JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

# AttenGene: A Deep Learning Model for Gene Selection in PDAC Classification Using Autoencoder and Attention Mechanism for Precision Oncology

**Dr. Farhana Haque**

Independent University, Bangladesh

## Abstract

The low survival rate of Pancreatic Ductal Adenocarcinoma (PDAC) is due to the fact that it is one of the deadliest forms of cancer and is typically detected at an advanced stage. The current diagnostic tools are ineffective and prone to error, and the issue of early diagnosis persists. To address these issues, the authors of this research proposal propose a new deep learning (DL) architecture called AttenGene, which incorporates a self-attention mechanism for selecting sparse genes and an autoencoder for learning unsupervised features. Even with very high-dimensional gene expression data, the suggested model can handle it, and it can reduce the amount of features without negatively impacting classification accuracy. There are less biologically significant genes in AttenGene, which means it may be more interpretable and perform better in classification than traditional classifiers like XGBoost and AE + CNN. Second, the model's simplicity and convenience, as well as the information it gives on potential biomarkers for PDAC diagnosis and treatment, will make it valuable in the clinical setting. Not only in PDAC but in other malignancies where the choice and classification of genes play a critical role, AttenGene stands as an important milestone in the field of precision oncology by being the first model to combine model performance with interpretability.

**Keywords:** *Pancreatic Ductal Adenocarcinoma (PDAC), Deep learning (DL), Gene selection, Autoencoder, Attention mechanism, Precision oncology, High-dimensional data.*

## 1. Introduction

Given that PDAC is frequently diagnosed at advanced stages and that there are currently no viable methods for early detection, it is one of the most difficult and deadly cancers with extremely low survival rates. The development of effective treatment approaches has been hindered by the clinical issue of early PDAC diagnosis. One potential approach is gene expression profiling, which can shed light on the molecular pathophysiology of cancer and help us find diagnostic and prognostic indicators. However, feature selection and classification become challenging due to the high dimensionality of gene expression data, which means that there are more genes than samples.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

Traditional machine learning (ML) models have a hard time processing such high-dimensional data without experiencing overfitting and interpretability problems, which further complicates any attempt to translate the results into clinical practice.

Models should be both very accurate and interpretable by selecting only the most physiologically significant genes, as the categorization of PDAC is becoming increasingly complex. Finding ways to efficiently interpret the high-dimensionality of gene expression data while simultaneously yielding unambiguous information on the molecular mechanisms behind the disease is becoming more important as access to multi-omics data increases. Present methods often prioritize informative qualities, which may include irrelevant or noisy data that obscures the model's meaning and makes it difficult to implement. Furthermore, the majority of current models fail to achieve the necessary balance between model simplicity and performance, frequently resulting in models that are too complicated to be used in clinical settings. A major improvement in PDAC diagnostic tools and treatment plans would be a novel framework that can automatically identify a limited set of highly relevant genes while maintaining high classification performance.

This research presents a novel approach to PDAC classification using the AttenGene model, a DL framework that unifies a feature-extraction autoencoder with a feature-selection attention system. The goal of the approach is to improve classification accuracy without sacrificing biological interpretability by identifying a small group of genes that provide useful information. To address the high-dimensionality data problem, this model employs autoencoders, which allow it to reduce the model's dimensions and attention mechanism to select a small group of genes, a performance against sparsity trade-off.

The key contributions of this work are the following:

- **AttenGene Model**: To achieve high PDAC classification accuracy, the AttenGene model is developed. It uses an attention mechanism in sparse gene selection in addition to unsupervised feature extraction via an autoencoder.
- **Gene Selection Efficiency**: The ability of the model to decrease the feature space without loss in discrimination allows it to achieve great classification performance with a limited set of genes (25). This is quite remarkable.
- **Interpretability and Biological Relevance**: Given the model's relative simplicity, it is possible to draw biological conclusions about PDAC and its few genes, many of which are recognized therapeutic targets in cancer biology.

Here is the outline for the rest of the paper: In Section 2, we survey the existing literature on PDAC classification using gene expression data, drawing attention to the shortcomings of existing methods and the necessity for models that are both highly effective and easy to understand. The research approach, including the AttenGene model, dataset, and metrics, is detailed in Section 3

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

of this paper. Section 4 presents the study's conclusions, compares the AttenGene model to XGBoost and AE + CNN, and details the selected genes' biological importance. Section 5 concludes the study by discussing its key findings, future directions for research, and potential applications to the diagnosis and treatment of PDAC.

## 2. Literature Review

Given the enormous amount of room for improvement in both the accuracy of diagnoses and the capacity to establish personalized treatment regimens, there has been a lot of recent interest in using ML approaches to the prediction and classification of various cancer kinds. A number of approaches, including feature selection, dimensionality reduction, and model interpretability, have been put forward to deal with the high-dimensional genomic data. Several deep learning (DL) and bio-inspired algorithms have been investigated as potential improvements to cancer diagnosis and prediction.

For cancer classification using sparse autoencoders and a Dual-Target Deep Neural Network, Bouazza developed the Adaptive Sparse Deep Feature Selection (ASDFS) technique to handle the high-dimensional genomic data input. Outperforming PCA-GWO, RFE, and mRMR-based approaches, the method under review achieved remarkable accuracies in ovarian cancer (99.9%), prostate cancer (100%), and lung cancer (99.8%). Additionally, a pathway enrichment study proved that the selected genes were relevant to biology. Concerning the noise and high dimensionality of microarray gene expression data, the author went on to suggest the DEGS-AGC (Deep Ensemble Gene Selection and Attention-Guided Classification) framework. The use of attention-based mechanisms to guide the generation of explainable predictions and sparse autoencoders for gene selection puts DEGS-AGC ahead of more conventional methods for providing efficient and personalized cancer diagnoses. Alkamli and Alshamlan constructed a failure to compare high-dimensional genomic data with hybrid bio-inspired algorithms and DL systems, using genes they had previously chosen, in order to classify cancer. They discovered that DL models required huge data sets and more processing power, but bio-inspired algorithms like the Grey Wolf Optimizer and Harris Hawks Optimization performed well with very few genes chosen. These algorithms were also computationally efficient and better at extracting features. When compared to the appropriateness of cancer applications, this was the efficacy of the trade-off characteristic richness that gives direction. To address the dimensionality problem with microarray gene expression data used for cancer classification, Bir-Jmelet al. presented the GFLASSO-LR (Generalized Fused LASSO logistic regression). Gene selection in conjunction with classifier training using dual penalties to ascertain gene relevance and redundancy, ultimately optimized by a sub-gradient algorithm that ensures global convergence, was their chosen strategy.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

Experimental results proved that GFLASSO-LR efficiently identified target gene subsets and that the classifier's performance was significantly improved with little effort. On multi-label breast cancer classification using scRNA-seq data, Cho and Qu compared GRU to Transformer, Low-Rank Adaptation (LoRA), and XGBoost models. While XGBoost was easier to understand, GRU-LoRA and Transformer outperformed it in terms of accuracy, precision, recall, and F1-score, especially when it came to rare subtypes. Additionally, the paper proved that DL may be fine-tuned and used to the field of oncology in order to uncover intricate gene-protein connections. Liu and Mei introduced Novel Drug Sensitivity Prediction (NDSP), a data-mining model that incorporates similarity network fusion, DL, and closely related medications in multi-omics data to address the issues of overfitting, heterogeneity, and interpretability in drug response prediction. Their model outperformed existing DL methods in predicting sensitivity to targeted and non-specific cancer drugs, as well as methylation levels, by integrating RNA sequence, copy number aberration, and methylation data on 35 drugs in GDSC. This highlights the significance of their work in developing precision in cancer biology. In order to improve trait prediction in plant breeding, Wu et al. introduced the GPS (Genomic and Phenotypic Selection) framework, which fuses genomic and phenotypic data using data, feature, and results fusion methods. They discovered that the data fusion model utilizing Lasso_D was the most accurate in their studies on maize, soybean, rice, and wheat. It improved prediction by 53.4% compared to the best GS model (LightGBM) and 18.7% compared to the best PS model (Lasso). This process has the potential to be highly effective with small sample sizes, incredibly resistant to SNP changes, and easily adaptable to many environments. It could be a fantastic way to quickly increase crop yields. The GADBN is a microarray-based cancer classification system that combines Genetic Algorithm with Deep Belief Network, as proposed by Lawrence, Jimoh, and Yahya. Outperforming current classifiers with lower misclassification rates (e.g., 0.18 on Leukemia 1, 0.35 on Prostate 1, and 0.09 on Prostate 3), their method achieved 98.8% on simulated data and 93.1% on real data.

