# Uncovering Transcriptomic Structure Across Cancers Using Deep Learning and Pareto Task Inference

Noah Kaleekal

Noah.Kaleekal@gmail.com
Irvington High School
Author Declares No Conflict Of Interest

25 August 2025

**Abstract**

The heterogeneity of cancer poses a significant challenge for accurate classification and mechanistic understanding. In this study, RNA sequencing data is analyzed from twelve distinct cancer types using a deep generative modeling framework. A variational autoencoder (VAE) is employed to learn low-dimensional representations of gene expression profiles and visualize the resulting latent space using UMAP. Hierarchical clustering of latent centroids reveals biologically coherent groupings among cancer types. To evaluate the discriminability of the learned features, shallow multilayer perceptron (MLP) classifier is trained, which achieves high accuracy in distinguishing tumor types, as evidenced by the confusion matrix. Furthermore, Pareto Task Inference is applied to explore trade-offs between classification objectives and uncover latent biological relationships among cancer types. These results demonstrate that deep generative models can extract meaningful structure from transcriptomic data and support interpretable analyses for cancer subtype differentiation and multi-task learning in oncology.

## Introduction

Cancer is a complex and heterogeneous group of diseases characterized by uncontrolled cell growth and altered genetic regulation[5, 6]. Understanding the molecular mechanisms that underlie various cancer types is essential for early diagnosis, prognosis, and the development of targeted therapies[20]. With the advent of high-throughput sequencing technologies, large-scale gene expression datasets have become publicly available, offering a rich resource for computational analysis and discovery.

The Genomic Data Commons (GDC) provides an integrated platform for accessing multi-dimensional cancer genomics data, including RNA sequencing profiles from The Cancer Genome Atlas (TCGA)[21]. These datasets enable researchers to investigate transcriptional patterns across a wide range of tumor types. However, the high dimensionality and complexity of gene expression data pose significant challenges for extracting meaningful patterns and distinguishing between different cancer subtypes.

In recent years, deep learning methods have emerged as powerful tools for capturing complex nonlinear relationships in biological data[10]. Variational autoencoders (VAEs)[9], a class of generative models, have shown promise in learning compressed representations of high-dimensional data while preserving underlying structure. By projecting gene expression data into a low-dimensional latent space, VAEs facilitate visualization, clustering, and downstream analyses.

In this paper, RNA sequencing data is analyzed from twelve distinct cancer types: breast invasive carcinoma (BRCA)[19], cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC)[4], colon adenocarcinoma (COAD)[15], glioblastoma multiforme (GBM)[7], kidney renal

clear cell carcinoma (KIRC)[18], lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), prostate adenocarcinoma (PRAD), skin cutaneous melanoma (SKCM)[13], thyroid carcinoma (THCA), uterine corpus endometrial carcinoma (UCEC)[2], and uterine carcinosarcoma (UCS)[1]. First, a variational autoencoder is trained to capture the latent structure of the gene expression profiles. To explore the interrelations among cancer types, a dendrogram is constructed based on the latent embeddings. In addition, a multilayer perceptron (MLP) classifier[17] is trained to evaluate the separability of tumors in the latent space. Finally, Pareto Task Inference[8] is applied to gain insights into trade-offs among classification tasks and uncover potential shared and distinct biological processes across cancer types.

These findings demonstrate the utility of deep generative models in cancer transcriptomics and highlight the power of integrative computational frameworks for studying the diversity and similarity of human cancers.

