# Evaluating the Generalization of Neural Network-Based Pan-Cancer Classification Models for Cohort-Specific Predictions

**Thomas Fontanari**[1,2], **Mariana Recamonde-Mendoza**[1,2]

[1]Institute of Informatics, Universidade Federal do Rio Grande do Sul (UFRGS),
Porto Alegre - RS, Brazil

[2]Bioinformatics Core, Hospital de Clínicas de Porto Alegre (HCPA),
Porto Alegre - RS, Brazil

{tvfontanari,mrmendoza}@inf.ufrgs.br

***Abstract.*** *This study develops and evaluates pan-cancer (PC) models for cohort-specific (CS) predictions using neural networks (NNs). We adopt a dual approach, including a method inspired by few-shot learning, aiming at improving the models' ability to distinguish between normal and tumorous tissues across diverse cohorts. The first approach trains a NN with comprehensive PC datasets containing 16 cancer types, comparing it against CS models on a target cohort, while the second analyzes whether PC models could generalize to smaller and unseen cohorts by training on 15 cohorts and evaluating on the excluded cohort. Our experiments show that PC models generally outperform CS models, even with limited sample sizes and class imbalances. Moreover, the few-shot approach successfully generalizes to other cancer types, highlighting its potential to advance personalized cancer diagnosis and treatment.*

## 1. Introduction

In the last decades, new genome sequencing technologies have greatly increased the availability of gene expression data. Besides enabling a deeper understanding of genomic diseases, they also make it possible to develop machine learning (ML) models that can potentially be used in clinical settings. The genomics of cancer, in particular, has been widely studied and modeled, as it is one of the leading causes of death worldwide. As an example, ML techniques have been utilized to create models capable of distinguishing cancerous tissues from normal ones, differentiating among cancer types or molecular subtypes, and predicting cancer survival or recurrence [Li et al. 2022].

Various works argue that gene expression data are difficult to model due to their high-dimensionality in comparison with the number of samples and possible non-linearities [Mostavi et al. 2021, Hanczar et al. 2022, Alharbi and Vakanski 2023]. These challenges are particularly pronounced in the context of rarer cancer types and subtypes, where limited sample sizes can hinder model performance. Consequently, various strategies have been proposed to enhance the robustness of models in small-sample settings, including transfer learning [Hanczar et al. 2022, Khorshed et al. 2020] and feature selection [Wang et al. 2022].

A well-established insight from biological research is that certain genes are implicated across multiple cancer types. For example, the tumor suppressor gene *TP53* is frequently mutated in a wide array of cancers [Chen et al. 2022]. This observation suggests that predictive models for specific cancer types could benefit from leveraging expression levels of genes that are relevant markers of the same task across multiple types of cancer. For instance, in the task of distinguishing normal from tumorous tissue samples, one might traditionally train a model exclusively on samples from the tissue of interest. However, an alternative approach involves training a model on a diverse set of samples from multiple cancer types, aiming to identify gene expression patterns that are broadly applicable across cancers. Such patterns, supported by a larger and more diverse dataset, could lead to more robust and accurate models.

Building on this idea, our work aims at improving model performance in small-sample scenarios by incorporating data from all available cancer types during training. Following prior studies [Hanczar et al. 2022], we refer to these models as pan-cancer (PC) models, given that they are trained using data from multiple cancer types. We compare the performance of PC models on specific cohort tasks with that of cohort-specific (CS) models, which are trained only on data from a single cohort. In this study, we focus on the task of distinguishing between normal and tumorous samples using neural networks (NN), which have demonstrated state-of-the-art results in cancer prediction based on gene expression data [Hanczar et al. 2022, Khalsan et al. 2022].

We evaluate PC and CS models under two scenarios. In the first scenario, PC models are trained using data from multiple cohorts, including the CS dataset. As expected, the PC models generally outperform their CS counterparts, with the extent of improvement influenced by the size and class imbalance of the CS dataset. In the second scenario, we exclude the CS dataset from the PC model's training data and evaluate its performance on that specific cohort. This allows us to assess the generalization capability of PC models to unseen cancer types. Notably, we find that in certain cases, PC models trained with this approach inspired by few-shot learning can generalize well to new cohorts.

In summary, the contributions of this work are twofold: (i) We provide evidence that NN-based PC models outperform CS models on the same tasks, offering a new strategy to enhance performance on small gene expression datasets (Section 4.1); and (ii) We demonstrate that NN-based PC models can, in some instances, successfully predict outcomes in cohorts not seen during training (Section 4.2)). While our experiments have focused on NN, we expect that the insights and conclusions drawn in our experiments may be applicable to other supervised learning algorithms.

