GMR
The Original

# A MULTI-STAGE FEATURE SELECTION AND HYBRID CNN–TRANSFORMER FRAMEWORK WITH BLOCKCHAIN-ENABLED CONSENT GOVERNANCE FOR EXPLAINABLE CLINICAL GENOMIC RISK ASSESSMENT

**Kashyap Dave[1*], Dr. Hitesh kumar M. Nimbark[2]**

[1,2] Gyanmanjari Institute of Technology, Gyanmanjari Innovative University, Bhavnagar, Gujarat, India
ORCID: 0009-0008-2621-795X[1], 0009-0006-7040-0458[2]
*Corresponding author: kashyapdave.be@gmail.com

**ABSTRACT**
High-throughput sequencing has made clinical genomic data abundant, but it has not made genomic artificial intelligence trustworthy. Gene-expression matrices are extremely wide and extremely shallow — tens of thousands of transcripts measured over a few hundred patients — a regime in which classifiers memorise rather than generalise, operate as opaque decision engines, and are trained on some of the most re-identifiable data a person can produce. This paper presents a complete methodological progression from an empirically grounded diagnosis of that failure to a governance-aware architecture designed to correct it. A published baseline pipeline was independently reproduced on the GSE68086 tumour-educated-platelet RNA-sequencing benchmark (57,736 transcripts; 173 samples; six cancer classes) using Support Vector Classification, Decision Tree, Multi-Layer Perceptron and one-dimensional Convolutional Neural Network models. All four classifiers exhibited severe train–test divergence: generalisation gaps of 18.3, 45.3, 41.3 and 37.6 percentage points respectively, with the best test accuracy reaching only 60.4 per cent despite 98.1 per cent training accuracy and a multi-class AUC of 0.89. Diagnostic analysis attributed this divergence to three specific and correctable design faults rather than to dataset difficulty alone: unsupervised Shannon-entropy gene ranking that optimises variability instead of class separability, synthetic minority oversampling applied before rather than within the train–test partition, and a purely convolutional inductive bias that cannot model the long-range gene–gene dependencies characteristic of co-regulated transcriptional programmes. Each fault is then addressed explicitly. A three-stage supervised gene-selection pipeline is formalised — univariate ANOVA F-test or mutual-information screening, minimum-Redundancy-Maximum-Relevance filtering, and cross-validated LASSO confirmation — and is selected through a structured comparison of filter, wrapper, embedded and hybrid families against redundancy handling, small-sample stability, computational cost and reproducibility. A hybrid one-dimensional CNN–Transformer classifier is then justified against four competing architectures, coupling convolutional local motif extraction with multi-head self-attention over the selected gene tokens. The predictive core is embedded in a layered deployable framework that adds SHAP-based and attention-based explanation, a blockchain smart-contract layer providing dynamic patient consent and immutable access auditing, and a clinician-facing decision-support interface. A systematic review of fifteen studies published in indexed journals during 2025–2026 confirms that no existing work simultaneously compares multiple feature-selection strategies, models both local and global gene interactions, exposes a structured explanation layer, and enforces consent governance. The contribution is therefore an evidence-linked architectural specification with a pre-registered progressive validation protocol, in which every component is traceable to a measured baseline failure.

**KEYWORDS:** Clinical genomic risk assessment; feature selection; mRMR; LASSO; hybrid CNN–Transformer; explainable artificial intelligence; SHAP; blockchain consent management; generalisation gap; high-dimensional low-sample-size data.

## 1. INTRODUCTION

Next-generation sequencing has compressed the cost of reading a human transcriptome by several orders of magnitude, and the consequence is a research environment in which the measurement is no longer the bottleneck. Public repositories now hold expression profiles for hundreds of thousands of patients across dozens of malignancies [1], and this abundance underpins the translation of genomic insight into targeted, patient-specific therapy [2]. The bottleneck has shifted upstream, into the analytical layer: deciding which of tens of thousands of measured transcripts actually carry a diagnostic signal, building models that learn biology rather than sampling noise, explaining those models to the clinicians expected to act on them, and governing the underlying data in a way that patients can accept. The last of these is not a peripheral concern, because genomic records are among the most persistently re-identifiable data a person can produce [3].

These four problems are usually studied in isolation, and that isolation is itself the problem. A gene-selection study optimises a subset without asking whether the downstream architecture can exploit it. An architecture study reports headline accuracy without reporting the gap between training and testing performance that determines whether the number will survive contact with a new cohort. An explainability study attaches post-hoc attributions to a model whose input features were never chosen for biological interpretability. A privacy study designs consent infrastructure with no predictive workload attached to it. Each contribution is locally sound; the composition is not, and the composition is what a hospital would have to deploy.

This paper takes the opposite route. It begins with a measurement rather than a proposal. A published genomic-classification pipeline [4] was reproduced end-to-end on a deliberately difficult public benchmark, and the resulting failure was characterised precisely enough to be actionable. That diagnosis then drives every subsequent design decision, so that the proposed architecture is not a plausible combination of fashionable components but a set of targeted responses to observed defects.

### 1.1 Motivation

Genomic expression matrices occupy a statistical regime that most machine-learning intuition is not calibrated for. Where a typical benchmark offers many samples and few features, a gene-expression study offers the reverse: in the dataset examined here, 57,736 transcripts are measured across 173 usable samples, a feature-to-sample ratio exceeding 300:1. In this regime the training data can almost always be separated perfectly by chance alone, so training accuracy carries essentially no information about generalisation. Compounding this, biologically related malignancies share large portions of their transcriptional programmes, so the classes are not cleanly separable even in principle. And unlike most data modalities, genomic records cannot be meaningfully anonymised — a genotype is a permanent identifier for the individual and a partial identifier for their relatives [3] — which means the governance question cannot be deferred to a later engineering phase.

The practical consequence is that accuracy alone is the wrong optimisation target for this domain. A model reporting 98 per cent training accuracy and 60 per cent test accuracy is not a strong model with a minor deployment caveat; it is a model that has learned the training set. Reporting only the former is common in the literature and actively misleading. This work therefore adopts the generalisation gap — the arithmetic difference between training and testing accuracy — as a first-class reported metric alongside conventional performance measures, and treats its reduction as the primary design objective.

### 1.2 Research Objectives

The study is organised around five objectives, each mapped to a specific section of this paper:

• RO1 — To investigate and comparatively analyse genomic feature-selection techniques for effective dimensionality reduction and identification of informative biomarkers from high-dimensional clinical genomic data.

• RO2 — To develop a hybrid CNN–Transformer framework for accurate and robust clinical genomic risk assessment through efficient feature representation and classification.

• RO3 — To incorporate Explainable AI techniques that improve model transparency and enhance the clinical interpretability of genomic risk predictions.

• RO4 — To design a blockchain-enabled consent-management layer for secure genomic data governance, ensuring patient privacy, ownership and controlled data access.

• RO5 — To develop a clinical decision-support interface that presents interpretable genomic classification output to healthcare professionals in an actionable form.

### 1.3 Contributions

The specific contributions of this paper are as follows.

• A reproducible baseline benchmark on GSE68086 across four classifier families, with the generalisation gap reported explicitly for every model rather than concealed behind aggregate accuracy.

• A root-cause analysis that attributes the observed divergence to three identifiable and correctable pipeline defects — unsupervised entropy-based gene ranking, oversampling applied outside the validation partition, and a convolution-only inductive bias — rather than to dataset difficulty in the abstract.

• A systematic literature review of fifteen studies published in indexed journals during 2025–2026, synthesised into a capability matrix that demonstrates the absence of any integrated solution across the six required dimensions.

• A formalised three-stage supervised feature-selection pipeline (ANOVA F-test or mutual information → mRMR → cross-validated LASSO), selected through explicit comparison of filter, wrapper, embedded and hybrid families against reproducibility and small-sample stability criteria.

• A hybrid 1D-CNN–Transformer architecture justified against four competing candidates, with the mathematical argument for why convolution alone is structurally unable to represent long-range gene–gene dependencies in an unordered feature space.

• A complete layered framework specification integrating prediction, explanation, blockchain-based consent governance and clinical decision support, accompanied by a staged validation protocol defined in advance of experimentation.

### 1.4 Organisation of the Paper

Section 2 presents the systematic literature review and derives the research gaps. Section 3 describes the dataset. Section 4 reports the baseline reproduction and its diagnostic analysis. Section 5 develops the feature-selection framework. Section 6 specifies the proposed architecture. Section 7 compares the framework against existing studies. Section 8 discusses results, Section 9 states limitations and future scope, and Section 10 concludes.

## 2. SYSTEMATIC LITERATURE REVIEW

A structured review was conducted to establish what the current state of the art achieves and, more importantly, what it does not attempt. The review deliberately restricts itself to very recent work so that the identified gaps reflect the present frontier rather than a historical one.

