Gene Expression-Based Cancer Classification Using Machine Learning for Precision Oncology

Shreejay Pandey[1], Sumathra Manokaran[2], A. H. Manjunatha Reddy[2,*]

[1]Department of Computer Science and Engineering,
R.V. College of Engineering, Bangalore 560059, Karnataka, India
[2]Department of Biotechnology,
R.V. College of Engineering, Bangalore 560059, Karnataka, India

*Corresponding author: ahmanjunatha@rvce.edu.in
Email: shreejaypandey.cs22@rvce.edu.in

***Abstract*—Gene expression–based cancer classification plays a critical role in precision oncology, enabling early diagnosis and more targeted treatment strategies. However, the extremely high dimensionality of transcriptomics data presents significant challenges for conventional analysis techniques. In this study, an interpretable machine learning framework is proposed for the classification of cancer subtypes using gene expression data obtained from The Cancer Genome Atlas (TCGA). Dimensionality reduction methods, including Principal Component Analysis (PCA) and autoencoder-based feature extraction, are employed to obtain compact and informative representations of the original dataset. These reduced features are then used to train supervised learning classifiers such as Support Vector Machine (SVM), Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN). Experimental results demonstrate that the proposed framework effectively improves classification performance, with the SVM model achieving the highest accuracy of 93.94%, followed by Random Forest with an accuracy of 92.42%. The findings indicate that combining dimensionality reduction techniques with traditional machine learning models significantly enhances predictive performance on high-dimensional gene expression data, highlighting the potential of the proposed framework to support automated cancer diagnosis in clinical settings.**

***Index Terms*—Gene Expression, Cancer Classification, Machine Learning, PCA, Autoencoder, SVM, Random Forest, Precision Oncology**

I. Introduction

Cancer remains one of the leading causes of mortality worldwide, and accurate classification of cancer subtypes is essential for early diagnosis, prognosis estimation and personalized treatment planning. Traditional diagnostic techniques rely heavily on histopathological examination of biopsy samples, which can be time-consuming and subject to inter-observer variability.

With the advent of high-throughput sequencing technologies, gene expression profiling has emerged as a powerful tool to capture molecular-level signatures of tumors by measuring the expression levels of thousands of genes simultaneously [1]. However, gene expression datasets are typically characterized by ultra-high dimensionality and relatively small sample sizes, which pose serious challenges for conventional statistical and machine learning models. To address these challenges, dimensionality reduction methods such as Principal Component Analysis (PCA) [2] and autoencoder-based models [3] have been proposed to compress high-dimensional biological data while preserving essential characteristics. These reduced feature representations enable more effective learning in high-dimensional settings.Supervised machine learning models, including Support Vector Machine (SVM)[4], Random Forest[5], Logistic Regression[6], and K-Nearest Neighbors (KNN)[7], have demonstrated strong performance in cancer subtype classification tasks. Comprehensive reviews of machine learning applications in oncology further highlight the effectiveness of classical classifiers and ensemble methods for gene expression analysis [8]. In addition, deep learning approaches have shown significant potential in capturing complex nonlinear patterns present in transcriptomic datasets [10].

In this work, a complete machine learning pipeline is developed for cancer classification using gene expression data. The pipeline integrates data preprocessing, dimensionality reduction, model training, evaluation and interpretability analysis, with the goal of producing a framework that is both accurate and clinically meaningful.

II. Related Work

Several studies have investigated the use of machine learning techniques for cancer classification using gene expression data.

Ainsworth et al. [8] provided a critical review of machine learning applications in oncology, highlighting the effectiveness of Support Vector Machines and ensemble models for high-dimensional biological data. Min et al. [10] explored deep learning approaches in bioinformatics, demonstrating that neural networks can uncover highly complex patterns within gene expression datasets. Similarly, Hinton and Salakhutdinov [3] demonstrated that autoencoders can effectively reduce dimensionality in extremely large feature spaces while preserving essential data characteristics.