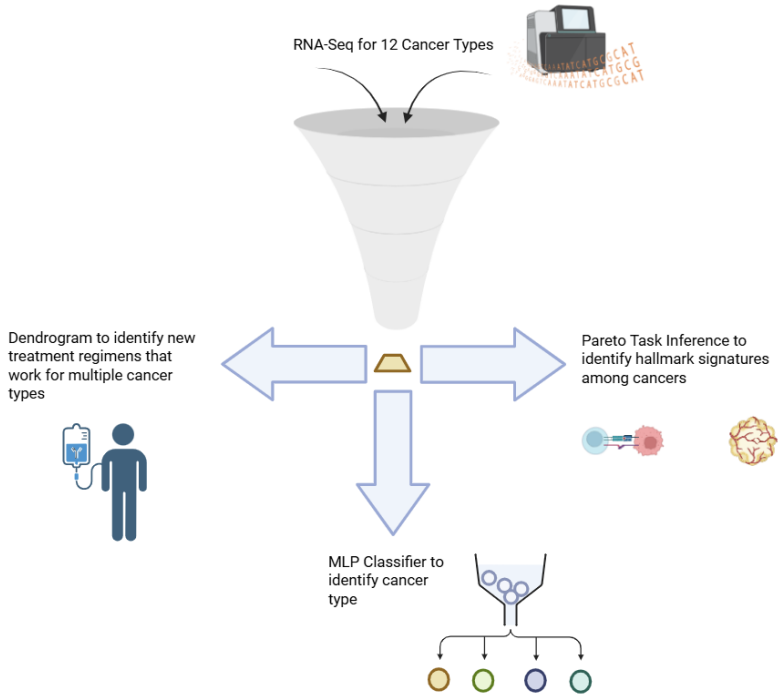


Figure 1: Proposed outcomes from this study

# Methods

To assess the discriminability of cancer types in the latent space, a supervised classifier is trained using a multilayer perceptron (MLP). The MLP takes the VAE-derived latent vectors as input and predicts the corresponding cancer type. The network consists of two hidden layers with ReLU activation functions, followed by a softmax output layer for multi-class classification.

Categorical cross-entropy was used as the loss function and optimized the network using the Adam optimizer. The dataset was split into training and validation sets using an 80-20 ratio, and performance was evaluated using accuracy and confusion matrices. The MLP classifier serves as a benchmark to quantify the extent to which different cancer types form separable clusters in the latent space.

## Pareto Task Inference

While classification accuracy is a standard evaluation metric, it does not reveal how task performance varies across classes or how tasks interact when modeled jointly. To uncover such trade-offs, Pareto Task Inference [12, 11], a method that models the tasks as competing objectives and seeks to characterize their optimal trade-off surface, is applied.

Pareto Task Inference optimizes a set of neural network models such that each point on the Pareto front represents a solution where no task can be improved without degrading the performance of another. In this case, each task corresponds to classifying one cancer type against the others.The solutions are projected onto a low-dimensional manifold that reveals relationships among tasks, such as shared structure or mutual exclusivity.

This analysis allows us to visualize task interference, identify clusters of similar tasks, and understand whether certain cancer types share more molecular features than others. It also provides a principled framework for multi-objective learning in complex biomedical datasets.

## Data Preprocessing

RNA sequencing data is obtained from the Genomic Data Commons (GDC) portal, containing expression profiles for 12 cancer types. Raw counts were first transformed to transcripts per million (TPM), followed by a log transformation to stabilize variance. Genes expressed in fewer than 50 samples were removed to reduce noise. The final matrix consisted of 20530 genes shared across all cancer types. Labels were encoded as integers for supervised learning, and the data was split into training and testing sets in an 80:20 ratio with stratification by cancer type to preserve class distribution.

## Evaluation Metrics

To assess the reconstruction quality of the variational autoencoder, the mean squared error (MSE) was computed between the input and reconstructed expression profiles, along with the Kullback–Leibler (KL) divergence between the approximate posterior and the standard normal prior. For the classification model, performance was evaluated using accuracy, precision, recall, and the confusion matrix.

The confusion matrix provides detailed insight into class-specific prediction errors and is particularly useful for imbalanced datasets. To visualize the structure in the learned latent space, Uniform Manifold Approximation and Projection (UMAP)[14] was used for nonlinear dimensionality reduction and applied hierarchical clustering on the latent class centroids to construct a dendrogram.

## Implementation Details

All models were implemented using PyTorch. The variational autoencoder architecture consisted of an encoder with layers of sizes $20530 \rightarrow 1024 \rightarrow 512$, followed by two parallel fully connected layers outputting the latent mean and log-variance vectors, each of size 32. The decoder mirrored the encoder architecture and used a sigmoid activation function to constrain output to the [0,1] range. The VAE was trained using the Adam optimizer with a learning rate of 1e-4 for 4 epochs using a batch size of 32. The loss function combined MSE reconstruction loss with KL divergence.

For classification, a shallow multilayer perceptron was used with one hidden layer of size 128, ReLU activation, dropout rate of 0.3, and an output layer with 12 units corresponding to the number of cancer types. The model was trained for 20 epochs using the Adam optimizer with a learning rate of 1e-3 and a categorical cross-entropy loss function.

