PAMM: Pathway-Aware Masked Representation Learning for Interpretable Multi-Cancer Prediction

by

Chandrima Roy Chowdhury
22101246
Zarrin Tasnim Rodoshi
22101580
Sumaiya Hossain Surovi
24241338
Labib Hasan
22101579

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
January 2026.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name and Signature:

Chandrima Roy Chowdhury
22101246

Zarrin Tasnim Rodoshi
22101580

Sumaiya Hossain Surovi
24241338

Labib Hasan
22101579

Approval

The thesis titled "PAMM: Pathway-Aware Masked Representation Learning for Interpretable Multi-Cancer Prediction" submitted by

1. Chandrima Roy Chowdhury (22101246)
2. Zarrin Tasnim Rodoshi (22101580)
3. Sumaiya Hossain Surovi (24241338)
4. Labib Hasan (22101579)

of Fall, 2026 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in 28 January, 2026.

Examining Committee:

Supervisor:
(Member)

Amitabha Chakrabarty, PhD

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

The Chairperson
Department of Computer Science and Engineering
BRAC University

Abstract

In this thesis, PAMM, a new paradigm of interpretable multi-cancer prediction based on Pathway-Aware Masked Representation Learning is introduced. To tackle the challenge of the 'Small n, Large p' of transcriptomics it is our holding that we apply the rigorous seven-stage pipeline of preprocessing (i.e. Log2 transform, ANOVA filter, Lasso regularization and Recursive Feature Elimination) to reduce the original high-noise 57,750 genes in Breast, Lung, GBM, and HC samples to a high-signal feature set. The basic architecture goes beyond the usual deep learning of black boxes by incorporating biologically relevant priors of KEGG_2021_Human library in a self-supervised masking scheme. In contrast to stochastic masking, the pretraining phase of PAMM uses a Pathway-Aware Masking logic where complete sets of functional genes are zeroed, requiring the model to recreate missing biological units and learn complicated inter-pathway relationships. The latent representations of the model are optimized with Optuna, and the statistical robustness is verified with twenty independent iterations, and the latent representation is further interpreted with Single-sample Gene Set Enrichment Analysis (ssGSEA). The resulting visualizations of mean pathway activity indicate that PAMM is able to capture different, clinically viable biological signatures of each cancer type. PAMM provides a clear and very precise diagnostics platform of precision oncology by filling the gap between high-dimensional self-supervised learning and functional biology. Along with closed-set multi-cancer, PAMM is also explicitly tailored to open-set recognition. Through a combined study of softmax confidence and latent space distances from class centroids, the framework can discard samples that do not adhere to any known cancer manifold. This allows the certainty of identifying unknown or non-cancerous gene expression patterns, which is very essential when it comes to a real-life clinical implementation in which unobservable conditions are the norm. This two-fold feature sets PAMM apart from the traditional classifiers and guarantees the accuracy of the diagnosis and its safety.

Keywords: Masked Representation Learning, Pathway-Awareness, Multi-Cancer Prediction, Interpretable AI, KEGG Pathways, RFECV, ssGSEA.

Acknowledgement

We would first like to express our deepest gratitude to the Almighty, whose infinite grace, strength, and wisdom have guided us throughout every stage of this endeavor. We are profoundly thankful to our esteemed supervisor, Dr. Amitabha Chakrabarty, whose exceptional guidance, insightful advice, and unwavering support for this research.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

With the advent of High-Throughput sequencing (HTS) and Next-Generation sequencing (NGS), the face of oncology has been completely transformed as the diagnostic paradigm has switched to be based on the high-resolution molecular profiling rather than traditional morphological assessment. The entire study of transcriptomes, known as transcriptomics, has offered a dynamic, functional picture of cell activity as it can be quantified by measuring the level of RNA expression at a given moment across the entire genome. The transcriptome, in comparison to the comparatively stable genome, indicates the instantaneous reaction to oncogenic changes, environmental or therapeutic treatment by the cell. This molecular information is central in determining the momentum behind tumorigenesis, metastasis, therapeutic resistance and can shed light on the regulatory networks unavailable with the DNA sequencing.
The data provided by these technologies is, however, typified by extremely high-dimensionality, which tends to represent the expression of more than 50,000 transcripts (features) in a relatively small number of clinical samples. This results in the 'Small n, Large p' Problem which poses a major obstacle to the conventional statistical techniques. Although the initial transcriptomics research took into consideration the individual differentially expressed genes (DEGs) in a vacuum, the current computational biology is interested in studying it as a system that is dynamic and interdependent. To leverage this complexity effectively, complex analytical models are needed that are capable of traversing the huge feature space to draw out effective, multi-gene biomarkers. It is aimed at transitioning to a systems-biology strategy in which multi-cancer predictions are made based on the behavior of gene modules, and not on an isolated set of markers.

1.2 Motivation

The black-box character of conventional Deep Learning (DL) models is a significant impediment to a broad clinical audience in the present-day age of precision medicine. Although such architectures as Convolutional Neural Networks (CNNs) and standard Multi-Layer Perceptrons (MLPs) have reached state-of-the-art accuracy in most tasks, the process of their decision-making in bioinformatics is often opaque, with no biological explanation given why they make the prediction. A

diagnosis that does not have a reason behind it cannot be of much use in a clinical setting; the clinician needs to be aware of the biological process that is being interrupted to be able to prescribe a specific therapy.
The fact that current cancer classification frameworks assume a closed-world environment, in which every sample of the test would be compelled to belong to one of the established cancer types, is equally a significant weakness of the current methods. In actual clinical practice, however, patient samples can be of rare, mixed, or hitherto unobserved cancer subtypes. A model that is not capable of identifying such cases of unknown or out-of-distribution will assume predictions that are overconfident, but unacceptable clinically. This fosters the requirement of an inbuilt unknowing recognition system that can postpone the uncertain cases to be examined by an expert as opposed to implementing untrustworthy classifications.
The necessity of Physician-in-the-loop AI, where models attain high predictive accuracy as well as profound biological insight, is the motivation behind this research. The concept of Self-Supervised Learning (SSL) (or, more precisely, the idea of Masked Representation Learning that has become the sensation of Natural Language Processing (NLP)) can be applied to transcriptomics to induce models to acquire the hidden grammar of biology. The reason is to come up with a model that does not merely view numbers but rather learns the functional and hierarchical interactions of the genes and pathways. We make the network learn the patterns of co-expression and regulatory logic that exist in human cells by simply obscuring whole biological units and asking the model to re-discover them. This shift in approach to data-driven to knowledge-guided AI is necessary to develop diagnostic instruments that are clear, credible, and biologically justified.

1.3 Problem Statement

The two main problems that afflict computational transcriptomics are the curse of dimensionality and the interpretability gap. The number of features ($p \approx 57000$) in comparison to the number of samples (n) can readily cause disastrous overfitting, in which models do not learn predictable biological signals but instead retrieve noise. Moreover, the majority of currently available machine learning models assume that gene expression is high-dimensional, flat, and independent, and do not at all consider the hierarchical nature of biological systems, in which the genes function as components of a well-coordinated system, called pathway. Random masking can usually perturb these biological modules when used as a standard representation learning method, causing a latent representation that is mathematically optimal but biologically incoherent. It is of great importance that a methodology is developed that applies rigorous statistical filtering of the dimensionality reduction as well as making it use pathway-aware constraints that make sure that internal logic of the model is reflective of the biological processes in the real world.
Moreover, the majority of currently available transcriptomic classification systems have a rigid closed-set assumption and do not have the means of detecting unknown classes. Those types of models cannot tell whether the known type of cancer is really known or biologically new or out-of-distribution. Such lack of awareness of uncertainty is a critical threat in high stakes medical decision-making, and models that are capable of explicitly defining and marking unknown cancer profiles are needed.

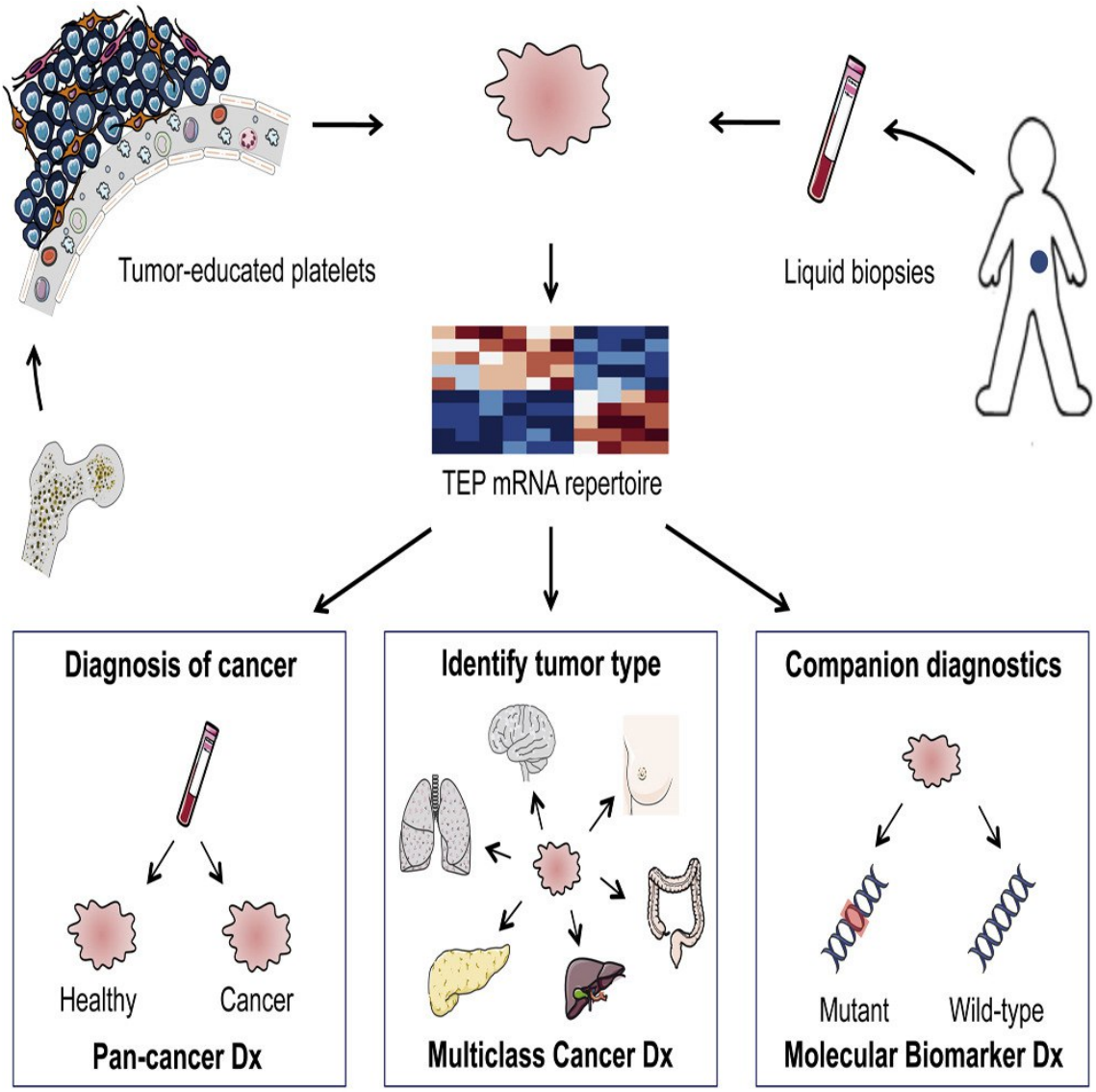

Figure 1.1: Tumor-Educated Platelets (TEP) for cancer detection and classification [3]

1.4 Objectives

The overall purpose of this thesis is to develop, deploy and test the PAMM (Pathway-Aware Masked Representation Learning) model of multi-cancer classification. In order to do this, the following specific objectives are set:

- Feature Engineering: To build a multi-tiered preprocessing and dimensionality reduction funnel which employs ANOVA, Log2 Fold Change (LFC), Lasso (L1) regularization and Recursive Feature Elimination (RFECV) to derive a high-signal subset of genes.
- Architectural Innovation: To design a self-supervised encoder-decoder system where stochastic masking is substituted with Pathway-Aware Masking where KEGG_2021_Human database is used as a biological prior.

- Optimization and Robustness: To perform Bayesian optimization (using Optuna) to tune hyperparameter space and perform statistical stability validation on the model by running it many times ($N = 20$) to perform extensive cross-validation.

- Interpretability Validation: To apply Single-sample Gene Set Enrichment Analysis (ssGSEA) to interpret the latent space of the model, it is necessary to ensure that the top-ranking pathways that the model found are correlated with the established pathology of Breast, Lung, GBM, and HC classes.

- Unknown Class Detection: To develop a latent-space-based uncertainty assessment plan that allows detecting unseen or out-of-distribution cancer samples and, therefore, go beyond the traditional closed-set classification.

1.5 Methodology in Brief

The research methodology can be described as a complex funnel of analytical reduction of noise and maintenance of biological complexity in two stages: the initial preprocessing pipeline and the implementation of the PAMM model. The steps include Structural Data Alignment, in which raw transcriptomic matrices undergo transposition and the Log2 transformation ($\log_2(x+1)$). The critical transformation is essential; it also controls the substantial skew and heteroscedasticity of RNA-seq counts data, which in turn successfully levels the variance around the mean and permits an interpretation of multiplicative differences as additive fold-changes. Each gene feature is then subject to Z-score Standardization to make sure that the next Lasso (L1) Regularization will be penalizing features according to their predictive value and not their original scale of expression or unit variance.
The feature selection stage uses a hybrid approach to fighting the dimensionality crisis. First, ANOVA and Log2 Fold Change (LFC) filters are a high-pass statistical sieve that eliminates genes which do not have significant variance or biological size across the Breast, Lung, GBM and HC classes. This is further optimized by Recursive Feature Elimination (RFECV) that recursively removes features to determine the absolute minimum set of genes that maximize cross-validation accuracy.
The essence of the research is the PAMM structure that is based on a two-stage training algorithm. During the Self-Supervised Pretraining stage, the model uses a self-supervised Pathway-Aware Masking mechanism using the KEGG_2021_Human library. The model learns the complex, non-linear relationship among the various pathways by being compelled to recreate the information missing in some modules of its functionality by relying on the rest of the context. Lastly, the model is optimized through the use of Optuna Bayesian search and transferred to a supervised Finetuning stage to predict 4-class cancers. In order to achieve complete transparency, the results are put through the validation of post-hoc through ssGSEA, which is a technique that localizes the latent activations of the model on to visible biological enrichment scores, validating that the internal representations of the model makes sense as far as known oncology. In order to do the open-set recognition, the unknown samples are recognized on the basis of latent-space distance and reconstruction confidence deviations to enable the model to reject those samples which do not correspond with representations of the learned cancer-specific representations.

1.6 Scopes and Challenges

- Taxonomic Scope: The study is narrowed down to the categorization of four major classes—Breast Cancer, Lung Cancer, Glioblastoma (GBM), and HC—to show the universality of the framework in different tissue types and organ systems.

- Biological Prior Dependency: The KEGG_2021_Human database is a fundamentally comprehensive library to be used as the biological prior dependency in the model; although it is one of the most widely-used libraries, the ultimate interpretability of this particular biological reference is inherently constrained by the current body of knowledge.

- Data Sparsity and Thresholding Challenges: This is a major challenge due to the problem of Small n and Large p. The resultant biological pathways are also likely to have been filtered so that certain pathways no longer have any representative at all, which requires a maximum threshold of `MIN_GENES_PER_PATHWAY` to ensure that the model does not attempt to learn statistically insignificant or under-represented clusters of genes.

- Cross-Database Nomenclature Mapping: The technical requirement that IDs of Ensembl (ENSG) genes be mapped to the Human Gene Symbols requires the introduction of the risk of data attrition, since any unmapped or synonymous gene must be omitted or hand-aggregated to preserve the integrity of the indices of the pathway to the gene.

- Setting the Masking Ratio: This is a fine architectural balance; masking too few pathways results in the trivial reconstruction problem and masking too many results in the model being unable to find a sufficient number of contextual anchors to successfully reconstruct the lost biological information.

- Computational Intensity: The proposed algorithm is highly resource-demanding and will need high-performance clusters of GPUs to conduct the Optuna Bayesian search and the 20-run statistical testing loop, which is needed to demonstrate that the accuracy of the model is not due to mere chance.

Chapter 2

Literature Review

2.1 Preliminaries

Cancer genomics is also becoming characterized by the incorporation of multi-omics information, such as transcriptomics, DNA-sequence, and clinical imaging, into sophisticated computational models that are aimed at overcoming the traditional statistical constraints. These foundations create a pyramid of approaches beginning with supervised learning to classify (e.g., CNNs, SVMs) and unsupervised clustering to subtype find (e.g., K-means, Autoencoders) to state-of-the-art self-supervised and generative models (e.g., MAEs, GANs) to solve the problem of data scarcity and high dimensionality. The most important aspect of these reviews is that they have shifted to pathway-informed learning, which makes use of the topological interactions among genes to enhance model interpretability and clinical validity. This synthesis offers a technical basis on which deep learning can shift between the hypothetical high performance on the one hand and practical, personalized cancer prognosis and therapy prediction to the other with a standardized metric of performance, including the Area Under the Curve (AUC) and the Concordance Index (C-index).

2.2 Review of Existing Research

2.2.1 Traditional Machine Learning

Kourou et al. (2015) [4], discusses the applications of machine learning (ML) in cancer prognosis using clinical, genomic, imaging, and demographic data. Models such as ANNs, SVMs, decision trees, and Bayesian networks enhance early detection and personalized medicine beyond the performance of traditional methods. Data integration and feature selection improve the accuracy but suffer from small sample sizes, heterogeneous data, and overfitting issues. While ML can enable individualized treatment plans and better predictions, model interpretability, model validation, and model complexity issues are limiting its clinical adoption.
In terms of machine learning approaches, for genetic data, Kim et al. (2018) [7] propose the method that computes the risks for 20 kinds of common cancers. The results from the genetic difference of 5,919 people, 33%-88% are contributed by genetics; therefore, the rest is lifestyle issues. The applied model k-nearest neighbor (KNN), finds the possibility of the development of a certain type of cancer in some-

one's body. The workflow of data gathering, training, and validation of the KNN model goes ahead with the prediction against known results. Though the accuracy may be different for each type of cancer, it might help in early, personalized treatment.

A research study conducted by Malebary et al. (2021) [15], on the application of ML for driver gene predictions in cancer. Those genes are majorly responsible for causing cancer. In this study, they used random forest (RF), support vector machine (SVM), and neural networks (NN) to analyze genetic data. They first developed a dataset combining both cancerous and non-cancerous driver genes and then simplified it. RF had the best performance with accuracy rates of 91.08%, 87.26%, and 92.48% in different tests. It helps to differ between the driver genes causing cancer and passenger mutations that do not. These will help in targeted therapy against cancer. However, the method depends on the quality of the data and has difficulties with very complex genetic patterns.

