**II. Specific Aims**

The long-term goal of our project is to discover which type(s) of cancer data and/or combinations of data are best for predicting the length of survival of a cancer patient after diagnosis. We also hope to discover the best ML algorithms for making these predictions. To accomplish this, we will:

1. Identify which types of data are most predictive of patient prognosis when using ML algorithms for these predictions, thus helping to inform decisions about data collection.
2. Evaluate whether predicting prognosis based on molecular data leads to better accuracy than predicting using clinical data. In addition, we want to evaluate whether combining these two data types is better than using either data type alone. This will show if the millions of dollars spent on producing molecular data truly yields benefits for cancer patient prognosis.
3. Combine all of the data types together and evaluate ways to standardize the data so that the ML algorithms can best deal with the variety and size of these combined data sets.

**III. Background and Significance**

Large volumes of recently produced clinical and molecular data provide unprecedented opportunities for physicians and researchers to perform precision medicine and make accurate cancer outcome predictions using ML[4](https://paperpile.com/c/HdROHr/NajQ). “Machine learning is not new to cancer research”[3](https://paperpile.com/c/HdROHr/NM0N). However, ML in cancer prognosis is still relatively new[3](https://paperpile.com/c/HdROHr/NM0N), and there seems to be incredible potential for ML in cancer prognosis with the rapid development of genomic, proteomic and imaging technologies. Molecular biomarkers such as somatic mutations, tumor proteins, and the chemical environment of the tumor may serve as powerful predictive indicators[3](https://paperpile.com/c/HdROHr/NM0N).

With this incredible potential, we anticipate that more researchers will study cancer prognosis through ML. One recurring issue, however, is that when researchers want to predict cancer patient outcomes and survival rates using molecular or clinical data, they typically have to prioritize what type of data to collect due to cost. Furthermore, few researchers have studied which types of molecular data and/or which combinations of such data provide the best predictions. This makes it imperative that we more fully understand whether certain ML algorithms work better for specific types of molecular data.

A few researchers have already begun exploring these questions. One group tested ML with DNA methylation, mRNA, microRNA and protein expression data with four cancer types to find which data was the most predictive. They found that combining molecular data with clinical variables greatly improved the accuracy of predictions. However, this group didn’t test multiple ML algorithms[1](https://paperpile.com/c/HdROHr/oQcs). Another group tried testing three ML algorithms to make breast cancer predictions with clinical data. They found that the SVM (Support Vector Machines) algorithm was the most accurate, but they only tested clinical data[5](https://paperpile.com/c/HdROHr/1RZR). In addition, my mentor, Dr. Piccolo, performed a study on glioblastoma multiforme prognosis with ML. He discovered a significantly more accurate technique for prognosis, but he didn’t vary the data type.[2](https://paperpile.com/c/HdROHr/fSvP)

In previous studies done in our lab, we’ve developed preliminary data that we hope to add to. We tested three cancer types with three ML algorithms and seven molecular data types. We found that in general, protein expression and gene expression data were the most predictive of length of survival after diagnosis, and DNA mutation data was the least. However, our studies didn’t test many cancer types, and they only worked as a pilot analysis. We also didn’t explore the impact of combining data sets.

My project would build upon these previous studies. However, instead of only testing a few cancer types with one ML algorithm, or testing multiple ML algorithms with one data type, I would test these in combinations. I would also combine molecular data with clinical data as stated in my aims. I plan to use 6 different molecular data sets with 5 different ML algorithms to test 14 different types of cancer (significantly more than what we have previously done). I also intend to do a more comprehensive analysis then the previous studies cited. My project will provide the scientific community with much-needed understanding regarding the algorithm and type and/or combination of types of data that yield the best cancer prognosis results.

Dr. Piccolo already has extensive expertise in cancer research, ML, and managing data sets. He developed a software toolbox called ML-Flex, which allows researchers to combine and standardize multiple data sets while also incorporating third party ML algorithms. In essence, he already developed the perfect tool for our project[6](https://paperpile.com/c/HdROHr/gge5).

