CS137 Project: Addressing Data Sparsity in Male Breast Cancer Diagnosis through Transfer Learning

Johnny Lai, Susie Li, Chelsie Wei

Nov. 11, 2023

1 Introduction

1.1 Motivation

This research seeks to investigate the application of transfer learning in the context of breast cancer diagnosis; particularly addressing the sparsity of genomic and proteomic data coming from male patients with breast cancer. Currently, less than 1% of all breast cancers occur in men, leading to a significant disparity between the publicly available genomic data collected from male patients versus female patients [Hopkins Medicine]. This is observable in the dataset available under NIH's Genomic Data Commons Portal (incorporates genome sequencing data from TCGA and GENIE) which comprises a substantial 6,124 female cases but only 45 male cases [NIH GDC].

While this wealth of data is advantageous for treating the majority group affected by breast cancer, it presents a significant hurdle when applying deep learning models - especially considering that male and female patients have inherently different cancer genome profiles [Zhu, Boutros 2021]. This underscores the importance of exploring applicable methods which can help address the sparsity of data and enhance the overall accuracy of deep learning models for this problem domain. Thus, we propose exploring the application of transfer learning, using data collected from female patients as the source domain and data collected from male patients as the target domain, to address the sparsity of data for male patients and improve the use of deep learning for classifying breast cancers in male patients.

1.2 Project Overview

Breast cancer diagnosis with deep learning has traditionally focused on genomics, histological, and imaging data [Nasser, Yusof 2023]. In this study, we will specifically concentrate on the use of proteomics and genomics for our input data. While we are still narrowing down our specific data sources, we are considering a few potential candidates. Our Genomic data will likely be sourced from querying the NIH Genomic Data Commons Portal, specifically for data from The Cancer Genome Atlas (TCGA) project and the Genomics Evidence Neoplasia Information Exchange (GENIE) project [NIH GDC Query]. We will be focusing on genomic data that informs on the composition of RNA as well as the exomes of cancerous breast specimens. The data we are focusing on comes in various formats (eg .maf, .bam files) - mainly consisting of sequencing reads, single nucleotide variation (SNV), and transcriptome profiling. These files can be parsed using the `biopython` libary, to identify genetic regions of interest.

Provided we are examining the use of transfer learning for breast cancer diagnosis, the problem

domain will concern the classification of breast cancer types. Based on the current data sources, the predictive labels will take the form of string labels that identify the cancer type associated with the provided genomic profile. These include (but are not limited to) more common cancer types like lobular carcinomas and ductal carcinomas as well as rarer breast cancer types like epithelial neoplasms and mucinous neoplasms. Considering the relatively substantial size of the input genomic data, we will also be considering alternative pre-processed datasets that help lessen the curse of dimensionality. For instance, some pre-processed datasets have condensed proteomic data to protein marker metrics classically indicative of breast cancer - such as HER2, ESR1, and PGR [Ösz, Lánczky, Győrffy 2021] e.g. in the UK Kaggle Dataset.

To evaluate the effectiveness of transfer learning, we will first partition the dataset into data obtained from male patients and that obtained from female patients. We will then develop preliminary versions of the tested models which can utilize processed features from the input genomic / proteomics data to predict the respective patient cancer types. Two trial types will be conducted: in the first, we will train and test the models exclusively on the genomic data collected from male patients. The performance metrics collected from these models will serve as the basis for comparison on whether transfer learning actually improves model performance. In the latter trial, we will leverage transfer learning - i.e. the models will be initially trained on data from female patients before extending training and testing to data from male patients [further details in Background section]. The comparative analysis between the two trials will examine differences in their performance metrics (e.g. their learning curves, accuracy, loss, precision, AUC, etc.) to provide insight on the impact of transfer learning on this problem domain.

1.3 Models

Since our project focuses on addressing gender data inequality in breast cancer research using transfer learning, we plan to employ and evaluate a variety of models that are suited for structured tabular data. Our primary choice includes classic Deep Neural Networks (DNNs). We will fine-tune the feedforward neural network architecture to identify intricate patterns in the female breast cancer dataset, which can be crucial for transfer learning to the male dataset. Additionally, we will employ and experiment with transformers. Although primarily used in natural language processing, TabNet has shown promise in handling tabular data, especially due to its proficiency in identifying long-range dependencies, which could be beneficial in understanding relational aspects of the data. Alongside these neural network-based models, we also plan to use XGBoost, a highly efficient and scalable implementation of gradient boosting. XGBoost is particularly renowned for its performance in structured data scenarios, making it an excellent choice for our tabular datasets. It's known for its speed and model performance, especially in medical datasets where precision is crucial [please see Section 3 Background for more details on the models].