### 2.1 Review Protocol

Searches were executed across Scopus, PubMed, IEEE Xplore, SpringerLink, ScienceDirect and publisher-hosted repositories (Nature Portfolio, PLOS, MDPI, PeerJ, JMIR) using combinations of the terms "gene expression classification", "feature selection", "cancer genomic risk", "hybrid deep learning", "transformer", "explainable AI" and "blockchain genomic consent". Inclusion required that a study (i) be published in a peer-reviewed, indexed journal between January 2025 and the review cut-off in 2026; (ii) address classification, prognosis or feature selection on high-dimensional molecular data; (iii) report a quantitative evaluation; and (iv) be available in full text. Conference abstracts, non-indexed venues, purely wet-laboratory studies and preprints without journal acceptance were excluded. Fifteen studies satisfied all criteria and form the review corpus summarised in Table 1.

Each study was extracted against a fixed schema: venue and year, dataset, feature-selection strategy, model family, reported performance, and stated limitations. A second extraction pass scored every study against six capability dimensions derived from the research objectives — comparative feature selection, local pattern modelling, global dependency modelling, structured explainability, clinical interface, and consent governance — producing the capability matrix in Table 2.

**Table 1. Systematic review of fifteen genomic classification and feature-selection studies published in indexed journals, 2025–2026.**

| Ref. | Study / Method | Journal (Year) | Dataset | FS / Model | Key finding and limitation |
|---|---|---|---|---|---|
| [4] | KalaiSelvi et al. — ML with IoT biosensors (baseline reference) | Scientific Reports (2025) | GEO GSE68086 (57,736 genes, 285 samples) | Shannon entropy + SVC / DT / MLP / CNN + STRING-KEGG | Couples expression profiling with simulated biosensor readout and pathway enrichment; entropy FS is unsupervised, biosensor modelling theoretical, no hybrid architecture or governance. |
| [5] | Dhamercherla et al. — Eagle Prey Optimization | Frontiers in Genetics (2025) | Microarray cancer sets (GEO) | Eagle Prey Optimization (wrapper) | Strong subset accuracy on microarray benchmarks; stochastic search is difficult to reproduce across runs and computationally heavy. |
| [6] | Yaqoob et al. — mRMR + SSO + WSVM | Multimedia Tools and Applications (2025) | Breast cancer RNA-seq / microarray | mRMR + Social Spider Optimization + weighted SVM | Hybrid filter–wrapper improves breast-cancer classification; wrapper stage is tuning-heavy and single-cancer scope. |
| [7] | Cherif et al. — SSO with mutual information | Iranian Journal of Computer Science (2026) | Colon, prostate, leukaemia, lymphoma | SSO + MIM / JMI / mRMR | High reported accuracy across four cohorts; swarm stage is non-deterministic and no generalisation-gap reporting. |
| [8] | Shi et al. — multiple filter-wrapper FS | PLOS ONE (2025) | 10 high-dimensional microarray cancer sets | Random-forest importance + dung-beetle wrapper | Process-optimisation mechanism improves subset quality across ten datasets; |

| Ref. | Study / Method | Journal (Year) | Dataset | FS / Model | Key finding and limitation |
|---|---|---|---|---|---|
| | | | | | wrapper stage remains computationally costly. |
| [9] | Sahu et al. — Binary Portia Spider + fast mRMR | Bioengineering (2025) | Cancer microarray | Binary Portia-spider optimisation + fast mRMR | Demonstrates that redundancy-aware filters accelerate metaheuristic search; custom metaheuristic imposes implementation burden. |
| [10] | Zhang et al. — GONF (gene-optimised neural framework) | Scientific Reports (2025) | TCGA, AHBA | mRMR + deep CNN over image-mapped expression | 97% (TCGA) and 95% (AHBA) with biologically interpretable genes; single FS method, CNN-only, no global gene modelling or XAI layer. |
| [11] | Shah et al. — hybrid PCA + mutual information | PLOS ONE (2026) | TCGA + ICGC (lung) | PCA + MI feature extraction + CNN | 98% accuracy with PPI hub-gene validation, benchmarked against ten FS methods; lung-only, CNN-only, no transformer, XAI or governance. |
| [12] | Jiang & Hassanpour — GexBERT | Scientific Reports (2025) | Large-scale TCGA (pan-cancer) | Pre-trained transformer, learned gene embeddings | Transformer pretraining yields strong performance from as few as 64 input genes; requires large-scale pretraining, no explicit FS, no local extractor. |
| [13] | EL kati et al. — hybrid GNN-Transformer | PeerJ Computer Science (2026) | Multi-omic TCGA (expression, methylation, pathways) | Pathway-guided + wavelet reduction | Pathway-guided attention gives interpretable multi-omic subtype classification; GNN rather than CNN front end, single FS strategy, no consent layer or GUI. |
| [14] | Li et al. — DeepInsight CNN | Health Information Science and Systems (2025) | TCGA breast, lung, colon | Class-activation-map multi-class FS | CAM-based multi-class gene selection outperforms SVM, LightGBM, NN and DT on F1; gene-to-image mapping is spatially arbitrary, CNN-only. |
| [15] | Reddy et al. — MoX | Journal of King Saud University – CIS (2025) | GSE85047 (neuroblastoma) | Random forest + dual-branch DL + SHAP | Explainable event-free-survival prediction from multi-omics; survival task rather than multi-class detection, no transformer, no governance. |
| [16] | Zeng et al. — MGAN-FB | Scientific Reports (2025) | Eight open-source gene microarray datasets | Feature bundling + multi-class GAN | Addresses high-dimension, small-sample and class imbalance jointly at feature and algorithm level; GAN training unstable, no XAI, no FS comparison. |
| [17] | Ameen et al. — stacked DL ensemble | JMIR Bioinformatics and Biotechnology (2025) | TCGA (32 types) + CPTAC / LinkedOmics | Multi-omics integration, stacked SVM/KNN/ANN /CNN/RF | Balanced accuracy of 0.96–0.98 across RNA-seq, mutation and methylation; heavy multi-omics input, no explicit FS comparison or transformer. |

| Ref. | Study / Method | Journal (Year) | Dataset | FS / Model | Key finding and limitation |
|---|---|---|---|---|---|
| [18] | Al Marouf et al. — explainable ML biomarkers | Cancers (2025) | Prostate tissue microarray (102 samples) | Random forest + SHAP | Identifies candidate prostate biomarkers with post-hoc attribution at 81.0% accuracy; very small cohort, single cancer, no generalisation analysis. |

**2.2 Synthesis of Findings**

Three patterns emerge from the corpus, and each is consequential.

The first concerns feature selection. Every study in the corpus adopts a single selection strategy and then defends it; none compares strategies within a common evaluation harness on the same data. The field has bifurcated into a metaheuristic branch [5]–[9] that pursues optimal subsets through stochastic search, and a filter or projection branch [10], [11] that prioritises speed and determinism. The metaheuristic branch reports the higher subset quality but pays for it in reproducibility: swarm and evolutionary searches are non-deterministic [5], [7], sensitive to initialisation, and rarely accompanied by seed-level variance reporting. For a clinical pipeline in which the selected gene panel is itself the clinically meaningful artefact, an unstable selector is a serious liability — a panel that changes between runs cannot be validated, assayed or regulated.

The second concerns architecture. The corpus splits cleanly into convolution-only models [10], [11], [14] and attention-only models [12], with a single graph-plus-attention hybrid [13] that uses a GNN rather than a convolutional front end. This split matters because the two families fail in complementary ways. Convolutional models extract local motifs efficiently and remain trainable at small sample sizes, but their receptive field is bounded and they impose an ordering assumption on gene indices that has no biological meaning — the position of a transcript in an expression matrix is an accident of annotation. Attention-only models represent arbitrary long-range dependencies without any ordering assumption, but their parameter count and quadratic attention cost make them acutely data-hungry; GexBERT [12] achieves its results only through large-scale pretraining on pan-cancer TCGA, an option unavailable to a study working with a few hundred samples. No study in the corpus combines convolutional local extraction with self-attention over a selected gene panel, which is precisely the configuration that would deliver both properties at a tractable parameter budget.

The third concerns deployability. Explainability appears in four studies [10], [11], [14], [15], and in three of those it is partial or incidental interpretability rather than a designed explanation layer; only [15] and [18] treat attribution as a first-class output. No study in the corpus provides a clinician-facing interface, and none addresses consent or data governance at all. This is not an oversight peculiar to these fifteen papers; it reflects a structural division of labour in which the genomic-AI literature and the health-data-governance literature cite different bodies of work and publish in different venues. The governance literature is itself substantial and active — covering the ethics of encrypted genomic sharing [19], privacy-preserving DNA-sequence ecosystems [20], owner-governed genomic data models [21], blockchain-and-machine-learning pipelines for next-generation sequencing data [22], systematic reviews of privacy preservation in blockchain-based health data sharing [23], [24], tokenised genomic data ownership [25], edge and off-chain storage architectures [26], [27], scalability mechanisms for smart-contract systems [28], and federated learning with privacy-preserving explanation [29] — yet none of this work is coupled to a predictive genomic model. The result is that no reviewed system could be deployed against identifiable patient genomes without an entirely separate governance implementation being retrofitted around it.