## 2. Related Works

Our research focuses on cancer classification models using gene expression data, particularly addressing how to better utilize available data in scenarios with limited samples, aligning with few-shot learning methodologies [Wang et al. 2020]. Specifically, we concentrate on distinguishing between tumorous and normal samples (tumor prediction) using cohorts of different cancer types.

Numerous studies have explored models for cancer genomics classification tasks. Several have concentrated on comparing different models and neural network architectures for specific tasks, such as tumor prediction, tissue-of-origin classification [Hanczar et al. 2022], and disease stage identification [Yu et al. 2019]. Common approaches include the use of (deep) NNs [Yu et al. 2019, Divate et al. 2022, Khalsan et al. 2022], convolutional neural networks (CNNs) [Mostavi et al. 2020, Mohammed et al. 2021], graph neural networks (GNNs) [Ramirez et al. 2020, Lee et al. 2020, Hayakawa et al. 2022], and transformers [Zhang et al. 2022] for cancer-related gene expression tasks. However, most of these works do not consider the integration of samples from different cohorts to enhance predictions on separate cohorts, distinguishing their focus from ours.

More closely related to our work are studies that employ few-shot learning techniques in the context of cancer genomics. Hanczar *et al.* [Hanczar et al. 2022] conducted an extensive investigation into various forms of transfer learning between different types of cancer, particularly comparing pan-cancer and cohort-specific models for tumor prediction tasks. An earlier study by Khorshed *et al.* [Khorshed et al. 2020] also examined transfer learning for cancer diagnosis. Our approach differs from these in that we do not fine-tune the model on the target cohort. Instead, we train the pan-cancer model using all available samples simultaneously and then evaluate its performance on individual cohorts, including unseen ones. This strategy demonstrates that pan-cancer models can learn generalized gene expression patterns that remain informative across multiple cohorts.

Another line of research has focused on metric learning models for distinguishing among tissue types. Mostavi *et al.* [Mostavi et al. 2021] developed a model using siamese networks to learn a distance function between tissue samples, enabling one-shot learning for cancer classification even when only a single sample is available. Metric learning has also been applied to single-cell RNA-seq data to differentiate between cell types across various experiments [Koh and Hoon 2021, Ma et al. 2022]. While these studies share similarities with our work in differentiating classes that were not present during training, our approach differs in that we aim to generalize a predictive model to different cohorts, making predictions for an unseen class, rather than differentiating between unseen classes.

## 3. Materials and Methods

The following sections describe the data collection and pre-processing processes, the steps involved in model training and evaluation, and our experimental approaches. Code and data are provided in our public repository[1].

### 3.1. Data Collection and Preprocessing

We obtained RNA-seq data for various cancer types from The Cancer Genome Atlas (TCGA)[2], focusing on the analysis of gene expression profiles. Specifically, we retrieved Fragments Per Kilobase of transcript per Million mapped reads Upper Quartile (FPKM-UQ) normalized data through the Xena Browser [Goldman et al. 2020] for each of the 33 available cancer cohorts. To ensure sufficient data for ML model training and validation, we filtered the cohorts to include only those with at least 10 Primary Tumor samples and 10 Normal Tissue samples. This criterion led to the inclusion of 16 cohorts in our study.

For each cohort, we applied additional filtering to improve data quality. Genes and samples with more than 20% missing values were excluded, following established practices in similar studies [Duan et al. 2021, Albaradei et al. 2021, Chaudhary et al. 2018]. To further reduce the dimensionality of the data and focus on biologically relevant features, we retained only protein-coding genes by mapping them to their corresponding *peptide IDs* using the STRINGdb API[3]. The FPKM-UQ values were then subjected to a

[1] https://github.com/thomasvf/generalization-pc

[2] https://www.cancer.gov/ccg/research/genome-sequencing/tcga

[3] https://string-db.org/cgi/help.pl?subpage=api

logarithmic transformation to normalize the data and approximate a Gaussian distribution, which is beneficial for many ML algorithms. After completing these preprocessing steps, each sample in our final dataset contained expression values for 14,133 genes. Table 1 provides a summary of the number of examples for each cohort and sample type, with cohorts abbreviated according to their TCGA study names[4].

