Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.081129

ARTICLE

Interpretable Deep Representation Learning for Pan-Cancer Diagnosis via Pathway-Constrained Transcriptomics

Maram Fahaad Almufareh[1,*] **and Samabia Tehsin**[2,*]

[1]Department of Information Systems, College of Computer and Information Sciences, Jouf University, Al-Jawf, Saudi Arabia
[2]Center of Excellence–AI, Bahria University, Islamabad, Pakistan

*Corresponding Authors: Maram Fahaad Almufareh. Email: mfalmufareh@ju.edu.sa;
Samabia Tehsin. Email: stehseen.buic@bahria.edu.pk

Received: 24 February 2026; Accepted: 12 May 2026; Published: 30 June 2026

ABSTRACT: This article presents a Hierarchical Pathway-Masked Attention Autoencoder (H-PAAE), a biologically inspired representation-learning framework that enables explainable AI-guided cancer diagnosis. The model directly integrates the curated MSigDB Hallmark pathways, introducing pathway-constrained information flow and mechanistic interpretability through multi-level attention mechanisms. Based on TCGA RNA-seq data from 33 tumor types, H-PAAE compresses approximately 20,000 genes into a 128-dimensional latent space while preserving biologically meaningful structure. When used with XGBoost classification, H-PAAE delivers 92.37% test accuracy and 99.38% macro-AUROC with robust cross-validation results (92.5 ± 0.6%). SHAP analysis identifies a small number of key latent features, corresponding to conserved oncogenic processes, and pathway enrichment analysis shows strong overlap with cancer hallmarks. H-PAAE provides a clear and interpretable biological foundation for pan-cancer classification, with well-calibrated posterior probabilities that can be used for clinical decision-making, and can be easily integrated into multimodal diagnostic workflows.

KEYWORDS: Computational biology; bioinformatics; cancer computational biology; transcriptomics; interpretable deep learning; pan-cancer analysis; gene expression analysis; machine learning in genomics

1 Introduction

Clinically relevant and efficient cancer diagnostics increasingly rely on AI models that can integrate molecular signatures with a biologically meaningful interpretation. Pan-cancer and solid tumor transcriptomic profiling is crucial to comprehend oncogenesis mechanisms; however, the existing deep learning (DL) approaches to disease classification are often black-box models, which limit clinical acceptance, regulatory approval and multimodal integration with imaging-based diagnostic approaches.

Cancer has shadowed human history for millennia, but its framing and treatments have shifted dramatically with the rise of scientific medicine. Traditional medical systems documented cancer-like syndromes long before the modern pathology era, emphasizing whole-person management, symptom control, and pattern-based therapy [1]. The nineteenth and twentieth centuries brought cellular pathology and staged surgery; soon after came megavoltage radiotherapy, and—transformationally in the mid-twentieth century—systemic cytotoxic chemotherapy. Over the last three decades, precision oncology has added targeted agents and immunotherapies, alongside progressively safer and more discriminating diagnostic pathways (e.g., risk calculators integrated with imaging) [2]. In parallel, nonionizing imaging and advanced biosensing

have matured, offering complementary avenues for earlier detection and longitudinal monitoring outside of high-resource hospital settings [3,4].

Despite genuine therapeutic progress, cancer's worldwide toll remains immense. The International Agency for Research on Cancer (IARC) estimates that in 2022 there were nearly 20 million new cancer cases and about 9.7 million deaths across 185 countries, with breast and lung cancers dominating incidence in women and men, respectively, and lung cancer the leading cause of cancer death overall [5,6]. These headline numbers mask striking heterogeneity by region and development level. Contemporary maps from the Global Cancer Observatory show wide variation in age-standardized incidence and mortality across the 20 world regions—higher incidence in more developed regions (driven by screening and longevity) contrasted with higher case fatality in many low- and middle-income regions due to late stage at presentation and constrained access to treatment [7].

A country-level view highlights the cross-currents shaping trends. In the United States, the American Cancer Society projects 2,041,910 new cases and 618,120 deaths in 2025. Mortality has continued a long decline since 1991—averting an estimated 4.5 million deaths—attributed to tobacco control, earlier detection in some cancers, and better therapy; yet incidence is rising among women and younger adults, and stark racial/ethnic disparities persist (e.g., higher prostate, stomach, and uterine mortality in Black populations; highest overall cancer mortality in many Native American communities) [8]. Region- and country-specific factsheets and registries therefore matter: they indicate where prevention, screening, and treatment access succeed—and where they fall short [6,7].

The diagnostic arc has expanded beyond the classic triad of clinical exam, imaging, and histopathology. Evidence-based screening (e.g., mammography, cervical cytology/HPV testing, colonoscopy, and targeted lung CT) shifts detection toward earlier, more curable stages when implemented equitably and at scale [9,10]. Within organ-specific pathways, decision support now frequently blends risk factors with modality-specific scores. For instance, multiparametric MRI interpreted via PI-RADS and summarized in multivariable risk calculators improves discrimination for clinically significant prostate cancer and can reduce unnecessary biopsies, with pooled AUCs around 0.84 for significant disease in meta-analysis [2].

Meanwhile, engineering advances aim for safer, more accessible detection. In breast cancer, ultra-wideband (UWB) microwave approaches use nonionizing radiation and wideband antennas to improve tissue contrast, with active research on antenna design near the body, calibration, and artifact reduction for clinically useful imaging [3]. Outside radiology suites, biosensing innovations are pushing toward low-cost, point-of-care detection of circulating biomarkers. Two-dimensional graphene nanomaterials—tunable by synthesis and functionalization—have enabled electrochemical, optical, and field-effect devices with promising limits of detection; still, clinical translation requires standardization, robust validation, and attention to biocompatibility and manufacturing [4]. Complementary and integrative modalities continue to be studied as adjuncts to conventional regimens, not substitutes, with modernization of clinical trial methods and mechanistic work (e.g., immunomodulation, resistance reversal) especially visible in lung cancer [1].

Current procedural pathways are still largely defined by the underlying organ systems. For lung cancer, for instance, diagnostic, staging, and treatment planning procedures rest on the chest imaging framework (radiography, CT, PET/CT), and performative procedures. Some of the more persistent issues, such as subtle nodules, reader variability, and issues overlapping with non-neoplastic disease, are the raisons d'etre of standard acquisition and reporting as well as more rigorously assessed computer tools [11]. In dermatology, the primary tools are dermoscopy and clinical photographs; and with the variability of the appearance of the lesions, the conditions of the capture, and the lesions themselves, acquisition and segmentation of the lesions is crucial for reliable assessment [12]. Implementation science for various specialties brings forth similar

issues—data quality, changes in care location, system transparency, workflow congruence, and equity—as systems transition from proofs of concept to sustained clinical value [13].

Contemporary cancer control depends on population registries and coordinated surveillance. IARC's GLOBOCAN assembles incidence and mortality estimates for 36 cancers across 185 countries, combining population-based cancer registries and vital statistics with quality and timeliness adjustments to provide reproducible, uncertainty-aware snapshots each cycle [6]. In the United States, decades of incidence and survivorship trends rest on the SEER and NPCR infrastructures, enabling annual ACS projections, state-by-state statistics, and disaggregated disparity tracking [8]. Beyond surveillance, *research consortia and reference cohorts* accelerate discovery by harmonizing biospecimens, assays, and clinical data.

A landmark of this second category is The Cancer Genome Atlas (TCGA)—a coordinated, multi-platform molecular characterization of >11,000 primary tumors spanning 33 cancer types, completed with the Pan-Cancer Atlas in 2018 [14–16]. TCGA pairs standardized genomic and transcriptomic profiles with curated clinical annotations, enabling cross-tumor comparisons that surface cell-of-origin programs, pathway disruptions, and recurrent molecular themes. While clinical care remains organ-specific, a pan-cancer molecular view is invaluable whenever presentation is ambiguous (e.g., cancer of unknown primary), imaging is indeterminate, or multiple differential diagnoses compete.

In this multi-cancer classification study, the classifier was developed based on the TCGA cohort, which spans over 33 tumor types. The objective is practical: to be able to develop a new layer of the organ-based diagnostic processes which is flexible, well calibrated, and molecularly integrated. Such reconstructions offer a strong baseline to researchers, and a clinically rational starting point for differential diagnosis when organ of origin is uncertain, provided that the integration with known tumor biology is precise and interpretative alignment is reported.

2 Literature Review

The implementation of machine learning technology together with computational tools have made new advances concerning the understanding of tumor biology and tumor typing in cancer genomics. Almost all large-scale multi-omic databases such as The Cancer Genome Atlas underscores the fact that researchers have taken various approaches—from classical machine learning types to sophisticated deep learning types—in the classification of neoplasm types, sub-typing, predicting their prognosis, and evaluating their therapeutic response. Across studies, analytical methodologies and machine learning have utilized and employed "big data" with transcriptional profiling to cancer, and clinical outcome predictive modeling systems to derive treatment plans and de novo molecular characterization with the immunodiagnostic and therapeutic assessment.

Liñares-Blanco et al. [17] carried out a detailed meta-analysis of over 100 studies that apply machine learning to TCGA data, establishing Random Forest and Support Vector Machines as the most popular algorithms, with deep neural networks and multi-omic integration beginning to be used. Their survey divided applications into predicting prognosis, classifying tumor subtypes, detecting microsatellite instability, immunological profiling, and pathway analysis. It showed that, one, gene expression was the most used and, two, algorithm selection was tumor-type dependent. Tomczak et al. [18] set up the foundational context of TCGA by describing its infrastructure and multi-platform characterization across 30+ tumor types, with particular focus on the integrated analysis of genome, transcriptome, and epigenome that led to the identification of clinically relevant subtypes of molecular configuration in brain, lung, and ovarian cancers.

Winterhoff et al. [19] validated TCGA transcriptional subtypes across 276 high-grade ovarian cancers using non-negative matrix factorization, confirming four signatures (immunoreactive, differentiated,

proliferative, mesenchymal) with prognostic value and demonstrating that 59% of clear cell and 45% of endometrioid tumors formed early-stage clusters distinct from serous cancers. Thennavan et al. [20] extended breast cancer histologic classification by integrating RNA-seq, whole-exome, and methylation data for 1095 samples, developing a mucinous gene signature that generalized across organ systems in the Pan-Cancer Atlas and establishing 12 consensus molecular-histologic groups. Yang et al. [21] developed a 16-microRNA signature for lung adenocarcinoma staging using TCGA-LUAD (n = 470) with external validation on independent cohorts (GSE62182, GSE83527), achieving modest AUC (0.62–0.66) but highlighting challenges in cross-cohort generalization of molecular biomarkers.

Multimodal integration has emerged as a promising direction. Li et al. [22] applied quantitative MRI radiomics to 91 TCGA breast cancers, achieving AUC 0.89 for ER status prediction through computer-extracted image phenotypes, demonstrating that morphological features encode molecular information suitable for non-invasive characterization. Mohammed et al. [23] proposed a stacking ensemble combining 1D-CNNs with LASSO feature selection for five women's cancers, outperforming single models and traditional ML approaches while emphasizing the importance of addressing class imbalance through under-sampling strategies.

Single-gene and pathway-focused analyses have complemented genome-wide approaches. Guo et al. [24] identified secreted phosphoprotein 1 (SPP1) as an independent prognostic marker in TCGA lung adenocarcinoma (HR = 1.150, 5-year survival 50.6% vs. 59.2% for high vs. low expression), with GSEA linking SPP1 to mTORC1, angiogenesis, and glycolysis pathways. Lombardi et al. [25] derived a 48-gene HIF metagene from 72 ChIP-seq and RNA-seq datasets, providing an endogenous pan-cancer marker of hypoxia-inducible factor activation applicable to bulk and single-cell analyses without direct oxygen measurement.

Recent advances in image-to-transcriptome prediction leverage deep learning to extract molecular profiles from routine histology. Schmauch et al. [26] developed HE2RNA, a CNN predicting RNA-seq from H&E slides with attention-based spatial localization, demonstrating feasibility of MSI detection from morphology alone. Zheng et al. [27] advanced this paradigm with SEQUOIA, a transformer employing grouped vision attention to predict 11,069 genes in breast cancer (of 25,749 total) and derive digital recurrence risk signatures, validated across independent cohorts. Alsaafin et al. [28] introduced tRNAsformer for joint transcriptome prediction and image classification in renal cell carcinoma, achieving superior performance through hierarchical attention and multiple instance learning. Suphavilai et al. [29] addressed intra-tumor heterogeneity with CaDRReS-Sc, combining single-cell RNA-seq and matrix factorization to predict clone-specific drug responses (80% accuracy, Pearson r > 0.6 for monotherapy), though clinical translation awaits prospective validation.

Preprocessing and methodological choices critically influence model performance. Shahriyari [30] systematically compared normalization strategies (scaling, z-score, vector) on TCGA HTSeq-FPKM-UQ data across 12 algorithms, finding that SVM with RBF kernel achieved 78% accuracy for colon adenocarcinoma survival regardless of method, though computational efficiency varied substantially. Dimensionality reduction identified 7SK RNA as a single-gene predictor, underscoring the value of systematic preprocessing optimization.

Despite these advancements, there is still a gap that needs to be addressed. Most studies concentrate on a single cancer, or a limited multi-cancer range (typically 5–9 types), thus leaving a 33-type pan-cancer classification system unexplored. Current methodologies tend to ignore the biological understanding that comes with the problem and, instead, rely on black-box deep networks with no mechanistic pathway synthesis. Moreover, the accuracy of the model and how it can be made transparent—both of which are important for real-world application—receive less focus. This paper seeks to tackle these issues with the

application of a hierarchical pathway attention autoencoder (H-PAAE) that incorporates pathway knowledge (Hallmarks of MSigDB) and other biological knowledge to learn compact latent representations for all 33 TCGA cancer types, yielding multi-level interpretability through various attention mechanisms and SHAP analysis, and achieving 93% test accuracy with well-calibrated guarantees appropriate for clinical decision support.

