MULTI-TASK VARIATIONAL AUTOENCODERS FOR BREAST CANCER DETECTION AND PAM50 SUBTYPING FROM GENE EXPRESSION WITH GENE PANEL COMPRESSION

Amir Abbas Alvand
Department of Mathematics & Computer Science
Amir Kabir University of Technology (Tehran Polytechnic)
Tehran, Iran
amir.alvand@aut.ac.ir

December 12, 2025

ABSTRACT

Handling breast cancer classification based on gene expression data is a high-dimensional classification task, and it can be significantly impacted by so-called dataset effects and risk of evaluation leakage. The structure of this thesis will be as follows. First, I replicate a published anomaly detection method based on a VAE and trained on normal samples for tumor detection making use of reconstruction loss. Note that, within the specific domain of breast cancer classification, a ROC-AUC-value around 0.91 on a separate test set can be achieved. Subsequently, I propose a multi-head VAE that learns a common latent space for two supervised learning tasks: (i) diagnosis (normal vs. tumor), and (ii) classification of PAM50 subtyping on labeled tumor samples, with missing labels masked. Using a leakage-resistant pipeline (training-only feature scaling), the multi-head VAE demonstrates excellent diagnostic accuracy on the test set (ROC-AUC ≈ 0.98, accuracy ≈ 0.98) and moderate subtype prediction accuracy (weighted F1 ≈ 0.81 on PAM50 subtyping classes), but an L1-regularized logistic regression approach shows near-ceiling diagnostic accuracy on task (i) with ROC-AUC ≈ 0.999, suggesting that task (i) is almost linearly separable on this specific data set. To improve interpretability and usefulness, I apply sparse linear models for gene panel reduction. I then assess classification accuracy and VAE-based anomaly detection with very aggressive compression. A 20-gene panel maintains excellent classification accuracy (ROC-AUC ≈ 0.99). To analyze domain shift, I show that GTEx normal samples are detectable with near-perfect ROC-AUC from TCGA normal samples. These findings confirm that linear models are adequate for highly discriminable classification problems but that VAEs are still useful for representation learning and subtype modeling and exploring accuracy-compression tradeoffs with small gene panels.

1 Introduction

Breast cancer remains one of the most prevalent malignancies worldwide, and early, accurate detection is crucial for improving patient outcomes. Beyond imaging and histopathology, transcriptomic gene expression profiling provides a molecular view of tumor biology by measuring the activity of thousands of genes simultaneously. This high-dimensional signal can encode both the presence of disease (tumor versus normal tissue) and clinically meaningful subtypes, which are known to correlate with prognosis and treatment response.

From a machine learning perspective, transcriptomic cancer detection is attractive but non-trivial. The feature space is extremely high-dimensional (tens of thousands of genes), sample sizes are limited relative to dimensionality, and measurements often come from heterogeneous sources with systematic differences caused by protocols, sequencing platforms, and cohort-specific preprocessing. These "batch effects" and cross-dataset shifts can inflate apparent performance and reduce out-of-cohort generalization. As a result, rigorous experimental design—especially correct train/validation/test separation and leakage prevention—is essential when evaluating models on gene expression data.[1]

1.1 Problem setting

This thesis focuses on breast cancer (BRCA) using transcriptomic gene expression data assembled from two widely used sources: GTEx (healthy tissue) and TCGA (tumor and matched-adjacent normal tissue). Each sample is represented by a vector of gene expression values across a common set of genes. Two prediction problems are considered:

1. Diagnosis: binary classification of normal vs tumor samples.
2. Molecular subtyping: multi-class classification of tumor samples into PAM50 subtypes (Basal, HER2-enriched, Luminal A, Luminal B, and Normal-like), a clinically established taxonomy that summarizes major breast cancer expression phenotypes.

While diagnosis is the first-order task, subtype prediction adds an additional layer of biological and clinical relevance: a model that can learn a compact representation supporting both tasks may be more informative than a single-task classifier. At the same time, practical applications often demand interpretability and feature reduction, motivating the study of compact "gene panels" that preserve predictive performance with far fewer genes than a full transcriptome.

1.2 Variational autoencoders for anomaly detection and representation learning

A central theme in modern biomedical machine learning is how to leverage generative or self-supervised models to learn robust representations. Variational autoencoders (VAEs) are a class of generative models that learn to encode high-dimensional inputs into a lower-dimensional latent space while regularizing this latent space to follow a simple prior distribution (typically a standard normal). The VAE is trained to reconstruct its input from the latent representation, with an additional KL-divergence term enforcing the prior. In principle, this encourages a smooth and structured latent space that can capture salient biological variation.

A notable application of VAEs in cancer genomics is anomaly detection. The idea is conceptually simple: train a model only on normal samples so that it learns the manifold of healthy gene expression. At inference time, tumor samples—being biologically different—should be reconstructed poorly and therefore yield higher reconstruction error. This turns cancer detection into a one-class or semi-supervised problem in which reconstruction error serves as an anomaly score. However, anomaly detection pipelines can be deceptively sensitive to evaluation choices (e.g., threshold selection, class mixture proportions, and leakage between splits), making rigorous validation design especially important.

1.3 Why baselines and leakage prevention matter

In high-dimensional biology, it is surprisingly easy to obtain inflated test performance by inadvertently using information from the test set during preprocessing or model selection. Two common sources of leakage are:

- Feature selection performed on the full dataset before splitting, which allows test samples to influence which genes are retained.
- Scaling/normalization fit on the full dataset, which similarly leaks test distribution information into training.
- Threshold tuning on the test set in anomaly detection, which effectively optimizes the decision rule on the evaluation data.

Because the goal of this thesis is not only to obtain strong performance but also to present scientifically defensible results, the entire experimental pipeline is designed to be leakage-resistant. Feature selection (variance filtering), scaling (MinMaxScaler), and threshold selection for anomaly detection are performed using training and validation data only, with the test set held out for final evaluation.

1.4 Thesis approach

This thesis proceeds in three stages:

1. Replication and critique of a published anomaly-detection VAE. I reimplement a VAE anomaly detection approach that trains on normal samples and uses reconstruction error to separate tumors from normals. The replication serves two purposes: it establishes a baseline and demonstrates how evaluation protocol choices influence performance, particularly when selecting thresholds for anomaly classification.
2. Multi-task (multi-head) VAE for diagnosis and PAM50 subtype prediction. Building on the representation-learning capability of VAEs, I introduce a multi-head VAE that combines a shared encoder/latent space and

decoder with two supervised prediction heads: one for diagnosis and one for PAM50 subtype. This allows the model to learn a single latent representation supporting both tasks. Missing subtype labels are handled by masking subtype loss contributions for samples without valid PAM50 annotations.

3. Gene panel discovery and compression experiments. Given that diagnosis may be highly separable in expression space, I evaluate strong linear baselines (L1-regularized logistic regression) and use them to derive compact, interpretable gene panels. I then assess (i) how much supervised performance can be preserved under compression and (ii) how different panels affect anomaly detection when the VAE is trained exclusively on normal samples.

1.5 Contributions

The main contributions of this thesis are:

- Leakage-resistant evaluation of VAE anomaly detection for breast cancer, including clean train/validation/test separation and validation-based threshold selection.
- A multi-head VAE architecture that jointly performs diagnosis and PAM50 subtyping using a shared latent representation and masked subtype supervision.
- A rigorous baseline comparison demonstrating near-ceiling diagnosis performance with L1 logistic regression, clarifying when complex models are (and are not) necessary.
- Sparse gene panel construction and analysis, including diagnosis-oriented and subtype-oriented panels, and evaluation of performance retention under severe feature compression.
- Quantification of domain shift between GTEx normals and TCGA normals, highlighting limitations for generalization and deployment.

A key practical theme is trade-offs: between generative modeling and discriminative baselines, between full-transcriptome models and compact panels, and between within-cohort performance and cross-source robustness.

1.6 Thesis outline

The remainder of this thesis is organized as follows. Section 2 reviews background on gene expression modeling, VAEs, anomaly detection, and PAM50 subtyping. Section 3 describes datasets and preprocessing, with emphasis on leakage prevention and split design. Section 4 presents the proposed models, baselines, and gene panel construction methods. Section 5 reports experimental results across diagnosis, subtype prediction, anomaly detection, and representation analysis. Section 6 discusses implications, domain shift, interpretability, and limitations. Section 7 concludes with a summary and directions for future work.

2 Background and Related Work

This Section summarizes the biological and machine-learning concepts needed to interpret the experiments and results of this thesis: transcriptomic gene expression data, PAM50 molecular subtypes, domain shift between cohorts, and the modeling approaches used (variational autoencoders, anomaly detection, and sparse linear baselines).

2.1 Transcriptomic gene expression for cancer detection

Gene expression profiling quantifies mRNA abundance across thousands of genes, providing a molecular snapshot of cellular state. In cancer, gene expression changes reflect processes such as proliferation, immune infiltration, metabolic reprogramming, and lineage identity. Compared to low-dimensional clinical biomarkers, transcriptomic profiles are high-dimensional and can encode both diagnostic signal (tumor vs. normal) and finer-grained biological differences among tumor subtypes.

In this thesis, each sample is represented as a vector $x \in \mathbb{R}^G$ of expression values across a shared set of genes. The high dimensionality ($G \approx 2 \times 10^4$ in the raw intersection) motivates careful preprocessing and model selection to avoid overfitting and to ensure scientifically defensible evaluation.

2.2 PAM50 molecular subtyping in breast cancer

Breast cancer is heterogeneous and commonly stratified into intrinsic molecular subtypes. The PAM50 taxonomy assigns tumors to five major subtypes: Basal-like, HER2-enriched, Luminal A, Luminal B, and Normal-like, each

associated with distinct expression programs and clinically relevant differences in prognosis and treatment response (e.g., endocrine sensitivity in luminal disease and HER2-targeted therapy for HER2-enriched tumors) [2, 3].

Subtype prediction can be framed as a multi-class classification task on tumor samples with available PAM50 labels. In practice, subtype datasets are often imbalanced (e.g., LumA is typically the largest class), which makes per-class evaluation important and motivates class-weighting or other imbalance-aware training strategies.

2.3 Domain shift and batch effects (GTEx vs. TCGA)

A persistent challenge in transcriptomic modeling is that data generated in different cohorts can differ systematically due to technical and procedural factors, including sample collection, sequencing protocols, library preparation, alignment/quantification pipelines, and tissue handling. Such differences can introduce *batch effects* and broader *domain shift*, where the marginal distribution of features differs between sources even for the same biological condition [4, 5].

This matters for two reasons. First, models may learn cohort-specific artifacts that inflate in-cohort performance while failing to generalize. Second, mixing sources (e.g., GTEx normals and TCGA normals) can change the effective decision boundary for diagnosis models and can distort anomaly detection thresholds. For these reasons, this thesis explicitly evaluates source separability among normals as an indicator of domain shift and discusses the implications for generalization.

2.4 Variational autoencoders

Variational autoencoders (VAEs) are latent-variable generative models that learn a probabilistic mapping between observations x and latent variables z [6, 7]. A VAE consists of an encoder $q_\phi(z \mid x)$ and a decoder $p_\theta(x \mid z)$. The encoder produces parameters of an approximate posterior (commonly Gaussian),

$$q_\phi(z \mid x) = \mathcal{N}\big(z; \mu_\phi(x), \mathrm{diag}(\sigma_\phi^2(x))\big), \tag{1}$$

and samples are drawn using the reparameterization trick:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{2}$$

Training typically maximizes the evidence lower bound (ELBO), equivalently minimizing the negative ELBO:

$$\mathcal{L}_{\mathrm{VAE}}(x) = \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[-\log p_\theta(x \mid z)\right]}_{\text{reconstruction term}} + \beta \underbrace{\mathrm{KL}\big(q_\phi(z \mid x) \,\|\, p(z)\big)}_{\text{regularization}}, \tag{3}$$

where $p(z)$ is typically a standard normal prior and β can be used to control the strength of latent regularization ("β-VAE") [8]. In this thesis, reconstruction is implemented via mean-squared error on scaled expression values, and the KL term encourages a structured latent space.