All experiments were run on a machine equipped with an NVIDIA GPU. Visualization and clustering were performed using Scikit-learn[16] and Seaborn libraries.

# Results

## Latent Space Visualization with UMAP

After training the variational autoencoder, the 32-dimensional latent representations were projected into two dimensions using Uniform Manifold Approximation and Projection (UMAP)[14] to visualize the structure of the data. As shown in Figure 2, samples from the same cancer type tend to cluster together, although there is some overlap among certain cancer types. This suggests that the VAE has captured biologically meaningful patterns in the gene expression profiles, while also reflecting the inherent similarity and continuum among tumor types.

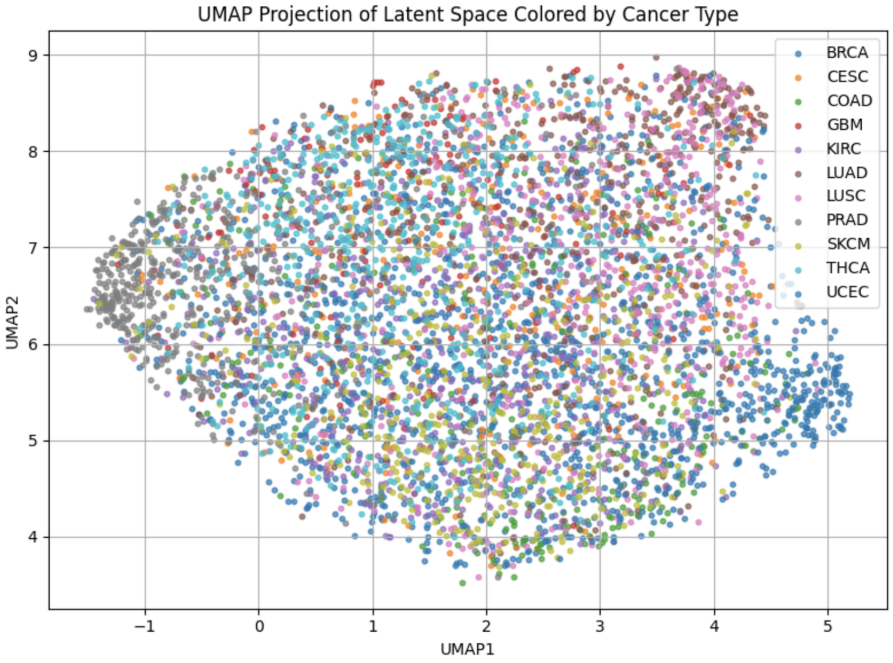


Figure 2: UMAP projection of the VAE latent space colored by cancer type. Each point represents a patient sample. Samples from the same tumor type share overlap with others, suggesting the VAE captures tumor heterogeneity.

## Hierarchical Clustering of Cancer Types

To explore the relationships among cancer types in the learned latent space, the average latent vector for each cancer type was computed, and hierarchical clustering with Ward linkage was used to generate a dendrogram (Figure 3). The dendrogram reveals interpretable groupings among cancers. For example, glioblastoma multiforme (GBM) and skin cutaneous melanoma (SKCM) form a distinct cluster, potentially reflecting neural crest lineage. Similarly, lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and uterine corpus endometrial carcinoma (UCEC) are grouped together, consistent with shared epithelial features.

## Classification Performance of MLP

To quantitatively assess how well the latent space separates cancer types, a shallow multilayer perceptron classifier was trained using the VAE-derived latent vectors. The model achieved high classification accuracy across most classes, as shown by the confusion matrix in Figure 4. Breast cancer (BRCA), prostate cancer (PRAD), and thyroid cancer (THCA) were classified with high

