Enhancing Breast Cancer Diagnosis: Transfer Learning Across Racially Sparse Genomic Data

Johnny Lai, Susie Li, Chelsie Wei

Dec. 20, 2023

1 Introduction

1.1 Motivation

This research seeks to investigate the application of transfer learning in the context of genomic based breast cancer diagnosis, particularly addressing the sparsity of available genomic data coming from ethnic minority patients with breast cancer. A recent survey of data availability across various cancer projects (including TARGET, OncoArray, and TCGA) revealed that an overwhelming majority ($\sim$ 92%) of available genomic cancer data was sourced from white patients, while the remaining 7% of data comprised Black / African American patients, Asian patients, and other ethnic minorities [19] [*Fig. 1*] .

This data inequality is analogously observable under breast cancer specific genomic data publicly available under NIH's Genomic Data Commons Portal (which includes genome sequencing data collected from cancerous breast samples under the TCGA project). This data source comprises a substantial 733 cases collected from white patients, but only 175 cases collected from black or African American patients and only 61 cases collected from Asian patients [21] [*Fig. 2*].

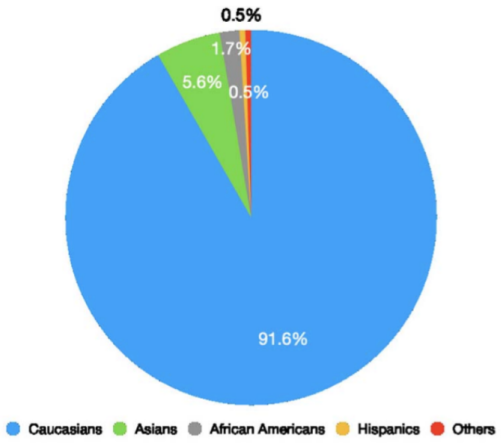

Figure 1: Survey Data Distribution

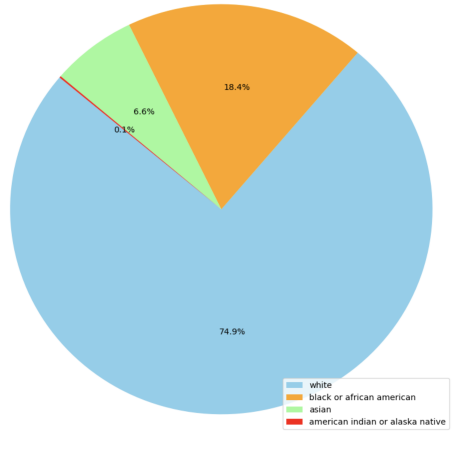

Figure 2: NIH GDC Data Distribution

While this wealth of data is advantageous for diagnosing and treating the majority group affected by breast cancer, it presents a significant hurdle when applying deep learning models - especially considering that the prevalence of genetic polymorphisms and mutations present racial differ-

ences across various cancer types [40]. This disparity in available breast cancer genomic data presents a plethora of concerns considering the possibility of skewing machine learning models with racial biases. 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 deep transfer learning, using data collected from white patients as the source domain and data collected from ethnic minorities as the target domain, to address the sparsity of data for ethnic minorities and improve the use of deep learning for classifying breast cancers in non-white patients.

1.2 Data Source

Breast cancer diagnosis with deep learning has traditionally focused on genomics, histological, and imaging data [20]. In this study, we specifically concentrated on the use of CNV genomics (copy number variations) for our input data. That is to say the objective for each of our tested models is to classify the specific subtype of breast cancer (e.g. lobular v.s. ductal) associated with the CNV data of a given cancerous sample. As aforementioned, our genomic data was sourced by querying the NIH Genomic Data Commons Portal, specifically for breast cancer CNV data from The Cancer Genome Atlas (TCGA) project [21].

As denoted in its name, Copy Number Variation files are obtained by high-throughput sequencing the DNA material of cancerous specimens and comparing it against a reference genome to identify differences in gene copy numbers. In the data obtained from the NIH GDC portal, the CNV data sample files contain copy number variation data on 60,623 unique genes (providing the chromosome number, start index, end index, copy number median and range of each genomic mutation). The outlined genomic variations in each file provide details on the magnitude and locus of specific deletions (loss of genetic material) or duplications (extra copies of genetic material) in the cancerous sample - often characteristic to the specific subtype of breast cancer that is associated [5].

The data that was utilized for this study stored the CNV data of each sample in the form of a .tsv (tab separated variable) file which could be read into a pandas dataframe for further processing. After downloading the selected CNV datafiles, the patientID for each sample had to be cross queried with the clinical data in the NIH GDC database to obtain the corresponding breast cancer subtype of each sample. These include (but are not limited to) more common breast 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, Principal Component Analysis was leveraged in some cases to reduce the dimensionality of the 60,623 genes in each data point.

1.3 Experimental Overview & Models

In short, the purpose of this study was to explore the use of transfer learning methods for mitigating effects of the racial data gaps previously observed in genomic breast cancer data. For the scope of this study, two techniques of transfer learning were primarily observed: (i) fine tuning transfer learning, and (ii) feature-based transfer learning. Three Deep Learning models were tested under these transfer learning methods: Classic DNN, LSTM, and Transformer.