Previous studies primarily focused on either traditional machine learning models or deep learning models independently. In contrast, this study integrates both dimensionality reduction

techniques and multiple classical classifiers into a unified framework, offering improved performance and interpretability for cancer subtype classification.

III. Methods

A. Data Provenance and Quality Disclosure

The gene expression dataset used in this study was obtained from The Cancer Genome Atlas (TCGA), a publicly available and highly curated cancer genomics database [1]. The data consist of RNA-Sequencing profiles collected from tumor samples across multiple cancer subtypes.

To ensure quality and consistency, several preprocessing steps were applied. Features (genes) with more than 30% missing values were removed. Remaining missing values were imputed using median imputation. The dataset was then standardized using z-score normalization so that each feature had zero mean and unit variance, reducing scale-related bias among genes.

To mitigate potential class imbalance, stratified sampling was employed to split the dataset into training and test sets, preserving the class distribution in both subsets. All models were trained and evaluated under these class-balanced conditions to ensure robust and fair performance across cancer subtypes.

TABLE I
HYPERPARAMETER SETTINGS OF THE APPLIED MACHINE LEARNING MODELS.

Model	Parameters
SVM	Kernel = RBF, C = 1.0, γ = auto
Random Forest	Trees = 100, Max depth = None
Logistic Regression	Solver = lbfgs, Max iterations = 500
KNN	K = 5, Distance = Euclidean

B. Dimensionality Reduction

1) Principal Component Analysis (PCA): Let the gene expression dataset be represented as a matrix $X \in R^{n \times d}$, where n is the number of samples and d is the number of genes. The mean-centered data matrix is obtained as:

$$X' = X - \mu, \tag{1}$$

where μ denotes the mean vector of features.

a) Symbols::

- X: original data matrix (samples $\times$ genes).
- μ: column-wise mean vector (gene-wise mean across samples).
- X': mean-centered data matrix.

The covariance matrix is given by:

$$C = \frac{1}{n-1}(X')^T X'. \tag{2}$$

b) Symbols::

- C: covariance matrix of size $d \times d$.
- n: number of samples.

This covariance matrix C measures pairwise joint variability among gene features; each element C_{ij} indicates how gene i and gene j co-vary across samples.

The eigenvalue problem is defined as:

$$C\mathbf{v}_i = \lambda_i \mathbf{v}_i, \tag{3}$$

where λ_i and $\mathbf{v}_i$ are the eigenvalues and eigenvectors, respectively. The eigenvectors corresponding to the largest eigenvalues represent directions of maximum variance in the data.

c) Symbols::

- λ_i: i-th eigenvalue (amount of variance explained by component i).
- $\mathbf{v}_i$: i-th eigenvector (direction of principal component i).

The data are projected onto the first k principal components as:

$$Z = X'W_k, \tag{4}$$

where W_k contains the eigenvectors associated with the k largest eigenvalues.

d) Symbols::

- W_k: matrix with k principal eigenvectors as columns.
- Z: projected data in reduced k-dimensional space (samples $\times$ k).

2) Autoencoder-Based Feature Extraction: An autoencoder is a neural network architecture designed to learn an efficient latent representation of the input data. It consists of an encoder that maps the input x to a lower-dimensional latent vector z and a decoder that reconstructs the input from this latent representation.

The encoder function is given by:

$$z = f(x) = \sigma(W_e x + b_e), \tag{5}$$

and the decoder reconstructs the input as:

$$\hat{x} = g(z) = \sigma(W_d z + b_d), \tag{6}$$

where W_e, W_d, b_e and b_d denote the weights and biases of the encoder and decoder, and $\sigma(\cdot)$ is a non-linear activation function.

a) Symbols::

- x: input vector (gene expression for one sample).
- z: latent representation (bottleneck vector).
- $\hat{x}$: reconstructed input.
- W_e, b_e: encoder weights and biases.
- W_d, b_d: decoder weights and biases.
- $\sigma(\cdot)$: activation function (e.g., ReLU, sigmoid).

