Mathematical Journal of Modelling and Forecasting
Vol. 4, No. 1, June 2026, pp. 1-12

Classification of Nicotine Treatment Response Based on Gene Expression Profiles Using Support Vector Machine and Gaussian Process Models

Rahmadi Yotenka[1,3], Adhitya Ronnie Effendie[1,*], Gunardi[1], Afiahayati[2], Aisha Ellany Midnova[1], Eugenia Rivanda Gita Flamboyan[1], Daninta Indiana Mahaputri[1], Leonardo[1], Nayla Revania Dewayani[1], and Muhammad Ahnaf Billie Chesta[1]

[1] Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia
[2] Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia.
[3] Department of Statistics, Universitas Islam Indonesia, Yogyakarta, Indonesia.
[*] adhityaronnie@ugm.ac.id

Abstract. Nicotine is known to impair endothelial function and increase cardiovascular risk through transcriptional dysregulation. This study investigates the gene expression response of human induced pluripotent stem cell (iPSC)-derived endothelial cells to nicotine exposure using the RNA-seq dataset GSE274506. The analysis was conducted on 40 samples, consisting of 20 nicotine-treated and 20 untreated/control samples, using a 5-fold outer stratified cross-validation with 3-fold inner cross-validation for model tuning, reflecting the limited sample size relative to the high-dimensional gene expression feature space. Differential expression analysis identified 46 significant genes, comprising 28 upregulated and 18 downregulated, indicating perturbations in G protein-coupled receptor (GPCR) signaling, calcium homeostasis, and inflammatory processes. Functional enrichment analyses based on Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome consistently revealed dominant involvement of GPCR signaling, cyclic adenosine monophosphate (cAMP) signaling, calcium signaling, and transient receptor potential (TRP)-related pathways, suggesting a coordinated molecular response to nicotine-induced stress. To discriminate between nicotine-treated and control samples, Support Vector Machine (SVM) and Gaussian Process Classification (GPC) models were evaluated. The linear SVM achieved the best and most stable performance, with an accuracy of 0.875, an F1-score of 0.881, and a G-mean of 0.861, outperforming SVM with radial basis function kernels, single-kernel GPC variants, and a multiple kernel learning (MKL) GPC model. These findings indicate that the underlying transcriptomic structure of the data is predominantly linear, favoring linear kernel-based classifiers in high-dimensional gene expression analysis.

Keywords: Nicotine exposure, Gene expression profiling, RNA-seq, Support Vector Machine, Gaussian Process Classification, iPSC.

1 Introduction

Researchers have extensively associated nicotine, the principal bioactive component in tobacco products and various next-generation nicotine delivery devices, with adverse cardiovascular effects. Vascular endothelial cell dysfunction, an essential initial step in atherosclerosis and vascular inflammation, represents one of the primary pathogenic effects of nicotine exposure. Various studies have demonstrated that nicotine can stimulate the α7 nAChR–ERK1/2–Snail signaling pathway in human aortic endothelial cells, leading to endothelium-to-mesenchymal transition (EndMT) [1]. Further investigations have shown that nicotine alters endothelial gene expression profiles, as exemplified by the upregulation of LGALS9, a gene associated with inflammation and cell adhesion following nicotine exposure [2]. Although these findings have uncovered important aspects of nicotine-induced endothelial dysfunction, the broader endothelial response transcriptome remains largely unexplored, particularly in genetically diverse human populations. Using induced pluripotent stem cell-derived endothelial cells (iPSC-ECs) alongside advances in transcriptome profiling techniques, such as RNA sequencing (RNA-seq), has enabled the simulation of nicotine's effects in genetically relevant human systems. Studies utilizing iPSC-ECs indicate that exposure to nicotine or electronic cigarette aerosols markedly alters both mRNA and lncRNA expression. This indicates that endothelial cells are highly susceptible to nicotine-induced molecular stress [3].

An iPSC-EC model sensitive to tobacco smoke was created by [4], who also demonstrated that the model reflects transcriptional and functional changes consistent with vascular toxicity. Additionally, prolonged nicotine exposure has been demonstrated to impair the therapeutic potential of iPSC-ECs in peripheral artery disease models, underlining nicotine's deleterious effects not only on vascular health but also on cell-based treatments [4]. Further emphasizing nicotine as a key regulator of these processes, [5] combined tissue-level transcriptome investigations of cigarette smoke exposure, revealing gene expression patterns linked to endothelial dysfunction and atherosclerosis. These studies show how RNA-seq data can be used to find molecular markers that indicate how the endothelium reacts to nicotine in a normal manner.

