Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3743~3751
e-ISSN: 2548-6861

Classification of Tumor and Normal Tissue Gene Expression in Lung Adenocarcinoma Using Support Vector Machine and Gaussian Process Classification

Rahmadi Yotenka[1]*,, Adhitya Ronnie Effendie[2]*, Rohmatul Fajriyah[3]****
* Department of Mathematics, Gadjah Mada University
** Department of Statistics, Universitas Islam Indonesia
rahmadiyotenka@mail.ugm.ac.id [1], adhityaronnie@ugm.ac.id [2], rfajriyah@uii.ac.id [3]

Article Info

Article history:

Received 2025-11-13
Revised 2025-12-01
Accepted 2025-12-10

Keyword:

Biomarker,
Gene expression,
GPC,
Lung adenocarcinoma,
SVM.

ABSTRACT

Lung adenocarcinoma (LUAD) is a major cause of cancer-related mortality worldwide. This study aims to identify potential LUAD biomarkers and develop robust classification models using the GSE151101 microarray dataset. Preprocessing included RMA normalization, ComBat batch-effect correction, and feature filtering based on annotation completeness, variability, and statistical significance. Support Vector Machine (SVM) and Gaussian Process Classification (GPC) models were constructed, with the polynomial GPC model achieving the best performance (accuracy 97.92%; F1-score 97.96%). Repeated 10-fold cross-validation confirmed its stability (mean accuracy 96.88%, SD ±1.97%, CV 2.03%), outperforming linear SVM, GPC-RBF, and Multiple Kernel Learning (MKL). Functional enrichment analysis showed that key discriminative genes; CDH13, CDKN2A, BCL2L11, MYL9, COL1A1, and AKT3; were enriched in pathways related to epithelial–mesenchymal transition, extracellular matrix remodelling, focal adhesion, PI3K/AKT signalling, and cell-cycle regulation, all of which are central to LUAD progression. In general, polynomial-kernel GPC is a stable and useful way to classify transcriptomes and rank biomarkers. Nevertheless, the translational potential of these signatures requires further validation in independent and clinically controlled cohorts.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

Lung cancer is a predominant cause of cancer-related mortality globally, with lung adenocarcinoma (LUAD) being the most prevalent subtype of non–small-cell lung cancer (NSCLC). The development of microarray-based gene expression technology allows for the concurrent analysis of thousands of genes, facilitating the identification of molecular biomarkers that can differentiate tumor tissues from normal tissues. This capability is essential for the development of early diagnostic tools and the advancement of personalized, precision-based cancer therapies [1][2][3].

A study conducted by [4] examined the GSE151101 gene expression dataset, consisting of 237 lung tissue samples from 126 LUAD patients. The findings indicated a correlation between aberrant expression of Y-chromosome genes and autosomal hypomethylation with poorer prognoses in male patients. The study primarily concentrated on biological mechanisms and did not consider the creation of predictive classification models for the accurate automation of tumor detection using gene expression profiles.

On the other hand, machine-learning techniques, particularly the Support Vector Machine (SVM), have been widely employed for the classification of cancer based on gene expression [5]. In [6] showed that multiclass SVM surpassed traditional methods, including Linear Discriminant Analysis (LDA) and k-Nearest Neighbor (KNN), in the classification of diverse cancer types based on gene expression data. This finding highlights the effectiveness of SVM in managing high-dimensional, small-sample datasets, which are typical of microarray data [7].

Classification performance can be improved through methods that separate classes while also offering probabilistic estimations and implicit feature selection, as demonstrated by Gaussian Process Classification (GPC) [8]. In [9] utilized Gaussian Process Classification (GPC) to identify gene-expression biomarkers by incorporating the Automatic Relevance Determination (ARD) parameter into the model's covariance function. This method allows the model to classify data and concurrently identify the key genes that influence class differentiation [8].

This study seeks to develop and compare Support Vector Machine (SVM) and Gaussian Process Classification (GPC) models utilizing gene expression data from GSE151101. This study aims to identify genes with potential as LUAD biomarkers, in addition to evaluating the classification performance between tumors and normal tissues, while considering the biological insights reported by [4]. The findings are anticipated to enhance both methodological approaches in applied statistics and practical applications in bioinformatics and cancer research.

II. METHOD

This research comprises seven primary stages: data collection, data preprocessing, feature selection, feature engineering, classification model development, performance assessment, and functional analysis and interpretation of results. The steps are depicted in Figure 1.

The data collecting phase was conducted to acquire gene expression datasets from the public repository Gene Expression Omnibus (GEO) under the code GSE151101, comprising 237 lung tissue samples (both tumor and normal) assessed using the Affymetrix Human Gene 1.1 ST Array platform. The data preparation phase seeks to ready the data for analytical purposes, encompassing background correction, quantile normalization, and summarization procedures.

Subsequently, feature filtering was conducted to eliminate duplicates, discard probes without gene annotations, and identify significant genes utilizing a t-test with Bonferroni correction. The feature engineering phase encompassed the identification of genes that satisfied statistical requirements and the partitioning of the dataset into training and testing subsets (80:20 ratio).