To benchmark performance, we ran an additional baseline trial for each model using no transfer

learning techniques. For this baseline trial, the selected model was only trained on data from white patients before being tested on the target domain. For further benchmark performance of the deep learning models, we ran two additional baseline models (SVM and XGBoost), both of which do not use deep neural networks. The comparative analysis of the peformance metrics (namely accuracy and loss) of the different model runs provided insight into the impact of transfer learning on this problem domain.

1.4 Metrics of Success

The success of this study was assessed via the impact of transfer learning on the models' performance (specifically, performance on the sparse data from ethnic minority patients). The baselines for metric comparison primarily came from model performance without transfer learning. Through the application of transfer learning to the three different deep learning models - DNN, LSTM, and Transformers - we discerned potential improvements in test accuracy and loss. By comparing the performance of models with and without transfer learning, we seeked to identify variations in effectiveness. We further inspected how transfer learning affected the three deep learning models differently, to hopefully provide insight into the nuanced impact of transfer learning on the various model architectures studied.

As another point of reference, we also collected the performance metrics for the SVM and XGBoost model, to abstract how deep learning holistically performs in comparison to other machine learning methods on this problem domain. We anticipated that the study would yield some understanding of how transfer learning has potential to improve deep learning performance in classifying breast cancers - contributing to the development of more effective oncological diagnostic tools.

1.5 Hypothesis

Our hypothesis proposed that both fine-tuning and feature-based transfer learning (by exposing target data during training) would enhance the accuracy of our baseline models trained exclusively on the source data. This improvement in accuracy was quantitatively assessed using the model accuracy metric.

2 Related Work

2.1 Machine Learning for Cancer Diagnosis

In recent years, researchers have identified Machine Learning (ML) tools to be extremely useful in cancer prognosis and diagnosis because of their ability to detect key features within complex data sets. Namely, Artificial Neural Networks, Bayesian Networks, Support Vector Machines (SVM), and Decision Trees are widely used in cancer research as predictive tools [13].

Liew et al. [15] combined Deep Learning and XGBoost, a gradient-boosted decision tree [23], to classify breast cancer histology images into binary malignant and benign results. They identified a way to avoid over-fitting caused by fine-tuning by replacing a Fully-Connected layer of CNN with XGBoost, where tree boosting had been effective in reducing over-fitting issues. Zhang et al. [38] used Copy Number Variation of 10 tumor data sets from TCGA. They utilized XGBoost as classifier and Extra Tree for dimension reduction. Results showed 0.8913 in overall accuracy in predicting tumor tissue of origin. Additionally, Yu Daping et al. [34] used XGBoost classifier to predict lung

cancer via, again, Copy Number Variations and identified places on one's chromosomes that are more easily subject to mutations like amplifications and deletions. Original contributors [4] of the XGBoost library noted in their 2016 paper that XGBoost is still successful when handling sparse data.

SVM had proved useful involving feature selection among discrete and subtle datasets, exemplified by the case of genomic data [10]. Segal et al. [26] used linear SVM to predict adult Soft Tissue Sarcoma subtypes from specific genes and found SVM's high sensitivity able to give out accurate subtype prediction. They trained with hold-one-out scheme, in which each training set is all but one group of sample, with the lone sample used in testing. Results confirmed SVM was able to identify trend between genomic data and tumor subtypes.

These studies were done on linear models without Deep Learning or with limited Deep Learning. Their successes showed 1). the genomic, namely Copy Number Variation, basis of tumor and cancer types and subtypes which allowed for accurate prediction from ML models and 2). various limitations, including over-fitting and sparse data, that led to model adjustments and outlier identification.

2.2 Transfer Learning

Transfer Learning (TL) is notably used to solve problems arising from processes of data collection. Realistically, data are not always readily available and may not be easily obtained. Additionally, data may not all belong in the same domain and feature space [25]. Weiss et al. [30] and Lu et al. [16] have conducted extensive survey on real-world applications of various Transfer Learning methods, noting the difficulty in training sparse, very little labeled data that belong in the same domain. Unlike traditional supervised and unsupervised learning [38], [26], TL assumes that domains, distributions, and tasks can be different [25].

Prior to Deep Learning's popularity, TL were mostly done on traditional machine learning methods [28]. In a 2014 paper Gönen & Margolin [9] outlined a Transfer Learning method using Bayes, involving kernel-based dimensionality reduction models, where the authors used it to successfully predict cancer gene mutations. It is also worth to mention that there exists the semi-supervised classification method which uses both labeled and unlabeled data, with the end goal of addressing insufficient "feature/label" pairs in the dataset [39]. Semi-supervised learning does not necessarily use Transfer Learning, but it deals with similar problem in limited data size. Zemmal et al. [36] and Al-Azzam et al. [2] both navigated various semi-supervised classification models that use relatively little training data to achieve high accuracy in cancer prediction. For example, Al-Azzam et al. used a S3VM (Semi-supervised Support Vector Machine) classifier.

However traditional, non-deep learning methods do not depend heavily on data/sample size, which is the case for Deep Transfer Learning (DTL). There seems to be an increasing emphasis to use TL on deep learning models, especially on problems with non-linear datasets [11]. Soekhoe et al. finds that, though TL helps with very small & non-linear datasets, it works especially well when freezing only the initial layers of the Neural Net the smaller the dataset [27]. This finding is in agreement with Yosinski et al.'s 2014 paper, where it confirmed that higher neurons in the network are more specialized than the first few layers, which tend to learn more generally at first [32]. This paper also emphasizes that the transferability between target and source domains is low when the domains are dissimilar. Of course, *how small* of a data and *how large* is the transferable gap between target and source domains are unique to each situation.