VAEs are attractive for transcriptomics because they can compress high-dimensional expression profiles into a lower-dimensional latent representation while retaining major axes of variation. This can support downstream tasks such as classification, visualization, and interpretability analyses (e.g., correlating latent dimensions with gene expression patterns).

2.5 Anomaly detection via reconstruction error

In anomaly detection, the goal is to identify inputs that deviate from the distribution of "normal" data. A common approach with autoencoders and VAEs is to train on inliers only and use reconstruction error as an anomaly score [9][10, 11, 12]. If the model learns to reconstruct normal samples well, then out-of-distribution samples (e.g., tumors) are expected to have larger reconstruction error.

Formally, given an input x and reconstruction $\hat{x}$, an anomaly score can be defined as a distance $s(x) = d(x, \hat{x})$ (e.g., L_2 or MSE). A threshold τ yields a binary decision:

$$\hat{y} = \begin{cases} 1 & \text{if } s(x) \geq \tau \quad \text{(anomaly)}, \\ 0 & \text{otherwise} \quad \text{(normal)}. \end{cases} \tag{4}$$

A key methodological point is that τ must be selected without using the test set (e.g., via validation), and performance should be reported both threshold-free (ROC-AUC) and threshold-dependent (accuracy/F1 at a fixed threshold). This thesis emphasizes validation-based threshold selection to avoid optimistic bias.

2.6 Sparse linear baselines and gene panels

Linear classifiers such as logistic regression are widely used in gene expression analysis because they are simple, stable, and often highly competitive. With L_1 regularization, logistic regression encourages sparsity in the coefficient vector, effectively performing embedded feature selection [13, 14]. This enables construction of compact gene panels:

$$\min_{w,b} \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + \exp\left(-y_i(w^\top x_i + b)\right)\right) + \lambda \|w\|_1. \tag{5}$$

In this thesis, L_1-regularized logistic regression serves two roles: (i) a strong diagnostic baseline for tumor vs. normal classification and (ii) a principled method for selecting sparse gene panels for diagnosis and subtype prediction. These panels are then used to study performance–compression trade-offs and to compare cancer-oriented versus normal-oriented feature selection in anomaly detection settings.

2.7 Evaluation metrics and experimental hygiene

Model evaluation is reported using both threshold-free and threshold-dependent metrics. For diagnosis and anomaly detection, ROC-AUC summarizes ranking quality independent of a chosen threshold. For fixed-threshold classification, precision, recall, and F1-score characterize trade-offs between false positives and false negatives, which is particularly important in medical contexts where error types have different consequences.

Because preprocessing steps can leak information across splits, this thesis follows a split-before-preprocess protocol: (i) the dataset is split into training/validation/test sets, (ii) feature selection is performed using training data only, (iii) scaling is fit on training data only and applied to validation/test, and (iv) thresholds (for anomaly detection) are tuned on validation data only. This design is essential for making claims that are scientifically defensible and reproducible.

3 Data and Preprocessing

This Section describes the datasets used in this thesis and the preprocessing pipeline applied prior to model training and evaluation. A central design goal is to prevent information leakage by ensuring that feature selection, scaling, and (for anomaly detection) threshold tuning are performed using training and validation data only.

3.1 Datasets

Gene expression data. This thesis uses transcriptomic gene expression measurements from two sources: (i) GTEx breast tissue samples representing healthy tissue, and (ii) TCGA breast cancer (BRCA) samples including both tumor and normal tissue. The expression data are provided as gene-by-sample matrices with comparable units (FPKM-like). The three main files used are:

- `breast-rsem-fpkm-gtex.txt` (GTEx breast normals),
- `brca-rsem-fpkm-tcga.txt` (TCGA BRCA normals),
- `brca-rsem-fpkm-tcga-t.txt` (TCGA BRCA tumors).

Only genes present in all sources are retained, resulting in an intersection of $G = 19{,}738$ genes in the combined dataset.

Subtype annotations. PAM50 subtype annotations for TCGA BRCA samples are provided in `PAM50_subtype.txt`. Subtypes include `Basal`, `Her2`, `LumA`, `LumB`, and `Normal`. Samples without a PAM50 label are treated as missing and represented as `NA`.

3.2 Sample harmonization and metadata

Patient ID normalization and duplicate handling. TCGA identifiers can occur multiple times due to technical replicates or multiple aliquots. To avoid data leakage across splits and to reduce duplicate-driven bias, TCGA sample identifiers are normalized to patient-level IDs by truncating to the first 12 characters. When multiple samples map to the same patient ID, expression vectors are aggregated by taking the per-gene mean. GTEx samples are retained as provided, with a source label indicating their origin.

Labels and masks. Each sample is assigned:

- a binary diagnosis label $y_{\text{diag}} \in \{0, 1\}$, where 0 denotes normal and 1 denotes tumor;
- a subtype label y_{sub} for PAM50 when available;
- a subtype mask $m_{\text{sub}} \in \{0, 1\}$, where $m_{\text{sub}} = 1$ indicates a valid PAM50 label and $m_{\text{sub}} = 0$ indicates missing (`NA`).

Subtype labels are integer-encoded using a label encoder. The subtype mask is used during training so that samples without subtype annotations contribute zero weight to the subtype loss.

3.3 Train/validation/test splitting

Supervised diagnosis and multi-task experiments. For supervised diagnosis and the multi-head VAE experiments, samples are split into disjoint training, validation, and test sets with stratification by diagnosis label to preserve the class ratio in each split. The test set is held out for final reporting. The validation set is used for early stopping and for model selection choices that require performance feedback.

Anomaly detection experiments. For anomaly detection, the training set is constructed using normal samples only (GTEx normals and TCGA normals). Validation and test sets are constructed as mixtures of normals and tumors. Thresholds on reconstruction error are tuned on the validation mixture (e.g., by maximizing validation F1 with cancer treated as the positive class), and all final performance metrics are reported on the held-out test mixture using the fixed threshold.

3.4 Feature selection

Variance-based gene filtering. To reduce dimensionality while avoiding leakage, variance-based feature selection is performed using the training set only. Genes are ranked by their variance across training samples, and the top K genes are retained (in this thesis, $K = 3000$ for the main experiments). The selected gene indices are then applied unchanged to the validation and test sets.

This split-before-select protocol prevents test samples from influencing which genes are used by the model and is especially important in high-dimensional transcriptomic settings where subtle leakage can yield optimistic performance estimates.

3.5 Scaling and missing values

Scaling. Because gene expression magnitudes can vary widely between genes, features are scaled using min–max normalization to map values into $[0, 1]$. The scaler is fit on the training data only and then applied to validation and test sets using the same fitted parameters.

Missing values. For most experiments, preprocessing produced no missing values after merging common genes. In the anomaly detection experiments, panel-specific matrices are checked for NaNs. When NaNs occur, they are imputed using per-gene medians computed on the training normals for that panel to maintain robustness without using test-set statistics.

3.6 Class imbalance handling

Diagnosis is imbalanced due to a larger number of tumor samples than normals. For supervised baselines, class weights are used where appropriate. For PAM50 subtype prediction, class imbalance is more pronounced (e.g., LumA is the dominant subtype). In the multi-head VAE, subtype loss contributions are weighted using inverse-frequency class weights computed on the training set tumors with valid subtype labels. Samples without subtype labels receive zero subtype weight via the subtype mask.

3.7 Summary

In summary, the data pipeline combines GTEx normals, TCGA normals, and TCGA tumors across a shared gene set, merges TCGA replicates at the patient level, and defines diagnosis and subtype labels with explicit masking of missing subtypes. The pipeline enforces leakage prevention by performing splitting prior to feature selection and scaling, and by tuning anomaly thresholds on validation data only. This experimental hygiene is a core requirement for producing reproducible and scientifically defensible conclusions from transcriptomic machine learning.

4 Methodology

This Section describes the models and experimental procedures used in this thesis. The overall goal is to evaluate (i) VAE-based anomaly detection trained on normal samples, (ii) a multi-head VAE that jointly predicts diagnosis and PAM50 subtype, and (iii) strong sparse linear baselines that also enable compact gene panel construction. Throughout, all model selection choices (including early stopping and anomaly thresholds) are made using training and validation data only.

4.1 Problem definitions and notation

Let $x \in \mathbb{R}^K$ denote a gene expression vector after feature selection (e.g., $K = 3000$ or a smaller gene panel). For diagnosis, each sample has a binary label $y_{\text{diag}} \in \{0, 1\}$ with 0 for normal and 1 for tumor. For subtype prediction, a subset of tumor samples has a PAM50 label $y_{\text{sub}} \in \{1, \ldots, C\}$ with $C = 5$ classes (Basal, Her2, LumA, LumB, Normal-like). A subtype mask $m_{\text{sub}} \in \{0, 1\}$ indicates whether a valid subtype label is present.

4.2 Baseline: paper-style VAE anomaly detector

Motivation. The replicated approach follows the anomaly detection paradigm: train a VAE primarily on normal samples to learn the normal expression manifold, then score tumors as anomalies based on reconstruction behavior.

Model. The anomaly detector is a VAE consisting of an encoder $q_\phi(z \mid x)$ and decoder $p_\theta(x \mid z)$ with a Gaussian latent space. The encoder outputs mean and log-variance vectors $(\mu_\phi(x), \log \sigma^2_\phi(x))$ and latent samples are drawn via the reparameterization trick.

Training objective. The model is trained by minimizing a VAE loss consisting of a reconstruction term and a KL regularization term:

$$\mathcal{L}_{\text{anomVAE}}(x) = \text{MSE}(x, \hat{x}) + \beta \, \text{KL}\big(q_\phi(z \mid x) \,\|\, \mathcal{N}(0, I)\big), \tag{6}$$

where $\hat{x}$ is the decoder reconstruction and β controls the strength of latent regularization.

Anomaly score and thresholding. After training on normals, each sample is assigned an anomaly score equal to its reconstruction error:

$$s(x) = \frac{1}{K} \sum_{j=1}^{K} (x_j - \hat{x}_j)^2. \tag{7}$$

A threshold τ yields a binary prediction $\hat{y}_{\text{diag}} = \mathbb{I}[s(x) \geq \tau]$. In this thesis, τ is selected on validation data (not on the test set), typically by scanning candidate thresholds and choosing the one that maximizes validation F1-score with cancer treated as the positive class. In addition, ROC-AUC is reported using $s(x)$ as a continuous score.

4.3 Multi-head VAE for joint diagnosis and PAM50 subtyping

Motivation. While anomaly detection focuses on reconstructing normals, many clinical settings provide tumor labels and subtype annotations. A multi-task model can exploit these labels while still learning a structured latent representation that supports visualization, interpretability, and compression.

Architecture overview. The multi-head VAE consists of:

- a shared encoder producing latent parameters $(\mu(x), \log \sigma^2(x))$;
- a latent sample z obtained by reparameterization;
- a decoder producing a reconstruction $\hat{x}$;
- a diagnosis head predicting $\hat{y}_{\text{diag}}$ from z;
- a subtype head predicting $\hat{y}_{\text{sub}}$ from z.

Encoder and decoder. The encoder maps x through fully connected layers with batch normalization[15] and dropout[16], then outputs $\mu(x)$ and $\log \sigma^2(x)$. The decoder mirrors this structure to reconstruct $\hat{x} \in [0, 1]^K$.

