for data collection. Describe only your project. Do not include work done by mentor or others.
      - **Risk and Safety:** Identify any potential risks and safety precautions needed.
      - **Data Analysis:** Describe the procedures you will use to analyze the data/results.
   d. **BIBLIOGRAPHY:** List major references (e.g. science journal articles, books, internet sites) from your literature review. If you plan to use vertebrate animals, one of these references must be an animal care reference.

Items 1–4 below are subject-specific guidelines for additional items to be included in your research plan/project summary as applicable.

1. Human participants research:
   a. Participants: Describe age range, gender, racial/ethnic composition of participants. Identify vulnerable populations (minors, pregnant women, prisoners, mentally disabled or economically disadvantaged).
   b. Recruitment: Where will you find your participants? How will they be invited to participate?
   c. Methods: What will participants be asked to do? Will you use any surveys, questionnaires or tests? If yes and not your own, how did you obtain? Did it require permissions? If so, explain. What is the frequency and length of time involved for each subject?
   d. Risk Assessment: What are the risks or potential discomforts (physical, psychological, time involved, social, legal, etc.) to participants? How will you minimize risks? List any benefits to society or participants.
   e. Protection of Privacy: Will identifiable information (e.g., names, telephone numbers, birth dates, email addresses) be collected? Will data be confidential/anonymous? If anonymous, describe how the data will be collected. If not anonymous, what procedures are in place for safeguarding confidentiality? Where will data be stored? Who will have access to the data? What will you do with the data after the study?
   f. Informed Consent Process: Describe how you will inform participants about the purpose of the study, what they will be asked to do, that their participation is voluntary and they have the right to stop at any time.

2. Vertebrate animal research:
   a. Discuss potential ALTERNATIVES to vertebrate animal use and present justification for use of vertebrates.
   b. Explain potential impact or contribution of this research.
   c. Detail all procedures to be used, including methods used to minimize potential discomfort, distress, pain and ~~injury~~ to the animals and detailed chemical concentrations and drug dosages.
   d. Detail animal numbers, species, strain, sex, age, source, etc., include justification of the numbers planned.
   e. Describe housing and oversight of daily care
   f. Discuss disposition of the animals at the termination of the study.

3. Potentially hazardous biological agents research:
   a. Give source of the organism and describe BSL assessment process and BSL determination.
   b. Detail safety precautions and discuss methods of disposal.

4. Hazardous chemicals, activities & devices:
   - Describe Risk Assessment process, supervision, safety precautions and methods of disposal.
   - Material Safety Data Sheets are not necessary to submit with paperwork.

Preethi Krishnamoorthy
Computational Biology and Bioinformatics

Research Plan

A Comparison of Machine Learning Methods in the Analysis of Lymphocyte Patterns in Cancer Research

1. Rationale

a. Cancer has a major impact on many people's lives all around the world. In the United States alone, approximately 38.4% of men and women will be diagnosed with cancer at some point during their lifetimes, with breast cancer being the most common cancer [2].

b. Tumor infiltrating lymphocytes (TILs) play an important role in the study, diagnosis, and treatment of cancer. TILs are a type of immune cell that has moved from the blood into a tumor to try and attack the cancer. Lymphocyte patterns and distribution in tissue specimens and their relation to tumor regions is very important, as studies have shown that tumor infiltrating lymphocyte patterns can be used as biomarkers to predict disease outcome and response to treatment [3].

c. Being able to utilize TIL images and develop resources to be able to accurately diagnose cancer is very important, as many cases of cancer are either misdiagnosed or diagnosed very late [7].

d. Contemporary digital microscopes can capture high resolution images of tissue specimens. Analyses of these images using computer algorithms enable the quantitative study and characterization of tumor infiltrating lymphocyte (TIL) patterns and the relationships between TILs and the disease state [5].

e. In recent years, deep learning has emerged as a powerful machine learning technique. Deep learning-based analysis methods have demonstrated impressive results in image classification and segmentation compared to other machine

learning approaches. Hence, deep learning is rapidly becoming a core image analysis tool in biomedical imaging research [1].

f. Several deep learning models (such as ResNet, VGG, and Inception) have been developed by the imaging and computer vision communities [6-8]. These models differ in their network designs and method parameters. It is important to systematically evaluate and understand the classification performance of different deep learning models in biomedical image analysis.

g. In this project, we carry out an experimental evaluation of several deep learning methods for detection and classification of tumor infiltrating lymphocytes in whole slide tissue images. We will use results generated in the work done by Saltz et al. [5] In that work, a deep learning model was developed and applied to generate TIL maps from a range of cancer types, including breast cancer. These maps were reviewed and refined by a group of pathologists. This process generated high quality TIL maps but is a labor intensive and expensive process as it involved manual review and refinement of TIL maps. Our goal in this project is to experimentally evaluate a set of state-of-the-art deep learning networks and study if they can achieve similar results to those obtained in the previous work. By studying the differences and similarities in analysis performance of these networks we will identify the network that is the best for detecting and classifying TILs in whole slide tissue images. We will also examine what aspects of the network allow it to produce the best results.

