Interpretable Surrogate Modeling for TCGA RNA-Seq Cancer Type Classification

Subhash Saravanan, Manish Ram

A Class Project Report

submitted in partial fulfillment of the

requirements of the class CSS 490/590

University of Washington, Bothell

August 23, 2025

Abstract

The successful application of deep learning models in genomics, particularly for cancer type classification using RNA-Seq data, has been hampered by their inherent "black box" nature, which limits clinical trust and adoption. This project addresses the critical gap between high predictive accuracy and model interpretability. We developed a high-performance 1D Convolutional Neural Network (CNN) for classifying 32 different cancer types from The Cancer Genome Atlas (TCGA) RNA-Seq data, achieving a final test accuracy of **95.4%**, supported by a 5-fold cross-validation accuracy of **95.02%**. To make the model's decisions transparent, we employed knowledge distillation to train an interpretable Soft Decision Tree (SDT) as a surrogate model. The SDT learns to mimic the complex decision boundaries of the teacher CNN by training on its probabilistic outputs (soft labels). The resulting SDT achieved a fidelity of 88.6% to the teacher model and an accuracy of 87.5% on true labels, significantly outperforming a standard decision tree which scored 45%. This approach successfully translates the predictive power of a complex neural network into a simple, human-readable set of if-then rules, or a "Decision Fingerprint," for each cancer type, thereby creating a more trustworthy and clinically applicable diagnostic tool.

1. Introduction

1.1 Research Question and Motivation

Gene expression profiling via RNA sequencing (RNA-Seq) has become a cornerstone of modern oncology. This technology provides vast, high-dimensional datasets that can be leveraged to classify tumors, predict patient outcomes, and uncover novel therapeutic targets. Deep learning models, such as Convolutional Neural Networks (CNNs), have demonstrated remarkable accuracy on this complex data. However, their clinical utility is severely limited by their lack of transparency. These "black box" models provide predictions without clear, human-understandable reasoning, making it difficult for clinicians to trust their outputs for critical diagnostic decisions.

Traditional machine learning models like logistic regression or standard decision trees are inherently interpretable but often falter when faced with the high dimensionality of RNA-Seq data, a phenomenon known as the "curse of dimensionality," leading to weaker predictive performance. This creates a critical trade-off between accuracy and interpretability. Clinicians require more than just a prediction; they need verifiable, rule-based logic that aligns with biological reasoning to have confidence in a model's output.

This project directly addresses this challenge with the central research question: ***Can we distill the knowledge from a complex, high-accuracy deep learning model into a simpler, interpretable surrogate model for TCGA RNA-Seq cancer classification without a significant loss of performance?*** Our objective is to develop a hybrid system that leverages the predictive strength of a CNN and translates its decision-making process into a set of explicit, human-readable rules that can be trusted and validated by medical professionals.

1.2 Background and State-of-the-Art

RNA-Seq technology captures the expression levels of thousands of genes simultaneously, offering a comprehensive snapshot of a tumor's biological state. This rich dataset forms a powerful foundation for understanding tumor biology and developing molecular classifiers. **The Cancer Genome Atlas (TCGA) is a landmark pan-cancer project**, having generated a massive public repository of RNA-Seq data across dozens of cancer types. This dataset enables large-scale computational analyses like the one presented here, where models are trained to distinguish between many different types of cancer simultaneously.

The state-of-the-art in this field has seen a rapid adoption of deep learning. **Mostavi et al. (2020) established a strong baseline**, demonstrating the effectiveness of various CNN architectures (1D, 2D, and Hybrid) for pan-cancer classification on TCGA data. Their work highlighted the ability of CNNs to automatically learn relevant features from raw gene expression values, outperforming many traditional methods. However, these models remained largely uninterpretable.

To bridge this interpretability gap, the concept of **knowledge distillation** has emerged. **Frosst and Hinton (2017)** introduced a novel approach to distill the knowledge from a trained neural network into a "Soft Decision Tree" (SDT). Unlike standard decision trees with hard splits, an SDT uses probabilistic routing at each node, allowing it to better approximate the smooth decision boundaries of a neural network. By training the SDT on the rich, probabilistic outputs (soft labels) of the teacher network, the simpler tree model learns to mimic its behavior, effectively translating its "knowledge" into a more transparent structure.

1.3 Project Overview

This project follows a systematic design to build and interpret a pan-cancer classification pipeline. The core steps of our methodology are illustrated in the workflow diagram below.

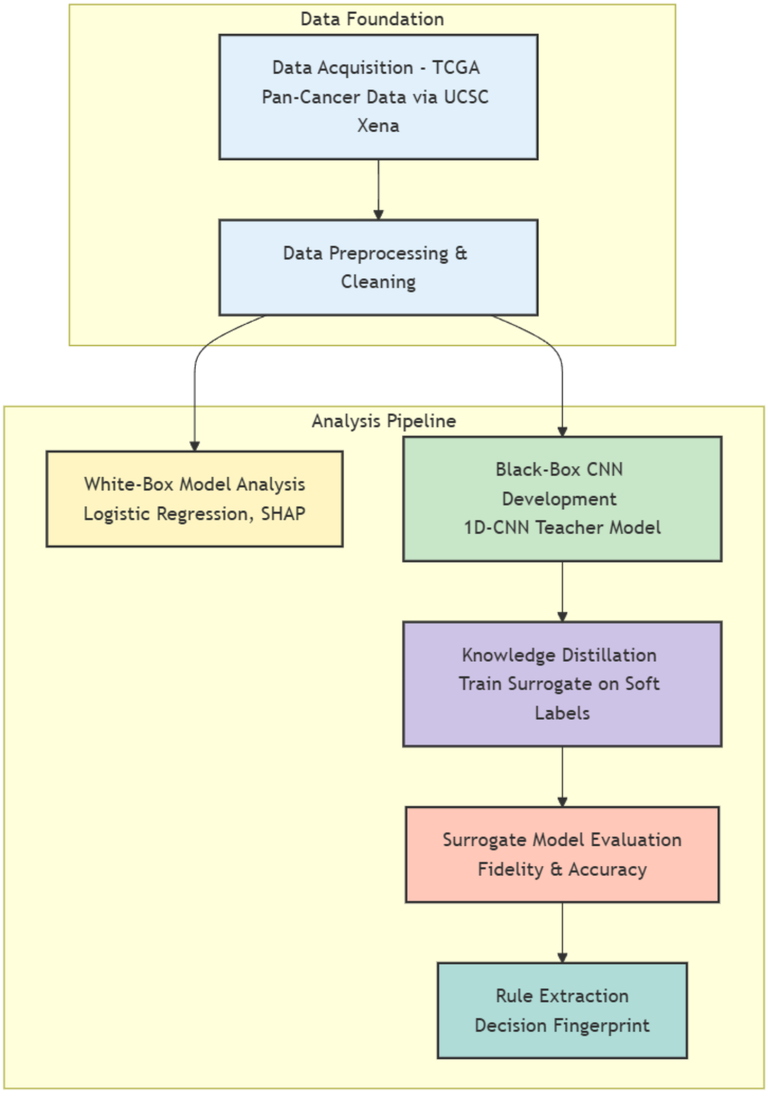
Data Foundation
Data Acquisition - TCGA Pan-Cancer Data via UCSC Xena
Data Preprocessing & Cleaning
Analysis Pipeline
White-Box Model Analysis Logistic Regression, SHAP
Black-Box CNN Development 1D-CNN Teacher Model
Knowledge Distillation Train Surrogate on Soft Labels
Surrogate Model Evaluation Fidelity & Accuracy
Rule Extraction Decision Fingerprint

Diagram 1: *The overall workflow of the project, from data acquisition to the extraction of interpretable rules.*

2. Related Work

The challenge of creating accurate and interpretable models for genomic data has been approached from various angles. Our work builds upon several key areas of research, primarily deep learning applications in cancer genomics, the emerging field of knowledge distillation for model explanation, and model-agnostic interpretability techniques.