Supervised heads. The diagnosis head is a small multilayer perceptron ending in a sigmoid output:

$$\hat{y}_{\text{diag}} = \sigma(f_{\text{diag}}(z)). \tag{8}$$

The subtype head ends in a softmax over C subtypes:

$$\hat{p}_{\text{sub}} = \text{softmax}(f_{\text{sub}}(z)). \tag{9}$$

Training losses. The VAE component is optimized using reconstruction loss and KL divergence:

$$\mathcal{L}_{\text{VAE}}(x) = \text{MSE}(x, \hat{x}) + \beta\,\text{KL}\big(q_\phi(z \mid x) \,\|\, \mathcal{N}(0, I)\big). \tag{10}$$

Diagnosis supervision uses binary cross-entropy:

$$\mathcal{L}_{\text{diag}}(y_{\text{diag}}, \hat{y}_{\text{diag}}) = -y_{\text{diag}} \log(\hat{y}_{\text{diag}}) - (1 - y_{\text{diag}}) \log(1 - \hat{y}_{\text{diag}}). \tag{11}$$

Subtype supervision uses sparse categorical cross-entropy:

$$\mathcal{L}_{\text{sub}}(y_{\text{sub}}, \hat{p}_{\text{sub}}) = -\log\big(\hat{p}_{\text{sub}}[y_{\text{sub}}]\big). \tag{12}$$

The overall objective combines these terms with task weights:

$$\mathcal{L} = \mathcal{L}_{\text{VAE}}(x) + \lambda_{\text{diag}}\,\mathcal{L}_{\text{diag}} + \lambda_{\text{sub}}\,m_{\text{sub}}\,w_{\text{sub}}(y_{\text{sub}})\,\mathcal{L}_{\text{sub}}, \tag{13}$$

where λ_{diag} and λ_{sub} control the relative contribution of supervised heads, m_{sub} masks missing subtype labels, and $w_{\text{sub}}(\cdot)$ are class weights computed from training tumors with valid PAM50 labels to mitigate imbalance.

Optimization and early stopping. Models are trained with mini-batch gradient descent using Adam[17]. Early stopping monitors validation loss and restores the best model checkpoint to reduce overfitting. This is especially important for VAEs trained on limited normal samples and for multi-task learning where different heads may converge at different rates.

4.4 Sparse linear baselines

Logistic regression for diagnosis. As a strong discriminative baseline, logistic regression is trained to predict diagnosis from the selected feature set:

$$p(y_{\text{diag}} = 1 \mid x) = \sigma(w^\top x + b). \tag{14}$$

Performance is reported using ROC-AUC and fixed-threshold classification metrics on the held-out test set.

L_1-regularized logistic regression (sparse). To obtain sparse models and gene panels, L_1-regularized logistic regression is trained by minimizing:

$$\min_{w,b} \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + \exp\left(-y_i(w^\top x_i + b)\right)\right) + \lambda \|w\|_1. \tag{15}$$

The L_1 penalty yields many zero coefficients; genes with the largest absolute coefficients define a compact panel.

Logistic regression for PAM50 subtyping. For subtype prediction, a multinomial (one-vs-rest) L_1 logistic regression model is trained on tumor samples with valid PAM50 labels. Per-class precision, recall, and F1-score are reported on a held-out tumor test set.

4.5 Gene panel construction

This thesis compares multiple panel construction strategies:

Cancer-oriented panel (diagnosis). A sparse L_1 logistic regression model is trained on the diagnosis task. The top K_{diag} genes by absolute coefficient magnitude form a diagnosis-oriented gene panel.

Subtype-oriented panel. A sparse multinomial L_1 logistic regression model is trained on tumor samples with PAM50 labels. For each subtype class, the top genes by absolute coefficient magnitude are collected and merged into a subtype-oriented gene panel.

Normal-oriented panels (for anomaly detection). Because anomaly detection trains on normals only, panels that capture stable normal structure may be advantageous. Normal-oriented panels are constructed using training-set statistics only, prioritizing genes with stable expression in normals and large differences from tumors, using a Δmean/variance style criterion. Panels of different sizes (e.g., $K = 20$ and $K = 100$) are evaluated.

Combined panels. To study joint diagnosis and subtype prediction under compression, a combined panel is formed by merging the diagnosis-oriented panel with the subtype-oriented panel. The multi-head VAE is retrained on this panel and evaluated on held-out test data.

4.6 Anomaly detection protocol for gene panels

For each panel, anomaly detection experiments follow a consistent protocol:

1. Restrict all datasets to the panel genes.
2. Construct a normals-only training set (GTEx normals plus TCGA normals, with a held-out subset reserved for normal validation and normal test).
3. Construct validation and test mixtures by combining held-out normals with tumor samples.
4. Fit scaling parameters on training normals only and apply to validation/test.
5. Train an anomaly VAE using normals only.
6. Compute reconstruction error scores on validation and select threshold τ by maximizing validation F1.
7. Evaluate on the test mixture using fixed τ and report ROC-AUC, confusion matrix, and classification metrics.

This protocol ensures that both representation learning and threshold selection are performed without using the test set.

4.7 Evaluation metrics

Diagnosis and anomaly detection. Performance is evaluated using ROC-AUC (threshold-free) and classification metrics at a fixed threshold: accuracy, precision, recall, and F1-score. Confusion matrices are reported to distinguish false positives (normals misclassified as cancer) from false negatives (tumors missed).

Subtype prediction. Subtype performance is reported on tumor samples with valid PAM50 labels using per-class precision, recall, and F1-score, along with macro-averaged and weighted-averaged summaries. Because subtype classes are imbalanced, weighted metrics and per-class reports are emphasized.

Representation analysis. To analyze latent structure, the thesis uses UMAP visualizations[18] of learned embeddings and quantitative separation via silhouette score[19]. Latent dimensions are correlated with diagnosis and with individual gene expression values using Pearson correlation, enabling interpretation of which latent factors capture diagnostic signal.

5 Experiments and Results

This Section reports the experimental results in a structured manner that matches the thesis narrative: (i) replication of a paper-style anomaly-detection VAE, (ii) the proposed multi-head VAE for diagnosis and PAM50 subtype prediction, (iii) strong sparse linear baselines and gene panel compression, (iv) anomaly detection performance across different gene panels, and (v) analyses of latent structure and domain shift.

5.1 Experimental setup

Unless stated otherwise, experiments use the preprocessing pipeline described in Section 3: splitting is performed prior to feature selection and scaling; the top $K = 3000$ genes are selected by training-set variance only; and min–max scaling is fit on the training set only and applied to validation and test sets. Performance is reported on held-out test sets that are not used for hyperparameter tuning or threshold selection.

For anomaly detection experiments, VAEs are trained on normal samples only. Reconstruction error is used as a continuous anomaly score and ROC-AUC is reported. When a binary decision threshold is required, the threshold is selected using validation data only (by maximizing validation F1 with cancer treated as the positive class), and the fixed threshold is then applied to the test set.

5.2 Replication: VAE anomaly detection (paper-style)

We first reimplemented a published anomaly detection setup in which a VAE is trained to model normal breast expression profiles and tumor samples are detected by increased reconstruction error[20].

Performance. On the held-out test set, the paper-style anomaly detector achieved ROC-AUC $= 0.912$. At the selected decision threshold, the model obtained an overall accuracy of 0.94. The classification report shows high specificity (normal recall $= 0.98$) and lower sensitivity for tumors (cancer recall $= 0.84$), indicating that some tumors overlap with normals in reconstruction error.

Table 1: Replication results for the paper-style anomaly detection VAE on the held-out test set.

Class	Precision	Recall	F1-score	Support
Normal (0)	0.94	0.98	0.96	60
Cancer (1)	0.95	0.84	0.89	25
Accuracy		0.94 (N=85)		
ROC-AUC		0.9117		

Reconstruction error distributions and ROC. Figure 1 visualizes the reconstruction error distributions for normal and tumor samples. Figure 2 shows the corresponding ROC curve when reconstruction error is treated as a continuous score. Figure 3 presents the confusion matrix at the selected threshold. Training and validation loss curves are shown in Figure 4.

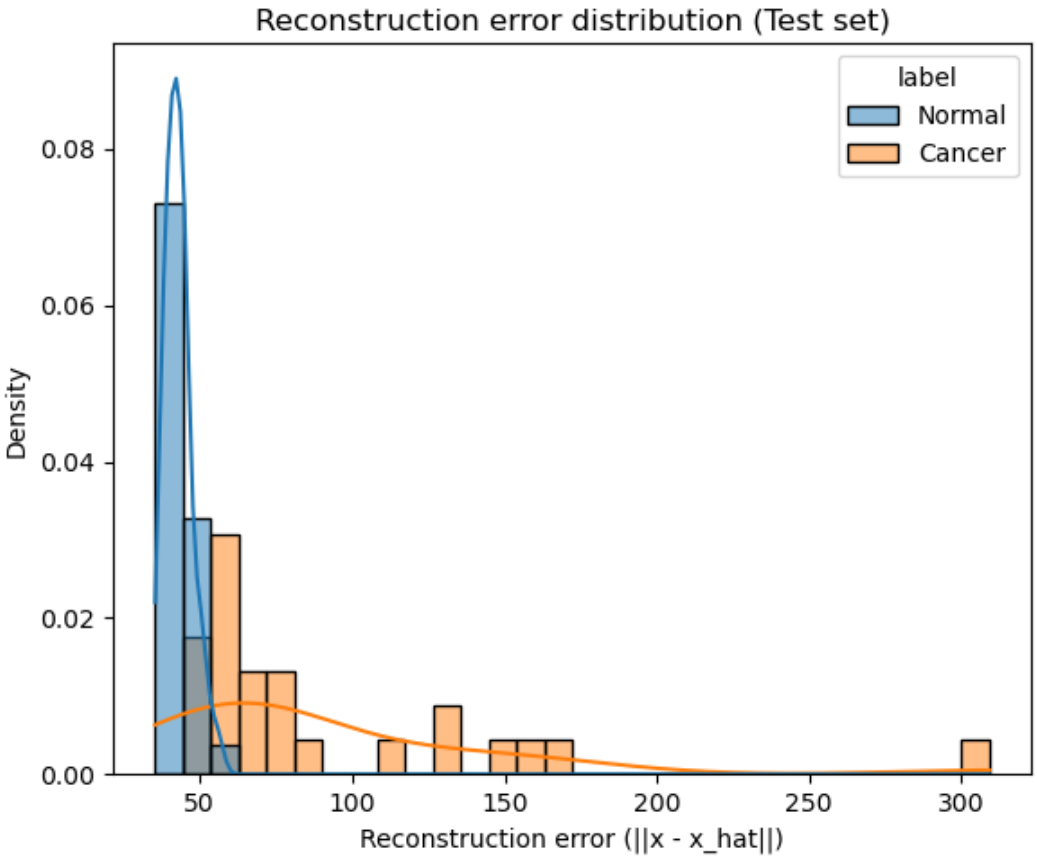

Figure 1: Replication (paper-style anomaly VAE): reconstruction error distributions for normals and tumors on the test set. Tumors exhibit higher reconstruction errors on average, but with overlap.