The model classification phase was executed utilizing the Support Vector Machine (SVM) and Gaussian Process Classification (GPC) methods with diverse kernel functions, such as Radial Basis Function (RBF), Polynomial, and Multiple Kernel Learning (MKL). The model's predictive performance was evaluated using accuracy, precision, recall, and F1-score derived from the confusion matrix, complemented by repeated 10-fold cross-validation to assess model stability and variability.

The concluding phase involves functional analysis and interpretation of data, aimed at elucidating the biological roles of the discriminatory genes identified by the optimal model using Gene Ontology (GO), KEGG, and Reactome analyses. This stage's data offer biological insights into the molecular underpinnings of lung cancer and establish a foundation for conclusions and future research objectives.

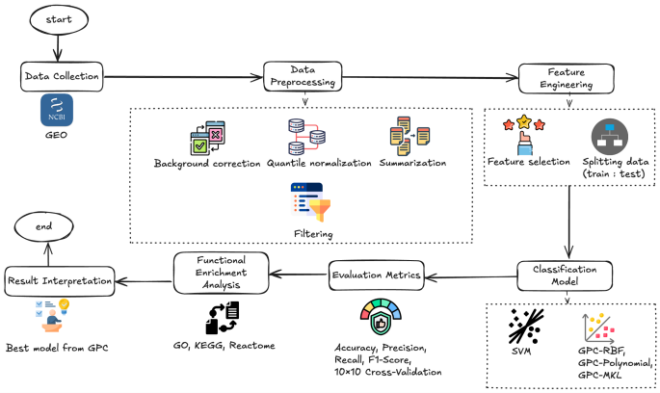


Figure 1. Gene Expression Classification Research Flowchart

A. Data and Research Variable

This research employed secondary data sourced from the Gene Expression Omnibus (GEO), a public resource administered by the National Center for Biotechnology Information (NCBI) in the United States. National Institutes of Health (NIH). The dataset GSE151101 comprises gene expression data derived from human lung tissue, produced using the Affymetrix Human Gene 1.1 ST Array platform. The dataset comprises 237 samples obtained from 126 individuals diagnosed with lung adenocarcinoma (LUAD), including 124 tumor samples and 113 normal samples, reflecting a somewhat equal distribution between the two categories. This specific class fraction is crucial for assessing potential overfitting concerns, as microarray investigations generally encompass high-dimensional feature spaces with relatively small sample numbers [4].

The Affymetrix platform utilizes probe-based microarray technology, wherein small DNA fragments are engineered to selectively hybridize with their corresponding mRNA sequences. Probes that target the same gene are organized into a probeset, and the fluorescence intensity from hybridization indicates the signal, which corresponds to the gene's expression level. Phenotypic data, encompassing tissue type, sex, and patient ID, was obtained from the GEO metadata.The dependent variable in this study was tissue categorization (tumor/normal), whereas the independent variables were expression levels from 1178100 normalized probesets. Data preparation was performed utilizing R and Bioconductor tools, including normalization, logarithmic transformation, and feature selection based on variance and statistical significance. Only probesets with valid genetic IDs, namely those linked to an Entrez Gene ID in NCBI, were preserved for subsequent categorization analysis.The dataset was randomly divided into two parts to build and test classification models. The training set had 80% of the samples (n = 189), and the testing set had the other 20% (n = 48).

B. Preprocessing

The objective of the preprocessing stage was to get the data on gene expression ready for the classification procedure so that it could be used in a reliable manner. The signal was normalized, the logarithmic transformation was performed, feature selection was performed, class labeling was performed, and statistical filtering of significant features was performed at this step. Detailed descriptions of the preprocessing processes that were carried out for this investigation are provided below [10][11].

- Background Correction. Using equation 1, this step removes the non-biological signal components coming from the microarray background.

$$PM_{ijn} = bg_{ijn} + s_{ijn} \quad (1)$$

where s_{ijn} is the actual signal with $s_{ijn} \sim Exp(\lambda)$, bg_{ijn} is the background signal with $bg_{ijn} \sim \mathcal{N}(\mu, \sigma^2)$, and PM_{ijn} is the signal seen at the probe.

- By projecting each column q_k to the reference distribution vector using rank-based averaging, as stated in equation 2, quantile normalization is carried out to align expression distributions across samples.

$$proj_d q_k = \left(\frac{1}{n}\sum_{j=1}^{n} q_{kj}, \ldots, \frac{1}{n}\sum_{j=1}^{n} q_{kj}\right) \quad (2)$$

- Use equation 3 to perform summarization by integrating several signals from probes in a single probeset into a single expression value per gene.

$$log_2\left(PM_{ijk}\right) = \alpha_{jk} + \beta_{ik} + \varepsilon_{ijk} \quad (3)$$

where ε_{ijk} is the residual error component, β_{ik} is the expression level of gene i, and α_{jk} denotes the probe affinity effect.

- Probesets were chosen based on genetic variability and identity. Duplicate ID probesets and probesets without Entrez IDs were removed. The interquartile range (IQR) value in equation 4 was used to assess expression variability; only features with an IQRi value > 0.5 were kept.

$$IQR_i = Q_3(x_i) - Q_1(x_i) \quad (4)$$

- While running a two-sample t-test to find any noteworthy variations between "tumor" and "normal" tissues. Equation 5 was utilized to account for non-homogeneous variation between groups using Welch's t-test methodology.

$$t_i = \bar{x}_{i1} - \bar{x}_{i0} \Big/ \sqrt{s_{i1}^2/n_1 + s_{i0}^2/n_0} \quad (5)$$

In this context, s_{i1}^2, s_{i0}^2 represent the variance values for the "tumor" and "normal" groups, respectively, while n_1, n_0 denote the sample sizes of these groups. Additionally, $\bar{x}_{i1}, \bar{x}_{i0}$ indicate the average expression values of the i-th gene in each group.