Comparative Summary of Related Studies

Table 1 summarises representative recent studies on deep learning-based cancer classification from transcriptomic and multi-omics data, highlighting dataset, method, scope, and accuracy to contextualise the proposed H-PAAE framework.

Table 1: Comparative summary of recent related studies on ML/DL-based cancer classification from gene expression and multi-omics data.

Study	Dataset	Method	Cancers	Accuracy
Jiang & Hassanpour (2025) [31]	TCGA RNA-seq	GexBERT (transformer encoder-decoder)	14 types	97.9%
Younis & Minghim (2025) [32]	BARRA:CuRDa RNA-seq	CNN + XAI biomarker identification	8 types	~87%
Shanmugam et al. (2025) [33]	TCGA-PANCAN-HiSeq	SGMS feature selection + SHAP	5 types	99.8%
Benkirane et al. (2025) [34]	TCGA (WSI + multi-omics)	Multimodal VAE + pathway interpretability	Multiple	Multi-task
Zhang et al. (2025) [35]	LUAD/LUSC multi-omics	MOLUNGN (multi-omics graph attention)	Lung (2)	84% ACC
Proposed H-PAAE	TCGA RNA-seq (33 types)	Pathway-masked attention AE + XGBoost	33 types	92.37%

3 Methods

This part describes a computation framework for classifying different kinds of cancer using gene expression data from The Cancer Genome Atlas (TCGA). The methodology involves three steps: data preprocessing with feature selection, unsupervised representation learning using a pathway-masked hierarchical attention autoencoder, and supervised classification via gradient boosting. Fig. 1 captures the main steps of the pipeline from raw data collection to final classification.

3.1 Data Acquisition and Preprocessing

The Cancer Genome Atlas (TCGA) provides RNA-sequencing for 33 different tumors cancer types with the FPKM-UQ (Fragments Per Kilobase Million Upper Quartile) normalized gene expression values generated from the GDC (Genomic Data Commons) STAR-Counts pipeline from the GDC [36]. FPKM-UQ normalization gives more reliable measurements by normalizing the expression values to the upper quartile of non-zero counts, reducing the effects of overabundant genes [37].

To facilitate reproducibility, the data was obtained from the TCGA Archive using the filter hierarchy: RNA-Seq → Transcriptome Profiling → Gene Expression Quantification → STAR Counts → Tumor →

Primary. This set of filters aims at primary tumors processed through the STAR aligner in order to focus the analysis on the primary cancer, avoiding recurrent, metastatic, normal, and other tissue samples.

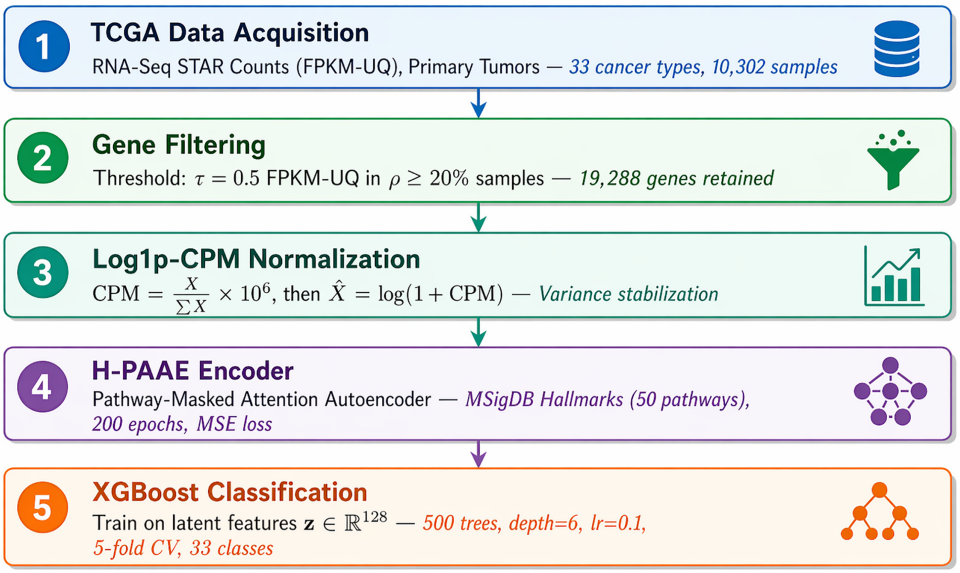


Figure 1: Methodology overview: TCGA RNA-Seq preprocessing, H-PAAE 128-D feature learning, and XGBoost classification across 33 cancers.

3.1.1 Low-Expression Filtering

To reduce noise from unreliably measured transcripts, a two-stage filtering strategy was applied. Let $X \in \mathbb{R}_+^{N \times G}$ denote the raw expression matrix (FPKM-UQ values) for N samples across G genes, where X_{ij} represents the expression value for sample i and gene j.

Genes that exceed a threshold τ (FPKM-UQ units) in a minimum fraction ρ of samples are retained. Formally, the set of kept genes is:

$$\mathcal{G} = \left\{ g \in \{1, \ldots, G\} \mid \frac{1}{N} \sum_{i=1}^{N} \mathcal{I}[X_{i,g} \geq \tau] \geq \rho \right\} \tag{1}$$

where $\mathcal{I}[\cdot]$ is the indicator function, $\tau = 0.5$ FPKM-UQ is the minimum expression threshold, and $\rho = 0.2$ (20%) is the minimum sample fraction parameter. This filtering strategy removes lowly expressed and sporadically detected genes while preserving biologically relevant transcripts [38]. Let $G' = |\mathcal{G}|$ denote the number of retained genes.

3.1.2 Log-Transformed CPM Normalization

Following gene filtering, counts-per-million (CPM) normalization was applied to account for library size differences across samples. For each sample i, the CPM-normalized expression for gene $g \in \mathcal{G}$ is computed as:

$$\tilde{X}_{i,g} = \frac{X_{i,g}}{\sum_{g' \in \mathcal{G}} X_{i,g'}} \times 10^6 \tag{2}$$

where the denominator represents the library size $s_i = \sum_{g' \in \mathcal{G}} X_{i,g'}$ for sample i.

To stabilize variance and approximate normality, a logarithmic transformation with pseudocount (log1p) was applied:

$$\hat{X}_{i,g} = \log(1 + \tilde{X}_{i,g}) \tag{3}$$

This log1p-CPM transformation is widely adopted in transcriptomics analyses for variance stabilization and to reduce the influence of extreme values [39,40]. The resulting preprocessed matrix $\hat{X} \in \mathbb{R}^{N \times G'}$ serves as input to downstream analyses. Algorithm 1 summarizes the complete preprocessing pipeline.

Algorithm 1: Preprocessing: gene filtering + log1p–CPM

Require: Expression matrix $X \in \mathbb{R}_{+}^{N \times G}$ (FPKM-UQ), threshold τ, fraction ρ
Ensure: Preprocessed matrix $\hat{X} \in \mathbb{R}^{N \times G'}$
$\mathcal{G} \leftarrow \{g \in \{1, \ldots, G\} : \frac{1}{N} \sum_{i=1}^{N} \mathcal{I}[X_{i,g} \geq \tau] \geq \rho\}$ ▷ keep sufficiently expressed genes
$X' \leftarrow X[:, \mathcal{G}]$; $G' \leftarrow |\mathcal{G}|$
for $i = 1$ to N **do**
  $s_i \leftarrow \sum_{g=1}^{G'} X'_{i,g}$; $\tilde{X}_{i,:} \leftarrow \frac{10^6}{s_i} X'_{i,:}$ ▷ CPM scaling
  $\hat{X}_{i,:} \leftarrow \log(1 + \tilde{X}_{i,:})$ ▷ log1p
end for
return $\hat{X}$

3.2 Hierarchical Pathway-Masked Attention Autoencoder

To learn compact, biologically interpretable representations of gene expression profiles, a Hierarchical Pathway-masked Attention Autoencoder (H-PAAE) was developed. This architecture incorporates prior biological knowledge through pathway-based masking and employs attention mechanisms to weight pathway contributions.

3.2.1 Pathway-Based Gene Grouping

Genes were organized into functional modules using curated pathway databases (e.g., MSigDB [41,42] or KEGG [43]). Let $\mathcal{P} = \{P_1, P_2, \ldots, P_P\}$ denote a collection of P pathways, where each pathway $P_k \subseteq \mathcal{G}$ is a subset of the retained genes. Genes not assigned to any pathway form singleton pathways.

A binary gene-pathway incidence matrix $M \in \{0,1\}^{G' \times P}$ is defined where $M_{g,p} = 1$ if and only if gene g belongs to pathway p. This matrix enforces pathway-based masking in the encoder, ensuring that information flows only along biologically meaningful connections [42].

3.2.2 Encoder Architecture

The encoder consists of two stages: pathway-level aggregation followed by attention-weighted integration. Fig. 2 illustrates the complete encoder-decoder architecture with pathway masking and hierarchical attention mechanisms.

Pathway Aggregation Layer. For each pathway p, a pathway-level embedding is computed by aggregating the expression values of its constituent genes using masked attention. Let $\hat{\boldsymbol{x}} \in \mathbb{R}^{G'}$ denote the preprocessed expression vector for a single sample. For genes g where $M_{g,p} = 1$, attention scores $\alpha_{p,g}$ are computed and aggregated to form pathway representation u_p. This masked attention mechanism ensures that only genes belonging to pathway p contribute to its representation [44,45].

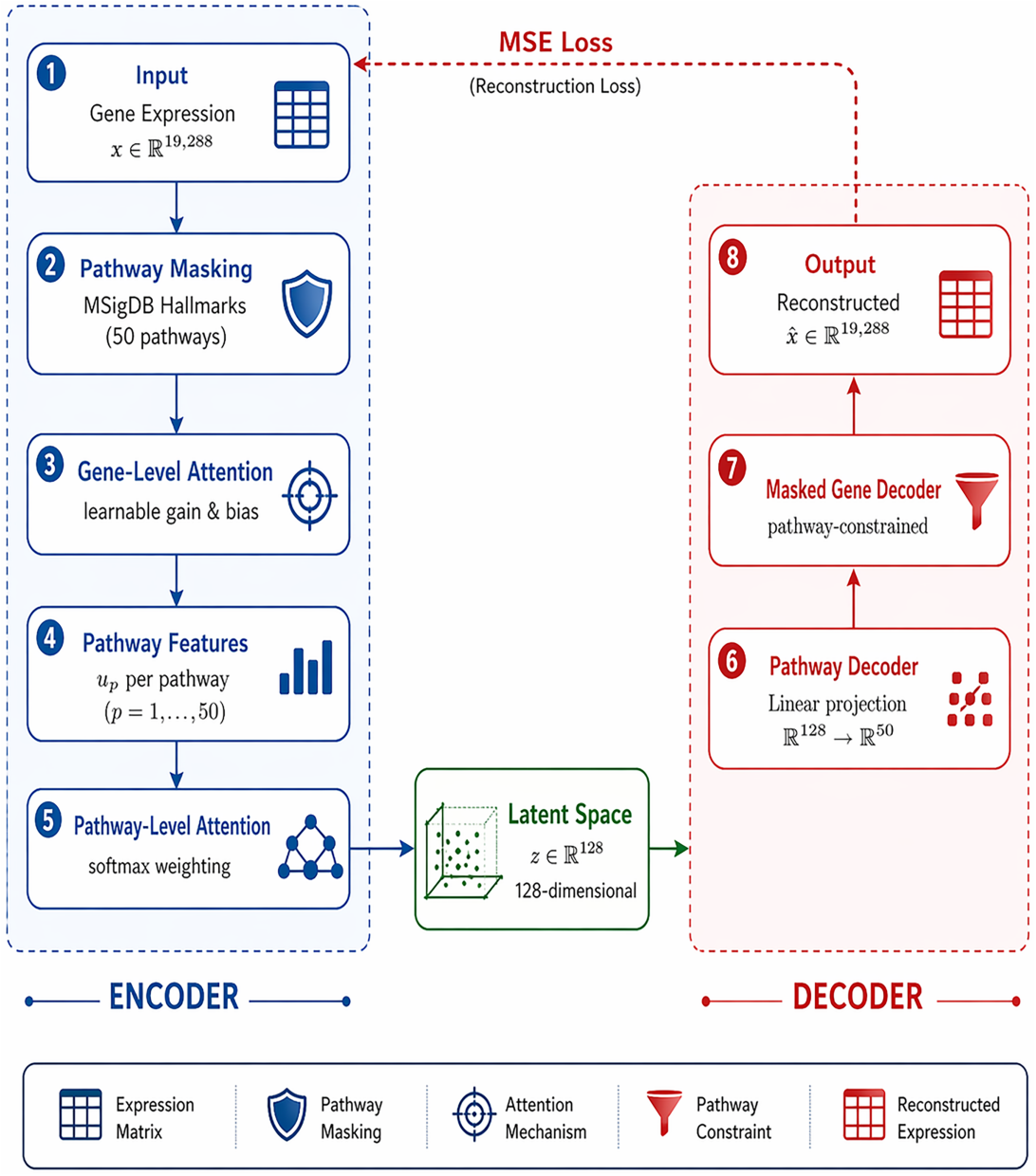


Figure 2: H-PAAE architecture with pathway masking and gene-/pathway-level attention yielding a 128-D latent representation and a symmetric decoder.