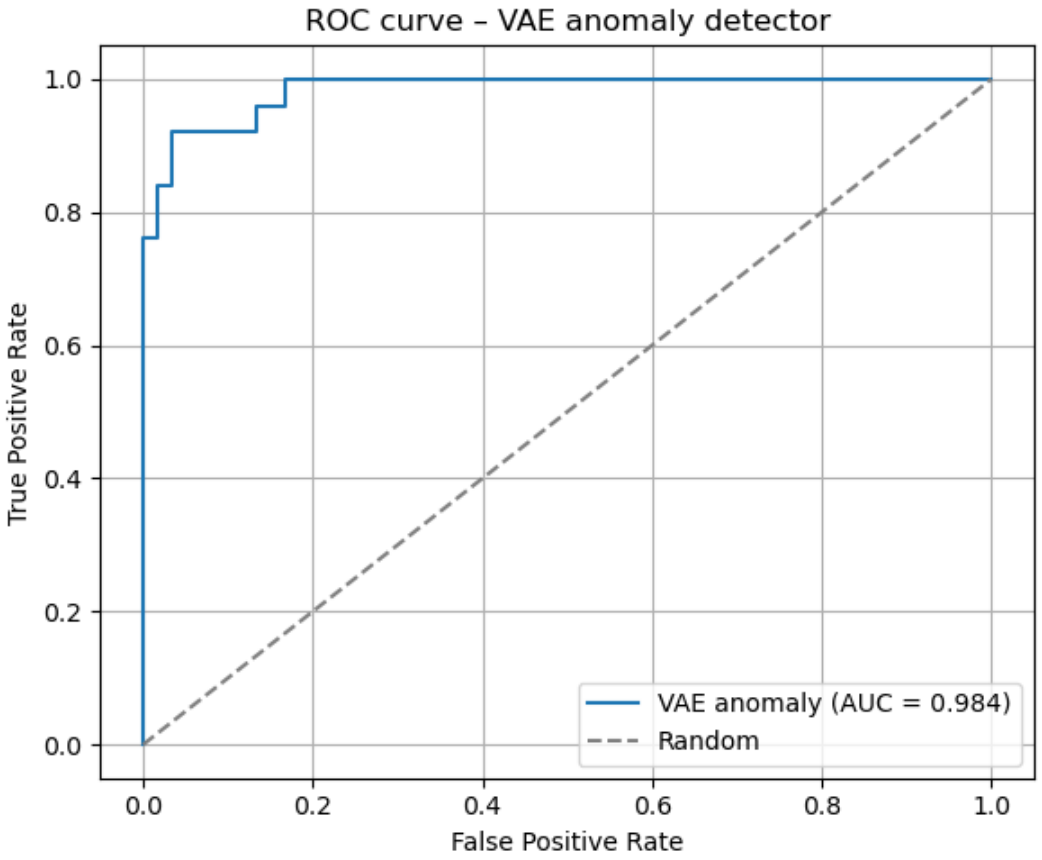

Figure 2: Replication (paper-style anomaly VAE): ROC curve using reconstruction error as anomaly score.

5.3 Proposed model: multi-head VAE for diagnosis and PAM50 subtyping

We next trained the proposed multi-head VAE on the combined dataset (GTEx normals + TCGA normals + TCGA tumors) using $K = 3000$ genes selected by training-set variance. The model jointly predicts diagnosis (normal vs cancer) and PAM50 subtype for labeled tumor samples, while ignoring missing subtype labels via a subtype mask.

Dataset summary and splits. After intersecting genes across sources, the combined dataset contained $G = 19{,}738$ common genes and $N = 1{,}164$ samples. The dataset comprised 198 normals and 966 tumors, with normals originating from GTEx (89) and TCGA (109). Data were split into training ($n = 581$), validation ($n = 233$), and test ($n = 350$) sets with stratification by diagnosis label. After feature selection and scaling, shapes were: train $(581, 3000)$, validation $(233, 3000)$, and test $(350, 3000)$.

5.3.1 Diagnosis performance

On the held-out test set, the multi-head VAE achieved accuracy $= 0.98$ and ROC-AUC $= 0.9816$ for diagnosis. The classification report indicates strong performance on both classes, with slightly lower precision/recall for the smaller normal class due to imbalance.

Table 2: Multi-head VAE diagnosis performance (normal vs cancer) on the held-out test set.

Class	Precision	Recall	F1-score	Support
Normal	0.92	0.97	0.94	60
Cancer	0.99	0.98	0.99	290
Accuracy		0.98 (N=350)		
ROC-AUC		0.9816		

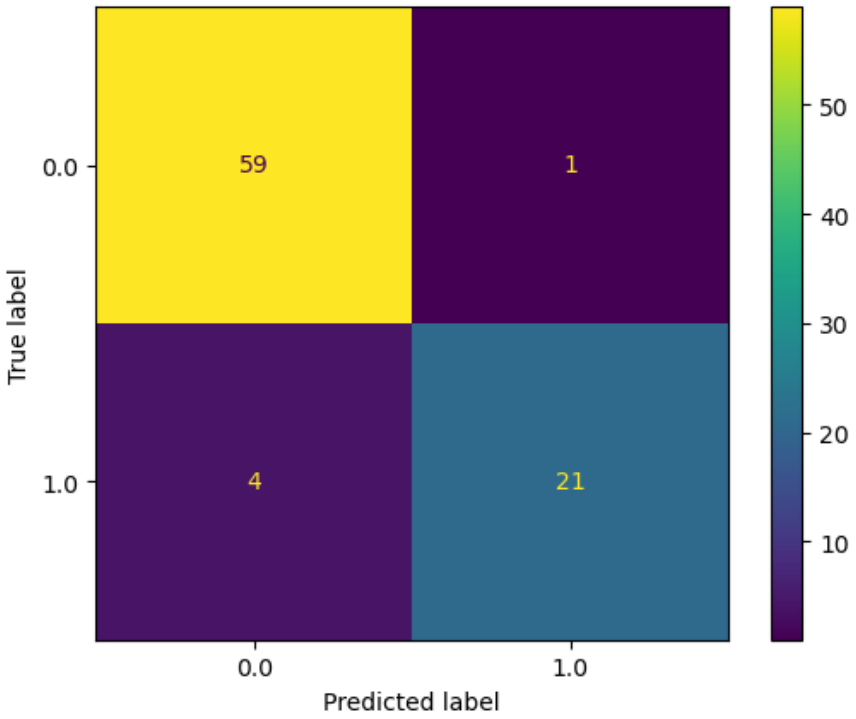

Figure 3: Replication (paper-style anomaly VAE): confusion matrix at the selected threshold (threshold tuned without using test data).

5.3.2 PAM50 subtype performance

Subtype evaluation is performed on tumor samples with valid PAM50 labels in the test set ($n = 284$). The multi-head VAE achieved weighted F1 $= 0.81$ and micro-average accuracy ≈ 0.80. Performance varied across subtypes, with the smallest class (Normal-like) showing the highest variance due to low support.

Table 3: Multi-head VAE PAM50 subtype classification on labeled tumor test samples ($n = 284$).

Subtype	Precision	Recall	F1-score
Basal	0.94	0.96	0.95
Her2	0.65	0.63	0.64
LumA	0.91	0.79	0.85
LumB	0.62	0.77	0.69
Normal-like	0.53	0.67	0.59
Micro avg / Accuracy		0.80	
Weighted avg		F1 = 0.81	

Multi-head training dynamics and diagnostic ROC. Figure 5 summarizes training dynamics and overall performance. The learned latent space is visualized in Figure 6, showing separation by diagnosis and partial organization by subtype.

5.4 Domain shift analysis: GTEx normals vs TCGA normals

To quantify domain shift between sources, we trained a classifier to distinguish GTEx normals from TCGA normals using expression features from normal samples only. The resulting performance was high (accuracy ≈ 0.97 on the normal-only test split and ROC-AUC $= 1.00$), indicating strong source-specific signal among samples that are nominally the same biological condition (normal breast tissue). This supports the interpretation that batch effects or cohort-specific differences are present and may influence downstream diagnosis or anomaly detection thresholds.

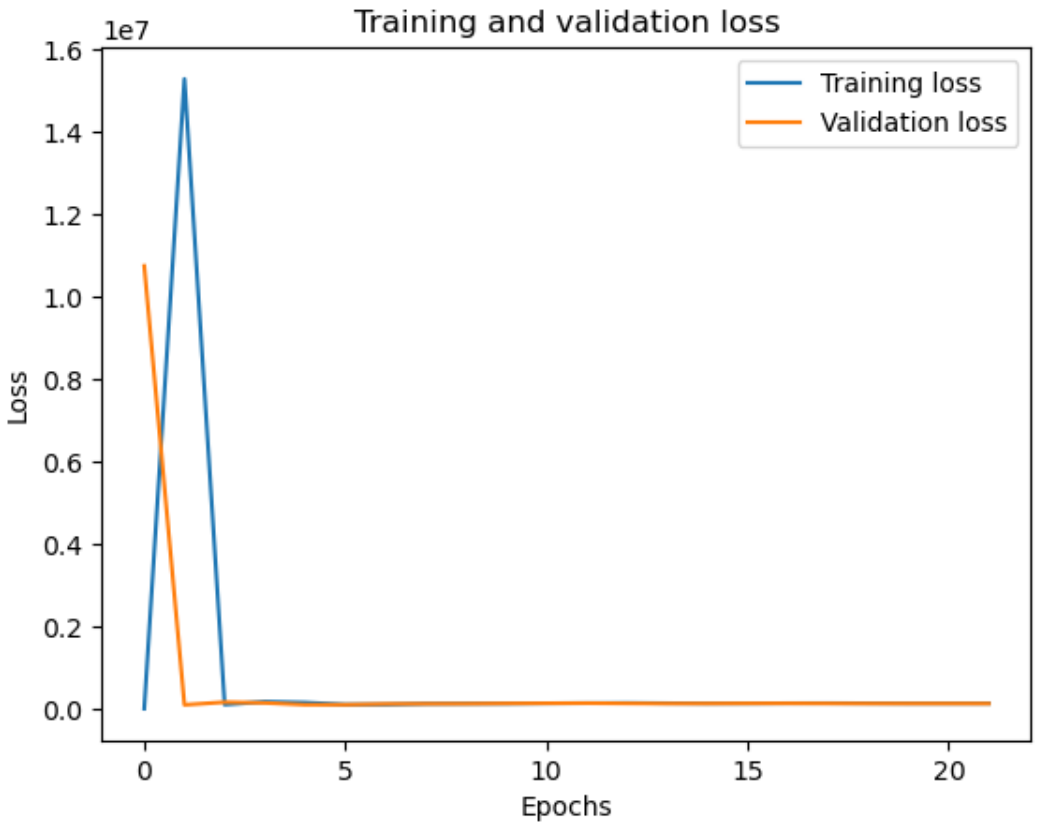

Figure 4: Replication (paper-style anomaly VAE): training and validation loss curves showing convergence.

In addition, we evaluated diagnosis errors by normal source: all GTEx normal test samples were classified correctly, while a small number of TCGA normals were misclassified as cancer. These results highlight that mixing normal sources may change the effective "normal manifold" learned by generative models and can introduce source-dependent errors.

5.5 Strong baselines: logistic regression for diagnosis

We evaluated logistic regression as a discriminative baseline for tumor vs normal classification. Using the same leakage-resistant preprocessing (training-only feature selection and scaling), logistic regression achieved near-ceiling performance (ROC-AUC $= 0.9994$, accuracy ≈ 0.99 on the held-out test set). This indicates that the diagnosis task is close to linearly separable in this dataset after preprocessing, and establishes that improvements should be argued primarily in terms of representation learning, multi-task learning, interpretability, and compression rather than raw diagnostic accuracy.

5.6 Gene panel compression and sparse feature selection

Because full-transcriptome models are expensive and difficult to interpret, we evaluated performance under gene panel compression. Sparse L_1-regularized logistic regression was used to identify diagnosis-oriented genes by selecting features with non-zero coefficients. A compact panel of 20 genes was formed using the top genes by absolute coefficient magnitude.

Diagnosis on a 20-gene panel. After restricting inputs to the 20-gene panel, the VAE-based diagnosis model maintained strong performance on the held-out test set (ROC-AUC $= 0.9922$, accuracy ≈ 0.98). This shows that diagnostic performance can be largely preserved under severe feature compression.