RNA-seq datasets are valuable, yet they present several analytical challenges. Gene expression data are usually high-dimensional because they examine thousands of genes in minimal sample sizes. The biological diversity of iPSC donors significantly complicates repeatability. Differential expression analysis can help identify genes that have changed significantly. However, it can be difficult to understand long lists of DEGs without situating them within biological pathways or underlying molecular mechanisms. Meanwhile, machine learning (ML) methods are increasingly used to detect patterns in transcriptome data that can differentiate among cell types, predict treatment efficacy, and group cells into categories. A recent systematic review found that supervised learning methods, such as Support Vector Machines (SVMs) and Gaussian process models, perform well in classifying RNA sequencing data. Integrative techniques such as DEG analysis, ML modeling, and pathway enrichment analysis (e.g., GO, KEGG, Reactome) offer a more comprehensive approach to correlating gene-level modifications with biological function, molecular pathways, and potential transcriptome biomarkers [6]. In this context, the present study uses the GSE274506 RNA-seq dataset to elucidate nicotine-induced changes in gene expression in genetically diverse human iPSC-derived endothelial cells. By combining differential expression analysis, multi-database pathway enrichment, and comparative supervised machine learning models, including SVM, single-kernel GPC, and MKL-based GPC, this study provides an integrated framework for linking nicotine-induced transcriptomic alterations to biological pathway interpretation and predictive classification performance. We use differential expression analysis, machine-learning classification (especially SVM and Gaussian Process Classification (GPC) models, including single-kernel and multiple-kernel learning (MKL) variants), and pathway enrichment analysis to identify important genes and molecular pathways that underlie nicotine's effects on the cardiovascular system. This integrative paradigm is expected to identify transcriptome biomarker candidates associated with nicotine-induced cardiovascular risk and provide novel insights into the molecular mechanisms governing endothelial responses to nicotine.

2 Materials and Methods

2.1. Dataset Construction and Preprocessing

The RNA-seq gene expression dataset GSE274506 was obtained from the NCBI Gene Expression Omnibus (GEO) platform via the supplementary file GSE274506_TPM_NicotineTreated.csv.gz, while sample metadata were extracted from GSE274506_series_matrix.txt.gz using Python-based processing. *Sample_characteristics_ch1* attribute as biological labels (TREATED and CONTROL). The expression matrix was imported in TPM format, and gene identifiers were set as row indices before downstream processing. To provide an overview of the dataset structure used in this study, Table 1 presents a compact representation of the sample-by-gene expression matrix, where each row corresponds to a sample, selected gene-expression columns represent TPM values, and the final column denotes the biological response label used for supervised classification.

Table 1. Example of the sample-by-gene expression matrix used for classification

No	Sample_Id	WASH7P	MIR6859-1	RP11-34P13.15	...	*y*
1	DW09	11.794458	1.819480	10160169	...	Nicotine
2	DW10	5.878388	2.867683	4.003361	...	Nicotine
3	DW11	10.799977	7.086646	2.872204	...	Nicotine
4	DW12	12.463361	10.724332	4.691515	...	Nicotine
5	DW13	10.104777	9.415275	2.119996	...	Nicotine
6	DW14	10.438339	4.730582	6.390984	...	Nicotine

7	DW15	12.660251	0	3.541517	...	Nicotine
8	DW16	8.080543	12.981319	3.069093	...	Nicotine
⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	DW54	10.136123	17.653236	9.539762	...	Untreated

Preprocessing was performed by filtering genes with expression levels greater than 1 TPM in at least 20% of samples to remove low-expression and noisy features. Expression values were then transformed using a $\log_2(\text{TPM} + 1)$ scheme to stabilize variance and reduce skewness. Subsequently, variance-based feature selection was applied by retaining the 1,000 most variable genes across all samples, a strategy commonly employed in transcriptomic classification studies. The resulting matrix was transposed to conform to the machine-learning input format (samples x features), and all features were standardized using StandardScaler to obtain zero-mean, unit-variance distributions [7]. This preprocessing pipeline yielded a high-dimensional, filtered, and normalized expression matrix suitable for downstream classification analyses using Support Vector Machines and Gaussian Process Classification.