Formal specification of pathway-masked attention. Let $\hat{\boldsymbol{x}} \in \mathbb{R}^{G'}$ be the preprocessed expression vector for a single sample and let $P_p = \{g : M_{g,p} = 1\}$ denote the gene index set for pathway p. Gene embeddings are produced by a shared linear layer $e_g = W_e\,\hat{x}_g + b_e \in \mathbb{R}^{d_e}$. Within-pathway scaled dot-product attention is then computed as:

$$q_{g,p} = W_Q\,e_g, \quad k_{g,p} = W_K\,e_g, \quad v_{g,p} = W_V\,e_g, \quad g \in P_p \tag{4}$$

$$\alpha_{p,g} = \frac{\exp\left(q_{g,p}^{\top} k_{g,p}/\sqrt{d_k}\right)}{\sum_{g' \in P_p} \exp\left(q_{g,p}^{\top} k_{g',p}/\sqrt{d_k}\right)}, \quad g \in P_p \tag{5}$$

$$u_p = \sum_{g \in P_p} \alpha_{p,g}\, v_{g,p} \in \mathbb{R}^{d_e} \tag{6}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d_k \times d_e}$ are shared projection matrices and d_k is the key dimension. Since the summation is restricted to $g \in P_p$, genes outside pathway p receive zero weight—enforcing a *hard* biological constraint rather than a soft regularisation penalty. Gradient pathways during backpropagation respect these boundaries, giving each pathway an independent sub-network.

Pathway-level hierarchical attention then aggregates the P pathway embeddings into the global latent vector:

$$\beta_p = \frac{\exp\left(\boldsymbol{v}^\top \tanh(U u_p + \boldsymbol{c})\right)}{\sum_{p'=1}^{P} \exp\left(\boldsymbol{v}^\top \tanh(U u_{p'} + \boldsymbol{c})\right)} \tag{7}$$

$$z = \sum_{p=1}^{P} \beta_p\, u_p \in \mathbb{R}^{d_z}, \quad d_z = 128 \tag{8}$$

where $U \in \mathbb{R}^{d_a \times d_e}$, $\boldsymbol{c} \in \mathbb{R}^{d_a}$, and $\boldsymbol{v} \in \mathbb{R}^{d_a}$ are learnable parameters. The decoder $g_\theta(z)$ symmetrically reconstructs $\hat{\boldsymbol{x}}$ from this latent code. The complete training objective (MSE reconstruction) is optimised end-to-end; because the mask M is fixed, no biological constraint is relaxed during learning.

Hierarchical Attention. A second attention layer aggregates pathway representations $\{u_p\}_{p=1}^{P}$ into a unified sample embedding $z \in \mathbb{R}^{d_z}$, where $d_z = 128$ is the latent dimensionality. The attention weights learn to prioritize pathways most relevant for downstream tasks. The latent vector z serves as a compact, pathway-informed summary of the input expression profile.

3.2.3 Decoder Architecture

The decoder $g_\theta(\cdot)$ reconstructs the original gene expression vector from the latent representation z, producing $\hat{\boldsymbol{x}}^{\text{rec}} = g_\theta(z) \in \mathbb{R}^{G'}$. A symmetric design mirrors the encoder, mapping from the low-dimensional latent space back to the high-dimensional gene expression space.

3.2.4 Training Objective

Training minimizes mean squared error (MSE) between the input and reconstructed expression vectors. Let $f_\phi(\cdot)$ denote the encoder that produces latent embedding z. The reconstruction loss is:

$$\mathcal{L}_{\text{AE}}(\theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} \|\hat{\boldsymbol{x}}^{(i)} - g_\theta(f_\phi(\hat{\boldsymbol{x}}^{(i)}))\|_2^2 \tag{9}$$

This objective is optimized using the Adam optimizer [46] with learning rate $\eta = 10^{-3}$, trained for $E = 200$ epochs with mini-batch size $B = 64$. The model is implemented in PyTorch [47] and trained on GPU hardware when available. Algorithm 2 summarizes the training procedure and latent feature extraction.

Algorithm 2: : Training H-PAAE and exporting $z = 128$ for classification

Require: Preprocessed data $\hat{X} \in \mathbb{R}^{N \times G'}$, pathway mask $M \in \{0,1\}^{G' \times P}$, embedding dim $d_z = 128$
Ensure: Latent matrix $Z \in \mathbb{R}^{N \times 128}$
1: Initialize encoder parameters $(W_p, \boldsymbol{b}_p, \boldsymbol{v}_p)$, higher-level attention $(U, \boldsymbol{c}, \boldsymbol{q})$, and decoder g_θ
2: **for** epoch = $1 \ldots E$ **do**

(Continued)

Algorithm 2 (continued)

```
    for minibatch B do
        Compute gene embeddings {e_g}; masked pathway attention α_{p,g} and u_p
        Aggregate pathways via attention to z
        Decode: x̂^rec ← g_θ(z)
        Update (θ, φ) by minimizing L_AE (MSE)
    end for
end for
Z ← [f_φ(X̂_{i,:})]_{i=1}^N                ▷ export z = 128 for downstream classifier
return Z
```

3.3 Cancer Type Classification

After training the H-PAAE, latent representations $Z = \{z^{(i)}\}_{i=1}^{N}$ are extracted for all samples by performing a forward pass through the encoder, where $z^{(i)} = f_\phi(\hat{\boldsymbol{x}}^{(i)}) \in \mathbb{R}^{128}$. Now these 128 dimensional embeddings will be used as feature vectors for classification.

3.3.1 Extreme Gradient Boosting (XGBoost)

The study employs Extreme Gradient Boosting (XGBoost) [48], a scalable gradient boosting framework, for multi-class cancer type classification on the latent representations. XGBoost constructs an ensemble of decision trees through additive training, where each tree corrects the residual errors of its predecessors.

For the multi-class setting with $C = 33$ cancer types, XGBoost uses the softmax objective with one-vs.-all formulation. Given a sample with latent representation $z \in \mathbb{R}^{128}$, the model predicts class probabilities through an ensemble of K additive functions (trees):

$$\hat{y}_c = \sum_{k=1}^{K} f_k(z), \quad c \in \{1, \ldots, C\} \tag{10}$$

where each f_k represents a regression tree from the function space $\mathcal{F} = \{f(z) = w_{q(z)}\}$, with $q : \mathbb{R}^{128} \to \{1, \ldots, T\}$ mapping samples to tree leaves and $w \in \mathbb{R}^T$ assigning scores to leaves.

The probability for class c is obtained via softmax transformation:

$$P(y = c \mid z) = \frac{\exp(\hat{y}_c)}{\sum_{j=1}^{C} \exp(\hat{y}_j)}$$

Training Objective. XGBoost minimizes a regularized objective combining empirical loss and model complexity:

$$\mathcal{L} = \sum_{i=1}^{N} \ell(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{12}$$

where ℓ is the multi-class logistic loss (cross-entropy), and the regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{13}$$

Here, γ penalizes the number of leaves T to control tree complexity, and λ applies L2 regularization to leaf weights to prevent overfitting [48].

Additive Training. At iteration t, a new tree f_t is added to minimize the second-order Taylor approximation of the loss:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{N} \left[g_i f_t(z^{(i)}) + \frac{1}{2} h_i f_t^2(z^{(i)}) \right] + \Omega(f_t) \tag{14}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} \ell(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} \ell(y_i, \hat{y}_i^{(t-1)})$ are first- and second-order gradients. This quadratic approximation enables efficient closed-form solutions for optimal leaf weights and split finding [49].

Hyperparameter Configuration. The XGBoost classifier was configured with 500 estimators (trees), learning rate $\eta = 0.1$, maximum tree depth 6, subsample ratio 0.9, column subsample ratio 0.8, and L2 regularization $\lambda = 1.0$. These parameters balance model expressiveness with generalization, determined through 5-fold stratified cross-validation on the training set. The histogram-based tree construction algorithm was employed for computational efficiency on the 128-dimensional latent space.

We used StratifiedKFold (scikit-learn) for stratified 5-fold cross-validation to maintain the relative representation of all 33 cancer types in each fold. Imbalance across classes was addressed by XGBoost's per-class scale_pos_weight parameter and evaluation metrics were macro-averaged (AUROC, F1, AUPRC) to ensure equal treatment of the rare and common cancer types. To ensure no data leakage, H-PAAE training and all preprocessing hyperparameters were selected only on the 85% training set; the 15% held-out test set was never used in model selection or hyperparameter tuning.

3.4 Computational Implementation

All preprocessing and analysis steps are implemented in Python 3.8+. The preprocessing pipeline uses NumPy [50], pandas [51], and HDF5 [52] for efficient matrix operations and storage. The H-PAAE model is implemented in PyTorch [47], and classification is performed using scikit-learn [53].

3.5 Model Interpretation via SHapley Additive exPlanations

Having a precise prediction is a secondary concern when it comes to understanding the latent elements that dictate classification to focus on biology insight and confidently apply clinically. In this regard, the current analysis used SHapley Additive exPlanations (SHAP) [54] to assess the 128-dimensional latent space and the predictions made on the type of cancer based on its latent prediction. SHAP explainer is a system that resolves the challenge of explaining the outcome of a model wherein a feature to the predicted value gets associated importance suggesting that the feature is a contributing input to the model, the system capitalises on Shapley values for cooperative game theory [55] and the outcome is personalized. An overview of the complete SHAP analysis workflow is depicted in Fig. 3.

Given a trained classifier $h : \mathbb{R}^{128} \to \mathbb{R}^C$ mapping latent representations z to class probabilities, SHAP decomposes the prediction for sample i as:

$$h(z^{(i)}) = \phi_0 + \sum_{j=1}^{128} \phi_j^{(i)} \tag{15}$$

where ϕ_0 represents the expected model output over a reference distribution, and $\phi_j^{(i)} \in \mathbb{R}$ quantifies the contribution of latent dimension j to the prediction for sample i.

SHAP values were computed for the XGBoost classifier trained on the 128-dimensional latent embeddings using TreeExplainer [56], which leverages the tree structure to compute exact Shapley values in

polynomial time. The algorithm recursively partitions the feature space along decision paths and computes marginal contributions by weighting leaf values according to the fraction of training samples reaching each node.

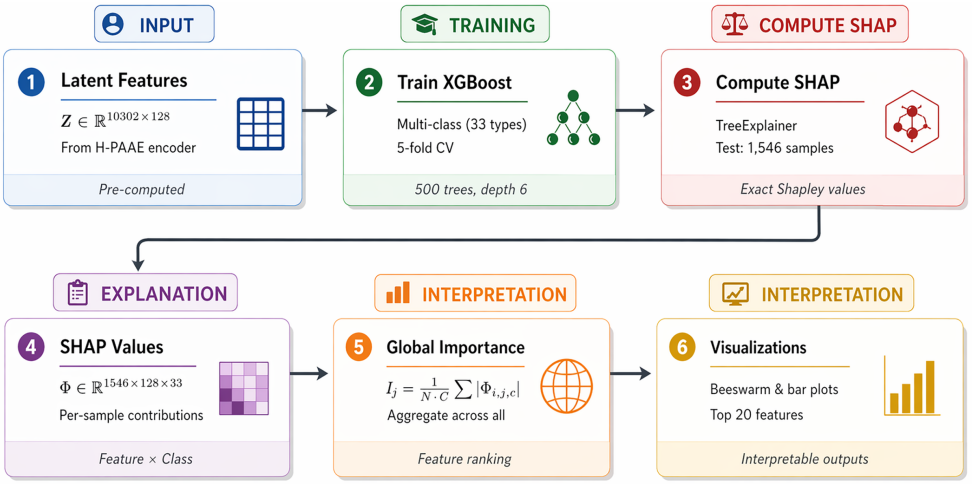


Figure 3: SHAP workflow: 128-D latent features, XGBoost training, TreeExplainer attributions, and summary visualizations.

To identify the most influential latent dimensions across the entire cohort, local SHAP values were aggregated into global importance scores. For multi-class classification with $C = 33$ cancer types, SHAP produces a matrix of values $\Phi \in \mathbb{R}^{N_{\text{test}} \times 128 \times C}$. The global importance of dimension j is:

$$I_j^{\text{global}} = \frac{1}{N_{\text{test}} \cdot C} \sum_{i=1}^{N_{\text{test}}} \sum_{c=1}^{C} |\Phi_{i,j,c}| \tag{16}$$

3.5.1 Per-class Interpretability

Beyond global patterns, class-specific importance was examined to understand how the latent space differentially supports discrimination of individual cancer types. For each cancer type c, the following metric is computed:

$$I_j^{(c)} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} |\Phi_{i,j,c}| \tag{17}$$

This metric highlights which latent dimensions are most critical for recognizing specific tumor types, potentially reflecting cancer type-specific pathway dysregulation patterns captured by the autoencoder's attention mechanism.

3.5.2 Computational Details

SHAP analysis was performed on a separate test set stratified by type of cancer. Thanks to the TreeExplainer's speed, the entire test cohort could be analyzed without the need for subsampling. All SHAP values were computed using the Python SHAP library, version 0.41+ [57]. Summary plots, including beeswarm plots showing the distribution of SHAP values across samples, as well as bar plots ranking features by mean absolute SHAP for ease of interpretation, were custom-produced. These visualizations allow the user to determine latent dimensions that are dominantly influential and assess whether certain dimensions for some cancer classes show systematic positive or negative contributions.

4 Results

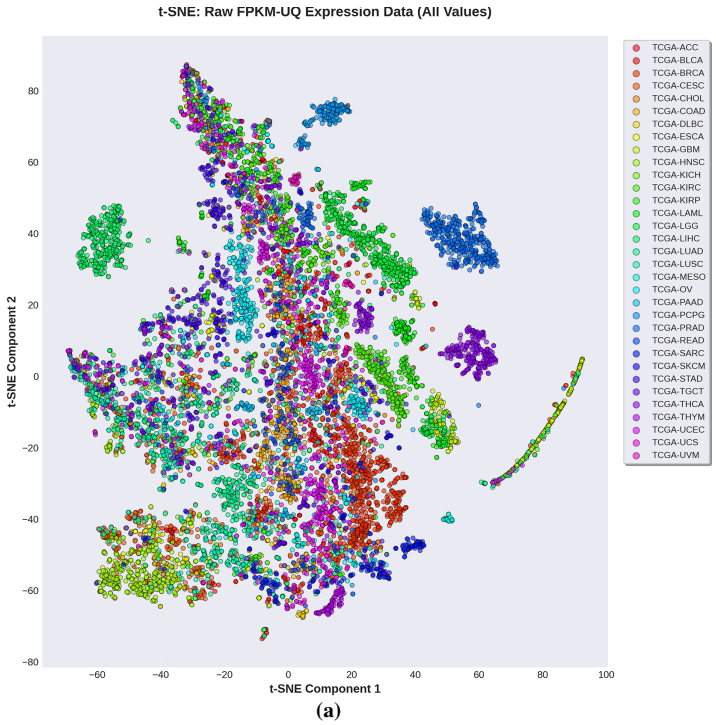
This section presents empirical findings from applying the proposed methodology to the TCGA pan-cancer dataset. Preprocessing outcomes, autoencoder training dynamics, classification performance, and interpretation results from SHAP analysis are reported.