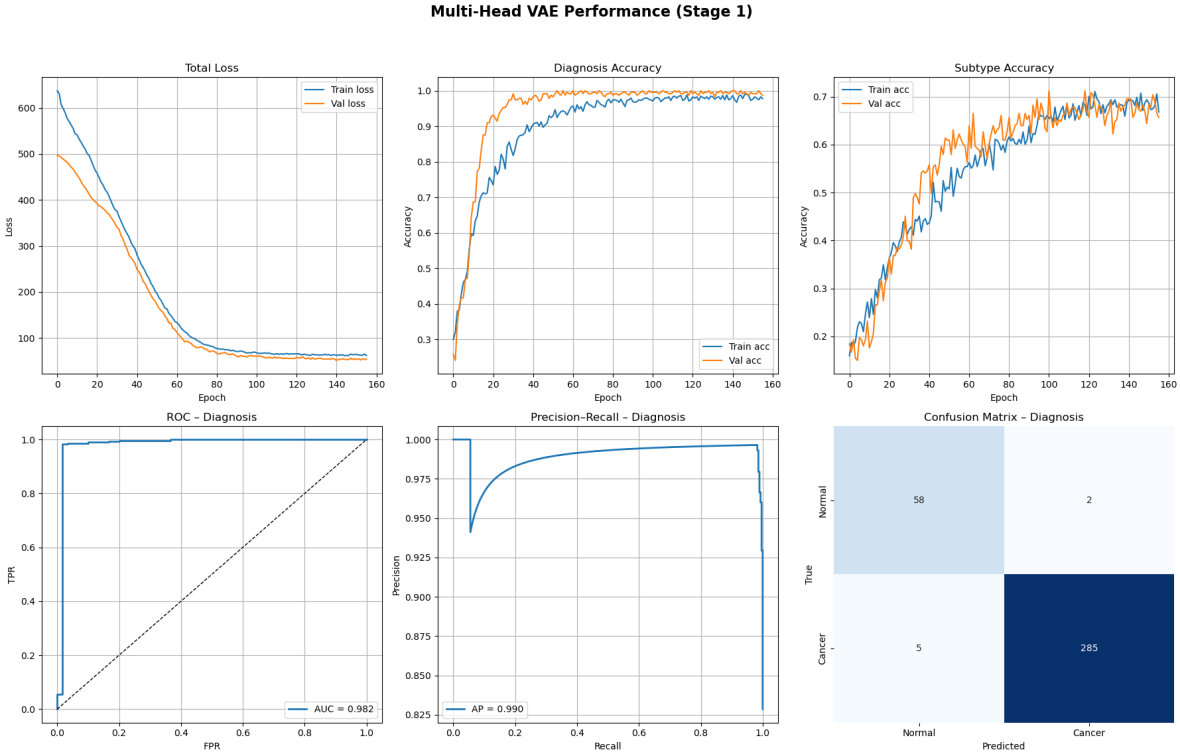

Figure 5: Multi-head VAE: training curves and performance summaries for diagnosis and subtype heads.

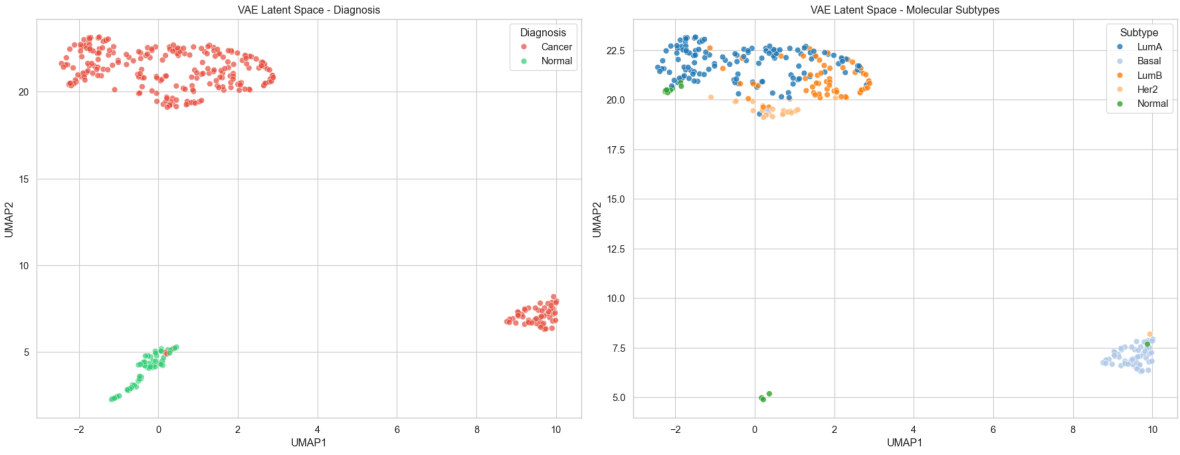

Figure 6: UMAP visualization of the multi-head VAE latent space. Coloring by diagnosis (normal vs tumor) indicates strong separation; coloring by PAM50 subtype shows partial clustering structure among tumors.

Subtype-oriented panels and combined panels. To support subtype prediction under compression, a subtype-oriented gene panel was constructed using L_1 multinomial logistic regression trained on tumor samples with valid PAM50 labels. Combining diagnosis-oriented and subtype-oriented genes produced a compact panel that preserved high diagnostic performance (ROC-AUC $= 0.9978$) and yielded subtype performance comparable to or slightly better than the full 3000-gene multi-head model (accuracy ≈ 0.81 on labeled tumor test samples).

5.7 Anomaly detection across gene panels

We evaluated reconstruction-error anomaly detection VAEs on several panels using a consistent protocol (train on normals only; tune threshold on validation; evaluate on held-out test mixture). Panels included a cancer-oriented 20-gene panel (selected by L_1 diagnosis coefficients) and normal-oriented panels (constructed using training-only Δmean/variance statistics), with sizes $K = 20$ and $K = 100$.

Panel comparison. Table 4 summarizes test ROC-AUC across panels. The normal-oriented 20-gene panel achieved the highest AUC (≈ 0.994), outperforming the cancer-oriented 20-gene panel (≈ 0.96). This supports the hypothesis that when training a VAE exclusively on normals, panels that capture stable normal structure facilitate learning the normal manifold and improve reconstruction-based anomaly scoring.

Table 4: Summary of anomaly detection performance across gene panels (train on normals only; validation thresholding; test ROC-AUC).

Panel	genes	Val best F1	Test ROC-AUC	Test size
Cancer-oriented (L1, $K = 20$)	20	0.94	0.96	54
Normal-oriented (Δmean/var, $K = 20$)	20	0.92	0.994	54
Normal-oriented (Δmean/var, $K = 100$)	100	0.93	0.989–0.993	54

Visualizations. Figure 7 compares test AUCs across panels. Figure 8 shows ROC curves for reconstruction-error anomaly detection. Figures 9–11 visualize UMAP projections of samples using different panels. Figures 12 and 13 provide additional insight into reconstruction error behavior and panel interpretability.

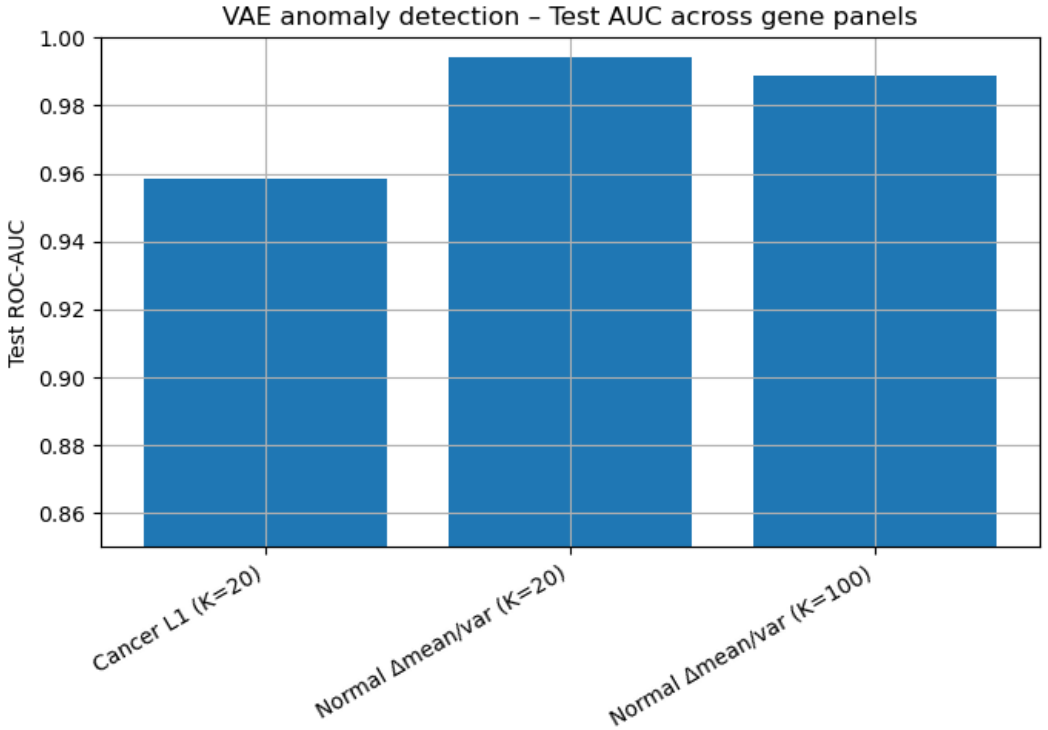

Figure 7: Test ROC-AUC across gene panels for reconstruction-error anomaly detection (VAE trained on normals only).

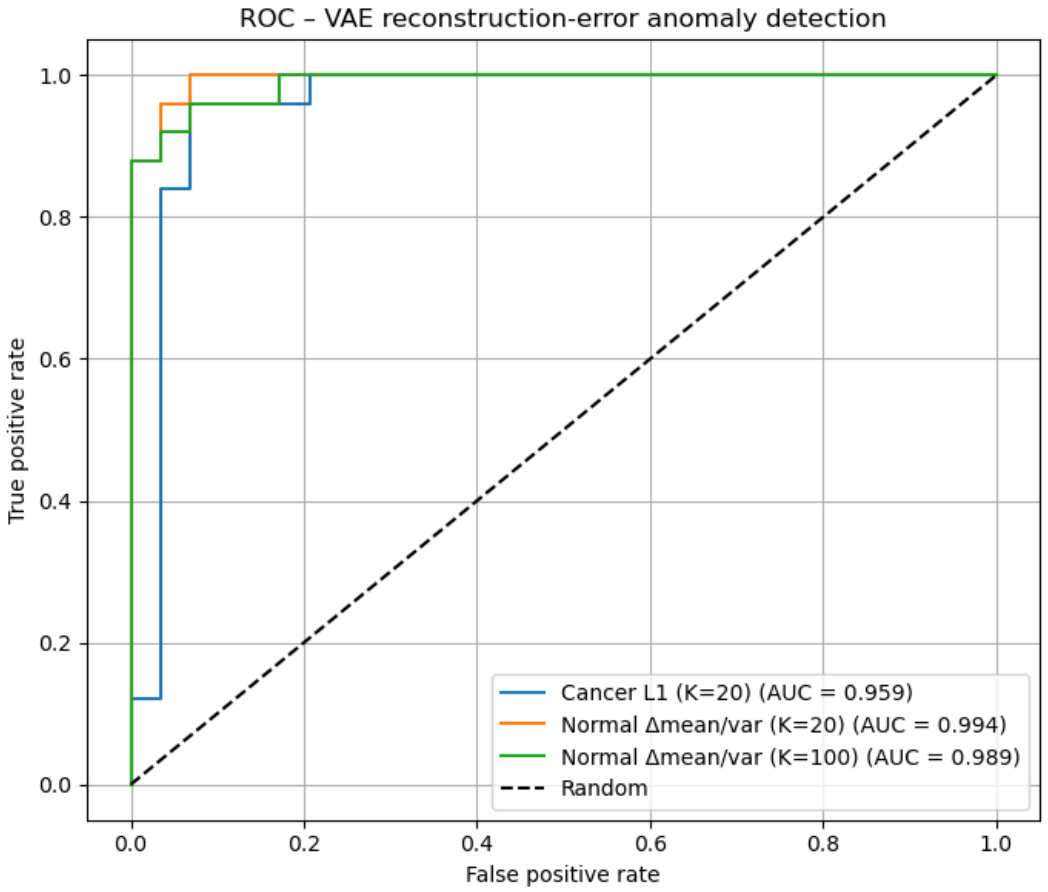

Figure 8: ROC curves for anomaly detection using reconstruction error across different gene panels.

