

# From Machine Learning to Biology: Stacked Ensemble Models and Gene Interpretation in Cancer Classification

**Author: Tianyi (Amy) Gao, Kehan (Audrey) Qian**

**Date: 9.2.2025**

## Abstract

Gene expression data provides valuable insight into the molecular mechanisms of cancer, but its high dimensionality poses challenges for accurate prediction and clinical interpretation. In this study, we developed a stacked ensemble framework for cancer classification using gene expression data. A **Support Vector Machine (SVM)** was employed as the base-learner, and meta-learners including **Decision Trees (DT)**, **Random Forests (RF)**, and **Logistic Regression (LR)** were trained on the resulting stacked feature sets. To address the high dimensionality of gene expression data, we applied **Recursive Feature Elimination (RFE)** to identify the most informative features, and the models were retrained on these reduced feature sets. Model performance was evaluated using weighted F1-scores and cross-validation.

Our primary goal was to visualize the genes selected by the models and examine their biological relevance to cancer. Using two cancer datasets, we applied three meta-learners: DT, RF and LR, and summarized the top 10 features selected by each learner. This study concentrated specifically on the interpretations of the **DT results**. In the liver cancer dataset, DT identified ECM1, TTC3, and EIF2S1, all of which are highly associated with hepatocellular carcinoma. In the renal cancer dataset, DT highlighted EHP1L1, a gene

strongly linked to renal cancer progression. While RF and LR also generated important predictive features, their biological interpretation will be explored in future work. By linking predictive performance with biological annotation, this framework not only improves classification accuracy but also provides clinically meaningful insights, bridging machine learning predictions with cancer biology.

## Introduction

Cancer remains one of leading global health challenges, and early detection and accurate classification are essential for improving treatment outcomes. Advances in machine learning now make it possible to analyze gene expression data, providing interpretable insights that support doctors in cancer diagnosis and treatment. However, gene expression datasets are characterized by thousands of features but relatively few samples, leading to overfitting and reduced generalizability. Previous works have applied classifiers such as SVM [\[1\]](#), Random forest [\[2\]](#), and boosting algorithms [\[3\]](#) to gene expression data, often yielding strong predictive performance and addressing the overfitting problem. While these machine learning methods achieve high accuracy, their results frequently lack interpretability, limiting their usefulness in clinical decision-making.

We present a stacked ensemble framework that uses SVM as the base learner and independently evaluates meta-learners such as Decision Trees (DT), Random Forests (RF), and Logistic Regression (LR). This framework builds on previous work and is motivated by the interpretability of tree-based models, which are straightforward for physicians and clinical researchers to understand. To address the overfitting challenge inherent in high-dimensional gene expression data, we incorporate Recursive Feature Elimination (RFE) to identify the most informative subsets of genes. The top-ranked genes from each model are then listed, and we analyzed the link from these genes with the cancer type.

# Methodology

## 1. Data Loading and Preprocessing

We use publicly available [CuMiDa](#) dataset, which provides multiple binary classification datasets consisting of normal and tumor samples. Each sample is represented by Gene Expression values measured through Affymetrix probe IDs (PROBEID). Each probe corresponds to a gene (or probe set) and captures the amount of mRNA present in a sample, with higher mRNA abundance indicating stronger gene expression. These values are recorded as signal intensities, which are often log-transformed for normalization.

The probe IDs can be mapped to gene symbols and gene names, allowing us to connect the features selected by the model to their biological functions. This mapping enables downstream interpretation of the most informative genes and their potential relevance to cancer.

## 2. Two-Level Stacked Ensemble Framework

### 2.1 Level - 0: Base Learner (SVM)

We trained a SVM on the original gene expression features. Two kernels were considered:

- Linear Kernel  $K(x_i, x_j) = x_i^T x_j$ . This operates in the original feature spaces; effective when classes are approximately linearly separable.
- RBF Kernel  $K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$ . This kernel implicitly maps data to high dimensional space; suitable for nonlinear decision boundaries. The RBF kernel requires tuning  $\gamma$  (influence radius) and additional  $C$  (regularization).

To create unbiased meta-features, we generated **out-of-fold (OOF) predictions** with cross validation predict: for each fold, the SVM was trained on  $k - 1$  folds and used to predict the held-out fold. Concatenating across folds yielded a prediction vector aligned with the training labels, where each entry was produced by a model that had not seen that sample. These predictions were then appended to the original features to form the **meta-training set**,

$$X_{train\ meta}$$

For the test set, the SVM was retrained on the entire training dataset and then used to generate predictions for all test samples. These predictions were appended to the original test features to construct the **meta-test set**,  $X_{test\ meta}$ .