2.1 Deep Learning for Pan-Cancer Classification

The use of CNNs for classifying cancer types from gene expression data was comprehensively benchmarked by **Mostavi et al. (2020)**. They developed 1D, 2D, and hybrid 2D CNN architectures and applied them to the TCGA pan-cancer RNA-Seq dataset. Their 1D-CNN, which treats the gene expression profile as a one-dimensional signal, proved to be highly effective, achieving over 95% accuracy. They also explored transforming the 1D gene vector into a 2D "pseudo-image" to leverage traditional computer vision architectures. Their work provides a critical foundation for our project, as we adopt their 1D-CNN architecture as our high-performance "teacher" model due to its excellent balance of accuracy and computational efficiency.

2.2 Knowledge Distillation for Model Interpretability

The core of our interpretability approach is based on the work of **Frosst and Hinton (2017)**, who introduced the concept of Soft Decision Trees. They proposed training a smaller, more transparent tree model by minimizing the cross-entropy between its output distribution and the "soft labels" (class probabilities) generated by a larger, pre-trained neural network. This process, a form of knowledge distillation, allows the tree to learn the nuanced decision boundaries of the teacher network. Our project directly implements this methodology, using the soft labels from our 1D-CNN to train an SDT, thereby creating a rule-based surrogate.

2.3 Model-Agnostic and Visual Interpretation Techniques

To validate the internal workings of our models, particularly the black-box CNN, we draw on techniques from model-agnostic interpretability. The SHAP framework, introduced by **Lundberg and Lee (2017)**, provides a unified approach to computing feature importances based on Shapley values from cooperative game theory. SHAP values explain the prediction of an instance by computing the contribution of each feature to the prediction. We utilize SHAP to analyze the feature attributions of our baseline white-box models and our teacher CNN, allowing us to identify key biomarkers and potential data biases, such as the influence of gender-specific genes.

Visual explanation techniques for CNNs also offer a path toward interpretability. **Selvaraju et al. (2017)** developed Gradient-weighted Class Activation Mapping (Grad-CAM), which uses the gradients flowing into the final convolutional layer to produce a coarse localization map highlighting important regions in an input image. While developed for images, this technique can be adapted to genomic data to verify that the CNN is focusing on biologically relevant genes, a potential avenue for our future work.

2.4 Advanced Distillation and Broader Applications

More advanced distillation methods have also been explored. **Kim et al. (2020)** proposed "feature-map distillation," which improves the fidelity of the student model by forcing it to learn the intermediate feature representations of the teacher network, not just the final output. **Liu et al. (2022)** added further refinements by introducing new evaluation metrics and regularization techniques to control the complexity of the distilled model. While our project focuses on the foundational output-based distillation of Frosst and Hinton, these works represent important future directions for improving surrogate model performance. Other works, such as that by **Zhang et al. (2021)**, have also explored distilling knowledge from complex models like BERT into simpler ones like decision trees for natural language processing tasks, showing the broad applicability of this concept.

2.5 Positioning the Current Work

Finally, numerous studies have applied machine learning to the TCGA dataset for various tasks, including survival prediction and subtype classification. Our work contributes to this body by focusing specifically on bridging the gap between the high accuracy of deep learning classifiers and the clinical need for interpretable, rule-based systems.

3. Method

3.1 Data Acquisition and Preprocessing

Our project utilized publicly available RNA-Seq gene expression and corresponding clinical phenotype data from **The Cancer Genome Atlas (TCGA) pan-cancer cohort**.

3.1.1 Data Source and Preparation

Initial data acquisition was attempted via the Genomic Data Commons (GDC) Data Portal. However, this approach presented significant computational challenges due to the large file sizes of the datasets. For instance, the dataset for Breast Invasive Carcinoma (BRCA) alone exceeded 2.5 GB. The considerable size and complexity of the raw data from the GDC portal led to performance issues and inefficiencies, particularly within the Google Colaboratory environment used for this analysis. To overcome these limitations, we subsequently utilized the UCSC Xena platform. The Xena browser provides pre-compiled and curated TCGA datasets that are more manageable in size and optimized for computational analysis, thereby streamlining the data acquisition process. The RNA-Seq gene expression data and the associated clinical phenotype data were downloaded as separate files from the Xena platform.

3.1.2 Data Preprocessing and Integration

The raw RNA-Seq dataset contained expression levels for 20,531 genes across 11,070 samples. The corresponding phenotype dataset included information for 12,068 samples, with three key features: 'Sample ID', 'Cancer Type' (detailing the specific malignancy), and 'Tumor Status' (classifying samples as 'Primary Tumor' or 'Solid Tissue Normal').

A data integration pipeline was developed to merge these two disparate datasets into a unified dataset. First, the RNA-Seq data matrix was transposed to orient the data with samples as rows and genes as columns. Subsequently, the transposed RNA-Seq matrix was merged with the phenotype dataset using the unique 'Sample ID' column, which served as the primary key common to both files. This merge operation resulted in a combined dataset consisting of 10,441 samples and 20,534 columns (20,531 gene features and 3 clinical features).

3.1.3 Final Dataset Curation

The final step involved data cleaning and curation. Samples containing a significant number of missing values (NaNs) were systematically removed from the merged dataset. During this process, it was observed that samples corresponding to Acute Myeloid Leukemia (LAML) had a disproportionately high incidence of missing values. Consequently, all LAML samples were excluded. This step resulted in a final dataset comprising 32 distinct cancer types.

In preparing the data for our models, **we followed the literature, specifically adopting preprocessing guidelines from Mostavi et al. (2020)**. This ensured our data was formatted in a way that was comparable to their baseline study, allowing for a more direct and meaningful comparison of model performance.

Dataset Combination & Filtering: To reduce dimensionality and remove noise, we applied:

1. **Data Normalization:** The gene expression data was standardized using `StandardScaler` to have a mean of 0 and a standard deviation of 1.
2. **Low-Variance Gene Filtering:** We filtered out genes with low variance across samples using the criteria: `(gene_mean >= 0.5) AND (gene_std >= 0.8)`. This reduced the feature space from 20,531 to **12,854** informative genes.

3.2 Class Distribution

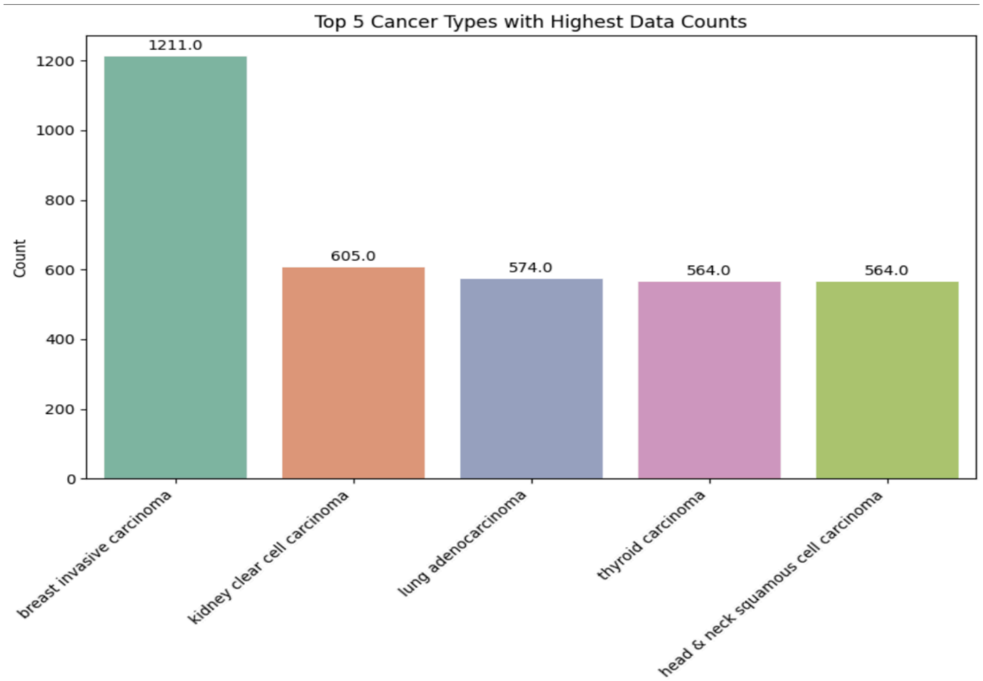

Diagram 2: *Bar chart showing the class distribution of the cancer types with more than 400 examples*

This distribution shows the cancer types with the highest sample counts. Breast carcinoma has the highest count with 1,211 samples, followed by kidney cell carcinoma (605) and lung carcinoma (547).