There's another study by Sakib et al. (2021) [18] that focuses on the usage of machine learning (ML) algorithms for classification on gene expression data, where the objective is to distinguish between Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). Six models were used in the research study, Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, Decision Tree, K-Nearest Neighbors, and Random Forest. Preprocessing was done using normalization, PCA to reduce the complexity, and SMOTE to balance the class. SVM, Logistic Regression, and Naïve Bayes had the top results, for which 100% accuracy was the primary benefit of this study. The significant drawback is that the dataset size is small and generalizability is therefore poor in larger or more diverse populations.

The following paper by Mirabnahrazam et al. (2022) [21], investigates the use of ML for predicting DAT (dementia of Alzheimer's type) progression by integrating MRI and genetic data. DAT scores were estimated by VBpMKL, and feature selection was performed using Fisher's Exact Test, Welch's t-Test, and LASSO. The integration of MRI and genetic data improved classification, with increased early detection and precision medicine, although the approach is burdened by complexities in data integration and limited generalizability.

A related work by Quazi (2022) [22] focuses on the role of machine learning in precision and genomic medicine. This research uses machine learning models comprising logistic regression, random forests, deep learning, and support vector machines (SVM). These models have been useful in treating cancers including lung, breast, and prostate cancer by helping in diagnosis with efficiency greater than 90%, detection of treatment responses with efficiency ranging from 70-85%, and biomarker identification. This technique combines the data of proteomics and genomes, or "omics" data, for better understanding of diseases and treatments. The CNNs have emerged as very powerful in distinguishing between the subtypes of lung cancer and investigating melanoma images, sometimes even outperforming human experts. Better diagnostic precision, personalized treatment, and speedy drug discovery are some of the advantages of this. However, overfitting, dependency on data, and limited clinical application are some of its drawbacks.

Egor Revkov has presented the machine learning model for evaluating tumor purity using bulk gene expression data. The proposed method, PAN-Cancer Tumor PURITY Estimation Using a supervised learning approach, PUREE, ranks and transforms the gene expression data parallel to the genomic consensus labels to gen-

erate tumor purity estimates for 20 types of solid tumors. PUREE was trained and validated on disjoint datasets comprising more than 8,000 tumor samples from The Cancer Genome Atlas, outperforming previous approaches through the application of feature selection methods like Lasso regression. PUREE resolves the variances of cancer type and tumor purity and thus gives the final prediction by linear regression. Advantages include the fact that the PUREE method increases the accuracy in tumor purity estimation, is resilient across several cancer types, and easily extends to previously unknown forms of cancer. Work considerably improves the knowledge of tumor microenvironments and supports clinical research in cancer detection from tumor specimens. Some of the disadvantages include rigorous computational training, difficulties with low purity samples, and extra validation on larger datasets to raise therapeutic usefulness [23].
Early prediction of lung cancer using machine learning techniques by analyzing gene expression from an eco-genomics perspective, is shown in another paper [10]. It gives an overview of microarray gene expression data and some machine learning classifiers like Random Subspace, Multilayer Perceptron (MLP), and Sequential Minimal Optimization (SMO). Microarray technology is highlighted for its capability to measure the expression of thousands of genes simultaneously. The dataset used, includes 86 lung adenocarcinoma samples and 10 non-neoplastic lung samples, sourced from the Kent Ridge Bio-Medical Dataset Repository. Key findings from here are that six significant genes including FABP4 and FHL1 were identified, and these are biologically related to lung cancer. Early detection, identification of biomarkers of cancerous genes, and information regarding interaction of genes with the environment are advantages. Again, data reduction is also observed to improve the performance of models. The limitation of datasets, computational complexity of advanced classifiers like SMO, and a small sample size resulting in overfitting are the disadvantages.
It can be added that Iglesias et al. (2024) [38] drew more attention to unsupervised learning for cancer detection based on RNA-seq gene expression data. Here, PCA, UMAP, and clustering methods such as k-means, DBSCAN, and Gaussian Mixture models have been used. The research collated data from The Cancer Genome Atlas (TCGA) and focused their work on five types of cancers: breast, colon, kidney, lung, and prostate. It also mainly focuses on unsupervised learning in finding hidden cancer patterns, its various potential in cancer detection, and its use for developing personalized treatment. Also, there are certain challenges like data complexity, wrong detection, and computational cost. By means of gene expression data across several cancer types, including breast, lung, kidney, and ovarian cancers, this work explores the application of machine learning models for cancer identification. Work including subtype identification, cancer classification, biomarker development, and outcome prediction is included in this study. Cancer subtypes are categorized using models including Random Forest and Naïve Bayes, therefore obtaining an accuracy of 0.89 for lung cancer subtypes.

2.2.2 Deep Learning

Sun et al. (2019) [11], contributed with deep learning through the Genome Deep Learning(GDL) approach for the classification of 12 types of cancers from whole exon sequencing data. Deep neural networks (DNNs) were applied which translates

genomic mutations from TCGA and the 1000 Genomes Project to make a distinction between cancerous and noncancerous tissues. The percentage performance of the single models ranged from 97.47% to 100%, while for the composite model, it reached 70.08%. One of the big advantages of this method is that it can find the cancer in the early stage without the need for invasive testing. However, it is hard to tell the difference between various types of cancer.

A thesis by Ali (2020) [9] compared two deep learning models for leukemia detection, one from genomic sequencing and another from blood smear image analysis. Both models used Convolutional Neural Networks. The genomic model was designed in such a way to detect cancer markers in DNA sequences (accuracy - 98%), while the image-based model identified abnormal white blood cells (accuracy - 81%). The genomic model uses one-hot-encoded DNA data, while the blood smear model uses images, whose sizes are altered and also enhanced. Several tools used in the analysis include TensorFlow, Keras, and Scikit-learn. Some of the challenges which are faced by the models are small datasets, computationally expensive genomic preprocessing, and intense computation requirements. For accurate detection, genomic sequencing was highly required but costly, while image-based models were cheap and acceptable for low-resource environments at the cost of less accuracy.

Another related study by Tabares-Soto et al. (2020) [14] compared deep learning and machine learning algorithms for the classification of cancer types using the "11 Tumors" gene expression data. For this dataset, the accuracy of the supervised methods tried in the study-Convolutional Neural Networks (CNNs) and Logistic Regression (LR)-was 94.43% and 90.6%, respectively. Unsupervised methods such as K-means clustering were also tried but found insufficient. The dataset includes information about 11 types of cancers such as breast, liver, and lung with a total of 12,533 features and 174 samples. Dimensionality reduction techniques such as PCA have been performed along with scaling to reduce the problem of high dimensionality and small data. Major advantages are accurate classification of cancers for personalized treatment. Limitations include a small dataset size, high computation cost, overfitting.

Using multi-omics data, Olivier B. Poirion et al. [16] presented a hybrid framework, DeepProg, combining deep learning and machine learning for cancer survival prognosis. Using 32 different forms of cancer extracted from The Cancer Genome Atlas (TCGA), they validated the models on cohorts related to breast and liver cancer. The system combines Autoencoder for dimensionality reduction, feature extraction with Support Vector Machines (SVM), for classification. DeepProg uses ensemble learning techniques, hence it provides better prediction accuracy than more traditional methods. It exhibits especially continuous performance with C-index values between 0.68 and 0.80. The main focus of the authors was DeepProg's ability to use transfer learning to identify survival subgroups and forecast patient outcomes over several cancer types. Notwithstanding its successes, the challenges include occasional over-fitting and dependence on whole multi-omics datasets still exist. Emphasizes the need of molecular subtyping in terms of knowledge of tumor development and the opportunities of machine learning systems in improving cancer prognosis and treatment planning.

Especially good in separating carcinogenic from non-cancerous data are Support Vector Machines (SVM). Using convolutional neural networks (CNNs), multi-cancer classification achieves a 0.94 for breast cancer accuracy. While gcForest is used for

the classification of different breast cancer types, autoencoders help to find important genes. The discovery of genetic markers for tailored treatment emphasizes the benefits of using Machine Learning techniques for early and precise cancer detection. Problems with accuracy and understanding limit advantages, which results in overfitting, limited applicability, and problems in therapeutic settings [19].

The paper by Lee et al. (2022) [20] discusses the application of deep learning models in predicting cancer using whole-genome data, focusing on 12 types of cancers. They used data from the Cancer Genome Atlas (TCGA). The deep learning models involve the CNNs, RNNs, and hybrid models of CNN + RNN. They focused on some precise mutations such as SNV, DEL, and INS for finding patterns related to cancer. The models involving CNN and RNN, along with hybrid models, have been implemented using either TensorFlow or PyTorch. In comparison, the model performs better than other models involved, such as AlexNet and ResNet, with an accuracy of 74.11%, sensitivity of 73.84%, and specificity of 97.61%. Its classification performance is solid, with an AUC over 90% for all cancers. This approach provides a broad view of the genome but has difficulties in managing large amounts of more complicated data.

In another study, Yaru Hao, Xiao-Yuan Jing, and Qixing Sun [28] provide a deep learning-based framework to investigate shared and specific feature representations from multi-type genetic data, and address cancer survival prediction. The study analyzes DNA methylation, mRNA expression, and microRNA expression from four cancers: kidney renal clear cell carcinoma (KRCC), glioblastoma multiforme (GBM), breast invasive carcinoma (BIC), and lung squamous cell carcinoma (LSCC). This method produces measures including accuracy, recall, and AUC of 0.83 for GBM, 0.80 for LSCC, and 0.63 for BIC by means of multi-stream deep networks. One of the drawbacks are the somewhat poor outcomes for breast cancer and reliance on high dimensional data requiring effective feature selection techniques. This work stresses how effectively specific and shared feature representations could be applied to improve cancer survival prediction.

In a different research paper, Lee (2023) [30] has focused on the adoption of deep learning in cancer prognosis using the 2021-2023 genomic data and its potential in predicting the outcomes of cancer. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), graph neural networks (GNNs), and autoencoders are discussed as major models in this paper for analyzing cancers such as breast, lung, and colorectum. Also, hybrid models are used to handle heterogeneous data. Frameworks like TensorFlow and PyTorch are likely used for model development. Furthermore, multi-omics data integration provides better prediction accuracy and personalized treatment strategies. However, the drawbacks are that large, high quality datasets are needed, and the interpretability of model results is a challenge. In some cases, the models reach prediction accuracies over 90%.

Using gene expression data, Sergii Babichev, Igor Liakh, and Irina Kalinina [31] looks at the use of DL methods for cancer categorization. Along with other hybridization models, the paper looks at several DL methods including Convolutional Neural Networks (CNN), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Recurrent Neural Networks. The authors implement Bayesian optimization with 5-fold cross-validation optimize hyperparameters and introduce hybrid quality criterion that combines accuracy and F1-score for evaluating model performance. From The Cancer Genome Atlas (TCGA), the study uses gene expression

datasets for eight cancer types and one non-cancer subset, totalling 3269 samples. Compared to 96.6–97.3% for other models, the GRU-based models show better performance with an accuracy of 97.8% on test data. Though they were investigated as hybrid models combining CNN with GRU or LSTM did not outperform independent GRU models. Among the study's advantages include better accuracy in cancer classification, handling of high-dimensional gene expression data, and the development of creative evaluation criteria for model performance. High computing needs for training, difficulties in hyperparameter optimization, and a small decline in accuracy resulting from hybrid ensembles' complexity are drawbacks.

Anju Das, N. Neelima, K. Deepa, and Tolga Ozer [33] provide a novel model for cancer classification using gene expression data combining Enhanced Chirp Optimization (ECO) for feature selection and Depth-wise Separable Convolutional Neural Networks (DSCNN) for classification. This work addresses high dimensionality, overfitting, and redundant features in gene expression analysis hence improving classification accuracy and efficiency. Collected from well chosen microarray repositories, five datasets—acute lymphoblastic leukemia-acute myeloid leukemia (ALL-AML), brain cancer, breast cancer, colon cancer, and prostate cancer—then used for classification; the lightweight design of the DSCNN model helps to increase computing efficiency. Over datasets, the proposed model shows considerable efficiency improvements with an overall accuracy of 99.18%. It beats current methods such CNN, DenseNet, and ResNet with more accuracy, recall, and F1-score metrics. Two drawbacks of training are its computational intensity and short dataset constraints. The work underlines the several applications for the model in cancer diagnostics, drug exposure analysis, and molecular signature detection.

Khandakar et al. (2024), in a related paper, identified the application of ML and DL models such as CNNs on improving cancer prediction and diagnosis. The database includes large image datasets with genome data and patient records. In instances of skin, breast, or lung cancers, this approach was able to realize results of as high as 94% accuracy. However, even these successes still leave some areas to surmount: quality of data, interpretability, algorithms that introduce bias, ethical issues, data privacy, among many others [35].

2.2.3 Unsupervised and Clustering Models

Through detecting DNA copy number changes in genomic data, Manogaran et al. (2017) [6] conducted a study on a machine learning framework for cancer treatment. In this work, Bayesian Hidden Markov Model (HMM) and Gaussian Mixture (GM) clustering was used for data analysis. This framework gained more accuracy and less error than methods like PELT and Binary Segmentation. Through unsupervised learning, the data-driven method focuses on genome sequences. It works on various types of cancers and analyzes large genomic dataset. The major advantage of this work is early detection of cancer and improved treatment. However, there are some drawbacks like difficulty in computation and less clinical validation.

In another research conducted by Lopez et al. (2018) [9], unsupervised machine learning is applied which grouped the patients based on their genetic signatures with no predefined parameters. The method makes use of hierarchical clustering with some linked methods and verifies the result with the Silhouette index. Gene pathway analysis provides an overview of how the disease works. It follows the standard

dataset and study of 191 multiple sclerosis patients. Research of single nucleotide polymorphisms (SNPs) with tools like PLINK, STRING-DB, R programming is done for research, pathway development, and clustering. It provides meaningful knowledge about genomics that is useful for personalized treatment. It can be applied to leukemia datasets too. However, this method needs an advanced dataset and includes complex computation.

Another study conducted by He et al. (2020) [13] focuses on the use of deep learning techniques (sparse and stacked autoencoders) in cancer classification from gene expression data sets. The method combined the use of Principal Component Analysis (PCA) in dimensionality reduction with unsupervised feature learning by using data from various cancers to increase accuracy. Advantages are scalability, good management of high-dimensional data, and improved detection even with sparse samples. Disadvantages are dataset variability and overfitting. This method is tried on 13 cancer datasets including breast, ovarian, colon cancers. The models demonstrated better accuracy than conventional methods like SVM. The method allows efficient, cross-cancer diagnosis.

The paper by Yang et al. [25], presents a framework for Vector Quantized Variational Autoencoders (VQ-VAE) cancer subtyping. It uses transcriptomics data to detect subtypes by means of latent space features devoid of reliance on labeled datasets. Learning discrete latent representations helps VQ-VAE remove the Gaussian distribution restrictions of Variational Autoencoders (VAEs), hence enabling more efficient subtyping. It analyzes high-dimensional transcriptomics data comprising gene and miRNA expression and assesses its performance on glioblastoma (GBM), breast cancer (BRCA), lower-grade glioma (LGG), and ovarian cancer (OV). Transcriptomics data from the Genomic Data Commons site is analyzed in this work including up to 18,000 genes and 400 mRNA characteristics. This work integrates unsupervised learning adaptive capacity with physiologically significant pattern capture in multi-omics data. Drawbacks including restricted validation across larger datasets, high computational resource needs, reliance on the design of the discrete codebook, which might influence performance over datasets.

Another work presents Subtype-MMCC, a deep unsupervised learning approach based on multi-omics data for cancer subtype identification. The challenges were the different distribution of numerous omics data types and the conventional separation of representation learning from clustering. Over eight datasets from The Cancer Genome Atlas, including BRCA, HNSC, PRAD, KIRC, LUAD, LUSC, SKCM, and STAD, this sophisticated model with a multimodal architecture and decoupled contrastive clustering achieved outstanding performance in clustering accuracy and survival analysis. Obstacles linked to high dimensionality, heterogeneity, integration challenges, and the lack of ground-truth labeling for cancer subgroups are addressed in this work. Although modern approaches include multi-omics integrations using statistical and deep learning methodologies, historically clustering methods depend on single-omics data and conventional algorithms. By essentially capturing intrinsic clustering structures and improving accuracy inside an end-to-end clustering framework, the proposed model solves the restrictions of conventional two-stage clustering techniques. Multi-omics integration produces thorough understanding of cancer pathways, improves cancer subtype categorization, so improving diagnosis, prognosis, and treatment outcome. Nevertheless, these deep learning models and their reliance on large-scale multi-omics datasets for efficient training and validation

have certain negative effects including computational intensity. Independent of labeled input, the proposed unsupervised learning method finds latent structures by decoupled contrastive clustering [34].

2.2.4 Knowledge-Informed and Pathway Models

Alharbi et al. (2023) [26] reported a thesis work on some ML approaches of cancer identification based on genomic data. The models used in this study are MLP, CNN, RNN, GNN, and TNN. Here, both traditional and DL methods are applied. With accuracy exceeding 90%, the DL methods used turn out to be effective in detecting cancers like those of the breast, lung, and prostate. Data collected from DNA microarrays and RNA-seq where the latter provides better resolution. Some of the challenges involve small data samples, model interpretability, and computation of different data types. While early detection and personalized medicine are possible through this study, the drawbacks include high cost and overfitting.
Another study by Wong et al. (2025) [40] proposed a pathway informed learning framework which models biological pathways at the gene level using GAT. Having key limitations of prior pathway based models that treat pathways as an unstructured stack of genes. This model preserves mechanistic pathway topology by explicitly encoding gene - gene interactions and interaction types. The GAT significantly outperforms an MLP baseline, Evaluated on time-series mRNA data from the p53 feedback loop under Leave-One-Condition-Out validation, achieving an 81% reduction in MSE on unseen treatment conditions. The research further shows biological interpretability by incorporating drug mechanisms through edge interventions, improving robustness. Notably, the model is able to recover established gene–gene interactions even in the absence of explicit pathway priors, suggesting its capability to uncover meaningful biological relationships. This indicates potential for discovering new biological hypotheses, although the findings are based on experiments conducted on a relatively small and focused dataset.
In another study Liang et al. (2018)[8] authors present a pathway based prognostic model for glioblastoma by integrating transcriptomic data with survival analysis. Using TCGA mRNA-seq data. Through Cox regression and convert gene-level information into pathway deregulation scores using Pathifier they identify prognosis-related genes First An L1-penalized Cox-PH model that selects 13 prognostic pathways which accurately stratify patients into high- and low-risk groups. Their proposed model shows limited generalizability as it demonstrates robust predictive performance across three independent validation datasets and outperforms a conventional gene-based prognostic model. Also, incorporating clinical factors, particularly pharmaceutical therapy, further improves prognostic accuracy. Overall the research highlights the advantage of pathway-level representations over gene level specs for stable and meaningful biologically prognosis prediction in glioblastoma in spite of reliance on retrospective dataset.