## 2.2 Level-1 Meta-Learner

We trained multiple meta-learners on the stacked feature set (original features combined with SVM out-of-fold predictions) in order to compare their performance:

### A Decision Tree (DT)

DT was trained on the meta features, allowing the ensemble to combine SVM's margin-based separation with tree-based nonlinearity and interactions. We evaluated two impurity measures:

- Gini impurity:  $Gini = 1 - \sum_{i=1}^C p_i^2$ . Fast to compute; often yields splits similar to entropy; can exhibit mild bias toward larger classes.
- Entropy/information Gain:  $Entropy = - \sum_{i=1}^C p_i \log_2(p_i)$ ; Information Gain = entropy before split – weighted entropy after split. More theoretically grounded; slightly slower due to logs; sometimes more sensitive to class distribution.

We choose to use Gini impurity for our model. Hyperparameters (e.g., `max_depth`, `min_sample_split`) were tuned to control variance and prevent overfitting. The meta-learner was evaluated with the same cross-validation scheme as the base learner to ensure fair comparison.

### **Random Forest(RF)**

RF extends the DT approach by building an ensemble of trees trained on bootstrap samples of the data, with random feature subsets considered at each split. This bagging approach reduces variance and improves generalization compared to a single DT, while still capturing nonlinear relationships.

- Each tree is trained on a random subset of samples and features, ensuring decorrelation.
- Predictions are aggregated by majority vote (classification).
- Feature importance can be derived from the average decrease in impurity across trees.

We tuned hyperparameters such as the number of trees (`n_estimators`), tree depth (`max_depth`), and class weights to handle class imbalance. By averaging across many trees, RF provides higher stability and robustness in high-dimensional gene expression data.

### **Logistic Regression(LG)**

Unlike tree-based models, it assumes a linear decision boundary in the feature space, making it simpler but still effective for high-dimensional data like gene expression. Even though LR is not a tree based model, we can still rank coefficients to determine importance of features.

We evaluate the regularization term:

- L1 Regularization (Lasso); Penalty term:  $\lambda \sum |w_i|$ . pushes some coefficients exactly to **zero**, effectively performing feature selection.
- L2 Regularization (Ridge); Penalty term:  $\lambda \sum w_i^2$ . Shrinks coefficients toward zero but **does not eliminate them**.

We selected L2 regularization for Logistic Regression because it stabilizes coefficients in the presence of correlated features, which is particularly important in gene expression data.

Unlike L1, which enforces sparsity by removing predictors, L2 retains all features while controlling overfitting, thereby preserving potentially relevant biological signals.

### 3. Feature Selection with Recursive Feature Elimination (RFE)

To reduce dimensionality and identify the most informative genes, we applied Recursive Feature Elimination (RFE) using a Decision Tree estimator (criterion = "gini", max\_depth = 3). RFE was evaluated across a predefined range of feature subset sizes, **from 25 to 500**.

For each candidate number of features k:

1. An RFE selector was initialized with n\_features\_to\_select = k and step = 0.1, eliminating 10% of the least important features at each iteration.
2. The RFE selector was fitted on the training data to rank feature importance.
3. Both training and test sets were transformed to retain only the top k features.
4. A Decision Tree classifier was trained on the reduced training set and used to predict the test set.

Model performance was assessed using the weighted F1 score and 5-fold cross-validation.

The F1 scores and cross-validation results were tracked for each subset size. The feature subset corresponding to the highest F1 score was selected as the optimal set of genes for downstream analysis.

We selected a **Decision Tree** as the estimator for RFE because tree-based models naturally provide feature importance rankings, making them well-suited for recursive elimination in high-dimensional gene expression data.

#### 4. Selected Top k features

From the RFE results, we identified the feature subset size that achieved the best performance. Specifically, we index the **top k features** (where k varied depending on the dataset) for downstream analysis. The RFE selector's support mask was used to extract the indices corresponding to these features, which were then mapped back to their probe IDs. This step allowed us to isolate the most informative genes from the high-dimensional dataset for biological interpretation.

#### 5. Retraining on Selected Features

After feature selection, the meta-training set was reduced to include only the top k features identified by RFE. **Final models** were then trained separately on this reduced feature set using **Decision Tree (DT)**, **Random Forest (RF)**, and **Logistic Regression (LR)**.

For the tree-based models (DT and RF), feature importances were extracted and ranked in descending order, highlighting the genes that contributed most strongly to classification. For Logistic Regression, feature coefficients were examined, with the magnitude and direction of each coefficient indicating the strength and type of association with tumor versus normal samples.