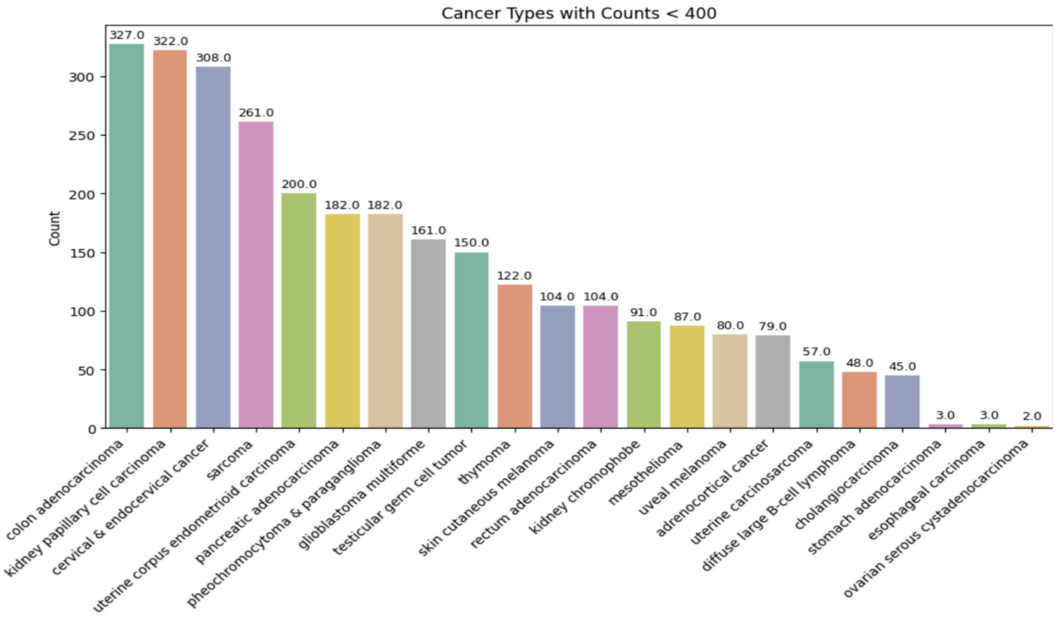

Diagram 3: *Bar chart showing the class distribution of the cancer types with less than 400 examples*

An exploratory analysis of the final dataset revealed a significant class imbalance across the 32 cancer cohorts. The distribution of samples per cancer type was highly skewed, with a mean of approximately 277 samples per class and a large standard deviation of 258. This high degree of variance indicates that while some cancer types were well-represented, others comprised only a small fraction of the total dataset. A number of cancer types contained fewer than ten samples, creating a long-tail distribution. For instance, the cohorts for stomach adenocarcinoma, esophageal carcinoma, and ovarian serous cystadenocarcinoma were severely underrepresented, containing only three samples each.

3.3 Experimental Setup

The preprocessed data was split into training (72%), validation (8%), and test (20%) sets using stratified sampling to preserve the distribution of the 32 cancer types. This is crucial given the severe class imbalance, which we addressed by applying class weights during model training to penalize errors on minority classes more heavily.

3.4 Model Architectures

3.4.1 White-Box Baseline Models:

Our experimental design began with establishing a performance and interpretability baseline using several well-understood "white-box" models. These models are inherently transparent, meaning their decision-making logic can be directly examined. We trained a **Logistic Regression** model as a linear baseline, a single **Decision Tree** to evaluate a simple rule-based approach, and a **Random Forest** to assess the performance of a more powerful ensemble method. These models were trained to benchmark the maximum achievable performance with traditional techniques and served as the initial subject of our feature importance analysis using SHAP, allowing us to gain preliminary insights into the dataset before moving to more complex architectures.

3.4.2 Black-Box Deep Learning Models:

We developed and tested three distinct CNN architectures to identify the most effective "teacher" model for our knowledge distillation task. Each model was designed to capture different aspects of the gene expression data. These models serve as the "black-box" component of our pipeline, prioritizing predictive accuracy over inherent interpretability.

The first and simplest architecture was a **1D-CNN**, also known as a point-wise convolutional model. This model treats the 12,854-gene expression profile as a one-dimensional signal. It utilizes 1D convolutions with a kernel size of 1, which function as point-wise feature extractors. This process is equivalent to learning a weighted combination of expression values for each gene independently before aggregating these learned features in deeper, fully connected layers. This architecture is the most efficient of the three, with the smallest model size at approximately 1 MB and the fastest training time. Its primary strength lies in its ability to efficiently detect localized signals and determine feature importance at the individual gene level. However, its main limitation is that it may miss complex spatial or cross-gene interactions that could be captured by treating the data as a 2D structure.

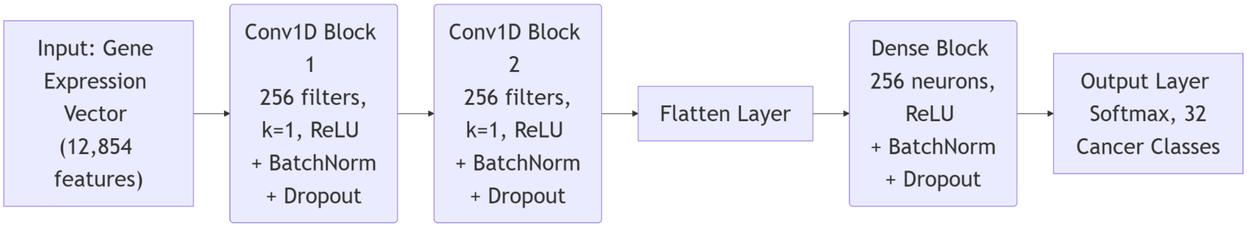
Input: Gene Expression Vector (12,854 features)
Conv1D Block 1 256 filters, k=1, ReLU + BatchNorm + Dropout
Conv1D Block 2 256 filters, k=1, ReLU + BatchNorm + Dropout
Flatten Layer
Dense Block 256 neurons, ReLU + BatchNorm + Dropout
Output Layer Softmax, 32 Cancer Classes

Diagram 4: *The architecture of the 1D-CNN.*

The second architecture we evaluated was a **2D-CNN**, which processes the gene expression data as a pseudo-image. This model first reshapes the flat 1D vector of 12,854 genes into a 2D grid, allowing the use of standard 2D convolutional filters. These filters scan small, local blocks of "genes" to learn potential spatial patterns and dependencies between genes that are arranged adjacently in the grid. This mid-size model, at approximately 4 MB, is slower to train than its 1D counterpart but possesses the unique strength of capturing potential spatial relationships within the data. Its principal limitation is that this spatial arrangement is artificial; its effectiveness hinges on whether functionally related genes happen to be co-located in the reshaped grid, which is not guaranteed.

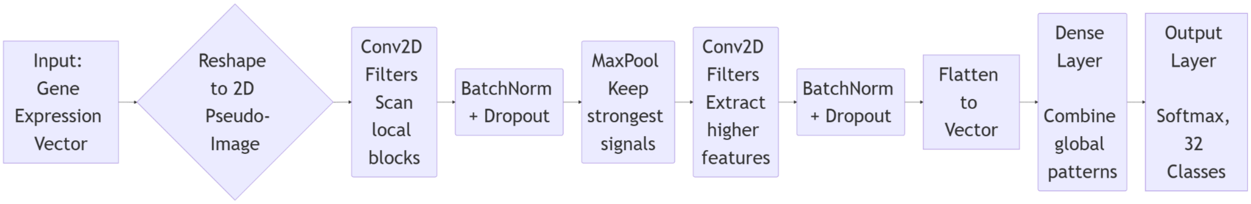
Input: Gene Expression Vector
Reshape to 2D Pseudo-Image
Conv2D Filters Scan local blocks
BatchNorm + Dropout
MaxPool Keep strongest signals
Conv2D Filters Extract higher features
BatchNorm + Dropout
Flatten to Vector
Dense Layer Combine global patterns
Output Layer Softmax, 32 Classes

Diagram 5: *The architecture of the 2D-CNN model.*

The third and most complex model was the **2D-Hybrid CNN**, a dual-branch architecture designed to learn both local (1D) and spatial (2D) features simultaneously. This model processes the input data in parallel through two distinct branches. The 1D branch operates on the original one-dimensional signal, while the 2D branch processes the reshaped pseudo-image. The feature maps from both branches are then flattened, concatenated into a single, unified vector, and fed into fully connected layers for the final classification. As the largest model (~6 MB), it is also the slowest to train. Its primary strength is its comprehensive approach, combining the point-wise feature extraction of the 1D-CNN with the spatial learning of the 2D-CNN, making it theoretically the most powerful at modeling complex cross-gene relationships. However, its increased complexity also elevates the risk of overfitting if not properly regularized.