- Testing on hundreds of genes at once raises the possibility of false positives, or Type I mistakes. The above issue was addressed by applying the Bonferroni procedure with equation 6 to make a p-value adjustment.

$$p_i^{adj} = min(p_i \cdot m, 1) \quad (6)$$

where m is the total number of tests conducted, p_i is the initial p-value for the i-th feature/gene, and p_i^{adj} is the p-value following adjustment. Genes and features that have $p_i^{adj} < \alpha = 0.0005$ are deemed statistically significant and are kept for additional examination. These genes exhibit notable variations in expression between "tumor" and "normal" tissues, making them promising candidates for use as biomarkers.

C. Support Vector Machine

One well-liked classification technique for high-dimensional data, such as gene expression data, is Support Vector Machine (SVM). SVM finds the best hyperplane in feature space with the largest margin between the two classes [12]. SVM seeks to distinguish between the two classes as much as feasible using a linear combination of gene expression in binary classification scenarios such as the one in this study (tumor vs. normal). Equation 7 represents the linear SVM objective function in primal form for binary classification situations using training data $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$ [13].

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad (7)$$

with the constraints:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0, \qquad i = 1, \ldots, n$$

where w is the weight vector, b is the bias, $\boldsymbol{\xi_i}$ are slack variables, and C is the regularization parameter that regulates the trade-off between the maximum margin and classification errors.

SVM maps data to a higher-dimensional space using a kernel function $K(x_i, x_j)$ for data that cannot be separated linearly. Because it is appropriate for the high-dimensional gene expression data structure and the small sample size, a linear kernel was employed in this investigation. Equation 8 describes the linear kernel, which is typically computationally efficient [12][14].

$$K(x, x') = x^{\mathrm{T}}x' \quad (8)$$

The SVM model's performance is greatly impacted by the choice of the C parameter. To find the ideal C value, tweaking was done in this study using k-fold cross-validation. Training data was split into k subsets for each candidate $C_j \in C$. k-1 subsets were used to train the model, while the remaining subset was used for validation. Equation 9 is used to compute the average validation error [12].

$$CV_{Err}(C_j) = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{|D_i|}\sum_{x \in D_i}\mathbb{I}\left(f_j^{(i)}(x) \neq y\right) \quad (9)$$

Nilai optimal dipilih sebagai yang meminimalkan sebagaimana persamaan 10. According to equation 10, the value of C (C^*) that minimizes $CV_{Err}(C_j)$ is the ideal value [12].

$$C^* = \arg\min_{C_j \in C} CV_{Err}(C_j) \quad (10)$$

where $f_j^{(i)}$ is the SVM model trained on fold i with parameters, $\mathbb{I}(\cdot)$ is the indicator function that is 1 if the

argument is true and 0 if it is false, y is the actual class label, and D_i is the validation data for fold i.

D. Gaussian Process Classification

For high-dimensional data, like gene expression data, Gaussian Process Classification (GPC), a non-parametric, probabilistic classification technique, works incredibly well. A Gaussian Process (GP), which is a collection of random functions with a multivariate normal distribution at each input point, is used by GPC to represent the distribution of the prediction function. This method generates uncertainty estimates in addition to class predictions [8][15].

GPC presupposes that a latent function $f(x) \sim \mathcal{GP}\big(0, k(x,x')\big)$ exists in binary classification. This latent function is transferred to class probabilities via a link function, such as the probit function [8]. Equation 11 defines the likelihood of a label $y_i = 1$ in the probit technique given the value of the latent function f_i.

$$P(y_i = 1|f_i) = \Phi(f_i) \tag{11}$$

where the usual normal distribution's cumulative distribution function ($\mathcal{N}(0,1)$) is represented by $\Phi(f_i)$.

The kernel function $k(x,x')$, which establishes the degree of similarity between data points, is the primary element of GPC. The Radial Basis Function (RBF) kernel is used in this study due to its smoothness and suitability for intricate patterns in biological data [16]. Equation 12 is followed by the RBF kernel.

$$K(x,x') = exp\left(-{\|x-x'\|^2}/{2\ell^2}\right) \tag{12}$$

where ℓ is the tuning-controlled length-scale parameter.

In addition to RBF, a polynomial kernel is employed, which adds non-linear correlations between features to the linear kernel [13]. Equation 13 is the generic form.

$$K(x,x') = \left(x_i^{\mathrm{T}} x' + c\right)^d \tag{13}$$

where c is the offset constant and d is the polynomial's degree.