**Table 1. Number of examples from each cohort and sample type in the TCGA pan-cancer datasets after the preprocessing.**

| Cohort | Primary Tumor | Solid Tissue Normal | Total |
|---|---|---|---|
| BLCA | 411 | 19 | 430 |
| BRCA | 1097 | 113 | 1210 |
| COAD | 469 | 41 | 510 |
| ESCA | 161 | 11 | 172 |
| HNSC | 500 | 44 | 544 |
| KICH | 65 | 24 | 89 |
| KIRC | 533 | 72 | 605 |
| KIRP | 288 | 32 | 320 |
| LIHC | 371 | 50 | 421 |
| LUAD | 524 | 59 | 583 |
| LUSC | 501 | 49 | 550 |
| PRAD | 498 | 52 | 550 |
| READ | 166 | 10 | 176 |
| STAD | 375 | 32 | 407 |
| THCA | 502 | 58 | 560 |
| UCEC | 547 | 35 | 582 |
| Total | 7008 | 701 | 7709 |

### 3.2. Model Training

Given our objective to evaluate the generalizability of pan-cancer (PC) models for cohort-specific (CS) predictions across both seen and unseen cohorts, we opted not to perform a broad comparative analysis of different learning algorithms. Instead, we focused our experiments on a 2-layer fully connected neural network, a learning algorithm that has demonstrated state-of-the-art performance in gene expression tasks related to cancer prediction [Hanczar et al. 2022, Yu et al. 2019].

Specifically, the NN architecture we employed consists of 256 hidden neurons, followed by a ReLU activation function, and a final sigmoidal output neuron that predicts whether a given sample is tumorous. We utilized the AdamW optimizer [Loshchilov and Hutter 2017], with hyperparameters tuned for each run and cohort through random search with 8 iterations. The only hyperparameters tuned were the learning rate and the weight decay of the optimizer. Given that gene expression data is generally imbalanced and can easily lead to overfitting if not handled properly, the models were trained using a weighted cross-entropy function. This variation of the cross-entropy increases the penalty associated with mistakes of the minority classes in proportion to the inverse of their sizes.

[4] https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations

### 3.3. Experimental Approaches

We designed two experiments to assess the generalization capabilities of models trained on multiple cohorts. The first experiment aimed to determine whether incorporating samples from additional cohorts (out-of-cohort samples) during training enhances the model's performance on a specific target cohort. For each target cohort, we first trained a PC model using samples from all 16 cohorts on the task of distinguishing between tumorous and normal samples (Figure 1-(a)). We then evaluated its performance on a test set consisting solely of samples from the target cohort. We compared this performance with that of a CS model trained exclusively on samples from the target cohort (Figure 1-(b)). Both models were evaluated on the same test set, ensuring no overlap with the training samples. This evaluation was repeated 16 times, with each cohort serving as the target cohort in turn. We note that the PC model was consistent across all 16 evaluations.

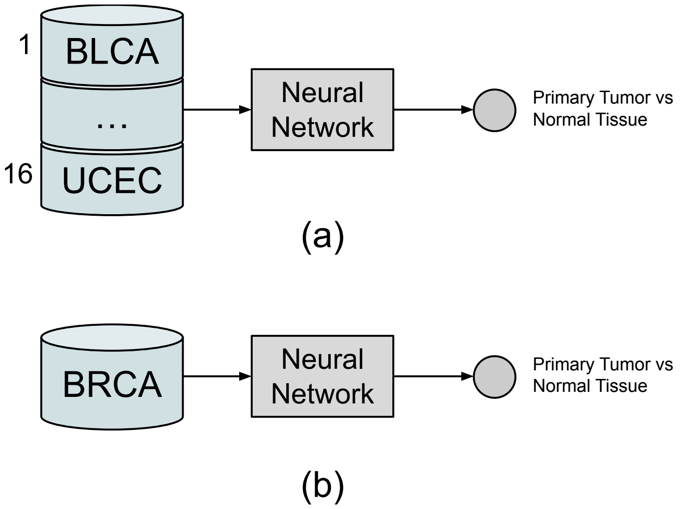

1
BLCA
...
16
UCEC
Neural Network
Primary Tumor vs Normal Tissue
(a)
BRCA
Neural Network
Primary Tumor vs Normal Tissue
(b)

**Figure 1. In the out-of-cohort samples experiment, the training cohorts were used for fitting the (a) pan-cancer and (b) cohort-specific models. Their performance was compared on a test set composed only of target cohort samples (BRCA in this example) to evaluate whether out-of-cohort samples improve the model performance on a cohort-specific task.**

In the second experiment, depicted in Figure 2, we evaluated whether the PC model could generalize to cohorts not included in the training set, that is, to unseen cohorts. This is inspired by the few-shot learning strategy that enables a pre-trained model to generalize over new categories of data. We trained the PC model on 15 of the 16 available cohorts and tested it on a cohort left out during training, referred to as the target cohort. This model's performance was again compared with that of a CS model trained solely on samples from the target cohort. Both models were tested on the same test set, with no overlap between the test and training samples. This evaluation was repeated 16 times, with each cohort serving as the target in turn.