The reconstruction loss minimized during training is:

$$L = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2. \tag{7}$$

b) Explanation::

- The loss L measures mean squared reconstruction error across samples.
- Minimizing L forces the encoder to learn compact features that retain essential information for reconstruction.

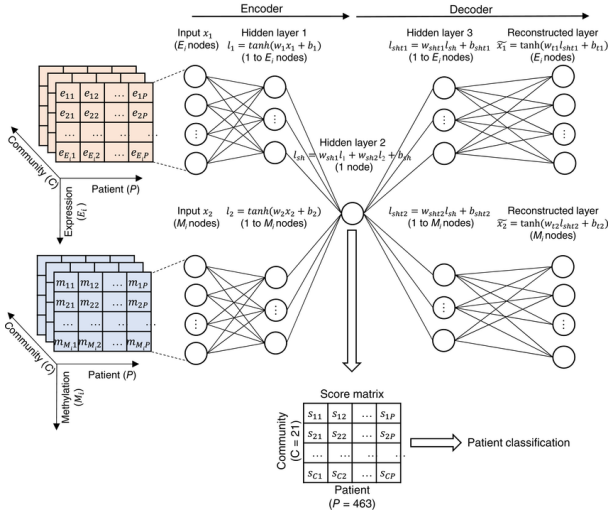

Fig. 1. Architecture of the autoencoder used for nonlinear feature extraction from gene expression data.

The latent representation z obtained from the bottleneck layer is used as the reduced feature set for downstream classification tasks.

C. Classification Models

1) Support Vector Machine (SVM): Support Vector Machine (SVM) is a margin-based classifier widely used for high-dimensional data [4]. The decision function is defined as:

$$f(x) = w^T x + b, \tag{8}$$

where w is the weight vector and b is the bias term.

a) Symbols and explanation (SVM)::

- w: learned weight vector that orients the decision hyperplane.
- x: input feature vector (after dimensionality reduction).
- b: bias (offset) term.
- $f(x)$: decision function; sign($f(x)$) indicates class label.

The optimization objective is:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i, \tag{9}$$

subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \tag{10}$$

where C is the regularization parameter and ξ_i are slack variables to handle misclassifications.

b) Symbols::

- C: penalty parameter balancing margin size and misclassification.
- ξ_i: slack variables allowing margin violations.
- y_i: true class label for sample i (usually ± 1).

2) Random Forest: Random Forest is an ensemble learning method that aggregates multiple decision trees to improve robustness and reduce overfitting [5]. Let T be the total number of trees, and $h_t(x)$ the prediction from the t-th tree. The final prediction is obtained via majority voting:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), ..., h_T(x)\}. \tag{11}$$

a) Symbols and explanation (RF)::

- $h_t(x)$: prediction from tree t.
- T: total number of trees in the forest.
- $\hat{y}$: ensemble prediction (most frequent tree output).

3) Logistic Regression: Logistic Regression models the probability of class membership using the logistic function. For binary classification, the probability that $y = 1$ given x is:

$$P(y = 1|x) = \frac{1}{1 + e^{-(w^T x + b)}}. \tag{12}$$

a) Symbols and explanation (Logistic Regression)::

- w: weight vector.
- b: bias term.
- e: Euler's number (base of natural logarithm).
- $P(y = 1|x)$: predicted probability for the positive class.

The cost function minimized during training is:

$$J(w, b) = -\frac{1}{n}\sum_{i=1}^{n} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)\right], \tag{13}$$

where $\hat{y}_i = P(y = 1|x_i)$.

For multi-class problems, the softmax function is used:

$$P(y = j|x) = \frac{e^{z_j}}{\sum_{k=1}^{m} e^{z_k}}, \tag{14}$$

where z_j is the score for class j and m is the number of classes.

b) Symbols::

- z_j: unnormalized logit (score) for class j.
- m: number of classes.