**Table 2. Capability matrix of reviewed studies against the six dimensions required for a deployable clinical genomic risk assessment system (✓ = present, ~ = partial, ✗ = absent).**

| Study | Multiple FS compared | CNN (local) | Transformer (global) | XAI layer | Clinical GUI | Consent / governance |
|---|---|---|---|---|---|---|
| Baseline reference, Scientific Reports (2025) [4] | ✗ | ✓ | ✗ | ~ | ✗ | ✗ |
| GONF, Scientific Reports (2025) [10] | ✗ | ✓ | ✗ | ~ | ✗ | ✗ |
| PCA-MI + CNN, PLOS ONE (2026) [11] | ✗ | ✓ | ✗ | ~ | ✗ | ✗ |
| GexBERT, Scientific Reports (2025) [12] | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |

| Study | Multiple FS compared | CNN (local) | Transformer (global) | XAI layer | Clinical GUI | Consent / governance |
|---|---|---|---|---|---|---|
| GNN-Transformer, PeerJ Comput. Sci. (2026) [13] | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ |
| DeepInsight CNN, Health Inf. Sci. Syst. (2025) [14] | ✗ | ✓ | ✗ | ~ | ✗ | ✗ |
| MoX, J. King Saud Univ. CIS (2025) [15] | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ |
| MGAN-FB, Scientific Reports (2025) [16] | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Stacked ensemble, JMIR Bioinform. Biotech. (2025) [17] | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ |
| Explainable ML biomarkers, Cancers (2025) [18] | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ |
| **Proposed framework** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

### 2.3 Research Gaps

Four gaps follow directly from the synthesis above and define the scope of the remainder of this paper.

Gap 1 — Local-only modelling and limited generalisability. Most reviewed studies build predictive models on single-centre or single-cohort datasets using architectures whose inductive bias captures only local structure [10], [11], [14]. Neither the evaluation design nor the model family supports a claim of robustness across diverse populations or acquisition platforms, and the generalisation gap that would expose the problem is almost never reported.

Gap 2 — Separation of clinical prediction from data governance. The literature addresses either AI-driven prediction [4]–[18] or privacy, consent and secure sharing [19]–[29], but very rarely both within one framework. There is consequently no end-to-end system that delivers a trustworthy prediction while simultaneously enforcing transparent, patient-controlled management of the genomic record that produced it.

Gap 3 — Weak explainability and limited clinical interpretability. Where explanation is present it is typically coarse and post-hoc [15], [18], attached to models whose input features were never selected for biological meaning. Such explanations do not give a clinician a defensible gene-level rationale for a risk prediction, which limits trust and therefore adoption.

Gap 4 — Absence of GUI-based clinical decision support. Reviewed work concentrates on algorithm development and offline evaluation. Interactive clinician-oriented interfaces that present risk scores alongside their explanations, and that operate only on data the patient has authorised, are effectively absent — leaving a translation gap between published models and routine practice.

## 3. DATASET AND PREPROCESSING

The GSE68086 tumour-educated platelet (TEP) dataset from the NCBI Gene Expression Omnibus [1] was selected as the experimental benchmark. TEP profiling exploits the fact that blood platelets sequester tumour-derived RNA, allowing a molecular readout of malignancy from a peripheral blood draw rather than a tissue biopsy. This makes the dataset directly relevant to non-invasive diagnostic workflows [4], which is why it was retained despite being substantially harder than the pan-cancer TCGA cohorts favoured elsewhere in the literature [12], [17].

The raw matrix contains 57,736 transcript rows across 285 sample columns. Samples were transposed to patient-major orientation and labelled by parsing the six malignancy identifiers present in the sample names — Breast, Colorectal (CRC), Glioblastoma (GBM), Liver, Lung and Pancreatic. Samples outside these six classes, including healthy controls and ambiguous annotations, were removed, yielding 173 labelled samples. The resulting feature-to-sample ratio of approximately 334:1 places the dataset firmly in the high-dimension, low-sample-size regime that [16] identifies as the central difficulty of microarray and RNA-sequencing classification.

**Table 3. Characteristics of the GSE68086 benchmark dataset.**

| Attribute | Description |
|---|---|
| Dataset | GEO – GSE68086 (Tumour-Educated Platelet RNA-sequencing dataset) |
| Source | NCBI Gene Expression Omnibus (GEO) |
| Accession URL | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68086 |

| Attribute | Description |
|---|---|
| Data type | Platelet-derived RNA expression counts |
| Raw dimensions | 57,736 transcripts × 285 samples |
| Samples after class filtering | 173 |
| Classes | 6 — Breast, CRC, GBM, Liver, Lung, Pancreatic |
| Feature-to-sample ratio | Approximately 334:1 |
| Representation | Numerical gene expression matrix, samples as rows |
| Task | Multi-class cancer-type classification |
| Principal challenges | Curse of dimensionality, severe overfitting risk, transcriptional overlap between biologically related malignancies, small and imbalanced class counts |

Two properties of this dataset deserve emphasis because they shape every result that follows. First, with 173 samples spread across six classes, several classes are represented by fewer than twenty patients, so a single misclassified sample moves class-level recall by several percentage points. Second, the malignancies are not transcriptionally independent: colorectal, liver and pancreatic tumours share substantial gastrointestinal programme overlap, so a portion of the confusion observed later is biological rather than algorithmic. The dataset is therefore an honest stress test rather than a favourable showcase — which is exactly the property required to expose generalisation failure.

## 4. BASELINE IMPLEMENTATION AND DIAGNOSTIC ANALYSIS

The first experimental phase of this study reproduced the pipeline of the reference work [4] independently, using the same dataset, the same feature-selection principle and the same four classifier families, in order to obtain a measured reference point rather than a cited one.

### 4.1 Replication Protocol

Genes were ranked by Shannon entropy computed over the discretised expression distribution of each transcript, and the twenty highest-entropy genes were retained, matching the reference configuration [4]. Features were standardised to zero mean and unit variance. Class imbalance was addressed with Synthetic Minority Oversampling (SMOTE), producing a balanced set of 264 samples, which was partitioned 80:20 with stratification into 211 training and 53 testing samples under a fixed random seed. Four classifiers were then trained on identical partitions: Support Vector Classification with probability estimation enabled, a Decision Tree at default settings, a Multi-Layer Perceptron with a single 128-unit hidden layer, and a one-dimensional Convolutional Neural Network. The CNN comprised a 32-filter convolution with kernel size 3, max pooling, a 64-filter convolution, flattening, a 128-unit dense layer with 0.4 dropout, and a six-way softmax output, optimised with Adam at a learning rate of 0.001 under sparse categorical cross-entropy, with early stopping on validation loss at a patience of eight epochs.

**Table 4. Experimental configuration for the baseline reproduction.**

| Parameter | Value |
|---|---|
| Feature selection | Shannon entropy, top 20 transcripts |
| Scaling | StandardScaler (zero mean, unit variance) |
| Class balancing | SMOTE (random_state = 42) |
| Train–test split | 80:20 stratified, random_state = 42 (211 train / 53 test) |
| SVC | RBF kernel, probability = True |
| Decision Tree | Default parameters (unpruned) |
| MLP | Hidden layer 128 units, max_iter = 500 |
| CNN layer 1 | Conv1D, 32 filters, kernel 3, ReLU |
| CNN layer 2 | Conv1D, 64 filters, kernel 3, ReLU |
| Dense / dropout | 128 units, dropout 0.4 |
| Output | Softmax over 6 classes |
| Optimiser / loss | Adam, lr = 0.001 / sparse categorical cross-entropy |
| Batch size / epochs | 16 / up to 100, early stopping patience 8 |
| Environment | Python, TensorFlow-Keras, Scikit-learn, imbalanced-learn; 16 GB RAM |

### 4.2 Baseline Results

Table 5 reports training accuracy, testing accuracy and the generalisation gap for all four classifiers. The gap is defined as the arithmetic difference between training and testing accuracy and is reported here as a primary result rather than a footnote.

**Table 5. Baseline comparative performance on GSE68086, with the generalisation gap reported explicitly.**

| Model | Training accuracy | Testing accuracy | Generalisation gap | Interpretation |
|---|---|---|---|---|
| SVC | 0.730 | 0.547 | 0.183 | Stable but underpowered — underfits |
| Decision Tree | 1.000 | 0.547 | 0.453 | Complete memorisation of training set |
| MLP | 0.960 | 0.547 | 0.413 | Unstable deep model, severe overfitting |
| CNN | 0.981 | 0.604 | 0.376 | Best test accuracy, still severe overfitting |

The pattern is unambiguous. Every model converges to essentially the same testing accuracy — 54.7 per cent for three of the four — while their training accuracies differ enormously. The Decision Tree achieves perfect training separation and generalises no better than the underfitted SVC, which is the signature of memorisation rather than learning. The CNN attains the best test accuracy at 60.4 per cent and the highest multi-class AUC at 0.89, but it does so with a 37.6-point gap, meaning that roughly two-fifths of its apparent capability does not survive the move to held-out data. An accuracy-only report of this experiment would have described a CNN reaching 98 per cent, which would have been a factually accurate and thoroughly misleading summary.