Multiple Kernel Linear, which incorporates several kernels, were also employed in this investigation. Equation 14 is the generic form [17].

$$K(x,x') = \sum_{m=1}^{M} d_m K_m\,(x,x'), \quad d_m \geq 0, \sum_{m=1}^{M} d_m = 1 \tag{14}$$

Predicting the class probability for a fresh sample x_* is the aim of GPC. Equation 15 is used to integrate the posterior of the latent function f over the training data D in order to do inference [18].

$$P(y_* = 1|x_*, D) = \int \Phi(f_*)\, P(f_*|D, x_*)\, df_* \tag{15}$$

E. Confusion Matrix

All classification goodness metrics, including accuracy, precision, and F1-score, are computed using the confusion matrix as the foundation. The number of accurate and inaccurate forecasts for each class is shown in this matrix. Table 1 is the confusion matrix for a binary instance in its generic form [5][10].

TABEL I
CONFUSION MATRIX

Actual	**Predicted**	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

The confusion matrix in Table 1 may be used to compute the following classification goodness measures:

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$, $Precision = \frac{TP}{TP+FP} \times 100\%$, $Recall = \frac{TP}{TP+FN} \times 100\%$, and $F1 - Score = \frac{Precision \times Recall}{Precision+Recall} \times 100\%$.

F. Differential Gene Expression

Differential gene expression analysis is a crucial method in transcriptomics designed to detect genes with significantly varying expression levels between experimental groups, such as treatment vs. control conditions. The fundamental purpose of DGE analysis is to identify genes that are differently expressed in the comparative contexts [19]. The theoretical basis of DGE is that environmental stressors or molecular disruptions, such as drug exposure or genetic mutation, modify transcriptional regulation, leading to quantifiable alterations in RNA abundance across genes [20]. By quantifying these alterations, DGE enables researchers to deduce functional pathways (with Gene Oncology, the Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome), discover biomarkers, and elucidate molecular processes behind particular biological responses [21][22].

III. RESULT AND DISCUSSION

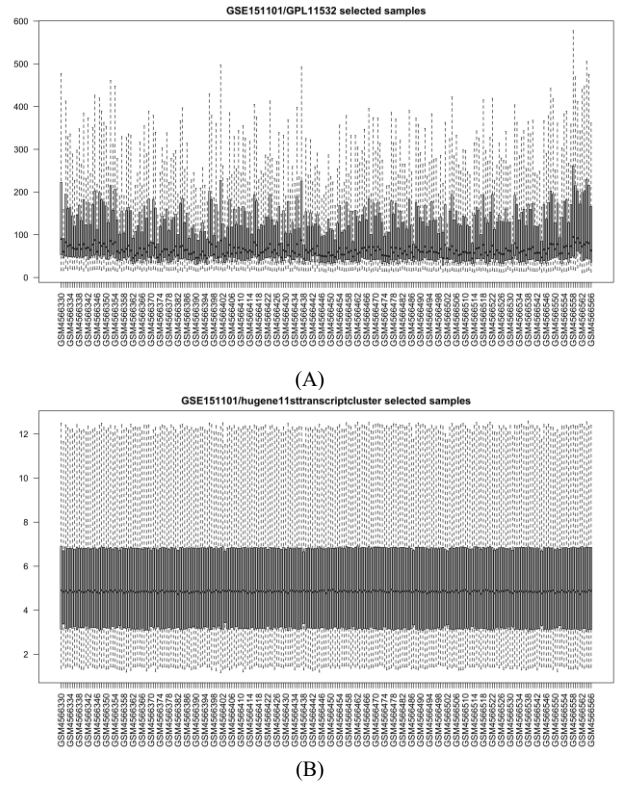
A. Preprocessing and Filtering

The Bioconductor ecosystem was used to process the HuGene 1.1 ST microarray data during the preprocessing phase. The pipeline was carried out using the oligo package, which is specifically designed for Gene ST array platforms. The primary objective of preprocessing is to eliminate undesired variation across samples, reduce noise, and minimize non-biological artifacts introduced by technical or

experimental conditions. Background correction, quantile normalization, and summarization were performed using the Robust Multi-array Average (RMA) procedure at the transcript-cluster level. Figure 2(A) illustrates that the inter-sample gene expression distribution appeared non-uniform prior to preprocessing. As shown in Figure 2(B), RMA normalization effectively reduced this deviation, producing more consistent intensity patterns across samples.

The GSE151101 dataset contains 33297 transcript features (GPL11532 platform) measured across 237 samples, comprising 124 tumor and 113 normal samples. Because GEO microarray datasets are inherently prone to batch effects, which cannot be fully addressed by RMA alone, an additional correction step was incorporated. Prior to batch adjustment, exploratory principal component analysis (PCA) revealed mild clustering driven by technical variation rather than biological class labels. To mitigate this, batch effect correction was performed using the ComBat algorithm from the sva package. After ComBat adjustment, PCA plots showed a clear improvement, with samples clustering mostly by biological condition (tumor vs. normal). This showed that unwanted technical variation had been successfully removed. This step ensured that downstream differential analysis and classification were not confounded by batch-associated noise.