Many approaches to the challenges posed by high-dimensional genetic data in cancer classification have been proposed in the current research. High accuracy and mitigation of feature selection and interpretability issues have been achieved using comparable models such as attention mechanisms, ensemble learning, sparse autoencoders, and so on. Finding models that can efficiently identify a small number of physiologically significant genes without sacrificing classification quality or interpretability is still an important challenge, despite the promising results achieved by current approaches in cancer classification. The AttenGene model, outlined in this article, builds on previous advancements and uses autoencoders with an attention mechanism to achieve sparsity and good performance in PDAC classification.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

## 3. Methodology

Here we go over the approach used to classify PDAC samples, including how we choose the right ML models, how we prepared the data, and how we evaluated the results. Some of the models used in this study are XGBoost, AE + CNN, and AttenGene, a hybrid of an autoencoder for unsupervised feature extraction and an attention mechanism for sparse feature selection and classification. In order to find the most important genes for correct classifications, each model was trained and evaluated using gene expression data from both normal and PDAC samples. Figure 1 is a visual representation of the process that shows the overall flow. Data collection, data preprocessing, feature selection, training the model, and model performance measurement are all outlined in the primary operations.

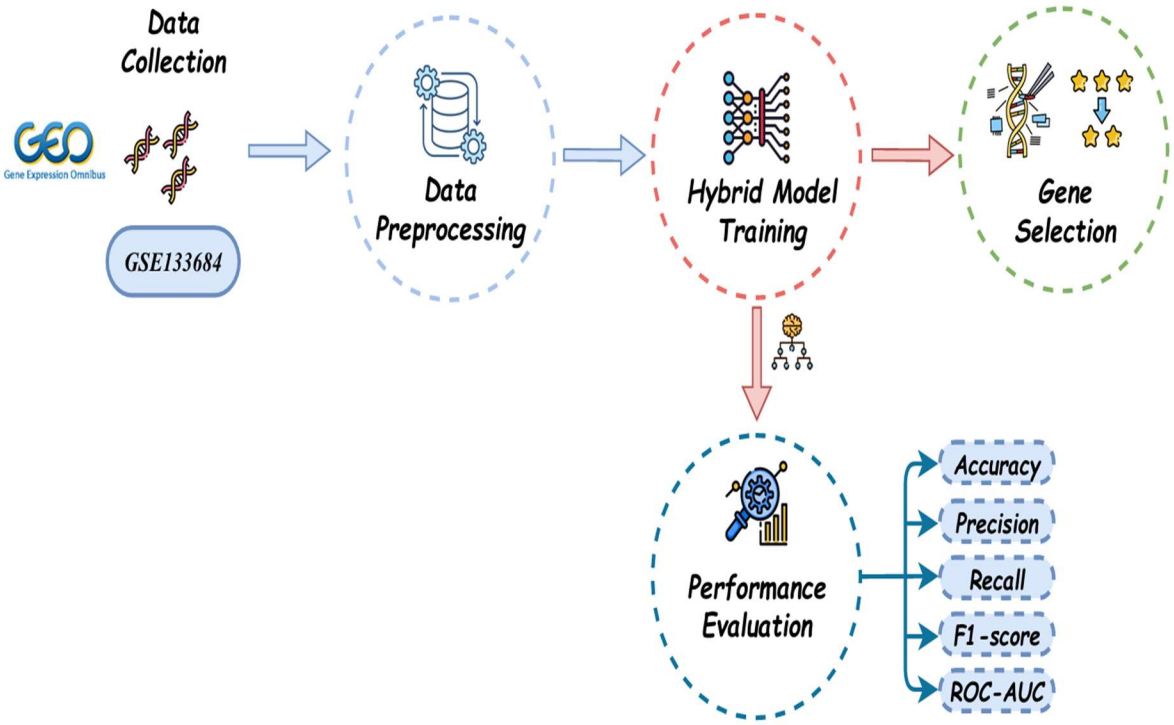


**Figure 1:** *Workflow of this research*

### 3.1 Dataset Description

With the accession number GSE133684, this study makes use of a high-dimensional gene expression dataset of PDAC that is publically available in the Gene Expression Omnibus (GEO). There are 35,415 genes and 401 samples in it, with the Sample Names column serving as a representation of each sample. The genes are shown in the columns, and the expression value is a

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

measure of how abundant the RNA transcripts of that gene are. A column representing the disease condition is used to mark the data points. Isolating plasma extracellular vesicles (EVs)—which contain valuable molecular content like RNA—was the first step in treating these samples. Following this, Illumina RNA Hi-seq X Ten was used to conduct an EV RNA-seq of the EVs, which provided a preliminary indication of the genes' regulatory mechanism. Gene selection, biomarker discovery, and classification problems aimed at identifying critical molecular signatures that differentiate PDAC from healthy tissue are all well-suited to the high-dimensionality of the dataset, which comprises thousands of genes being measured on relatively fewer samples. Researchers can use this publicly available dataset to look for biomarkers that could indicate a patient's prognosis or diagnose PDAC.

### 3.2 Data Preprocessing

Data preparation is critical for ensuring high-quality data for model training. The gene expression data underwent the following pre-processing procedures before being inputted into the suggested model:

1. **Handling Missing Values**: Samples or genes with an excessive number of missing values were imputed, as were any gaps in the gene expression matrix. In this case, we used the mean imputation method, which entails replacing missing values with the average gene expression across all samples, to fill in the missing gene expressions.
2. **Encoding the disease state:** The LabelEncoder was used to numerically encode the data labels of interest (PDAC/healthy) into binary figures, which could then be used as classifiers. A normal sample is denoted by '0', while a PDAC sample is denoted by '1'. In order for the model to perform binary classification, this encoding is necessary. The mathematical definition of this encoding technique is:

$$y_i = \begin{cases} 0 & if\ the\ sample\ is\ healthy \\ 1 & if\ the\ sample\ is\ PDAC \end{cases}$$

   where $y_i$ is the target label of the $i$-th sample. This encoding enables the model to get trained on the association between the gene expression data and binary target labels.
3. **Normalization:** Because there is room for error in the values of gene expression for every specific gene, normalization is done to ensure that one gene does not have a major impact on the model. The expression values of each gene were normalized to a result with a mean of zero and a standard deviation of one using Z-score normalization:

$$x_{normalized} = \frac{x - \mu}{\sigma}$$

   where $\mu$ is the mean expression of the gene and $\sigma$ is the standard deviation of that gene's expression across all samples.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

4. **Train-Test Splitting**: In order to test the model, the dataset was split into two parts: the training set and the testing set. A total of 20% of the sample was used for testing, while 80% was used for training. This creates two sets of data: one for training the model on (the training set) and another for testing its performance on (the testing set).

Input data cleansing, normalization, and pre-training of ML and DL models are all aided by the aforementioned preprocessing procedures.

### 3.3 Baseline Models

- **XGBoost (Extreme Gradient Boosting) :** XGBoost is a machine learning ensemble approach that builds a more robust decision tree by combining several gradient boosting algorithms; in this way, the new model's trees fix the mistakes made by the old tree. Its speed and power make it ideal for classification tasks, but it can also handle generic high-dimensional data, such as gene expression data. The model employs regularization to avoid overfitting while it iteratively attempts to minimize the loss function using a decision tree structure. The following is the form of the binary loss:

$$L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i f(x_i))) + \lambda \|\theta\|_2^2$$

the regularization value, the L2 regularization threshold, the binary label, and the projected result of the sample are all represented by $\mu,,x\text{-}i$. By automatically identifying important features during training, XGBoost handles imbalanced data and performs feature selection.