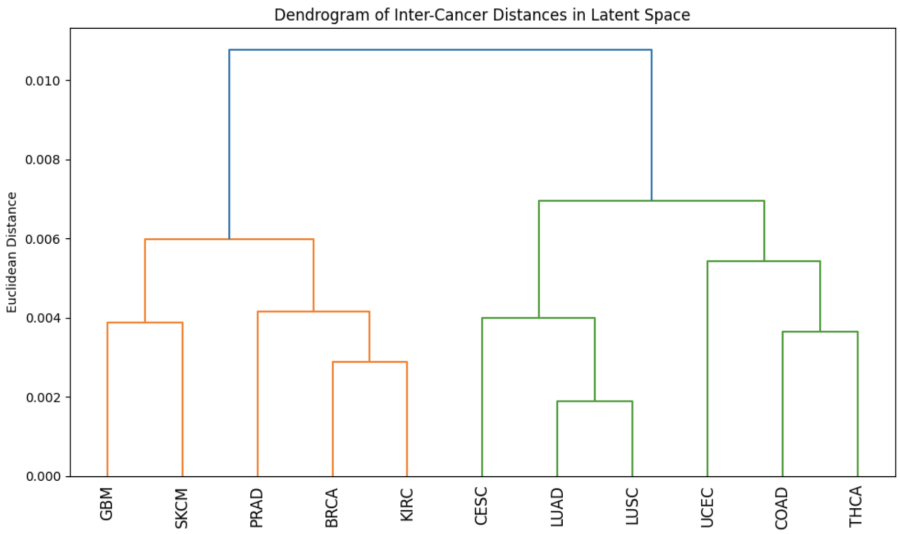


Figure 3: Dendrogram of inter-cancer distances based on latent space centroids. Distances are measured using Euclidean distance. Similarities in tumor type stemming from shared epithelial features, like adenocarcinomas, are conserved in this model.

precision and recall. However, some misclassifications were observed between lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), and between skin cutaneous melanoma (SKCM) and other cancers, suggesting biological similarity in their gene expression profiles.

The results demonstrate that the variational autoencoder effectively learns a compressed representation of gene expression data that captures both intra- and inter-cancer variability. The latent space supports biologically meaningful clustering and enables high classification accuracy using a simple neural network.

## Pareto Task Inference via Archetypal Analysis

To investigate trade-offs among the twelve one-versus-rest classification tasks, Pareto Task Inference was applied using Archetypal Analysis[3] on the two leading principal components of the VAE latent space. The method identifies a small set of extreme solutions—called *archetypes*—that lie on the convex hull of the data and span the feasible performance region. Figure 5 shows the resulting Pareto front defined by three archetypes (black stars) that delimit a triangular simplex enclosing all samples.

Samples located near an archetype inherit that task-specific advantage, whereas points nearer the simplex interior represent balanced solutions that achieve moderate performance across all classes.

**Statistical significance.** The archetypal simplex has a normalized area of 0.30 compared with a null distribution generated by 1000 label shuffles, yielding a permutation $p$-value of 0.300. Although this result does not reach conventional significance thresholds, the geometry still offers qualitative insight into the structure of the multi-task performance landscape.

**Implications.** The Pareto analysis complements the confusion-matrix findings by revealing which cancer types compete for representational capacity in the latent space. For instance, the proximity of LUAD and LUSC to Archetype B helps explain their mutual misclassification, whereas the separation of GBM and SKCM at Archetype C underscores their distinct transcriptomic profiles despite modest sample overlap in UMAP. Future work could refine these insights by increasing