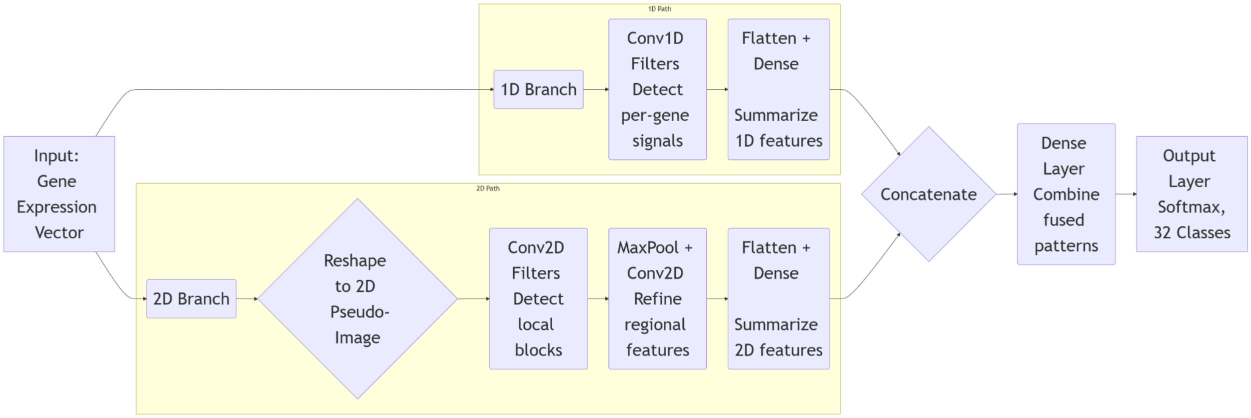
1D Path
Input: Gene Expression Vector
1D Branch
Conv1D Filters Detect per-gene signals
Flatten + Dense Summarize 1D features
2D Path
2D Branch
Reshape to 2D Pseudo-Image
Conv2D Filters Detect local blocks
MaxPool + Conv2D Refine regional features
Flatten + Dense Summarize 2D features
Concatenate
Dense Layer Combine fused patterns
Output Layer Softmax, 32 Classes

***Diagram 6:** The architecture of the 2D-Hybrid CNN model.*

3.4.3 Interpretable Surrogate Model (Soft Decision Tree)

To serve as the interpretable "student" model in our knowledge distillation pipeline, we selected a **Soft Decision Tree (SDT)**. This architecture was chosen because, unlike a standard decision tree with rigid, axis-aligned splits, an SDT is designed to approximate the smooth and complex decision boundaries of a neural network. Its structure is a binary tree of a predefined depth, where each internal node functions as a probabilistic router. Instead of making a hard "yes" or "no" decision, each node applies a sigmoid function to a linear combination of the input features. The output of this function determines the probability of a given sample traversing to the right or left child.

Consequently, every sample travels down all possible paths to the leaves, and its final prediction is a weighted average of the class distributions stored at each leaf node. The weights are the path probabilities calculated along the way. This probabilistic framework allows the SDT to achieve better generalization and more accurately mimic the behavior of a neural network. While slightly less intuitive than a hard tree, its hierarchical, rule-based nature makes it vastly more transparent than the teacher CNN, providing a **global interpretation** of the complex model's overall decision logic.

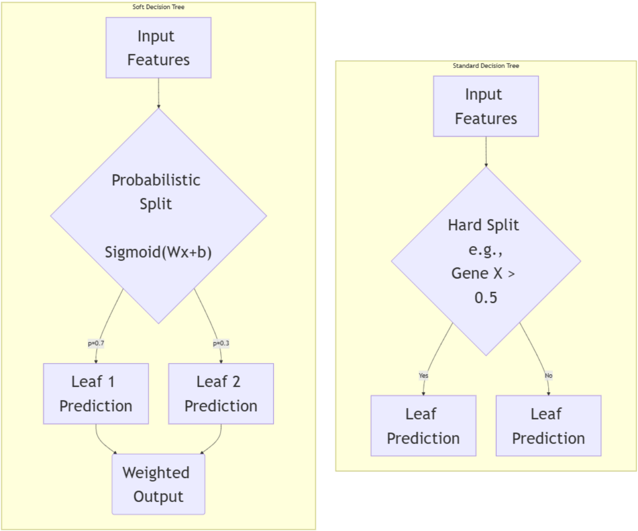

Diagram 7: *Comparison of a standard decision tree versus a Soft Decision Tree with probabilistic routing.*

3.5 Knowledge Distillation and Rule Extraction

The core of our method lies in transferring the learned function of the 1D-CNN into the interpretable SDT. This process, known as knowledge distillation, begins with training the 1D-CNN teacher on the original data. Once trained, this network is used to generate "soft labels"—full class probability distributions—for the entire training set. The Soft Decision Tree is then trained on the same input features but uses the CNN's soft labels as the target, minimizing the cross-entropy between their respective distributions. This encourages the simpler tree model to mimic the nuanced decision function of its more complex teacher. Finally, the learned weights at each internal node of the trained SDT are analyzed to identify the most influential genes, allowing for the construction of a human-readable "Decision Fingerprint" by tracing the most probable path for any given class.

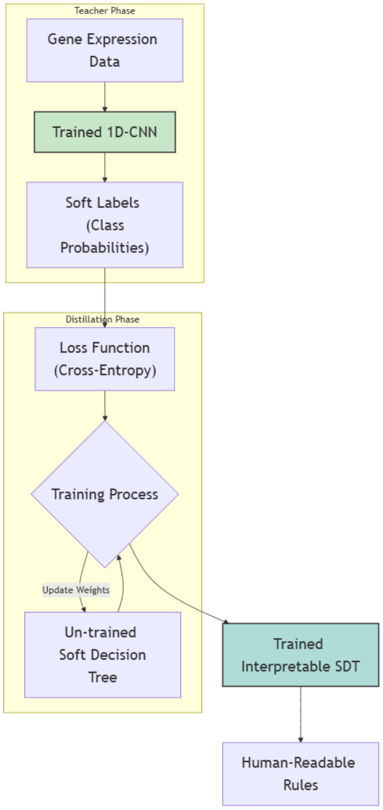
Teacher Phase
Gene Expression Data
Trained 1D-CNN
Soft Labels (Class Probabilities)
Distillation Phase
Loss Function (Cross-Entropy)
Training Process
Update Weights
Un-trained Soft Decision Tree
Trained Interpretable SDT
Human-Readable Rules

Diagram 8: *The knowledge distillation process where the SDT (student) learns from the CNN's (teacher) soft labels.*

4. Results and Analysis

This project successfully developed and evaluated several models to classify 32 different cancer types from TCGA RNA-Seq data, focusing on bridging the gap between high-accuracy "black box" models and interpretable "white box" models.