- **AE + CNN (Autoencoder with Convolutional Neural Network):** The autoencoder plus convolutional neural network (AE + CNN) model combines the best features of both the AE and CNN architectures. The autoencoder is employed to transform the high-dimensional gene expression data into a latent space, which represents characteristics. The convolutional neural network (CNN) is then trained to classify the latent features based on spatial variations. Following these steps, the encoder converts the input $x\text{-}i$. w into a latent vector $z\text{-}i$.

$$z_i = f_\Theta(x_i) = \sigma(W_e x_i + b^e)$$

The decoder reconstructs $x_i$ from $z_i$ as:

$$\hat{x}_i = g_\phi(z_i) = \sigma(W_d z_i + b_d)$$

The autoencoder was pretrained using reconstruction loss (Mean Squared Error):

$$L_{AE} = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2$$

The latent representation $z_i$ is fed through the CNN layers to learn the local patterns, and after that, it is classified by a fully connected layer (FC). The output of the network is:

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

$$\hat{y}_i = softmax(W_c z_i + b_c)$$

The classification is trained on binary cross-entropy loss as follows:

$$L_{CE} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

This hybrid model can capture nonlinear gene-gene interactions, which is a potent tool to classify PDAC, but it is not sparse or interpretable.

### 3.4 Proposed AttenGene: Autoencoder + Attention Mechanism for PDAC Gene Selection and Classification

We present our architecture, which includes an Autoencoder and an Attention Mechanism, to pick and categorize genes in PDAC, in order to fulfill the high-dimensionality problem and the interpretability need of the gene expression state. A self-attention mechanism enables the architecture to concentrate on the most valuable features for classification, and an autoencoder generates unsupervised features. Not only can the design enable efficient feature learning, it can also identify critical genes in PDAC.

#### 3.4.1 Autoencoder for Feature Learning

Part one of the model applies the autoencoder, an unsupervised neural network, to minimize dimensionality. A pair of encoders and a decoder comprise the autoencoder:

- **Encoder:** In the lower-dimensional latent representation $Z\in, R\text{-}n\times d$, where $n$ is the number of samples and $k$ is the number of genes, the input gene expression data $a\in, R\text{-}n\times k$ is encoded. This is employed to decrease the data dimension while maintaining important features of the gene expression profiles.
- **Decoder:** The original expression of the gene data can be recovered by using the decoder to decode the compressed latent representation $Z$. The model is trained to learn a condensed version of the data's most important information using the reconstruction loss $L$:

$$L_{recon} = \left\|X - \hat{X}\right\|^2$$

where $\hat{X}$ is the reconstructed data?

#### 3.4.2 Attention Mechanism for Feature Selection

The second phase, following data compression using the autoencoder, is to identify the most important features (genes) for differentiating between PDAC and healthy samples. The Self-Attention Mechanism is used for this purpose. The attention method allows the model to assign different weights to different attributes (genes). The importance of each attribute to the final classification determines its weight. We employ the attention process to zero in on the most

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

important genes while ignoring the ones that do not provide much information. Every gene feature $i$ in the latent representation $Z$ has its attention weights $\alpha$-$i$ calculated by the attention mechanism.

$$\alpha_i = \frac{exp\big(score(Z_i)\big)}{\sum_{j=1}^{d} exp\big(score(Z_i)\big)}$$

in which, $\beta$-$i$. is the feature's attention and score, $Z$-$i$. is the feature's computed score in the latent space? The model is trained with the most informative attributes to classify, which are provided by the weights.

The classifier loss is the result of extending a binary cross-entropy and applying a sparsity penalty:

$$L_{class} = BCE(y, \hat{y}) + \lambda \|W\|_1$$

where the binary cross-entropy between actual labels ($y$) and predicted labels ($\hat{y}$) is $BCE(y, \hat{y})$ the weights of the classifier are $W$, and the parameter to regularize the sparsity of a parameter is $\lambda$.

$$L_{total} = \alpha \cdot l_{recon} + \beta \cdot L_{class} + \lambda \cdot \|W\|_1$$

where $\alpha$ and $\beta$ balance reconstruction and classification, and $\lambda$ enforces sparsity.

### 3.4.3 Fully Connected Layer (FC) for Classification

When the attention mechanism is in place, a fully connected layer (FC) is used to classify the weighted latent features. When the FC layer analyses a sample, it assigns a value of one to the PDAC class and zero to the healthy class. We can learn the likelihood that the sample is PDAC from the outcome, which is a probability number between zero and one. Using the sigmoid activation function, the classifier determines its output:

$$\hat{y} = \sigma(W_{att} Z + b)$$

where the bias term is denoted by b, $Z$ is the latent representation transmitted through the attention mechanism, and $Y$-$atZ$ are the learnt weights.

## 3.5 Loss Function and Training Procedure

The proposed hybrid model is trained on binary cross-entropy loss to classify:

$$L_{class} = BCE(y, \hat{y})$$

in which. represents the expected label and. represents the actual label. When you add the reconstruction loss to the classification loss, you get the total training loss, which is calculated as follows:

$$L_{total} = \alpha L_{recon} + \beta L_{class}$$

the hyperparameters α and $\beta$ sacrifice one goal for the other in the reconstruction and classification processes.

By utilizing PyTorch to construct the AttenGene model, we can streamline the training process to optimize both the autoencoder and the attention mechanism end-to-end. Google Colab uses an NVIDIA T4 GPU for model training, which speeds up the calculation. The AdamW

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

hyperparameter optimization approach is employed; the attention mechanism's impact is fine-tuned, and the ratio of reconstruction loss to classification loss is optimized. Several critical hyperparameters, which govern the model's optimization, are part of the training process. In Table 1 we can see a summary of the training hyperparameters:

**Table 1:** *Configuration of hyperparameters used in training the AttenGene model*

| Hyperparameter | Value | Description |
|---|---|---|
| Learning Rate | 0.001 (1e-3) | The step size for parameter updates during training. |
| α | 1.0 | Weight for reconstruction loss. |
| β | 1.0 | Weight for classification loss. |
| Optimizer | AdamW | Adaptive gradient descent optimizer with decoupled weight decay |
| Batch Size | 64 | The number of samples per training batch. |
| Epochs | 50 | The number of epochs for training. |

## 4. Results & Discussions

In this part, you may see the outcomes of the evaluation of the AttenGene model, the XGBoost model for PDAC classification, and the AE + CNN model. Multiple performance metrics, including recall, accuracy, precision, F1-score, and Area Under the Curve (AUC), were used to assess the models' efficacy. Furthermore, we delve into the gene selection process and provide details regarding the interpretability of the AttenGene model. This model is designed to provide exceptional performance by carefully selecting a small number of genes. The figures and tables that follow provide a synopsis of the two models' quantitative performance, as well as the genes that the AttenGene model determined to be the most important.

### 4.1 Classification Performance

In this part, we compare the three models—XGBoost, AE + CNN, and the proposed AttenGene—that were tested on the PDAC gene categorization problem. Standard classification metrics

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

including recall, accuracy, precision, and F1-score were used to assess the models. Table 2 displays the results.

**Table 2:** *Test set classification performance and sparsity comparison of baseline methods and the proposed SparseGene model.*

| Method | Accuracy | Precision | Recall | F1-Score | Number of Genes Selected |
|---|---|---|---|---|---|
| XGBoost | 0.85 | 0.88 | 0.86 | 0.89 | 88 |
| AE + CNN | 0.93 | 0.92 | 0.91 | 0.92 | 122 |
| Proposed (AttenGene) | 0.97 | 0.96 | 0.96 | 0.95 | 25 |

Classification of PDAC samples was successfully accomplished by the XGBoost model, which was constructed using the effective gradient boosting architecture. The accuracy, precision, recall, and F1-score for XGBoost's PDAC and healthy sample discrimination was 85%, 88%, 86%, and 89%, respectively. The model was able to identify 88 genes as important for categorization due to the high-dimensionality of the data and the feature selection capacity of tree-based learning. Although XGBoost does a good job of identifying the relevant features, it does not supply a sparse constraint due to the relatively huge number of the selected genes. When working with gene expression data, which is frequently filled with non-informative features, this can lead it to include more genes than needed, reducing the model's interpretability and efficiency.