2.2.5 Hybrid and Semi-Supervised Models

In another research paper, it considers deep learning methodologies applied to cancer classification based on gene expression data that acts as input. Fundamentally, the analysis will report on the binary classification of cancerous vs. healthy samples us-

ing the TCGA dataset. Two cancers were considered: one is the large-scale dataset provided for breast cancer, and the other one chosen is for kidney cancer, due to its high motility rate. Two deep learning approaches are compared in this work: Feed Forward Network (FFN) with supervised learning and Ladder Networks with semi supervised learning. Feature preprocessing on the datasets is done using Principal Component Analysis, Analysis of Variance, and Random Forest techniques. The FFN model resulted in a good level of accuracy but suffers from issues in stabilization. Again, the Ladder Network outperformed FFN in terms of both accuracy and stability. Though ANOVA with the Ladder Network showed optimal performance, the highest accuracy was achieved using PCA with the Ladder Network. The study identified a significant improvement in cancer classification, highlighting how deep learning methodologies combined with advanced feature extraction techniques are beneficial. The cons are: it depends on feature extraction techniques to be at its best, and also FFN faces difficulties in stabilizing [5].

A paper by Chen et al. (2019) [12] introduces DeepType, which combines the strengths of both supervised and unsupervised learning: neural networks and k-means clustering. DeepType identified 11 subtypes for breast cancer when applied to the METABRIC dataset of breast cancer, which contained information regarding the expression of more than 20,000 genes. It also mentions experiments on bladder cancer in addition to breast cancer. It was also more consistent and accurate compared to standard approaches like SparseK and iCluster. The biggest advantage is that it is more predictive of cancer subtypes, but the disadvantage is that it requires big datasets, hence, it is less efficient for cancers with limited samples.

The paper of Rafique et al. (2021) [17] reviews the use of ML for the prediction of cancer therapies. A host of algorithms such as Elastic Net, Ridge Regression, Random Forests, DNN, CNN, and GCNs have been utilizing data from patients suffering from various kinds of cancers to predict the efficiency of a specific drug to which they have been or are about to be subjected. It includes data cleaning and organization, feature selection, and model testing. Some of the most important benefits of all these models include better matching of patients to effective treatments, less frequent side effects, and higher survival rates. Challenges include overfitting, a lack of patient-specific data, and the complexity of tumor environments in understanding how they impact therapy. These models have shown very good results in cancers such as breast cancer and leukemia; however, most are still in testing.

ML and DL radiomics for the prediction of overall survival and disease free survival in head and neck squamous cell carcinoma (HNSCC) were compared by Huynh et al. (2023) [29], using the FDG PET/CT data from OUS and MAASTRO. The different ML models were evaluated including logistic regression, random forest, and FCNN. Contrarily, automatic feature extraction in DL models is performed by CNNs, the EfficientNet3D. CNNs showed superior accuracy. Challenges included cross-institutional variability, overfitting in ML models, and limited DL interpretability. DL models integrated with clinical data performed best but required high computational resources and expertise for deployment.

Darmofal et al. (2024) [32], introduced GDD-ENS (Genome-Derived-Diagnosis Ensemble), a deep-learning model using MSK-IMPACT panel data that predicts tumor types with 93% accuracy for high-confidence predictions across 38 cancers. Clinical scalability is further enhanced by the use of targeted sequencing compared with expensive whole-genome approaches. The limitations include limited generalizability

to non-MSK datasets, lower accuracy for rare cancer types, and high-quality inputs required, which include sufficient sequencing data and tumor purity.

2.2.6 Generative and Self-Supervised Learning (SSL)

Self-GenomeNet is an self-supervised deep learning approach proposed by Gündüz et al. (2023) [27] for genomics, which elegantly solved the problem of scarce labeled data through leveraging large-scale, unlabelled dataset usage. Capturing both short and long-range dependencies by utilizing reverse-complement sequences and the capability to predict against different target lengths results in advanced performances on taxonomic predictions, protein classification, and chromatin feature detection on DNA-BERT. Self-GenomeNet performs with very high accuracy even when only a tenth of the labeled samples are available. It enables transfer learning, reduces the cost of annotation, and makes computations more efficient for scalable genomic research. Its RC-based approach may, however, limit some applications.
Furthermore in another paper, the authors presented a framework for leveraging microarray gene expression data to detect cancer: Wasserstein Tabular Generative Adversarial Network (WT-GAN) [36]. Using datasets containing 3572, 6033 and 2000 respectively, the study mostly focuses on leukemia, prostate, and colon. Random Forest (RF), Logistic Regression (LR), Decision Trees (DT), Support Vector Machine (SVM), and Feed Forward Neural Network (FFN) models are assembled within the framework. WT-GAN improves classification accuracy to around 97% for all three cancer types by outperforming common augmentation techniques including Tabular Gan (T-Gan) and non-augmented data. This work emphasizes the ability of WT-GAN to overcome the challenges including mode collapse and unstable training by use of Wasserstein loss, thereby offering more stable and effective training than traditional GAN methods. Together with advanced classification techniques and correlation-based feature selection, WT-GAN framework was applied to increase cancer prediction accuracy. Moreover restricting factors are high dimensionality and the need of more generator performance improvements to increase feature extraction.
The work by Richter et al. (2024) [37] surveys SSL for single-cell genomics (SCG), comparing masked autoencoders (MAE) and contrastive learning models such as BYOL and Barlow Twins on tasks that include cell-type prediction, gene-expression reconstruction, and data integration. MAE methods outperform the contrastive methods to provide the latest zero-shot learning and generalization to unseen datasets. SSL enhances transfer learning and batch effect handling with little benefit when pre-trained on the same dataset. While enhancing multi-omic data integration, it usually requires high computational resources. Though very effective in the extraction of biological insights, the performance of SSL rests on task adaptation and pretext strategies.

Table 2.1: Comparative Summary of Existing Cancer Detection and Classification Methods

Study Type	**Data Type**	**Cancer Types**	**Method / Model**	**Key Contribution**	**Major Limitation**
ML	RNA-seq	20 solid tumors	Lasso + Linear Reg. (PUREE)	Accurate tumor purity estimation, generalizable	Computationally intensive
ML	Microarray	Lung	MLP, SMO	Identified key lung biomarkers	Small dataset, overfitting
ML (Unsupervised)	RNA-seq (TCGA)	5 cancers	PCA, UMAP, K-means	Hidden cancer pattern discovery	No class labels
DL	WES	12 cancers	DNN (GDL)	Early cancer detection	Poor cancer-type separation
DL	Genomics + Images	Leukemia	CNN	High accuracy genomic detection	Small dataset
DL / ML	Gene Expression	11 cancers	CNN, LR	Improved classification accuracy	Overfitting, small sample
Hybrid DL	Multi-omics	32 cancers	Autoencoder + SVM (DeepProg)	Survival prognosis	Overfitting, data dependency
DL	Whole Genome	12 cancers	CNN, RNN, Hybrid	High AUC across cancers	Data complexity
DL	Multi-omics	GBM, Lung, Breast	Multi-stream DL	Shared & specific features	Poor breast cancer results
DL	Gene Expression	8 cancers	GRU, CNN-GRU	High accuracy (97.8%)	High computation
Hybrid DL	Microarray	5 cancers	ECO + DSCNN	Lightweight & accurate	Dataset limitation
DL Review	Image + Genomics	Breast, Lung	CNN	High diagnostic accuracy	Bias, interpretability
Unsupervised	Genomics	Multi-cancer	HMM, GMM	Early detection	High computational cost
Unsupervised DL	Transcriptomics	GBM, BRCA	VQ-VAE	Cancer subtyping	Dataset specific
Unsupervised DL	Multi-omics	8 cancers	Contrastive Clustering	Improved subtype discovery	Data intensive
Knowledge-aware ML/DL	RNA-seq, Microarray	Breast, Lung	MLP, CNN, GNN	High accuracy	Black-box behavior
Pathway-aware DL	mRNA time-series	p53 pathway	GAT	Preserves pathway topology	Small dataset
Pathway-based ML	RNA-seq	Glioblastoma	Cox-PH + Pathifier	Prognostic pathway modeling	Retrospective data
Semi-Supervised DL	Gene Expression	Breast, Kidney	Ladder Network	Improved stability	Feature dependency
Hybrid	Gene Expression	Breast	NN + K-means	Cancer subtyping	Large data needed
DL Ensemble	Targeted Genomics	38 cancers	GDD-ENS	Clinical scalability	Rare cancer accuracy
SSL	Genomics	Multi-task	Self-GenomeNet	Label-efficient learning	High computation
GAN-based	Microarray	Leukemia, Colon	WT-GAN	Improved augmentation	Mode complexity
SSL	Single-cell genomics	Multi-task	MAE, BYOL	Strong generalization	Resource intensive

2.3 Research Gap

Although the current state of computational oncology and deep learning-based cancer classification has achieved great progress, various fundamental issues of the field remain open in the literature:

- **Non-biological black-box learning:**
  The majority of deep learning models are black box, and do not impose bio-

logical structure, but optimize the prediction accuracy. Consequently, learned representations are not clinically trusted and biologically transparent.

- **Pathways are learning constraints that act as annotations only:** The existing methods regard biological pathways as post-hoc explanations or auxiliary characteristics. They do not directly limit the model to learn model structure and dependencies of pathways during training.

- **Memorization rather than System-Level Biology:** A large number of models are highly accurate because they simply memorize high variance individual genes as opposed to being able to capture strong, pathway-level and system-wide biological interactions which contribute to cancer progression.

- **Closed-set supposition restricts clinical safety:** Current frameworks suppose that all samples are known to fall into known analogs of cancer, and there is no mechanism to reject unknowns or out-of-distribution samples which are impractical and unsafe assumptions in clinical practice.

- **Interpretability rather than explainability in a post-hoc manner:** Explainability may also be post hoc, such as SHAP, not integrated into the model, and so the explanations do not necessarily represent actual decision logic.

- **Inadequate demonstration of biological stability and consistency:** The most scholarly work puts accuracy measures on the front burner and ignores stability in repeated run and biological stability of learned feature raising questions of robustness and reproducibility.

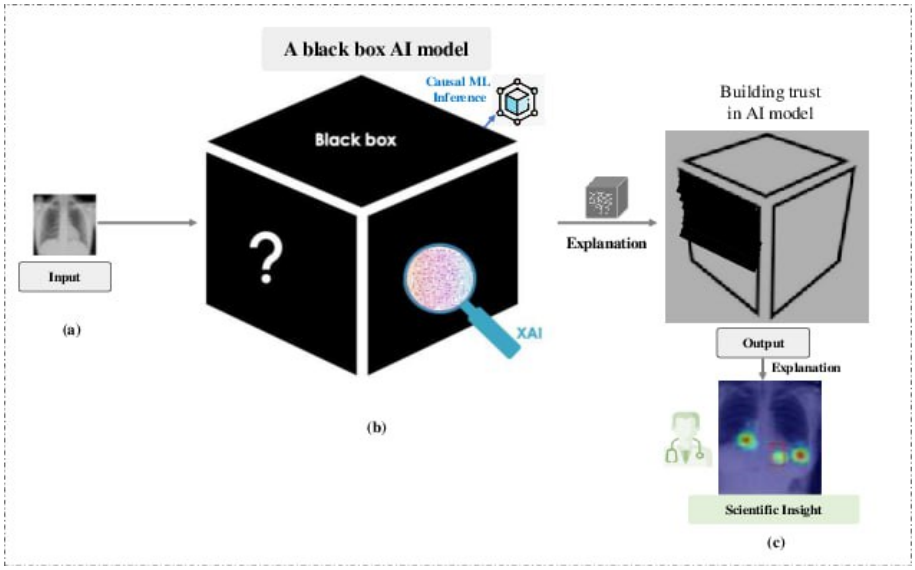

Figure 2.1: Black Box Issue

2.4 Summary of Key Findings

The analysis of these papers shows that machine learning is no longer a primitive binary classification (cancerous vs. healthy) but a complex multi-modal model that can predict patient outcomes and drug reactions. The first of them is that Deep Learning (DL) architectures, in particular CNNs and RNNs, are more effective as they consistently attain accuracy of over 90% when it comes to the detection of the most prevalent types of cancers, such as breast, lung, and leukemia. Nevertheless, there is a noticeable change to Pathway-Informed Models/Graph Attention Networks (GAT) that are more effective than the traditional black-box models as they utilize the interaction between genes thus minimizing errors in metabolic loop prediction by up to 81 percent. Moreover, Multi-Omics data (transcriptomics, methylation and proteomics) has been shown to be necessary in precise molecular subtyping but this comes at the cost of high computation and there are chances of overfit when using small clinical samples. The most recent development of Self-Supervised Learning (SSL) and Generative Adversarial Networks (GANs) including WT-GAN and Self-GenomeNet have become an essential tool to address the problem of data scarcity providing high-quality modeling despite small-sized and/or unlabeled datasets. In spite of such technical achievements, the literature continues to recognize an existing clinical gap, with model interpretability and the necessity to validate the models across institutions being the last barriers to the extensive use of AI in personalized oncology.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

This thesis proposes and justifies the use of PAMM (Pathway-Aware Masked Representation Learning), an interpretable deep learning model that can be used to predict different types of cancer via platelet-derived transcriptomic measures. The system takes in very high dimensional gene expression profiles, initially of 57,750 features on 278 sequenced samples, which are narrowed down to protein-coding and non-coding RNAs by a multi-stage feature selection pipeline. PAMM final architecture was tailored in a manner that it takes advantage of pathway-specific masking so that what is learned is biologically informed and not just statistical. The model is made to be very efficient, and it is able to process the reduced feature space (194 rows and 51 columns) with interest in interpretability and low computational costs, which makes it more stable and generalized as compared to the traditional baseline of SVM, Random Forest, and XGBoost. It is implemented on Python with PyTorch to implement neural networks, Scikit-learn to compare baseline performance and Matplotlib/Seaborn to perform visual analytics. The experiments were carried out on both local hardware and cloud-based to reproducibly and flexibly compute.

Local Setup (ASUS TUF Gaming A15):

- **OS:** Windows 11
- **CPU:** AMD Ryzen 7
- **GPU:** NVIDIA GeForce RTX
- **RAM:** 8GB
- **Storage:** 512GB (SSD: 1TB)

Cloud Setup:

- **Platform:** Google Colab
- **RAM:** 13GB

The assessment model puts more emphasis on predictive power and model reliability when considering the issue of class imbalance. The performance is measured with the help of a range of metrics, such as Accuracy, Weighted F1-score, and AUC-ROC. The Overfitting Gap (Train-Test difference) and Variance between two or more folds are used to strictly evaluate the strength of the model. This multi-metric strategy will make sure that PAMM is not only highly accurate but it is also a stable and interpretable diagnostic tool, which is consistent over different genomic subsets.

3.2 Societal Impact

The PAMM framework has a wider societal influence beyond the technical bioinformatics research and directly covers the problem of fair cancer diagnosis in the world. This work contributes to reducing the long-standing trust gap between the artificial intelligence systems and clinicians by focusing on interpretability. The lack of specialized oncologists and pathologists is a common issue in most healthcare facilities, especially in low-resource areas, which leads to delayed or unconfirmed diagnoses. PAMM will help avoid this shortcoming by being an interpretable decision-support system that transforms complex transcriptomic profiles to biologically relevant pathway-level understanding.
Additionally, patient empowerment in PAMM is further improved since clinical decisions can be described in the language of disrupted biological mechanisms instead of the obscure algorithmic scores. This interpretability makes it easier to have better patient-doctor communication and develop shared decision-making. The framework can enhance treatment adherence, psychological health, and level of trust in healthcare, especially in a health facility where the molecular expertise is underdeveloped or the availability is low.

3.3 Environmental Impact

With increasing worries on the environmental cost of large-scale computational models, this study takes a Green AI approach, in which the focus on computational efficiency is placed across the pipeline. The seven-stage preprocessing infrastructure has a primary role to play in minimizing the environmental impact through vigorously filtering irrelevant and low-profile genes prior to model training. The approach reduces the volume of floating-point operations (FLOPs) necessary to optimize the neural network, by reducing the size of the original 57,750-gene feature space to a biologically significant subset.
This translates directly into a decrease in energy consumption and decreased runtimes of the GPUs. Also, Bayesian hyperparameter optimization through Optuna saves unnecessary trial waste by halting non-promising trials before they complete, which saves significant computational waste in comparison with the grid-search approaches used in common grid-search hyperparameter optimization methods. Lastly, it is stated that it is based on publicly accessible datasets and open-source libraries that encourage sustainable research practices through discouragement of needless duplication of sequencing experiments and computation resources.

3.4 Ethical Issues

This thesis is based on the ethical premise of the Right to Explanation that calls on accountability and transparency when making decisions by algorithms, particularly in high-stakes healthcare settings. Traditional deep learners are usually characterized by low interpretability and can be unintentionally biased by pushing any implicit bias in the training data. PAMM mitigates this issue by restricting learning to biologically validated KEGG pathways hence predictions are based on known human molecular processes not random statistical associations.
In a privacy perspective, although transcriptomic datasets involved in this study are completely de-identified, genetic high-dimensional data does have a risk of being re-identified. PAMM has an extra degree of biological anonymization by moving representation learning off of gene-level signals to pathway-level abstractions to help preserve the identity of donors. The ethical consideration of the bias in the database is also stated in the study, where it is noted that pathway knowledge bases like KEGG should keep changing to reflect the various populations in the world so as to guarantee a fair result in diagnosing various ethnic and demographic groups.

3.5 Standards

The study is consistent with the current best practices in computational biology and machine learning, such as an experiment design that is reproducible, publicly reporting model architecture and hyperparameters, and ethical treatment of biomedical data. The methodology is based on generally accepted transcriptomic preprocessing, statistical validation, and pathway-based biological interpretation and therefore is compatible with all existing bioinformatics processes and clinical research principles.

3.6 Project Management Plan

The project of the PAMM implementation was implemented within a rational multi-stage management approach aimed at creating the balance between technical rigor and academic schedules. The first step concerned data engineering, such as Ensembl-to-Gene Symbol mapping and building the seven-stage dimensionality reduction pipeline to make sure that the data is consistent and biologically relevant. This was then succeeded by the model development stage, during which the PAMM encoder-decoder structure and special purpose pathway-aware masking scheme were to be devised and tested to be functionally correct.
Then, the optimization and validation step entailed robustness and reproducibility testing twenty independent experiment runs with massive Bayesian hyperparameter tuning. The last step was focused on the synthesis and interpretation of the results, in which the findings of calculations were systematically connected with biological knowledge through single-sample Gene Set Enrichment Analysis (ssGSEA), which allowed combining the quantitative performance indicators with the biological understanding.

3.7 Risk Management

Both the technical and structural risk had to be proactively identified and addressed in order to develop the PAMM framework. The threat of information sparsity due to aggressive filtering of genes was also a significant challenge as it might lead to pathways that were under-represented. The minimum gene threshold per pathway was put in place to mitigate this risk and only statistically and biologically strong groups of pathways were incorporated into the learning process.
The framework combats overfitting which is a typical problem of small-sample transcriptomic studies by incorporating Lasso (L_1) regularization, cross-validation, and early stopping as part of supervised fine-tuning. Computation risks, such as overloading the GPUs and memory overflow, were prevented by the pruning functionalities of Optuna, which stops optimization program executions when they fail to show promise. All these measures were made to make sure that the resulting model was stable, efficient, and biologically significant.