PCA Plot Analysis:

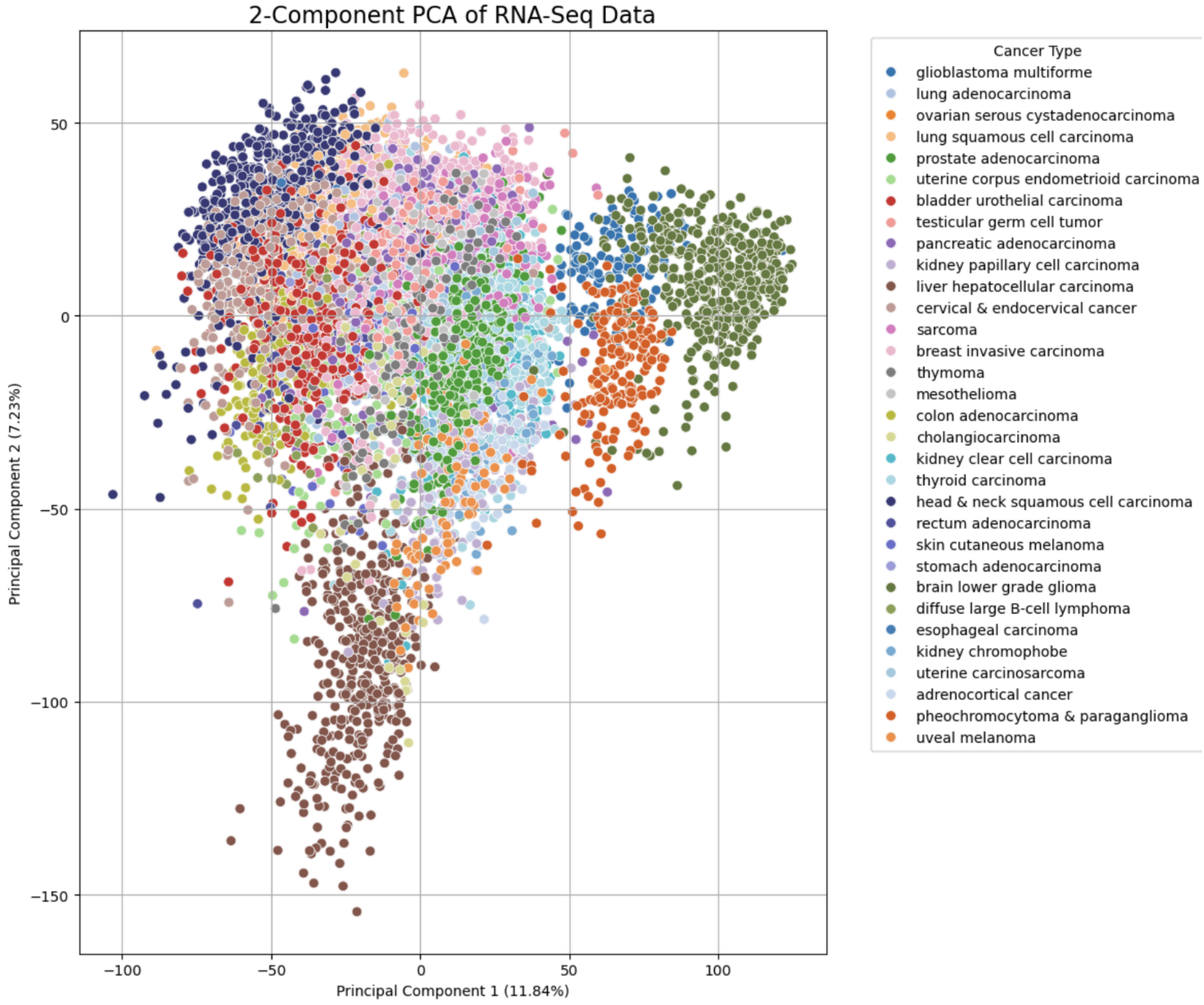

Diagram 9: *PCA Plot for RNA-Seq data after cleaning*

Cluster Formation: There are some noticeable clusters of cancer types visible in the plot, suggesting that gene expression profiles can, to some extent, distinguish between different cancers. For example, Liver hepatocellular (brown) carcinoma and Glioblastoma multiforme (dark blue) and head &neck squamous cell carcinoma(purple) form clusters.

Class Overlap**:** Despite some clustering, there is significant overlap between many of the classes on the 2D plot. This indicates that the first two principal components are not sufficient to separate all 32 cancer types cleanly. This would mean that while a 2D plot is not able to cluster the data in higher dimension is able to separate the classes more clearly and this would mean that a Hyperplace in higher dimension more than 2-D can separate the cancer clusters.

This would mean it would really be useful to include a Logistic Regression for testing as the fundamental goal of a Logistic Regression model is to find the best line, plane, or hyperplane that separates the different classes in the feature space.

4.1 White Box Model Performance

Initial experiments were conducted using interpretable models: Random Forest and Decision Trees. The performance of these models was evaluated under two conditions: with a maximum tree depth of 5 and with an uncapped maximum depth. We included logistic regression based on the findings from the PCA plot from the previous section.

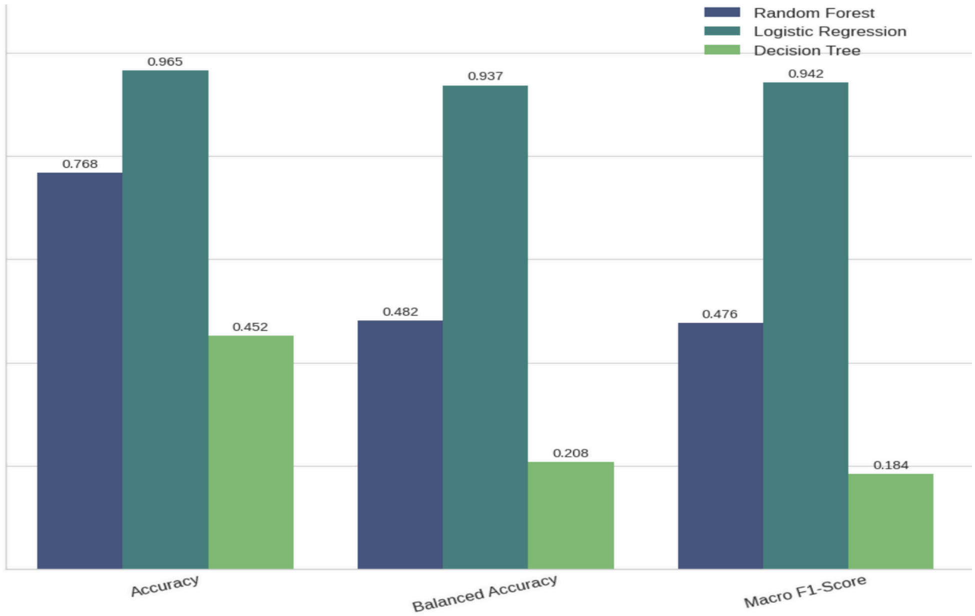

Diagram 10: *Performance Metric for Tree Depth = 5*

With the tree depth limited to 5, the models exhibited poor performance due to underfitting. In this scenario, Logistic Regression was the clear top performer, achieving a Balanced Accuracy of 0.937 and a Macro F1-Score of 0.942. The Decision Tree performed the worst, with a Balanced Accuracy of only 0.208 and a Macro F1-Score of 0.184.

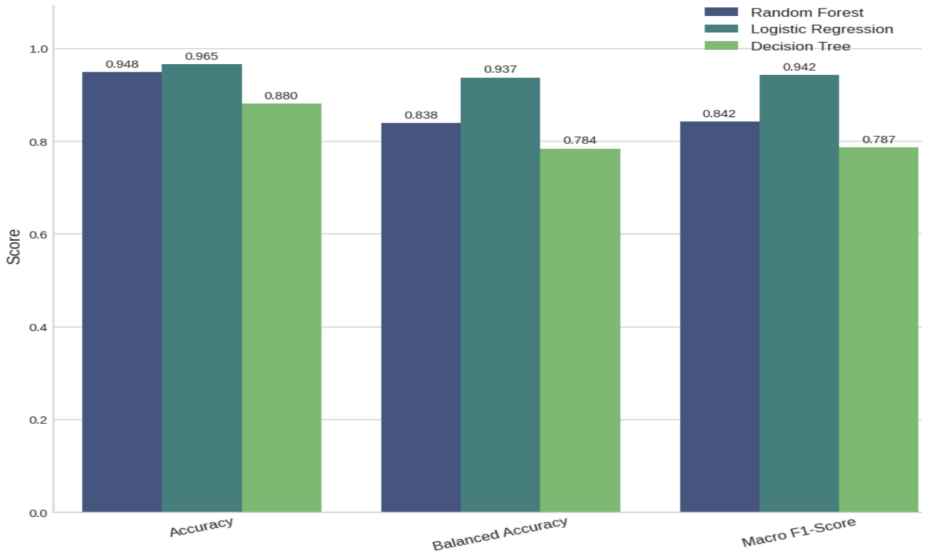

Diagram 11: *Performance Metric for Tree Depth = Uncapped*

When the tree depth was uncapped, performance improved significantly for the tree-based models. The Random Forest model achieved a Balanced Accuracy of 0.838 and a Macro F1-Score of 0.842. The Decision Tree model's scores increased substantially to a Balanced Accuracy of 0.784 and a Macro F1-Score of 0.787. Throughout both experiments, Logistic Regression's scores remained the highest, and Random Forest generally performed better than Decision Trees.