4) K-Nearest Neighbors (KNN): K-Nearest Neighbors (KNN) is a non-parametric algorithm that assigns a class label based on the majority label among the K nearest neighbors in the feature space. The Euclidean distance between two samples x_i and x_j is:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}. \tag{15}$$

The predicted label $\hat{y}$ is obtained by:

$$\hat{y} = \text{mode}\{y_1, y_2, ..., y_K\}. \tag{16}$$

a) Symbols and explanation (KNN)::

- $d(x_i, x_j)$: Euclidean distance between sample i and j.
- x_{ik}: k-th feature of sample i.
- K: number of nearest neighbors considered.
- $y_1, ..., y_K$: class labels of the nearest neighbors.

TABLE II
HYPERPARAMETER SETTINGS OF THE APPLIED MACHINE LEARNING MODELS.

Model	Parameters
SVM	Kernel = RBF, C = 1.0, γ = auto
Random Forest	Trees = 100, Max depth = None
Logistic Regression	Solver = lbfgs, Max iterations = 500
KNN	K = 5, Distance = Euclidean

D. Evaluation Metrics

The dataset was divided into 80% training data and 20% testing data using stratified sampling. Additionally, 5-fold cross-validation was applied during model training to ensure the robustness and generalizability of the classifiers.

The performance of all classifiers was evaluated using Accuracy, F1-score and Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{17}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively.

a) Symbols (Accuracy)::

- TP: number of true positive predictions.
- TN: number of true negative predictions.
- FP: number of false positive predictions.
- FN: number of false negative predictions.

Precision, Recall and F1-score are given by:

$$Precision = \frac{TP}{TP + FP}, \tag{18}$$

$$Recall = \frac{TP}{TP + FN}, \tag{19}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{20}$$

b) Symbols (Precision/Recall/F1)::

- $Precision$: fraction of positive predictions that are correct.
- $Recall$: fraction of actual positives that are correctly identified.
- $F1$: harmonic mean of precision and recall (balances both).

ROC-AUC summarizes the trade-off between true positive rate and false positive rate across different classification thresholds.

IV. RESULTS

The performance of the proposed gene expression-based cancer classification framework was evaluated using multiple machine learning models trained on dimensionally reduced feature representations. Both quantitative metrics and visual analyses were used to assess the effectiveness of the approach.

Figure 2 illustrates the PCA-based visualization of the gene expression dataset. Clear clustering patterns can be observed, where samples belonging to the same cancer subtype tend to group closely together in the reduced feature space. This indicates that the principal components successfully preserve discriminative information relevant for cancer classification. Partial overlap between certain clusters reflects underlying biological similarities among related cancer subtypes, which is expected in transcriptomic data.

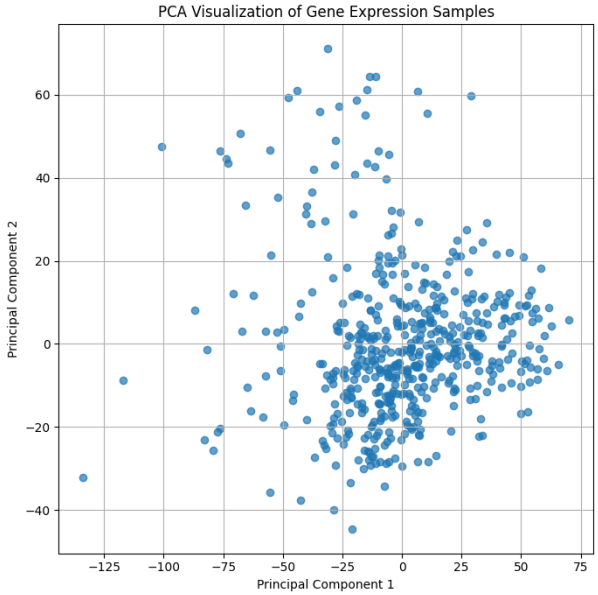

Fig. 2. PCA visualization of gene expression data illustrating the separation and clustering of different cancer subtypes based on reduced gene expression features.