Across models, the top-ranked features were highlighted for interpretation, providing a focused list of genes most strongly associated with the classification tasks.

## 6. Gene Annotation Using R

To translate the selected probe IDs into biologically interpretable features, we used the **Affymetrix Human Genome U133 Plus 2.0 Array annotation data** [hgu133plus2.db](#) annotation package in R. Probe IDs were queried with the `select()` function, which returns the corresponding **gene symbols (SYMBOL)** and **gene names (GENENAME)**. For example, the probe ID *1553611\_s\_at* was mapped to the gene symbol *KLHL35* (*kelch like family member 35*). This step enabled us to interpret model-selected probes in terms of known genes and their biological functions.

## 7. Literature-Based Gene Interpretation

To investigate the biological relevance of the selected genes, we queried **OpenAI's GPT-5 API** for functional descriptions and reported associations with cancer. A custom function was implemented to iterate through the gene list, submit queries, and collect responses, with error handling and rate limiting to ensure reliability. In addition, we cross-referenced results with publicly available resources such as the [GEPIA database](#), which provides gene-specific expression profiles and annotations. This combined approach enabled systematic interpretation of the selected genes in the context of cancer biology.

# Result

## Dataset 1: Liver Cancer

-Sample size: 357

-Feature size: 22278

Classification report of Level-0: Base Learner(SVM)

	Accuray	Precision	Recall	F-1 Score
SVM Linear	0.95	0.95	0.95	0.95
SVM Kernel	0.95	0.95	0.95	0.95

Most important features from Decision Tree(feature importance>0)

Feature Index	Symbol	Feature Name	Importance
209365_s_at	ECM1	extracellular matrix protein 1	0.898400
208664_s_at	TTC3	tetratricopeptide repeat domain 3	0.061777
201143_s_at	EIF2S1	eukaryotic translation initiation factor 2 subunit alpha	0.031854
200957_s_at	SSRP1	structure specific recognition protein 1	0.007969

Top 10 important features from Random Forest

Feature Index	Symbol	Feature Name	Importance
209365_s_at	ECM1	extracellular matrix protein 1	0.133370
205019_s_at	VIPR1	vasoactive intestinal peptide receptor 1	0.087328
218002_s_at	CXCL14	C-X-C motif chemokine ligand 14	0.083794
201293_x_at	PPIA	peptidylprolyl isomerase A	0.070867

220114_s_at	STAB2	stabilin 2	0.066750
204428_s_at	LCAT	lecithin-cholesterol acyltransferase	0.060458
200910_at	CCT3	chaperonin containing TCP1 subunit 3	0.057108
203316_s_at	SNRPE	small nuclear ribonucleoprotein polypeptide E	0.044443
203554_x_at	PTTG1	PTTG1 regulator of sister chromatid separation, securin	0.037653
207804_s_at	FCN2	ficolin 2	0.036256

Top 10 important features from Logistic Regression

Feature Index	Symbol	Feature Name	Importance
220491_at	HAMP	hepcidin antimicrobial peptide	0.067096
217521_at	HAL	histidine ammonia-lyase	0.048428
205695_at	SDS	serine dehydratase	0.047097
211745_x_at	HBA1	hemoglobin subunit alpha 1	0.043790
211745_x_at	HBA2	hemoglobin subunit alpha 2	0.043790
209116_x_at	HBB	hemoglobin subunit beta	0.043246
206643_at	HAL	histidine ammonia-lyase	0.042616
209458_x_at	HBA1	hemoglobin subunit alpha 1	0.042348
209458_x_at	HBA2	hemoglobin subunit alpha 2	0.042348
203824_at	TSPAN8	tetraspanin 8	0.041441
214414_x_at	HBA1	hemoglobin subunit alpha 1	0.040911

214414_x_at	HBA2	hemoglobin subunit alpha 2	0.040911
211696_x_at	HBB	hemoglobin subunit beta	0.040420