2.3 DTL, Breast Cancer & Racial Diversity

Cancer, by nature, is a heterogeneous disease. Every cancer cell possesses unique genotypes and phenotypes that govern its biological attributes, resulting in distinct cancer subtypes and behaviors [7]. As outlined above in section 2.1 as well as past biology research found Copy Number Variations to associate with breast cancer risk [14] [6]. Copy Number Variations are variations, such as insertions and deletions of genes, in DNA chromosomes that may contribute to mutations and therefore diseases. For example, mutations in BRCA1 or BRCA2 are main contributors to developing heritable breast cancer [14].

In the past many have used DTL to look between various ethnicities and racial groups in clinical fields to investigate the inequality in biomedical data; genome projects mostly collect data from people of European ancestry [29] [8]. One big question is how transferable are biomedical data from one ethnicity/race to another ethnicity/race. When one talks about race in a clinical setting one talks about the difference in cultural, socioeconomic, religious properties in addition to some biological difference [24] [31]. Yedjou et al. [31] found Black women tend to get more aggressive breast cancer, and Ozdemir et al. [24] concludes that people with African ancestry has an increase in stem cell populations in healthy breast cancer. The heterogeneousity of breast cancer deems it to not only have biological basis but also environmental basis, and it varies between race [18] [17].

Some used various Transfer Learning methods, including DTL, to see whether this difference manifests in clinical image data, but none of Copy Number Variations [3] [12], including a survey by Yu et al. [35]. It is important to note that transfer learning in the visual domain is characteristically different from transfer learning done in text based domains [22]. Similarly, there are people who investigated on cancer and Copy Number Variations through TL, but none look between races [37]. An intersection of Deep Transfer Learning, Race, and Breast Cancer genomic data (Copy Number Variations) is missing its literature.

3 Background

In this section, we provide a detailed overview of XGBoost as well as the two transfer learning methods used: Fine-tuning based TL, and Feature-based TL.

3.1 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 [23].

3.2 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 when target data is sparse. An overview of the workflow of a transfer learning task involves: 1) training on the source task, 2) transfer of learned knowledge across domains, and 3) training/fine-tuning on the target task. By

exposing a pre-trained model to the target domain via transfer learning, the model often performs better on the target task.

Under our use case, transfer learning was used a method to potentially improve an existing workflow on using CNV genomic data to predict breast cancer prognoses. Thus, the source task for this study involved breast cancer subtype classification of genomic samples from white patients. The knowledge (i.e. weights, layers, features extracted) were then transferred as a starting point for continual training in the target domain (classifying data from ethnic minority patients). The two forms of transfer learning tested in this study are detailed below. For the purpose of demonstration, these paradigms will be outlined using our DNN model architecture.

Fine-Tuning based Transfer Learning

Fine-tuning based Transfer Learning operates by first training the selected model on source data for a set number of epochs. Training is then continued for another set number of epochs - but in this stage the training data is swapped out for data from the target domain. In this fine tuning stage, all of the model's parameters (e.g. weights, biases, layers) are updated to fit the target domain. Then, the new model is tested on new data from the target domain to benchmark performance of the transfer learning model [*Fig. 3*].

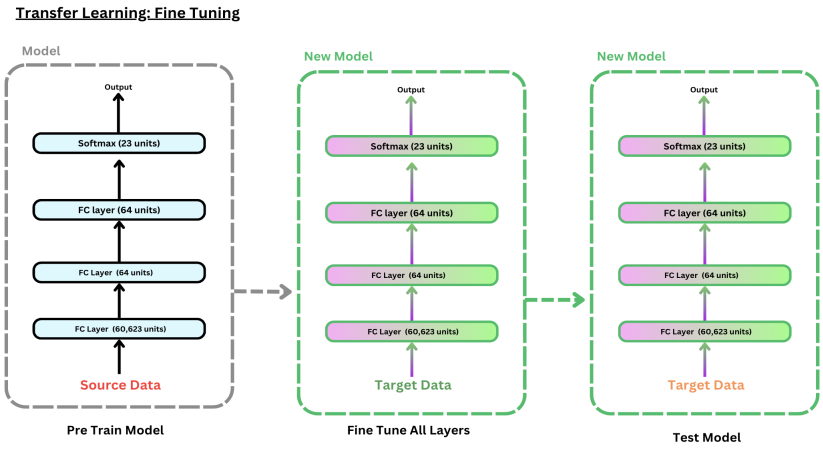

Figure 3: Fine-tuning based TL

Featured-based Transfer Learning

Feature-based Transfer Learning also begins by first training the selected model on source data for a set number of epochs. However, before training is continued a portion of the layer's are "frozen" (usually, all layers / connections except for the last classifier layer). So once the model continues training on the target domain, it can only update parameters found in the layers which are not frozen. Thus, this method is called "feature-based" TL because the final model leverages the pre-learned features from the initial model to continue fine tuning towards the target domain. At the end, the new model is tested on new data from the target domain to benchmark performance of the transfer learning model [*Fig. 4*].