5.8 Latent space and interpretability analyses

To analyze representation quality, we compared embeddings from PCA, a standard autoencoder (AE), and a VAE using silhouette score. The VAE achieved the highest silhouette score (PCA: 0.235, AE: 0.326, VAE: 0.398), suggesting improved separation in the learned latent space. Figure 14 illustrates the corresponding latent space comparison.

We further quantified how individual latent dimensions relate to diagnosis and gene expression. Several latent dimensions were strongly correlated with the diagnosis label (e.g., $|r| > 0.3$), indicating that diagnostic signal concentrates in a subset of latent factors. Figures 15 and 16 summarize these correlation analyses.

5.9 Summary of key findings

Across experiments, we observe: (i) VAE anomaly detection is feasible but sensitive to evaluation protocol and benefits from validation-based thresholding; (ii) the multi-head VAE achieves strong diagnosis performance and moderate PAM50 subtype performance, while providing a structured latent representation for visualization and analysis; (iii) diagnosis in this dataset is near-ceiling with logistic regression, emphasizing the importance of baselines; (iv) compact gene panels can preserve diagnosis and subtype performance under heavy compression; and (v) normal-oriented panels substantially improve reconstruction-error anomaly detection when the VAE is trained on normals only. These findings motivate the discussion in Section 6 regarding domain shift, interpretability, and the practical role of generative models relative to simpler discriminative baselines.

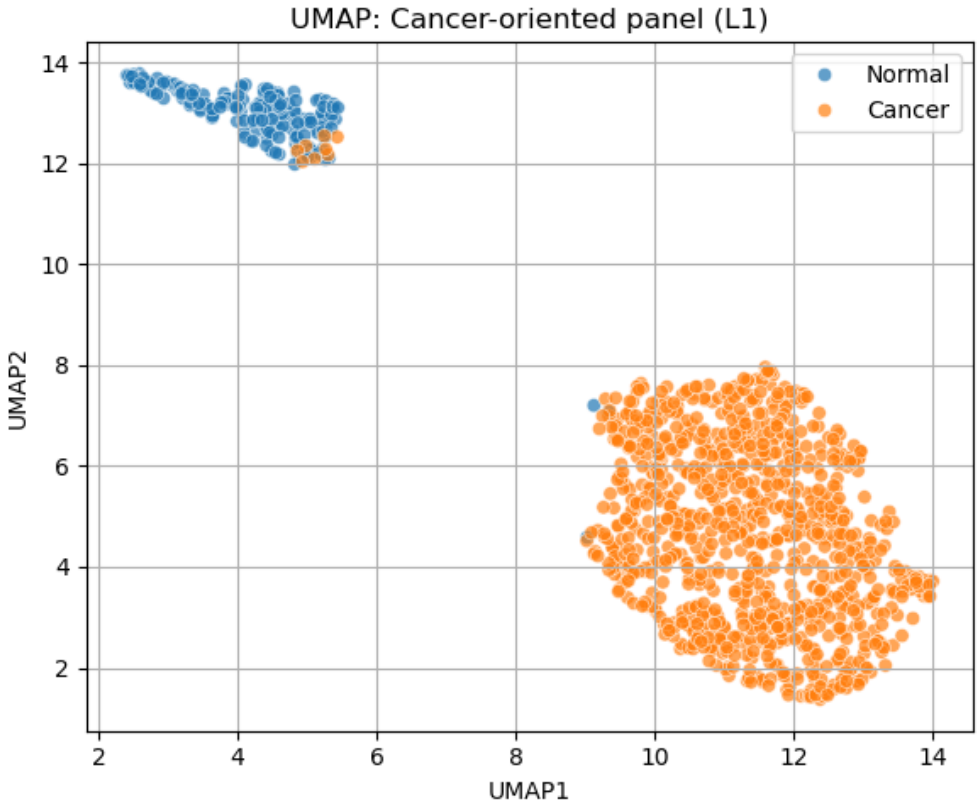

Figure 9: UMAP visualization using the cancer-oriented 20-gene panel.

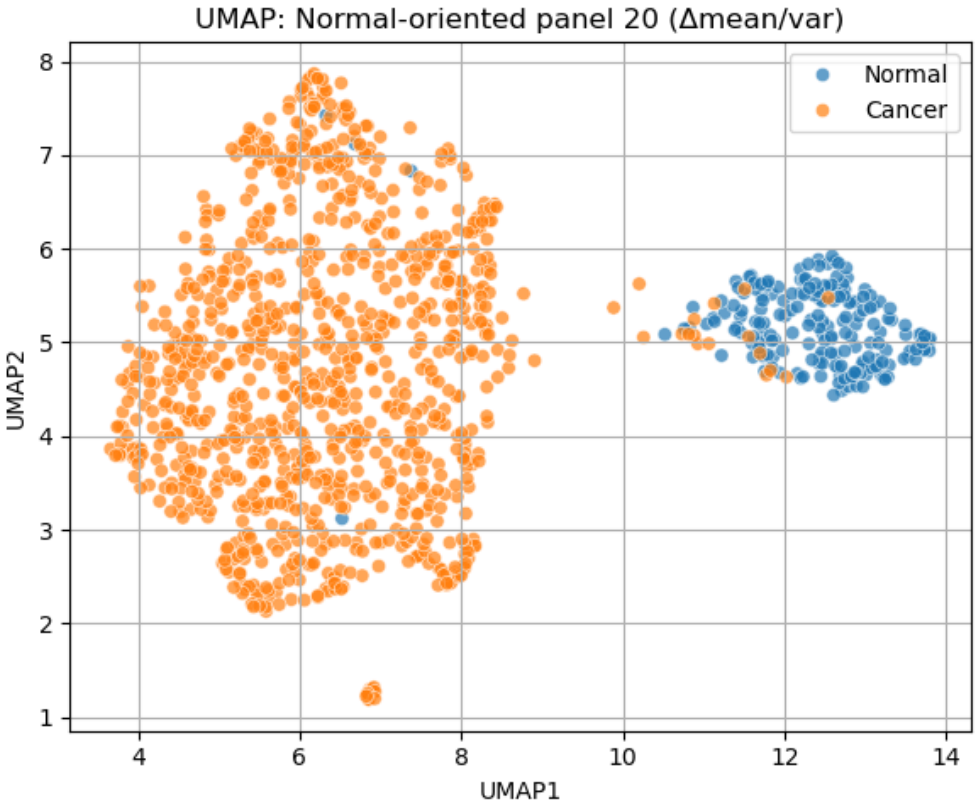

Figure 10: UMAP visualization using the normal-oriented 20-gene panel (Δmean/var).

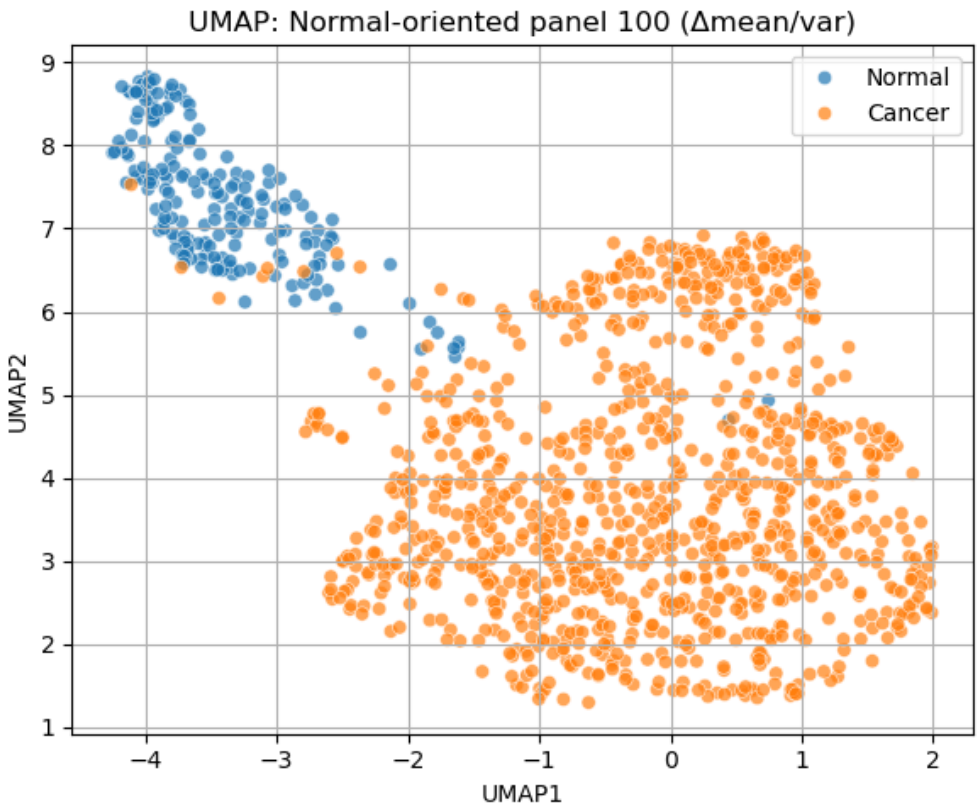

Figure 11: UMAP visualization using the normal-oriented 100-gene panel (Δmean/var).

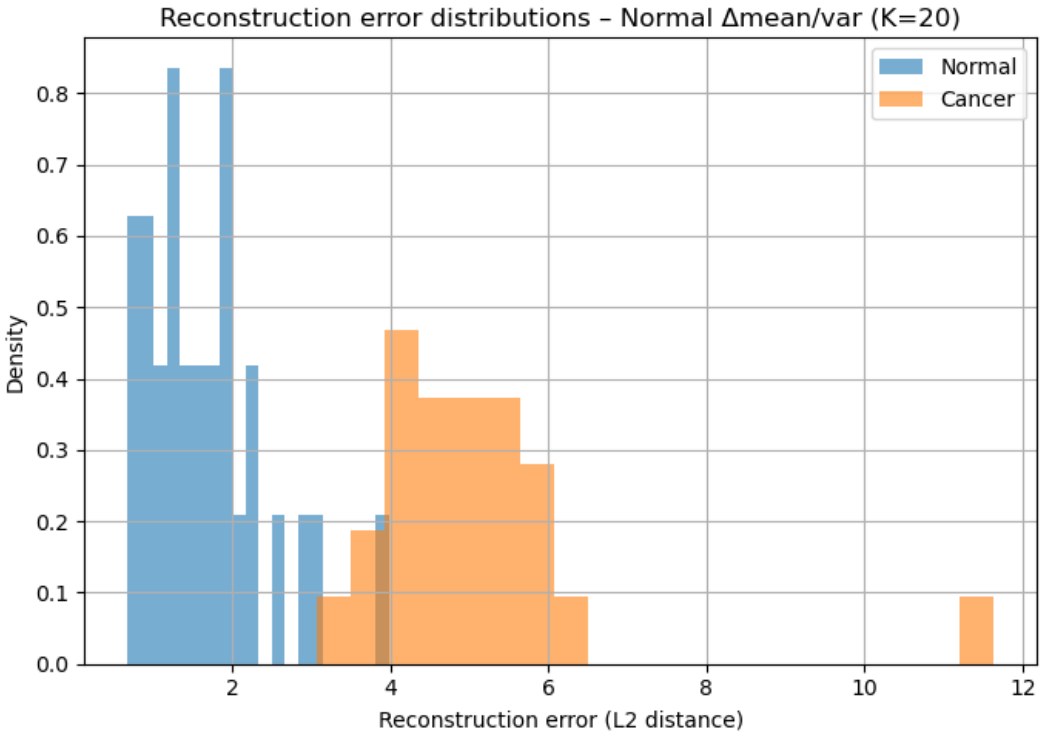

Figure 12: Reconstruction error distributions for anomaly detection using the normal-oriented 20-gene panel.