4.1.1 Logistic Regression SHAP value Analysis

The initial analysis revealed a significant bias in the model, which learned to use gender-specific genes as powerful shortcuts for classification instead of identifying cancer-specific biological markers.

For Testicular Germ Cell Tumor: The model's predictions were based entirely on genes from the Y chromosome, such as **'RPS4Y1', 'DDX3Y', and 'KDM5D'**. This indicates the model simply identified the patient as male rather than analyzing the tumor's specific gene expression profile.

For Breast Invasive Carcinoma: Predictions were similarly biased, dominated by the female marker '**XIST' and Y-chromosome genes like 'RPS4Y1' and 'DDX3Y'.** The model was leveraging the patient's sex as the primary predictive feature.

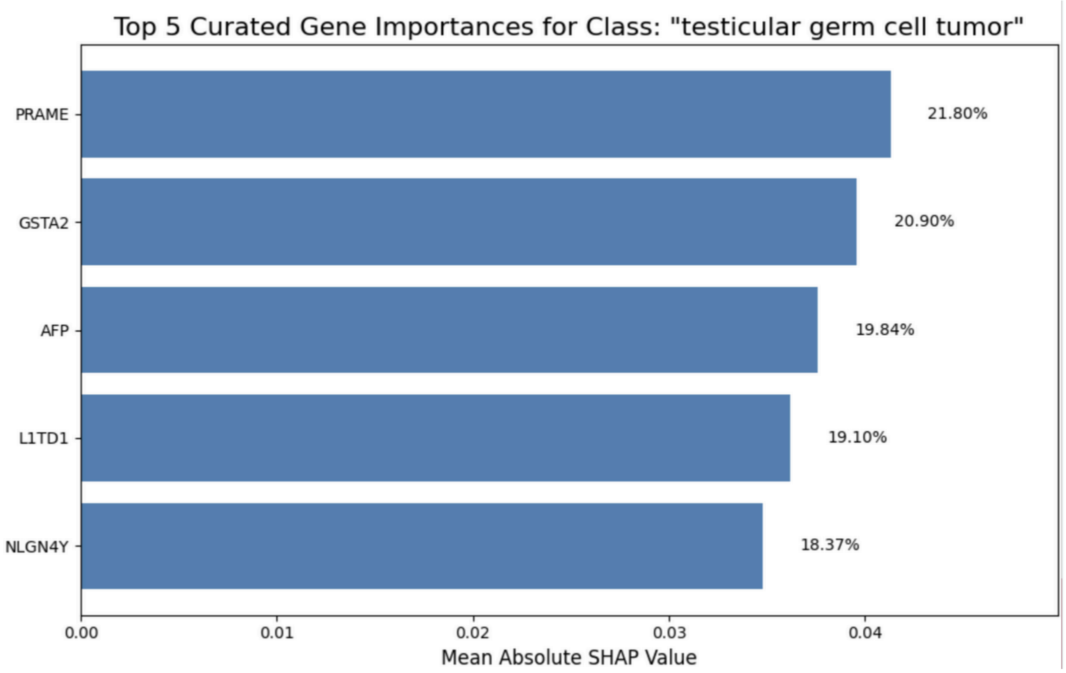

Diagram 12: *SHAP value for top 5 genes expressed in Testicular Tumor (After X&Y Genes removed)*

After the biased, gender-specific genes were removed, the model was forced to learn from more subtle, clinically relevant biological signals in the data.For Testicular Germ Cell Tumor: The model identified **'PRAME'**, a known cancer-testis antigen, and **'AFP'** (Alpha-fetoprotein), which is a clinical biomarker used in diagnosis.

For Breast Invasive Carcinoma: The model highlighted important genes such as **'SCGB2A2'**, which is expressed in breast tissue, and **'AZGP1'**, a protein secreted by breast epithelial cells.

4.2 Black-Box Teacher Model Performance

After establishing the performance of our white-box models, we evaluated the three deep learning architectures using 5-fold cross-validation to identify the best teacher model. The results are summarized in Table 2.

Model	**Accuracy**	**Balanced Accuracy**	**Precision**	**Recall**	**F1-Score**	**MCC**
1D-CNN	**0.9502**	0.8998	0.9524	0.9502	**0.9474**	**0.9473**
2D-CNN	0.9411	0.9031	0.9483	0.9411	0.9402	0.9378
2D-Hybrid	0.9559	0.9167	0.9572	0.9559	0.9547	0.9533

Table 1: *5-Fold Cross-Validation Performance of CNN Models.*

The analysis showed that all three CNN models performed exceptionally well, significantly outperforming the interpretable baseline models in terms of balanced accuracy and F1-score. The 2D-Hybrid model achieved the highest overall accuracy and MCC score, likely due to its ability to capture both point-wise and spatial features. However, the **1D-CNN** was highly competitive, with an accuracy of 95.02% and an F1-score of 94.74%, while being significantly smaller (~1 MB vs. ~6 MB) and faster to train. Given its excellent balance of high performance and computational efficiency, we selected the **1D-CNN** as our "teacher" model for the knowledge distillation phase. On the held-out test set, the final trained 1D-CNN achieved a top-line accuracy of **95.4%**.

4.2.1 Feature Importance in the 1D-CNN (SHAP Analysis)

To gain insight into the "black-box" 1D-CNN, we employed SHAP (SHapley Additive exPlanations) to analyze its feature importances. The analysis revealed that the model learned to prioritize several known cancer-associated genes. For instance, in classifying breast cancer, it assigned high importance to genes like **A1CF** (which may promote cell proliferation), **CYP2C9** (involved in metabolizing breast cancer drugs like tamoxifen), and **RBFOX1** (a potential tumor suppressor often reduced in breast cancer tissues). The model also highlighted lesser-known genes of interest, such as **EZHIP** and **UBE2Q2P2**, suggesting potential new avenues for investigation.

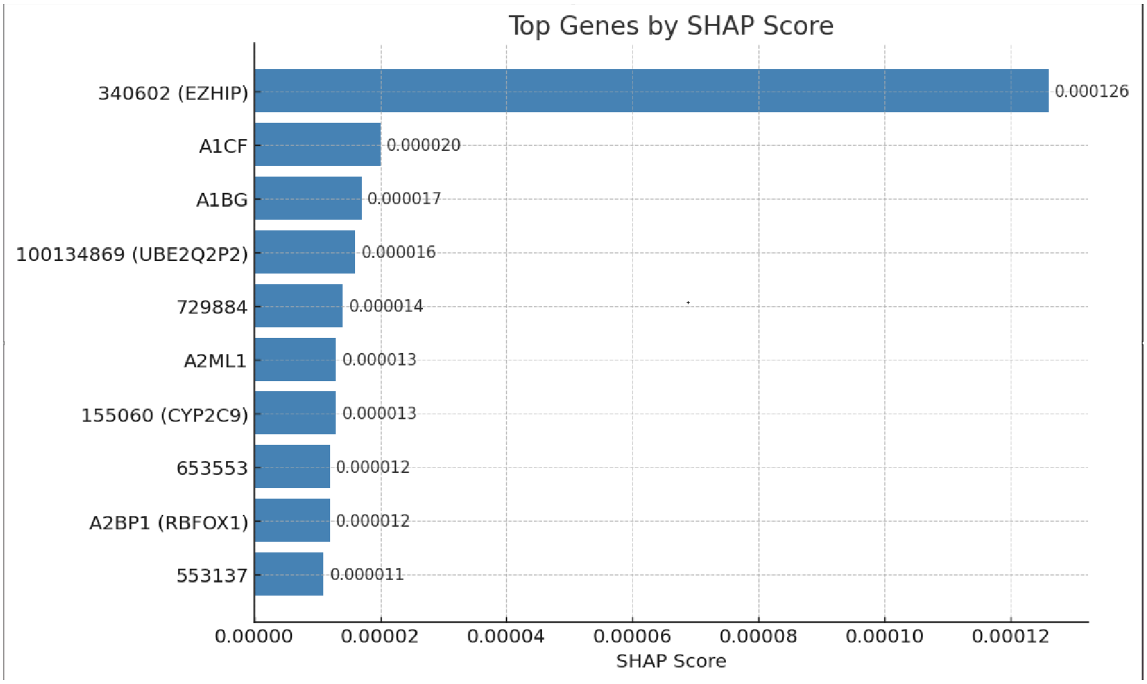

Diagram 13: *SHAP value for top genes expressed in Breast Invasive Carcinoma*

However, a notable finding was the absence of several major, well-established cancer biomarkers from the top-ranked features, including **BRCA1**, **BRCA2**, **HER2**, and **TP53**. This "missing markers" phenomenon is a critical insight, suggesting that either the variance-based feature filtering step removed them or the model's architecture learned to rely on a different combination of features to make its predictions. This underscores the need for interpretable models to validate and question the reasoning of even high-performing black-box systems.