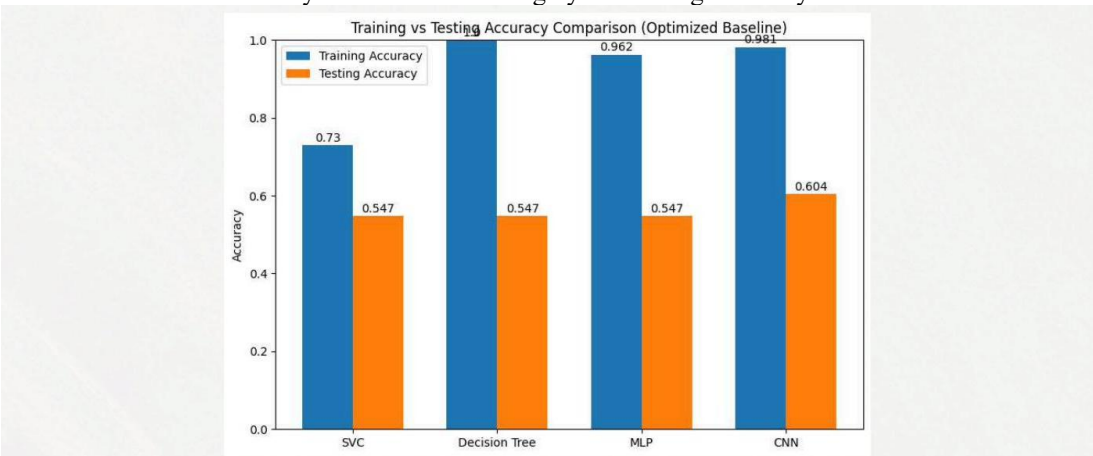


**Figure 1. Training and testing accuracy across the four baseline models, showing the generalisation gap that aggregate accuracy reporting conceals.**

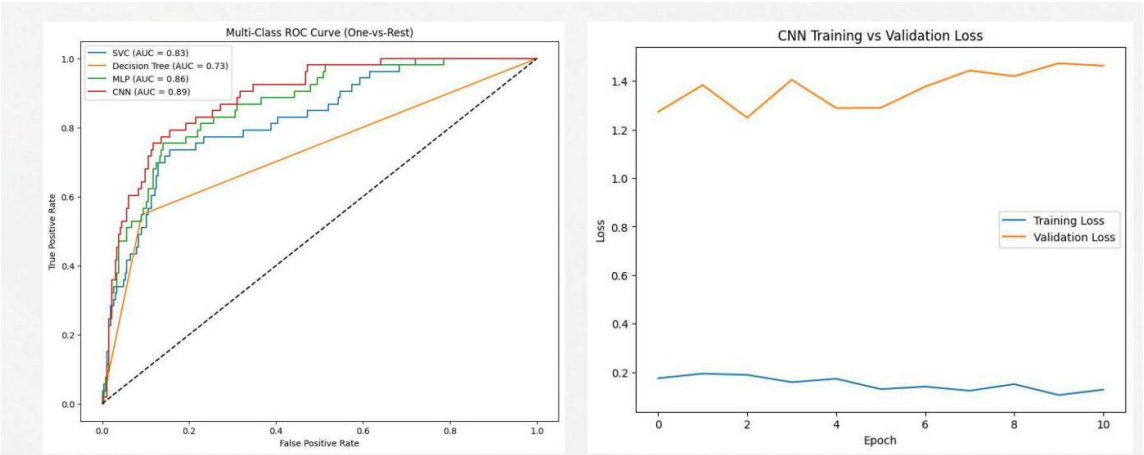


**Figure 2. Multi-class ROC-AUC curves for the baseline models (left) and CNN training versus validation loss trajectories (right), showing discriminative capability alongside loss divergence**.

The loss trajectories in Figure 2 make the mechanism visible. Training loss declines smoothly toward zero while validation loss plateaus early and then oscillates without improvement — the classic divergence signature of a model fitting sample-specific noise once the genuine signal has been exhausted. The ROC curves are more encouraging: an AUC of 0.89 indicates that the model has learned real class-discriminative structure, and that the low accuracy reflects poor decision-boundary calibration in a crowded six-class space rather than an absence of learnable signal. This distinction matters for the design that follows, because it implies the problem is representational and regularisation-related, not a claim that platelet RNA cannot separate these malignancies.

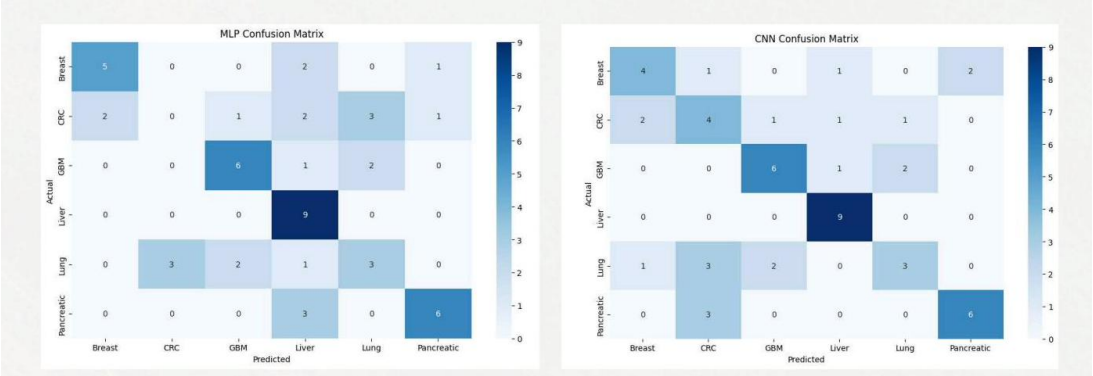


**Figure 3. Confusion matrices for the MLP and CNN models, showing class-level misclassification concentrated between biologically related malignancies.**

The confusion matrices confirm that errors are structured rather than random. Misclassification concentrates between malignancies sharing transcriptional programmes, particularly among the gastrointestinal group, while glioblastoma — transcriptionally distinct from the epithelial cancers — is separated more reliably. Structured error of this kind indicates that the model has learned biologically meaningful boundaries but lacks the representational capacity to resolve fine distinctions between overlapping programmes.

### 4.3 Diagnostic Analysis: Three Correctable Defects

Attributing this failure to "high dimensionality" would be true and useless. Closer examination of the pipeline identifies three specific defects, each of which admits a concrete remedy.

**Defect 1 — Unsupervised feature selection**

Shannon entropy measures the dispersion of a transcript's expression distribution without reference to the class label. A gene that varies wildly for reasons unrelated to malignancy — batch effects, platelet activation state, sequencing depth — scores highly, while a gene with modest but perfectly class-separating variation scores poorly. Selecting the twenty highest-entropy transcripts therefore optimises variability, not discriminative power, and there is no reason to expect the resulting panel to be diagnostically informative. This is the single largest identifiable defect in the baseline and motivates the supervised selection framework of Section 5, in line with the class-aware selection strategies adopted in [6], [8], [11].

**Defect 2 — Oversampling applied outside the validation partition**

In the baseline pipeline SMOTE was applied to the full dataset before the train–test split. Because SMOTE synthesises minority-class points by interpolating between existing neighbours, synthetic samples derived from a test-set point can end up in the training set. The test partition is consequently not independent, and the reported testing accuracy is optimistically biased. The practical implication is stark: the true generalisation gaps are likely wider than the values in Table 5, not narrower. Correcting this requires that all resampling, scaling and feature-selection operations be fitted inside each cross-validation fold on training data alone.

**Defect 3 — Convolution-only inductive bias**

A one-dimensional convolution assumes that adjacent positions in the input vector are related, and that relationships are local and translation-invariant. Neither assumption holds for gene expression. The ordering of transcripts in an expression matrix follows annotation identifiers, not biology; two genes adjacent in the matrix are no more likely to be co-regulated than two genes ten thousand positions apart. Convolution therefore imposes a spurious ordering prior while structurally excluding the long-range co-regulation relationships that define transcriptional programmes — the very dependencies that attention-based models such as [12] and [13] are designed to capture. Increasing depth extends the receptive field only slowly and at a parameter cost that a 173-sample dataset cannot support. The architectural remedy is developed in Section 6.

## 5. FEATURE SELECTION FRAMEWORK

Feature selection is the highest-leverage intervention available in this setting. Reducing 57,736 transcripts to a few dozen simultaneously lowers the effective capacity of the model, removes noise-dominated dimensions, cuts training cost, and — if the selection is biologically principled — produces a gene panel that is itself an

interpretable clinical artefact [10], [11]. Selection is therefore treated here as a designed subsystem rather than a preprocessing convenience.

### 5.1 Comparison of Selection Families

Candidate methods were assessed against five criteria that matter specifically in this regime: whether the method removes redundancy between correlated genes, computational cost, stability at small sample size, implementation and reproducibility burden, and fitness for the pipeline's role.