Filtering of expression features was subsequently conducted using the nsFilter function from the genefilter package. The filtering criteria included (i) removal of features and control probes lacking complete annotation, (ii) elimination of duplicated probe set identifiers, and (iii) application of an interquartile range (IQR) threshold to retain only genes exhibiting sufficient variability. From the initial 33297 probesets, 9425 features were retained after filtering based on annotation completeness, uniqueness, and variability (IQR > 0.5).To ensure transparency and reproducibility in the feature-selection process, a statistically rigorous differential expression analysis was performed. Because many patients contributed paired tumor–normal samples, the analysis accounted for within-patient dependence by including patient ID as a blocking factor. Differential testing was carried out using a two-sample Welch's t-test computed across all genes via the multtest package. To control for multiple testing, which is critical in high-dimensional microarray data, the Bonferroni correction was applied. Genes were considered significantly differentially expressed if their adjusted p-value was below 0.0005. This procedure yielded 5378 candidate genes for model development. These selected genes formed the feature set used to construct the primary classification models, Support Vector Machine (SVM) and Gaussian Process Classification (GPC), ensuring that the downstream machine learning analysis was based on statistically robust and biologically relevant predictors.



(A)

(B)

Figure 2. GSE151101 Expression Profile (A) Before Preprocessing and (B) After Preprocessing

B. Classification Model

A classification model was created utilizing GSE151101 gene expression data and many machine learning methods. Grid search across cost parameters (0.001–100) using 10-fold cross-validation improved the first model, a linear kernel Support Vector Machine (SVM). The ideal cost value (C = 0.001) has 3.21% cross-validation error. This model showed significant linear separability of gene expression patterns, reaching 95.83% accuracy on the independent test set and 99.47% accuracy on the training set, with perfect precision for the normal class and perfect recall for the tumor class. Many kernel functions were then implemented in Gaussian Process Classification (GPC). The degree three polynomial kernel performed best, with 97.92% accuracy and a 97.96% F1-score. The Radial Basis Function (RBF) kernel had an accuracy equivalent to SVM, while in the Multiple Kernel Learning (MKL) configuration, RBF and polynomial kernels had equal test accuracy (97.92%) with balanced precision and recall. The Random Forest classifier had 95.83% accuracy and 100% precision as a non-kernel benchmark. The main model performance comparison is in Table 2.

TABLE 2
COMPARATIVE PERFORMANCE OF CLASSIFICATION MODELS

Model	Accuracy	Precision	F1-Score
SVM	95.83%	92.00%	97.96%
GPC-RBF	95.83%	95.80%	95.83%
GPC-Polynomial	97.92%	100.00%	97.96%
GPC-MKL	97.92%	100.00%	95.83%
Random Forest	95.83%	100.00%	95.83%

The high test-set accuracy observed across multiple models does not adequately address the limitations of single-split performance in high-dimensional microarray data, which exhibit significant sensitivity to variations in data partitioning. In response to this concern, a full stability test was done using repeated 10-fold cross-validation. This assessment measured robustness through mean accuracy, standard deviation (SD), 95% confidence intervals (CI), and coefficient of variation (CV). Table 3 indicates that the GPC-Polynomial and GPC-MKL models exhibited the greatest stability, evidenced by low variance (SD ≤ 2.0%) and the lowest coefficient of variation (≤ 2.1%). In contrast, the GPC-RBF model demonstrated the highest variability (SD = 4.61%; CV = 5.23%), suggesting reduced reliability despite comparable test-set accuracy.The results indicate that the GPC-Polynomial model demonstrates both accuracy and statistical stability across repeated resampling, thereby reinforcing its position as the optimal classifier for LUAD gene expression data.

TABLE 3
STABILITY EVALUATION OF CLASSIFICATION MODELS

Model	Mean Accuracy	SD	95% CI	CV
SVM	96.12%	±2.30%	[95.48, 96.76]	2.39%
GPC-Poly	96.88%	±1.97%	[96.33, 97.42]	2.03%
GPC-RBF	88.15%	±4.61%	[86.87, 89.42]	5.23%
GPC-MKL	96.54%	±1.01%	[96.03, 97.05]	1.05%
Random Forest	96.54%	±1.69%	[95.69, 97.39]	1.75%

Only two of the 25 normal samples were mistakenly labeled as tumors, according to the confusion matrix derived from the prediction results on the test data for the GPC model with a polynomial kernel. On the other hand, every one of the 23 tumor samples was effectively identified as a tumor. Figure 3 illustrates this confusion matrix and highlights how accurate the model is in identifying tumor gene expression patterns.

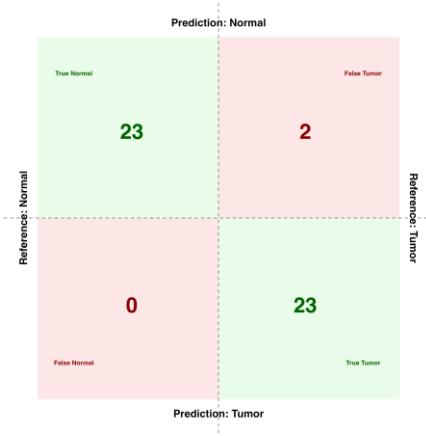


Figure 3. Confusion Matrix for GPC-Polynomial Classification on Test Set

C. *Functional Enrichment Analysis*

Functional enrichment analysis examined the biological significance of the most distinguishing genes identified by the classification model. GO and KEGG studies were performed utilizing the clusterProfiler and org.Hs.eg.db packages inside R/Bioconductor. The investigation concentrated on the highest-ranked genes generated by the Gaussian Process Classification (GPC) model utilizing a third-degree polynomial kernel, which exhibited the most superior performance. Enrichment findings were deemed significant at an adjusted p-value of less than 0.05, employing the Benjamini–Hochberg correction. GO enrichment analysis indicated that the distinguishing genes were mostly associated with extracellular matrix remodeling, cytoskeletal structure, wound healing, and cell adhesion mechanisms (Table 4). These basic features are pivotal to the evolution of lung adenocarcinoma (LUAD), notably via epithelial–mesenchymal transition (EMT), augmented migratory ability, and aberrant cell–matrix interactions.