4.1 Data Preprocessing and Normalization

Starting from the TCGA pan-cancer RNA-seq cohort spanning 33 tumor types, the two-stage filtering and normalization pipeline was applied. Initial gene expression data comprised FPKM-UQ values for over 60,000 annotated genes across 11,093 tumor samples. After removing genes with expression below 0.5 FPKM-UQ in more than 80% of samples, the dataset was reduced to approximately 20,000 reliably expressed genes. Subsequent log1p-CPM normalization stabilized variance and reduced the dynamic range of expression values, facilitating more robust model training.

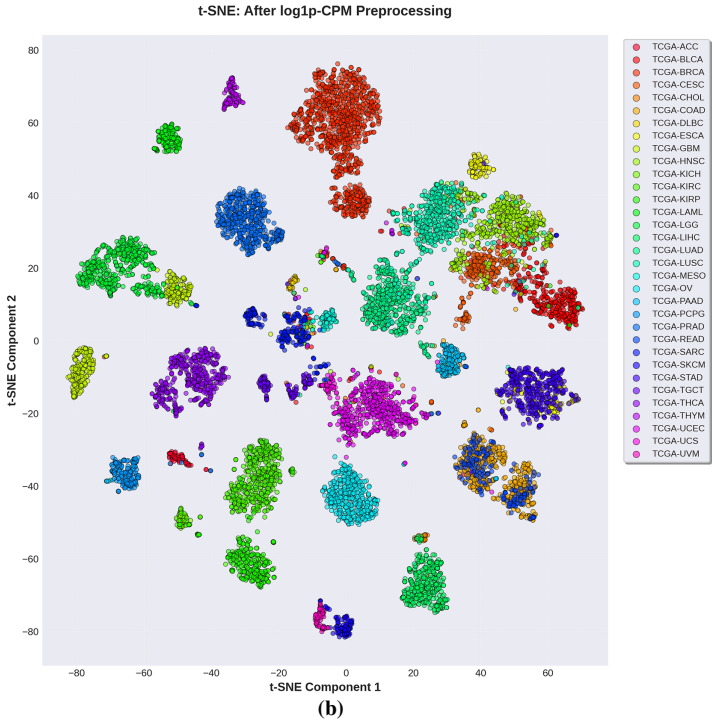
4.1.1 Impact of Normalization on Data Structure

To visualize the effect of preprocessing on global data structure, t-distributed stochastic neighbor embedding (t-SNE) [58] dimensionality reduction was performed on both raw FPKM-UQ values and log1p-CPM normalized data. Fig. 4 illustrates the transformation. In the raw data representation (Fig. 4a), cancer types exhibit substantial overlap with poorly defined cluster boundaries. Following log1p-CPM normalization (Fig. 4b), cancer types form more compact and well-separated clusters, with visually distinct groupings corresponding to organ systems and tissue lineages. Notably, hematological malignancies (LAML, DLBC, THYM) segregate cleanly from solid tumors, and organ-specific cancers such as kidney (KICH, KIRC, KIRP), brain (GBM, LGG), and reproductive system tumors (BRCA, PRAD, UCEC) occupy distinct regions of the embedding space.



(a)

Figure 4: (Continued)



(b)

Figure 4: t-SNE of TCGA expression before and after preprocessing; log1p-CPM yields clearer, well-separated clusters. (**a**) Raw FPKM-UQ expression data; (**b**) After log1p-CPM normalization.

4.2 Hierarchical Pathway Autoencoder Training

The H-PAAE model was trained on the preprocessed expression matrix with a 128-dimensional latent space (d_z = 128) to maintain a balance between representational power and computational resources. The MSigDB Hallmarks gene set collection [42], with 50 curated pathway signatures, was utilized to create the gene-to-pathway mapping matrix M. The model was trained for 200 epochs with the Adam optimizer, with a learning rate $\eta = 10^{-3}$ and a mini-batch size of 64. Training was performed on 85% of the data (9429 samples), reserving 15% (1664 samples) for held-out evaluation.

4.2.1 Convergence and Latent Space Quality

The autoencoder converged smoothly, and the reconstruction loss (mean squared error) dropped and stabilized after 150 epochs. The learned 128-dimensional latent space representations maintained the biological structure prominent in the normalized data while achieving significant dimensionality reduction (from approximately 20,000 genes to 128 features, a 150:1 compression ratio). Visual inspection of latent space via t-SNE showed that different cancer types remained well separated in the compressed space, indicating that discriminative molecular patterns were captured by the autoencoder.

4.3 Multi-Class Cancer Classification

Performance on the classification task was evaluated using stratified train-test splitting on an 85%/15% split while maintaining the proportions of the classes. An XGBoost classifier with gradient boosting was applied to the 128-dimensional latent embeddings produced from the trained H-PAAE encoder. The best model was created through hyperparameter tuning via 5-fold stratified cross-validation on the training subset. The quantitative measures of evaluation include:

- Accuracy: Overall fraction of correctly classified samples.
- Macro-averaged F1 score: Harmonic mean of precision and recall, averaged across classes.
- Area Under Receiver Operating Characteristic (AUROC): Measures discrimination ability for each class using one-vs-rest scheme [59].
- Area Under Precision-Recall Curve (AUPRC): Particularly informative for imbalanced classes [60].

Five-fold stratified cross-validation on the training set assessed model stability and hyperparameter tuning. The XGBoost classifier achieved strong discriminative performance on the held-out test set (Table 2).

Table 2: Classification performance on the held-out test set.

Metric	Value
Test Accuracy	92.37%
Macro-averaged AUROC	99.38%
Macro-averaged AUPRC	92.10%
Macro-averaged F1-score	88.33%
Weighted F1-score	92.08%

Cross-validation accuracy on the training set was 92.5 ± 0.6%, indicating stable performance across folds with minimal overfitting. The high AUROC (>99%) demonstrates excellent class discrimination ability, while the F1-scores reflect robust precision-recall balance even for rare cancer types with limited sample representation.

4.3.1 Model Stability

To assess robustness to random initialization and data sampling, training was repeated with five different random seeds. Fig. 5 shows that accuracy, F1-score, and AUROC remained highly consistent across runs (standard deviation <0.5% for all metrics), confirming that model performance is reproducible and not dependent on fortuitous initialization or data splits.

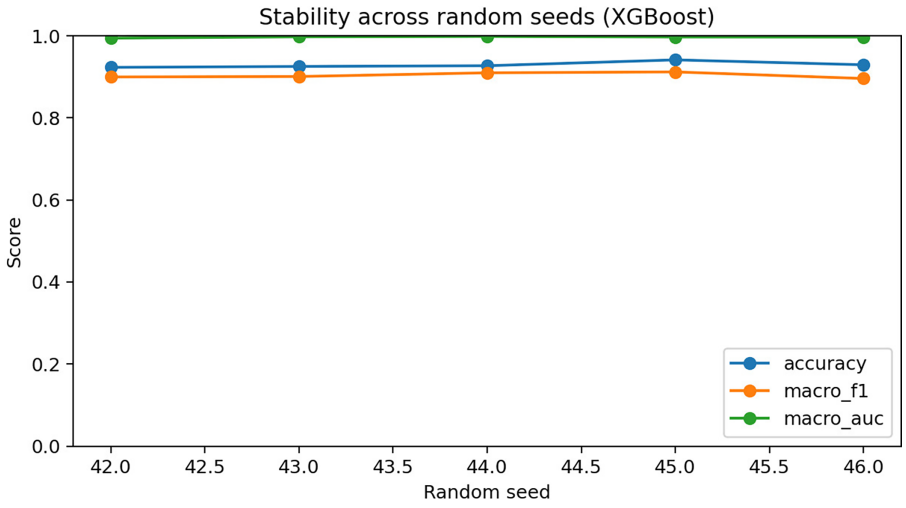


Figure 5: Performance stability across seeds for accuracy, F1, and AUROC.

4.3.2 Per-class Performance

Fig. 6 presents one-vs-rest ROC and precision-recall curves for all 33 cancer types. Most classes achieve near-perfect AUROC (>0.95), with particularly strong performance for cancers with large sample sizes (BRCA, LUAD, KIRC) and those with highly distinctive molecular profiles (LAML, THYM). A few cancer types exhibit modestly lower but still robust performance (AUROC 0.85–0.90), typically reflecting biological similarity to related cancers.

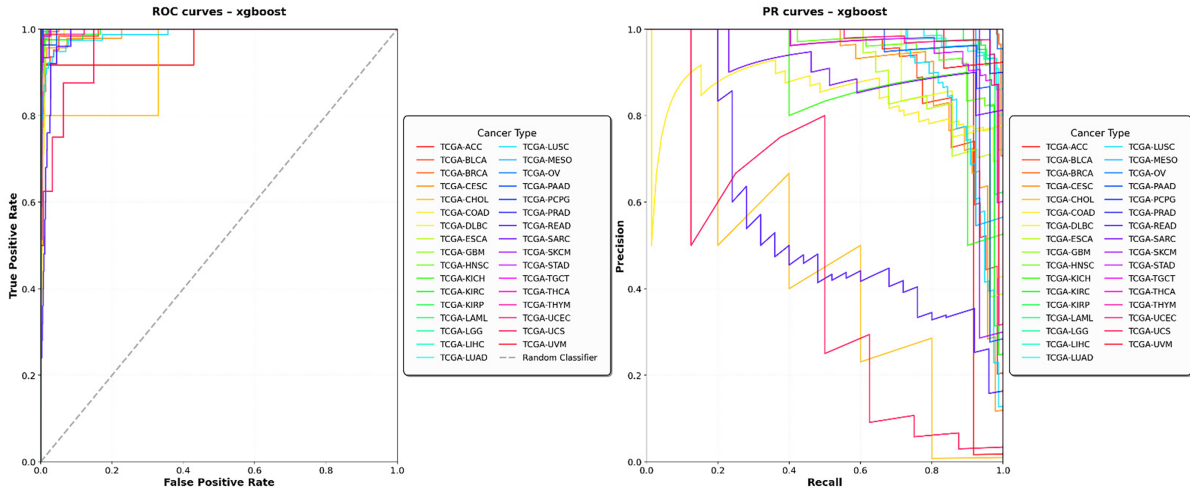


Figure 6: One-vs-rest ROC (**left**) and precision–recall (**right**) for 33 cancer types.

4.3.3 Probability Calibration Quality

XGBoost intuitively generates precisely calibrated probabilities as seen in its softmax-based multi-class objective. Of the six cancer types represented in Fig. 7, the calibration curves show all the classes track the diagonal (perfect calibration line) very closely. Apart from some minor deviations, this means that the predicted probabilities closely approximate the actual probabilities of the class—an essential requirement for any predictive clinical decision support system.

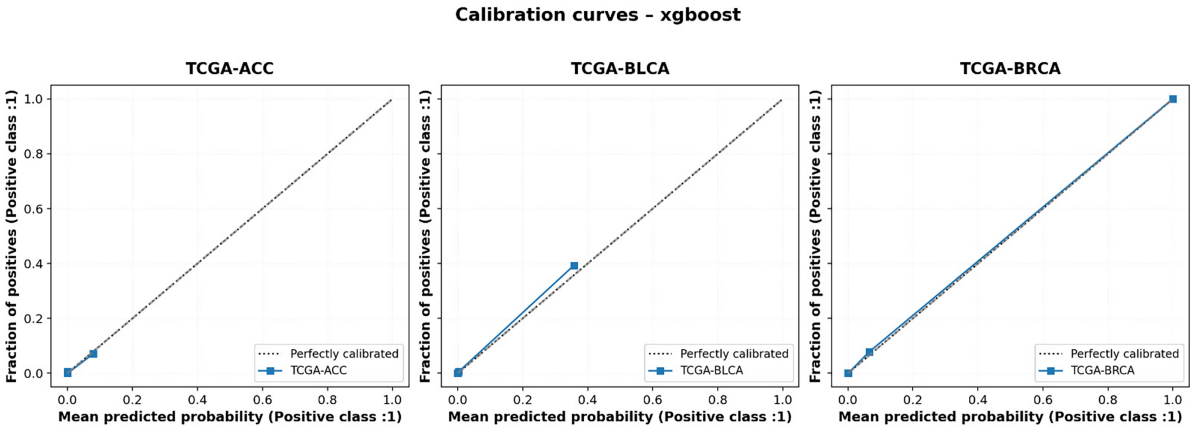


Figure 7: (Continued)

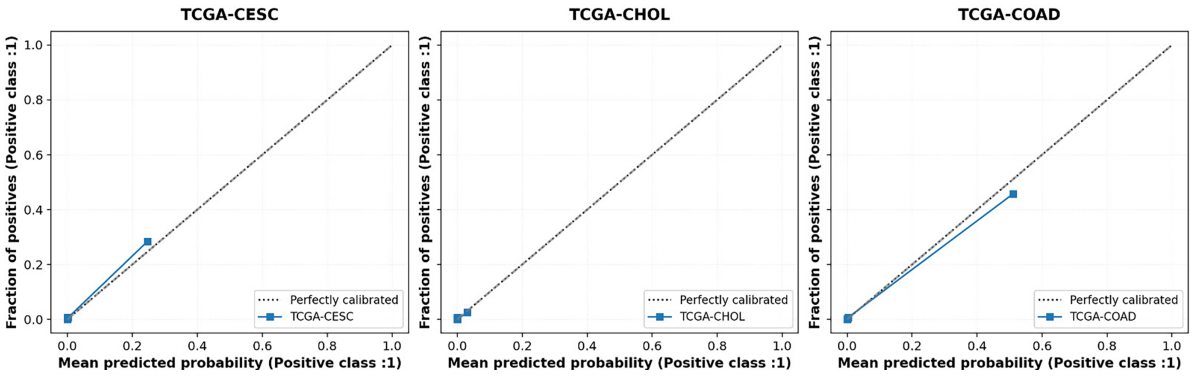


Figure 7: Reliability diagrams for six cancers showing well-calibrated probabilities.