Feature description for Decision Tree

Feature Name	Function	Impact on Disease
ECM1	The human gene ECM1 (Extracellular Matrix Protein 1) encodes a glycoprotein involved in the regulation of various biological processes, including cellular growth, differentiation, and migration. It is a key component of the extracellular matrix, playing a crucial role in maintaining the structural integrity of tissues and organs. ECM1 also participates in angiogenesis, lymphangiogenesis, and wound healing. It has been identified as a binding partner for several molecules, including perlecan, fibulin, and laminin, thereby influencing cell adhesion, proliferation, and signaling.	Alterations in the ECM1 gene have been associated with the progression and metastasis of liver cancer. Overexpression of ECM1 has been observed in hepatocellular carcinoma (HCC), the most common type of primary liver cancer. This overexpression is correlated with poor prognosis, increased tumor size, vascular invasion, and higher recurrence rates. ECM1 promotes tumorigenesis by enhancing cell proliferation, migration, and invasion, and by inhibiting apoptosis. Moreover, ECM1 may facilitate the formation of a conducive tumor microenvironment by modulating the extracellular matrix and interacting with other signaling molecules. Therefore, ECM1 is not only a potential biomarker for HCC but also a promising therapeutic target. Further research into the role of ECM1 in liver cancer could provide new insights into the molecular mechanisms of HCC development and progression, and contribute to the development of novel therapeutic strategies.
TTC3	The human TTC3 gene, short for Tetratricopeptide Repeat Domain 3, encodes a protein that is a member of the tetratricopeptide repeat (TPR) family. The TPR motif is a protein-protein interacting module involved in a myriad of cellular processes. The TTC3 protein is known to interact with the NEDD4 family of ubiquitin-protein ligases, specifically NEDD4 and NEDD4L. It facilitates the ubiquitination and subsequent proteasomal degradation of target proteins, thereby playing a critical role in the regulation of protein turnover. Additionally, TTC3 is involved in the regulation of the PI3K-Akt signaling pathway, which is	In relation to liver cancer, aberrations in the TTC3 gene may contribute to tumorigenesis. The PI3K-Akt signaling pathway, which TTC3 helps regulate, is often dysregulated in various cancers, including hepatocellular carcinoma (HCC), the most common type of liver cancer. Overactivation of this pathway can lead to increased cell proliferation and survival, promoting tumor growth. Additionally, the role of TTC3 in protein degradation could also be relevant in cancer. Abnormal protein turnover can lead to the accumulation of damaged or misfolded proteins, which can contribute to cellular stress and potentially cancer development. Therefore, alterations in the TTC3 gene could affect these processes, leading to the initiation or progression of liver cancer. Further research is needed to elucidate the exact mechanisms by which TTC3 may contribute to liver cancer and to explore its potential as a therapeutic target.

	crucial for many cellular processes such as growth, proliferation, and survival.	
EIF2S1	<p>The human gene EIF2S1, also known as Eukaryotic Translation Initiation Factor 2 Subunit Alpha, is a critical component in protein synthesis. This gene encodes the alpha subunit of eukaryotic translation initiation factor 2 (EIF2), which plays a key role in the early steps of protein synthesis. Specifically, it forms a ternary complex with GTP and initiator tRNA and then associates with the 40S ribosomal subunit to form a pre-initiation complex. EIF2 is also involved in the regulation of protein synthesis in response to diverse cellular stresses, including nutrient deprivation, viral infection, and heat shock.</p>	<p>EIF2S1 has been implicated in the pathogenesis of liver cancer, particularly hepatocellular carcinoma (HCC), through its role in protein synthesis and stress response. Aberrant activation of EIF2S1 can lead to dysregulated protein synthesis, contributing to the uncontrolled cellular growth and proliferation characteristic of cancer. Moreover, the stress response function of EIF2S1 can enable cancer cells to survive and thrive in the adverse conditions often found within tumors, such as hypoxia, nutrient deprivation, and metabolic stress. Studies have shown that overexpression of EIF2S1 is associated with poor prognosis in HCC patients, suggesting a potential role for this gene in disease progression. Thus, EIF2S1 represents a potential therapeutic target in liver cancer, and further research into its function and regulation could yield valuable insights into the molecular mechanisms underlying this disease.</p>
SSRP1	<p>The human gene 'SSRP1' (Structure Specific Recognition Protein 1) is a crucial component of the Facilitates Chromatin Transcription (FACT) complex, a general chromatin factor that acts to reorganize nucleosomes. SSRP1 is involved in several cellular processes including transcription elongation, DNA replication, and DNA repair. The gene product interacts with various transcription factors and modulates their activity. Additionally, SSRP1 is implicated in the processes of chromatin assembly and disassembly, playing a vital role in maintaining genome stability.</p>	<p>Alterations in the expression level or mutations in the SSRP1 gene have been linked with the pathogenesis of various types of cancers, including liver cancer. Overexpression of SSRP1 has been observed in hepatocellular carcinoma (HCC), the most common type of primary liver cancer. High expression of SSRP1 is associated with poor prognosis and reduced survival rate in HCC patients. SSRP1 promotes cancer cell proliferation, migration, and invasion, and inhibits apoptosis, thereby facilitating tumor growth and metastasis. Moreover, SSRP1 has been found to be involved in chemoresistance, making it a potential therapeutic target for the treatment of liver cancer. Further research is needed to elucidate the precise mechanisms through which SSRP1 contributes to liver cancer pathogenesis and progression.</p>