1.4 Metrics of Success

The success of the project will assess the impact of transfer learning on the models' performance (specifically, performance on the sparse data from male patients). The baseline for metric comparison will come from the classic deep neural network's performance without transfer learning. Through the application of transfer learning to the three different models - the basic deep neural network, XGBoost, and Transformers - we aim to discern improvements in test accuracy, AUC, recall, and precision. By comparing the performance of models with and without transfer learning, we seek to identify variations in effectiveness.

We will also inspect how transfer learning affects the three different models differently, hopefully providing insights into the nuanced impact of transfer learning on the various modeling architectures that we will be studying. We anticipate that this study will yield some understanding of how transfer learning can enhance the performance of deep learning models in classifying breast cancers - contributing to the development of more effective oncological diagnostic tools.

1.5 Hypothesis

Our hypothesis posits that employing transfer learning, specifically by fine-tuning a model initially trained on female data, will enhance the accuracy of our baseline model that was trained exclusively on male data. This improvement in accuracy will be quantitatively assessed using the Area Under the Receiver Operating Characteristic (AUROC) metric.

2 Related Work

Male Breast Cancer (BC) represents a relatively small fraction of the total global BC cases, comprising just 1%; the rarity of male BC meant many of its data had to be extrapolated from female BC patients [Gucalp, et al]. Although male and female BC have similar epidemiological compositions, emerging research indicates that they may not exhibit identical molecular and clinicopathologic characteristics [Gucalp, et al].

In a clinical setting, both male and female BC are notable for their heterogeneity nature – the prognosis and treatment of each breast cancer tumor depend on a unique combination of intrinsic (genomic and proteomic) labels and the extrinsic, micro-environment of the cancer cells [Lüönd, Fabiana et al]. Intrinsic and extrinsic markers together determine the specific subtype of BC, and such heterogeneity proved to be difficult in pinning down patient-specific treatment due to the variety of causal factors [Dagogo-Jack and Shaw]. Protemic and Genomic markers, especially, can provide clues to identify BC subtypes [Neagu, et al]. As early as 1990s, some researchers were able to identify groups of genes prominent in BC families, BRCA1 and BRCA2 [Ford, et al.]. Recent advances in proteomic technologies allowed biologists to study proteomics and genomic markers in BC patients altogether [Tyanova, et al]. Due to the difficult process in determining the subtype of BC in male and females, as described above, many turned to machine learning as a predictive tool for its ability to minimize false positive rate, subjective treatments, etc. [Hiramatsu, et al.].

Our research question differs from existing research since to our knowledge there has been no published or ongoing studies addressing the data inequality (sparsity) between female and male breast cancer data using transfer learning. While deep learning has been used for cancer sub-type prediction based on genomic data [Chen, Yang, et al.], few of these studies have focused on transfer learning for addresssing the aforementioned data sparsity. Although transfer learning has been applied in oncology to address certain data scarcity and data inequality issues (Toseef, 2022), its source and target domains are different ethnicities, while we focus on breast cancer for male and female as the target and source domains. Past literature on Breast Cancer prediction via transfer learning dealt mostly with image data, for example using pictures of ultrasound, mammography, and CT [Ayana, et al.]. Some used transfer learning to detect certain types of BC in females and males by training on histopathological images (microscopic examination of tissues), however none to compare between female and male BC traits [Mahmud, et al.]. Relatively few studies looked at textual/tabular data of BC patients, specifically through perspectives of proteomics and genomics. It is important to note that transfer learning in the visual domain is characteristically

different from transfer learning done in text based domains because of the different pre-training techniques and distribution shifts we might concern ourselves with [Neyshabur, et al.].

3 Background

In this section, we provide an overview of Transformers, XGBoost, and transfer learning, which are concepts integral to the methodology and design of this project.

3.1 Transformers

Transformers are a type of neural network architecture that differ from traditional neural network architectures by relying on self-attention mechanisms. These mechanisms allow the model to weigh the importance of different parts of the input data differently, making transformers good at handling sequential data and capturing long-range dependencies. Their scalability and efficiency in processing sequences have led to their adaptation beyond language tasks, including applications in tabular data [Transformers, NVIDIA].

3.2 XGBoost

XGBoost (eXtreme Gradient Boosting) is a scalable and efficient implementation of gradient boosting machines. It is a decision tree-based ensemble machine learning algorithm that uses a gradient boosting framework. XGBoost is particularly useful for structured data like classification and regression tasks due to its advantage in handling large and complex datasets efficiently and providing robust results. It works by building a series of decision trees, each correcting the errors of the previous ones, and combining their outputs to make more accurate predictions [XGBoost, NVIDIA].