Ultimately, as a result of our study, we anticipate that cancer prognosis will be more efficient, more accurate, and cheaper. Physicians will be able to better stratify patients into risk groups and provide the appropriate treatments necessary. The computational predictive methods used in research will improve as researchers use the right data with the correct software, and our understanding of the fundamental cancer survival rate factors will increase as we learn which types of data have the best cancer markers.

**IV. Research Design and Methods**

All of my work will be performed on the BYU supercomputer due to the highly intensive computational aspects of this project. I have prior experience working with the supercomputer.

The first thing I will need to do to achieve our long-term goals is to make the data uniform. This involves scaling all data so that it falls into the range -1 to +1. This prevents factors with higher numbers (such as age) from being emphasized too heavily by the ML algorithms. To do this, I will use the ML-Flex tool developed by Dr. Piccolo. This will also involve putting the data into the same file format. To facilitate this, I will use Python modules and an accompanying command line tool called ShapeShifter, which allows researchers to merge, compress, parse, and filter file formats from one to the other. I previously worked on the coding and development of this tool, so I have significant experience with it.

After standardizing the data, I will test each specific ML algorithm on each type of data. The algorithms we will use include Support Vector Machines[7](https://paperpile.com/c/HdROHr/kXYn), Random Forests[8](https://paperpile.com/c/HdROHr/2HLG), Artificial Neural Networks (Residual Networks)[9](https://paperpile.com/c/HdROHr/9vDY), k-Nearest Neighbors[10](https://paperpile.com/c/HdROHr/63cP), and Gradient Boosting Machines[11](https://paperpile.com/c/HdROHr/gwqN). All five of these algorithms are state-of-the-art and show promise in biomedical research, but have not been thoroughly tested in this context.

The molecular data types will be somatic mutations, DNA methylation, RNA expression, microRNA expression, protein expression, copy number variants. We will get the data for all of this from The Cancer Genome Atlas[12](https://paperpile.com/c/HdROHr/dQhY). One can access this database at <https://cancergenome.nih.gov/>. Dr. Piccolo has worked extensively with this data as well as with these ML algorithms in his lab. We will also test clinical data’s ability to allow ML algorithms to predict cancer patient outcome. The clinical data we plan to use includes age at diagnosis, sex, stage, histological features, etc. It will be interesting to look at data types separately, but it will likely be even more interesting to look at them in combinations. One data type in isolation may not be great at predicting cancer patient outcomes but combined with another data type could give great results.

The cancer types we will test include Glioblastoma, Serous Ovarian Carcinoma, Lung Adenocarcinoma, Lung Squamous Cell Carcinoma, Breast Carcinoma, Prostate Adenocarcinoma, Skin Melanoma, Colon Adenocarcinoma, Bladder Carcinoma, and Sarcoma. As we perform our ML benchmark evaluation we will measure what the program actually predicts in terms of the length of survival after diagnosis. Each benchmark evaluation will include a comparison of the program’s predictions with the actual outcome. Each combination of ML algorithm, classification scheme, and biological data type will be run 10 times to get a more detailed picture of the algorithm’s performance. We will use the area under the receiver operating characteristic curve (AUROC) as a measure of how accurately each iteration is able to predict the given class.

We will test the combinations of data in the following way: First, we will use one ML algorithm and test it on each type of data for each cancer type. Each type of data and cancer type will be run 10 times using cross validation. We will then test the next ML algorithm following the same procedure until all of our ML algorithms, data types, and cancer types have been thoroughly tested on their own.

Afterwards we will begin to mix multiple types of data together in the following way. First, we will test each type of molecular data combined with clinical data. We will follow the previous procedure which includes testing each algorithm with each cancer type and data type ten times. Then, whichever form of molecular data performs the best in combination with clinical data, we will combine with all other forms of data and perform our benchmark analysis. Finally, we will combine all of the types of data together to see which algorithm does the best job of supporting the mixture of data types. This will help us determine if having more types of data ultimately improves results. It will also be interesting to see which ML algorithm can best handle the huge combination of data. One key aspect of this analysis will be to evaluate techniques for “standardizing” the data so that different types of data are more comparable.