4.3 Surrogate Model Evaluation

The primary goal of this project was to successfully distill the knowledge from the 95.4% accurate 1D-CNN into an interpretable Soft Decision Tree (SDT) of depth 5. The performance of this distillation process is measured by three key metrics:

- **Teacher Accuracy:** 95.4% (The performance of the 1D-CNN on the test set).
- **Soft Decision Tree Fidelity:** 88.6% (The SDT correctly predicted the CNN's output on 88.6% of test samples).
- **Soft Decision Tree Accuracy:** 87.5% (The SDT's accuracy on the true labels of the test set).

The high fidelity of 88.6% is a strong indicator that the SDT successfully learned to approximate the complex decision boundaries of the teacher network. Its final accuracy of 87.5% on the ground truth labels is remarkably high for a simple, tree-based model. This represents a massive improvement over a standard decision tree of the same depth, which achieved only **45%** accuracy on this dataset. This result demonstrates the power of knowledge distillation for creating models that are both accurate and interpretable, successfully bridging the gap between the two.

The full structure of the trained Soft Decision Tree is visualized in Diagram 7. While the complete tree appears complex due to its depth, its true interpretability lies not in viewing the entire structure at once, but in its ability to be deconstructed into simple, linear paths for each classification decision, as shown in the next section.

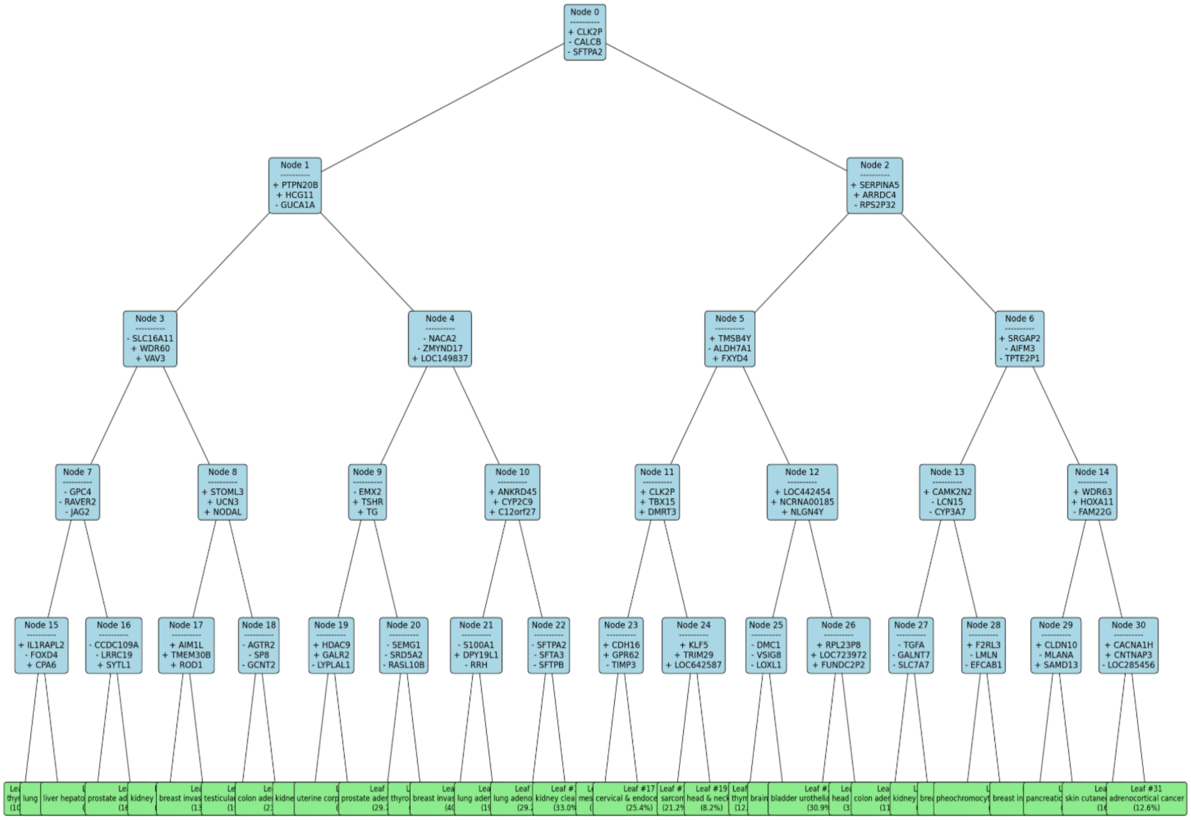
Node 0 + CLK2P - CALCB - SFTPA2
Node 1 + PTPN20B + HCG11 - GUCA1A
Node 2 + SERPINA5 + ARRDC4 - RPS2P32
Node 3 - SLC16A11 + WDR60 + VAV3
Node 4 - NACA2 - ZMYND17 + LOC149837
Node 5 + TMSB4Y - ALDH7A1 + FXYD4
Node 6 + SRGAP2 - AIFM3 - TPTE2P1
Node 7 - GPC4 - RAVER2 - JAG2
Node 8 + STOML3 + UCN3 + NODAL
Node 9 - EMX2 + TSHR + TG
Node 10 + ANKRD45 + CYP2C9 + C12orf27
Node 11 + CLK2P + TBX15 + DMRT3
Node 12 + LOC442454 + NCRNA00185 + NLGN4Y
Node 13 + CAMK2N2 - LCN15 - CYP3A7
Node 14 + WDR63 + HOXA11 - FAM22G
Node 15 + IL1RAPL2 - FOXD4 + CPA6
Node 16 - CCDC109A - LRRC19 + SYTL1
Node 17 + AIM1L + TMEM30B + ROD1
Node 18 - AGTR2 - SP8 - GCNT2
Node 19 + HDAC9 + GALR2 - LYPLAL1
Node 20 - SEMG1 - SRD5A2 - RASL10B
Node 21 - S100A1 + DPY19L1 - RRH
Node 22 - SFTPA2 - SFTA3 - SFTPB
Node 23 + CDH16 + GPR62 - TIMP3
Node 24 + KLF5 + TRIM29 + LOC642587
Node 25 - DMC1 - VSIG8 - LOXL1
Node 26 + RPL23P8 + LOC723972 + FUNDC2P2
Node 27 - TGFA - GALNT7 - SLC7A7
Node 28 + F2RL3 - LMLN - EFCAB1
Node 29 + CLDN10 - MLANA + SAMD13
Node 30 + CACNA1H + CNTNAP3 - LOC285456
lung
liver hepato
prostate ad
kidney
breast invas
testicular
colon aden
uterine corp
prostate aden (29.7
thyro
breast invas (40
lung aden (19
lung adeno (29.2
Leaf #1 kidney clea (33.0%
Leaf #17 cervical & endoce (25.4%)
sarcom (21.2%
Leaf #19 head & neck (8.2%)
thym (12.
brain
bladder urothelia (30.9%
head
colon aden
kidney
pheochromocyt
breast in
pancreatic
skin cutane
Leaf #31 adrenocortical cancer (12.6%)

Diagram 14: *Visualized Soft Decision Tree trained from the soft labels from 1D-CNN model*

4.4 Interpretable Rules from the Soft Decision Tree

The trained SDT provides a transparent, hierarchical set of rules for classification. Each internal node learns a linear separator, and the genes with the highest absolute weights at that node represent the most important features for its routing decision. By tracing the most probable path for a given cancer type from the root to a leaf, we can generate a "Decision Fingerprint."