**Table 6. Comparison of filter, wrapper, embedded and hybrid feature-selection families for high-dimensional genomic data.**

| Method | Family | Removes redundancy | Compute cost | Stability (small n) | Implementation burden | Fitness for this pipeline |
|---|---|---|---|---|---|---|
| Univariate filter (ANOVA F / MI) | Filter | No | Very low | High | Trivial | Stage 1 — adopted |
| mRMR | Filter | Yes | Low | High | Easy | Stage 2 — adopted |
| LASSO / Elastic Net | Embedded | Yes | Low | Medium | Easy | Stage 3 validator — adopted |
| Tree importance / RFE | Embedded | Partial | Medium | Medium | Moderate | Optional comparator |
| Evolutionary / swarm search | Wrapper | Yes | High | Low | Hard | Rejected — non-reproducible |
| Ensemble / hybrid selectors | Hybrid | Yes | High | High | Hard | Rejected — disproportionate complexity |
| Autoencoder / deep selection | Embedded | Yes | High | Low | Moderate | Rejected — data-hungry |

The rejection of wrapper-based metaheuristics warrants explanation, since [5]–[9] report strong results with exactly these methods. The objection is not to their accuracy but to their stability. Evolutionary and swarm searches are stochastic; run twice with different seeds on 173 samples, they return materially different gene panels. When the selected panel is intended to become a clinical biomarker set — something to be assayed, validated prospectively and potentially regulated — a selector that cannot reproduce its own output is unusable regardless of its reported accuracy. Deterministic filters trade a small amount of subset optimality for reproducibility, and in this application that is the correct trade.

### 5.2 The Three-Stage Selection Pipeline

The adopted pipeline applies three complementary criteria in sequence, each removing a different class of uninformative feature.

**Stage 1 — Supervised univariate screening**

The full transcript set is reduced by ranking each gene independently on its relationship to the class label. For continuous expression against categorical class, the ANOVA F-statistic compares between-class to within-class variance:

$$F = [\ \Sigma_k\, n_k(\bar{x}_k - \bar{x})^2 / (K-1)\ ] / [\ \Sigma_k\, \Sigma_i\, (x_{ik} - \bar{x}_k)^2 / (N-K)\ ]$$

where K is the number of classes, n_k and $\bar{x}$_k the size and mean of class k, and N the total sample count. Where relationships are expected to be non-linear or non-monotonic, mutual information $I(X;Y) = \Sigma\ p(x,y) \log[p(x,y)/(p(x)p(y))]$ is used instead, since it captures arbitrary statistical dependence rather than only mean separation; this is the screening principle also adopted by [7] and [11]. This stage is computationally trivial, requires no model fitting, and reduces the candidate set from tens of thousands to a few hundred while — crucially, unlike entropy ranking — using the class label at every step.

**Stage 2 — Redundancy elimination via mRMR**

Univariate screening has a systematic blind spot: co-expressed genes carry near-identical information and therefore receive near-identical scores, so the top-ranked set is typically saturated with members of one or two co-regulated modules. Since co-expression is pervasive in transcriptomic data, this failure mode is the rule rather than the exception. Minimum-Redundancy-Maximum-Relevance, applied successfully in [6], [9] and [10], corrects it by scoring candidates against both their relevance to the target and their redundancy with genes already selected:

$$g_j = \arg\max_{g_j \in \Omega - S}\ [\ I(g_j; y) - (1/|S|)\ \Sigma_{g_i \in S}\, I(g_j; g_i)\ ]$$

where S is the already-selected set and Ω the candidate pool. Genes are added greedily, each chosen to maximise new information rather than to duplicate existing information. The result is a compact panel spanning multiple biological programmes rather than one deeply sampled module — which is both statistically more efficient and biologically more informative.

**Stage 3 — Cross-validated LASSO confirmation**

The final stage validates the selected panel within a predictive model rather than in isolation. L1-regularised regression minimises

$\min_{\beta} \{ (1/2N) \Sigma_i L(y_i, \beta^T x_i) + \lambda \Sigma_j |\beta_j| \}$

The L1 penalty drives coefficients of weakly contributing genes exactly to zero, and λ is chosen by cross-validation rather than assumed. Genes surviving this stage are those that are individually class-relevant, mutually non-redundant, and jointly predictive within a fitted model. A gene selected by the first two stages but eliminated here was relevant in isolation yet redundant once the panel was assembled — precisely the case that univariate screening cannot detect.

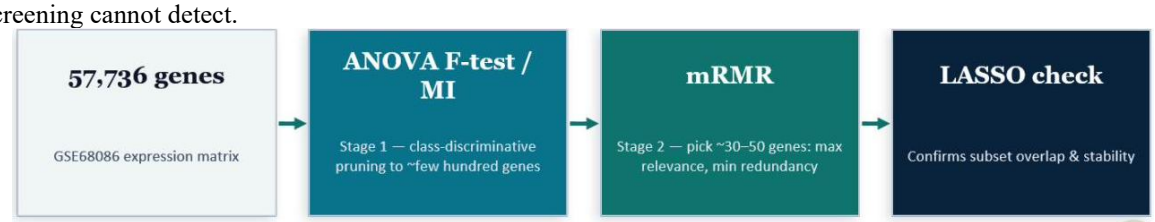


**Figure 4. Proposed gene-selection strategy combining ANOVA F-test or mutual information screening, mRMR redundancy filtering, and cross-validated LASSO confirmation.**

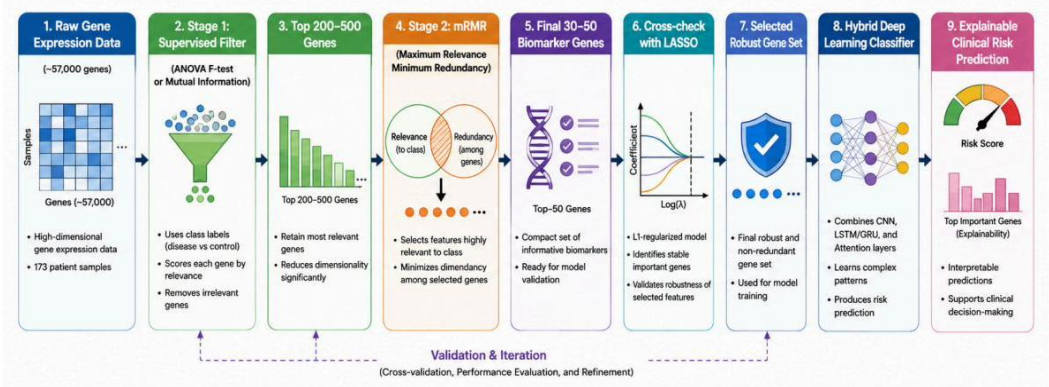


**Figure 5. Two-stage supervised feature-selection framework positioned within the end-to-end classification pipeline for high-dimensional genomic data.**

### 5.3 Justification and Leakage Control

The pipeline is class-aware at every stage, which directly repairs Defect 1: no gene enters the panel without demonstrating a relationship to the diagnostic label. It is redundancy-aware, which addresses the co-expression saturation endemic to expression data. It is composed entirely of deterministic operations with well-characterised behaviour at small sample sizes, so the selected panel is reproducible across runs and auditable. And it is buildable with standard tooling — SelectKBest, an mRMR implementation and cross-validated LASSO — with no custom metaheuristic requiring independent verification, which distinguishes it from [5], [9] and matters for a pipeline that must eventually be validated by parties other than its authors.

Equally important is where the pipeline is fitted. To repair Defect 2, all three selection stages, together with scaling and any class-balancing operation, are fitted exclusively on the training partition inside each cross-validation fold and then applied unchanged to the held-out fold. No information from held-out samples influences which genes are selected. Selection stability is quantified across folds by Jaccard overlap of the selected panels, so that panel reproducibility becomes a reported result rather than an assumption. This is a stricter protocol than the baseline and than most of the reviewed corpus, and it should be expected to yield lower nominal accuracy than leakage-permissive alternatives — a difference that reflects honest measurement rather than inferior method.

## 6. PROPOSED FRAMEWORK

### 6.1 Architecture Selection

Five candidate architectures were evaluated against the requirements established by the diagnostic analysis: capacity for local pattern extraction, capacity for global dependency modelling, absence of a spurious ordering prior over unordered gene features, overfitting risk at n = 173, and compatibility with a structured explanation layer.

**Table 7. Comparative evaluation of candidate deep learning architectures for clinical genomic risk assessment.**

| Candidate | Local patterns | Global gene interactions | Bias for unordered genes | Overfit risk at n = 173 | Explainability | Verdict |
|---|---|---|---|---|---|---|
| Single CNN (baseline) [4], [10], [11] | Yes | No | Weak | Medium | Low | Insufficient |
| CNN-LSTM / CNN-BiLSTM | Yes | Sequential only | Weak | High | Low | Fallback |
| 1D-CNN + self-attention | Yes | Yes | Strong | Medium | High | Selected |
| Transformer only [12] | No | Yes | Strong | Very high | High | Too data-hungry |
| Graph neural network [13] | Via graph | Yes | Strong | Medium | Medium | Future work |

The hybrid configuration is selected because it is the only candidate satisfying all five criteria simultaneously. A pure transformer would model global dependencies well but cannot be trained from scratch on 173 samples — GexBERT [12] demonstrates this by requiring pan-cancer pretraining to reach its reported performance. Recurrent hybrids impose a sequential ordering assumption that is no more biologically justified than the convolutional one. Graph neural networks are attractive in principle [13] but require a prior interaction graph whose construction introduces its own assumptions and error sources; they are retained as a defined extension rather than the initial configuration.