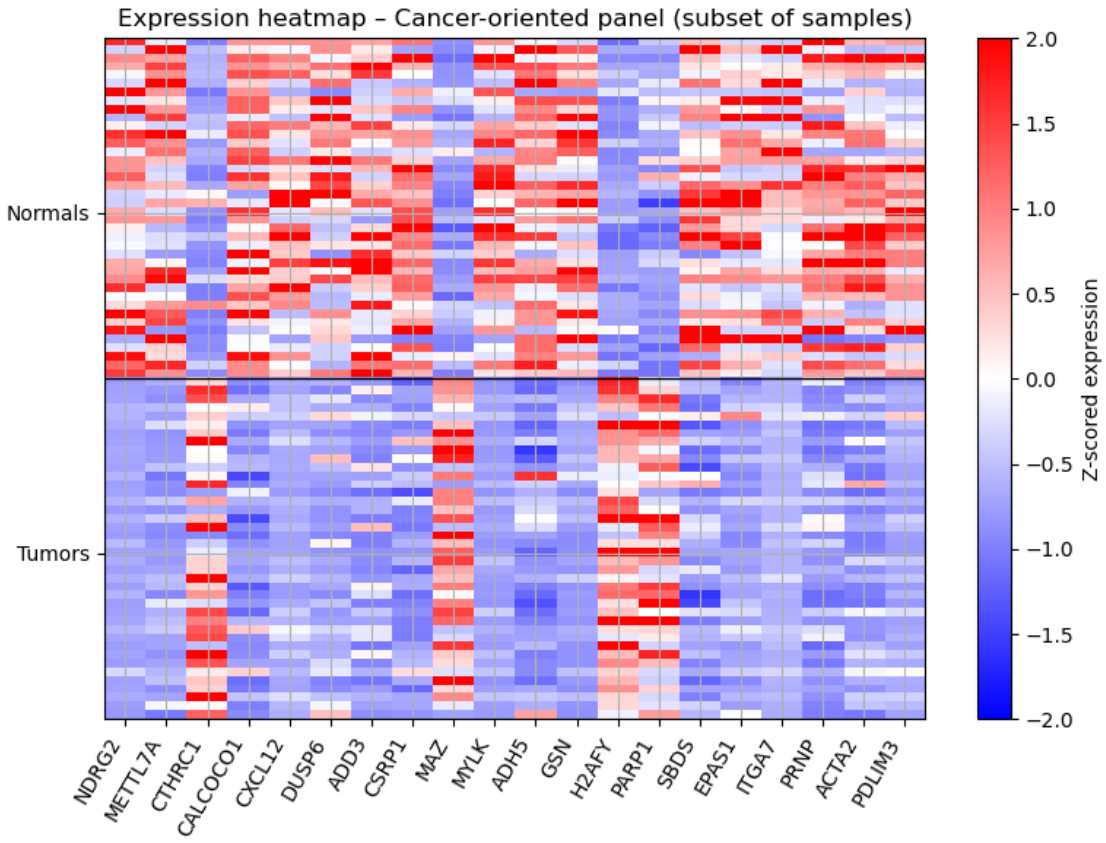

Figure 13: Heatmap of expression values for the cancer-oriented panel, illustrating interpretable differences between normal and tumor samples.

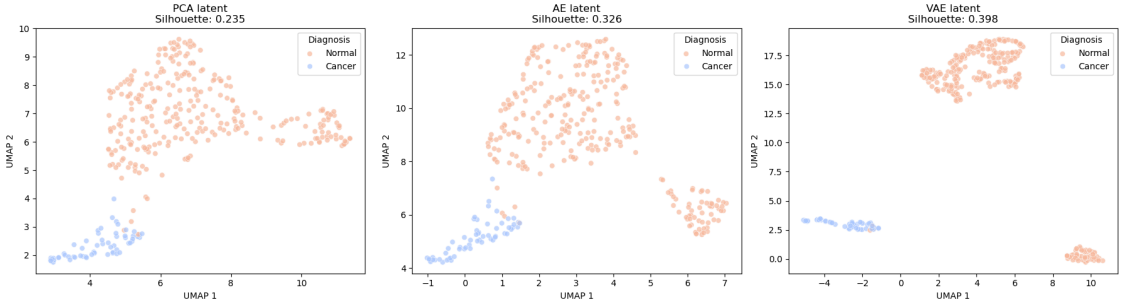

Figure 14: Latent space comparison across PCA, AE, and VAE embeddings. The VAE yields the strongest separation by diagnosis according to silhouette score.

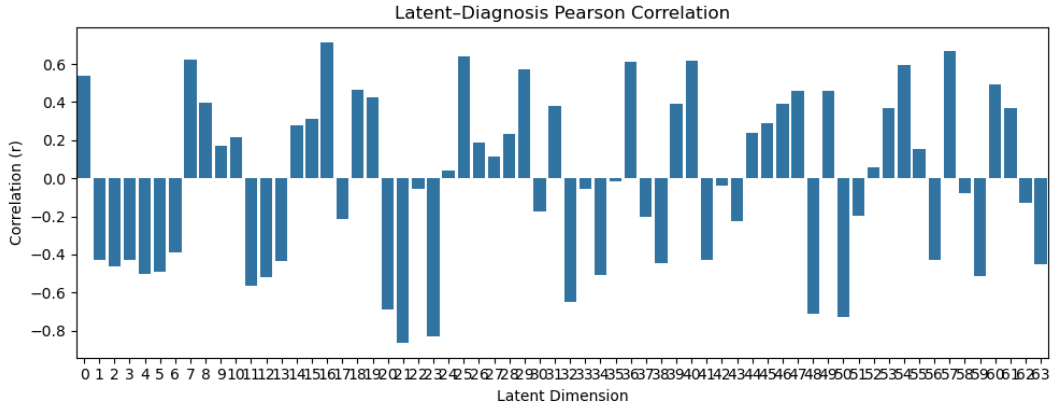

Figure 15: Pearson correlation between each latent dimension and the diagnosis label. Multiple dimensions show strong association with diagnosis.

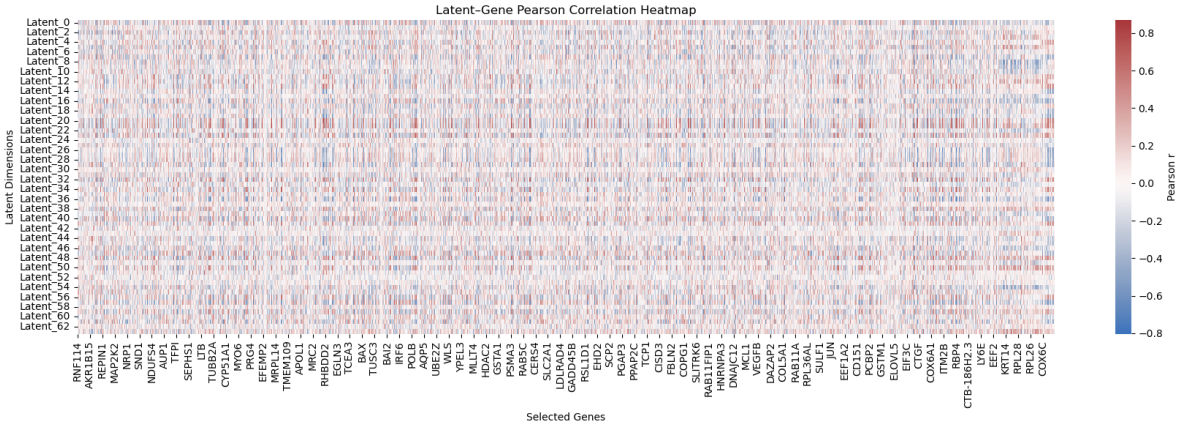

Figure 16: Heatmap of correlations between latent dimensions and gene expression values for selected genes, supporting interpretability of latent factors.

6 Discussion

This Section interprets the results, connects them to the thesis goals, and discusses what the findings imply for modeling transcriptomic breast cancer data in practice. A central theme is that raw diagnostic performance alone is not sufficient for justifying model complexity: strong baselines, domain shift, and interpretability constraints must be considered.

6.1 Replication of anomaly-detection VAE: what works and what is fragile

The replication experiments confirm the core intuition behind reconstruction-based anomaly detection: a VAE trained on normal samples tends to assign higher reconstruction error to tumor samples. This produces a meaningful ROC-AUC and demonstrates that reconstruction error can serve as an anomaly score for cancer detection. However, the classification report indicates that tumor recall is noticeably lower than normal recall, reflecting overlap between tumor and normal reconstruction errors. This overlap is expected in biological data: tumors are heterogeneous, and some tumor samples can remain close to normal tissue along certain expression axes (e.g., contamination with normal tissue, tumor purity variation, or subtype-dependent similarity).

At the same time, anomaly detection pipelines are particularly sensitive to evaluation protocol. Unlike supervised classifiers, anomaly detectors require an explicit decision threshold on the anomaly score, and this threshold can be inadvertently tuned on the test set if experimental hygiene is not enforced. In this thesis, thresholds were selected using validation data only, and test performance was reported using a fixed threshold to avoid optimistic bias. The replication therefore supports two conclusions: (i) reconstruction-based scoring is feasible for breast cancer detection, and (ii) careful split and threshold design is essential for scientifically defensible evaluation.

6.2 Multi-head VAE: strong diagnosis, moderate subtyping, and representation value

The multi-head VAE achieved strong diagnosis performance (high accuracy and ROC-AUC) while simultaneously predicting PAM50 subtype among labeled tumor samples. Subtype prediction performance was moderate and heterogeneous across classes. This pattern is consistent with two properties of PAM50 classification: (i) class imbalance (e.g., LumA dominating the distribution) and (ii) intrinsic biological overlap between subtypes, especially between LumA and LumB, which share many expression programs but differ in proliferation-related and hormonal signaling patterns. As a result, subtype boundaries are less clean than the binary tumor-vs-normal boundary.

A key benefit of the multi-head design is not merely predictive accuracy but the learned latent representation. Representation analyses showed improved separation in the VAE latent space compared to PCA and a standard autoencoder, suggesting that the VAE objective produces a more structured embedding. In addition, latent dimensions exhibited meaningful correlations with diagnosis and with groups of genes, indicating that diagnostic signal is concentrated in specific latent factors. These findings support the view that the VAE learns a compressed representation that can be used for visualization and exploratory interpretation, which is valuable even when simpler models can match or exceed diagnosis accuracy.

6.3 Why logistic regression is near-ceiling on diagnosis

The logistic regression baseline achieved near-perfect ROC-AUC on diagnosis, substantially higher than the anomaly-detection VAE and slightly higher than the multi-head VAE. This result strongly suggests that, after preprocessing, the tumor vs normal decision boundary is close to linearly separable in the chosen feature space. In such cases, linear models are difficult to beat because they match the functional form needed for separation while requiring less capacity and less optimization complexity than deep generative models.

Importantly, this does not imply that the VAE approach is "wrong" or unnecessary. Instead, it reframes the scientific claim: in this dataset, the main value of VAEs is not superior diagnosis accuracy but (i) learning a shared representation for multiple tasks, (ii) enabling structured low-dimensional embeddings, and (iii) supporting principled compression experiments. The baseline comparison is therefore essential for honest interpretation: deep models should be justified by additional goals beyond raw classification performance.

6.4 Gene panels: interpretability and performance–compression trade-offs

A practical motivation for transcriptomic modeling is the development of compact, interpretable gene panels that reduce measurement cost and improve clinical feasibility. Sparse L_1 logistic regression provides a principled mechanism for selecting such panels by identifying a small set of genes with non-zero coefficients. In this thesis, diagnostic

performance remained high even when restricting inputs to a 20-gene panel, indicating that much of the diagnostic signal is concentrated in a small subset of genes.