2.2. Differentially Expressed Genes

A fundamental transcriptomic method for identifying genes whose expression varies under different experimental conditions is called Differentially Expressed Genes (DEGs) analysis. It is predicated on the idea that biological perturbations cause systematic changes in RNA abundance that can be statistically detected and biologically interpreted [8].

Overdispersed discrete variables with a negative binomial distribution are used to represent RNA-seq count data. The observed read count for gene *i* in sample *j* is modeled as

$$k_{ij} \sim NB(\mu_{ij}\, \alpha_i);\ \mu_{ij} = s_j\, q_{ij} \tag{1}$$

where α_i is the gene-specific dispersion parameter and μ_{ij} is the anticipated expression level. A sample-specific size factor s_j is used to accommodate sequencing depth so that q_{ij} is the real underlying expression level.

A log-linear model is used to evaluate differential expression between conditions:

$$log(\mu_{ij}) = \beta_{i,0} + \beta_{i,1}\, x_j \tag{2}$$

where x_j is the experimental condition and $\beta_{i,1}$ is the log_2 fold change evaluated under the null hypothesis $H_0 : \beta_{i,1} = 0$. Either the likelihood ratio test or the Wald statistic is used for statistical inference [9], [10].

The Benjamini--Hochberg false discovery rate adjustment,

$$p_{(i)}^{adj} = min_{k \geq i} \left(\frac{n}{k}\, p_{(k)}\right) \tag{3}$$

where n is the total number of hypotheses tested, k denotes the rank position among the ordered p-values, and $p_{(k)}$ is the raw p-value at rank k. This adjustment is used to control multiple testing and account for sequencing depth and compositional bias [9], [10].

2.3. Pathway Enrichment Analysis

Compared with a reference background, pathway enrichment analysis assesses whether genes from a selected list are overrepresented in specific biological pathways. Three popular resources are the subject of this study: Reactome, KEGG, and Gene Ontology (GO).

Over-representation analysis utilizing the hypergeometric distribution is the foundation of GO enrichment [11]. Given *K* annotated genes in a backdrop of size *N*, the probability of finding at least *k* annotated genes in a gene set of size *n* is [12], [13]:

$$P(X \geq k) = 1 - \sum_{i=0}^{k-1} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}. \quad (4)$$

An enrichment score is defined as

$$ES_{KEGG}(neighbors_g, KEGG_j) = -log_{10}\left(\sum_{x=k}^{n} \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}\right). \quad (5)$$

It is frequently used to describe KEGG enrichment, which uses a similar hypergeometric framework. In this case, *k* represents the overlap size, *m* the route size, *n* the query size, and *N* the background size [14].

Reactome enrichment also uses a hypergeometric model [15], [16], with an enrichment score calculated as

$$ES_{Reactome} = log_{10}(P(X \geq k)) \quad (6)$$

and pathway significance is defined as

$$P(X \geq k) = \sum_{i=k}^{\min(n,m)} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}. \quad (7)$$

The Benjamini--Hochberg correction, as defined in Equation (3), is used to control multiple testing. Impact analysis, which weights gene contributions based on their network positions, can further integrate pathway topology. Impact analysis, which weights gene contributions based on their network positions, can further integrate pathway topology. This is calculated as

$$Impact = \frac{\sum_{i=1}^{k} PF_i}{\sqrt{N_{pathway}}} \quad (8)$$

where $N_{pathway}$ is the total number of genes in the pathway and PF_i is the perturbation factor of gene *i* based on its centrality and amplitude of differential expression [17].

2.4. Support Vector Machine

Support Vector Machine (SVM) is a kernel-based classification technique that maximizes the margin between two classes to find the best separating hyperplane [18]. The definition of binary classification in the soft-margin formulation is given by

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

subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0, \qquad i = 1, \ldots, n \quad (10)$$

where *C* regulates the trade-off between margin maximization and misclassification, and ξ_i indicates a mapping to a high-dimensional feature space [18].

The dual formulation allows for non-linear classification using linear and radial basis function (RBF) kernels. SVM only uses support vectors in this formulation, and the decision function is stated by a kernel function [19], [20],

$$K(x_i, x_j) = x_i^T x_j \quad (11)$$

For high-dimensional, small-sample transcriptomic data, such as RNA-seq, SVM remains highly successful due to its margin stability, sparsity, and structural risk minimization [21], [22]. The performance of probabilistic Gaussian Process Classification models is evaluated in this work, with SVM as the baseline classifier.