### 6.2 Hybrid CNN–Transformer Design

The classifier operates on the reduced panel of d genes produced by the selection pipeline, treating each selected gene as a token rather than as a position in an ordered sequence. A one-dimensional convolutional block first extracts local co-expression motifs and projects each gene into a learned embedding of dimension d_model. Because the panel has already been reduced from tens of thousands of transcripts to a few dozen, this block operates at a parameter budget compatible with the available sample count.

The resulting token embeddings are passed to a multi-head self-attention encoder, which computes, for each attention head,

$$\text{Attention}(Q, K, V) = \text{softmax}( QK^{T} / \sqrt{d_k} ) V$$

with queries, keys and values derived by learned projection from the gene embeddings. The decisive property is that every gene attends to every other gene in a single layer, at a cost independent of their separation in the input vector. This is exactly the capability that convolution structurally lacks and that Defect 3 identified as missing. Because gene order carries no meaning, positional encoding is deliberately omitted, making the encoder permutation-equivariant over the panel — the model cannot exploit annotation order because it cannot perceive it. Multiple heads allow distinct co-regulation patterns to be represented in parallel: one head may attend to a proliferation programme while another tracks an immune-infiltration signature. The attended representation is pooled and passed through a dense classification head to produce a six-way risk distribution. The complementary pairing of a convolutional local branch with a global attention branch has proved effective in adjacent biomedical classification domains [30], and the present design transfers that principle to unordered transcriptomic features.

Regularisation is treated as a first-class design element rather than an afterthought, given that the baseline failure was fundamentally one of overfitting. Attention and residual dropout, layer normalisation, weight decay, early stopping on a fold-internal validation split, and a deliberately shallow encoder depth are all applied. Model selection is performed under repeated stratified k-fold cross-validation with the generalisation gap reported per fold, so that a configuration achieving high mean accuracy with unstable fold-level behaviour is not preferred over a lower-accuracy configuration with consistent behaviour.

### 6.3 System Architecture

The predictive core is embedded within a layered system designed for clinical deployment rather than offline benchmarking. Figure 6 presents the complete architecture.

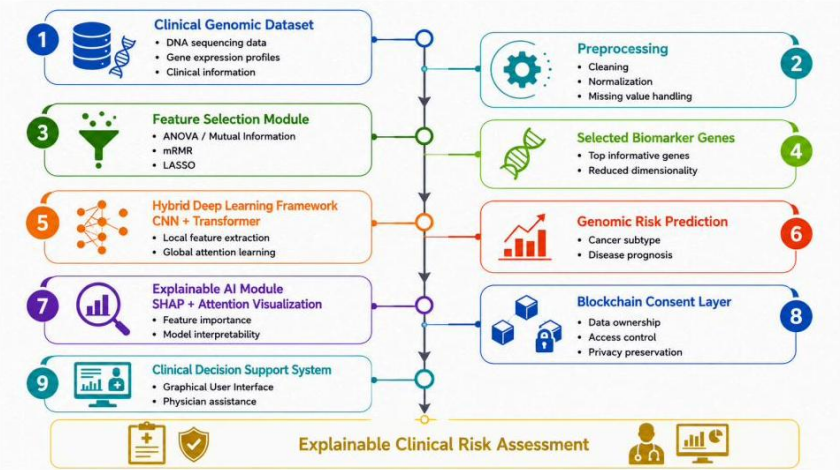


**Figure 6. Overall architecture of the proposed blockchain-enabled explainable hybrid deep learning framework for clinical genomic risk assessment.**

**Module 1 — Feature selection**

ANOVA F-test or mutual-information screening, mRMR redundancy filtering, and cross-validated LASSO confirmation, fitted fold-internally as specified in Section 5.

**Module 2 — Hybrid deep learning**

Convolutional local feature extraction followed by multi-head self-attention for long-range dependency modelling, terminating in multi-class risk classification.

**Module 3 — Explainable AI layer**

The explanation layer converts each prediction into a gene-level rationale a clinician can inspect, and it is designed rather than bolted on. SHAP attributions, as employed for biomarker identification in [15] and [18], provide per-patient, per-gene contribution values with an additive-consistency guarantee, answering the question of which transcripts drove this particular prediction. Attention maps extracted from the encoder provide a complementary model-internal view of which gene pairs the network treated as jointly informative, in the spirit of the pathway-guided attention of [13]. Aggregated biomarker importance across the cohort identifies transcripts that are consistently influential, supporting downstream biological validation against pathway databases [4], [11]. The two explanation sources are deliberately independent — post-hoc attribution and model-internal attention can disagree, and where they disagree the explanation is flagged as low-confidence rather than presented as authoritative. Explanation fidelity is itself evaluated, using deletion and insertion curves that measure whether removing highly attributed genes actually degrades the prediction.

**Module 4 — Blockchain governance layer**

The scope of this layer is defined narrowly and deliberately. Blockchain does not participate in the model computation, does not improve predictive accuracy, and makes no claim to do so; overstating its role is a common weakness in the literature and is avoided here. Its function is dynamic consent management: a patient grants or revokes permission for specific uses of their genomic record through a smart contract, and every access request is checked against the current on-chain consent state before data is released [20], [21]. Raw genomic sequences are never written on-chain — only hashes, access-policy state and immutable audit events — with the underlying records held in encrypted off-chain storage, following the hybrid on-chain/off-chain pattern established in [26] and [27] and the layer-2 scalability strategies analysed in [28]. The properties this delivers are consent granularity, revocability, tamper-evident access auditing and verifiable data provenance [22], [23], [25]. These are governance properties, not performance properties, and they are what makes the predictive system deployable against identifiable patient genomes [3], [19], [24].

**Module 5 — Clinical decision support system**

The interface layer presents the predicted risk distribution, the associated explanation panel, and the contributing gene set to a clinician in a form suited to a consultation rather than a research notebook. Critically, it is consent-aware: the interface executes only against data for which blockchain-granted authorisation is currently valid, and every inference is logged as an auditable access event. This closes the loop between the governance layer and the point of clinical use.

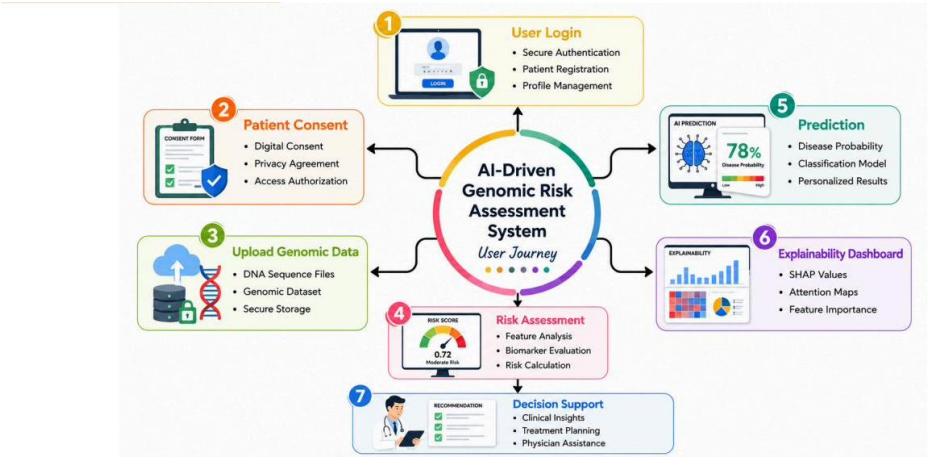


**Figure 7. Clinical decision support system user journey for explainable genomic risk assessment, from consent verification through prediction to explanation review.**

### 6.4 Progressive Validation Protocol

The framework is validated in defined stages rather than assessed as a monolith, so that any performance change can be attributed to a specific component. This protocol is stated in advance of experimentation.

• Stage A — Selection ablation. The three-stage supervised pipeline is compared against entropy ranking, PCA, univariate filtering alone and mRMR alone, holding the classifier fixed. Panel stability across folds is reported alongside accuracy.

• Stage B — Architecture ablation. The hybrid CNN–Transformer is compared against CNN-only, transformer-only and CNN-BiLSTM configurations, holding the selected panel fixed. The generalisation gap is the primary comparison metric.

• Stage C — Explanation validation. SHAP and attention explanations are compared for agreement, evaluated for fidelity via deletion and insertion curves, and assessed for biological plausibility against pathway-enrichment databases.

• Stage D — Governance integration. Consent enforcement is validated functionally through grant, revoke and audit scenarios, with smart-contract latency and throughput measured under representative access loads [28], [29].

• Stage E — External transfer. The frozen pipeline is applied to an independent cohort, with the cross-cohort accuracy drop reported as the definitive robustness measure.

## 7. COMPARISON WITH EXISTING STUDIES

Table 8 positions the present work against representative recent studies. The comparison is presented with an explicit caveat: accuracy figures across rows are not directly comparable, because the datasets differ in size, class count, class separability and validation rigour by margins large enough to dominate any methodological difference.