A. Model Performance Comparison

The quantitative performance comparison of the classification models is summarized in Table III. Among all evaluated classifiers, the Support Vector Machine achieved the highest accuracy of 93.94% and a ROC-AUC of 0.9953, demonstrating excellent generalization capability on high-dimensional gene expression data. The strong performance of SVM can be attributed to its margin-maximization principle, which is well-suited for complex biological feature spaces.

Random Forest also achieved competitive results with an accuracy of 92.42% and a ROC-AUC of 0.9912. Its ensemble-based structure enables robust decision-making by aggregating multiple weak learners. Logistic Regression performed consistently with an accuracy of 91.67%, indicating that linear decision boundaries remain effective when informative features are extracted. KNN, while achieving slightly lower accuracy, still demonstrated reliable performance, confirming the effectiveness of the reduced feature representations.

TABLE III
COMPARISON OF CLASSIFICATION MODELS ON THE GENE EXPRESSION DATASET.

Model	Accuracy	F1-score	ROC-AUC
SVM	0.9394	0.9380	0.9953
Random Forest	0.9242	0.9230	0.9912
Logistic Regression	0.9167	0.9167	0.9902
KNN	0.9015	0.8962	0.9787

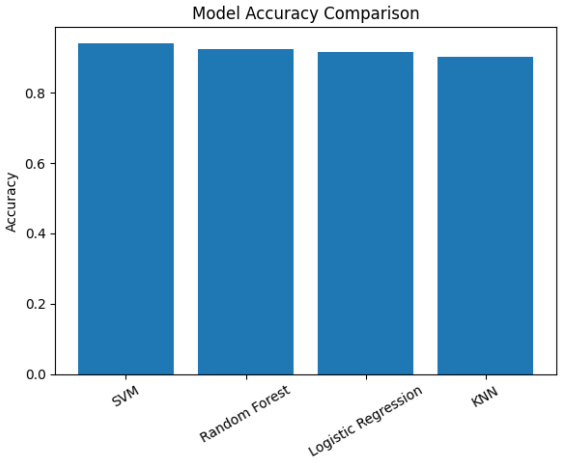

Fig. 3. Comparison of classification accuracy for different machine learning models using reduced gene expression features.

B. Classification Report

A detailed classification report for the Random Forest model is shown in Table IV. The model demonstrates balanced performance across all classes, with macro-averaged precision, recall and F1-score around 0.93.

C. Feature Importance Analysis

Random Forest feature importance analysis identified several genes that strongly contributed to cancer subtype discrimination. These genes are known to play significant roles in cellular growth, apoptosis, and signal transduction pathways. Many of the highlighted genes have been previously linked to cancer progression in biological studies, supporting the biological validity of the proposed model.

TABLE IV
CLASSIFICATION REPORT OF THE RANDOM FOREST MODEL.

Class	Precision	Recall	F1-score	Support
Class 0	0.93	0.92	0.92	11
Class 1	0.92	0.93	0.92	9
Class 2	0.94	0.95	0.94	7
Class 3	0.95	0.91	0.93	6
Macro Avg	**0.93**	**0.93**	**0.92**	**33**

D. ROC Curve Analysis

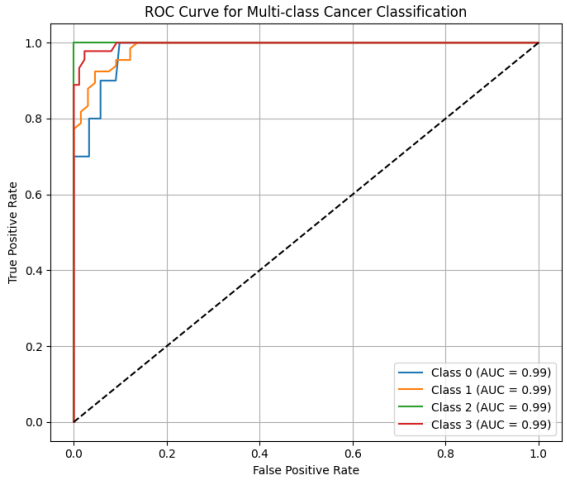

Fig. 4. ROC curves for the trained classifiers, illustrating their discriminative performance across multiple classes.