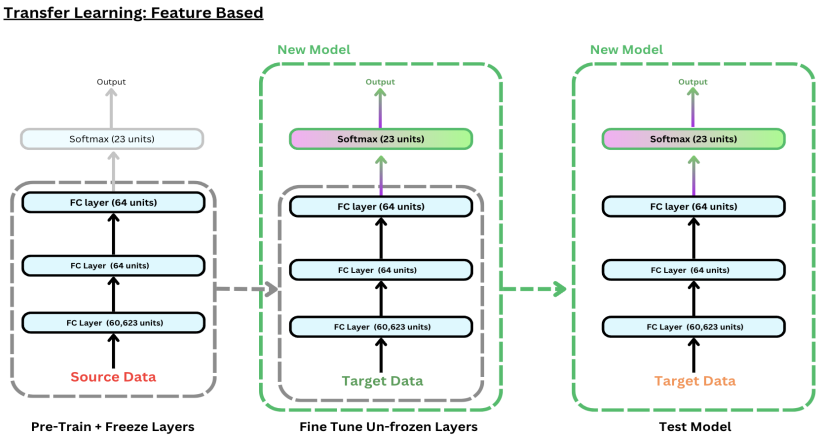

Figure 4: Feature-based TL

4 Methodology

4.1 Data Partitioning

To evaluate the effectiveness of transfer learning, we first partitioned the collected data by race; specifically, using the samples collected from white patients as the source dataset and those collected from non-white patients as the target dataset. The samples collected from non-white patients were further divided into samples from black / african american patients and samples from asian patients (other races were excluded due to extreme sparsity in the available dataset). From this point, the samples in the black/african american and asian categories were further split in half (first half reserved for fine tuning the models in the transfer learning stage, and the latter half used explicity for testing model accuracy / loss).

4.2 Data Processing

After properly partitioning the data, the data was then prepared for model input.

First, all null values and irrelevant features were filtered out from the dataset. The only features we were concerned with for this study were chromosome number, median copy number, start index, and end index. Thus at this point, each sample input came in the form of a (60623, 4) matrix encoding the aforementioned features. [note: each CNV file maps these features for 60,623 unique genes in the sample genome].

After this inital pre-processing, the absolute difference between start and end index was computed to obtain the gene length of each tracked mutation. Restructuring these two columns output a matrix size of (60623, 3). The next step involved one hot encoding the chromosome number (1 of 24 unique possibilities) as well as the breast cancer labels (1 of 23 unique diagnoses). At the end, the columns for gene length and median copy number were normalized between 0 and 1.

After all processing was finished, every data point came in the form of a (60623,26) matrix encoding the relevant CNV data; likewise, the corresponding labels came in the form of a 23 length vector denoting the breast cancer subtype associated.

4.3 Model Details

As previously mentioned, the three main deep learning models being tested under transfer learning in this study included: (1) the classic DNN, (2) LSTM, and (3) the Transformer. Note that that individual sample inputs to each model came in the form of a (60623,26) matrix.

Deep Neural Network architecture: For this study, the Deep Neural Network Architecture came in the form of 3 fully connected layers followed by a softmax classifier layer. The input layer had 60,623 input units for the CNV data. The two layers in the middle had 64 units each and the output layer had 23 units.

Transformer Model: For this study, the Transformer Model Architecture came in the form of a PCA module, 3 consecutive encoder blocks, and a softmax classifier layer. The input first utilized the PCA module to reduce the (60623,26) matrix into a (25,26) matrix - this was done to cut down computational complexity to a feasible threshold. Within the 3 encoder blocks, the fully connected layer contained 128 units to process output of the self attention sublayer. The output layer had 23 units.

Long-Short Term Memory Model: For this study, the LSTM Architecture came in the form of a PCA module, a 64 unit LSTM layer, followed by a softmax classifier layer. The input first utilized the PCA module to reduce the (60623,26) matrix into a (25,26) matrix - this was done to cut down computational complexity to a feasible threshold. The LSTM layer contained 64 units to process input for each gene passed in. The final softmax layer had 23 units to classify the output aggregated across the LSTM cells.

Note: These architectures have been visualized below [*Fig. 5*]