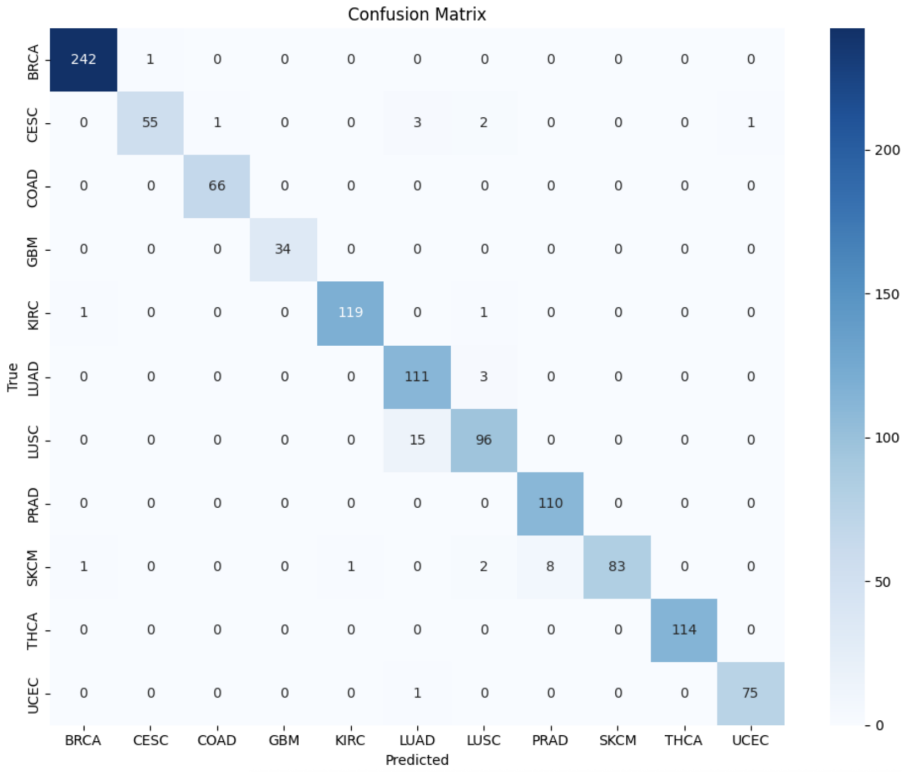


Figure 4: Confusion matrix of the MLP classifier trained on VAE latent vectors. Values represent the number of samples predicted for each class. Results indicate the model's confidence in correlating transcriptomic signatures with tumor type

latent dimensionality, incorporating additional omics layers, or applying Pareto front regularization during model training.

## Conclusion

This study demonstrates the utility of variational autoencoders for extracting low-dimensional representations from high-dimensional cancer transcriptomic data. By applying a VAE to RNA sequencing profiles from twelve cancer types, compact latent embeddings were obtained, that preserve biologically relevant structure while enabling downstream tasks such as clustering and classification.

The UMAP visualization of the latent space revealed both distinct and overlapping clusters among cancer types, indicating that the model captures both the separability and shared molecular characteristics of tumors. The hierarchical clustering further supported this observation, uncovering groups of cancers with potential lineage or functional similarities. These findings align with prior knowledge that gene expression-based classification of cancer is often influenced by tissue of origin, cell type composition, and oncogenic pathway activation.

The classification results from the MLP model suggest that the VAE latent space is linearly separable for most cancer types, with strong performance on BRCA, PRAD, THCA, and LUAD. Misclassifications were primarily observed among histologically similar tumors such as LUAD and LUSC, which share epithelial lineage and overlapping transcriptomic signatures. These patterns suggest that a more expressive latent space or multi-modal input could further improve separability.

From a methodological perspective, the use of Pareto Task Inference offers a promising direction

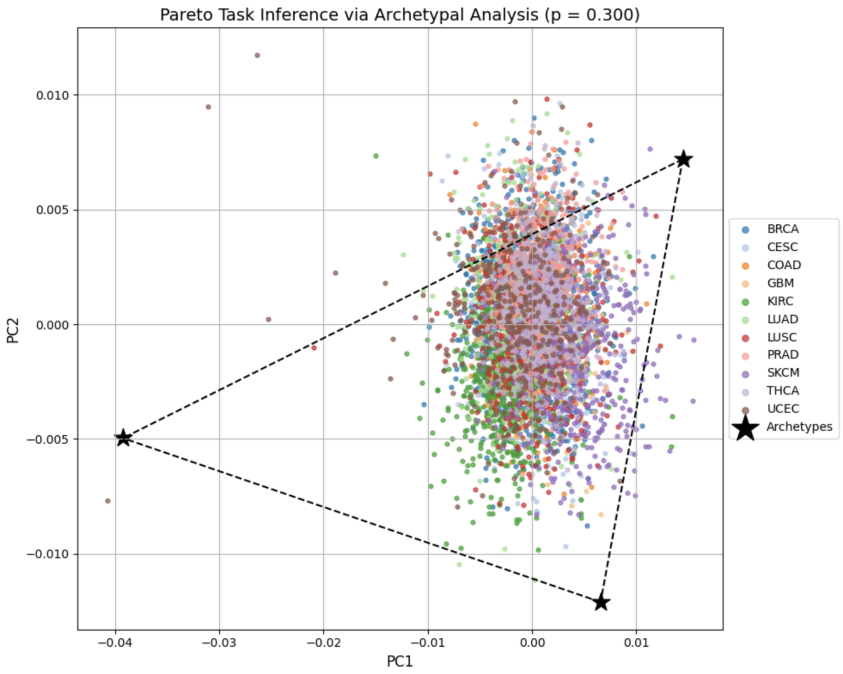


Figure 5: Pareto Task Inference via Archetypal Analysis in the latent PC1–PC2 plane. Colored points correspond to individual samples, while black stars mark the inferred archetypes. The dashed lines outline the Pareto front; $p = 0.300$ denotes the permutation-test probability that a simplex of equal or greater area would arise by chance under task reshuffling.