### 3.4. Performance Evaluation and Correlation Analysis

Both experiments were evaluated through 5-times repeated holdout and we focused primarily on the macro-average F1-score, which computes the average between the F1-scores of each class in the task (in our experiments, the Tumor and Non-tumor classes).

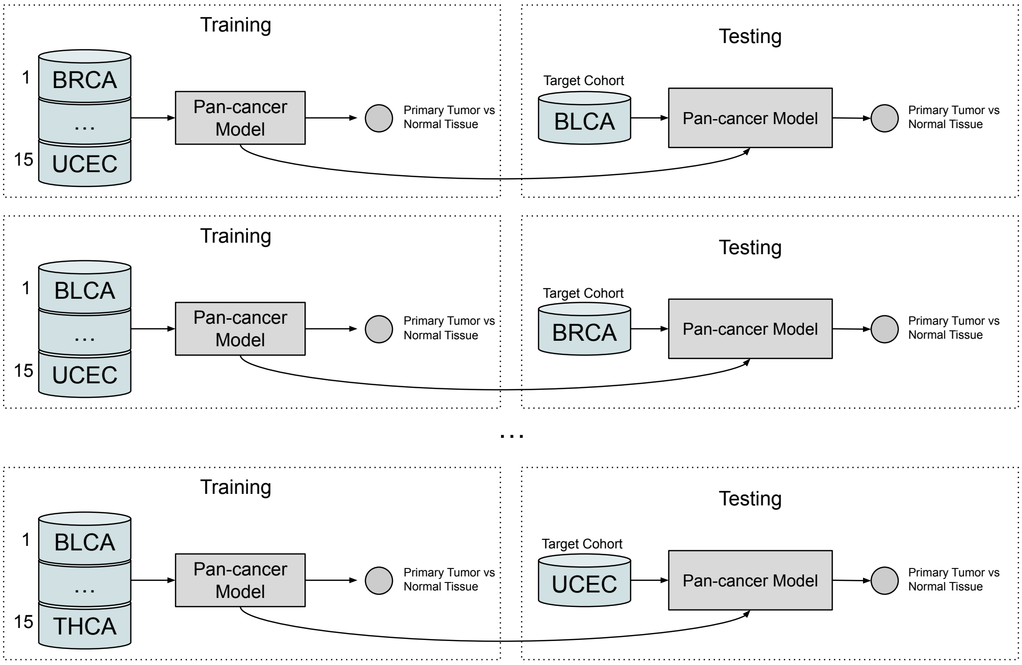


**Figure 2. In the unseen cohorts experiment, for each of the available cohorts, we trained a PC model using all samples but those of a specific target cohort. Then, we evaluate this model only on samples belonging to the target cohort.**

To draw conclusions about the effects of sample size and dataset imbalance on the scores, we used the Pearson correlation coefficient. Furthermore, we computed the two-sided p-value for the null hypothesis that a correlation as great as the ones observed could have been obtained if the variables were not actually correlated. This value is computed based on the probability density function of the correlation coefficient $r$ obtained by assuming that the two variables are normally distributed and independent. More details about its computation can be found in *scipy*'s documentation[5].

## 4. Results

In the following, we present the results obtained from the two experiments: the out-of-cohort samples experiment and the unseen cohort experiment.

### 4.1. Out-of-Cohort Samples Experiment

Figure 3 shows the results of the first experiment, comparing the performance of the PC model trained with all the available cohorts (blue bars) to that of CS models (orange bars). The results of our experiments demonstrate that PC models generally outperform CS models in predicting whether a tissue sample is tumorous or normal across a variety of cancer cohorts. Specifically, the PC model achieved higher or equivalent F1-scores in 15 out of the 16 cohorts, with the most significant improvement observed in the ESCA cohort, where the PC model outperformed the CS model by nearly 20%. This suggests that integrating data from multiple cohorts can substantially enhance model performance, particularly in cohorts with limited or imbalanced samples.