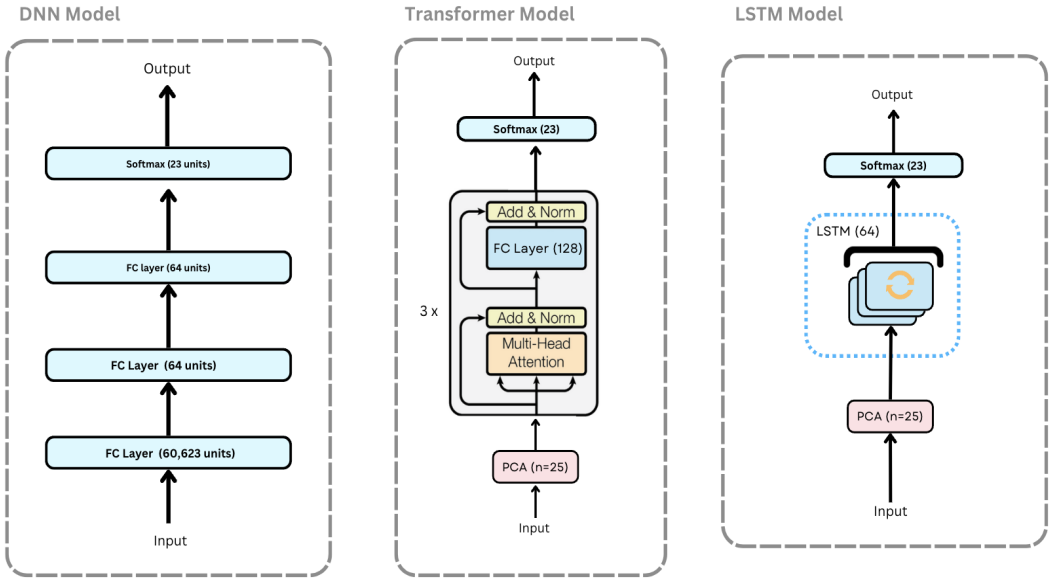

Figure 5: DL Model Architectures

4.4 Experimental Setup

For the purposes of this study, 5 models were observed under the baseline condition. Of the 5, the 3 deep learning models were additionally observed under the two transfer learning techniques. Thus, a total of 11 experimental conditions were observed. The conditions are outlined below:

- **Baseline Models:** DNN, LSTM, Transformer, SVM, XGBoost
- **Fine-Tuning TL Models:** DNN, LSTM, Transformer
- **Feature-based TL Models:** DNN, LSTM, Transformer

For each trial run, the model's test accuracy and loss on the target data-set was collected after training. In order to procure a more accurate sense of each model's performance, the accuracy metrics were averaged over 50 trial runs for each of the 11 experimental conditions.

Ludwig AI's documentation suggests keeping the number of training epochs between 3 and 10 when fine-tuning [1]. To keep training duration consistent across all deep learning models, the following criteria were held for all models:

- **Baseline Models** were trained on source data for 8 consecutive epochs, prior to testing on target data.
- **Fine-Tuning TL Models** were first trained on source data for 8 consecutive epochs before continued training on target data for 4 extra epochs. The model was then tested on target data.
- **Feature-based TL Models** were first trained on source data for 8 consecutive epochs before freezing layers (all layers frozen except for last classifier layer). The model then continued training on target data for 4 extra epochs before testing on target data.

The primary metric used for benchmarking model performance was test accuracy on the target domain, averaged across 50 trials runs for each experimental condition.

5 Results

After running the outlined experimental trials, the following results were obtained. Results below depict average test accuracy metrics aggregated across 50 trials runs. Baseline model performance is highlighted in yellow. Experimental conditions where test accuracy improved are highlighted in green, and conditions where test accuracy regressed are highlighted in red. Fig. 6 shows experimental results with black and african american patients as the target dataset, and Fig. 7 for asian patients as the target dataset.

Accuracy Results Table [Black + AA Patients]

Model/Technique	Baseline	Fine Tuning	Feature Based
DNN	82.05%	78.2%	86.33%
LSTM	79.49%	73.48%	82.05%
Transformer	79.05%	85.05%	85.05%
XG Boost	92.18%	N/A	N/A
SVM	75.64%	N/A	N/A

Figure 6: Results for Black and African American Patients

Accuracy Results Table [Asian Patients]

Model/Technique	Baseline	Fine Tuning	Feature Based
DNN	82.14%	77.14%	87.43%
LSTM	80.24%	74.57%	83.43%
Transformer	80.14%	86.34%	86.34%
XG Boost	95.24%	N/A	N/A
SVM	67.86%	N/A	N/A

Figure 7: Results for Asian Patients

There are a few observable trends in the provided results above [*Fig. 6* and *Fig. 7*].

First and foremost, Feature-Based Transfer Learning saw slight improvements to model performance from the baseline metric across all deep learning models. The change in accuracy saw an improvement of $\sim 5\%$ across the tested models. This trend was observable across both sample target domains (black / african american patients and asian patients).

On the other hand, Fine-Tuning based Transfer Learning yielded more ambiguous results. While fine-tuning based transfer learning slightly improved test accuracy for the Transformer model, it ultimately decreased test accuracy for the DNN and LSTM model. These changes in accuracy (both improvement and regression) were also in the range of $\sim 5\%$. This trend was observable across both sample target domains (black / african american patients and asian patients).

While Transfer Learning had a generally discernable impact on performance of the deep learning models, the baseline XG Boost model outperformed all experimental conditions with a test accuracy above 90% on both target domains. However, the SVM baseline model significantly underperformed all experimental conditions with a test accuracy below 76% for both target domains.

Holistically, all improvements and reductions in test accuracy under Transfer Learning techniques only yielded a change of about $\sim 5\%$. Though these changes are notable, they are not necessarily statistically significant - and thus require further testing.

6 Discussion

6.1 Effects of Transfer Learning

We suspect that Feature-Based Transfer Learning enhanced test accuracy across all deep learning models because it allows the model more control over which learned features it utilizes for the fine tuned prediction. By freezing layers after training on the source domain, the generated features from the CNV data are refined to smaller subset - which may make it easier for the model to leverage whatever abstracted features were learned from training on the source dataset.