for understanding trade-offs between classification objectives in multi-label biomedical tasks. It provides an interpretable framework for identifying shared and distinct biological mechanisms across diseases, especially in scenarios where multiple diagnostic or therapeutic goals must be balanced.

Despite the encouraging results, there are limitations to address in future work. First, the current analysis is limited to gene expression data; incorporating additional omics layers such as DNA methylation, mutation profiles, or histopathology images could enrich the model's understanding of tumor biology. Second, the VAE architecture can be enhanced using more advanced priors or hierarchical structures to better capture discrete and continuous sources of variation. Finally, clinical translation would require external validation on independent cohorts and integration into a diagnostic decision support framework.

This work highlights how unsupervised and supervised deep learning models can be combined to analyze cancer heterogeneity and build interpretable latent representations for downstream clinical applications. With further development, such approaches hold promise for improving patient stratification, biomarker discovery, and precision oncology.

# References

[1] Leigh A Cantrell, Stephanie V Blank, and Linda R Duska. "Uterine carcinosarcoma: a review of the literature". In: *Gynecologic Oncology* 137.3 (2015), pp. 581–588.

[2] Philip B Clement and Robert H Young. "Endometrioid carcinoma of the uterine corpus: a review of its pathology with emphasis on recent advances and problematic aspects". In: *Advances in anatomic pathology* 9.3 (2002), pp. 145–184.

[3] Adele Cutler and Leo Breiman. "Archetypal analysis". In: *Technometrics* 36.4 (1994), pp. 338–347.

[4] Lilian T Gien, Marie-Claude Beauchemin, and Gillian Thomas. "Adenocarcinoma: a unique cervical cancer". In: *Gynecologic oncology* 116.1 (2010), pp. 140–146.

[5] Douglas Hanahan. "Hallmarks of cancer: new dimensions". In: *Cancer discovery* 12.1 (2022), pp. 31–46.

[6] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation". In: *cell* 144.5 (2011), pp. 646–674.

[7] Farina Hanif et al. "Glioblastoma multiforme: a review of its epidemiology and pathogenesis through clinical presentation and treatment". In: *Asian Pacific journal of cancer prevention: APJCP* 18.1 (2017), p. 3.

[8] Yuval Hart et al. "Inferring biological tasks using Pareto analysis of high-dimensional data". In: *Nature methods* 12.3 (2015), pp. 233–235.

[9] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[11] Shu Liao et al. "Feature selection with Pareto optimization for multi-class classification". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2014), pp. 528–535.

[12] Xianxu Lin et al. "Pareto Multi-Task Learning". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 21946–21958.

[13] Georgina V Long et al. "Cutaneous melanoma". In: *The Lancet* 402.10400 (2023), pp. 485–502.

[14] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[15] Ioannis Nesseris et al. "Cutaneous metastasis of colon adenocarcinoma: case report and review of the literature". In: *Anais Brasileiros de Dermatologia* 88.6 Suppl 1 (2013), pp. 56–58.

[16] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[17] Marius-Constantin Popescu et al. "Multilayer perceptron and neural networks". In: *WSEAS Transactions on Circuits and Systems* 8.7 (2009), pp. 579–588.

[18] Tracy L Rose and William Y Kim. "Renal cell carcinoma: a review". In: *Jama* 332.12 (2024), pp. 1001–1010.

[19] Eric Ka Ho Shea, Valerie Cui Yun Koh, and Puay Hoon Tan. "Invasive breast cancer: Current perspectives and emerging views". In: *Pathology international* 70.5 (2020), pp. 242–252.

[20] Gregory P Way and Casey S Greene. "Machine learning detects pan-cancer RNA expression signatures". In: *BioRxiv* (2018), p. 372383.

[21] John N Weinstein et al. "The Cancer Genome Atlas Pan-Cancer analysis project". In: *Nature Genetics* 45.10 (2013), pp. 1113–1120.