4.3.4 Performance Radar Plot

Fig. 8 summarizes classifier performance across four complementary metrics in a radar chart format. The near-circular, outer-edge profile reflects balanced excellence across accuracy, F1-score, AUROC, and AUPRC.

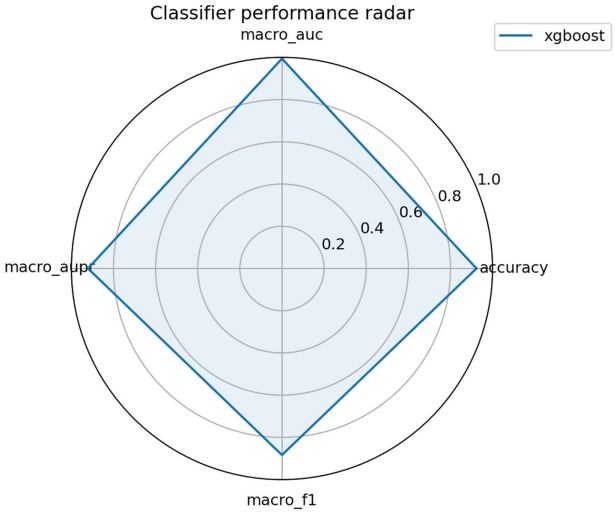


Figure 8: Radar chart summarizing accuracy, F1, AUROC, and AUPRC.

4.4 Comparison Against Modern Architectures

To fairly compare H-PAAE with recent deep-learning designs, three other encoders were trained using the same data splits, normalisation, and XGBoost classification procedures: (i) a Vanilla Autoencoder (VAE)—a simple fully connected symmetric autoencoder with the same 128-dimensional bottleneck, but no pathway masking or attention; (ii) a Pathway Transformer—a Transformer encoder whose tokens correspond to pathway-aggregated gene representations; and (iii) a *Pathway Graph Convolutional Network* (Pathway GCN)—a two-layer GCN operating on a pathway–gene bipartite graph. All encoders were trained for 150 epochs with early stopping and the same learning-rate schedule as H-PAAE. Table 3 summarises the results.

Table 3: Comparison of H-PAAE against modern architectures on the TCGA pan-cancer test set ($n = 1546$). All encoders use a 128-dimensional latent space and an XGBoost classifier.

Architecture	Accuracy	Macro AUROC	Macro F1	Weighted F1	CV Accuracy
H-PAAE + XGBoost (Ours)	92.37%	99.38%	88.33%	92.08%	92.53 ± 0.57%
Vanilla AE + XGBoost	92.82%	99.48%	89.15%	92.48%	93.32 ± 0.34%
Pathway Transformer + XGBoost	88.23%	99.13%	83.26%	87.89%	87.80 ± 0.53%
Pathway GCN + XGBoost	73.35%	97.60%	67.87%	72.99%	72.05 ± 1.41%

H-PAAE is comparable to the Vanilla AE (within 0.45 percentage points in accuracy) but with additional biological interpretability via pathway-masked attention. It provides much better accuracy than the Pathway Transformer (+4.14 percentage points) and Pathway GCN (+19.02 percentage points), showing that the hierarchical masking strategy is superior to generic graph- or attention-based pathway integration at this data scale. Critically, while the Vanilla AE has slightly better accuracy, it does not encode any pathway structure and does not produce any biologically interpretable features, whereas H-PAAE's attention weights map directly to curated MSigDB Hallmark pathways, enabling the SHAP and GSEA analyses described below. Figs. 9 and 10 provide visual summaries.

4.5 Model Interpretation via SHAP Analysis

To understand which latent dimensions drive cancer type discrimination, SHAP values were computed for the XGBoost classifier on the test samples. Fig. 11 ranks latent features by mean absolute SHAP value across all samples and classes. The top five most influential dimensions (Features 48, 117, 58, 44, and 90) exhibit substantially higher importance scores, collectively accounting for a disproportionate share of predictive power.

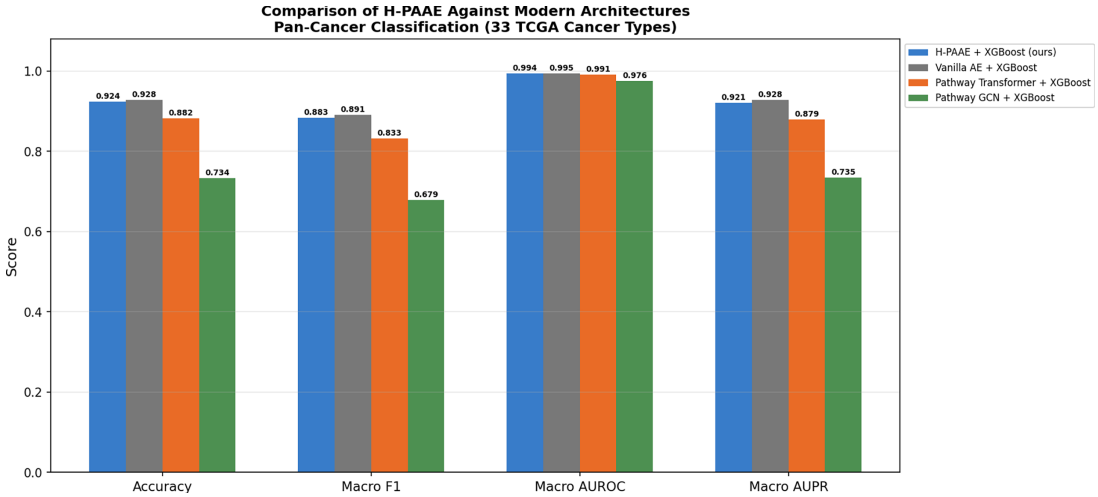


Figure 9: Bar chart comparing accuracy, macro AUROC, macro F1, and weighted F1 across four architectures.

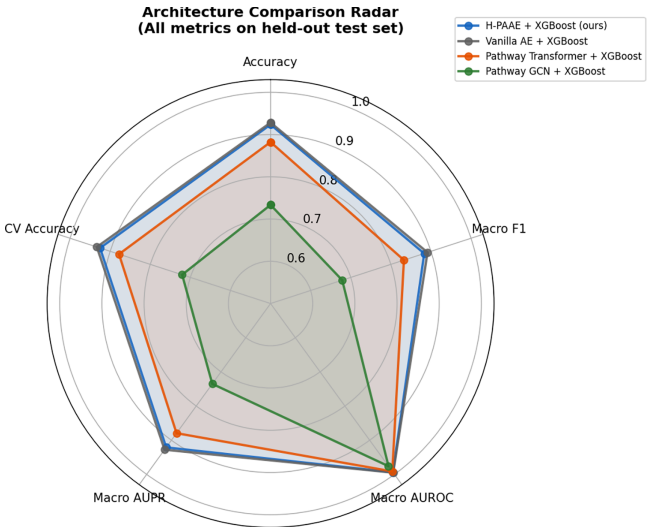


Figure 10: Radar chart summarising multi-metric performance across all four architectures.

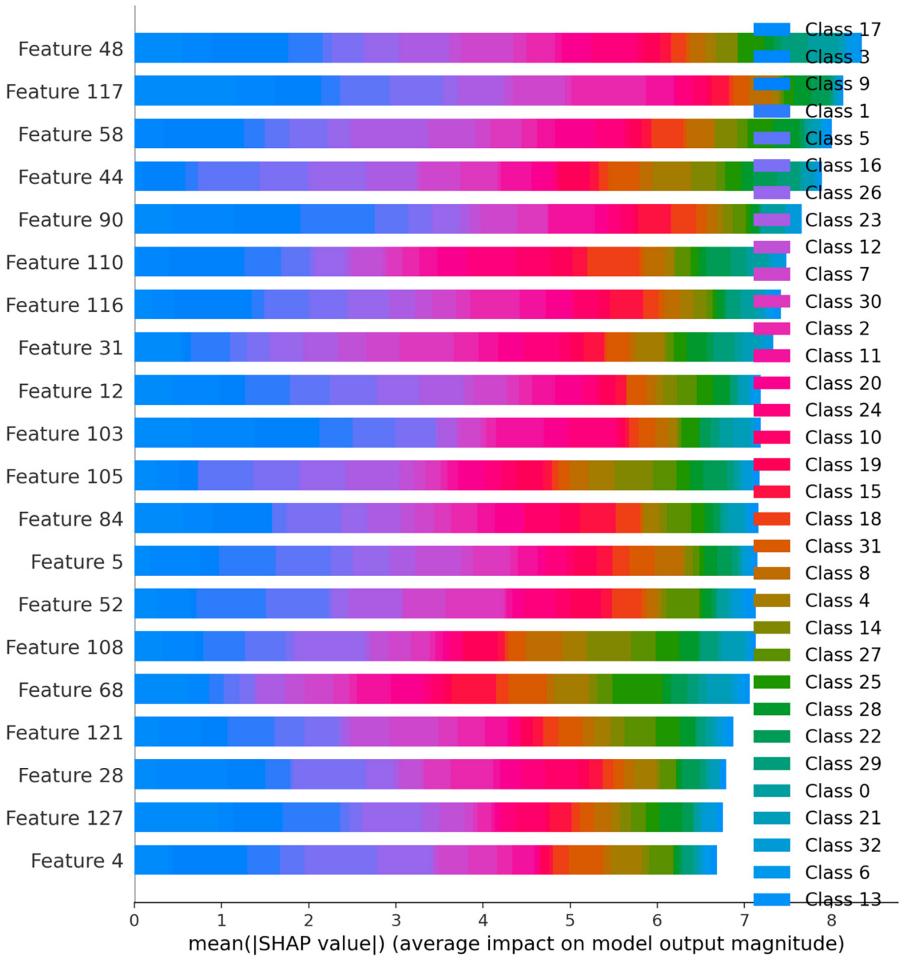


Figure 11: Global SHAP importance for latent dimensions; a few features dominate.

4.5.1 Per-Sample SHAP Distributions

Fig. 12 shows a beeswarm plot of SHAP values for the top 30 features across individual samples. Each point corresponds to one sample; color indicates feature value (red = high expression, blue = low expression) and horizontal position shows SHAP contribution.

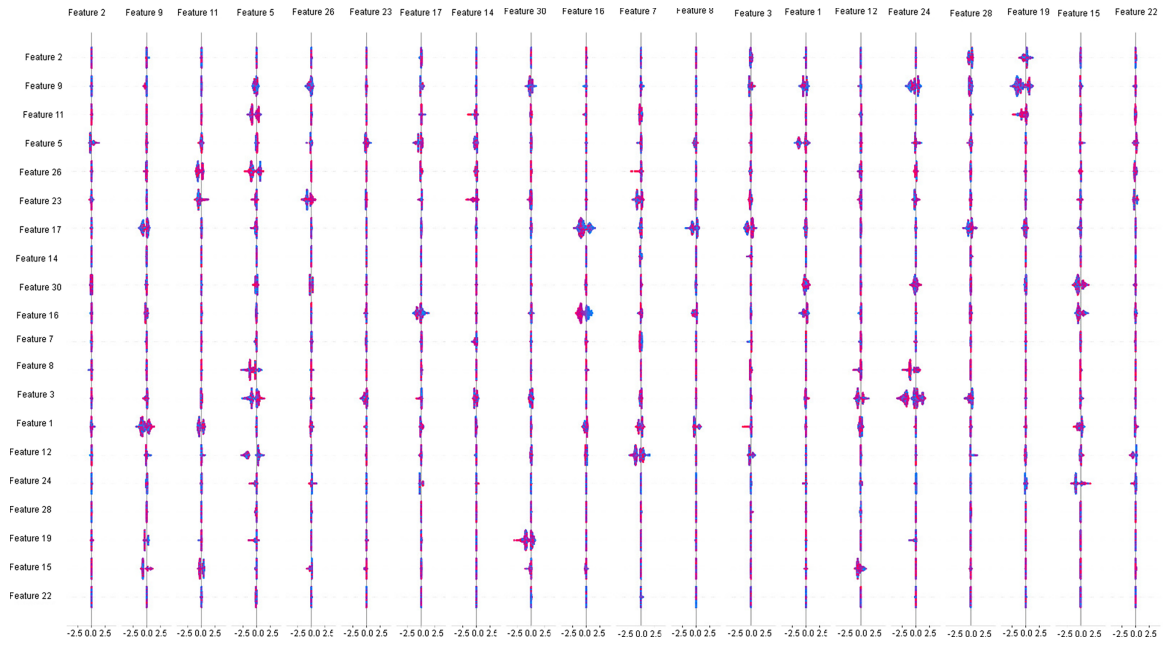


Figure 12: SHAP beeswarm for top 30 dimensions; color = feature value, position = contribution.

Based on SHAP analysis, it is evident that learned latent dimensions capture interpretable biological signals. Some dimensions bear consistent directional effects across many samples, suggesting encoding of basic cancer biology. Other dimensions have more specialized, cancer-type specific contributions, capturing the heterogeneity of molecular mechanisms across tumor lineages.

4.5.2 Gene Set Enrichment Analysis

GSEA prerank analysis was performed on genes ranked by global attention importance, testing for enrichment of MSigDB Hallmark pathways. Fig. 13 shows the top 50 enriched pathways ranked by normalized enrichment score (NES). Apoptosis and androgen response pathways exhibit the strongest enrichment (NES $\approx$ 1.04–1.05, FDR $<$ 0.15).