The figures below [*Fig. 8* and *Fig. 9*] display the loss and accuracy learning curves for an example trial run of the DNN model under feature-based transfer learning. Prior to the transfer learning phase (switch denoted by the gray dotted line), most of the loss reduction was already handled. Using the learned features from the source data as a starting point for continued training likely explains the upward trajectory of performance. However, after the initial training phase, further substantial progress seems limited.

This behavior is similarly observable in the accuracy learning curve where most of the learning occurs prior to the transfer learning phase, and improvement after switching to the target domain seems positive but relatively slow and limited. This likely explains why feature-based transfer learning yielded some observable improvement across all 3 deep learning models, but only ever so slightly ($\sim 5\%$).

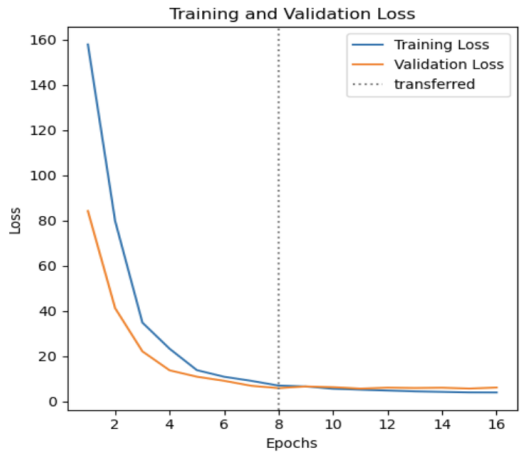

Figure 8: DNN Trial Run Loss

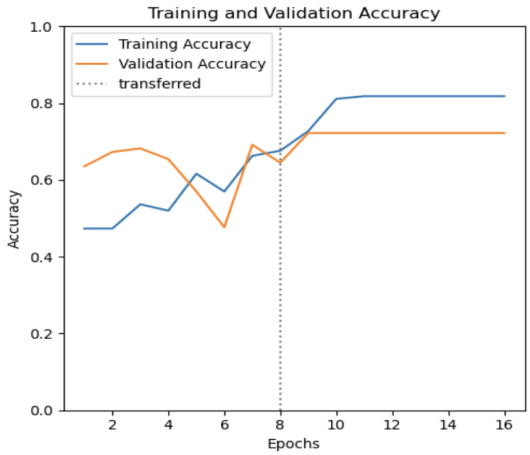

Figure 9: DNN Trial Run Accuracy

On the other hand, a majority of the tested deep learning models encountered setbacks in performance under Fine-Tuning based Transfer Learning. This regression in test accuracy leads us to speculate that transfer learning under fine-tuning based TL actually causes extreme overfitting in the model. This is observable in the loss learning curve taken from the following example trial run of the LSTM model under fine-tuning based TL [*Fig. 10*].

In this figure, the gap in between the validation and training loss curves exponentially grows in the fine-tuning phase. This leads us to believe that the effects of over-fitting are exacerbated during the fine-tuning phase - likely because allowing further weight / bias updates and modifications across the entire model architecture can over-amplify the model's capacity when fine-tuning. A 2014 study into the usability of deep transfer learning demonstrated that if the target dataset is small and the number of parameters is large, fine-tuning may result in overfitting - so it is

Figure 10: LSTM Trial Run Loss

often better if the features are left frozen [33]. Thus, fine-tuning based TL may ultimately prevent our models from reaping the same benefits of knowledge transfer that are observed when freezing model layers in featured-based TL. This potential for extreme overfitting allowed under fine-tuning based TL likely caused the deep learning models to regress in performance from the baseline metrics.

While performance regressed with this technique under the DNN and LSTM model, the Transformer model's accuracy was largely unaffected by the over-fitting. We attribute this experimental difference to the residual connections present in the Transformer's encoder layers. Architecturally, these connections serve as barrier against the regression of learned information, ensuring that the model's learning trajectory only improves through the training process. Thus, it is likely that the design of the Transformer model allowed its continual enhancement in performance through fine-tuning based TL.

In the case that fine-tuning based TL does cause excessive over fitting, it is important to consider the following possibility. Although there are demonstrable racial differences in the kinds of cancerous genomic mutations that occur [40], these differences may not be significant enough to warrant continual training / fine tuning without freezing layers (as doing so, would ultimately cause the over-fitting that was observed). To say the least, the experimental results under fine-tuning based transfer learning warrants further investigation into the racial disparities in genomic mutation data.

6.2 Feasibility of Deep Transfer Learning

Upon further inspection of the experimental results, it is important to take note of how the deep learning models (with and without transfer learning) performed in comparison to XGBoost and SVM. As previously noted, SVM severely underperformed all the tested models while XGBoost outperformed all of them (with above 90% accuracy on both target domains). This may possibly suggest that Deep Learning and the application of Transfer Learning may not be the best subset of machine learning models for this use case. While models like the Transformer and LSTM are able to capture relationships in between the genes outlined in the CNV data (e.g. through self attention and use of hidden state / long term memory), the structure of genomic mutation data may be better processed via decision tree type structures. Although SVM underperformed, it

is likely that XG Boost's use of gradient boosting and ensemble learning allowed it pick up on mutation patterns that otherwise would have gone unrecognized via deep learning models. That being said, this line of questioning would also need further investigation for confirmation.