## Dataset 2: Renal Cancer

-Sample size: 143

-Feature size: 54676

Classification report of Level-0: Base Learner(SVM)

	Accuray	Precision	Recall	F-1 Score
SVM Linear	0.88	0.91	0.88	0.88
SVM Kernel	0.93	0.93	0.93	0.93

Most important features from Decision Tree(feature importance>0)

Feature Index	Symbol	Feature Name	Importance
91703_at	EHBP1L1	EH domain binding protein 1 like 1	1

Top 10 important features from Random Forest

Feature Index	Symbol	Feature Name	Importance
91703_at	EHBP1L1	EH domain binding protein 1 like 1	0.130381
1553989_a_at	ATP6V1C2	ATPase H+ transporting V1 subunit C2	0.079437
244216_at	NA	NA	0.079050
1554571_at	APBB1IP	amyloid beta precursor protein binding family B member 1 interacting protein	0.070569
35820_at	GM2A	ganglioside GM2 activator	0.069618
40273_at	SPHK2	sphingosine kinase 2	0.059999
244337_at	LOC105378515	uncharacterized LOC105378515	0.050190
65517_at	AP1M2	adaptor related protein complex 1 subunit mu 2	0.049407
31835_at	HRG	histidine rich glycoprotein	0.044592
59625_at	NOL3	nucleolar protein 3	0.040377

### Top 10 important features from Logistic Regression

Feature Index	Symbol	Feature Name	Importance
211298_s_at	ALB	albumin	0.375859
238287_at	SLC7A13	solute carrier family 7 member 13	0.280400
229916_at	ENPP6	ectonucleotide pyrophosphatase/phosphodiesterase 6	0.251260
206025_s_at	TNFAIP6	TNF alpha induced protein 6	0.205507
206026_s_at	TNFAIP6	TNF alpha induced protein 6	0.194697
206134_at	ADAMDEC1	ADAM like decysin 1	0.166195
212992_at	AHNAK2	AHNAK nucleoprotein 2	0.150942
204416_x_at	APOC1	apolipoprotein C1	0.150584
201313_at	ENO2	enolase 2	0.144837
202238_s_at	NNMT	nicotinamide N-methyltransferase	0.141993

### Feature description for Decision Tree

Feature Name	Function	Impact on Disease
EHBP1L1	The human gene EHBP1L1 (EH Domain Binding Protein 1 Like 1) encodes a protein that plays a crucial role in endocytic trafficking and cytoskeleton remodeling. It is involved in clathrin-mediated endocytosis, a process that internalizes extracellular molecules and plasma membrane proteins into the cell. EHBP1L1 also interacts with other proteins such as Rab8 and Rab11, which are essential for membrane trafficking and recycling. Additionally, this gene has been implicated in the regulation of cell shape and polarity, which are critical for	In the context of renal cancer, alterations in the EHBP1L1 gene may contribute to tumorigenesis and disease progression. Abnormal endocytic trafficking and cytoskeleton remodeling can lead to dysregulation of cell growth and division, potentially resulting in uncontrolled proliferation, a hallmark of cancer. Furthermore, changes in cell shape and polarity can affect cell migration, potentially promoting invasion and metastasis, which are characteristic features of advanced renal cancer. Therefore, understanding the role of EHBP1L1 in these processes may provide valuable insights into the molecular mechanisms underlying renal cancer and may lead to the development of novel therapeutic strategies. However, further research is needed to elucidate the precise role of EHBP1L1 in renal cancer.

	various cellular functions, including cell migration and division.	
--	--	--

## Conclusion

By employing meta-learning approaches such as Decision Trees (DT), Random Forests (RF), and Logistic Regression (LR), we identified several proteins from distinct biological pathways that demonstrate strong predictive power in distinguishing cancer patients from cognitively normal individuals. However, the sets of genes selected varied substantially across models, with limited overlap. In addition, we incorporated large language models (LLMs) into the framework, which facilitated biological interpretation: the identified proteins aligned closely with the cancer subtypes, and their relevance could be clearly contextualized with the support of AI-driven insights.

## Reference

1. Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 906–914.
2. Ram M, Najafi A, Shakeri MT. Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest. *Iran J Pathol*. 2017 Oct 1;12(4):339–347.
3. Dettling M, Bühlmann P. Boosting for tumor classification with gene expression data. *Bioinformatics*. 2003 Jun 12;19(9):1061–1069. doi:10.1093/bioinformatics/btf867. PMID: 12801866.