**Table 8. Comparison of the present work with representative recent genomic classification studies.**

| Study | Dataset | Method | Accuracy (%) | Generalisation reporting |
|---|---|---|---|---|
| KalaiSelvi et al. (2025) [4] | GEO expression + IoT biosensor data | Entropy FS + SVC/DT/MLP/CNN + PPI-KEGG | High (reported) | Not assessed — no train–test gap analysis |
| Zhang et al. — GONF (2025) [10] | TCGA, AHBA microarray | mRMR + deep CNN | 97 / 95 | Not assessed; single split; image-conversion overfitting risk |
| Shah et al. (2026) [11] | TCGA + ICGC (lung) | PCA + MI + CNN | 98 | Single cancer; stable curves but no cross-dataset gap analysis |
| Jiang & Hassanpour — GexBERT (2025) [12] | TCGA (pan-cancer) | Pre-trained transformer | ≈94–98 | Strong from limited gene subsets; needs large-scale pretraining; small-sample behaviour not analysed |
| Ameen et al. (2025) [17] | TCGA 32 types + CPTAC | Stacked multi-omics DL ensemble | 96–98 | Cross-omics validation; no train–test gap reported |

| Study | Dataset | Method | Accuracy (%) | Generalisation reporting |
|---|---|---|---|---|
| Al Marouf et al. (2025) [18] | Prostate microarray (102 samples) | Random forest + SHAP | 81.0 | Small cohort; single cancer; no generalisation-gap analysis |
| This work (baseline reproduction) | GSE68086 — TEP RNA-seq, 6 classes, 173 samples | Entropy FS + SVC/DT/MLP/CNN | 60.4 test / 98.1 train | Explicitly quantified: 18.3–45.3 pt gaps across four models |
| This work (proposed framework) | GSE68086, with defined external-cohort transfer | ANOVA/MI + mRMR + LASSO → CNN-Transformer + XAI + consent governance + clinical GUI | Under staged evaluation | Gap adopted as primary optimisation target; leakage-free fold-internal protocol |

The comparison exposes a consistent structural pattern in the literature. Reported accuracies cluster between 94 and 98 per cent [10]–[12], [17], but they are obtained predominantly on large pan-cancer cohorts or single-cancer binary tasks, and the train–test gap, AUC stability and class-wise robustness that determine real-world reliability are almost never reported. Where a study evaluates on a small, multi-class, transcriptionally overlapping cohort under a leakage-free protocol — as here, and as partially reflected in the small-cohort result of [18] — the achievable accuracy is substantially lower. That difference reflects evaluation honesty rather than methodological inferiority, and treating the two as comparable would be a category error.

The more meaningful comparison is the capability matrix of Table 2. No reviewed study compares multiple feature-selection strategies within a common harness; none combines convolutional local extraction with self-attention over a selected panel; none provides both a designed explanation layer and a clinician-facing interface; and none addresses consent governance at all. The contribution claimed here is integration under a rigorous evaluation protocol, not a higher number on an easier benchmark.

## 8. RESULTS AND DISCUSSION

The empirical results reported in this paper are those of the baseline reproduction, and their value is diagnostic rather than competitive. Three findings warrant discussion.

First, the convergence of all four classifiers to near-identical test accuracy despite training accuracies spanning 0.73 to 1.00 is more informative than any individual score. It indicates that the ceiling is imposed by the information content of the entropy-selected feature set and the sample-to-class ratio, not by the capacity of the classifiers. Adding capacity — moving from SVC to MLP to CNN — increases training accuracy by twenty-five points and test accuracy by six, which is the empirical signature of a representation bottleneck upstream of the classifier. This is the strongest available evidence that feature selection, not model complexity, is where the leverage lies, and it is consistent with the gains reported when class-aware selection is substituted for unsupervised ranking [10], [11].

Second, the divergence between accuracy and AUC is diagnostically important. A model with 60.4 per cent accuracy and 0.89 multi-class AUC is ranking samples correctly far more often than it is labelling them correctly. In a six-class problem with overlapping transcriptional programmes, this points to decision-threshold and calibration failure rather than absence of learnable signal, and it supports the expectation that better representation — a redundancy-free panel plus an architecture capable of modelling gene co-regulation [12], [13] — can convert existing ranking ability into classification accuracy. Had the AUC also been near chance, the appropriate conclusion would have been that platelet RNA cannot separate these malignancies at this sample size, and the entire architectural programme would have been unjustified.

Third, the leakage identified in the baseline protocol has implications beyond this study. Applying SMOTE before partitioning is a common pattern in the applied genomics literature, and where it occurs the reported test performance is optimistically biased by an amount that is rarely quantified. Since the baseline already exhibits gaps of 18 to 45 points under a leakage-permissive protocol, the true gaps are wider still. This strengthens rather than weakens the argument of this paper: the problem being addressed is larger than the published numbers suggest, and any subsequent comparison between the corrected pipeline and the baseline must hold the validation protocol constant to be meaningful.

A methodological point follows from these observations. The generalisation gap deserves routine reporting in high-dimensional biomedical machine learning, in the same way that confidence intervals are expected in clinical statistics. It is trivial to compute, it requires no additional experimentation, and it converts an unfalsifiable accuracy claim into an interpretable statement about model reliability. Its near-total absence from the reviewed corpus [4]–[18] is a reporting gap that the present work argues should close.

## 9. LIMITATIONS AND FUTURE SCOPE

### 9.1 Limitations

The limitations of this work are stated plainly, since a paper arguing for evaluation honesty cannot exempt itself.

• The proposed hybrid architecture, explanation layer, governance layer and clinical interface are specified and justified but not yet experimentally validated. The empirical content of this paper is confined to the baseline reproduction; the comparative results of Stages A through E remain to be produced.

• With 173 samples across six classes, statistical power is limited and confidence intervals on any accuracy estimate are correspondingly wide. Small differences between configurations may not be distinguishable from fold-level noise.

• Evaluation to date uses a single cohort. Transcriptional profiles are sensitive to acquisition platform, batch and population, so single-cohort performance is a weak predictor of cross-cohort performance. External validation is deferred to Stage E.

• Transcriptional overlap between biologically related malignancies places an irreducible ceiling on separability that no architecture can remove. Part of the observed confusion is biological rather than algorithmic.

• The blockchain layer is an architectural specification rather than a deployed implementation; its throughput, latency and cost characteristics under realistic clinical access patterns are not yet measured [28].

• Explanation quality has not been assessed with clinicians. Fidelity metrics measure whether an explanation reflects the model, not whether it is useful at the point of care, and only a user study can establish the latter.

### 9.2 Future Scope

Work proceeds along five defined lines. The immediate priority is executing the selection and architecture ablations of Stages A and B, since these determine whether the central hypothesis — that supervised, redundancy-aware selection combined with attention-based dependency modelling narrows the generalisation gap — holds empirically. Second, external validation on an independent cohort will establish whether the pipeline transfers, with the cross-cohort accuracy drop reported as the definitive robustness measure. Third, the explanation layer will be validated for fidelity and for biological plausibility against pathway-enrichment analysis [4], [11], and subsequently assessed with clinical users. Fourth, the consent layer will move from specification to a working smart-contract implementation with measured performance under representative access loads [26]–[28]. Fifth, the graph neural network configuration deferred in Section 6.1 will be revisited once prior interaction graphs of sufficient quality can be constructed, since incorporating known protein–protein interaction structure is a principled way to inject biological prior knowledge that attention must otherwise learn from limited data [13].

A longer-term direction concerns federated evaluation. The governance layer developed here manages consent for centralised access, but the same architecture extends naturally to federated training in which models travel to institutional data rather than the reverse [29]. Combining consent-governed federation with the explanation layer would address the multi-centre generalisation problem and the privacy problem through a single mechanism, and represents the most substantive extension of the present framework.

## 10. CONCLUSION

This paper has traced a complete methodological progression from measured failure to designed remedy in clinical genomic risk assessment. The baseline reproduction applied a published pipeline [4] to the GSE68086 tumour-educated-platelet benchmark and established, across four classifier families, that high-dimensional genomic classification fails in a specific and characterisable way: generalisation gaps of 18.3 to 45.3 percentage points, a best test accuracy of 60.4 per cent against 98.1 per cent training accuracy, and a multi-class AUC of 0.89 indicating that real discriminative signal exists but is not being converted into reliable classification.

The contribution of that experiment is not the numbers themselves but the attribution. Three correctable defects were identified: unsupervised entropy-based gene ranking that optimises variability rather than class separability; synthetic oversampling applied before partitioning, which compromises test-set independence and biases reported performance optimistically; and a convolution-only inductive bias that imposes a meaningless ordering prior over gene features while structurally excluding the long-range co-regulation relationships that define transcriptional programmes.

Each defect is answered by a specific design decision. A three-stage supervised selection pipeline — ANOVA F-test or mutual-information screening, mRMR redundancy filtering, cross-validated LASSO confirmation — replaces entropy ranking, selected on reproducibility and small-sample stability grounds after explicit comparison of filter, wrapper, embedded and hybrid families [5]–[11]. Fold-internal fitting of every preprocessing operation eliminates the leakage path. A hybrid CNN–Transformer classifier, chosen over four competing candidates, couples convolutional local motif extraction with permutation-equivariant multi-head self-attention over the selected gene panel, providing exactly the global dependency modelling that convolution cannot supply [12], [13], [30] at a parameter budget the sample size can support.