From a holistic standpoint, the answer to whether Transfer Learning can help mitigate effects of the sparsity in ethnic minority genomic data is a little more nuanced. Some techniques (e.g. feature-based transfer learning) showed some potential for improving model performance on genomic based breast cancer prognosis, while others (e.g. fine-tuning based TL) resulted in less consistent results across model types. While the use of deep transfer learning for this problem domain shows promise, further investigation will be needed to better understand whether it is a feasible solution to this task.

7 Conclusion

In conclusion, our study presents a set of nuanced findings in the realm of deep transfer learning within the cancer data domain. Referencing previous studies, we initially hypothesized that transfer learning would lead to significant improvements in model performance on the target domain with sparser data. However, the results, while indicating slight improvements, fell short of statistical significance. This suggests that the effectiveness of deep transfer learning is intricately linked to the model type and architecture, learning method, as well as the distribution of the dataset. Despite the modest improvements, these findings highlight the potential of deep transfer learning, warranting further exploration to fully understand its capabilities and limitations.

One of the most intriguing aspects of this study, as shown in Figure 6 and 7, is the comparison of deep neural network models against XGBoost, a non-neural network gradient boosting technique. XGBoost outperformed all the deep learning models under experimental conditions. This observation challenges our initial motivation, which leaned towards the advantages of deep learning in this domain. It propels us to reconsider whether deep learning is indeed the optimal approach or if alternative methods, like XGBoost, offer more effective solutions.

Delving deeper into the specific strategies employed, we observed a consistent trend across models: feature-based learning methods were effective. This approach, which involves freezing layers, allowed for a more controlled and focused feature set, leading to improved generalization. In contrast, models subjected to fine-tuning exhibited a decline in performance, which we attribute to overfitting. This aligns with findings from prior findings on deep transfer learning [33], suggesting that an abundance of trainable parameters coupled with sparse target data increases the risk of overfitting. Thus, freezing layers, which limits the number of parameters, emerges as a more viable strategy in certain scenarios.

Remarkably, the transformer model demonstrated a degree of resilience to the negative impacts of fine-tuning, likely due to its unique architecture, including residual connections in its encoder. These connections potentially mitigated overfitting effects, highlighting an architectural advantage of the transformer model.

In conclusion, this study opens up new avenues for future research in deep transfer learning for medical data with potentially different distributions. While deep learning shows promise, its current application in this domain may not be the most effective. Further investigations should aim to address the complexities observed and explore the potential solution synthesizing traditional machine learning techniques, like XGBoost, and deep learning models. Such explorations could

lead to the development of hybrid models that leverage the strengths of both approaches, potentially offering more robust solutions in this problem domain, thus benefiting more patients through improved clinical breast cancer diagnoses.

Contribution

Breakdown of Work Items:

- **Writing the proposal:** all equally contributed - Jonathan Lai (33.33%), Susie Li (33.33%), Chelsie Wei (33.33%)
- **Coding**
  - Jonathan Lai (50%) - Data Processing, Clinical Data Analysis, DNN, LSTM, and Transformer Models (baseline + transfer)
  - Susie Li (25%) - SVM, XGBoost Baseline Models
  - Chelsie Wei (25%) - Data Processing, Further Exploration
- **Running experiments and collecting data:** Jonathan Lai (25%), Susie Li (25%), Chelsie Wei (50%)
- **Discussions (of project ideas and experiment results):** Jonathan Lai (33.33%), Susie Li (33.33%), Chelsie Wei (33.33%)
- **Writing the final report:**
  - Jonathan Lai (60%) Introduction, Background, Methodology, Results, Discussion
  - Susie Li (20%) Background, Results, Discussion, Conclusion
  - Chelsie Wei (20%) Introduction, Related Works, Discussion, References

Resources

Github Link: Transfer Learning for Racial Sparsity in Genomic Breast Cancer Data

References

[1] Ludwig AI. Fine-tuning pretrained models, 2023. Last accessed 16 December 2023.

[2] Nosayba Al-Azzam and Ibrahem Shatnawi. Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer. *Annals of Medicine and Surgery*, 62:53–64, 2021.

[3] Michal Byra, Michael Galperin, Haydee Ojeda-Fournier, Linda Olson, Mary O'Boyle, Christopher Comstock, and Michael Andre. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. *Medical physics*, 46(2):746–755, 2019.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[5] Ph.D. Daniel A. Gilchrist. Copy number variation (cnv), 2023. Last accessed 14 December 2023.

[6] Joe Dennis, Jonathan P Tyrer, Logan C Walker, Kyriaki Michailidou, Leila Dorling, Manjeet K Bolla, Qin Wang, Thomas U Ahearn, Irene L Andrulis, Hoda Anton-Culver, et al. Rare germline copy number variants (cnvs) and breast cancer risk. *Communications biology*, 5(1):65, 2022.

[7] R Fisher, L Pusztai, and C Swanton. Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer*, 108(3):479–485, Feb 2013.

[8] Yan Gao and Yan Cui. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nature communications*, 11(1):5131, 2020.

[9] Mehmet Gönen and Adam Margolin. Kernelized bayesian transfer learning. 28(1).

[10] Shujun Huang, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, and Wayne Xu. Applications of support vector machine (svm) learning in cancer genomics. *Cancer Genomics Proteomics*, 15(1):41–51, Jan-Feb 2018.