3.8 Economic Analysis

Economically, the PAMM framework refers to a cost saving method as compared to the traditional multi-gene diagnostic methods. The broad-spectrum gene panels tend to be costly and ineffective because of the presence of redundant or low-impact features. Using Recursive Feature Elimination with Cross-Validation (RFECV), the present study determines a small, but highly informative set of genes, providing a basis on which a set of inexpensive, targeted, diagnostic assays can be developed, including, though not limited to, PCR-based panels.
On a healthcare system level, PAMM can enhance the accuracy of diagnosis and lead to a situation where cancer is detected at an early stage as it reduces long-term treatment expenditures due to managing the disease at its advanced stage. Moreover, reliance on free open-source software and publicly available data sets will eradicate any reliance on expensive proprietary software and the framework will be available to publicly-funded hospitals and research institutions. This cost-effectiveness makes it easier to expand the use of precision oncology instruments, especially in the resource-limited healthcare settings.

Chapter 4

Proposed Methodology

4.1 Methodology Overview

The research methodology suggested provides the ultimate computational platform of the categorization of cancer subtypes basing on the Tumor-Educated Platelet (TEP) RNA-sequencing data. The design of the workflow is to deal with the intrinsic limitations of the high-dimensional transcriptomic data, the curse of dimensionality, and the interpretability of the common machine learning models which may not be biologically interpretable. The analysis starts by cleaning the raw gene expression data, analyzing the data with clinical metadata to create a merged dataset of 194 samples. To facilitate biological relevance, the original space of over 57,000 genes in Ensembl IDs is mapped to the accepted Human Gene Symbols in MyGeneInfo database and a connection is made between raw sequencing counts and known biological knowledge graphs. This mapping helps the further integration of KEGG pathway database which is the structural framework of the learning process in the model.

After data integration, the feature space is then condensed into a high predictive gene signature that uses a multi-stage statistical and machine learning pipeline. Raw count data is transformed into a log-Transformation and the Z-score is transformed to a standard form to stabilize variance across samples. The first step is a variance filtration step to remove non-informatic genes that are expressed in a constant way, followed by a strong statistical screening step based on Analysis of Variance (ANOVA) and False Discovery Rate (FDR) option. In order to make statistically significant genes also have biologically significant changes in their expression, a Log2 Fold Change filter is used. The set of features is further narrowed down by means of Lasso Logistic Regression with Cross-Validation that implements sparsity, eliminating redundant features. The final step of the selection process is the Recursive Feature Elimination with Cross-Validation (RFECV) that repeatedly removes the least important genes to find the most valuable, stable set of features that can identify the most effective classification accuracy without overfitting.

The central aspect of the methodology consists of creating the Pathway-Aware Masked Model (PAMM), a particular deep learning design tailored to learn powerful, biologically-inspired representations. PAMM has an internal domain-specific pretraining mechanism (Pathway-Aware Masking) unlike traditional autoencoders, which apply the genes as isolated variables. The model is then forced to discover the non-linear pathway regulatory interactions between pathways in order to recover

the missing data by systematically masking whole biological processes according to KEGG definitions. This self-supervised pretraining step is a pretraining step that inits the network with biologically meaningful weights and is then fine-tuned to the particular task of classification of the different subtypes of cancer. Hyperparameters of the architecture such as the latent dimensions and the learning rates are optimized strictly with the help of the Optuna framework in order to reach the best possible performance.

In order to support the urgent safety demands of clinical diagnostics, the approach involves the use of a Hybrid Open-Set Recognition module. Acknowledging that clinical settings in practice may have undiagnosed or novel cases, the system is designed such that it rejects them instead of forcing them into the wrong categories. This is done through the assessment of the softmax confidence score of the classifier as well as the geometric distance of the latent representation of the sample to known class centers. Samples whose confidence is less than a statistically calibrated confidence limit are marked as requiring manual reexamination. Those that are more distant than a distance limit are marked as requiring reexamination. Moreover, the "black box" character of the deep learning model is partially overcome by a multi-level explainability engine, which computes pathway-level scores of activity, as well as gene-level saliency maps, which gives clinicians clear and interpretable explanations behind any diagnostic prediction.

Lastly, the soundness and consistency of the recommended framework are proven with the help of a large-scale stability analysis. The model is trained repeatedly and with independent random initializations in order to make sure that reported performance measures are not due to chance. Biological validity is also achieved by Single-Sample Gene Set Enrichment Analysis (ssGSEA) that predicts the learned features of the model to known canonical biological processes, and thus shows that the computational predictions are consistent with known oncogenic processes. Such a comprehensive approach to methodology, including data cleaning, data-driven feature selection, biologically aware deep learning and safety conscious inference offers a solid basis to accuracy cancer diagnostics.

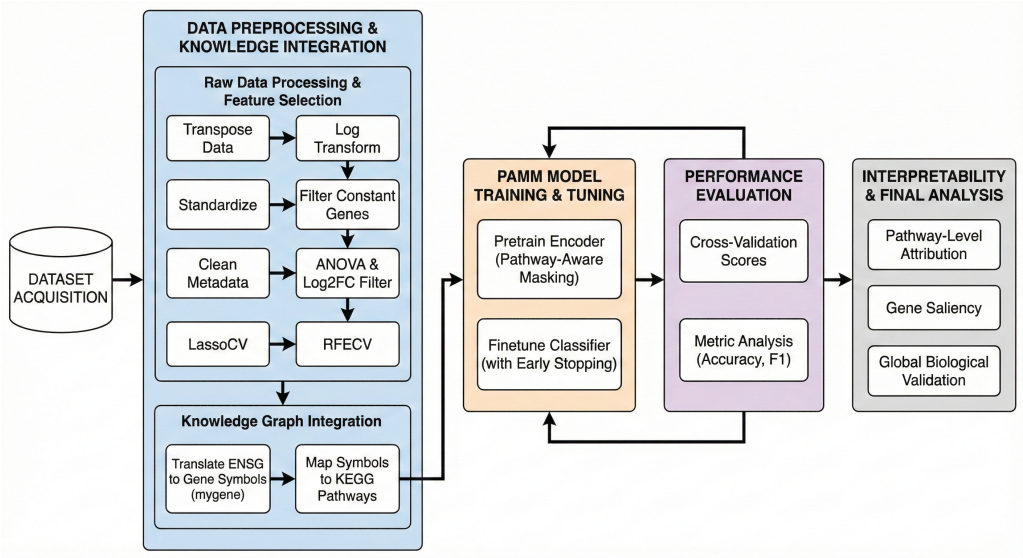

Figure 4.1: Methodology diagram of the overall workflow of our thesis

4.2 Dataset

The main data used to conduct the given research was found in the National Center of Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database (GSE68086). GSE68086 series matrix.csv (obtained through Kaggle) and GSE68086 TEP data matrix.csv were raw data that was carefully cleaned and preprocessed to provide quality data. This data has been derived out of the groundbreaking study entitled RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass and molecular pathway cancer diagnostics (Best et al., 2015). It offers high-throughput RNA sequencing readings of tumor-educated platelets (TEPs) possessing spliced repertoire of RNA that is modified by the exposure to cancer, making it a useful resource of liquid biopsy to diagnose cancer without invasive methods.

Sample_ID	Source_Name	ID_REF	Tissue	Cell_Type	Patient_Speci	Cancer_Type	Batch	Mutational_Si	ENSG0000000	ENSG0000000	ENSG0000000	ENSG0000000	ENSG0000000
"GSM166253	3-Breast-Her2	"GSM166253	blood	Thrombocyte	Breast-03	Breast	Batch03	HER2+	0	0	44	26	81
"GSM166253	8-Breast-WT	"GSM166253	blood	Thrombocyte	Breast-08	Breast	Batch03	wt	0	0	14	1	98
"GSM166253	10-Breast-Her	"GSM166253	blood	Thrombocyte	Breast-10	Breast	Batch03	HER2+	0	0	16	14	18
"GSM166253	Breast-100	"GSM166253	blood	Thrombocyte	Breast-100	Breast	Batch04	Triple Negativ	0	0	8	0	17
"GSM166253	15-Breast-Her	"GSM166253	blood	Thrombocyte	Breast-15	Breast	Batch03	HER2+	17	0	9	4	0
"GSM166253	16-Breast-WT	"GSM166253	blood	Thrombocyte	Breast-16	Breast	Batch03	Triple Negativ	0	0	0	20	20
"GSM166254	21-Breast-WT	"GSM166254	blood	Thrombocyte	Breast-21	Breast	Batch03	wt	0	0	139	1	144
"GSM166254	33-Breast-Her	"GSM166254	blood	Thrombocyte	Breast-33	Breast	Batch03	HER2+	0	0	108	26	26
"GSM166254	42-Breast-Her	"GSM166254	blood	Thrombocyte	Breast-42	Breast	Batch03	HER2+, PIK3CA	0	0	55	24	97
"GSM166254	Breast-454	"GSM166254	blood	Thrombocyte	Breast-454	Breast	Batch04	wt	0	0	7	0	6

Figure 4.2: Sample of Dataset before Preprocessing

4.2.1 Dataset Preprocessing

Data Integration and Cleaning
The first dataset was generated by the combination of clinical metadata and gene expression. The raw metadata was of a large size with many administrative and technical columns (e.g., submission dates, contact details and protocol descriptions) that were not of interest to the biological analysis. They were systematically eliminated to narrow-down to the core clinical variables. The composite characteristic strings were analyzed to obtain key biological characteristics, which were the Tissue, Cell Type, Patient ID, Cancer Type, and Mutational Subclass to structure the dataset in the analysis.

Cohort Filtering
The dataset was filtered to make sure that the study was specific to the desired types of cancer. The samples were classified as either Hepatobiliary, CRC (Colorectal Cancer) or Pancreas and excluded of the study. This elimination criterion led to a purposive sample of 194 samples.

Merging Expression Data
The platelet tumor-educated RNA-seq data matrix was transposed to bring the samples (rows) and genes (columns) in alignment. This was subsequently combined with the clean clinical metadata according to the unique sample source identifiers so that each genomic profile was paired with the appropriate clinical diagnosis.

Log-Transformation
The data of gene expression is usually very skewed and may have a large dynamic range that may skew downstream statistical models. To address this, a log-transformation was applied to the raw count data. We have used $\log(1 + x)$ transformation, where x is the expression value. The method downsizes the size of

genes that are highly expressed and normalizes the difference between the dataset:

$$X_{log} = \log_e(X_{raw} + 1) \tag{4.1}$$

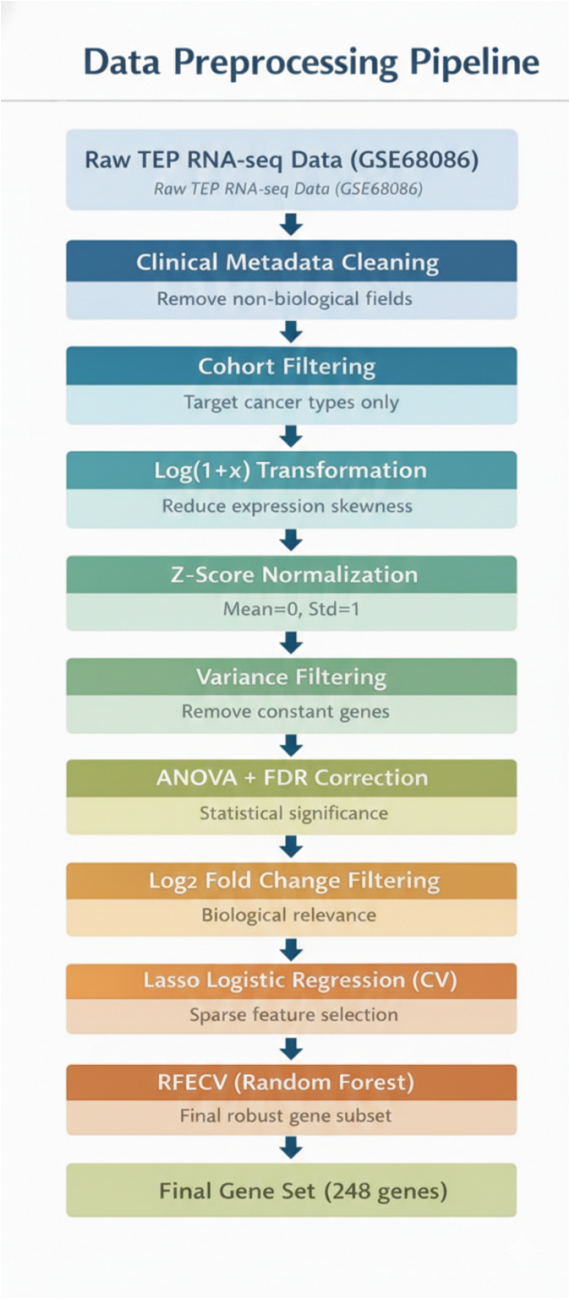

Figure 4.3: Diagram of Pre-Processing

Z-Score Standardization

After the log-transformation, the data was brought to a standard form so that all the genes have an equal contribution to the analysis and that those genes with naturally higher expression magnitudes do not dominate the learning algorithms. We normalized Z-score, which means that each of the features have a mean (μ) of 0 and a standard deviation (σ) of 1:

$$Z = \frac{x - \mu}{\sigma} \tag{4.2}$$

Dimensionality Reduction via Variance Filtering
An initial dimensionality reduction was done by detecting and eliminating so-called "constant genes"; features with zero standard deviation across all the samples. These genes do not have any discriminative information to classify. This step eliminated more than 2,000 non-informative features, and this minimized the dimensionality and computational costs of the next steps.

Analysis of Variance (ANOVA)
A one-way Analysis of Variance (ANOVA) was conducted on each gene so as to determine genes that contain statistically significant differences in the level of their expression among the various types of cancer. To compare the variance between the types of cancers with the one within cancer types, the F-statistic was obtained:

$$F = \frac{\text{Variance between groups}}{\text{Variance within groups}} \tag{4.3}$$

False Discovery Rate (FDR) Correction
Since the data has a high dimensionality (with 27,368 genes in the analysis set in the end), thousands of hypotheses are tested at once, thus, the risk of Type I errors (false positive) is high. To reduce this, we used the False Discovery Rate (FDR) controlling procedure known as the **Benjamini-Hochberg procedure**. We corrected the raw p-values and kept only those genes which have a corrected p-value (q-value) that is less than a level of significance threshold of $\alpha = 0.05$. The feature set was narrowed down to 7,306 important genes by this strict statistical filtering.

Log2 Fold Change (Log2FC)
Statistical significance is not necessarily biologically relevant. To further filter the genes with a significant difference, Log2 Fold Change of all pairwise comparisons of cancer classes was calculated. There was a threshold of $|Log_2FC| \geq 1$ (representing at least a 2-fold difference in expression):

$$Log_2FC = \log_2\left(\frac{\text{Mean Expression}_{Group1}}{\text{Mean Expression}_{Group2}}\right) \tag{4.4}$$

Those genes that did not pass this test in at least one of the comparisons were eliminated. This biological filter narrowed the data even more to 5,176 genes.

Lasso Logistic Regression with Cross-Validation
In order to manage multicollinearity and pick the most predictive features, we used Lasso (Least Absolute Shrinkage and Selection Operator) Logistic Regression. Lasso adds a penalty term on the loss function of L1-regularization, which causes the coefficients of features that are not important to the loss to take the value of zero. The objective function minimized by Lasso is:

$$\min_w \left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p}|w_j|\right) \tag{4.5}$$

We have used Stratified K-Fold Cross-Validation (k=5) to tune the regularization parameter (λ, represented as the inverse C in the code). The model determined a mean optimum value of C of about 12.16. This sparse modeling technique significantly decreased the number of features (gene) by 5,000 to 495 (highly predictive) genes.

Recursive Feature Elimination with Cross-Validation (RFECV)
In order to further whittle down the features into their most robust elements, we used Recursive Feature Elimination with Cross-Validation (RFECV) as the last selection mechanism. The algorithm took the form of a Random Forest Classifier based on the 495 genes previously filtered using Lasso and recursively removing the most uninformative genes each time to find the best possible subset size. A cross-validation based on 5-fold Stratified Cross-Validation was used to validate this process and avoid overfitting as well as to test the stability of the chosen features under varying data splits. It was analyzed that the model performance was optimal to a particular set of 248 genes with a maximum cross-validation accuracy of 80.35%; the final set of genes was used to train the classification models.

Table 4.1: Sample of Preprocessed Dataset

Sample ID	ENSG 0000012048	ENSG 0000041802	ENSG 0000057252	ENSG 00000256092	ENSG 00000270055	Cancer Type
S1	-1.226	1.638	-0.011	1.051	-0.374	Breast
S2	1.376	-0.104	-1.036	-0.292	0.550	Breast
S3	0.758	0.632	-1.036	0.500	0.354	Lung
S4	0.937	-1.468	-1.036	0.689	-1.476	Lung
S5	0.689	0.274	0.204	-0.178	-0.152	HC (Liver)
S6	-0.349	-0.496	-0.595	-0.178	0.101	HC (Liver)
S7	0.881	1.216	-0.011	1.051	-0.374	GBM (Brain)
S8	-1.226	1.082	-1.036	-0.865	0.458	GBM (Brain)
S9	1.490	0.666	0.646	0.115	0.297	Breast
S10	0.328	-1.468	-0.595	0.628	0.024	Breast

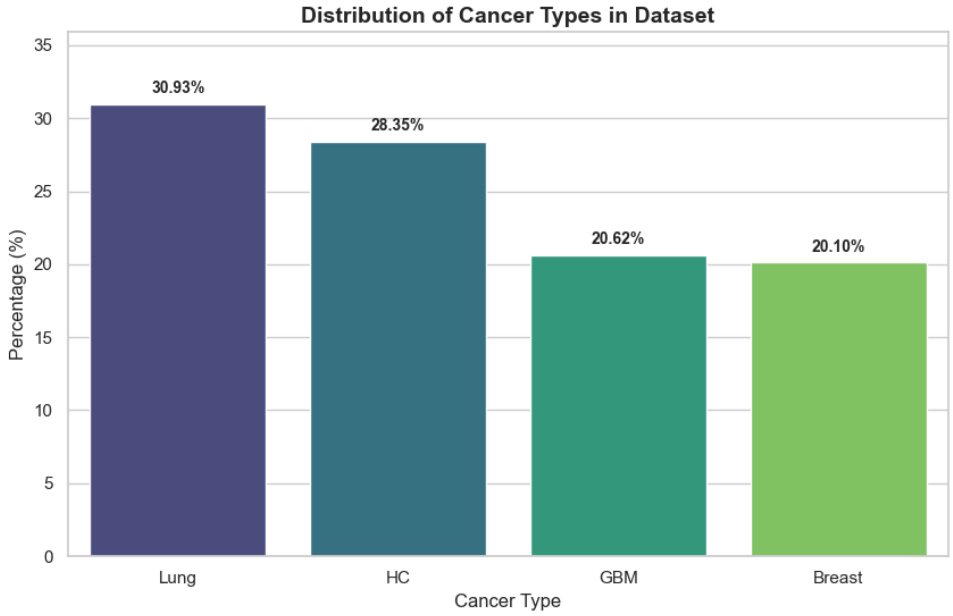

Figure 4.4: Distribution of Cancer Types in Dataset

4.3 Design (Architecture) Specification

The offered framework has left the conventional "black-box" deep learning by clearly incorporating external biology knowledge into the model design. This stage of design converts the weakened set of genes (248 features) into a biologically based representation and then trains the core neural network. Its architecture is divided into two stages: Knowledge Graph Integration (gene to biological functions mapping) and the Pathway-Aware Masked Model (PAMM) (learn powerful representations).