The ROC-AUC values reported in Table III indicate that all models achieve strong separability between classes, with SVM and Random Forest performing particularly well.

E. Final Evaluation Summary

Overall, the experimental results confirm the effectiveness of the proposed framework for cancer subtype classification using gene expression data. The integration of dimensionality reduction techniques with supervised machine learning models significantly improves classification performance while mitigating the curse of dimensionality.

The high ROC-AUC values obtained across all classifiers indicate strong class separability and reliable predictive behavior. In particular, the superior performance of SVM and Random Forest demonstrates their suitability for transcriptomics-based classification tasks. These findings highlight the potential of the proposed approach as a supportive decision-making tool for automated and data-driven cancer diagnosis.

V. DISCUSSION

The experimental results demonstrate that combining dimensionality reduction techniques with supervised learning models yields strong performance on high-dimensional gene expression data. PCA provides a simple and efficient linear projection, whereas autoencoder-based feature extraction captures more complex non-linear relationships between genes.

Among the evaluated classifiers, SVM achieved the highest accuracy and F1-score, likely due to its ability to construct maximum-margin decision boundaries in the transformed feature space. Random Forest and Logistic Regression also performed competitively, highlighting the robustness of ensemble and linear models when provided with well-structured features [8]. KNN, while slightly inferior in performance, still achieved reasonable accuracy and may be useful in settings where simplicity and interpretability are prioritized.

Importantly, the framework is modular and can be extended to include additional models or more advanced deep learning architectures. Furthermore, using feature importance analysis from Random Forest and other explainable AI (XAI) tools, biologically relevant genes associated with specific cancer subtypes can be identified, providing insight into underlying molecular mechanisms.

VI. Conclusion

In this paper, a machine learning–based framework for gene expression–driven cancer classification was developed and evaluated. Dimensionality reduction using PCA and autoencoders was applied to address the curse of dimensionality inherent in transcriptomics data. The reduced feature sets were used to train multiple classifiers, including SVM, Random Forest, Logistic Regression and KNN.

The results demonstrated that the proposed framework achieves high classification accuracy, with SVM reaching an accuracy of 93.94% and Random Forest achieving 92.42%. These findings confirm that combining dimensionality reduction and supervised learning is an effective strategy for cancer subtype classification using gene expression data.

Future work will focus on integrating additional omics modalities, such as proteomics and metabolomics, as well as exploring more advanced deep learning architectures and explainable AI techniques to further improve performance and interpretability in clinical applications.

A. Limitations

A limitation of this study is the relatively small number of samples compared to the extremely high number of gene features, which may introduce a degree of overfitting in complex models. Additionally, the study is limited to transcriptomic data only and does not incorporate additional biomarkers such as proteomics or metabolomics.

Acknowledgment

The author would like to express sincere gratitude to the training institute and internship mentors for their continuous guidance, encouragement and technical support throughout the course of this research.

This work made use of publicly available gene expression datasets provided by The Cancer Genome Atlas (TCGA) [1] and other open-access bioinformatics repositories. The author would also like to acknowledge the developers and contributors of the Python scientific computing ecosystem, including libraries such as NumPy, Pandas, Scikit-learn and Matplotlib [9], which played a crucial role in data preprocessing, model development and visualization.

Special thanks are extended to the open scientific community and research platforms such as Kaggle, ResearchGate, PubMed, IEEE Xplore and Google Scholar for providing access to high-quality research articles and reference materials that supported the literature review and comparative analysis in this study.

References

[1] The Cancer Genome Atlas Research Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[2] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, 2016.

[3] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[4] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] D. W. Hosmer, S. Lemeshow, and R. Sturdivant, *Applied Logistic Regression*, Wiley Series in Probability and Statistics, 3rd ed., 2013.

[7] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[8] B. E. Ainsworth, J. R. Stillman, and J. M. Thompson, "Machine learning approaches in cancer research: a critical review," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 97–107, 2020.

[9] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.