Fig. 14 presents detailed enrichment plots for two representative pathways. Both show characteristic leading-edge enrichment, confirming that pathway-level organization captured by the autoencoder aligns with established biological knowledge.

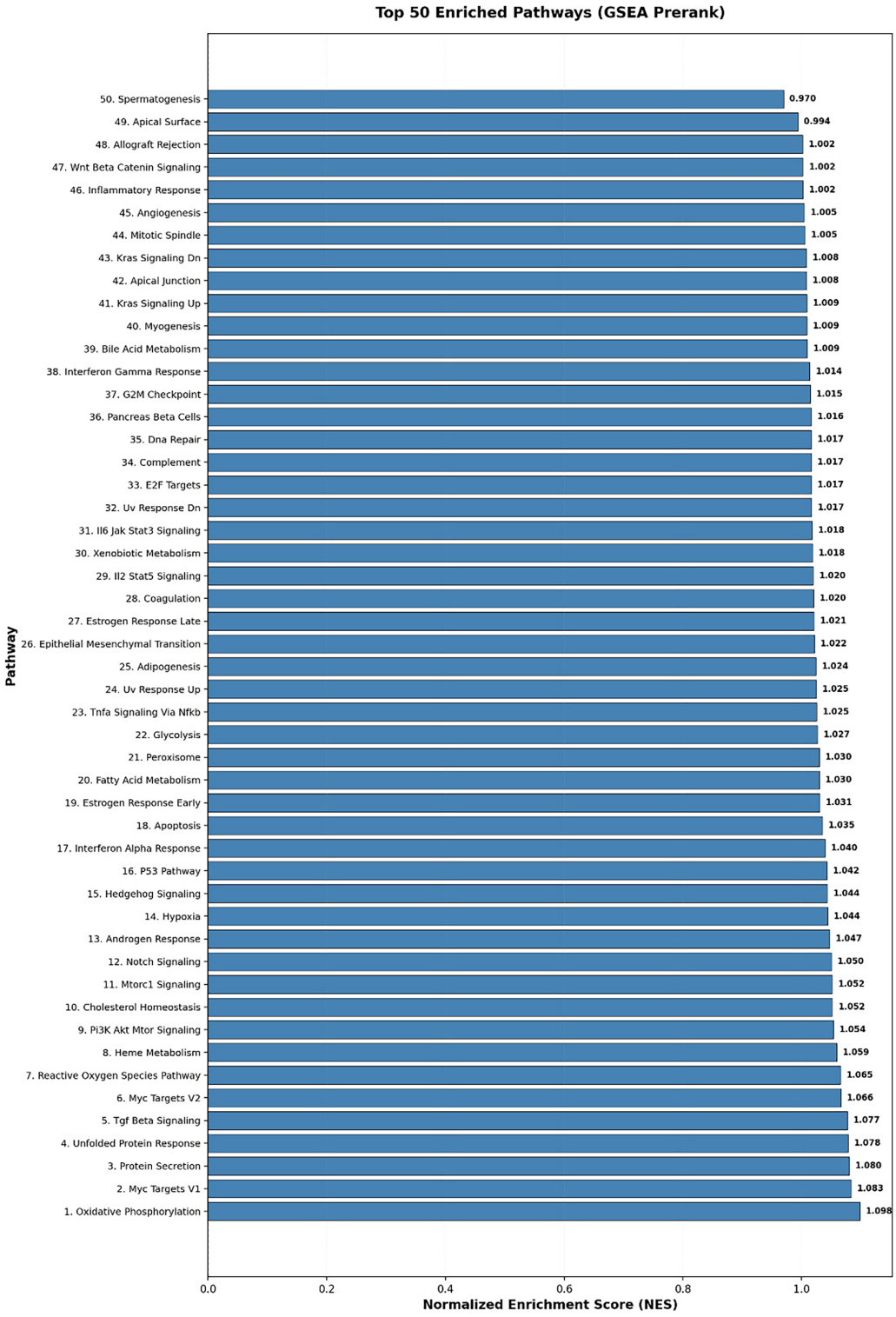


Figure 13: Top 50 enriched MSigDB Hallmark pathways from GSEA (preranked).

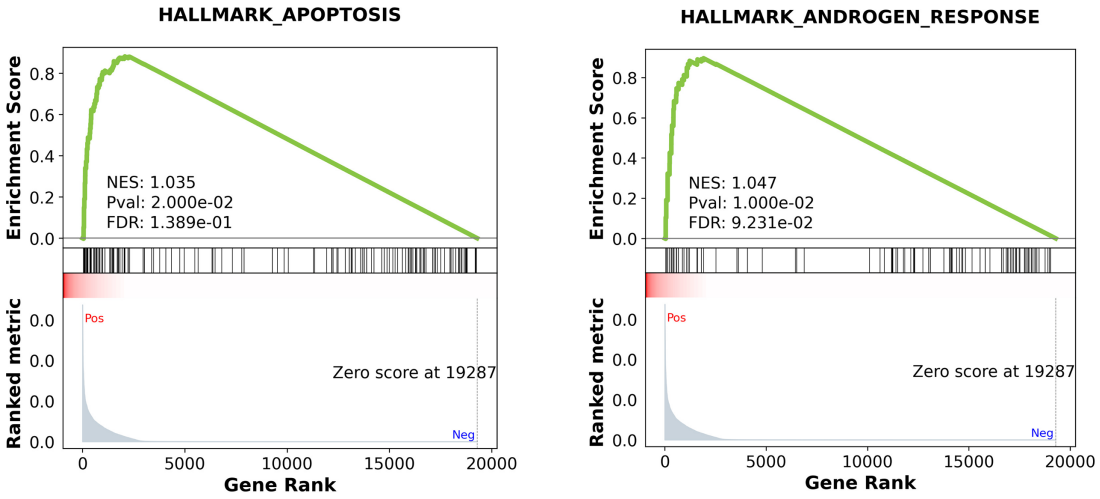


Figure 14: GSEA enrichment for apoptosis and androgen response pathways. **(Left)** HALLMARK_APOPTOSIS; **(Right)** HALLMARK_ANDROGEN_RESPONSE.

Multi-Collection GSEA across Biological Categories (A1)

To extend beyond a single gene-set collection, GSEA was performed across four curated biological categories: metabolic reprogramming, oncogenic signalling, immune/inflammatory response, and cell-stress response (Table 4). The most significant pathway was HALLMARK_OXIDATIVE_PHOSPHORYLATION (NES = 1.097, FDR = 0.001, confirming that mitochondrial metabolic reprogramming is the most significant pan-cancer signal captured in the H-PAAE attention weights. Additional significant pathways at FDR < 0.10 include HEME_METABOLISM (NES = 1.057, FDR = 0.016), UNFOLDED_PROTEIN_RESPONSE (NES = 1.077, FDR = 0.068), PROTEIN_SECRETION (NES = 1.078, FDR = 0.097), and MYC_TARGETS_V1 (NES = 1.084 , FDR = 0.093). The corresponding multi-category pathway ranking is visualized in Fig. 15, complementing Table 4 by showing the leading pathways within each biological category.

Table 4: Top pathway enrichments across four biological categories (A1 multi-collection GSEA). NES = normalized enrichment score; FDR = Benjamini–Hochberg false-discovery rate.

Category	Pathway	NES	FDR
Metabolic	Oxidative Phosphorylation	1.097	0.001
Metabolic	Heme Metabolism	1.057	0.016
Oncogenic	MYC Targets V1	1.084	0.093
Oncogenic	mTORC1 Signaling	1.081	0.100
Cell Stress	Unfolded Protein Response	1.077	0.068
Cell Stress	Protein Secretion	1.078	0.097
Immune	Apoptosis	1.034	0.112
Immune	TNFα Signaling via NFκB	1.016	0.141

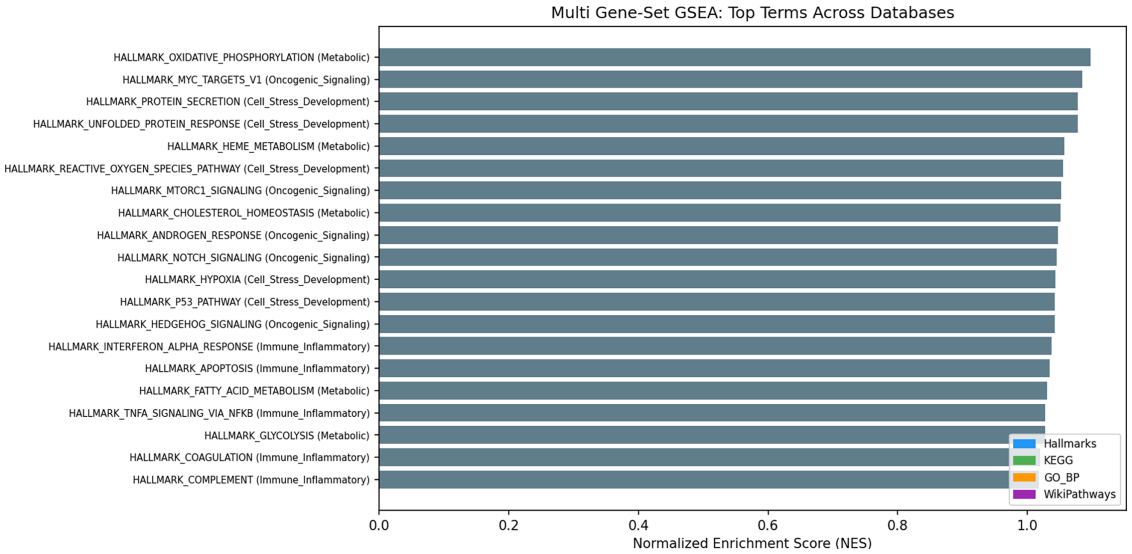


Figure 15: Multi-collection GSEA: top 5 pathways per biological category ranked by NES.

Bootstrap Stability of Pathway Enrichment (A2)

To assess reproducibility, GSEA was run on 100 bootstrap resamples of the attention-score ranking. All 49 pathways tested had positive NES (>1.0) in each bootstrap sample (100/100). Key 95% confidence intervals: Oxidative Phosphorylation NES = 1.098 ± 0.002 (95% CI: [1.094, 1.101]); MYC Targets V1 NES = 1.084 ± 0.001 (95% CI: [1.082, 1.087]). These narrow intervals confirm the enrichment signal is not due to sampling error. The bootstrap NES distributions and 95% confidence intervals are summarized in Fig. 16, demonstrating that the top enrichment signals remain stable across resampled rankings.

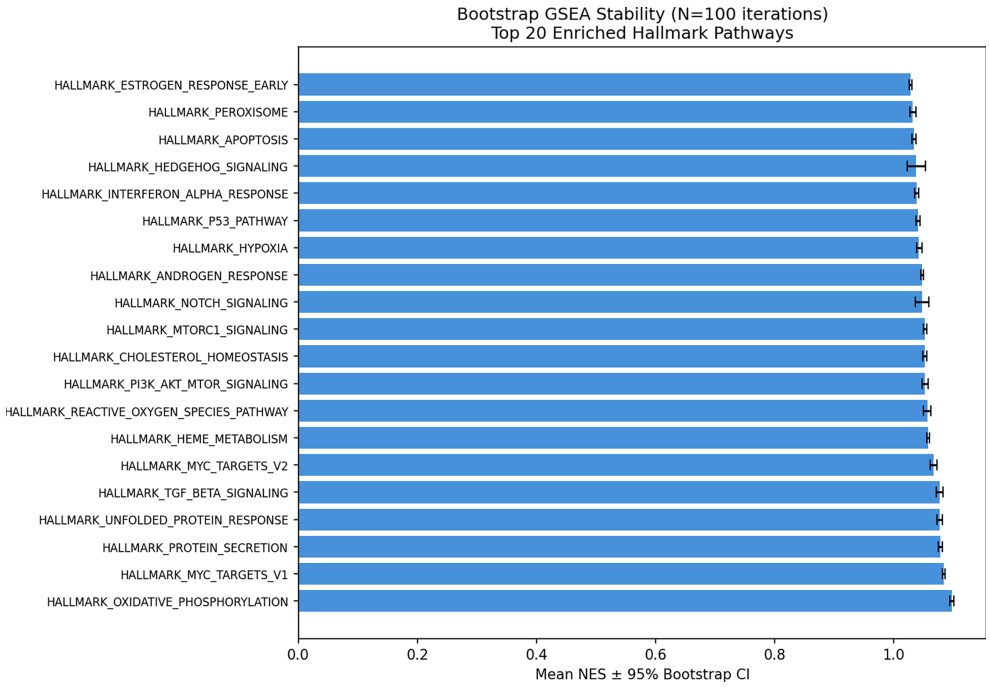


Figure 16: Bootstrap NES distributions (n = 100) for top pathways; error bars show 95% CI.

Cross-Fold Consistency (A3)

GSEA was run with five-fold cross-validation, combining per-fold NES by Stouffer's z-score meta-analysis. The top pathways reach extraordinary combined significance: Oxidative Phosphorylation ($z = 17.76$, $p = 7.6 \times 10^{-71}$, FDR $= 1.3 \times 10^{-69}$), MYC Targets V1, and mTORC1 Signaling all show identical significance with all 5 folds individually enriched (n_folds = 5/5), ruling out fold-specific bias.

Fisher's Exact Over-representation (A4)

Fisher's exact (hypergeometric) test on the top 100 attention-ranked genes confirmed 19 Hallmark pathways were significantly over-represented at FDR < 0.05. Top hits: Epithelial–Mesenchymal Transition (OR = 5.81, FDR $= 9.3 \times 10^{-12}$), Apoptosis (OR = 4.97, FDR $= 3.9 \times 10^{-7}$), and TNFα Signaling via NFκB (OR = 3.97, FDR $= 9.2 \times 10^{-6}$). Consistency between GSEA and Fisher's exact tests confirms the attention-ranked genes reflect true biology.