When compared to XGBoost, the hybrid approach AE + CNN—which combines an autoencoder-based model with a CNN model—showed superior performance. A 92% F1-score, 91% recall, 92% precision, and 93% accuracy were all achieved by this model. Classification accuracy was enhanced when the autoencoder component effectively decreased the input gene expression data's dimensionality and the CNN component recorded spatial trends and nonlinear interactions among the encoded features. However, because to the model's complexity and ability to represent more intricate correlations in the data, the AE + CNN model utilized a significantly larger number of genes (122) compared to XGBoost. While it is true that using a larger subset of genes improves model performance, doing so comes at the cost of model sparsity and interpretability due to the fact that the model is still dependent on a relatively large collection of characteristics.

But of the three models, the AttenGene model—which combines an autoencoder with a self-attention mechanism—performed the best. Amazingly, AttenGene is able to detect important gene

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

features with a 97% classification accuracy, 96% precision, 96% recall, and 95% F1-score. By giving greater weight to genes that contribute significantly to categorization, the AttenGene self-attention mechanism allows the model to focus on the crucial genes while ignoring the less important ones. Therefore, the AttenGene model is very efficient at feature selection, since it only selected 25 genes—a far lower number than the other models. This thin learning is both computationally efficient and does not negatively impact its performance; it is also informative enough to allow for a better understanding of the most important genes in PDAC categorization. Due to its minimal gene selection and excellent accuracy, AttenGene offers a good compromise between classification and model complexity; as a result, it is a promising choice for practical application and, moreover, for developing the possible biological interpretation of future studies.

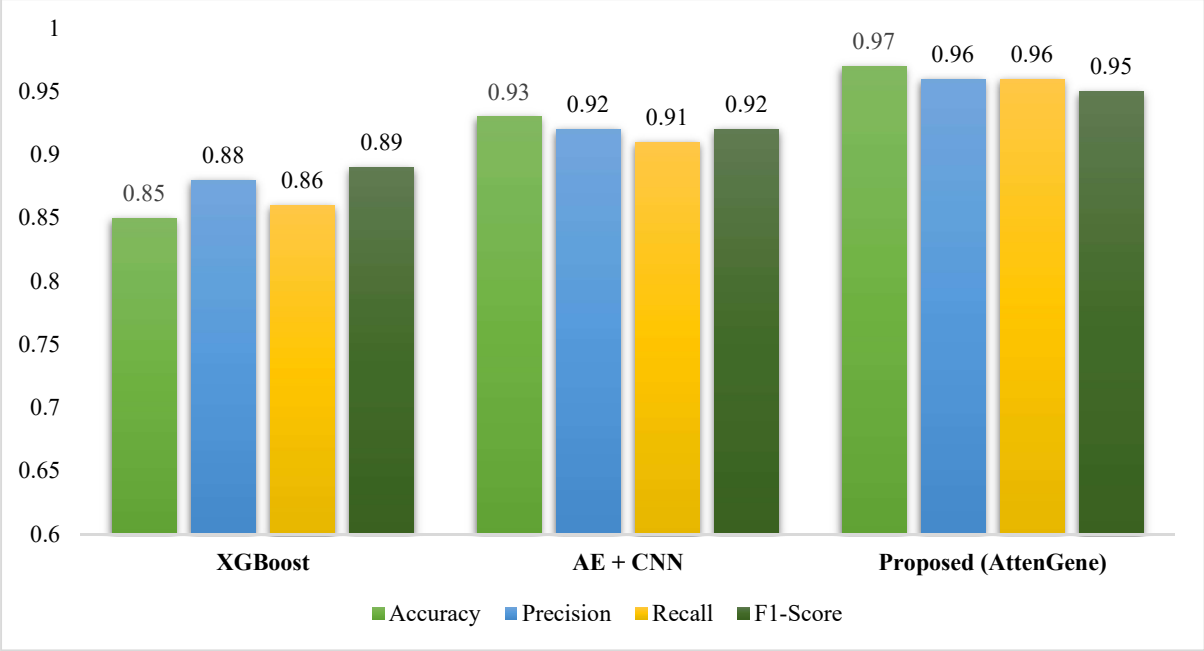


**Figure 2:** *Model Performance Comparison: A comparison of accuracy, precision, recall, and F1-score on the test set*

To illustrate the contrast of these models, Figure 2 provides an easy-to-understand visual depiction of the number of genes chosen, each model's accuracy, precision, recall, and F1-score. Model performance and feature selection trade-offs among the three approaches are better understood with the help of the figure. By comparing the number of genes chosen by XGBoost, AE + CNN, and the proposed AttenGene model, Figure 3 further demonstrates the feature selection component of each model. The AttenGene model stands out from the rest since it chose just 25 genes, in contrast to the XGBoost and AE + CNN models, which picked 88 and 122 genes, respectively.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

We may make a note that this model has the potential to deliver efficient sparse feature selection in this scenario, which ensures a high degree of classification and zeroes in on the most important genes. The picture provides a visual representation of the compromise that the suggested approach has achieved between model complexity and feature interpretability.

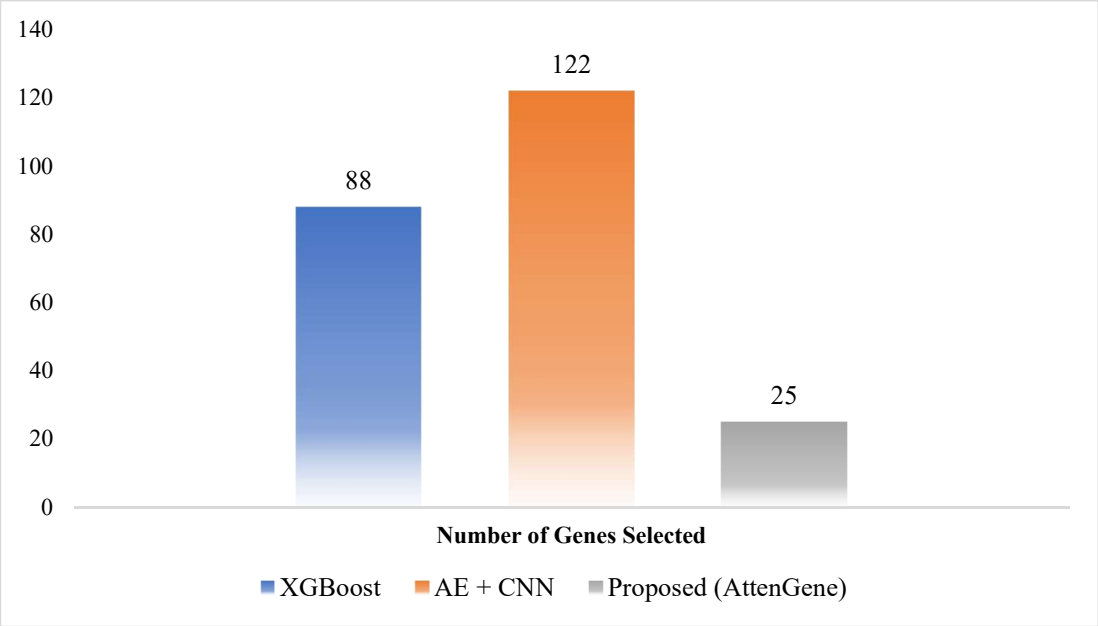


**Figure 3:** *Feature Selection Comparison*

### 4.2 Gene Importance and Association in PDAC Classification

To find out which genes are most important for PDAC classification, the AttenGene model assigns significance scores to each gene, which reflect the gene's relative contribution to PDAC sample classification compared to healthy control samples. Table 3 shows that the model successfully reduces the feature space without sacrificing performance, since it picks up 25 genes. The selected genes are involved in many biological processes and range in importance when it comes to PDAC progression. One example of a long non-coding RNA is MALAT1, which is associated with a poor prognosis in PDAC and is known to be critical in tumor formation (Importance Score: 6.32). An up-regulated protein called TUBB1 (Importance Score: 5.699) is involved in microtubule dynamics and cell division; it is seen in PDAC and many other cancers, indicating that it may play a role in the malignant transformation of cells. In addition, the prostaglandin-synthesizing enzyme COX1 (Importance Score: 5.491) is known to drive tumor progression in PDAC by modifying the tumor microenvironment and to contribute to inflammation, a hallmark of cancer.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