This investigation revealed many genes with well-established mechanistic functions in lung adenocarcinoma (LUAD) etiology. LUAD promoter hypermethylation silences CDH13 (H-cadherin), a tumor suppressor, leading to poor cell–cell adhesion and accelerated epithelial–mesenchymal transition (EMT), which aids tumor invasion and metastasis [23] [24]. COL1A1 (Collagen Type I Alpha 1 Chain), a major structural component of the extracellular matrix (ECM), contributes to stromal remodeling by increasing ECM stiffness, activating integrin; FAK signaling, and promoting LUAD cell migration. Recent studies have identified COL1A1 as a prognostic biomarker signature in LUAD [25]. In addition, AKT3, a major node in the PI3K/AKT oncogenic axis, improves tumor-cell survival, metabolic plasticity, and apoptosis resistance. Prognostic investigations of AKT isoforms have shown its dysregulation in LUAD [26]. LUAD is linked to uncontrolled proliferation and poor outcomes due to the loss or epigenetic silencing of the tumor suppressor CDKN2A (p16^INK4A), which governs G1-S cell-cycle progression. BCL2L11 (BIM), a pro-apoptotic mediator of the intrinsic mitochondrial system, is

essential to programmed cell death. Reduced BIM expression promotes tumor persistence and therapeutic resistance, especially in EGFR-mutated LUAD. MYL9 (Myosin Light Chain 9), a regulator of cytoskeletal contractility and cell motility, supports actin–myosin dynamics that drive metastatic dissemination. Recent studies indicate that MYL9 is down-regulated in non-small-cell lung cancer (NSCLC) and may suppress EMT in lung cancer cells [27]. These genes constitute a cohesive network of biological processes; EMT, ECM remodeling, PI3K/AKT signaling, cell-cycle dysregulation, apoptosis evasion, and cytoskeletal reorganization, that closely match LUAD development molecular pathways [28].

These gene-level findings reveal that the categorization approach prioritized statistically important, physiologically relevant, and mechanistically integrated LUAD pathogenesis genes.KEGG pathway analysis verified these genes' participation in carcinogenic processes such as ECM; receptor interaction, focal adhesion, PI3K/AKT signaling, Rap1 signaling, and the cell cycle (Table 5). The enrichment of neutrophil extracellular trap (NET) creation pathways (hsa04613) shows tumor-intrinsic transcriptional programs and the inflammatory tumor microenvironment may interact, a mechanism increasingly linked to LUAD aggressiveness. The enrichment patterns show that the GPC model's discriminative genes exhibit physiologically coherent pathways that match LUAD molecular pathophysiology. CDH13, COL1A1, AKT3, CDKN2A, BCL2L11, and MYL9 are promising biomarker validation candidates, consistent with prior LUAD biomarker panels and molecular investigations.

. TABLE 4
GO RESULT: BIOLOGICAL PROCESS OF GPC CLASSIFICATION GENE POLYNOMIAL KERNEL

GO ID	GeneRatio	Count	p.adjust	LeadingGenes
GO:0031589	175/4895	175	0.00	BCL2L11, CDH13, SORBS3, CDK5, CDKN2A
GO:0030198	156/4895	156	0.00	ADAM8, ANGPTL7, RAMP2, FBLN5, PRDX4
GO:0043062	156/4895	156	0.00	ADAM8, ANGPTL7, RAMP2, FBLN5, PRDX4
GO:0042060	184/4895	184	0.00	SH2B3, CDKN1A, MYL9, VAV3
GO:1901987	192/4895	192	0.00	CDK4, PSME3, CDK5, CTDSPL, CDK7
GO:0007160	114/4895	114	0.00	BCL2L11, CDH13, CDK5, CDKN2A
GO:0043410	185/4895	185	0.00	CD24, ADAM8, SORBS3, SPRY2, RAMP3
GO:0007264	203/4895	203	0.00	SH2D3C, SH2D3A, CDH13, RASA4, WASF2

TABLE 5
KEGG PATHWAY ENRICHMENT ANALYSIS RESULTS

ID	GeneRatio	Count	p.adjust	LeadingGenes
hsa04820	0.042	105	1.00e-07	MYL9, NEBL, LDB3, COL1A1, COL1A2
hsa04613	0.035	87	4.50e-06	AKT3, PPIF, ATG7, CLCN4, H2BC26
hsa04110	0.029	73	5.40e-06	CDK4, CDK7, CDKN1A, CDKN2A, NDC80
hsa04517	0.048	120	1.38e-05	AKT3, CLEC4M, MYL9, VAV3, ARPC1A
hsa04510	0.032	81	5.29e-04	AKT3, LAMC3, MYL9, VAV3, COL1A1
hsa04512	0.016	40	2.06e-03	LAMC3, COL1A1, COL1A2, COL4A3, COL4A4