**New Methodology/Novel Concepts, approaches, tools or technologies:** A new methodology and approach that has never been used before will be to combine the data. We will do this in two different ways: 1) merge the data and 2) apply the algorithm to each type of data separately. (If time permits, I will use dimensionality reduction and then combine eigenvectors).

**Potential Difficulties and Limitations:** There are many types of data to test and many ML algorithms to compare. Therefore, the workload is a major obstacle. To overcome this obstacle, I will do a pilot analysis with a small set of data, then scale it up to all combinations of data and algorithms. Then we will have the interpretation of the results. The types of data are also not standardized. Therefore, I will need to put significant work into making the data standard.

**Time Table:** The time table for our project is as follows: From April 27 to May 4, I will finish standardizing all the data types. Most of the data is ready to go, but some of it will still need to be normalized. Then from May 4 to May 11, I will do a "pilot phase" in which I test 2 algorithms on 2 data types for 2 cancer types to develop the computational pipeline. Then from May 13 to May 25, I will execute my analysis at full scale on the Supercomputer. This benchmark analysis will include the 14 cancer types, the 6 data types, and the 5 algorithms separately. Then from May 27 to June 8, I will merge the data and perform a benchmark of each type of molecular data combined with clinical data. Then from June 10 to June 22, I will determine which type of molecular data combined with clinical data performs best. I will then take that combination of data sets, and combine it with each of our other molecular data sets and perform a benchmark analysis. Then from June 24 to August 6, I will merge all the data sets together and test each algorithm and cancer type with the merged data set. From August 6 to August 9 (or the end of the fellowship), I will begin writing up a paper on the results from our experiment.

**Hazardous Procedure/Vertebrate Animals or Human Subjects:** We don’t anticipate any Hazardous Procedures associated with our project because all work will be completed in a dry lab. There will be no need to use vertebrate animals or human subjects.

VI. Literature Cited:

1. [Yuan, Y. *et al.* Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.* **32**, 644–652 (2014).](http://paperpile.com/b/HdROHr/oQcs)

2. [Piccolo, S. R. & Frey, L. J. Clinical and molecular models of glioblastoma multiforme survival. *Int. J. Data Min. Bioinform.* **7**, 245–265 (2013).](http://paperpile.com/b/HdROHr/fSvP)

3. [Cruz, J. A. & Wishart, D. S. Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Inform.* **2**, 117693510600200030 (2006).](http://paperpile.com/b/HdROHr/NM0N)

4. [Huang, C., Mezencev, R., McDonald, J. F. & Vannberg, F. Open source machine-learning algorithms for the prediction of optimal cancer drug therapies. *PLoS One* **12**, e0186906 (2017).](http://paperpile.com/b/HdROHr/NajQ)

5. [Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M. & Razavi, A. R. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *Journal of Health & Medical Informatics* **4**, 1–3 (2013).](http://paperpile.com/b/HdROHr/1RZR)

6. [Piccolo, S. R. & Frey, L. J. ML-Flex: A Flexible Toolbox for Performing Classification Analyses In Parallel. *J. Mach. Learn. Res.* **13**, 555–559 (2012).](http://paperpile.com/b/HdROHr/gge5)

7. [Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).](http://paperpile.com/b/HdROHr/kXYn)

8. [Ho, T. K. Random decision forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* **1**, 278–282 vol.1 (1995).](http://paperpile.com/b/HdROHr/2HLG)

9. [van Gerven, M. & Bohte, S. Editorial: Artificial Neural Networks as Models of Neural Information Processing. *Front. Comput. Neurosci.* **11**, 114 (2017).](http://paperpile.com/b/HdROHr/9vDY)

10. [Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **46**, 175–185 (1992).](http://paperpile.com/b/HdROHr/63cP)

11. [Li, C. A Gentle introduction to gradient boosting. *URL: http://www. ccs. neu. edu/home/vip/teach/MLcourse/4\_ boosting/slides/gradient\_boosting. pdf* (2016).](http://paperpile.com/b/HdROHr/gwqN)

12. [Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130**, 453–459 (2017).](http://paperpile.com/b/HdROHr/dQhY)