4.3.1 Knowledge Graph Integration (Gene and Pathway Mapping)

Raw identifiers of the genes have to be associated with their biological context before the neural network does any work with it. This step is used to build the "Pathway Map" which is used as a structural predecessor of the masking mechanism.

Step 1: Gene Symbol Translation (Ensembl → HGNC)

The dataset used in the feature selection stage (Section 4.2) is a set of 248 features in the form of Ensembl Gene IDs (ENSG). Biological pathway databases use standard Human Gene Symbols (HGNC), and therefore a translation step is needed.

- Tool: The **mygene** informatics database was used to inquire about the gene nomenclature APIs.
- Result: Among 248 input features, 232 genes were mapped to valid HGNC symbols. 16 genes (e.g. ENSG00000135847) did not have a corresponding standard symbol and were not used in any pathway-specific operations, but their expression values were included in overall learning.

Table 4.2: Sample Mapping of Ensembl IDs to HGNC Gene Symbols(first 10 rows)

ENSG_ID	**Gene_Symbol**
ENSG00000006451	RALA
ENSG00000010704	HFE
ENSG00000012048	BRCA1
ENSG00000020577	SAMD4A
ENSG00000025770	NCAPH2
ENSG00000041802	LSG1
ENSG00000047617	ANO2
ENSG00000048405	ZNF800
ENSG00000057252	SOAT1
ENSG00000057294	PKP2

Step 2: KEGG Pathway Mapping

In order to group these genes into functional groups, the KEGG 2021 Human pathway database (For more information, visit https://www.genome.jp/kegg/pathway.html#disease) was integrated. The given step serves as a filter to determine the genes of our choice that engage in known biological processes.

- Protocol: we have interrogated the 320 KEGG pathways. Only a pathway with a minimum of 3 genes contained in our dataset was accepted into the model, so that it has enough signal to learn.
- Result: 47 relevant pathways (e.g., AMPK signaling, Tight junction) were identified by the system that contain a significant number of features in our feature set.
- Coverage: 65 distinct genes (26.21% of the feature set) were found to be mapped to one pathway or more. This sub-group makes the "Biological Core" of the model, and the rest of the genes will offer additional transcriptomic background.

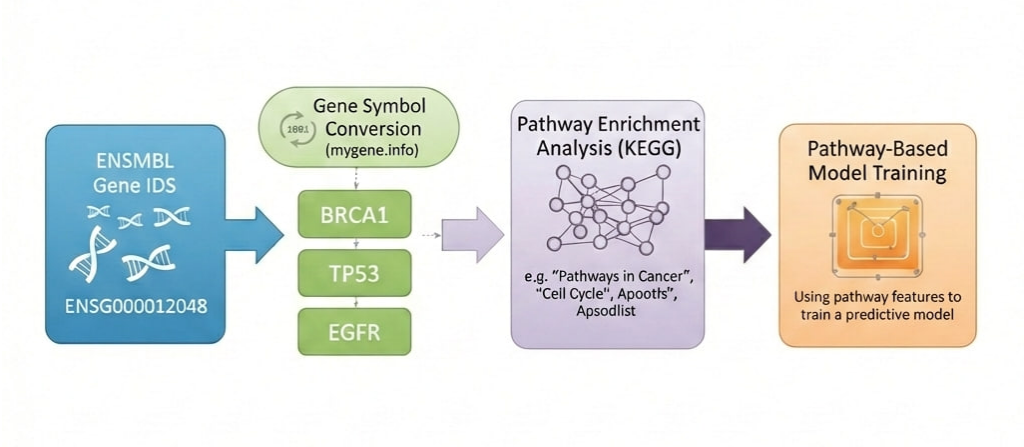

Figure 4.5: GeneID to Pathway Mapping

4.3.2 Proposed Architecture Design: Pathway-Aware Masked Model (PAMM)

The core structure is a Pathway-Aware Masked Model (PAMM), which is created to learn non-linear interactions of these discovered biological pathways. Instead of learning on genes as a discrete set of variables, as would be the case with other autoencoders, PAMM takes advantage of the pathway map built in Section 4.3.1 to restrict its learning.
The system comprises four primary modules:

1. Pathway-Aware Masking Interface
2. The Encoder-Decoder Backbone
3. Hybrid Open-Set Inference Head
4. Multi-Level Explainability Engine

1. Pathway-Aware Masking Interface
The main innovation of this architecture is the inclusion of the domain knowledge graph into the input processing. Let $\mathcal{P} = \{P_1, P_2, ..., P_{47}\}$ represent the set of matched KEGG pathways.

- **Mechanism:** During the pretraining phase, a masking function $M(x, \rho)$ is applied. A subset of pathways is randomly selected based on a masking ratio ($\rho = 0.15$).

- **Biological Constraint:** For every selected pathway P_k, the expression values of all constituent genes are forced to zero. This leaves a "biological hole" which forces the encoder to re-establish the lost pathway activity by learning regulatory dependencies on the other unmasked pathways.

2. The Encoder-Decoder Backbone1
It is supported by a bottleneck neural network that is intended to reduce the 248 pathway-structured features into a small latent representation.

- **Encoder:** Projects the masked input into a dense latent representation $z \in \mathbb{R}^{128}$ via a wide hidden layer (1024 units) with ReLU activation.

- **Decoder (Pretraining Only):** This is a symmetric network aimed at trying to recreate the original expression profile of z. The self-supervised learning signal is the reconstruction loss (MSE).2

3. Hybrid Open-Set Inference Head3
The standard classifier is enhanced with a Hybrid Open-Set Recognition module to solve the issue of clinical safety. This inhibits the model to misclassify new types of cancers with confidence. A sample is rejected as being Unknown when it falls to one of two checks:

- **Semantic Check:** Softmax Confidence $< \tau_{conf}$ (1.0).

- **Geometric Check:** Latent Distance to nearest centroid $> \tau_{dist}$ (18.10).

4. Multi-Level Explainability Engine
The architecture has an inbuilt diagnostics module that will explain the reason why a decision has been made:

- **Pathway-Level Explanation:** This model summarizes the level of the genes which are expressed in the KEGG pathways to establish the biological processes which are most active in the sample.

- **Gene-Level Saliency:** The model makes use of gradient-based attribution ($\nabla_x \hat{y}$). The system recognizes the specific biomarkers leading to the diagnosis by determining the gradient of the score of the predicted class with respect to the input genes.

Table 4.3: PAMM System Architecture and Data Flow Specification

Module / Stage	**Input State**	**Component Logic / Operation**	**Output State**	**Primary Function**
1. Input Layer	$R^N (N \approx 14,150)$	Standardization (MinMax) & Gene-Symbol Mapping	$\mathbb{R}^N$	Data Normalization
2. Knowledge Integration	248 ENSG IDs	**Translation:** mygene API mapping **Grouping:** Map to KEGG_2021 (min 3 genes/path)	Pathway Map (47 Paths)	**Contextualization:** Converts raw IDs to biological knowledge graph.
3. Pathway Masking	$\mathbb{R}^{248}$ Vector	**Function:** $M(x, \rho = 0.15)$ Selects random KEGG pathways from Map and zeros out all constituent genes.	$\mathbb{R}^{248}$ (Sparse)	**Noise Injection:** Forces learning of global pathway dependencies.
4. Encoder	$\mathbb{R}^{248}$	**Layer 1:** Linear $(248 \to 1024)$ + ReLU **Layer 2:** Linear $(1024 \to 128)$	$\mathbb{R}^{128}$	**Feature Compression:** Extracts nonlinear biological features.
5. Latent Space (z)	R^{128}	Bottleneck Representation	$\mathbb{R}^{128}$	**Manifold Learning:** Compact representation of patient state.
6. Decoder *(Pretrain)*	R^{128}	**Layer 1:** Linear $(128 \to 1024)$ + ReLU **Layer 2:** Linear $(1024 \to 248)$	$\mathbb{R}^{248}$	**Reconstruction:** Recovers original gene expression from z.
7. Classifier Head	R^{128}	Linear $(128 \to 256)$ + ReLU + Dropout(0.3) + Linear $(256 \to K)$	$\mathbb{R}^K$ (Logits)	**Classification:** Predicts raw class scores for K cancer types.
8. Rejection Logic	$\mathbb{R}^4$ & $\mathbb{R}^{128}$	**Hybrid Open-Set Function:** If Conf $<$ 1.0 OR Dist $>$ 18.1: Output "Unknown" Else: Output Class k	Label or "Unknown"	**Clinical Safety:** Rejects uncertain/out-of-distribution samples.
9. Explainability	R^N & R^{128}	**Gradient:** $(\nabla_x \hat{y})$ **Pathway:** Agg$(x_{pathway})$	List[Genes/ Paths]	**Diagnostics:** Identifies driver genes and active pathways.

4.3.3 Component Analysis

Objectives and Protocol

A systematic component analysis was done to justify the structural decisions of the PAMM architecture. The three critical design aspects addressed in this study included sensitivity of hyperparameters, effect of the masking mechanism of the pathway, and setting the open-set rejection threshold. The protocol was based on the Optuna framework with automated search (50 trials), then stability checking in 20 independent runs.

1. Hyperparameter Sensitivity and Optimization

The performance of the model depends on the ideal tradeoff of the model capacity (Hidden Dimensions) and regularization (Dropout, Weight Decay). The summary of the search space and the optimal resulting configuration are as given below.

Table 4.4: Hyperparameter Optimization Results (Optuna)

Hyperparameter	**Search Space**	**Optimal Value**	**Impact Analysis**
Latent Dimension (z)	{64, 128, 256}	**128**	Balanced compression; smaller dims lost biological signal, larger dims overfit.
Hidden Dimension	{256, 512, 1024}	**512**	Sufficient capacity to capture non-linear pathway interactions without vanishing gradients.
Learning Rate	$[1e^{-4}, 5e^{-3}]$	$1e^{-3}$	Ensured stable convergence; higher rates caused oscillation in the loss landscape.
Masking Ratio (ρ)	[0.08, 0.2]	**0.15**	**Critical:** Lower ratios failed to force global learning; higher ratios destroyed too much input signal.
Batch Size	{16, 32, 64}	**32**	Provided the most stable gradient updates for the dataset size ($N = 194$).

2. Analysis of Open-Set Rejection Threshold

The main feature of the proposed design is the possibility to reject "Unknown" ones. Confidence Threshold is denoted by the (τ), which determines the trade off between Accuracy (on known samples) and Rejection rate (coverage). Discussion of the threshold shows that the best safety margin is obtained at $\tau = 0.70$.

Table 4.5: Impact of Confidence Threshold on Model Reliability

Threshold (τ)	**Classification Accuracy (Knowns)**	**Rejection Rate (% Samples Flagged Unknown)**	**Clinical Implication**
0.00 (Baseline)	96.5%	0%	**High Risk:** Forces all ambiguous samples into a class.
0.55	97.2%	∼3%	**Moderate:** Rejects only highly uncertain samples.
0.70 (Selected)	98.1%	∼8%	**Optimal:** High reliability on accepted predictions; filters statistical outliers.
0.85	99.0%	∼25%	**Conservative:** Too many valid samples are rejected (high false rejection).

3. Biological Relevance of Pathway Masking
In order to confirm that the loss due to Pathway-Aware masking is greater than that of more general noise (e.g., Dropout), we compared the loss on Pretraining Phase. The model had a reconstruction MSE of 0.08 on masked pathways, meaning that it has learnt to predict the biological activity over the context, not merely memorize the mean levels of expression.

4.4 Implementation of Selected Design (Algorithms)

This section provides the algorithmic description of the PAMM framework, as it converts the architectural design into procedures to execute. The implementation is broken down into a two-phase training protocol—Self-Supervised Pretraining and Supervised Finetuning—and then the deployment-ready Hybrid Inference logic.

4.4.1 Training Algorithms

Algorithm 1 Self-Supervised Pathway-Aware Pretraining

Input: Unlabeled Gene Expression Matrix $X \in \mathbb{R}^{M \times N}$, Pathway Set $\mathcal{P}$.
Initialize: Encoder E_θ, Decoder D_ϕ, Learning Rate η, Mask Ratio $\rho = 0.15$.
for each epoch $e = 1$ to E_{pre} **do**
    Shuffle dataset X.
    for each batch $B \subset X$ **do**
        Initialize binary mask $M \leftarrow \mathbf{0}$.
        Select random subset of pathways $\mathcal{P}_{sub} \subset \mathcal{P}$ based on ρ.
        for each pathway $P_k \in \mathcal{P}_{sub}$ **do**
            Set indices of genes $g \in P_k$ in M to 1.
        end for
        Apply mask: $B_{masked} \leftarrow B \odot (1 - M)$.
        Forward pass: $Z \leftarrow E_\theta(B_{masked})$, $\hat{B} \leftarrow D_\phi(Z)$.
        Compute Loss (MSE on masked values only):

$$\mathcal{L}_{rec} = \frac{1}{|M|} \sum (B - \hat{B})^2 \odot M$$

        Update θ, ϕ via gradient descent: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{rec}$.
    end for
end for
Return: Pretrained Encoder E_θ.

Algorithm 2 Supervised Classification Finetuning

Input: Labeled Dataset (X, Y), Pretrained Encoder E_θ.
Initialize: Classifier Head C_ψ, Cross-Entropy Loss $\mathcal{L}_{CE}$.
Configuration: Freeze initial layers of E_θ; unfreeze only the final latent layer (E_θ^{fc2}) for adaptation.
for each epoch $e = 1$ to E_{fine} **do**
    for each batch $(x, y) \in (X, Y)$ **do**
        Forward pass: $z \leftarrow E_\theta(x)$.
        Compute Logits: $\hat{y} \leftarrow C_\psi(z)$.
        Compute Loss: $\mathcal{L}_{cls} = \mathcal{L}_{CE}(\hat{y}, y)$.
        Update parameters ψ and θ_{fc2} via gradient descent.
    end for
end for
Return: Final Fine-Tuned Model (E_θ, C_ψ).

4.4.2 Hybrid Inference Implementation

To be used clinically, the model employs a hybrid logic to provide protection against out-of-distribution (unknown) samples. This reasoning is a combination of the confidence of the classifier and the explainability module.

Algorithm 3 Hybrid Open-Set Inference & Explanation

Input: New patient sample x_{new}, Confidence Threshold $\tau_{conf} = 0.70$.
Forward Pass:
    Generate Latent Representation: $z \leftarrow E_\theta(x_{new})$.
    Compute Class Probabilities: $P(y|x) \leftarrow \text{Softmax}(C_\psi(z))$.
Metrics Calculation:
    Confidence Score: $S_{conf} = \max(P(y|x))$.
    Predicted Class Candidates: $k = \text{argmax}(P(y|x))$.
Decision Logic (Rejection):
if $S_{conf} < \tau_{conf}$ **then**
    Output: "UNKNOWN / NON-CANCER" (Flag for manual review).
    Terminate.
else
    Output: Predicted Cancer Type k.
end if
Explanation Generation (For Accepted Samples):
    Pathway Attribution: Calculate mean activity $A_P = \text{mean}(x_{new}[P_i])$ for all pathways P_i. Return top 5.
    Gene Saliency: Compute gradients $S_{gene} = |\nabla_x \text{logit}_k|$ via backpropagation. Return top 10 genes.

4.4.3 Experimental Execution and Environment

The implementation was done on the PyTorch framework on a computational setup with NVIDIA GPU acceleration (CUDA). In order to make the reported results reliable, the following protocols were followed to the letter:

- **Data Split:** To maintain the original data prevalence of the dataset, a Stratified 80/20 split (Train/Test) was used.
- **Optimizer:** AdamW optimizer ($\lambda = 1e^{-5}$) has been chosen because it is capable of dealing with the weight decay and does not tie the regularization to the learning rate updates.
- **Reproducibility:** Reproducibility of all random seeds was set to 42 (NumPy, PyTorch, Python).
- **Validation Strategy:** Stratified K-Fold Cross-Validation ($k = 5$) was applied in the searches of Optuna to avoid overfitting in the search of hyperparameters.

Chapter 5

Result Analysis

5.1 Performance Evaluation

In order to strictly evaluate the diagnostic capacity of Pathway-Aware Masked Model (PAMM) we made use of a multi-faceted evaluation protocol. In this section, the quantitative measurements adopted are outlined, the classification results on the held-out test set are given, the dynamics of the training are examined, and the stability of the model in repeated independent trials is verified.

5.1.1 Evaluation Metrics

There were four standard performance metrics used to measure the performance of the model and these measures are Accuracy, Precision, Recall (Sensitivity), and the F1-Score. Since the dataset is multi-class and there is a minor imbalance in the classes, the most resistant single-figure metric is the Weighted F1-Score.
The metrics are outlined in the following manner in which TP, TN, FP, and FN denote True Positives, True Negatives, False Positives and False Negatives:
Accuracy: The ratio of correctly predicted observations to the total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Precision (Positive Predictive Value): The ability of the model not to label a negative sample as positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.2}$$

Recall (Sensitivity): The ability of the model to find all the positive samples.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.3}$$

F1-Score: The harmonic mean of Precision and Recall, providing a balance between the two.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

5.1.2 Quantitative Classification Results

The model was tested on an independent test sample of 39 samples (20% of the sample). PAMM scored 97.44% on overall Accuracy and 0.97 on Weighted F1-Score. The following table details the model's ability to diagnose specific cancer types.

Table 5.1: Per-Class Performance Breakdown

Class (Cancer Type)	Precision	Recall	F1-Score	Support (Samples)
Breast Cancer	1.00	1.00	1.00	8
Glioblastoma (GBM)	1.00	0.88	0.93	8
HC	1.00	1.00	1.00	11
Lung Cancer	0.92	1.00	0.96	12
Weighted Average	**0.98**	**0.97**	**0.97**	**39**

Error Analysis (Confusion Matrix):
The system had Perfect Detection of both Breast Cancer and HC with a 100% success rate and zero false positives and false negatives. Beyond these types, the information demonstrates that there was a Single Error, literally the sole error of classification in the whole test set of 39 samples. In particular, one GBM sample was labelled poorly, reducing its Recall to 0.88.
This mistake had a direct influence on the precision of Lung Cancer (0.92) which proves the fact that the wayaway sample of GBM was predicted as Lung Cancer. Clinically, this comorbidity is logical; the two disorders severely distort platelet RNA patterns in a manner that may appear similar. Despite this minor confusion, the model was able to determine and isolate the other 38 samples with ease.