IV. CONCLUSION

This study developed and compared Support Vector Machine (SVM) and Gaussian Process Classification (GPC) models to differentiate between tumor and normal lung tissues utilizing the GSE151101 dataset. The GPC model utilizing a polynomial kernel exhibited superior performance, attaining a test accuracy of 97.92% and an F1-score of 97.96%. Repeated 10×10 cross-validation demonstrated that this model exhibited both accuracy on a single test split and statistical stability, achieving one of the lowest variability metrics among the assessed models (mean CV accuracy =

96.88%, SD = ±1.97%, CV = 2.03%). The findings demonstrate the effectiveness of GPC-Polynomial in managing high-dimensional microarray data when compared to linear SVM and GPC-RBF.The biological analysis indicated that the discriminative genes selected by the optimal model are closely associated with established mechanisms of LUAD pathogenesis. Gene Ontology (GO) analysis revealed enrichment in epithelial–mesenchymal transition (EMT), wound healing, extracellular matrix (ECM) remodeling, cell–substrate adhesion, and cell-cycle regulation; essential processes that contribute to tumor migration, invasion, and uncontrolled proliferation. Several genes, including CDH13, CDKN2A, BCL2L11, MYL9, and SORBS3, play important roles in these pathways. KEGG pathway enrichment analysis corroborated this finding by identifying participation in focal adhesion, ECM–receptor interaction, PI3K/AKT signaling, cell-cycle progression, and neutrophil extracellular trap formation (NETs). Pathways are influenced by genes including AKT3, CDK4, CDK7, COL1A1, COL1A2, and LAMC3, which recent studies on LUAD have identified as factors contributing to tumor aggressiveness and microenvironmental remodeling.

The results indicate that Gaussian Process Classification utilizing a polynomial kernel is an effective and stable method for modeling high-dimensional transcriptomic data. Nonetheless, the implementation of this method requires careful interpretation from a translational perspective. The identified genes, namely CDH13, CDKN2A, BCL2L11, MYL9, COL1A1, and AKT3, exhibit biological significance and potential as biomarkers for LUAD. However, their application in clinical diagnostics necessitates comprehensive validation across independent cohorts, assessment of inter-population variability, evaluation of cross-platform robustness (microarray versus RNA-seq), and standardization in laboratory practices. This study should not be seen as a useful diagnostic tool, but rather as a computational framework for finding biomarkers and coming up with new ideas.This research advances the methodology of machine learning in transcriptomic classification and offers biologically relevant insights into the molecular mechanisms of LUAD. Future research should include multi-center datasets, prospective validation, and integrative modeling to facilitate the translation of these findings into clinical practice.

REFERENCES

[1] T. I. A. Mohamed, A. E. Ezugwu, J. V. Fonou-Dombeu, M. Mohammed, J. Greeff, and M. K. Elbashir, "A novel feature selection algorithm for identifying hub genes in lung cancer," *Sci. Rep.*, vol. 13, no. 1, p. 21671, Dec. 2023, doi: 10.1038/s41598-023-48953-1.

[2] C. Zhang *et al.*, "Identification of lncRNA, miRNA and mRNA expression profiles and ceRNA Networks in small cell lung cancer," *BMC Genomics*, vol. 24, no. 1, p. 217, Apr. 2023, doi: 10.1186/s12864-023-09306-4.

[3] D. Wu, Y. Liu, J. Liu, L. Ma, and X. Tong, "Myeloid cell differentiation-related gene signature for predicting clinical outcome, immune microenvironment, and treatment response in lung adenocarcinoma," *Sci. Rep.*, vol. 14, no. 1, p. 17460, July 2024, doi: 10.1038/s41598-024-68111-5.

[4] S. A. G. Willis-Owen *et al.*, "Y disruption, autosomal hypomethylation and poor male lung cancer survival," *Sci. Rep.*, vol. 11, no. 1, p. 12453, June 2021, doi: 10.1038/s41598-021-91907-8.

[5] M. F. Azhari and R. Fajriyah, "Idektifikasi Gen Marker Pbmcs Ischemic Stroke Menggunakan Analisis Bioinformatika Dan Support Vector Machine," no. 1, 2024.

[6] Universitas Gadjah Mada, V. Sutanto, Z. Sukma, Business Intelligence Data Engineering Division, A. Afiahayati, and Universitas Gadjah Mada, "Predicting Secondary Structure of Protein Using Hybrid of Convolutional Neural Network and Support Vector Machine," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 1, pp. 232–243, Feb. 2021, doi: 10.22266/ijies2021.0228.23.

[7] N. Jiang *et al.*, "Identification of endoplasmic reticulum stress genes in human stroke based on bioinformatics and machine learning," *Neurobiol. Dis.*, vol. 199, p. 106583, Sept. 2024, doi: 10.1016/j.nbd.2024.106583.

[8] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, 3. print. in Adaptive computation and machine learning. Cambridge, Mass.: MIT Press, 2008.

[9] C. Hardcastle, R. O'Mullan, R. Arróyave, and B. Vela, "Physics-informed Gaussian process classification for constraint-aware alloy design," *Digit. Discov.*, vol. 4, no. 7, pp. 1884–1900, 2025, doi: 10.1039/D5DD00084J.