2. Research Question(s), Hypotheses, Engineering Goal(S), Expected Outcomes

a. Research Questions

i. Which deep learning or machine learning method will classify TIL probability maps and summary statistics (cluster indices) into different immune subtypes with the greatest accuracy?

ii. How similar are the probability maps generated prior to this research by the Inception and VGG deep learning models?

b. **Hypotheses / Expected Outcomes**

i. The pretrained convolutional neural networks will classify TIL maps the best because pretrained models have initial layers that represent features found in almost all images and can be used for different types of images; these pretrained layers will allow the network to classify TIL maps with a greater accuracy.

ii. There will be differences in the results from the three different networks due to the different architectures of each deep learning network; however, the differences will not be very high.

c. Goals

i. The objective of this research is to develop various deep learning and machine learning models to classify patients into immune subtypes C1 and C2. To do this, TIL probability maps and summary statistics from tumors of various cancer types will be used with deep learning and machine learning methods. Then, we will compare the test accuracies of these methods and determine the best method for classification.

ii. Another goal of this research is to compute the dice coefficient from TIL maps generated by the Inception and VGG networks to determine how similar these two networks are.

3. Research Methods

a. Procedures

i. First, algorithms for classification will be created, and trained and tested with images from multiple cancer types. The algorithms that will be created are a fully connected neural network, a CNN, modified pretrained models, random forest, SVM and K means clustering. These algorithms will be implemented using the tensorflow library in Python.

ii. The fully connected neural network, CNN, and modified pretrained models will use TIL probability maps for training and testing.

iii. Random forest, SVM, and K means clustering will use cluster indices (ball hall, c index, Banfeld Raftery, determinant ratio, and leukocyte fraction) for training and testing.

iv. The test accuracies of all methods will be compared, and the method with the highest classification accuracy will be determined.

1. To determine the classification accuracy of K means clustering, the Rand index will be used. The Rand index is a measure of accuracy of the clustering algorithm. The equation for calculating the Rand index is $$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

v. Next, we will use the dice coefficient to compare patch-level classifications from TIL maps generated by the Inception and VGG networks. The Dice coefficient computes similarity between two sets. In our case, given a whole slide tissue image, one of the sets will be the set of image patches classified as TIL positive by one deep learning network and the other set will be the set of patches classified as TIL positive by the other deep learning network. The Dice coefficient is defined as $DSC = \frac{2|X \cap Y|}{|X| + |Y|}$, where X and Y are two sets being compared. The Dice coefficient will be implemented in Python.

b. Risk and Safety

i. There are no hazardous chemicals or biological agents that will be used in this project.

c. Data Analysis

i. The data analysis will be comparing the dice coefficient values for results generated from the two networks to determine the similarity between the

datasets, and comparing the test accuracies of the machine learning and deep learning methods.

References

1. Bi, W. L., Hosny, A. , Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A. , Allison, T. , Arnaout, O. , Abbosh, C. , Dunn, I. F., Mak, R. H., Tamimi, R. M., Tempany, C. M., Swanton, C. , Hoffmann, U. , Schwartz, L. H., Gillies, R. J., Huang, R. Y. and Aerts, H. J. (2019), Artificial intelligence in cancer imaging: Clinical challenges and applications. CA A Cancer J Clin, 69: 127-157. doi:10.3322/caac.21552
2. Cancer Statistics. (2018, April 27). Retrieved from NIH: National Cancer Institute website: https://www.cancer.gov/about-cancer/understanding/statistics
3. Dieci, M. V., Radosevic-Robin, N., Fineberg, S., Eynden, G. V. D., Ternes, N., Penault-Llorca, F., . . . Salgado, R. (2018). Update on tumor-infiltrating lymphocytes (TILs) in breast cancer, including recommendations to assess TILs in residual disease after neoadjuvant therapy and in carcinoma in situ: A report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer. *Seminars in Cancer Biology*, *52*. https://doi.org/10.1016/j.semcancer.2017.10.003
4. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).
5. Saltz, J., Gupta, R., Hou, L., Kurc, T., et al. 2018. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell reports, 23(1), pp.181-193
6. Siegel, R. L., Miller, K. D. and Jemal, A. (2019), Cancer statistics, 2019. CA A Cancer J Clin, 69: 7-34. doi:10.3322/caac.21551
7. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In Proceedings of International Conference on Learning Representations, 2015.

8. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-First AAAI Conference on Artificial Intelligence 2017 Feb 12.

Items 1-4

4. **Human participants research- N/A**
5. **Vertebrate animals research- N/A**
6. **Potentially hazardous biological agents- N/A**
7. **Hazardous chemicals, activities, and devices- N/A**

NO ADDENDUMS EXIST