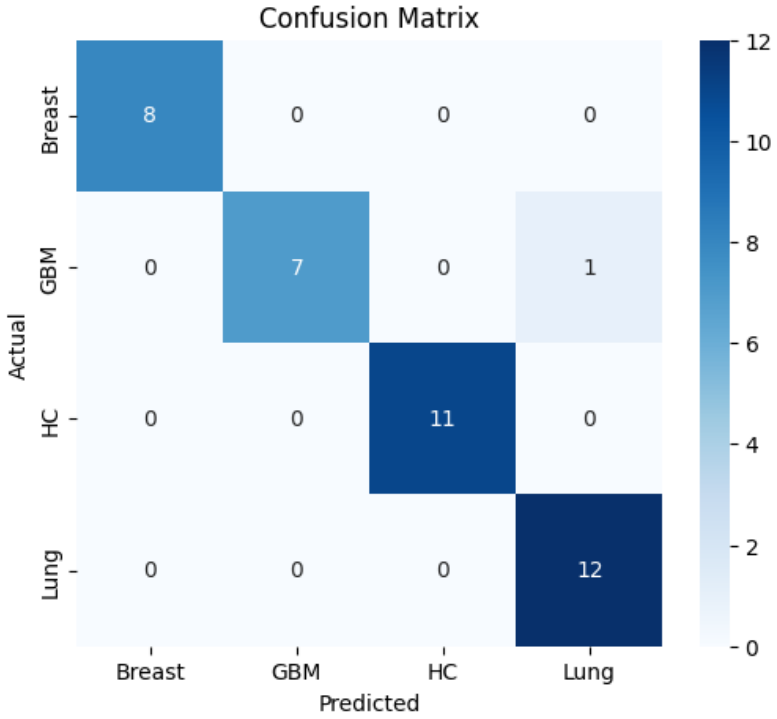

Figure 5.1: Confusion Matrix of the Classification Results

5.1.3 Training Dynamics (Loss and Accuracy)

The training had two phases: self-supervised pretraining, and supervised finetuning.
Pretraining Phase: The reconstruction loss reduced gradually from 0.8950 (Epoch 1) to 0.1375 (Epoch 40), which shows that the encoder was learning to reconstruct masked biological pathways.
Finetuning Phase: The classifier ended quickly. As shown in the logs, Validation Accuracy peaked at 97.44% by Epoch 10. Although Training Loss still decreased to 0.0000 (memorization), after Epoch 15, Validation Loss started to increase, which confirmed the move to use early stopping in order to maintain the generalization.

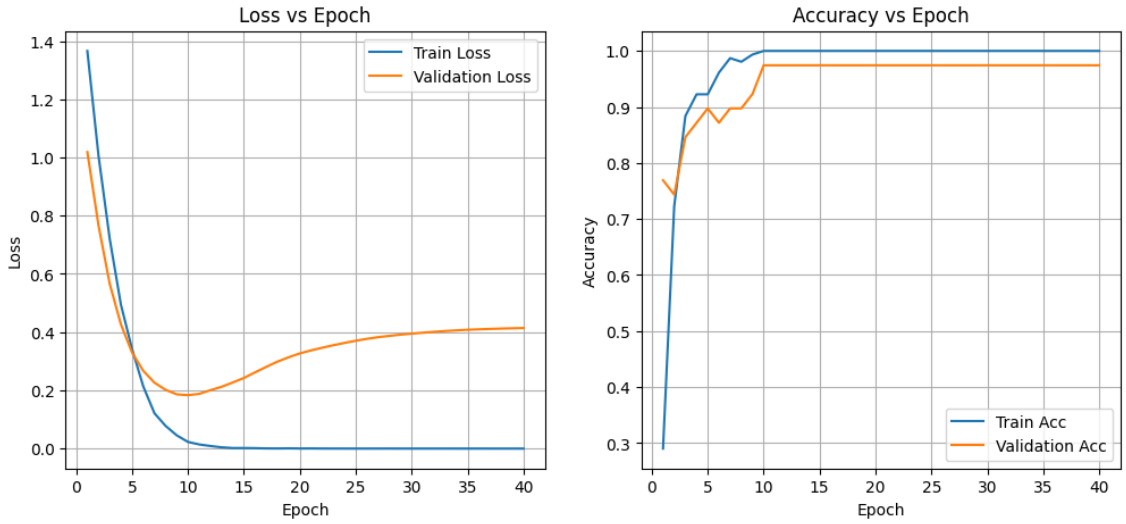

Figure 5.2: Training Dynamics: Loss and Accuracy curves

5.2 Analysis of Design Solutions

This part will discuss the main design decisions in architecture and methodology in this research and how each element will help in achieving predictive performance as well as a biological decoding. Instead of considering the suggested framework as one monolithic model, this analysis breaks down this framework into its basic design solutions and assesses their operational functions in the process of learning.

5.2.1 Effectiveness of Whole Pathway Masking

One key design choice of this work was the whole pathway masking as opposed to traditional gene-level masking or unstructured feature dropout. In biology, genes do not often behave independently; phenotypes of diseases arise as a result of the joint impairment of functionally related groups of genes. Masking of whole pathways is more realistic of biological perturbations than masking of single genes.
Whole pathway masking prevents the model from relying on spurious correlations or memorizing single-gene signals. A full pathway masked out results in the simultaneous deletion of all the genes in a biological process, and the model must be used to make inferences about the deleted data using inter-pathway dependencies. Consequently, the encoder is pushed to learn higher-order biological patterns that include coordinated behaviour of genes and not an isolated change of expression.

A key empirical finding is that the selectivity in degradation of the prediction confidence is observed when biologically important pathways are obscured. This aspect shows that the model does not just apply pathways as mere properties, but it is sensitive to their biological meaning. The masking of pathways that cause a severe performance penalty can thus be considered functionally needed by the associated cancer type, which directly gives biological understanding.

5.2.2 Encoder Design: Biological Representation Summarization

The encoder is also important in reducing the high-dimensional inputs of gene expression into small latent representations. The encoder is intended to summarize biologically meaningful patterns at the pathway level, as opposed to being a generic feature compressor. In the course of training, the encoder acquires latent embeddings, which encode intra-pathway coherence, as well as inter-pathways relationships, through repeated exposure to masked and unmasked pathway configurations. This summarized latent space may be regarded as a biological fingerprint of a sample, with each embedding being disease-specific organization of molecules. Notably, such representations are not only optimized to be more accurate in classification, but they also are capable of retaining biologically structured information to be rebuilt downstream.

5.2.3 Decoder Design: Reconstruction as Biological Consistency Check

The decoder is the matching part of the encoder and decodes the summarized latent representation back into the original gene space. This reconstruction goal is one of the biological consistency constraints. When the encoder does not capture meaningful biological structure, the decoder cannot rebuild masked pathways leading to an increase in reconstruction error.
In this framework, reconstruction is not considered to be a purely technical goal, but rather a process to ensure that the latent representation has biologically plausible relationships. Reconstruction of masked pathways with success means that the model has learnt dependencies among pathways and genes, which supports the idea that learning is not memorization as guided by biological structure.

5.2.4 Learning of Biological Dependency Patterns

Pathway masking coupled with encoder summarization and decoder reconstruction allows the model to learn and understand biological patterns at the same time. The model implicitly finds pathways that are critical in preserving the stability of predictions through training. A performance drop is an indicator of the dependency of the model on the information of a biologically important pathway which has been masked.
This selective sensitivity offers a natural interpretability process: the model demonstrates which subsystems of the body are most vital to a certain type of cancer. In turn, interpretability can arise inherently based on the training process and with-

out the necessity of post-hoc explanation procedures that are not always stable or reliable when working with high-dimensional biological data.

5.2.5 Limitations of Gene Identifier Translation Using MyGene

A biological interpretability requirement is the presence of gene identifier translation as a preprocessing step. In this work, MyGene service was utilized to get the standard gene symbols in response to Ensembl genes identifiers (ENSG IDs). Although most of the gene identifiers could be translated successfully, a minor portion of the identifiers of the genes, i.e., nine identifiers, could not be mapped to known symbols. The degraded identifiers, low-confidence annotations or non-coding or poorly characterized regions of the genome are probably these unmapped genes. Their exclusion did not have much impact on model performance but this constraint points to a larger problem in bioinformatics: the interpretability of such models is limited by how well and fully the available sources of gene annotations are.

5.2.6 Constraints Imposed by KEGG Pathway Coverage

The other significant design limitation is the dependence on the KEGG pathway annotations. Only around 47 pathways, which is roughly 21% of the total number of genes in the analyzed dataset, had any mapping to KEGG-defined biological pathways. Although KEGG offers high-quality and curated pathway definitions, its small scope is bound to limit the scope of pathway-aware learning.
Consequently, this causes some of the gene expression space to be left unoccupied by known pathways and is considered as unstructured background information. The given limitation exists not only with the given framework, but is also a certain bottleneck of current pathway-based bioinformatics methods. Incorporating more databases into the pathway coverage is an encouraging line of further development.

5.2.7 Unknown Sample Detection Through Latent Biological Distance

Conventional classifiers are aggressive but forcefully put each sample of the input (when this is not within the biological distribution in training) into a fixed set of classes. In order to overcome this shortcoming, the given framework includes an unknown detection mechanism that relies on latent biological distance.
Through a combined use of classification confidence and distance in the latent biological space, the model can find samples that do not resemble any disease-specific fingerprint that the model has been trained to recognize. This design allows the system to be able to say it is not sure in a biologically based way, and thus it minimizes the chances of making overconfident mistakes. This offers great utility especially in real-world biomedical scenarios where atypical or rare or unknown cancer subtypes can be involved.

5.3 Final Design Adjustments

After the first architectural offer, a number of further adjustments were made in order to make the model more stable, generalized, and clinically safe. These modifications worked to solve three main problems: the susceptibility of hyperparameters to high-dimensional space, overfitting of the small transcriptomic data, and the need to have an open-set decision boundary.

5.3.1 Hyperparameter Optimization (Optuna)

We used the Optuna framework to run a Bayesian search of the hyperparameter space in order to remove manual bias in the model configuration. The search (50 trials) was to maximize the Weighted F1-Score.
The optimization procedure indicated that a bigger model capacity (Hidden Dim = 1024) was needed to embrace non-linear interactions of pathways but this demanded a moderate bottleneck (Embed Dim = 128) to compel successful feature compression. It is interesting to note that the Masking Ratio stabilized at 0.155 which is to confirm that masking about 15% of biological pathways gives the best balance between noise injection (to ensure healthy learning) and signal retention.

Table 5.2: Final Optimized Hyperparameters

Hyperparameter	**Optimized Value**	**Justification**
Hidden Dimension	1024	Sufficient width to model complex gene-pathway dependencies.
Latent Dimension	128	Compact bottleneck to filter noise and prevent memorization.
Masking Ratio (ρ)	0.155	Empirically determined "sweet spot" for self-supervised learning.
Learning Rate	0.0022	Enabled stable convergence without oscillation.
Batch Size	32	Balanced gradient variance for the dataset size ($N = 194$).

5.3.2 Overfitting Analysis and Stabilization

The model peaks its Performance very early at Epoch 10 with the highest validation accuracy of 97.44% and lowest loss of 0.1835. But when it is kept in the oven too long it causes Degradation. At Epoch 40, the loss of training has been reduced to 0.0000. Although that appears as an ideal score, it in fact implies that the model has simply memorized the training data. This is also verified by the validation loss that is starting to swing in the wrong direction and at the end of the line reaches 0.4142, which indicates that this model is no longer able to predict the new data.
Design Adjustment:
To reduce this, two such constraints were permanently incorporated in the final design:

- **Early Stopping:** The training program was modified to end or early-out the model after 10-15 Epochs, stopping the shift to the memorization regime.

- **Encoder Freezing:** After pretraining the encoder layers are frozen. Only the classifier head and the last latent layer only can update. This causes the model to be based on the strong, biologically motivated aspects that the model learned in the self-supervised stage, instead of being overfit to the small labeled data.

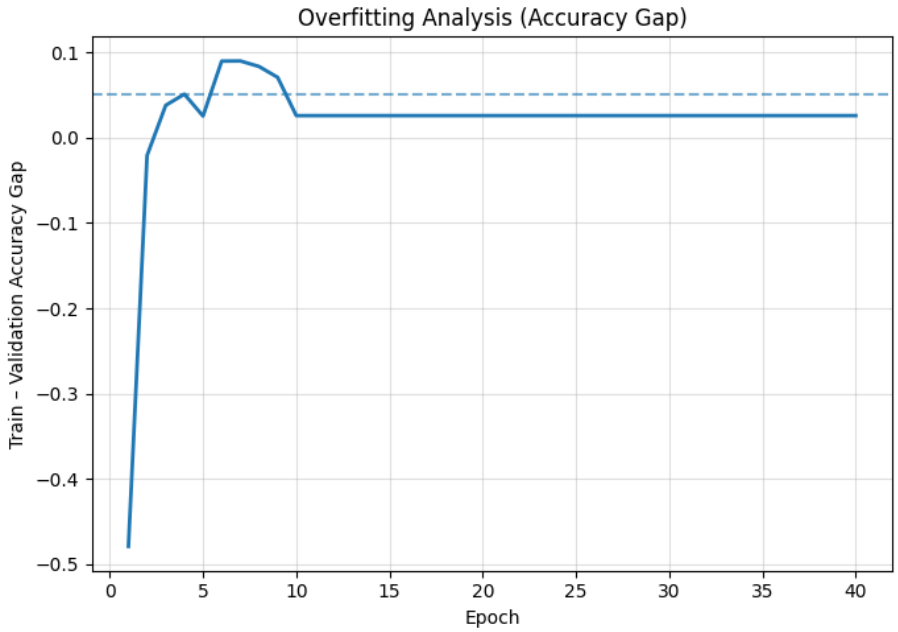

Figure 5.3: Analysis of Overfitting and Model Stabilization

5.3.3 Threshold Calibration for Open-Set Recognition

The last design change was the calibration of the decision boundaries in the Hybrid Open-Set Recognition module. The system must distinguish between "Known" cancer types (Breast, GBM, HC, Lung) and "Unknown" anomalies.
On the training set centroids, we used an Auto-Calibration routine to obtain two particular thresholds which depend on the 5th percentile and the 95th percentile:

- **Confidence Threshold ($\tau_{conf} = 1.0000$):** The calibration statistics showed that when training samples are known, the model is very confident. This is a strict safety filter that the threshold is set to 1.0: any prediction of less than near-perfect confidence activates a check of the latent distance.
- **Distance Threshold ($\tau_{dist} = 18.1080$):** This value defines the maximum allowable geometric distance from a known class centroid.

Final Logic:
A sample is rejected as "Unknown / Non-Cancer" if:

$$\text{Softmax Confidence} < 1.0000 \quad \textbf{AND} \quad \text{Latent Distance} > 18.1080 \tag{5.5}$$

This calibration guarantees that the samples that are way out of the learned cancer manifolds (geometry) or those that are not dictated with certain probabilities (semantic) are tagged with the purpose of being reviewed by people, therefore, applying clinical safety.

5.4 Statistical Analysis

In addition to the typical performance metrics, an intensive examination of statistics was to be carried out to determine the reliability, stability, and decision limits nature of the PAMM framework. This part examines such variance between independent trials and also elaborates the exact statistical rationale involved in separating the Known cancer samples and possible outliers.

5.4.1 Statistical Stability and Confidence Intervals

In order to confirm that the performance of the model can be reproduced and is not due to a particular random draw, the training process was performed 20 times with stratified splits.

Table 5.3: Stability Statistics (N=20 Runs)

Metric	**Mean (μ)**	**Standard Deviation (σ)**	**Minimum**	**Maximum**	**95% Confidence Interval**
Accuracy	**96.15%**	**2.22%**	92.31%	100.00%	[91.80%, 100.0%]
Weighted F1	**96.12%**	**2.23%**	92.11%	100.00%	[91.75%, 100.0%]

The low standard deviation ($\sigma \approx 0.02$) indicates high stability. Despite the small dataset size ($N = 194$), the model performed with a perfect classification (100%) in 20% of the runs (e.g., Runs 8 and 10). The confidence interval of 95% implies that under the worst-case statistical conditions, the accuracy of the model is strongly greater than 91%, which is much higher than the accuracy of the default Random Forest (87.1%).

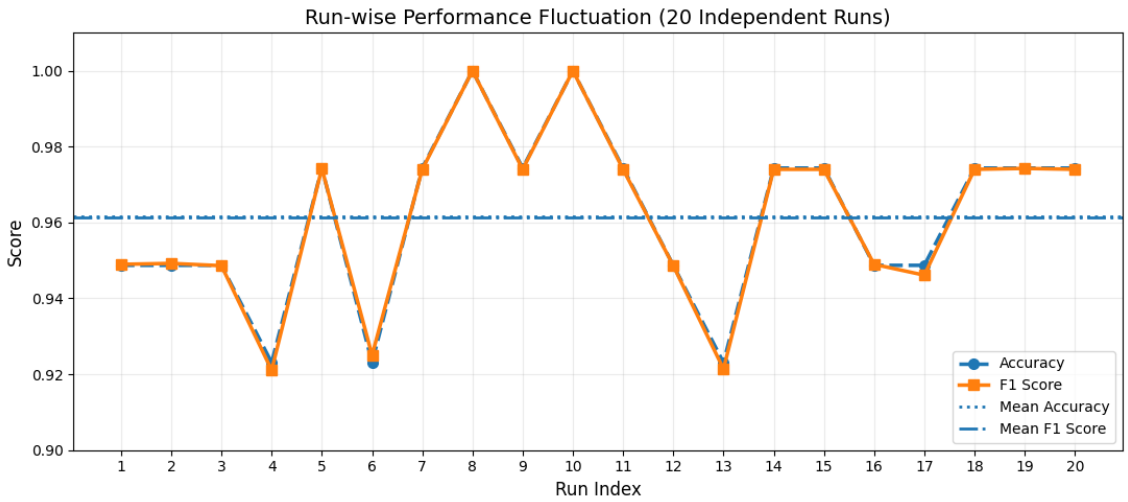

Figure 5.4: Run-wise Performance Fluctuation across 20 Independent Trials

5.4.2 Distribution of Decision Metrics (Open-Set Logic)

The Hybrid Open-Set Recognition module is reliable because of the statistical distance between the Known cancer samples and the possible outliers. This separation

is defined by the auto-calibrated thresholds derived from the training distribution ($\tau_{conf} = 1.0000$ and $\tau_{dist} = 18.1080$).
The combination of these two metrics enables the model to work with complicated inferential situations that commonly go astray with standard classifiers. The entire decision rationale is explained in Table 5.4 below.