[5] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html

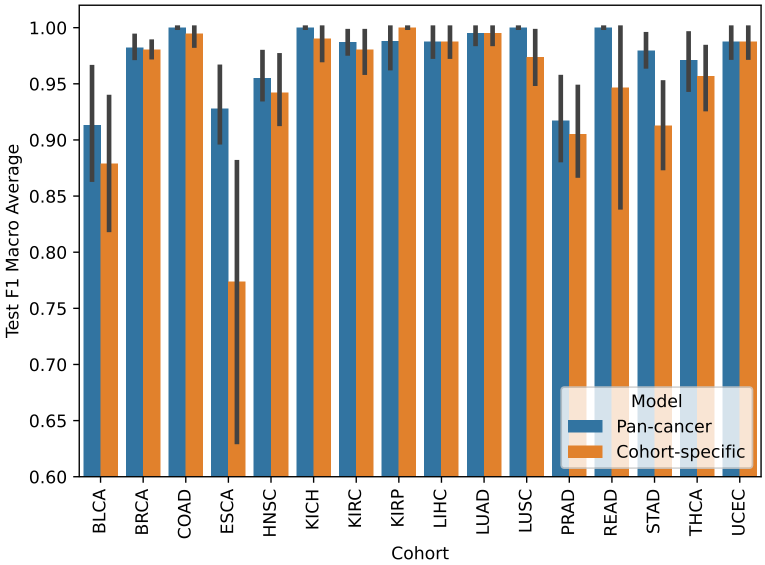
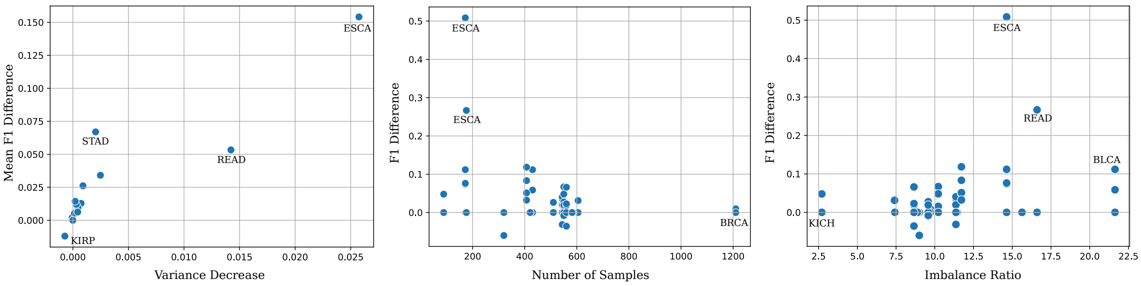


**Figure 3. Test F1-score of the pan-cancer model trained using all 16 available cohorts (blue bars) compared with the models trained only on the target cohort (orange bars), per cohort. In all but one case, the performance of the pan-cancer model was better than or equal to that of the cohort-specific.**

Further analysis revealed that the STAD and READ cohorts also benefited from the pan-cancer approach, showing F1-score improvements of $7.3\%$ and $5.6\%$, respectively. In contrast, for the LIHC, LUAD, and UCEC cohorts, the performance of the PC and CS models was identical, indicating that the inclusion of out-of-cohort data did not provide a significant advantage in these cases. Interestingly, the KIRP cohort was the only cohort where the CS model slightly outperformed the PC model, suggesting that in some cases, a focused, cohort-specific approach may be more effective.

In all these cases, we can see that the variance of the scores was high for the CS models, which led to an overall worst average result. This observation is made explicit in Figure 4-(a), where the mean performance gain for each cohort is shown against the variance difference between the PC and CS model results. The two variables show a Pearson correlation of $0.91$ (p-value $< 0.001$). In three cases (LIHC, LUAD, and UCEC) the performance was exactly the same for the PC and CS models, and for KIRP the CS model performance was actually $1.2\%$ higher. Interestingly, we observe that the variance of the CS KIRP model was smaller than its variance with the PC model.

As discussed previously, we expected that the major gains of the PC approach would come for datasets that have fewer samples and are more imbalanced. To evaluate that, we computed the difference between the PC and CS F1-scores for each run and cohort and plotted them against the sample size and the imbalance ratio in Figures 4-(b) and 4-(c), respectively. The imbalance ratio was calculated as the number of samples in the majority class over that in the minority class. The Pearson correlation coefficients for the sample size and class imbalanced were, respectively, $-0.243$ (p-value of $0.029$) and $0.236$ (p-value $0.035$). This indicates that there is indeed a small but significant relation between the benefit of using out-of-cohort samples and the size and imbalance of the cohorts.