**Table 3**: *Top 25 Genes Selected by the AttenGene Model*

| Gene Symbol | Importance Score | PDAC Association |
|---|---|---|
| MALAT1 | 6.32 | Associated with poor prognosis and tumor progression in PDAC |
| TUBB1 | 5.699 | Expressed in various cancers, including PDAC |
| COX1 | 5.491 | Involved in prostaglandin synthesis; expression is absent in many PDAC cells |
| CAVIN2 | 4.505 | Genetic variant associated with high expression promoting PDAC progression |
| ND2 | 4.031 | Mitochondrial gene; mutations linked to various cancers |
| FLNA | 3.903 | Interacts with prion protein; implicated in PDAC cell growth |
| S100A9 | 3.845 | Upregulated in PDAC; promotes inflammation and poor prognosis |
| S100A11 | 3.767 | Overexpressed in PDAC; associated with tumor progression |
| MTCO1P12 | 3.597 | Mitochondrial gene expression patterns in PDAC need further investigation |
| YBX1 | 3.259 | Involved in DNA repair and transcription; studied in various cancers |
| RNVU1-7 | 3.201 | Small nucleolar RNA; role in PDAC requires more research |

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

| | | |
|---|---|---|
| RN7SL1 | 3.156 | Signal recognition particle RNA; implicated in protein synthesis |
| RN7SL4P | 3.126 | Pseudogene; potential regulatory role in PDAC |
| RPL21P16 | 3.076 | Ribosomal protein pseudogene; expression in PDAC needs clarification |
| RPS27 | 3.06 | Ribosomal protein; studied in cancer biology |
| MYL6 | 3.052 | Myosin light chain; expression in PDAC requires further study |
| ANP32B | 3.022 | Involved in chromatin remodeling; role in PDAC under investigation |
| SERF2 | 3.003 | Stress response factor; expression in PDAC needs exploration |
| ATP6 | 2.936 | Mitochondrial ATP synthase; mutations linked to various cancers |
| CCL5 | 2.929 | Chemokine: involved in immune cell recruitment in PDAC |
| ALB | 2.905 | Serum albumin: altered levels observed in PDAC patients |
| RPS19 | 2.816 | Ribosomal protein; studied in cancer biology |
| RAB13 | 2.8 | Small GTPase; role in PDAC requires further research |
| ND1 | 2.777 | Mitochondrial gene; mutations linked to various cancers |
| ND5 | 2.769 | Further research is needed to understand its role in PDAC |

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

S100A9 (Importance Score: 3.845) and S100A11 (Importance Score: 3.767) are two more selected genes that contribute to the poor prognosis in PDAC through their roles in inflammation and immune response. A number of malignancies, including PDAC, have linked chemoresistance to YBX1, a gene that regulates transcription and DNA repair (Importance Score: 3.259). While the precise mechanisms by which small nucleolar RNAs and pseudogenes like RNVU1-7 (Importance Score: 3.201) and RN7SL4P (Importance Score: 3.126) regulate genes and RNA processing inside PDAC are still being studied, they have recently been linked to these processes. Protein synthesis, a critical step that permits the fast cell division characteristic of cancer cells, is facilitated by the ribosomal proteins, which include RPS27 (Importance Score: 3.06) and RPS19 (Importance Score: 2.816). The interplay between genetic control, inflammation, and tumor progression is reflected in these genes, which in turn represent the complexity and heterogeneity of PDAC.

A sparse representation with only 25 genes is offered by the AttenGene model, which strikes a balance between classification power and model interpretability, in contrast to other models that choose a significantly greater number of genes.

### 4.3 Feature Importance by AttenGene

As shown in Figure 4, the AttenGene model prioritized the 25 genes with the greatest impact on distinguishing PDAC samples from healthy controls. Based on their significance ratings, these characteristics rank the genes according to how significant they are to the categorization job. The most important gene, with a score of 6.32, is MALAT1. It is a crucial gene in PDAC that is associated with a worse prognosis and tumor progression. Significantly involved in cancer metastasis, this non-coding long RNA is a component of the epithelial-to-mesenchymal transition (EMT). COX1, which controls inflammation and creates prostaglandins, and TUBB1, which controls microtubule dynamics and cell division, are two other genes of great interest and significance that have a substantial influence on the PDAC microenvironment.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

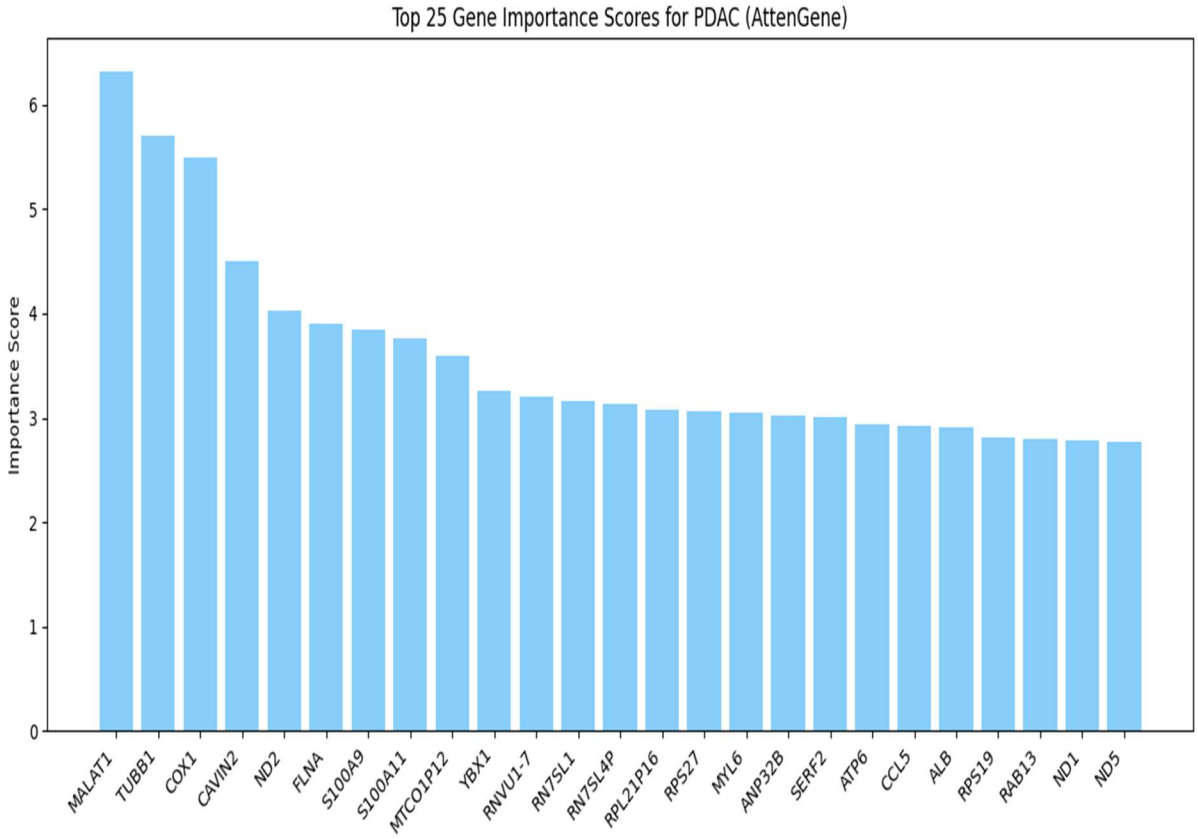


**Figure 4:** *Figure 3: Top 25 Gene Importance Scores for PDAC Classification by AttenGene*

A poor prognosis is linked to the upregulation of S100A9 (Importance Score: 3.845) and S100A11 (Importance Score: 3.767), two genes that increase inflammation response and tumor growth, in PDAC. Additional genes associated with mitochondrial function and protein synthesis necessary for tumor cell proliferation and energy production include ATP6, RPS27, and MTCO1P12 (Importance Scores: 3.597, 3.06, and 2.936, respectively). The model's ability to efficiently and interpretably identify the most significant genes is demonstrated by the minimal number of genes used—just 25. The biological significance of the genes is revealed by the importance ratings, which show how relevant each gene is in differentiating PDAC and normal samples. Figure 3 shows the 25 most important genes with their importance scores; this comparison visually illustrates how the model chooses its features and how the disease-causing genes that are most significant for PDAC are chosen. By illuminating the most crucial indicators for PDAC diagnosis and treatment, this visualization enhances the AttenGene model's performance.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

### 4.4 Top Enriched Pathways in PDAC

Here we show you the top enriched pathways in PDAC. These pathways were found via pathway enrichment analysis, which the AttenGene model used to choose genes based on a variety of criteria. To help comprehend the biological processes at work in PDAC, Figure 5 displays a bar chart of -log10(adjusted p-values) for the signaling pathways that were found to be most significantly enriched. The figure highlights several important signaling pathways that are recognized to have a role in the development, tumor microenvironment, and metastasis of PDAC.