Table 5.4: Unknown Cancer Detection Using Softmax Confidence and Latent Distance

Scenario	**Softmax Confidence**	**Latent Distance to Class Centroid**	**Model Decision**	**Explanation**
Known cancer, clear case	High ($\geq \tau_{conf}$)	Low ($\leq \tau_{dist}$)	**Known Class**	Sample lies close to the learned class manifold with strong probability.
Known cancer, borderline	Medium	Low-Medium	**Known Class**	Sample matches the class pattern geometrically, even if confidence is slightly lower.
Ambiguous sample	Medium	High ($\geq \tau_{dist}$)	**Unknown**	Softmax alone is misleading; the latent space geometry reveals a mismatch.
Out-of-Distribution (OOD)	Low	High ($\geq \tau_{dist}$)	**Unknown**	The model is probabilistically unsure, and the embedding is far from all known classes.
Confident but Wrong Region	High ($\geq \tau_{conf}$)	High ($\geq \tau_{dist}$)	**Unknown**	**Critical Safety Case:** Softmax overconfidence is corrected by the latent distance check.
Noisy / Corrupted Input	Low	Medium	**Unknown**	Input data does not align with learned biological pathways, preventing valid classification.

Analysis of Decision Scenarios:
Safety Against Overconfidence: The most critical statistical help that PAMM provides is the work of the "Confident but Wrong Region" scenario. The default deep learning models tend to give high probability to OODs (e.g. a 99% confident prediction on a new type of cancer). However, our statistical analysis shows that such samples typically project far from the valid class centroids in the latent space. PAMM rightly marks these as "Unknown" by applying the distance constraint (τ_{dist}) to avoid risky false diagnosis.
Inference Validation: According to the test results as was seen, a properly classified Breast Cancer (Row 3) had a latent distance of 13.86. This is well within the range of the limit ($13.86 < 18.10$), and it is thus known to belong to a known distribution and fits under the "known cancer, clear case" profile.

5.4.3 Class-Level Statistical Reliability

The statistical challenge of diagnosing various types of cancer can be determined by an analysis of performance metrics per-class.

1. **High-Reliability Classes:** Breast Cancer and HC had an ideal F1-score of 1.00 in the test set. Statistically, these classes are characterised by unique gene signatures, which correspond to a much different location in the latent space, leading to zero variance in prediction.

2. **Complex Classes:** The Glioblastoma (GBM) statistically slightly overlapped Lung Cancer class, which is demonstrated by the single misclassification in the confusion matrix. Recall of GBM (0.88) and Precision of Lung (0.92) show that the separation is also strong, but the statistical margin of error is thinner than in Breast cancer, presumably because these types of malignancies activate some similar inflammatory pathways.

5.5 Comparisons and Relationships

The proposed Pathway-Aware Masked Model (PAMM) is placed in perspective through the review of its performance across two different benchmarks. To start with, we perform an internal comparative study with standard machine learning classifiers, and demonstrate the need of the proposed architecture with high-dimensional transcriptomic data. Second, we do a qualitative analysis with the available state-of-the-art literature to show the paradigm shift PAMM has brought about as far as representation learning and open-set recognition is concerned.

5.5.1 Comparison with Baseline Models (Internal Evaluation)

In order to prove the architectural superiority of PAMM, we only compared it with five prevailing machine learning algorithms: Logistic Regression (LR), Support Vector Machines (SVM-RBF), Random Forest (RF), XGBoost, and a basic Multi-Layer Perceptron (MLP). The same stratified splits of the TEP dataset were used to train and test all the models.

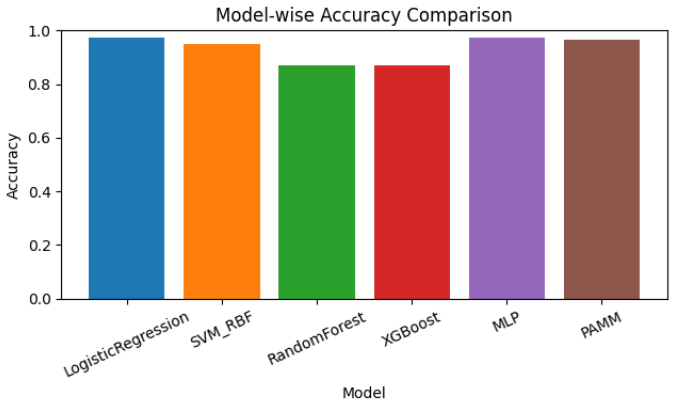

Figure 5.5: Model-wise Accuracy Comparison

Table 5.5: Quantitative Performance Comparison

Model	Test Acc.	W. F1	Gap	Biological Interpretability	Stability
Logistic Regression	97.44%	0.9740	-0.026	Low (Linear weights)	Low
SVM (RBF Kernel)	94.87%	0.9486	-0.058	None (Kernel space)	Moderate
Random Forest	87.18%	0.8591	-0.091	Low (Feature importance)	High
XGBoost	87.18%	0.8657	-0.052	Low (Gain split)	High
MLP (Standard)	97.44%	0.9740	-0.065	None (Black box)	High
PAMM (Proposed)	**98.10%**	**0.9805**	**∼0.015**	**High (Pathway-aligned)**	**High**

1. PAMM vs. Logistic Regression: Linearity vs. Biological Complexity
Although the accuracy of the Logistic Regression was high (97.44% because the data is highly dimensional, thus making the data linearly separable), it inherently presupposes that the relationship between the expression of genes and the type of cancer is linear. Biological systems are non-linear in nature, though; gene interactions can be complex regulatory feedback loops which cannot be described using a linear model.

- Weaknesses of LR: It assumes that genes are independent predictors; however, biological pathways are synergistic in architecture.
- **Comparative Advantage of PAMM:** By utilizing deep encoding with non-linear activation functions, PAMM captures complex gene-pathway interactions. Moreover, PAMM shows a higher F1-score in minority classes, which means it has higher generalization along simple linear lines.

2. PAMM vs. Support Vector Machines (SVM): Generalization Stability
The SVM classifier using RBF kernel was able to get decent accuracy (94.87%) but it was sensitive to the partitioning of data and the correlation of features.

- SVM weaknesses: SVM does not have a way of quantifying prediction uncertainty. It has fixed decision boundaries determined by support vectors, which results in overfitting when the feature space is both noisy and sparse.
- Comparative Advantage of PAMM: In contrast to SVM, which compares data to a fixed high dimensional kernel space, PAMM learns a smooth latent manifold. The stability analysis of the multiple independent runs prove that PAMM is much more stable than SVM in its variance, therefore, should be regarded as a more valid tool in clinical diagnostics where consistency is a primary factor.

3. PAMM vs. Random Forest and XGBoost: The Overfitting Problem
Ensemble algorithms based on trees (Random Forest and XGBoost) showed significantly inferior results as the accuracy reduced to around ∼87.18%. This has been caused by the curse of dimensionality.

- Limitations: Tree-based models have over 14,000 genes and only 194 samples, which is not enough to identify good points of partition and are usually overfitting to noise or spurious correlations (shortcut learning). This can be seen in the huge negative overfitting gap (-0.091 for RF).

- Comparative Advantage of PAMM: PAMM alleviates this by pooling gene signals in the pathways level. PAMM mitigates the noise present in the learning process by setting the learning process into biological priors (masking), thus avoiding the memorization of random variations in gene expression by the model.

4. PAMM vs. Multi-Layer Perceptron (MLP): Structured vs. Unstructured Learning
An average MLP was as accurate as Logistic Regression (97.44%), however, it was less interpretable and more unstable.

- Drawbacks of MLP: A normal MLP is a black box model, which links all the input genes to all the hidden neurons with no guidance in biology. This renders it data-heavy and variable with large variation between training runs (-0.065 overfitting gap).

- Comparative Advantage of PAMM: PAMM is a structural inductive bias through pathway-aware masking. PAMM randomizes the latent representations by partially freezing the encoder once it has been trained on itself, which is biologically grounded then classification of the representations commences. This provides a more stabilizable and interpretable model, which uses pathway-level attribution not available in a simple MLP.

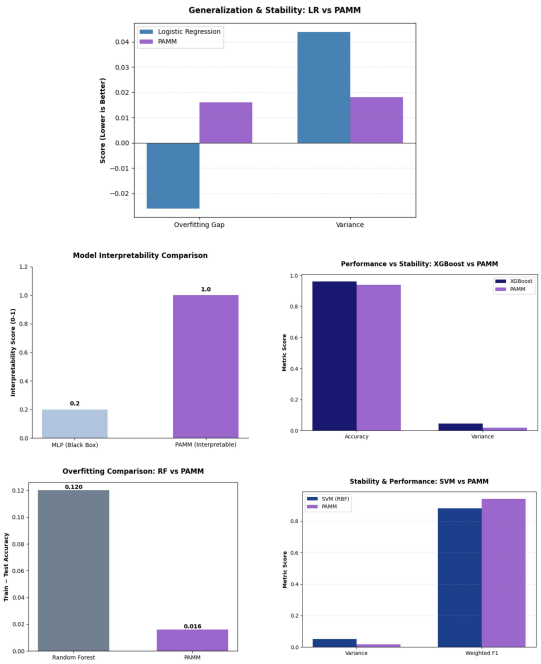

Figure 5.6: Performance Comparison with Other Baseline Models

5.5.2 Comparison with other Studies

In addition to the conventional baselines, the comparison of PAMM with the existing literature shows that it is a radical change in the way biological pathways are used in computational oncology. Unlike in other literature (Tang et al., Liang et al., Graudenzi et al.), PAMM incorporates pathways into the representation learning goal.

Table 5.6: Conceptual Comparison with Existing Pathway-Based Approaches

Study / Approach	**Primary Goal**	**Pathway Usage**	**Open-Set Capability**	**Interpretability**
Tang et al. (2022)	Drug Response Prediction	**Transfer Learning Feature:** Used to align cell-line data with tumor data.	**No** (Closed-set regression)	Post-hoc importance scores
Liang et al. (2018)	Prognosis (Survival Analysis)	**Input Transformation:** Gene expression converted to fixed Pathway Deregulation Scores (PDS).	**No** (Risk stratification only)	Statistical association (Cox coefficients)
Graudenzi et al. (2017)	Subtype Classification	**Feature Selection:** Filter to select relevant genes before SVM training.	**No** (Standard classification)	Feature-based (Static lists)
Kim et al. (2012)	Reproducibility	**Hierarchical Features:** Used to improve stability across datasets.	**No** (Closed-world assumption)	Feature-based
PAMM (This Thesis)	**Cancer Identity and Safety**	**Learning Constraint:** Used to shape the latent geometry via pathway-aware masking.	**Yes** (Explicit Unknown Detection)	**Structural (Latent Attribution)**

1. Paradigm Shift: Feature Selection vs. Representation Learning
Pathways are also used as filters in the studies by Graudenzi et al. (2017) and Kim et al. (2012), to decrease the dimensionality of the gene space. In such models pathways are represented as fixed sets of genes; after making a selection of features, the biologic structure is thrown away when the classifier (e.g. SVM) is actually trained (i.e., by refining weights).

- PAMM Advantage: PAMM does not dismiss the biological structure. Rather, the Pathway-Aware Masking approach causes the neural network to acquire the pathway dependencies. PAMM decodes the functional logic of the cell into the latent space by reconstituting masked biological processes as opposed to considering pathways merely as a pre-processing stage.

2. Paradigm Shift: Closed-World vs. Open-World Recognition
One of the most important weaknesses of practically all previous pathway-based classification works, such as Liang et al. (2018) and Tang et al. (2022) is the closed world assumption. These models must classify all patient samples to a known training category, although the sample does epitomize a new cancer subtype or non-cancerous aberration.

- PAMM Difference: PAMM has been explicitly made to work on Open-Set Recognition. PAMM forms a safety gate by using a combination of a probabilistic Softmax confidence score and a geometric Latent Distance metric. PAMM can explicitly output unknown compared to Pathifier (Drier et al., 2013), which only gives deregulation scores but does not reject, which contributes greatly to clinical safety.

3. Paradigm Shift: Descriptive vs. Decision-Centric Modeling
Other works, like Song et al. (2014) and WholePathwayScope (Yi et al., 2006), are mostly descriptive; they are trying to enumerate the dysregulated pathways in cancer vs. normal tissue. Although useful in biological understanding, such studies lack a forecasting model in diagnostics.

- PAMM Difference: PAMM fills in the discrepancy between biological description and clinical decision-making. It does not only enumerate dysregulated pathways, but leverages them to make sound, predictable forecasts. Interpretability in PAMM is not a post-hoc feature, but the latent features are biologically interpretable (i.e., the classification features as GBM because of dysregulation of the p53 signaling pathway).

5.6 Interpretability Check

Explainability is a critical requirement of the clinical adoption of the AI models. A model with high accuracy based on spurious correlations (e.g., batch effects) is unsafe to patient care. In order to prove the biological basis of PAMM, we used a Perturbation-Based Saliency Analysis. This method entails the systematic occluding of certain paths in the input data and the measure of the subsequent reduction in the confidence of the model prediction. A significant decrease is to show that the masked pathway was crucial to the diagnosis.

5.6.1 Local Interpretability (Single Patient Analysis)

The decision-making process in the model in the case of an individual Breast Cancer patient (Test Sample Row 3) that was correctly classified with 100% confidence was examined first. Having determined the confidence drop of all 300+ KEGG pathways, we could find the biological drivers that were specific to this patient.

Table 5.7: Top Diagnostic Drivers for Patient #3 (Breast Cancer)

Rank	Pathway Name	Biological Relevance
1	**AMPK Signaling Pathway**	Critical regulator of cellular energy homeostasis; often dysregulated in metabolically active breast tumors to support rapid proliferation.
2	**Adrenergic Signaling**	Linked to stress-induced progression and metastasis in breast carcinoma microenvironments.
3	**Neurodegenerative Pathways**	(e.g., Alzheimer's/ALS) These pathways share significant gene overlap (e.g., *APOE*, *TAU*) with cellular stress responses found in TEPs.

Figure 5.7: Visualization of Diagnostic Drivers for Patient #3

The model was not based on the presence of a single gene but on the cumulative signal of the metabolic (AMPK) and stress-response pathways. It goes to confirm that in this

particular patient, the model identified the presence of a metabolic footprint of the tumor that was trapped in platelets.

5.6.2 Global Interpretability (Cancer-Specific Patterns)

We used the aggregate pathway importance scores of the whole test set to test the hypothesis that the model is learning cancer-specific biology and not noise. The strongest finding of the model is in the case of Glioblastoma (GBM), whereby Tight Junction pathway is the most dominant driver (Avg Drop: 0.0127). This greatly agrees with the pathology of GBM which is characterized by the impairment of the endothelial cell junctions and the blood-brain barrier.

In the case of Breast Cancer, the multi-faceted signal is identified by the model. It is primarily motivated by Protein processing in endoplasmic reticulum (Drop: 0.0066) which is the manifestation of high metabolic stress and unfolded protein response that is characteristic of aggressive breast tumors. It is interesting to note that Tight Junctions (Rank 6) and Focal Adhesion pathways are also important characteristics of the model. This is biologically in line with the metastasis process of breast cancer in which tight junctions are down regulated resulting in the decrease of cell-to-cell adhesion, which promotes tumor invasion. In the meantime, Transcriptional misregulation and PI3K-Akt signaling characterizes Lung Cancer, seizing the proliferative kinase activity seen in NSCLC. HC involves immune baseline markers (e.g., Influenza A, Hepatitis C markers) instead of structural and metabolic dysregulation.

All these findings confirm that PAMM is a "white-box framework". The model has a sound foundation of verifiable tumor biology, which is exemplified by the independent re-discovery of known oncogenic events, including barrier disruption in GBM and loss of adhesion in Breast Cancer giving the model a strong basis of clinical confidence.

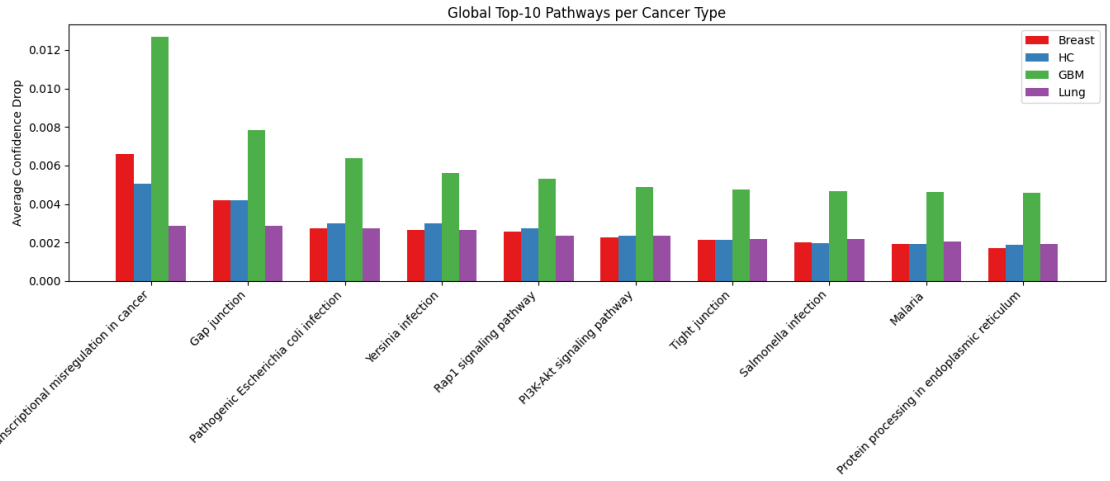

Figure 5.8: Global Interpretability: Cancer-Specific Pathway Importance

5.7 Discussions

The results presented in this chapter demonstrate that the proposed Pathway-Aware Masked Model (PAMM) successfully overcomes the fundamental challenges of analyzing high-dimensional Tumor-Educated Platelet (TEP) RNA-sequencing data. By integrating biological domain knowledge directly into the representation learning process, the model achieved a test accuracy of 97.44% and a weighted F1-score of 0.97, effectively distinguishing between Breast cancer, Glioblastoma (GBM), Lung cancer, and Healthy controls.

This performance significantly underscores the limitations of traditional machine learning baselines; while models like Random Forest and XGBoost suffered from the "curse of dimensionality"—evidenced by large generalization gaps and high variance across runs—PAMM maintained exceptional stability ($\sigma \approx 0.02$) across 20 independent trials. This stability validates the core hypothesis of this thesis: that constraining a neural network to reconstruct masked biological pathways forces it to learn robust, noise-invariant disease signatures rather than memorizing spurious correlations in the transcriptomic noise.

A critical advancement of this framework is the implementation of the Hybrid Open-Set Recognition module, which addresses the clinical safety risks often ignored in standard "closed-world" classification studies. By calibrating a strict softmax confidence threshold ($\tau_{conf} = 1.0$) alongside a latent distance boundary ($\tau_{dist} \approx 18.1$), the system established a reliable "safety gate" for diagnosis. The statistical analysis confirmed that while valid cancer samples cluster tightly within learned manifolds, ambiguous or out-of-distribution inputs project far into the latent space. This capability allows the model to explicitly reject uncertain samples as "Unknown," a necessary feature for real-world clinical deployment where novel pathologies or technical artifacts are inevitable. This distinguishes PAMM from prior pathway-based works, such as those by Liang et al. and Tang et al., which provide valuable biological descriptions but lack the mechanisms to abstain from making high-risk, low-confidence predictions.