For subtype prediction, a subtype-oriented panel constructed from sparse multinomial logistic regression captured subtype-discriminative genes and enabled competitive performance in the compressed setting. When combining diagnosis- and subtype-oriented genes into a single compact panel and retraining the multi-head VAE, both diagnosis and subtype performance remained strong. This suggests that multi-task representation learning can be compatible with aggressive feature reduction, and that panel construction can be guided by complementary objectives: diagnosis separability and subtype discriminability.

6.5 Why normal-oriented panels can improve anomaly detection

A notable finding is that, for reconstruction-error anomaly detection, normal-oriented panels outperformed the cancer-oriented panel, even when both panels contained the same number of genes. This can be explained by the training objective of the anomaly VAE: it is optimized to model the distribution of normals. Features that are stable and consistently expressed among normals make the normal manifold easier to learn and reduce reconstruction variance on inliers. In contrast, cancer-oriented genes are selected to maximize discriminative separation between tumor and normal in a supervised setting; such genes may include patterns that are highly variable even among normals or may encode tumor-specific structure that the normal-trained VAE cannot reconstruct reliably. Consequently, reconstruction error may become noisier and less aligned with the "distance from normal" concept.

This finding highlights an important conceptual distinction: *features optimal for supervised discrimination are not necessarily optimal for one-class generative modeling*. In anomaly detection, the choice of features should be aligned with the modeling goal (learning normals) rather than the downstream classification goal alone.

6.6 Domain shift: implications for generalization and deployment

The domain shift analysis shows that GTEx normals and TCGA normals can be distinguished with very high accuracy and near-perfect ROC-AUC. Since both represent normal breast tissue, such separability strongly suggests the presence of dataset-specific signal (batch effects or cohort differences). This has major implications:

- **Risk of shortcut learning.** A diagnosis model trained on a mixture of GTEx and TCGA data may partially rely on cohort artifacts correlated with labels (e.g., GTEx contributing many normals while TCGA contributes all tumors), inflating measured performance.
- **Threshold instability for anomaly detection.** In reconstruction-based anomaly detection, the learned "normal" manifold may depend strongly on which source dominates the normal training set. This can shift reconstruction error distributions and change the operating point of a fixed threshold.
- **Limited external validity.** High in-cohort accuracy does not guarantee performance on new hospitals, new sequencing pipelines, or prospective cohorts. Domain shift must therefore be treated as a primary limitation.

These results motivate future work on harmonization and domain adaptation, such as batch correction strategies, explicit domain-invariant representation learning, or evaluation on independent external datasets.

6.7 Summary

In summary, the experiments support five main conclusions. First, reconstruction-error VAEs can detect tumors as anomalies, but their performance and operating thresholds are sensitive to evaluation protocol. Second, the proposed multi-head VAE provides a single latent representation supporting both diagnosis and PAM50 subtype prediction, and the learned latent space exhibits meaningful structure. Third, diagnosis in this dataset is near-linearly separable, as evidenced by near-ceiling logistic regression performance, implying that deep generative models should be justified by goals beyond accuracy alone. Fourth, compact gene panels can preserve high diagnosis and competitive subtype performance, enabling practical compression and interpretability. Finally, substantial domain shift exists between GTEx and TCGA normals, limiting generalization and highlighting the need for careful cross-cohort validation and harmonization in transcriptomic machine learning.

7 Limitations

This thesis presents a leakage-resistant evaluation of VAE-based anomaly detection, a multi-head VAE for diagnosis and PAM50 subtyping, and gene panel compression experiments. Despite strong results, several limitations restrict interpretation and generalizability.

7.1 Domain shift and batch effects

A central limitation is the presence of substantial domain shift between GTEx normals and TCGA normals, demonstrated by near-perfect separability of source labels among normal samples. This indicates dataset-specific signal unrelated to biology (e.g., protocol and preprocessing differences). As a result, diagnostic performance estimates may be inflated if models partially rely on cohort artifacts correlated with the diagnosis label. In real deployment scenarios, such shortcuts may not transfer across hospitals, sequencing pipelines, or prospective cohorts.

7.2 Scope limited to breast cancer and the available cohorts

All experiments focus on breast cancer (BRCA) and the specific GTEx/TCGA data files available in this project. The results therefore do not establish general performance across tissues, cancer types, or additional cohorts. The original paper evaluated multiple tissues, whereas this thesis prioritizes depth of methodological analysis and multi-task extension in one tissue. Extending conclusions beyond BRCA requires additional validation.

7.3 Limited external validation

Models were evaluated using internal train/validation/test splits derived from the combined dataset. Although strict split discipline and validation-only thresholding were used, no completely independent external cohort was available for testing. Therefore, reported metrics represent within-dataset generalization rather than true external validity. An important next step would be evaluation on an independent breast cohort processed with a different pipeline.

7.4 Subtype label quality, missingness, and class imbalance

PAM50 subtype labels were missing for a subset of samples and were handled via a subtype mask. While this prevents mislabeled or unlabeled samples from contaminating subtype supervision, it reduces the effective sample size for subtype training and evaluation. In addition, PAM50 classes are imbalanced (e.g., LumA is dominant, Normal-like has low support), which increases variance of per-class estimates. In particular, metrics for the smallest subtype class should be interpreted cautiously.

7.5 Threshold dependence in anomaly detection

Anomaly detection requires selecting a threshold on reconstruction error to produce binary decisions. Although this thesis selects thresholds on validation data only, the operating point remains sensitive to the class mixture used in validation and to the intended clinical trade-off between false positives and false negatives. ROC-AUC provides a threshold-free measure, but deployed systems require explicit threshold setting, which should be calibrated on representative validation data for the target deployment environment.

7.6 Gene panel stability and biological interpretation

Gene panels were derived using sparse linear models and training-set statistics, producing compact and interpretable sets of genes. However, panel selection can be sensitive to sampling variation, regularization strength, and dataset composition. While cross-validation and shuffled-label controls support the robustness of the diagnosis baseline, a full stability analysis of selected genes (e.g., bootstrapped selection frequencies) was not performed for all panels. Furthermore, coefficient magnitude or selection by a predictive model does not imply causal biological relevance; selected genes should be interpreted as predictive markers in this dataset rather than definitive mechanistic drivers.

7.7 Modeling and preprocessing assumptions

Several modeling choices may influence results, including the choice of min–max scaling, the use of variance-based feature selection, and the use of mean aggregation for duplicate TCGA patient IDs. These are reasonable and common choices, but alternative approaches (e.g., log-transformations, quantile normalization, or explicit batch correction) could change both separability and learned representations. Additionally, reconstruction-based modeling assumes that the

chosen scaling and reconstruction loss (MSE) are appropriate for transcriptomic data, which may not perfectly capture count-based noise characteristics.

7.8 Summary

In summary, the most important limitations are domain shift between cohorts, lack of independent external validation, subtype label imbalance and missingness, threshold dependence in anomaly detection, and limited analysis of gene panel stability. These limitations motivate the future work proposed in the conclusion.

8 Conclusion and Future Work

This thesis investigated breast cancer detection and PAM50 molecular subtyping from transcriptomic gene expression through three complementary lenses: replication of VAE-based anomaly detection, a multi-head VAE for joint supervised learning, and sparse gene panel construction with compression experiments.

8.1 Conclusion

The experiments support five main conclusions:

1. **Anomaly detection with VAEs is feasible but evaluation-sensitive.** A VAE trained on normal samples can detect tumors using reconstruction error as an anomaly score, but performance depends strongly on split discipline and validation-based threshold selection.
2. **A multi-head VAE enables joint diagnosis and subtyping with a shared latent representation.** The proposed architecture achieved strong diagnosis performance and competitive PAM50 subtype prediction while learning a structured latent space useful for visualization and interpretability analyses.
3. **Diagnosis is near-linearly separable in this dataset.** Logistic regression achieved near-ceiling diagnostic ROC-AUC, indicating that improved diagnostic accuracy alone is not a sufficient justification for model complexity in this setting.
4. **Compact gene panels preserve much of the predictive signal.** Sparse linear models identify compact panels that maintain high diagnostic performance, and combined diagnosis–subtype panels allow multi-task models to operate under severe feature compression.
5. **Domain shift is a dominant concern for real-world generalization.** GTEx and TCGA normals are highly separable, implying that cohort artifacts can influence model behavior and that external validation is necessary before deployment-oriented claims can be made.

Overall, the results suggest that while simple linear models are sufficient for diagnosis in highly separable settings, VAEs remain valuable for learning shared representations across tasks, exploring latent biological structure, and studying compression and anomaly detection behavior under constrained feature sets.

8.2 Future work

Several directions would strengthen and extend this work:

External validation and cross-cohort evaluation. Evaluate all models on an independent breast cancer cohort processed with a distinct pipeline. This is essential for assessing true generalization and quantifying performance degradation under domain shift.

Batch correction and domain-invariant modeling. Incorporate harmonization methods (e.g., batch correction or domain-adversarial training) to reduce source-specific artifacts. Compare diagnosis and anomaly performance before and after correction to measure how much of the predictive signal is technical versus biological.

Panel stability analysis. Quantify gene panel stability via bootstrapping or repeated resampling, reporting selection frequencies and identifying a "core" set of robust genes. This would strengthen interpretability claims and improve reproducibility.

Calibration of anomaly thresholds. Study how anomaly thresholds change as the class mixture or the source composition of normals changes, and evaluate clinically motivated operating points (e.g., high sensitivity screening vs high specificity confirmation).

Biological interpretation of latent factors. Extend interpretability by testing enrichment of genes correlated with latent dimensions using pathway databases, and relate latent factors to known breast cancer biology (e.g., estrogen signaling, proliferation, immune infiltration).

Extension beyond BRCA. Apply the same methodology to additional tissues or cancer types to test whether the conclusions (linear separability, panel compression behavior, and anomaly detection trends) generalize beyond breast cancer.

8.3 Closing remarks

Transcriptomic modeling offers a powerful lens on cancer biology, but it also amplifies common pitfalls in machine learning evaluation due to high dimensionality and cohort heterogeneity. By combining rigorous experimental hygiene, strong baselines, multi-task representation learning, and panel-based compression, this thesis provides an honest and reproducible assessment of where VAEs help—and where simpler methods already suffice—for breast cancer detection and PAM50 molecular subtyping.

References

[1] Qingguo Wang, Joshua Armenia, Chao Zhang, Alexander V. Penson, Ed Reznik, Liguo Zhang, et al. Unifying cancer and normal RNA sequencing data from different sources. *Scientific Data*, 5(1), 2018. doi: 10.1038/sdata.2018.61.

[2] Joel S. Parker, Michael Mullins, Maggie C. U. Cheang, Samuel Leung, David Voduc, Tom Vickery, Sherri Davies, Claire Fauron, Xiaohui He, Zhen Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, 2009. doi: 10.1200/JCO.2008.18.1370.

[3] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

[4] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007. doi: 10.1093/biostatistics/kxj037.

[5] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, 2009. URL https://mitpress.mit.edu/9780262170055/dataset-shift-in-machine-learning/.

[6] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013.

[7] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1278–1286, 2014. URL https://proceedings.mlr.press/v32/rezende14.html.

[8] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

[9] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey, 2019. URL https://arxiv.org/abs/1901.03407.

[10] The GTEx Consortium. The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580–585, 2013. doi: 10.1038/ng.2653.

[11] GTEx Consortium. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015. doi: 10.1126/science.1262110.

[12] The Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013. doi: 10.1038/ng.2764.

[13] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[14] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. URL https://arxiv.org/abs/1502.03167.

[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL https://www.jmlr.org/papers/v15/srivastava14a.html.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

[18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. URL https://arxiv.org/abs/1802.03426.

[19] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

[20] I. T. Sado, A. Fitime, F. B. Pelap, H. A. Tinku, and T. B. Meudje Bouetou. Early multi-cancer detection through deep learning: An anomaly detection approach using variational autoencoder. *Journal of Biomedical Informatics*, 160:104751, 2024. doi: 10.1016/j.jbi.2024.104751.