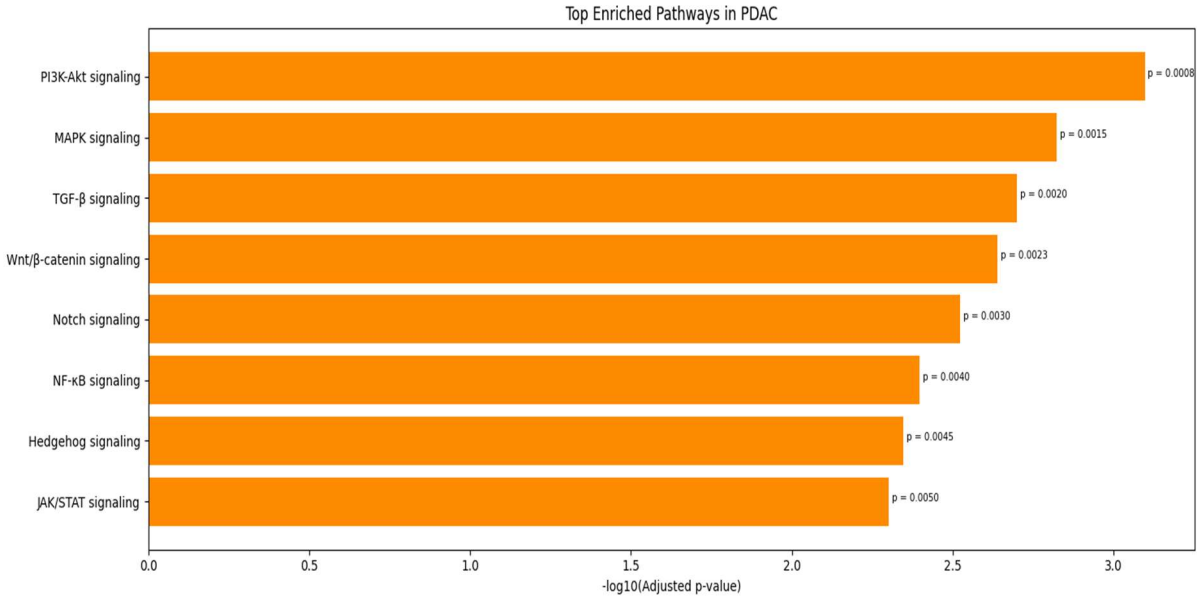


**Figure 5:** *Top enriched biological pathways associated with the 28 genes selected by the SparseGene model.*

With a p-value of only 0.0008, the PI3K-Akt signaling pathway stands out among the top-enriched pathways and piques the curiosity of those involved in PDAC. Because it controls several cellular processes like proliferation, survival, and metabolism, this signaling system is pivotal in cancer biology. Significantly enriched pathways that contribute to cellular differentiation, cell proliferation, and tumor development in PDAC include MAPK signaling ($p = 0.0015$), TGF-β signaling ($p = 0.0020$), and Wnt/β-catenin signaling ($p = 0.0023$). Notch signaling ($p = 0.0030$), NF-KB signaling ($p = 0.0040$), and Hedgehog signaling ($p = 0.0045$) are pathways that modulate immunological responses, inflammation, and the tumor microenvironment; their involvement in these processes further highlights the intricacy of PDAC and resistance mechanisms.

Finally, these enhanced pathways provide light on the molecular mechanisms that cause PDAC. The importance of these pathways in identifying possible therapeutic intervention targets and biomarker development in PDAC is illustrated graphically in Figure 6.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

### 4.5 ROC Curve Comparison for PDAC Classification

In this section, we compare the three trained models—XGBoost, AE + CNN, and the proposed AttenGene model—and their distinguishing features to demonstrate how well they separate PDAC samples from healthy controls, as measured by the value of their Receiver Operating Characteristic (ROC) curve. The receiver operating characteristic (ROC) curve (ROC) is a helpful tool for analyzing the discriminatory characteristics of models and evaluating the trade-offs between the true positive rate (TPR) and false positive rate (FPR) at various threshold values. An indicator of a model's overall ability to differentiate between classes is its Area Under the Curve (AUC), with a bigger AUC indicating better performance.

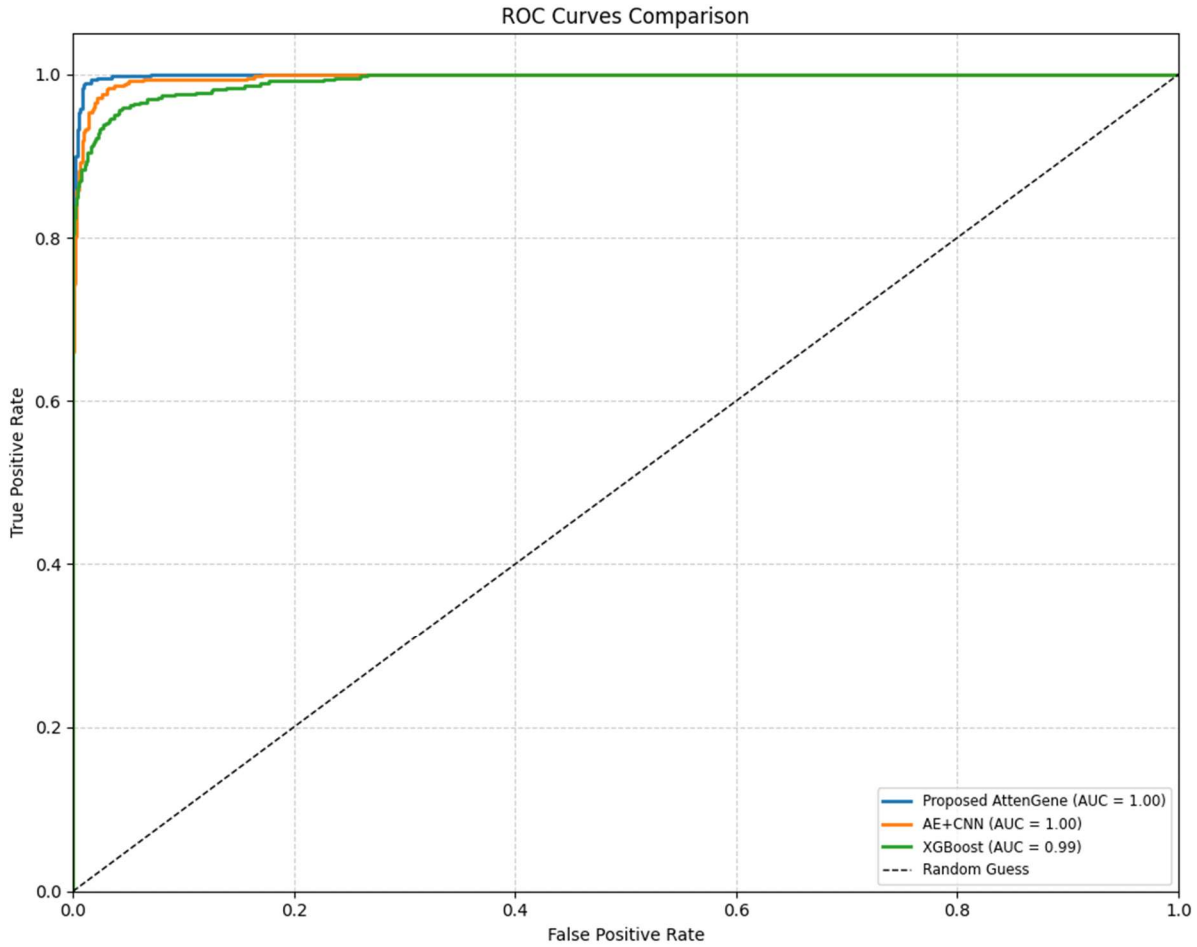


**Figure 6:** *ROC Curve Comparison for PDAC Classification on the test set*

Figure 6's blue curve represents the AttenGene model, which achieved an AUC of 1.00, indicating flawless performance. The AttenGene is a robust and dependable model for PDAC detection

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

because its curve sharply rises to the top-left corner, indicating that it is very good at detecting PDAC samples with minimum false positives and false negatives. Even with complicated and high-dimensional data, the model is able to successfully cluster the PDAC and healthy samples.