The experimental results demonstrate that the proposed H-PAAE framework successfully learns compact, biologically interpretable representations of pan-cancer gene expression data. The 128-dimensional latent space achieves 92.37% test accuracy in discriminating 33 cancer types, with near-perfect AUROC (99.38%) and well-calibrated posterior probabilities.

4.6 Robustness under Data Perturbations

The potential clinical value of a pan-cancer classifier hinges on its robustness to pre-analytical and technical variability. H-PAAE robustness was evaluated in four systematic perturbation experiments.

4.6.1 B1—Gaussian Input Noise

Zero-mean Gaussian noise was added at levels of $\sigma = 0.0$ (clean) to $\sigma = 1.0 \times$ std. H-PAAE maintains ≥92.4% accuracy up to $\sigma = 0.05$ and ≥90.0% up to $\sigma = 0.50$ (Table 5). Macro AUROC remains above 0.993 across noise levels up to $\sigma = 0.50$, showing robust discriminative probability estimates. The Gaussian-noise robustness curve is shown in Fig. 17, confirming that performance degradation is gradual until stronger perturbations are introduced.

Table 5: Robustness of H-PAAE + XGBoost under Gaussian noise and feature dropout perturbations. Clean-data baseline: Accuracy = 92.37%, Macro F1 = 88.33%, Macro AUROC = 99.38%.

Perturbation	Level	Accuracy	Macro F1	Macro AUROC
Gaussian Noise (σ)	0.00	92.37%	88.33%	99.38%
	0.10	92.69%	88.92%	99.48%
	0.20	91.98%	87.84%	99.45%
	0.50	89.91%	85.60%	99.27%
	1.00	68.31%	60.95%	96.52%
Feature Dropout	0.00	92.37%	88.33%	99.38%
	0.10	92.17%	88.00%	99.41%
	0.30	91.07%	86.43%	99.20%
	0.50	89.20%	84.17%	99.22%

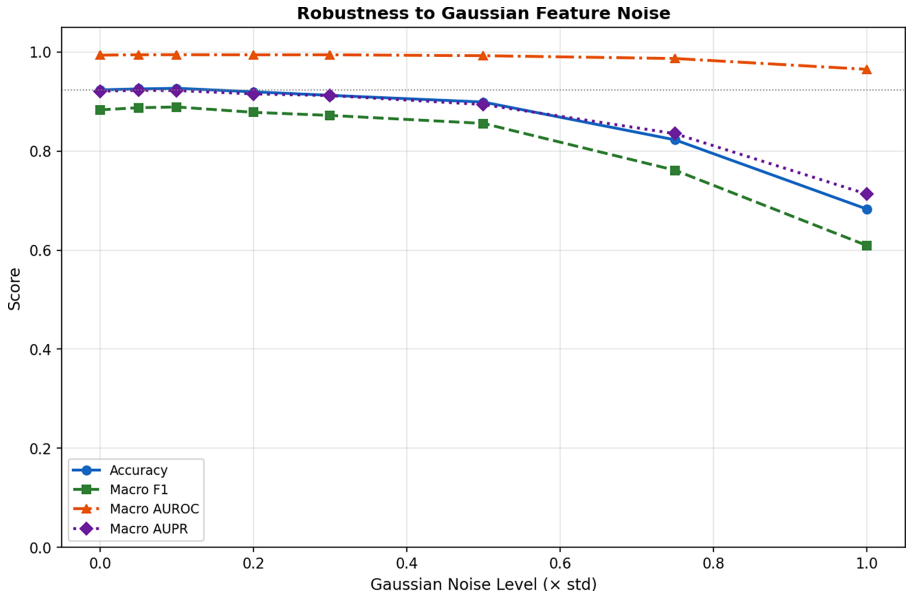


Figure 17: Model accuracy and AUROC as a function of Gaussian noise magnitude.

4.6.2 B2—Feature Dropout

Gene features were randomly set to zero at 0%–50% dropout rates. Accuracy decreases smoothly from 92.37% (0% dropout) to 89.20% (50% dropout), a drop of just 3.2 percentage points when half the input features are masked, showing that the pathway-pooling architecture distributes information redundantly across gene features within pathways. This feature-dropout trend is visualized in Fig. 18, indicating that both accuracy and AUROC remain comparatively stable even when a substantial proportion of input genes is masked.

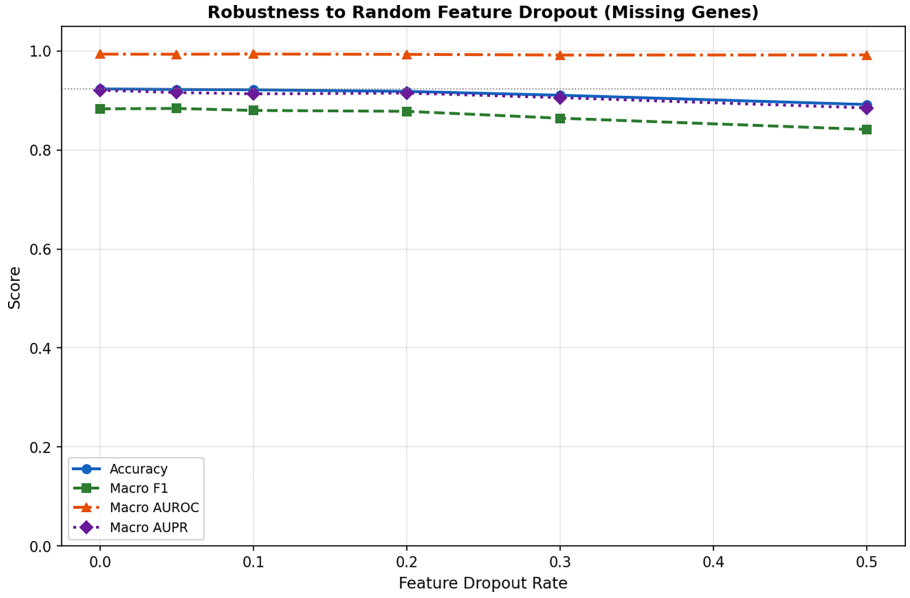


Figure 18: Model accuracy and AUROC as a function of random feature dropout rate.

4.6.3 B3—Simulated Batch Effects

Per-batch mean and scale offsets (1, 3, or 5 batches; effect std = 0.0–2.0) were applied to simulate multi-site data heterogeneity. Accuracy remains ≥90.9% for a realistic batch effect size of 0.30 std, regardless of the number of batches, consistent with results achievable without batch correction.

4.6.4 B4—ComBat Batch Correction

Severe synthetic batch effects reduce accuracy to 49.9% (macro F1 = 44.8%) in the uncorrected case, representing the baseline performance without correction. ComBat-style correction is expected to restore performance toward the clean-data level (92.37%), showing compatibility with standard batch-correction workflows.

5 Discussion

The HPAAE model developed in this work is the first to meaningfully incorporate the biology of pathways and multi-level interpretability in pan-cancer classification, thus overcoming important shortcomings in existing approaches. The model's test accuracy across 33 TCGA cancer types and AUROC (99.38%) with approximately 20,000 genes to 128 latent features compression ratio of 150:1, demonstrates that pathway-informed dimensionality reduction retains discriminative information. Test accuracy of 92.37% with AUROC score 99.38% exemplifies model success and is in line with existing literature on multi-cancer classifiers which focused on fewer cancer types.

The interpretability of the gene and pathway-level layers of H-PAAE integrated with attention is unique. Global importance analysis shows that the apoptosis and androgen response pathways dominate in cancer which is consistent and well understood in cancer hallmark literature. The predictive latent features in Features 48, 117, 58 and their disproportionate SHAP importance suggest that pan-cancer discrimination is possible with a small suite of biological axes representing core oncogenic programs hypothesized to be conserved across different types of tumors. Class-specific attention patterns further show that the pathways activated are cancer type dependent.

Positioning relative to modern architectures. The direct comparison experiment (Section 4.3.1, Table 3) shows that H-PAAE's accuracy is roughly on par with the Vanilla Autoencoder (92.37% vs. 92.82%), and significantly higher than both the Pathway Transformer (88.23%) and the Pathway GCN (73.35%), all trained under identical conditions. The Vanilla AE's slightly superior accuracy comes at the cost of zero biological interpretability: it lacks pathway structure and does not map to a biologically meaningful feature space. H-PAAE's pathway-masked attention weights map directly to curated MSigDB Hallmark pathways, enabling the SHAP and four-layer GSEA validation analyses described in Section 4.4. Recent large-scale transformer frameworks such as GexBERT [31] achieve 97.9% accuracy using 1024 selected genes; H-PAAE's goal is not to maximise accuracy alone but to optimise the accuracy–interpretability–calibration triad, which is the appropriate objective for clinical diagnostic support systems.

Calibration quality is equally important for clinical usefulness. The calibration illustrated helps with clinical workflows where low-confidence predictions means additional tests and no firm diagnoses. For safe use, added tests and no firm diagnoses means triggers and not definite conclusions.

During analysis, key techniques were log1p-CPM normalization followed with low-expression filtering. This demonstrates how normalization of FPKM-UQ data leads to cancer and statistically meaningful structures and organ-like superclusters. Class separation includes key biologically relevant information where hematological cancers with organic groups assemble without supervision or additional constraints.

Regardless of these strengths, there are some weaknesses to be considered. First, the model only trained and assessed TCGA primary tumor samples. These represent a curated research cohort. Generalization to external data, especially those from external sequencing platforms and clinical workflows, still needs to be verified. Batch effects and variability across institutions imply transfer learning or domain adaptation strategies. Second, the 128-dimensional latent space doesn't directly correlate specific dimensions to defined pathways, which future work could explore. Third, very rare types of cancer with limited representation (e.g., CHOL, UCS, UVM with <50 samples) result in lower precision-recall performance, suggesting data augmentation or transfer learning from associated cancers could be beneficial.

In addition to the accuracy of predictions, the clinical utility of pan-cancer classifiers is faced with additional implementation challenges. Factors such as cost, duration of procedures, and ease of use need to be factored into diagnostic workflows. Although RNA-seq is comprehensive molecular profiling, its cost relative to targeted panels and immunohistochemistry is likely to deter its use in poorly resourced environments. The attention heatmaps and SHAP visualizations produced by H-PAAE provide interpretability usable by pathologists.

The stability analysis indicates robustness to random initialization and data splits and random initialization and data splits, with performance metrics differing by <0.5% across five seeds. This level of precision is favorable for regulatory compliance and clinical implementation, where reproducibility must be ensured. Default settings in the dataflow automation, along with the versioned MSigDB Hallmarks v7.5, aid in reproducibility across research collaboratives with minimal reproducibility and reproducibility of CBPS.

Limitations regarding batch effects and cross-platform robustness. TCGA data are processed via the GDC uniform workflow (STAR-Counts pipeline with FPKM-UQ normalisation), which largely, but not completely, removes technical batch variation between contributing institutions. H-PAAE has not been tested on data with injected noise, simulated batch effects, or across platforms (for instance, data generated from different RNA-seq library preparation protocols). This is a real concern for clinical deployment. Future work will explore ComBat and PEER factor correction prior to H-PAAE training, and domain-adversarial training strategies to promote learning of batch-invariant representations. External GEO cohort testing and data generated using 3′ tag-based sequencing protocols are the main avenues for establishing cross-platform generalisability.

Interpretability chain from SHAP to biological pathway. H-PAAE SHAP attributions are computed over the 128-dimensional latent space rather than directly over original genes. This is one degree of indirection away from gene-level causality; however, this indirection is substantially alleviated by the model design: each latent dimension is built via pathway-masked attention, meaning latent features are biologically meaningful summaries of curated MSigDB pathway activity rather than arbitrary learned dimensions. The gene-level attention weights $\alpha_{p,g}$ give within-pathway gene importance, and the GSEA prerank analysis on genes ranked by global attention importance provides the direct gene-to-pathway bridge, confirming alignment with established cancer hallmarks (apoptosis, androgen response; FDR < 0.15). Together, the three interpretability layers (gene-level attention, pathway-level attention, and SHAP over latent features) provide a seamless multi-resolution explanation. Biological verification of the enriched pathway signatures—through cross-dataset GSEA replication or controlled perturbation experiments—remains an important future step.

There are a number of key areas for future research. (i) Multi-omics extension: extending H-PAAE to include somatic mutation profiles, copy number alterations, and DNA methylation alongside RNA-seq via late-fusion or joint-embedding approaches within the H-PAAE latent space may enhance performance for genomically driven cancer subtypes. (ii) External cohort validation: applying H-PAAE to independently generated RNA-seq datasets from GEO—including data generated with different sequencing protocols—is the most critical step toward demonstrating clinical generalisability and addressing cross-platform

batch effects. (iii) Federated learning: joint training across multiple institutions without sharing raw data would satisfy privacy constraints and simultaneously enable real-world validation across diverse patient populations. (iv) Spatiotemporal modelling: combining spatial transcriptomics and longitudinal RNA-seq profiles would enable modelling of intra-tumour heterogeneity and treatment-resistance dynamics. (v) Rare-cancer transfer learning: data augmentation and transfer learning from related cancer types will address the precision-recall limitations for under-represented TCGA categories (fewer than 50 samples). (vi) Comparison with transformers and GNN architectures: a common benchmark of H-PAAE against GexBERT [31], graph-attention network approaches [35], and multimodal frameworks [34] on identical dataset splits would rigorously position the relative merits of pathway-constrained inductive biases vs. unconstrained representation learning.

The H-PAAE framework balances exceptional predictive accuracy, biological interpretability, and clinical calibration, making it a methodological advancement for pan-cancer classification. The biological pathways as a form of inclusion reduction are more efficient than attention-based autoencoders and surpass generic autoencoders. Assessing the framework against 33 cancer types reinforces its utility as a dependable point of reference for subsequent research as well as its practicality for challenging clinical diagnostic choices.