[11] Mohammadreza Iman, Hamid Reza Arabnia, and Khaled Rasheed. A review of deep transfer learning and recent advancements. *Technologies*, 11(2):40, 2023.

[12] Fan Jiang, Hui Liu, Shaode Yu, and Yaoqin Xie. Breast mass lesion classification in mammograms by transfer learning. In *Proceedings of the 5th international conference on bioinformatics and computational biology*, pages 59–62, 2017.

[13] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. 13:8–17.

[14] Mahalakshmi Kumaran, Carol E Cass, Kathryn Graham, John R Mackey, Roland Hubaux, Wan Lam, Yutaka Yasui, and Sambasivarao Damaraju. Germline copy number variations are associated with breast cancer risk and prognosis. *Scientific reports*, 7(1):14621, 2017.

[15] Xin Yu Liew, Nazia Hameed, and Jeremie Clos. An investigation of XGBoost-based algorithm for breast cancer classification. 6:100154.

[16] Jie Lu, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, and Guangquan Zhang. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80:14–23, 2015.

[17] Fabiana Lüönd, Stefanie Tiede, and Gerhard Christofori. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *British journal of cancer*, 125(2):164–175, 2021.

[18] Zohre Momenimovahed and Hamid Salehiniya. Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer: Targets and Therapy*, pages 151–164, 2019.

[19] Ka-Chun Wong Muhammad Toseef, Xiangtao Li. Reducing healthcare disparities using multiple multiethnic data distributions with fine-tuning of transfer learning, 2022. Last accessed 14 December 2023.

[20] Maged Nasser and Umi Kalsom Yusof. Deep learning based methods for breast cancer diagnosis: A systematic review and future direction. *Diagnostics (Basel)*, 13(1), January 2023.

[21] National Cancer Institute. Gdc data portal, 2023. Last accessed 11 November 2023.

[22] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 512–523. Curran Associates, Inc., 2020.

[23] NVIDIA. What is xgboost?, 2023. Last accessed 11 November 2023.

[24] Berna C Özdemir and Gian-Paolo Dotto. Racial differences in cancer susceptibility and survival: more than the color of the skin? *Trends in cancer*, 3(3):181–197, 2017.

[25] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[26] Neil H Segal, Paul Pavlidis, Cristina R Antonescu, Robert G Maki, William S Noble, Diann DeSantis, James M Woodruff, Jonathan J Lewis, Murray F Brennan, Alan N Houghton, and Carlos Cordon-Cardo. Classification and subtype prediction of adult soft tissue sarcoma by functional genomics. *Am J Pathol*, 163(2):691–700, Aug 2003.

[27] Deepak Soekhoe, Peter Van Der Putten, and Aske Plaat. On the impact of data set size in transfer learning using deep neural networks. In *Advances in Intelligent Data Analysis XV: 15th International Symposium, IDA 2016, Stockholm, Sweden, October 13-15, 2016, Proceedings 15*, pages 50–60. Springer, 2016.

[28] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 270–279. Springer, 2018.

[29] Muhammad Toseef, Xiangtao Li, and Ka-Chun Wong. Reducing healthcare disparities using multiple multiethnic data distributions with fine-tuning of transfer learning. *Briefings in Bioinformatics*, 23(3):bbac078, 2022.

[30] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. 3(1):9.

[31] Clement G Yedjou, Jennifer N Sims, Lucio Miele, Felicite Noubissi, Leroy Lowe, Duber D Fonseca, Richard A Alo, Marinelle Payton, and Paul B Tchounwou. Health and racial disparity in breast cancer. *Breast cancer metastasis and drug resistance: challenges and progress*, pages 31–49, 2019.

[32] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[33] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[34] Daping Yu, Zhidong Liu, Chongyu Su, Yi Han, XinChun Duan, Rui Zhang, Xiaoshuang Liu, Yang Yang, and Shaofa Xu. Copy number variation in plasma as a tool for lung cancer prediction using extreme gradient boosting (xgboost) classifier. *Thorac Cancer*, 11(1):95–102, Jan 2020.

[35] Xiang Yu, Jian Wang, Qing-Qi Hong, Raja Teku, Shui-Hua Wang, and Yu-Dong Zhang. Transfer learning for medical images analyses: A survey. *Neurocomputing*, 489:230–254, 2022.

[36] Nawel Zemmal, Nabiha Azizi, Nilanjan Dey, and Mokhtar Sellami. Adaptive semi supervised support vector machine semi supervised learning with features cooperation for breast cancer classification. *Journal of Medical Imaging and Health Informatics*, 6(1):53–62, 2016.

[37] Huanan Zhang, Ze Tian, and Rui Kuang. Transfer learning across cancers on dna copy number variation analysis. In *2013 IEEE 13th International Conference on Data Mining*, pages 1283–1288. IEEE, 2013.

[38] Yulin Zhang, Tong Feng, Shudong Wang, Ruyi Dong, Jialiang Yang, Jionglong Su, and Bo Wang. A novel XGBoost method to identify cancer tissue-of-origin based on copy number variations. 11.

[39] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[40] Berna C. Özdemir and Gian-Paolo Dotto. Racial differences in cancer susceptibility and survival: More than the color of the skin? *Trends Cancer 2017*, 162(11), November 2017.