(a) (b) (c)

**Figure 4. Effect of variables of interest in the F1-score difference between the PC and CS models. The mean F1-difference for each cohort is plotted against the decrease in variance in (a), whereas the F1-difference for each run and cohorts is plotted against the number of samples and the imbalance ratio in (b) and (c).**

## 4.2. Unseen Cohorts Experiment

We present the F1-scores for the unseen cohort experiment in Figure 5. In this experiment, the PC model was trained on all available datasets except for the target cohort, totaling 15 cohorts. This approach differs from the previous experiment, where the PC model was trained using all cohorts. Results are provided in a decreasing order of performance for the PC model, with the COAD cohort achieving the best result. Despite the exclusion of the target cohort from the training set, the PC model exhibits strong performance across various cancer types, including READ, HNSC, and BLCA. The fact that the model maintains competitive F1-scores in these cases demonstrates its robust generalization capability, even under the few-shot learning paradigm. This is particularly notable because it suggests that PC models can successfully transfer knowledge from related cancer types, mitigating the challenges posed by limited or unseen data.

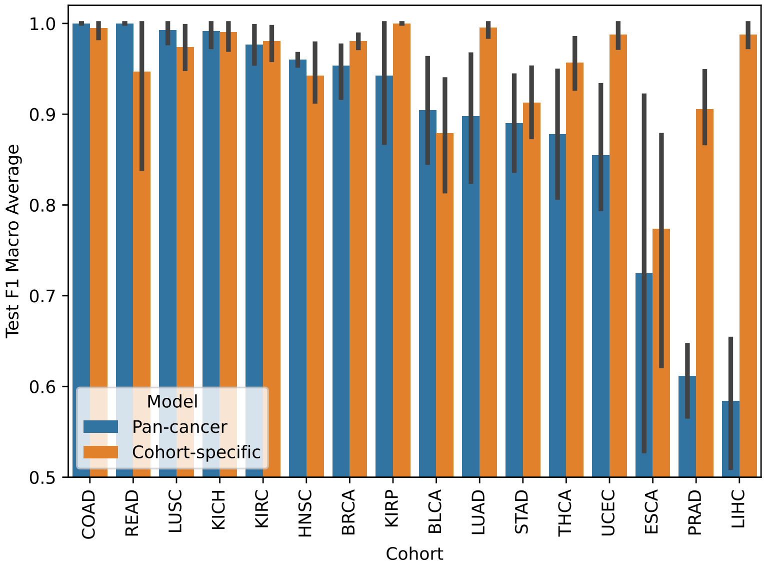


**Figure 5. Test F1-score of the pan-cancer model trained using all cohorts except the target (blue bars) compared with the models trained only on the target cohort (orange bars), per cohort, on the tumor prediction task.**

Nonetheless, there are specific instances where the CS model performs clearly better, as seen in the LIHC and PRAD cohorts. This suggests that in certain cancers, the unique characteristics of the cohort may require more focused learning that a broad, PC model might not fully capture. This variability highlights a limitation of the few-shot learning approach when the target cohort has distinct features not well-represented by other cohorts. Moreover, our results underscore the importance of understanding the specific characteristics of the target cohort, as certain cancer types may benefit more from a CS approach.

As before, we analyzed whether variations in performance correlated with factors such as sample size, class imbalance, or changes in variance. However, no statistically significant correlations were found, with Pearson correlation p-values of 0.627, 0.068, and 0.285 for sample size, imbalance, and variance change, respectively. A more compelling hypothesis is that the ability of the PC model to generalize to an unseen cohort may depend on the anatomical and molecular similarity between the unobserved cohort and those included in the training set. This is most evident in the COAD and READ cohorts. COAD and READ, both types of colorectal cancer (CRC), share similar molecular mechanisms and are anatomically close, as supported by the literature [Zuo et al. 2019]. In this experiment, the PC model achieved higher performance on COAD and READ than the CS models, with results comparable to those obtained when both cohorts were included in the PC model's training set.

A similar pattern is observed with the kidney cancer datasets (KICH, KIRC, and KIRP). KICH and KIRC, in particular, performed on par with the CS models, although KIRP showed a slight drop in performance. Despite this, the PC model's performance remained competitive. Interestingly, in some cases, the PC model demonstrated strong performance, even in the absence of training samples from any cohorts that are closely related to the unobserved cohort. For example, models trained without HNSC and BLCA still outperformed their cohort-specific counterparts, despite no obvious anatomical proximity to other cohorts in the training set. However, for certain cohorts like LIHC and PRAD, there was a noticeable decline in performance, indicating that the effectiveness of the pan-cancer approach may vary depending on the specific characteristics of the unseen cohort.