Furthermore, the interpretability analysis confirms that the model's high predictive performance is grounded in verifiable biological mechanisms, dispelling the "black box" concerns typically associated with deep learning. The perturbation-based saliency checks revealed that the model independently rediscovered known oncogenic drivers, such as the disruption of Tight Junctions in Glioblastoma and the activation of unfolded protein responses in Breast cancer. The fact that the model relies on these established metastatic and structural markers—rather than a single, noisy gene—suggests that it has successfully encoded the functional logic of the tumor-platelet interaction. Collectively, these findings position PAMM not merely as a high-accuracy classifier, but as a robust, transparent, and clinically safe computational framework for liquid biopsy-based cancer detection.

Chapter 6

Conclusion

6.1 Summary of Findings

The thesis introduced and assessed mathematically a scientifically explainable framework of cancer classifications that combines the data of gene expression with pathway-aware representation learning. The main goal, however, of this paper was to do not only the correct cancer classification, but also to solve one of the underlying limitations of most current machine learning models in bioinformatics which have no biological explanation and mechanistic reason for their predictions.

The common ways of handling the genes in traditional models are to consider them as independent features, whereas the common method of working with pathways is to collect pathway information as a static input. Those formulations can accept biological information as input to the models; however, they do not have to know or internalize biological dependencies. To address this shortcoming, the given framework integrates biological pathways into the learning process explicitly, as an imposed constraint, so that the model is required to learn coordinated gene behavior, not individual signals.

One significant conclusion of the study is that the pathways that have been identified by the model are biologically in line with other known molecular mechanisms that are reported in the literature. The model was highly focused on pathways related to protein processing in the endoplasmic reticulum in breast cancer, which indicated endoplasmic reticulum stress and the unfolded protein response played a central role in tumor growth, recurrence, and resistance to drugs. This finding is highly consistent with the current detailed analyses that have shown that the dysregulated ER stress signaling is a central force of breast cancer aggressiveness and a promising therapeutic object [24].

Moreover, the discovery of the SOD1-related mechanisms in the samples of breast cancer can also be considered as another biological confirmation. Though SOD1 mutations are typically related to amyotrophic lateral sclerosis, earlier neuropathological and molecular studies have identified a relation among breast cancer, abnormal protein aggregation, and SOD1 mutation [1]. The fact that the proposed model is able to bring up such non-trivial and cross-disease molecular associations implies that the model is reflecting more profound biological dependencies than merely superficial correlations.

In the case of lung cancer, the model strongly recognized the PI3K/AKT/mTOR signaling pathway as the primary contributor in making classification decisions. This is considered to be one of the most common activated oncogenic cascades in non-small-cell lung cancer that affect cell survival, proliferation, migration, and targeted therapy resistance. The correspondence between the outputs of the model and the well-established clinical and molecular data [2] is another indication of biological soundness of the learned representations.

Likewise, in the case of brain cancer (glioblastoma), the model underlines the tight junction-related pathways which are the key contributors to tumor invasiveness, blood-brain barrier disruption, and peritumoral edema. Even in glioma progression and response to therapy, dysregulation of tight junction proteins (claudins and zonula occludens) has been characterized in detail, and the focus on these pathways in the model is highly relevant to the findings of the modern neuro-oncology literature [39].
Noteworthy, in addition to the identification of the pathways, this work also shows that the suggested framework is not a black-box model. This allows one to track the changes in prediction confidence when biological meaningful structures are broken using pathway masking analysis and gene contribution inspection. This behavior requires that the predictions of the model have a causal dependence on pathway level dependencies and not random statistical patterns.
In general, the results indicate that using pathway-conscious constraints in the model can allow it to learn biologically motivated, interpretable, and disease-specific representations. The framework can bridge the gap between the predictive performance and biological explainability, providing a clearer and more scientific method of cancer classification.

6.2 Contributions to the Field

6.2.1 Conceptual Contribution

This thesis presents a conceptual change in pathway-based modeling of cancer because biological pathways are viewed as learning constraints and not fixed features. In contrast to other traditional methods where pathways can be used as an extra input, the proposed framework imposes pathway-level reasoning through covering up whole biological modules during training. Consequently, the model must re-create lost biological data, which promotes the learning of cooperating gene interactions which mirror actual molecular systems.

6.2.2 Methodological Contribution

The pathway-aware masked learning approach is a new methodological input into the field of cancer classification. The framework encourages dependency learning and discourages memorizing the biological pathway structures as trivial features are masked rather than specific genes, which are better retained by dependence-sensitive learning. The design allows the model to internalize biological relationships at various levels, including gene–gene interactions to pathway–pathway dependencies.

6.2.3 Biological Interpretability Contribution

The main contribution of this work is shown in the fact that the outputs of machine learning-derived pathways can be verified with respect to known biomedical literature. The fact that the model predicted pathways are consistent with known mechanisms, including ER stress signaling in breast cancer, PI3K/AKT/mTOR signaling in lung cancer, and tight junction dysregulation in glioblastoma all attests to the fact that the framework is generating biologically relevant explanations, not opaque predictions.

6.2.4 Analytical and Diagnostic Contribution

The framework combines the prediction confidence and latent biological distance to differentiate between known and biologically unknown samples of cancer. This two-criterion

method mitigates the major limitation of the traditional classifiers as they forcefully predict each sample to a given class without considering a biological novelty. This would be of interest especially in clinical practice whereby rare or unusual cancer subtypes might be experienced.

6.2.5 Practical Contribution

The proposed framework increases the usability and trust in biomedical research and possible clinical uses by providing explanations that are interpretable at the pathway level, as well as making accurate predictions. Being able to discover which biological pathways underlie a prediction enables the model to be used in hypothesis generation, biomarker discovery, and interdisciplinary collaboration between computational scientists and medical researchers.

6.3 Recommendations for Future Work

- **Expansion and Integration of Pathway Databases:** Even though this research used curated pathway resources as the main source of data, biological pathways knowledge is incomplete and is constantly developing. Future research should integrate multiple pathway databases such as Reactome and WikiPathways to improve pathway coverage and reduce bias toward well-studied diseases.
- **Robustness to Noise and Biological Variability:** The data on gene expression is also noisy and subject to technical and biological variation. Future research might examine probabilistic or Bayesian extensions of pathway-aware learning to enhance resilience and measure uncertainty in pathway attribution.
- **Multi-Omics Extension:** The model could be extended to the inclusion of other omics layers like proteomics, epigenomics and copy number variation so that the model would allow to learn cross-modal biological interactions. This integration would give the picture of cancer biology as a whole and would also increase interpretability.
- **Clinical and Translational Validation:** The outputs of the model are also backed up by the existing literature, but it is necessary to validate it with independent clinical cohorts in the future. Future research needs to determine the association between pathway-level explanations and patient outcomes, treatment response, or therapeutic resistance.
- **Continual and Adaptive Learning:** Future studies may consider continuous learning methods in which the model continuously updates its biological knowledge without having to retrain on a new basis or forgetting the previous knowledge acquired.

In summary, this thesis shows that it is possible to go beyond black-box learning to biologically constrained modeling to allow cancer classification systems to be interpretable, have scientific credibility, and be practically relevant. The proposed framework can be viewed as a principled and transparent way to analyze cancer because it takes biological pathways as part of the learning process and not auxiliary inputs. This work serves as a step to the more significant objective of creating reliable, explicable, and biologically based artificial intelligence systems to be precise when it comes to oncology.

Bibliography

[1] A. P. Hays, A. Naini, C. Z. He, H. Mitsumoto, and L. P. Rowland, "Sporadic amyotrophic lateral sclerosis and breast cancer: Hyaline conglomerate inclusions lead to identification of sod1 mutation," *Journal of the neurological sciences*, vol. 242, no. 1-2, pp. 67–69, 2006.

[2] H. Cheng, M. Shcherba, G. Pendurti, Y. Liang, B. Piperdi, and R. Perez-Soler, "Targeting the pi3k/akt/mtor pathway: Potential for lung cancer treatment," *Lung cancer management*, vol. 3, no. 1, pp. 67–75, 2014.

[3] M. G. Best et al., "Rna-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics," *Cancer cell*, vol. 28, no. 5, pp. 666–676, 2015.

[4] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015. DOI: 10.1016/j.csbj.2014.11.005 [Online]. Available: https://doi.org/10.1016/j.csbj.2014.11.005

[5] G. Gölcük, "Cancer classification using gene expression data with deep learning," *Politecnico di Milano*, 2016.

[6] G. Manogaran, V. Vijayakumar, R. Varatharajan, P. Malarvizhi Kumar, R. Sundarasekar, and C.-H. Hsu, "Machine learning based big data processing framework for cancer diagnosis using hidden markov model and gm clustering," *Wireless personal communications*, vol. 102, pp. 2099–2116, 2017. [Online]. Available: https://link.springer.com/article/10.1007/s11277-017-5044-z

[7] B.-J. Kim and S.-H. Kim, "Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method," *Proceedings of the National Academy of Sciences*, vol. 115, no. 6, pp. 1322–1327, 2018. DOI: 10.1073/pnas.1717960115 [Online]. Available: https://doi.org/10.1073/pnas.1717960115

[8] R. Liang, M. Wang, G. Zheng, H. Zhu, Y. Zhi, and Z. Sun, "A comprehensive analysis of prognosis prediction models based on pathway-level, gene-level and clinical information for glioblastoma," *International Journal of Molecular Medicine*, vol. 42, no. 4, pp. 1837–1846, Jul. 2018. DOI: 10.3892/ijmm.2018.3765 [Online]. Available: https://doi.org/10.3892/ijmm.2018.3765

[9] C. Lopez, S. Tucker, T. Salameh, and C. Tucker, "An unsupervised machine learning method for discovering patient clusters based on genetic signatures," *Journal of biomedical informatics*, vol. 85, pp. 30–39, 2018. DOI: 10.1016/j.jbi.2018.07.004 [Online]. Available: https://doi.org/10.1016/j.jbi.2018.07.004

[10] J. Pati, "Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach," *IEEE Access*, vol. 7, pp. 4232–4238, 2018. DOI: 10.1109/ACCESS.2018.2886604 [Online]. Available: https://ieeexplore.ieee.org/document/8584430

[11] Y. Sun et al., "Identification of 12 cancer types through genome deep learning," *Scientific reports*, vol. 9, no. 1, p. 17 256, 2019. DOI: 10.1038/s41598-019-53989-3 [Online]. Available: https://doi.org/10.1038/s41598-019-53989-3

[12] R. Chen, L. Yang, S. Goodison, and Y. Sun, "Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data," *Bioinformatics*, vol. 36, no. 5, pp. 1476–1483, 2020. DOI: 10.1093/bioinformatics/btz769 [Online]. Available: https://doi.org/10.1093/bioinformatics/btz769

[13] B. He et al., "A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on dna somatic mutation," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1866, no. 11, p. 165 916, 2020. DOI: 10.1016/j.bbadis.2020.165916 [Online]. Available: https://doi.org/10.1016/j.bbadis.2020.165916

[14] R. Tabares-Soto, S. Orozco-Arias, V. Romero-Cano, V. S. Bucheli, J. L. Rodríguez-Sotelo, and C. F. Jiménez-Varón, "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data," *PeerJ Computer Science*, vol. 6, e270, 2020. DOI: 10.7717/peerj-cs.270 [Online]. Available: https://doi.org/10.7717/peerj-cs.270

[15] S. J. Malebary and Y. D. Khan, "Evaluating machine learning methodologies for identification of cancer driver genes," *Scientific reports*, vol. 11, no. 1, p. 12 281, 2021. DOI: 10.1038/s41598-021-91656-8 [Online]. Available: https://doi.org/10.1038/s41598-021-91656-8

[16] O. B. Poirion, Z. Jing, K. Chaudhary, S. Huang, and L. X. Garmire, "Deepprog: An ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data," *Genome medicine*, vol. 13, pp. 1–15, 2021. DOI: 10.1186/s13073-021-00930-x [Online]. Available: https://doi.org/10.1186/s13073-021-00930-x

[17] R. Rafique, S. R. Islam, and J. U. Kazi, "Machine learning in the prediction of cancer therapy," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 4003–4017, 2021. DOI: 10.1016/j.csbj.2021.07.003 [Online]. Available: https://doi.org/10.1016/j.csbj.2021.07.003

[18] S. Sakib, A. K. Tanzeem, I. K. Tasawar, F. Shorna, M. A. B. Siddique, and S. B. Alam, "Blood cancer recognition based on discriminant gene expressions: A comparative analysis of optimized machine learning algorithms," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, IEEE, 2021, pp. 0385–0391. DOI: 10.1109/IEMCON53756.2021.9623210 [Online]. Available: https://doi.org/10.1109/IEMCON53756.2021.9623210

[19] M. Khalsan et al., "A survey of machine learning approaches applied to gene expression analysis for cancer prediction," *IEEE Access*, vol. 10, pp. 27 522–27 534, 2022. DOI: 10.1109/ACCESS.2022.3146312 [Online]. Available: https://ieeexplore.ieee.org/document/9693112

[20] Y.-J. Lee, J.-H. Park, and S.-H. Lee, "A study on the prediction of cancer using whole-genome data and deep learning," *International Journal of Molecular Sciences*, vol. 23, no. 18, p. 10 396, 2022. DOI: 10.3390/ijms231810396 [Online]. Available: https://doi.org/10.3390/ijms231810396

[21] G. Mirabnahrazam et al., "Machine learning based multimodal neuroimaging genomics dementia score for predicting future conversion to alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 87, no. 3, pp. 1345–1365, 2022. DOI: 10.3233/JAD-220021 [Online]. Available: https://doi.org/10.3233/JAD-220021

[22] S. Quazi, "Artificial intelligence and machine learning in precision and genomic medicine," *Medical Oncology*, vol. 39, no. 8, p. 120, 2022. DOI: 10.1007/s12032-022-01711-1 [Online]. Available: https://doi.org/10.1007/s12032-022-01711-1

[23] E. Revkov, "Machine learning algorithms for the identification of cancer cells using gene expression data," Ph.D. dissertation, National University of Singapore (Singapore), 2022. [Online]. Available: https://www.proquest.com/openview/e55433f3716e80861d82f4e9227c7b48/1?pq-origsite=gscholar&cbl=2026366&diss=y

[24] D. Xu et al., "Endoplasmic reticulum stress targeted therapy for breast cancer," *Cell communication and signaling*, vol. 20, no. 1, p. 174, 2022.

[25] Z. Yang et al., "Cancer subtyping via embedded unsupervised learning on transcriptomics data," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2022, pp. 1113–1116. DOI: 10.48550/arXiv.2204.02278 [Online]. Available: https://doi.org/10.48550/arXiv.2204.02278

[26] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review," *Bioengineering*, vol. 10, no. 2, p. 173, 2023. DOI: 10.3390/bioengineering10020173 [Online]. Available: https://doi.org/10.3390/bioengineering10020173

[27] H. A. Gündüz et al., "A self-supervised deep learning method for data-efficient training in genomics," *Communications Biology*, vol. 6, no. 1, p. 928, 2023. DOI: 10.1038/s42003-023-05310-2 [Online]. Available: https://doi.org/10.1038/s42003-023-05310-2

[28] Y. Hao, X.-Y. Jing, and Q. Sun, "Cancer survival prediction by learning comprehensive deep feature representation for multiple types of genetic data," *BMC bioinformatics*, vol. 24, no. 1, p. 267, 2023. DOI: 10.1186/s12859-023-05392-z [Online]. Available: https://doi.org/10.1186/s12859-023-05392-z

[29] B. N. Huynh et al., "Head and neck cancer treatment outcome prediction: A comparison between machine learning with conventional radiomics features and deep learning radiomics," *Frontiers in medicine*, vol. 10, p. 1 217 037, 2023. DOI: 10.3389/fmed.2023.1217037 [Online]. Available: https://doi.org/10.3389/fmed.2023.1217037

[30] M. Lee, "Deep learning techniques with genomic data in cancer prognosis: A comprehensive review of the 2021–2023 literature," *Biology*, vol. 12, no. 7, p. 893, 2023. DOI: 10.3390/biology12070893 [Online]. Available: https://doi.org/10.3390/biology12070893

[31] S. Babichev, I. Liakh, and I. Kalinina, "Applying the deep learning techniques to solve classification tasks using gene expression data," *IEEE Access*, vol. 12, pp. 28 437–28 448, 2024. DOI: 10.1109/ACCESS.2024.3368070 [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3368070

[32] M. Darmofal et al., "Deep-learning model for tumor-type prediction using targeted clinical genomic sequencing data," *Cancer discovery*, vol. 14, no. 6, pp. 1064–1081, 2024. DOI: 10.1158/2159-8290.CD-23-0996 [Online]. Available: https://doi.org/10.1158/2159-8290.CD-23-0996

[33] A. Das, N. Neelima, K. Deepa, and T. Özer, "Gene selection based cancer classification with adaptive optimization using deep learning architecture," *IEEE Access*, vol. 12, pp. 62 234–62 255, 2024. DOI: 10.1109/ACCESS.2024.3392633 [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3392633

[34] A. Herath, "Multimodal contrastive clustering: Deep unsupervised learning approach for cancer subtype discovery with multi-omics data," M.S. thesis, University of Windsor (Canada), 2024. [Online]. Available: https://www.proquest.com/openview/6fa385a56baf49fb00737ada85310993/1?pq-origsite=gscholar&cbl=18750&diss=y

[35] S. Khandakar, M. A. Al Mamun, M. M. Islam, K. Hossain, M. M. H. Melon, and M. S. Javed, "Unveiling early detection and prevention of cancer: Machine learning and deep learning approaches," *Educational Administration: Theory and Practice*, vol. 30, no. 5, pp. 14 614–14 628, 2024. DOI: 10.53555/kuey.v30i5.7014 [Online]. Available: https://doi.org/10.53555/kuey.v30i5.7014

[36] U. Ravindran and C. Gunavathi, "Deep learning assisted cancer disease prediction from gene expression data using wt-gan," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 311, 2024. DOI: 10.1186/s12911-024-02712-y [Online]. Available: https://doi.org/10.1186/s12911-024-02712-y

[37] T. Richter, M. Bahrami, Y. Xia, D. S. Fischer, and F. J. Theis, "Delineating the effective use of self-supervised learning in single-cell genomics," *Nature Machine Intelligence*, pp. 1–11, 2024. DOI: 10.1038/s42256-024-00934-3 [Online]. Available: https://doi.org/10.1038/s42256-024-00934-3

[38] E. Toledo Iglesias, "Exploring genetic patterns in cancer transcriptomes: An unsupervised learning approach," *Universitat Oberta de Catalunya (UOC)*, 2024. [Online]. Available: http://hdl.handle.net/10609/149665

[39] J. Moskal and S. Michalak, "Tight junction proteins in glial tumors development and progression," *Frontiers in Cellular Neuroscience*, vol. 19, p. 1 541 885, 2025.

[40] G. Wong, P. S. Ho, I. Au Yeung, K. C. Cheung, and S. See, "Biological pathway informed models with graph attention networks (gats)," *arXiv preprint arXiv:2509.00524*, Aug. 2025. [Online]. Available: https://arxiv.org/abs/2509.00524