[10] R. Fajriyah, H. A. Isnandar, and A. Arifuddin, "Gene Markers Identification Of Acute Myocardial Infarction Disease Based On Genomic Profiling Through Extreme Gradient Boosting (XGBoost)," *MEDIA Stat.*, vol. 17, no. 1, pp. 69–80, Oct. 2024, doi: 10.14710/medstat.17.1.69-80.

[11] J. P. Debnath *et al.*, "Identification of potential biomarkers for 2022 Mpox virus infection: a transcriptomic network analysis and machine learning approach," *Sci. Rep.*, vol. 15, no. 1, p. 2922, Jan. 2025, doi: 10.1038/s41598-024-80519-7.

[12] K. P. Murphy, *Machine learning: a probabilistic perspective*, 4. print. (fixed many typos). in Adaptive computation and machine learning series. Cambridge, Mass.: MIT Press, 2013.

[13] K.-L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, "Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions," *Mathematics*, vol. 12, no. 24, p. 3935, Dec. 2024, doi: 10.3390/math12243935.

[14] J. Wu and C. Hicks, "Breast Cancer Type Classification Using Machine Learning," *J. Pers. Med.*, vol. 11, no. 2, p. 61, Jan. 2021, doi: 10.3390/jpm11020061.

[15] A. Banerjee, D. Dunson, and S. Tokdar, "Efficient Gaussian Process Regression for Large Data Sets," June 29, 2011, *arXiv*: arXiv:1106.5779. doi: 10.48550/arXiv.1106.5779.

[16] N. Amaya-Tejera, M. Gamarra, J. I. Vélez, and E. Zurek, "A distance-based kernel for classification via Support Vector Machines," *Front. Artif. Intell.*, vol. 7, p. 1287875, Feb. 2024, doi: 10.3389/frai.2024.1287875.

[17] M. Gonen, E. Alpaydın, B. E. Tr, and B. E. Tr, "Multiple Kernel Learning Algorithms".

[18] L. Wang, H. Wang, and G. Fu, "Multiple Kernel Learning With Minority Oversampling for Classifying Imbalanced Data," *IEEE Access*, vol. 9, pp. 565–580, 2021, doi: 10.1109/ACCESS.2020.3046604.

[19] D. Rosati *et al.*, "Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review," *Comput. Struct. Biotechnol. J.*, vol. 23, pp. 1154–1168, Dec. 2024, doi: 10.1016/j.csbj.2024.02.018.

[20] P. Yang, P. Feng, G. Tian, G. Zhao, G. Yuan, and Y. Pan, "Integrative machine learning and bioinformatics analysis unveil key genes for precise glioma classification and prognosis evaluation," *Comput. Biol. Chem.*, vol. 119, p. 108510, Dec. 2025, doi: 10.1016/j.compbiolchem.2025.108510.

[21] V. Alur, V. Raju, B. Vastrad, C. Vastrad, S. Kavatagimath, and S. Kottursheti, "Bioinformatics Analysis of Next Generation Sequencing Data Identifies Molecular Biomarkers Associated With Type 2 Diabetes Mellitus," *Clin. Med. Insights Endocrinol. Diabetes*,

vol. 16, p. 11795514231155635, Jan. 2023, doi: 10.1177/11795514231155635.

[22] A. De Falco, Z. Dezso, F. Ceccarelli, L. Cerulo, A. Ciaramella, and M. Ceccarelli, "Adaptive one-class Gaussian processes allow accurate prioritization of oncology drug targets," *Bioinformatics*, vol. 37, no. 10, pp. 1420–1427, June 2021, doi: 10.1093/bioinformatics/btaa968.

[23] W. Pu *et al.*, "Aberrant methylation of *CDH13* can be a diagnostic biomarker for lung adenocarcinoma," *J. Cancer*, vol. 7, no. 15, pp. 2280–2289, 2016, doi: 10.7150/jca.15758.

[24] J. Magenheim *et al.*, "Universal lung epithelium DNA methylation markers for detection of lung damage in liquid biopsies," *Eur. Respir. J.*, vol. 60, no. 5, p. 2103056, Nov. 2022, doi: 10.1183/13993003.03056-2021.

[25] H. Devos, J. Zoidakis, M. G. Roubelakis, A. Latosinska, and A. Vlahou, "Reviewing the Regulators of COL1A1," *Int. J. Mol. Sci.*, vol. 24, no. 12, p. 10004, Jan. 2023, doi: 10.3390/ijms241210004.

[26] S. Khurana, A. P. Singh, A. Kumar, and R. Nema, "Prognostic value of AKT isoforms in non-small cell lung adenocarcinoma," *J. Biomed. Res.*, vol. 37, no. 3, pp. 225–228, May 2023, doi: 10.7555/JBR.36.20220138.

[27] "MYL9 binding with MYO19 suppresses epithelial-mesenchymal transition in non-small-cell lung cancer." Accessed: Nov. 19, 2025. [Online]. Available: https://journals.physiology.org/doi/epdf/10.1152/physiolgenomics.00119.2024

[28] L. Hou, T. Lin, Y. Wang, B. Liu, and M. Wang, "Collagen type 1 alpha 1 chain is a novel predictive biomarker of poor progression-free survival and chemoresistance in metastatic lung cancer," *J. Cancer*, vol. 12, no. 19, pp. 5723–5731, July 2021, doi: 10.7150/jca.59723.