## 5. Discussion and Limitations

Our results provide compelling evidence that incorporating samples from diverse tissue types can enhance performance in cohort-specific classification tasks and enable accurate classification in new, unseen cohorts. This aligns with findings from other studies, which have shown that transfer learning can significantly improve classification accuracy, particularly when the amount of available data is limited [Hanczar et al. 2022, Khorshed et al. 2020]. By leveraging data from multiple cohorts, our approach offers a promising strategy for building robust models from gene expression datasets that have few samples.

However, our study has several limitations that warrant discussion. Firstly, we focused exclusively on the task of tumor prediction, specifically distinguishing between normal and tumorous samples. While this is a critical task in cancer genomics, there are other gene expression-based tasks that are equally important in clinical settings and could benefit from a similar pan-cancer approach. For example, prognosis prediction, which involves predicting survival outcomes, cancer recurrence, and metastasis identification, are all clinically relevant tasks where the identification of biomarkers is crucial. Expanding our methodology to these tasks could further validate the utility of pan-cancer models.

Another limitation of our study is that we used only RNA-seq data from TCGA. While TCGA provides a rich and comprehensive dataset, it is important to explore whether the same benefits can be observed when generalizing to more heterogeneous data sources. For instance, would integrating TCGA data with datasets from the Gene Expression Omnibus (GEO)[6] yield similar improvements?

Finally, in this paper we restricted our experiments to models based on NNs. Our decision to focus exclusively on NNs stems from their proven effectiveness in handling the high dimensionality and complexity of gene expression data. NNs are particularly adept at capturing intricate patterns, which is essential for accurately distinguishing between normal and tumorous tissues. Additionally, NNs' capacity to learn from large and diverse datasets makes them well-suited for pan-cancer models. However, the underlying principles and strategies we explore—such as leveraging pan-cancer data for enhanced generalization—are not inherently dependent on this specific algorithm. We anticipate that the insights gained from our study could be applicable to other supervised learning algorithms as well, suggesting that the potential benefits of our approach may extend beyond NNs, with broader implications for the development of robust predictive models in cancer genomics.

## 6. Conclusion and Future Work

In this paper, we dealt with the generalization capability of PC models. In particular, we aimed at answering whether introducing samples from different cohorts improve cohort-specific performance, and if the pan-cancer models were able to generalize to cohorts that were not used in training. We obtained encouraging results in both cases. In particular, we observed that the benefits for cohort-specific tasks were more substantial when the original cohort contained fewer samples and was more imbalanced, even when correcting imbalance using weighted cross-entropy. The PC models were also able to perform classification on unseen cohorts, specially when cohorts sharing the same tissue were used to train the model. However, our results also underscore the importance of considering cohort-specific characteristics when deciding between a pan-cancer and a cohort-specific approach, as the benefits of broader data integration may vary depending on the specific nature of the cohort.

Despite these results, there are some important questions that should be explored in future work. Specifically, we hope to include a wider set of cohort-specific classification tasks besides that of tumor prediction, and we wish to also analyze if our results would also be true when trying to generalizing to and from more heterogeneous sources.

## Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and by grants from the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul - FAPERGS [21/2551-0002052-0 and 22/2551-0000390-7 (Project CIARS)] and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [308075/2021-8].

[6] https://www.ncbi.nlm.nih.gov/geo/

## References

[Albaradei et al. 2021] Albaradei, S., Napolitano, F., Thafar, M. A., Gojobori, T., Essack, M., and Gao, X. (2021). MetaCancer: A deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. *Computational and Structural Biotechnology Journal*, 19:4404–4411.

[Alharbi and Vakanski 2023] Alharbi, F. and Vakanski, A. (2023). Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering*, 10(2):173.

[Chaudhary et al. 2018] Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2018). Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 24(6):1248–1259.

[Chen et al. 2022] Chen, X., Zhang, T., Su, W., Dou, Z., Zhao, D., Jin, X., Lei, H., Wang, J., Xie, X., Cheng, B., Li, Q., Zhang, H., and Di, C. (2022). Mutant p53 in cancer: from molecular mechanism to therapeutic modulation. *Cell Death & Disease*, 13(11):974.

[Divate et al. 2022] Divate, M., Tyagi, A., Richard, D. J., Prasad, P. A., Gowda, H., and Nagaraj, S. H. (2022). Deep learning-based pan-cancer classification model reveals tissue-of-origin specific gene expression signatures. *Cancers*, 14(5):1185.