2.5. Gaussian Process Classification

Gaussian Process Classification (GPC) is a non-parametric, probabilistic classification method that performs very well for high-dimensional data, such as gene expression data [23]. Let $x_i \in \mathbb{R}^p$ denote the input feature vector of the *i*-th sample, where *p* is the number of features. GPC represents the distribution of the prediction function using a Gaussian Process (GP), which is a set of random functions that are multivariate normal at each input point. In addition to class predictions, this approach produces uncertainty estimates [24]. In binary classification, GPC assumes the existence of a latent function with a zero-mean Gaussian process prior. $f(\cdot) \sim \mathcal{GP}\big(0, k(\cdot,\cdot)\big)$, where $k\big(x_i, x_j\big)$ denotes the covariance function between two input feature vectors x_i and x_j. A link function, such as the probit function, is used to convert this latent function into class probabilities. Given the value of the latent function $f_i = f(x_i)$, Equation (12) defines the probability of a label. $y_i = 1$ in the probit approach [23].

$$P(y_i = 1|f_i) = \Phi(f_i) \tag{12}$$

where $\Phi(f_i)$ represents the cumulative distribution function of the standard normal distribution ($\mathcal{N}(0,1)$).

The main component of GPC is the kernel function. $k\big(x_i, x_j\big)$, which determines the degree of similarity between data points. Due to its ability to regulate the smoothness of the latent function via an explicit parameter v, the Matern kernel is employed as the covariance function in this study's Gaussian Process model. The Matern kernel allows modeling of functions with finite degrees of differentiability, which is often more realistic for biological systems than the squared-exponential (RBF) kernel, which assumes infinitely differentiable functions [25]. The Matern covariance between two input feature vectors x_i and x_j is defined as

$$k_{Matérn}\big(x_i, x_j\big) \;=\; \sigma^2 \,\frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\, r_{ij}}{\rho}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\, r_{ij}}{\rho}\right) \tag{13}$$

where $r_{ij} = \big\|x_i - x_j\big\|_2$ is the Euclidean distance between the two input vectors, σ^2 is signal variance, ρ is the length-scale parameter, ν regulates smoothness, and K_ν is the modified Bessel function of the second kind. The Matern kernel is a versatile generalization that encompasses both rough and smooth function classes since it converges to the RBF kernel as $\nu \to \infty$.

Apart from the Matern family, the Rational Quadratic (RQ) kernel is another alternative covariance function useful for adaptable, noise-tolerant modeling. This kernel may capture changes at multiple resolutions in gene expression data, as it can be viewed as a scale-mixing of squared-exponential kernels with varying length scales [26]. With the general form

$$k_{RQ}\big(x_i, x_j\big) \;=\; \sigma^2 \left(1 + \frac{r_{ij}^2}{2\alpha\rho^2}\right)^{-\alpha}, \tag{14}$$

The RQ kernel's sensitivity to variations in input distance is modulated by the parameter *C*. The RQ kernel offers greater flexibility in capturing diverse biological patterns, including both smooth regulatory trends and abrupt transcriptome alterations, compared with the RBF kernel, which assumes a single typical smoothness.

where $r_{ij} = \big\|x_i - x_j\big\|_2$, σ^2 is the signal variance, ρ is the length-scale parameter, and $\alpha > 0$ controls the relative weighting of different length scales. In contrast to the RBF kernel, which assumes a single typical smoothness, the RQ kernel offers greater flexibility in representing diverse biological patterns, including both smooth regulatory trends and more abrupt transcriptomic alterations.

This study also used Multiple Kernel Learning, which combines many kernels [27]. The general form is

$$k_{MKL}\big(x_i, x_j\big) = \sum_{m=1}^{M} d_m\, K_m\big(x_i, x_j\big), d_m \;\geq\; 0, \sum_{m=1}^{M} d_m \;=\; 1. \tag{15}$$

The goal of GPC is to predict the class probability for a new sample x_* [28]. To do inference, the posterior of the latent function $f_* = f(x_*)$ is integrated over the training data *D* using

$$P(y_* \;=\; 1 \mid x_*, D) \;=\; \int \Phi(f_*) p(f_* \mid D, x_*)\, df_*\,. \tag{16}$$

3 Results and Discussion

3.1. Preprocessing and Differentially Expressed Genes

Before differential expression analysis, RNA-seq data from GSE274506 were preprocessed. This stage included constructing the raw count matrix, checking for consistency between sample names and metadata, defining sample groups based on TREATED and CONTROL conditions, and filtering genes with very low expression or zero counts across all samples. After the data were prepared, pyDESeq2 was used to normalize for differences in sequencing depth across samples using the median-of-ratios approach, resulting in data more suitable for downstream statistical analysis.