3.3 Transfer Learning

Transfer learning is a technique for leveraging pre-existing knowledge from one task to improve learning in a related but different task, especially useful in scenarios of data scarcity. An overview of the workflow of a transfer learning task involves 1) training on the source task, 2) transfer of knowledge across domains, and 3) training/fine-tuning on the target task. Specifically, we begin by training a classic deep neural network on the source task (female breast cancer data), and the model learns the weights of the model parameters that minimize a loss function specific to this task. Then, the knowledge (weights, layers, and/or features extracted) is transferred to the target task (predicting male breast cancer). There are multiple approaches to this step, such as using the model from the source task as a feature extractor for the target task, for example using the output of a layer from the source model as an input for the new model trained on the target task. Finally, we train and fine-tune the transferred model on the target task data (male breast cancer patient data). Depending on the training and testing performance, the last step can be iterated and applied to all or some selected layers of the models [Transfer learning].

4 References

[1] Ayana G, Dese K, Choe SW. Transfer Learning in Breast Cancer Diagnoses via Ultrasound Imaging. Cancers (Basel). 2021 Feb 10;13(4):738. doi: 10.3390/cancers13040738. PMID: 33578891; PMCID: PMC7916666.

[2] Behravan H, Hartikainen JM, Tengström M, Kosma VM, Mannermaa A. Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. Sci Rep. 2020 Jul 6;10(1):11044. doi: 10.1038/s41598-020-66907-9. PMID: 32632202; PMCID: PMC7338351.

[3] Dagogo-Jack, I., Shaw, A. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 15, 81–94 (2018). https://doi.org/10.1038/nrclinonc.2017.166

[4] Ford, D et al. "Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium." American journal of human genetics vol. 62,3 (1998): 676-89. doi:10.1086/301749

[5] Gucalp, A., Traina, T.A., Eisner, J.R. et al. Male breast cancer: a disease distinct from female breast cancer. Breast Cancer Res Treat 173, 37–48 (2019). https://doi.org/10.1007/s10549-018-4921-9

[6] Lüönd, Fabiana et al. "Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression." British journal of cancer vol. 125,2 (2021): 164-175. doi:10.1038/s41416-021-01328-7

[7] Merritt, Rick. "What Is a Transformer Model?" NVIDIA Blog, 25 Mar. 2022, blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/. Neagu, Anca-Narcisa et al. "Proteomics and its applications in breast cancer." American journal of cancer research vol. 11,9 4006-4049. 15 Sep. 2021

[8] Neyshabur, Behnam, Hanie Sedghi, and Chiyuan Zhang. "What is being transferred in transfer learning?." Advances in neural information processing systems 33 (2020): 512-523.

[9] Nvidia. "What Is XGBoost?" NVIDIA Data Science Glossary, www.nvidia.com/en-us/glossary/data-science/xgboost/.

[10] Sahin C, Ucpinar BA, Mut DT, Yilmaz O, Ucak R, Kaya C, Tanik C. Male Breast Cancer with Radiological and Histopathological Findings. Sisli Etfal Hastan Tip Bul. 2020 Aug 22;54(3):375-379. doi: 10.14744/SEMB.2020.01643. PMID: 33312039; PMCID: PMC7729722.

[11] Tyanova, S., Albrechtsen, R., Kronqvist, P. et al. Proteomic maps of breast cancer subtypes. Nat Commun 7, 10259 (2016). https://doi.org/10.1038/ncomms10259

[12] "What Is Transfer Learning? A Guide for Deep Learning | Built In." Builtin.com, builtin.com/data-science/transfer-learning. Accessed 12 Nov. 2023.

[13] Ősz, Á., Lánczky, A. Győrffy, B. Survival analysis in breast cancer using proteomic data from four independent datasets. Sci Rep 11, 16787 (2021). https://doi.org/10.1038/s41598-021-96340-5

[14] Zhu, Chenghao, and Paul C Boutros. "Sex Differences in Cancer Genomes: Much Learned, More Unknown." Endocrinology, vol. 162, no. 11, 17 Aug. 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC8439393/, https://doi.org/10.1210/endocr/bqab170. Accessed 12 Nov. 2023.

[15] Yuya Hiramatsu, Chisako Muramatsu, Hironobu Kobayashi, Takeshi Hara, Hiroshi Fujita, "Automated detection of masses on whole breast volume ultrasound scanner: false positive reduction using deep convolutional neural network," Proc. SPIE 10134, Medical Imaging 2017: Computer-Aided Diagnosis, 101342S (3 March 2017); https://doi.org/10.1117/12.2254581

[16] Mahmud, Md Ishtyaq, et al. "A Deep Analysis of Transfer Learning Based Breast Cancer Detection Using Histopathology Images." arXiv.Org, 11 Apr. 2023, doi.org/10.48550/arXiv.2304.05022.

[17] Chen, R., Yang, L., Goodison, S., Sun, Y. (2020). Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. Bioinformatics (Oxford, England), 36(5), 1476–1483. https://doi.org/10.1093/bioinformatics/btz769

[18] Toseef, M., Li, X., Wong, K.-C. (2022). Reducing healthcare disparities using multiple multiethnic data distributions with fine-tuning of transfer learning. Briefings in Bioinformatics, 23(3), bbac078. https://doi.org/10.1093/bib/bbac078