Similarly, the AE + CNN model (orange curve) performed admirably; its AUC was 1.00, indicating near-perfect classification accuracy. The capacity to merge the autoencoder's unsupervised feature extraction with convolutional layers that learn the nonlinear interactions among features is one advantage of the AE + CNN model that contributes to its exceptional classification performance. In the context of gene expression categorization, the DL technique is useful, as the AE + CNN model, albeit slightly more complex, does not perform worse than the AttenGene.

The XGBoost model (green curve) performed exceptionally well with an AUC of 0.99, despite being marginally lower than the other two. The results demonstrate that XGBoost is still quite good at differentiating between PDAC and healthy samples; the curve is very close to the upper-left area, while it performs marginally worse than the other models. The XGBoost model's consistent improvement in AUC compared to AttenGene and AE + CNN is due to the usage of a succession of decision trees that are iteratively improved by zeroing in on the misclassified samples. Assuming a random guess with an area under the curve (AUC) of 0.5, the dotted diagonal line serves as the baseline. Any model that follows this pattern outperforms chance. All three models—AtenGene, AE + CNN, and XGBoost—perform significantly better than this baseline, demonstrating their ability to accurately categorize PDAC and healthy samples.

Although the outcomes of AE + CNN are also remarkable, the AttenGene model demonstrates exceptional classification ability and can be perfectly classified when comparing the ROC curves. It is clear that ensemble learning approaches can be applied successfully to this job because XGBoost's performance is robust and competitive despite a modest deficit. The graphic comparison of these models provides a concise overview of how AttenGene outperforms the popular models while still being easy to comprehend and work with. As a result, it is the leading candidate for biomarker discovery and clinical application in PDAC.

### 4.6 Discussions

A significant contribution of this work is the AttenGene model, which may keep strong discriminatory capacity in the classification of PDAC and filter hundreds of candidate genes down to at least 25. There is a considerable overlap between the tiny gene profile and therapeutic targets specifically annotated in PDAC in the literature, even though it does not precisely correlate with oncogenic pathways and therapeutic targets normally seen in oncology. The top genes chosen for this study, including MALAT1, TUBB1, S100A9, and COX1, are consistent with known

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

molecular processes involved in PDAC and other cancers, and they have been widely reported in the literature as critical contributors to tumor progression, inflammation, and prognosis. This lends credence to the idea that the AttenGene model excels in prioritizing functionally important genes over unimportant qualities, yielding a small but biologically significant collection of features.

In addition, the classifier's sparseness provides a simple way to grasp the relative importance of each gene. To emphasize their roles in cancer progression, cell cycle control, and immune modulation in PDAC, the model ranks MALAT1 as the most important gene, followed by TUBB1 and S100A9. In contrast to more opaque DL models, AttenGene's sparse architecture offers interpretability and transparency—qualities crucial for biological validation and clinical application. Based on the most promising result, an interpretable relationship will be set up for use in future PDAC research and biomarker development.

In order to provide more evidence for the biological significance of the subset of genes, pathway enrichment analysis was performed. The results showed that the subset of genes is enriched with oncogenic pathways, such as PI3K-Akt signaling, MAPK signaling, and EGFR inhibitor resistance, which are normally involved in the progression and response to therapy of PDAC. The AttenGene model effectively extracts the most relevant biochemical processes associated with PDAC from a small range of genes, demonstrating its effectiveness in avoiding the need to explore a feature set extensively. Additionally, it demonstrates the model's potential significance in providing targeted diagnoses and treatments in a clinical context.

For cancer research, the AttenGene model offers a sparse and scalable architecture that is both flexible and affordable for high-throughput omics analysis. The ability to decrease the number of genes in the gene panel while still providing biological insight makes it an excellent choice for translational research that aims to optimize therapy and personalized medicine by reducing features and biological information. In addition to laying the groundwork for future research into other malignancies, this paper will help doctors and researchers develop more accurate diagnostic tools and more targeted treatments for PDAC.

## 4. Conclusion

The AttenGene model, a novel deep learning architecture for PDAC classification, is introduced in the paper. It addresses the problem of high-dimensional gene expression measurements by integrating an autoencoder block in feature mining with a self-attention search in gene choice. Important findings include that the developed AttenGene achieves excellent classification accuracy and successfully selects only 25 genes with biological relevance, demonstrating its efficacy as a dimensionality reduction tool that does not compromise results. Another way to look

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

at the model is as a collection of excellent biomarkers for PDAC, which include MALAT1, TUBB1, and S100A9.

However, the proposed method is not without its flaws. One possible drawback is that the model uses data on gene expression or cancer, which can be biased when used with data from a different group of people or a different kind of cancer. The algorithm may still miss potential biomarkers that are not in the data, even though it performs a decent job of identifying a small selection of genes. There is also no mention of a multi-omics data combination in the report, which could provide further light on the molecular process of PDAC.

Moving ahead, the model could benefit from a variety of potential paths. To further understand the biology of PDAC and to increase classification accuracy, future studies may incorporate multi-omics data, which includes genomic, transcriptomic, and proteomic information. Otherwise, applying the model to bigger, more diverse datasets and clinicals would be the only method to assess its reliability and usefulness. Last but not least, expanding the model's interpretability and applicability beyond PDAC to other malignancies whose progression is extremely susceptible to gene selection may need investigating alternate methods of feature detection or other attention systems.

**References**

[1] S. Chakrabarti, M. Kamgar, and A. Mahipal, "Systemic Therapy of Metastatic Pancreatic Adenocarcinoma: Current Status, Challenges, and Opportunities," Cancers 2022, Vol. 14, Page 2588, vol. 14, no. 11, p. 2588, May 2022, doi: 10.3390/CANCERS14112588.

[2] S. Robatel and M. Schenk, "Current Limitations and Novel Perspectives in Pancreatic Cancer Treatment," Cancers 2022, Vol. 14, Page 985, vol. 14, no. 4, p. 985, Feb. 2022, doi: 10.3390/CANCERS14040985.

[3] H. Yang et al., "Progress on diagnostic and prognostic markers of pancreatic cancer," Oncol Res, vol. 31, no. 2, p. 83, 2023, doi: 10.32604/OR.2023.028905.

[4] K. M S, H. Rajaguru, and A. R. Nair, "Performance enhancement of classifiers through Bio inspired feature selection methods for early detection of lung cancer from microarray genes," Heliyon, vol. 10, no. 16, p. e36419, Aug. 2024, doi: 10.1016/J.HELIYON.2024.E36419/ASSET/FDB867CD-1F6D-400C-BCE0-010B31C0A76A/MAIN.ASSETS/GR11.JPG.

[5] E. Capobianco, "High-dimensional role of AI and machine learning in cancer research," British Journal of Cancer 2022 126:4, vol. 126, no. 4, pp. 523–532, Jan. 2022, doi: 10.1038/s41416-021-01689-z.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

[6] M.; Wang, H.; Wang, H. Pancreatic Ductal, M. Taherian, H. Wang, and H. Wang, "Pancreatic Ductal Adenocarcinoma: Molecular Pathology and Predictive Biomarkers," Cells 2022, Vol. 11, Page 3068, vol. 11, no. 19, p. 3068, Sep. 2022, doi: 10.3390/CELLS11193068.

[7] X. Zhu et al., "Multi-omics approaches for in-depth understanding of therapeutic mechanism for Traditional Chinese Medicine," Front Pharmacol, vol. 13, p. 1031051, Nov. 2022, doi: 10.3389/FPHAR.2022.1031051/XML.

