**Final Project Report**

**Denna Saji, Prem Almeida, Uma Dixit, Yogesh Chaudhari, Sheng Li**

CSCI 8950 Machine Learning

Department of Computer Science

University of Georgia

Athens, GA 30605

**Abstract**

Prognosis prediction is an extremely crucial step when treating a patient with breast cancer because accurate prediction can help the patients and the doctors to decide if any further treatments are necessary and are the medical costs involved worth it. In this study, we attempt to reassess a multimodal deep neural network (MDNN) method and improve the performance of the model.We have improved on the model by feeding it data that is extracted from an autoencoder feature selector. A new modality in the form of RNA ZScores was also added to the MDNN. We compared the benchmark results from the previous study with different models that have variations in the type of autoencoder used and number of modalities involved.

**1 Background**

* 1. **Literature Review**

With breast cancer being one of the most highly aggressive cancers, especially in females, it is crucial that a method to predict the survivability of the patient accurately. The American Cancer Society found case estimates to be near 250,000 of invasive breast cancer and about 40,000 death approximations in 2017. This disease is mainly quantified by varied molecular feature, clinical behavior, morphological appearance and disparate response to therapy. Thus, attempting to increase the accuracy of prognostication is critical in helping patients learn more about their life expectancy as well as helping clinicians make more informed decisions for the benefit of the patient.

With the recent developments in gene expression patterns analysis and microarrays there have been many contributions to the overall knowledgebase of molecular signatures of breast cancer that are based purely on gene expression patterns. One of the most influential studies in this field shows breast cancer prognosis through gene expression profiles in which they identify 70 gene signatures from 98 primary breast cancer patients by clustering the gene expression profile data and correlating them with the values they calculated from the prognostics.

The dataset which was used in the original paper consists of 25,000 gene expressions per patient however for the purpose of this study we have chosen the top 1227 gene expressions

* PCA
* PGM

**2 Methods and Materials**

**2.1 Materials**

**-** Metabric now has 2500 but we are using 1904

There are 1904 common patients across Clinical, CNA, Gene Expression, RNA ZScores

**2.2 Feature Selection**

- MRMR (still following the original paper, table 2) CNA 200, Gene 400, Clinical 19

- Denoising autoencoder

- Deep multimodal autoencoder

- CNA 900, clinical data is 19, gene expression 1,227, RNA 1227

**2.3 Prediction Models for Single and Multidimensional Dataset**

- Original paper dataset and find the accuracy

- Adding a multimodal with dataset and finding the accuracy

- changed the parameters and added RNA

**2.4 Experimental Design**

- Make a block diagram (add autoencoder block separately, MRMR)

**2.5 Other**

**3 Results**

**3.1 Comparison**

**3.2 Validation**

**4 Discussion and Conclusion**

**References**

[1] Extracting High-Quality Features From Biomedical

Datasets Using Multimodal Autoencoders (Dmitry Kazhdan)

[2] A Multimodal Deep Neural Network for Human

Breast Cancer Prognosis Prediction by

Integrating Multi-Dimensional Data (Dongdong Sun, Minghui Wang, and Ao Li)

<https://towardsdatascience.com/denoising-autoencoders-explained-dbb82467fc2>

Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters, Russnes H.G., Lingjaerde O.C., Borresen-Dale A.-L., Caldas C. (2017)  *American Journal of Pathology*,  187  (10) , pp. 2152-2162.