The diagram below illustrates the Decision Fingerprint for **Breast Invasive Carcinoma**, which corresponds to the path to Leaf #12 in our trained tree. This step-by-step logic provides a clear, verifiable rationale for the model's prediction, using combinations of supporting and opposing gene sets at each stage, making it far more trustworthy for clinical review than a single probability score from a black-box model.

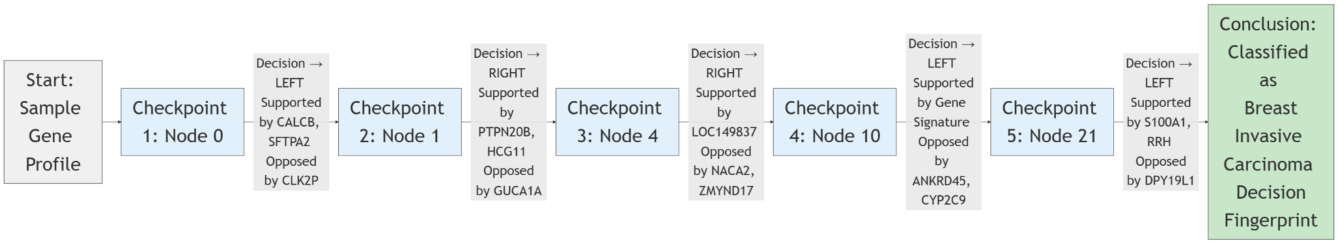
Start: Sample Gene Profile
Checkpoint 1: Node 0
Decision → LEFT Supported by CALCB, SFTPA2 Opposed by CLK2P
Checkpoint 2: Node 1
Decision → RIGHT Supported by PTPN20B, HCG11 Opposed by GUCA1A
Checkpoint 3: Node 4
Decision → RIGHT Supported by LOC149837 Opposed by NACA2, ZMYND17
Checkpoint 4: Node 10
Decision → LEFT Supported by Gene Signature Opposed by ANKRD45, CYP2C9
Checkpoint 5: Node 21
Decision → LEFT Supported by S100A1, RRH Opposed by DPY19L1
Conclusion: Classified as Breast Invasive Carcinoma Decision Fingerprint

Diagram 16: *Interpretable rule path for classifying Breast Invasive Carcinoma extracted from the SDT.*

5. Conclusion

5.1 Summary of Findings and Limitations

The primary objective of this project was to derive interpretable, human-readable rules from complex RNA-seq gene expression data for the purpose of pan-cancer classification. Our investigation confirmed the well-known trade-off between model performance and interpretability. We demonstrated that while deep learning models like Convolutional Neural Networks achieve high predictive accuracy (over 95%), their inherent "black box" nature limits their clinical utility, as their decision-making processes are not transparent. Conversely, traditional interpretable models such as Decision Trees were found to be more transparent but failed to deliver sufficient predictive accuracy on this high-dimensional dataset, especially when their complexity was constrained to maintain interpretability.

To address this challenge, we successfully implemented a knowledge distillation strategy, using a high-performing 1D-CNN as a "teacher" model to train a simpler, interpretable Soft Decision

Tree (SDT) as a surrogate. This approach proved highly effective. The resulting surrogate model achieved a fidelity of 88.6% to the teacher CNN, indicating it successfully approximated the teacher's complex decision function, and an absolute accuracy of 87.5% on the true labels. This represents a dramatic improvement over a standard decision tree of the same depth, which scored only 45% accuracy. The final model produces transparent, step-by-step rule paths, or "Decision Fingerprints," for cancer prediction, creating a trustworthy and clinically relevant tool that aligns with the needs of medical professionals.

Despite the project's success, a key limitation is the inherent trade-off between the teacher model's high accuracy and the surrogate model's complete interpretability. While the Soft Decision Tree's fidelity was high, the information loss of approximately 11% signifies that there are nuanced patterns and complex inter-gene relationships learned by the CNN that the simpler, hierarchical structure of the tree cannot fully capture. Furthermore, our feature importance analysis of the 1D-CNN revealed that some major cancer biomarkers, such as BRCA1 and TP53, were not among the top-ranked features. This "missing markers" phenomenon may be a result of our initial variance-based feature filtering or an indication that the CNN learned to rely on alternative, though equally predictive, biological signals. This highlights the ongoing challenge of ensuring that even accurate models are focusing on the most clinically established features.

5.2 Future Work

Based on the findings and limitations of this project, several avenues for future research are recommended to further enhance the model's clinical applicability and interpretability.

A valuable next step would be a **Comparative Interpretability Analysis**. This would involve a deep dive into the differences observed in SHAP values between the high-performing Logistic Regression model and the 1D-CNN. Such an investigation would help illuminate how different model architectures—linear versus non-linear—prioritize features and learn from genomic data. Understanding these differences could reveal which model type is better suited for uncovering specific kinds of biological insights and whether the CNN is identifying novel, non-linear interactions that the simpler model misses.

Furthermore, a significant improvement could be achieved through **Focused Modeling on High-Prevalence Cancers**. The current pan-cancer approach is challenged by severe class imbalance. A future iteration of this project would involve training a new teacher-student pipeline exclusively on the top 5-7 cancer types that have a substantial number of samples (e.g., >400). By eliminating the long-tail of underrepresented classes, the teacher CNN could potentially achieve even higher classification accuracy within this specific subset. Consequently, the knowledge distilled into the surrogate Soft Decision Tree would likely be more robust, leading to the generation of more stable and statistically significant interpretable rules for these specific, high-prevalence cancers.

Another critical extension would be the **Inclusion of Normal Tissue** samples in the training data. The current model is designed as a classifier to distinguish between different cancer types. By incorporating non-cancerous "Normal" tissue, the model could be trained not only to differentiate malignancies but also to distinguish cancerous tissue from healthy tissue. This would significantly enhance its clinical applicability, moving it from a subtyping tool to a potential diagnostic aid and allowing for the identification of gene expression patterns that are unique to tumorigenesis itself.

To further probe the inner workings of the teacher model, **Advanced CNN Interpretation** techniques should be employed. Methods like saliency maps, particularly Gradient-weighted Class Activation Mapping (Grad-CAM), could be used to directly visualize and interpret the predictions of the CNN models on the 2D pseudo-images of gene expression. This would provide a visual heatmap of which "gene hotspots" the model is focusing on, serving as a powerful tool to verify that its attention aligns with biologically relevant genes and pathways, thereby complementing the rule-based explanations from the surrogate model.

Finally, this project lays the groundwork for a more extensive **Cross-Cancer Gene Expression** analysis. The interpretable rules and feature importance rankings generated by our models can be systematically analyzed to identify genes that are consistently deemed important across multiple cancer types. Investigating these shared genes could reveal common biological pathways and vulnerabilities, potentially leading to the discovery of pan-cancer biomarkers or therapeutic targets that could have a broad impact across different oncological disciplines.

References

[1] E. Mostavi, Y. Chiu, Y. Huang, and P. C. Chen, “Convolutional neural network models for cancer type prediction based on gene expression,” BMC Medical Genomics, 2020.

[2] N. Frosst and G. Hinton, “Distilling a neural network into a soft decision tree,” arXiv:1711.09784, 2017.

[3] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” arXiv:1503.02531, 2015.

[4] The Cancer Genome Atlas Research Network, “The Cancer Genome Atlas Pan-Cancer analysis project,” Nature Genetics, 2013.

[5] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” NeurIPS, 2017.

[6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” ICML, 2017.

[7] B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” Biochimica et Biophysica Acta, 1975.

[8] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The balanced accuracy and its posterior distribution,” ICPR, 2010.

[9] M. J. Goldman et al., “Visualizing and interpreting cancer genomics data via the Xena platform,” Nature Biotechnology, 2020.

[10] R. L. Grossman et al., “Toward a shared vision for cancer genomic data,” New England Journal of Medicine, 2016.

[11] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[12] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. 32nd Int. Conf. Machine Learning (ICML), 2015.