Around this predictive core, the framework adds a designed explanation layer combining SHAP attribution with model-internal attention [15], [18], a blockchain smart-contract layer providing dynamic patient consent and immutable access auditing with a deliberately narrow and honestly stated scope [19]–[28], and a consent-aware clinical decision-support interface. A systematic review of fifteen studies published in indexed journals during

2025 and 2026 confirms that no existing work integrates comparative feature selection, local and global dependency modelling, structured explainability, clinical interfacing and consent governance within one system. The framework awaits the staged experimental validation defined in Section 6.4, and this paper makes no claim to have demonstrated its performance. What it does establish is that every component of the proposed architecture is traceable to a measured failure rather than to methodological fashion, and that the design is falsifiable through a validation protocol specified in advance. The broader argument is methodological: in a domain where a model can reach 98 per cent training accuracy and 60 per cent test accuracy on the same data, the generalisation gap belongs in every results table. Its absence from all fifteen reviewed studies is a reporting convention that clinically oriented genomic machine learning can no longer afford.

## DECLARATIONS

**Data availability.** The GSE68086 dataset is publicly available from the NCBI Gene Expression Omnibus at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68086. No new patient data were generated in this study.

**Conflict of interest.** The authors declare no competing interests.

**Funding.** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Ethical approval.** This study analysed only publicly available, de-identified datasets and did not involve human participants or animals directly; institutional ethical approval was therefore not required.

## REFERENCES

[1] National Genomics Data Center Members and Partners. (2025). Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2025. Nucleic Acids Research, 53(D1), D30–D44.
[2] Jamalinia, M., & Weiskirchen, R. (2025). Advances in personalized medicine: Translating genomic insights into targeted therapies for cancer treatment. Annals of Translational Medicine, 13(2), 18.
[3] Annan, R., Noland, J., Perkins, K., Yuan, X., Roy, K., & Qingge, L. (2025). Genomic privacy and security in the era of artificial intelligence and quantum computing. Discover Computing, 28(1), 108.
[4] KalaiSelvi, B., Babu, R. G., Cho, J., & Yoo, S. (2025). Gene expression profiling and predictive modeling of cancer biomarkers using machine learning and IoT-enabled biosensors. Scientific Reports, 15. https://doi.org/10.1038/s41598-025-30366-x
[5] Dhamercherla, S., Reddy Edla, D., & Dara, S. (2025). Cancer classification in high-dimensional microarray gene expressions by feature selection using eagle prey optimization. Frontiers in Genetics, 16, 1528810. https://doi.org/10.3389/fgene.2025.1528810
[6] Yaqoob, A., Verma, N. K., & Aziz, R. M. (2025). Improving breast cancer classification with mRMR + SSO + WSVM: A hybrid approach. Multimedia Tools and Applications, 84, 26041–26066. https://doi.org/10.1007/s11042-024-20146-6
[7] Cherif, C., Maiza, M., & Taleb-Ahmed, A. (2026). Social spider optimization with mutual information for high-accuracy cancer classification: A hybrid gene selection framework. Iranian Journal of Computer Science, 9, 4. https://doi.org/10.1007/s42044-025-00372-3
[8] Shi, Y., Zheng, Y., & Bai, X. (2025). A multiple filter-wrapper feature selection algorithm based on process optimization mechanism for high-dimensional omics data analysis. PLOS ONE, 20(12), e0338051. https://doi.org/10.1371/journal.pone.0338051
[9] Sahu, B., Panigrahi, A., Pati, A., Das, M. N., Jain, P., Sahoo, G., & Liu, H. (2025). Novel hybrid feature selection using binary Portia spider optimization algorithm and fast mRMR. Bioengineering, 12(3), 291. https://doi.org/10.3390/bioengineering12030291
[10] Zhang, Y., Chen, J., & Zhang, C. (2025). Integrating image processing with deep convolutional neural networks for gene selection and cancer classification using microarray data. Scientific Reports, 15(1), 39465. https://doi.org/10.1038/s41598-025-23147-z
[11] Shah, S. N. A., Issar, K., & Parveen, R. (2026). A hybrid feature extraction framework combining PCA and mutual information for gene expression-based lung cancer classification. PLOS ONE, 21(2), e0342160. https://doi.org/10.1371/journal.pone.0342160
[12] Jiang, S., & Hassanpour, S. (2025). Transformer-based representation learning for robust gene expression modeling and cancer prognosis. Scientific Reports, 15(1), 37581. https://doi.org/10.1038/s41598-025-14949-2
[13] EL kati, Y., Wang, S., & Ali, T. A. A. (2026). Hybrid GNN-Transformer model for multi-omic cancer classification with interpretable pathway-driven feature selection. PeerJ Computer Science, 12, e3415. https://doi.org/10.7717/peerj-cs.3415
[14] Li, C., Yan, Y., Lin, W., et al. (2025). Enhancing cancer subtype classification through convolutional neural networks: A DeepInsight analysis of TCGA gene expression data. Health Information Science and Systems, 13, 33. https://doi.org/10.1007/s13755-025-00349-3

[15] Reddy, C. K., Kaza, V. S., Daduvy, A., Shuaib, M., Alshanketi, F., & Alam, S. (2025). MoX: An explainable hybrid deep-learning model for integrating multi-omics data to predict event-free survival in neuroblastoma prognosis. Journal of King Saud University – Computer and Information Sciences, 37(9). https://doi.org/10.1007/s44443-025-00003-8
[16] Zeng, Y., Zhang, Y., Xiao, Z., & Sui, H. (2025). A multi-classification deep neural network for cancer type identification from high-dimension, small-sample and imbalanced gene microarray data. Scientific Reports, 15(1), 5239. https://doi.org/10.1038/s41598-025-89475-2
[17] Ameen, A., Alganmi, N., & Bajnaid, N. (2025). Stacked deep learning ensemble for multiomics cancer type classification: Development and validation study. JMIR Bioinformatics and Biotechnology, 6, e70709. https://doi.org/10.2196/70709
[18] Al Marouf, A., Rokne, J. G., & Alhajj, R. (2025). Identification of potential biomarkers in prostate cancer microarray gene expression leveraging explainable machine learning classifiers. Cancers, 17(23), 3853. https://doi.org/10.3390/cancers17233853
[19] Ahmed, S. A., & Hrzic, R. (2025). Blockchain and homomorphic encryption for genomic and health data sharing: An ethical perspective. Ethics, Medicine and Public Health, 33, 101127.
[20] Nguyen, T. T. A., Hsieh, Y. H., Tseng, C. H., Lin, Y. C., & Yuan, S. M. (2025). Blockchain-enabled privacy-preserving ecosystem for DNA sequence sharing. Applied Sciences, 15(6), 3193. https://doi.org/10.3390/app15063193
[21] Zhang, J., Zhou, Y., Ren, Y., Au, M. H., Chow, K. H., Chen, L., & Luo, R. (2025). Toward owner governance in genomic data privacy with Governome. Cell Reports Methods, 5(9).
[22] Morsi, S. A., Husain, H. I., Nair, R., Alkahtani, H., Abdulsahib, G. M., & Aldhyani, T. H. (2026). A privacy-preserving blockchain and machine learning framework for secure next-generation genomic data analysis. International Journal of Advances in Soft Computing and Applications, 18(1).
[23] Li, K., Lohachab, A., Dumontier, M., & Urovi, V. (2025). Privacy preservation in blockchain-based healthcare data sharing: A systematic review. Peer-to-Peer Networking and Applications, 18(6), 1–53.
[24] Kasralikar, P., Polu, O. R., Chamarthi, B., Rupavath, R. V. S. S. B., Patel, S., & Tumati, R. (2025). Blockchain for securing AI-driven healthcare systems: A systematic review and future research perspectives. Cureus, 17(4).
[25] Musamih, A., Salah, K., Jayaraman, R., Ellahham, S., Omar, M., & Yaqoob, I. (2025). Blockchain and NFT-based solution for genomic data management, sharing, and monetization. IEEE Access.
[26] Mohanta, B. K., Awad, A. I., Dehury, M. K., Mohapatra, H., & Khan, M. K. (2025). Protecting IoT-enabled healthcare data at the edge: Integrating blockchain, AES, and off-chain decentralized storage. IEEE Internet of Things Journal, 12(11), 15333–15347.
[27] Tiwari, K., & Kumar, S. (2025). A healthcare data management system: Blockchain-enabled IPFS providing algorithmic solutions for increased privacy-preserving scalability and interoperability. The Journal of Supercomputing, 81(8), 895.
[28] Vo, K. T., Nguyen, H. T., & Nguyen-Hoang, T. A. (2025). An integrated decision-making framework for enhancing blockchain scalability and efficiency: Layer-2, off-chain storage, and smart contracts. IEEE Access.
[29] Bhardwaj, T., & Sumangali, K. (2025). A federated incremental blockchain framework with privacy-preserving XAI optimization for securing healthcare data. Scientific Reports, 15. https://doi.org/10.1038/s41598-025-21852-3
[30] Wang, X., Wang, G., Li, L., Zou, H., & Cui, J. (2025). MFF-ClassificationNet: CNN-Transformer hybrid with multi-feature fusion for breast cancer histopathology classification. Biosensors, 15(11), 718. https://doi.org/10.3390/bios15110718