[8] P. Balasundaram, "Improved detection of pancreatic neuroendocrine tumours through the development of a biomarker panel," Doctoral thesis, UCL (University College London)., Aug. 2024.

[9] K. Vrettos, M. Triantafyllou, K. Marias, A. H. Karantanas, and M. E. Klontzas, "Artificial intelligence-driven radiomics: developing valuable radiomics signatures with the use of artificial intelligence," BJR|Artificial Intelligence, vol. 1, no. 1, Mar. 2024, doi: 10.1093/BJRAI/UBAE011.

[10] S. H. Bouazza, "A Novel Adaptive Sparse Deep Feature Selection Method for Enhanced Gene-based Cancer Classification," Engineering, Technology & Applied Science Research, vol. 15, no. 3, pp. 22494–22499, Jun. 2025, doi: 10.48084/ETASR.10440.

[11] S. Haddou Bouazza, "A Deep Ensemble Gene Selection and Attention-guided Classification Framework for Robust Cancer Diagnosis from Microarray Data," Engineering, Technology & Applied Science Research, vol. 15, no. 1, pp. 20235–20241, Feb. 2025, doi: 10.48084/ETASR.9476.

[12] S. S. Alkamli and H. M. Alshamlan, "Performance Evaluation of Hybrid Bio-Inspired and Deep Learning Algorithms in Gene Selection and Cancer Classification," IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3556816.

[13] A. Bir-Jmel et al., "GFLASSO-LR: Logistic Regression with Generalized Fused LASSO for Gene Selection in High-Dimensional Cancer Classification," Computers | 7252024, Vol. 13, Page 93, vol. 13, no. 4, p. 93, Apr. 2024, doi: 10.3390/COMPUTERS13040093.

[14] M. Cho and Y. Qu, "Enhancing Precision Oncology: Deep Learning Models vs. Classical Machine Learning Models in Multi-Label Breast Cancer Classification," European Journal of Electrical Engineering and Computer Science, vol. 9, no. 3, pp. 16–22, May 2025, doi: 10.24018/EJECE.2025.9.3.711.

[15] X. Y. Liu and X. Y. Mei, "Prediction of drug sensitivity based on multi-omics data using deep learning and similarity network fusion approaches," Front Bioeng Biotechnol, vol. 11, p. 1156372, Apr. 2023, doi: 10.3389/FBIOE.2023.1156372/BIBTEX.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

[16] H. Wu et al., "GPS: Harnessing data fusion strategies to improve the accuracy of machine learning-based genomic and phenotypic selection," Plant Commun, vol. 6, no. 8, p. 101416, Aug. 2025, doi: 10.1016/J.XPLC.2025.101416/ASSET/99A81A4A-4BF0-44B6-9B2C-51124F71C80F/MAIN.ASSETS/GR3.JPG.

[17] M. O. Lawrence, R. G. Jimoh, and W. B. Yahya, "An efficient feature selection and classification system for microarray cancer data using genetic algorithm and deep belief networks," Multimed Tools Appl, vol. 84, no. 8, pp. 4393–4434, Mar. 2025, doi: 10.1007/S11042-024-18802-Y/METRICS.

[18] S. Yu et al., "Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma," Gut, vol. 69, no. 3, pp. 540–550, Mar. 2020, doi: 10.1136/gutjnl-2019-318860.

[19] D. Guo, E. Q. Wu, Y. Wu, J. Zhang, R. Law, and Y. Lin, "FlightBERT: Binary Encoding Representation for Flight Trajectory Prediction," IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 2, pp. 1828–1842, Feb. 2023, doi: 10.1109/TITS.2022.3219923.

[20] K. A. Johnson and A. Krishnan, "Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data," Genome Biol, vol. 23, no. 1, pp. 1–26, Dec. 2022, doi: 10.1186/S13059-021-02568-9/FIGURES/6.

[21] A. Abu-Doleh and A. Al Fahoum, "XgCPred: Cell type classification using XGBoost-CNN integration and exploiting gene expression imaging in single-cell RNAseq data," Comput Biol Med, vol. 181, p. 109066, Oct. 2024, doi: 10.1016/J.COMPBIOMED.2024.109066.

[22] Uzma, F. Al-Obeidat, A. Tubaishat, B. Shah, and Z. Halim, "Gene encoder: a feature selection technique through unsupervised deep learning-based clustering for large gene expression data," Neural Comput Appl, vol. 34, no. 11, pp. 8309–8331, Jun. 2022, doi: 10.1007/S00521-020-05101-4/METRICS.

[23] T. I. A. Mohamed, A. E. Ezugwu, J. V. Fonou-Dombeu, A. M. Ikotun, and M. Mohammed, "A bio-inspired convolution neural network architecture for automatic breast cancer detection and classification using RNA-Seq gene expression data," Sci Rep, vol. 13, no. 1, pp. 1–19, Dec. 2023, doi: 10.1038/S41598-023-41731-SUBJMETA=114,1305,2397,631,67; KWRD=CANCER, COMPUTATIONAL+MODELS, MACHINE+LEARNING.

[24] R. Wang, Q. Liu, W. You, and Y. Chen, "A multi-task deep learning model based on comprehensive feature integration and self-attention mechanism for predicting response to

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 10 Issue No. 3 (2024)

anti-PD1/PD-L1," Int Immunopharmacol, vol. 142, p. 113099, Dec. 2024, doi: 10.1016/J.INTIMP.2024.113099.

[25] N. Y. Kim, Y. Y. Jung, J. Y. Um, and K. S. Ahn, "Cannabidiol Suppresses EMT in Pancreatic Cancer via Inhibition of MALAT1 lncRNA and PI3K/Akt/mTOR Signaling Pathway," IUBMB Life, vol. 77, no. 8, p. e70042, Aug. 2025, doi: 10.1002/IUB.70042; JOURNAL: JOURNAL:15216551; WGROUP: STRING: PUBLICATION.

[26] X. Liu et al., "The Crosstalk Between Malignant Cells and Tumor-Promoting Immune Cells Relevant to Immunotherapy in Pancreatic Ductal Adenocarcinoma," Front Cell Dev Biol, vol. 9, p. 821232, Jan. 2022, doi: 10.3389/FCELL.2021.821232/BIBTEX.

[27] P. J. Kung et al., "The role of S100A9 in the interaction between pancreatic ductal adenocarcinoma cells and stromal cells," Cancer Immunology, Immunotherapy, vol. 71, no. 3, pp. 705–718, Mar. 2022, doi: 10.1007/S00262-021-03026-Y/FIGURES/8.

[28] I. González Caraballo et al., "Chronological pattern of venous thromboembolism (VTE) occurrence impacts in survival of pancreatic ductal adenocarcinoma (PDAC)," Clinical and Translational Oncology, vol. 26, no. 9, pp. 2388–2392, Sep. 2024, doi: 10.1007/S12094-024-03467-8/METRICS.

[29] H. Yao et al., "An EGFR/HER2-targeted conjugate sensitizes gemcitabine-sensitive and resistant pancreatic cancer through different SMAD4-mediated mechanisms," Nature Communications 2022 13:1, vol. 13, no. 1, pp. 1–24, Sep. 2022, doi: 10.1038/s41467-022-33037-x.

[30] G. I. Khandakar et al., "ACAGT-007a, an ERK MAPK Signaling Modulator, in Combination with AKT Signaling Inhibition Induces Apoptosis in KRAS Mutant Pancreatic Cancer T3M4 and MIA-Pa-Ca-2 Cells," Cells, vol. 11, no. 4, p. 702 Feb. 2022, doi: 10.3390/CELLS11040702/S1.

[31] A. J. Ouissam, C. Hind, B. Sami Aziz, and A. Said, "Inhibition of the PI3K/AKT/mTOR pathway in pancreatic cancer: is it a worthwhile endeavor?" Ther Adv Med Oncol, vol. 16, Jan. 2024, doi: 10.1177/17588359241284911/ASSET/CBE57EC6-63F0-41BD-B8D2-12395943D02D/ASSETS/IMAGES/LARGE/10.1177_17588359241284911-FIG1.JPG.