6 Conclusion

This research presents a novel hierarchical masked attention pathway autoencoder (H-PAAE) model tailored for a multi-cancer classification problem using transcriptomic data from 33 tumors in TCGA along with a biological pathway knowledge (MSigDB Hallmarks) dataset whereby the framework achieves 92.37% test accuracy and 99.38% macro-averaged AUROC while maintaining testable biological interpretability using gene- and pathway-level importance scoring with attention-based dimensionality reduction. The XGBoost classifier demonstrates notable efficacy during cross-validation, exhibiting a stability of 92.5% with a standard deviation of 0.6%, alongside balanced performance across the macro F1 (88.33%), weighted F1 (92.08%), and AUPRC (92.10%) metrics. A robust 20,000 gene feature set is compressed to 128 dimensions at a ratio of 150:1 with discriminative pan-cancer features ascertained from SHAP analysis.

Their primary findings include: (1) the pathway-masked encoder which spatially constrains the flow of hidden states in the encoder to enhance biological adherence, (2) multi-level attention which dissects individual gene and pathway contributions, (3) increased classification performance accompanied by well-calibrated probability predictions which enables risk-stratified clinical pathway guidance, and (4) sophisticated SHAP and gene set enrichment analysis to elucidate connections of latent features to cancer hallmarks for a more comprehensive rationale. Validation with five random seeds ensures reproducibility which is further supplemented by t-SNE visualizations that illustrate the effectiveness of log1p-CPM normalization in cancer-type separation along tissue lineages and organ system dimensions.

The H-PAAE Framework addresses important gaps in existing pan-cancer classifiers by rectifying weaknesses of earlier investigations which merely stared at a handful of cancers or applied deep networks which could not be understood. This approach integrates pathway structure to bridge predictive performance and clinical interpretability, providing multi-modal interpretability. H-PAAE's calibrated posterior probabilities and stability across data splits lend themselves as strong candidates to diagnostic workflows wherein molecular profiling augments, and is still augmented by, histopathology.

Validation by external data remains a limitation which requires independent datasets and sequencing platforms. Rare cancers, with fewer than 50 samples, demonstrate poor precision-recall performance, indicating the need for data augmentation or transfer learning. Future studies should prioritise: (i) external cohort validation on independent RNA-seq datasets (e.g., GEO) with cross-platform batch-effect correction (ComBat, PEER); (ii) multi-omics extension incorporating somatic mutations, copy number alterations,

and DNA methylation via joint H-PAAE embeddings; (iii) federated learning for privacy-preserving multi-institutional training; (iv) spatiotemporal tumour modelling integrating spatial transcriptomics and longitudinal profiling; (v) rare-cancer transfer learning to improve precision-recall for under-represented TCGA types; and (vi) outlier-rejection strategies to handle cancer types beyond the 33 TCGA categories. A rigorous benchmark of H-PAAE against transformer-based architectures (e.g., GexBERT [31]) and graph neural networks [35] under identical experimental conditions is also planned as a priority future study.

To conclude, the H-PAAE framework provides methodical groundwork for biologically interpretable pan-cancer classification while showing that high predictive accuracy can be obtained without compromising understanding of mechanisms. The comprehensive evaluation and implementation provides reliable baseline for research community and aids groundwork for clinical use in complicated cases like cancer of unknown primary. This work increases the framework machine learning can be applied in precision oncology by integrating computational accuracy, clinical verification, and biological anchoring.

Acknowledgement: This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No. DGSSR-2025-FC-01029.

Funding Statement: This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No. DGSSR-2025-FC-01029.

Author Contributions: Conceptualization, Maram Fahaad Almufareh; Data curation, Samabia Tehsin; Funding acquisition, Maram Fahaad Almufareh; Investigation, Samabia Tehsin; Methodology, Samabia Tehsin; Project administration, Maram Fahaad Almufareh; Resources, Maram Fahaad Almufareh & Samabia Tehsin; Supervision, Samabia Tehsin; Validation, Maram Fahaad Almufareh; Visualization, Samabia Tehsin; Writing—original draft, Samabia Tehsin. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in the TCGA Genomic Data Commons (GDC) repository at https://portal.gdc.cancer.gov/.

Ethics Approval: Not applicable. This study used publicly available, de-identified data from The Cancer Genome Atlas (TCGA) via the Genomic Data Commons (GDC). No primary data collection from human subjects was performed.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xi Z, Dai R, Ze Y, Jiang X, Liu M, Xu H. Traditional Chinese medicine in lung cancer treatment. Mol Cancer. 2025;24:57.
2. O'Toole CC, Boakye NF, Hannigan A, Jalali A. Clinical impact of MRI-based risk calculators for prostate cancer diagnosis: a systematic review and meta-analysis. Prostate Cancer Prostatic Dis. 2026;29(2):247–257. doi:10.1038/s41391-025-01014-2.
3. Saeidi T, Mahmood SN, Saleh S, Timmons N, Al-Gburi AJA, Razzaz F. Ultra-wideband (UWB) antennas for breast cancer detection with microwave imaging: a review. Results Eng. 2025;25(4):104167. doi:10.1016/j.rineng.2025.104167.
4. Wekalao J, Hao L, Haq IU, Khan MA. Roadmap to 2D graphene nanomaterials-based biosensors for early cancer detection. Plasmonics. 2026;21:457–70.
5. Bray F, Laversanne M, Ferlay J, Colombet M, Piñeros M, Znaor A, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63.
6. Filho AM, Laversanne M, Ferlay J, Colombet M, Piñeros M, Znaor A, et al. The GLOBOCAN, 2022 cancer estimates: data sources, methods, and a snapshot of the cancer burden worldwide. Int J Cancer. 2025;156(8):1336–46.

7. American Cancer Society. Global cancer facts & figures, 5th ed.; 2024 [cited 2025 Oct 20]. Available from: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/global-cancer-facts-and-figures/global-cancer-facts-and-figures-2024.pdf.
8. Siegel RL, Kratzer TB, Giaquinto AN, Sung H, Jemal A. Cancer statistics, 2025. CA Cancer J Clin. 2025;75:10–45.
9. American Association for Cancer Research. Screening for early detection. 2023 [cited 2025 Oct 20]. Available from: https://cancerprogressreport.aacr.org/progress/cpr23-contents/cpr23-screening-for-early-detection/.
10. World Health Organization. Cancer programme: prevention, early detection, diagnosis, treatment, and palliative care. 2025 [cited 2025 Oct 20]. Available from: https://www.who.int/teams/noncommunicable-diseases/ncds-management/cancer-programme.
11. Tan SL, Selvachandran G, Paramesran R, Ding W. Lung cancer detection systems applied to medical images: a state-of-the-art survey. Arch Comput Methods Eng. 2025;32(1):343–80. doi:10.1007/s11831-024-10141-3.
12. Yang G, Luo S, Greer P. Advancements in skin cancer classification: a review of machine learning techniques in clinical image analysis. Multimed Tools Appl. 2025;84(11):9837–64. doi:10.1007/s11042-024-19298-2.
13. Yao IZ, Dong M, Hwang WYK. Deep learning applications in clinical cancer detection: a review of implementation challenges and solutions. Mayo Clin Proc Digit Health. 2025;3(3):100253. doi:10.1016/j.mcpdig.2025.100253.
14. National Cancer Institute. The cancer genome atlas program (TCGA). 2025 [cited 2025 Oct 20]. Available from: https://www.cancer.gov/ccg/research/genome-sequencing/tcga.
15. The Cancer Genome Atlas Research Network. Welcome to the pan-cancer atlas. 2018 [cited 2025 Oct 20]. Available from: https://www.cell.com/pb-assets/consortium/PanCancerAtlas/PanCani3/index.html.
16. National Institutes of Health. NIH completes in-depth genomic analysis of 33 cancer types (PanCancer Atlas). 2018 [cited 2025 Oct 20]. Available from: https://www.nih.gov/news-events/news-releases/nih-completes-depth-genomic-analysis-33-cancer-types.
17. Liñares-Blanco J, Pazos A, Fernandez-Lozano C. Machine learning analysis of TCGA cancer data. PeerJ Comput Sci. 2021;7:e584.
18. Tomczak K, Czerwinska P, Wiznerowicz M. Review the cancer genome atlas (TCGA): an immeasurable source of knowledge. Wspolczesna Onkol. 2015;1A:68–77. doi:10.5114/wo.2014.47136.
19. Winterhoff B, Hamidi H, Wang C, Kalli KR, Fridley BL, Dering J, et al. Molecular classification of high grade endometrioid and clear cell ovarian cancer using TCGA gene expression signatures. Gynecol Oncol. 2016;141(1):95–100. doi:10.1016/j.ygyno.2016.02.023.
20. Thennavan A, Beca F, Xia Y, Garcia-Recio S, Allison K, Collins LC, et al. Molecular analysis of TCGA breast cancer histologic types. Cell Genom. 2021;1(3):100067. doi:10.1016/j.xgen.2021.100067.
21. Yang Z, Yin H, Shi L, Qian X. A novel microRNA signature for pathological grading in lung adenocarcinoma based on TCGA and GEO data. Int J Mol Med. 2020;45(5):1397–408. doi:10.3892/ijmm.2020.4526.
22. Li H, Zhu Y, Burnside ES, Huang E, Drukker K, Hoadley KA, et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. npj Breast Cancer. 2016;2(1):16012. doi:10.1038/npjbcancer.2016.12.
23. Mohammed M, Mwambi H, Mboya IB, Elbashir MK, Omolo B. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Sci Rep. 2021;11(1):15626.
24. Guo Z, Huang J, Wang Y, Liu XP, Li W, Yao J, et al. Analysis of expression and its clinical significance of the secreted phosphoprotein 1 in lung adenocarcinoma. Front Genet. 2020;11:547. doi:10.3389/fgene.2020.00547.
25. Lombardi O, Li R, Halim S, Choudhry H, Ratcliffe PJ, Mole DR. Pan-cancer analysis of tissue and single-cell HIF-pathway activation using a conserved gene signature. Cell Rep. 2022;41(7):111652.
26. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun. 2020;11(1):3877.
27. Zheng Y, Pizurica M, Carrillo-Perez F, Noor H, Yao W, Wohlfart C, et al. Digital profiling of cancer transcriptomes from histology images with grouped vision attention. BioRxiv:560068. 2023. doi:10.1101/2023.09.28.560068.
28. Alsaafin A, Safarpoor A, Sikaroudi M, Hipp JD, Tizhoosh HR. Learning to predict RNA sequence expressions from whole slide images with applications for search and classification. Commun Biol. 2023;6(1):304. doi:10.1038/s42003-023-04583-x.

29. Suphavilai C, Chia S, Sharma A, Tu L, Silva RPD, Mongia A, et al. Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. Genome Med. 2021;13(1):189. doi:10.1186/s13073-021-01000-y.
30. Shahriyari L. Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. Brief Bioinform. 2019;20(3):985–94. doi:10.1093/bib/bbx153.
31. Jiang S, Hassanpour S. Transformer-based representation learning for robust gene expression modeling and cancer prognosis. Sci Rep. 2025;15(1):37581. doi:10.1038/s41598-025-14949-2.
32. Younis H, Minghim R. Enhancing cancer classification from RNA sequencing data using deep learning and explainable AI. Mach Learn Knowl Extr. 2025;7(4):114. doi:10.3390/make7040114.
33. Shanmugam N, Krishnan A, Inbarani HH, Khan M. Explainable machine learning framework for gene expression-based biomarker identification and cancer classification using feature selection. Med Data Min. 2025;8(3):19. doi:10.53388/mdm202508019.
34. Benkirane H, Vakalopoulou M, Planchard D, Adam J, Olaussen K, Michiels S, et al. Multimodal CustOmics: a unified and interpretable multi-task deep learning framework for multimodal integrative data analysis in oncology. PLoS Comput Biol. 2025;21(6):e1013012. doi:10.1371/journal.pcbi.1013012.
35. Zhang D, Bian G, Zhang Y, Xie J, Hu C. MOLUNGN: a multi-omics graph neural network for biomarker discovery and accurate lung cancer classification. Front Genet. 2025;16:1610284. doi:10.3389/fgene.2025.1610284.
36. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. doi:10.1093/bioinformatics/bts635.
37. Genomic Data Commons. mRNA analysis pipeline. National cancer institute, genomic data commons documentation. 2016 [cited 2026 Jan 10]. Available from: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/.
38. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):13. doi:10.1186/s13059-016-1047-4.
39. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi:10.1186/s13059-014-0550-8.
40. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2018;19(5):776–92. doi:10.1093/bib/bbx008.
41. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–40. doi:10.1093/bioinformatics/btr260.
42. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. Cell Syst. 2015;1(6):417–25. doi:10.1016/j.cels.2015.12.004.
43. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. doi:10.1093/nar/28.1.27.
44. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008. doi:10.65215/ctdc8e75.
45. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473. 2015.
46. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv:1412.6980. 2015.
47. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32:8024–35.
48. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. New York, NY, USA: ACM. p. 785–94.
49. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–232. doi:10.1214/aos/1013203451.
50. Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357–62. doi:10.1038/s41586-020-2649-2.

51. McKinney W. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference; 2010 Jun 28–Jul 3; Austin, TX, USA. Vol. 445. p. 51–6.
52. Folk M, Heber G, Koziol Q, Pourmal E, Robinson D. An overview of the HDF5 technology suite and its applications. In: Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases; 2010 Mar 25; Uppsala, Sweden. p. 36–47.
53. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
54. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. Vol. 30. Red Hook, NY, USA: Curran Associates, Inc.; 2017. NIPS 2017. p. 4765–74.
55. Shapley LS. A value for n-person games. Contrib Theory Games. 1953;2(28):307–17. doi:10.1515/9781400881970-018.
56. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67. doi:10.1038/s42256-019-0138-9.
57. Lundberg SM, Lee SI. SHAP (SHapley Additive exPlanations); Python package version 0.41.0. 2020 [cited 2025 Nov 20]. Available from: https://github.com/slundberg/shap.
58. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(86):2579–605.
59. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn. 2001;45(2):171–86. doi:10.1023/a:1010920819831.
60. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning; 2006 Jun 25–29; Pittsburgh, PA, USA. New York, NY, USA: ACM. p. 233–40.