[Duan et al. 2021] Duan, R., Gao, L., Gao, Y., Hu, Y., Xu, H., Huang, M., Song, K., Wang, H., Dong, Y., Jiang, C., Zhang, C., and Jia, S. (2021). Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLoS Computational Biology*, 17(8):1–33.

[Goldman et al. 2020] Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., Zhu, J., and Haussler, D. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology*, 38(6):675–678.

[Hanczar et al. 2022] Hanczar, B., Bourgeais, V., and Zehraoui, F. (2022). Assessment of deep learning and transfer learning for cancer prediction based on gene expression data. *BMC Bioinformatics*, 23(1):262.

[Hayakawa et al. 2022] Hayakawa, J., Seki, T., Kawazoe, Y., and Ohe, K. (2022). Pathway importance by graph convolutional network and Shapley additive explanations in gene expression phenotype of diffuse large B-cell lymphoma. *PLOS ONE*, 17(6):e0269570.

[Khalsan et al. 2022] Khalsan, M., Machado, L. R., Al-Shamery, E. S., Ajit, S., Anthony, K., Mu, M., and Agyeman, M. O. (2022). A survey of machine learning approaches applied to gene expression analysis for cancer prediction. *IEEE Access*, 10:27522–27534.

[Khorshed et al. 2020] Khorshed, T., Moustafa, M. N., and Rafea, A. (2020). Deep Learning for Multi-Tissue Cancer Classification of Gene Expressions (GeneXNet). *IEEE Access*, 8:90615–90629.

[Koh and Hoon 2021] Koh, W. and Hoon, S. (2021). MapCell: Learning a Comparative Cell Type Distance Metric With Siamese Neural Nets With Applications Toward Cell-Type Identification Across Experimental Datasets. *Frontiers in Cell and Developmental Biology*, 9:767897.

[Lee et al. 2020] Lee, S., Lim, S., Lee, T., Sung, I., and Kim, S. (2020). Cancer subtype classification and modeling by pathway attention and propagation. *Bioinformatics*, 36(12):3818–3824.

[Li et al. 2022] Li, R., Li, L., Xu, Y., and Yang, J. (2022). Machine learning meets omics: applications and perspectives. *Briefings in Bioinformatics*, 23(1):bbab460.

[Loshchilov and Hutter 2017] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[Ma et al. 2022] Ma, Z., Lu, Y. Y., Wang, Y., Lin, R., Yang, Z., Zhang, F., and Wang, Y. (2022). Metric learning for comparing genomic data with triplet network. *Briefings in Bioinformatics*, 23(5):bbac345.

[Mohammed et al. 2021] Mohammed, M., Mwambi, H., Mboya, I. B., Elbashir, M. K., and Omolo, B. (2021). A stacking ensemble deep learning approach to cancer type classification based on TCGA data. *Scientific Reports*, 11(1):1–22.

[Mostavi et al. 2021] Mostavi, M., Chiu, Y.-C., Chen, Y., and Huang, Y. (2021). Cancer-Siamese: one-shot learning for predicting primary and metastatic tumor types unseen during model training. *BMC Bioinformatics*, 22(1):244.

[Mostavi et al. 2020] Mostavi, M., Chiu, Y.-C., Huang, Y., and Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics*, 13:1–13.

[Ramirez et al. 2020] Ramirez, R., Chiu, Y.-C., Hererra, A., Mostavi, M., Ramirez, J., Chen, Y., Huang, Y., and Jin, Y.-F. (2020). Classification of cancer types using graph convolutional neural networks. *Frontiers in Physics*, 8:203.

[Wang et al. 2022] Wang, A., Liu, H., Yang, J., and Chen, G. (2022). Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data. *Computers in Biology and Medicine*, 142(33):105208.

[Wang et al. 2020] Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a Few Examples: A Survey on Few-Shot Learning. *ACM Computing Surveys*, 53(3).

[Yu et al. 2019] Yu, H., Samuels, D. C., Zhao, Y.-y., and Guo, Y. (2019). Architectures and accuracy of artificial neural network for disease classification from omics data. *BMC Genomics*, 20(1).

[Zhang et al. 2022] Zhang, T.-H., Hasib, M. M., Chiu, Y.-C., Han, Z.-F., Jin, Y.-F., Flores, M., Chen, Y., and Huang, Y. (2022). Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions. *Cancers*, 14(19):4763.

[Zuo et al. 2019] Zuo, S., Dai, G., and Ren, X. (2019). Identification of a 6-gene signature predicting prognosis for colorectal cancer. *Cancer Cell International*, 19(1):1–15.