Between the TREATED and CONTROL groups, differential expression analysis revealed 46 significant DEGs (adjusted p-value < 0.05, $|\log_2 FC| > 1$), including 18 down-regulated and 28 up-regulated genes. A negative binomial model with the design formula condition was used in the study, and the Benjamini-Hochberg (FDR) approach was used to account for multiple testing.

RP11-109L13.1 ($\log_2 FC = 1.46$, adj. p = 0.00075), KCNE4 ($\log_2 FC = 1.30$, adj. p = 0.00075), and DES ($\log_2 FC = 1.46$, adj. p = 0.0023) were among the top up-regulated genes. IGFBP1 ($\log_2 FC = -1.39$, adj. p = 0.00034), VIPR1 ($\log_2 FC = -1.00$, adj. p = 0.00289), and PRND ($\log_2 FC = -1.52$, adj. p = 0.0088) were among the significant down-regulated genes. These statistically significant expression changes provide strong potential biomarkers for further research, indicating significant transcriptome changes in response to therapy.

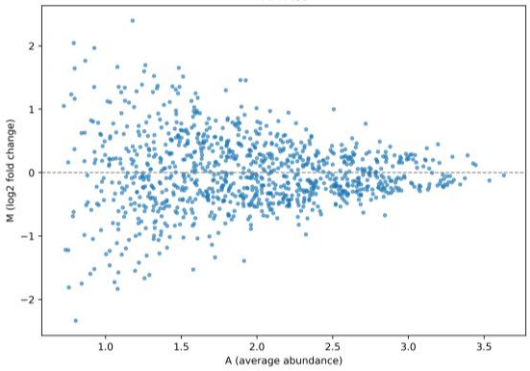

Fig 1. A MA plot of DEGs for the GSE274506 dataset

The differential expression analysis findings from the GSE274506 dataset are shown in the MA plot in Fig. 1 above. The y-axis (M) shows the $\log_2$ fold change between comparison conditions, while the x-axis (A) shows the average expression abundance ($\log_2$ scale). Each point represents a different gene. The symmetrical distribution of data points around the horizontal zero line indicates that most genes do not exhibit significant variation in expression across environments. However, several genes are candidates for increases (higher expression) or decreases (lower expression) when they deviate from zero in either direction.

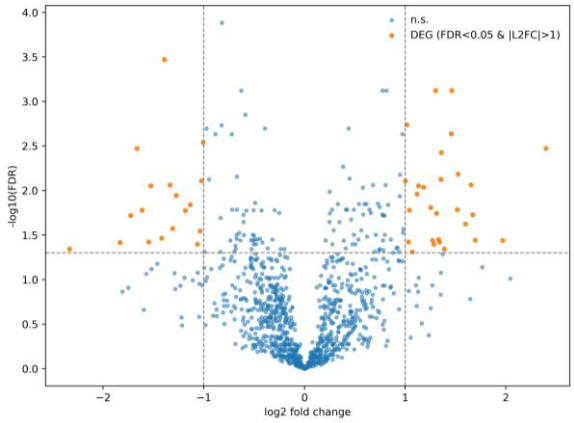

Fig 2. A volcano plot of DEGs for the GSE274506 dataset

A volcano plot showing gene distribution by statistical significance (–log10(FDR)) and the magnitude of the expression change (log_2 fold change) is shown in Fig 2 above. One gene is represented by each point, with statistically non-significant genes being represented by blue and significantly differentially expressed genes (FDR < 0.05 and | log_2FC | > 1) being shown by orange. Genes on the left have negative log_2 fold changes, indicating downregulation, whereas those on the right have positive log2 fold changes.

A heatmap of the 46 most substantially differentially expressed genes (top DEGs) found by differential expression analysis of the GSE274506 dataset is shown in Fig 3 above. Gene upregulation is indicated by red, while downregulation is indicated by blue. The z-score transformation is used to normalize expression data. The two distinct clusters formed by the TREATED and CONTROL groups indicate that the therapy significantly affects gene regulation, and the identified genes most likely play a key role in the biological processes underlying the treatment response.

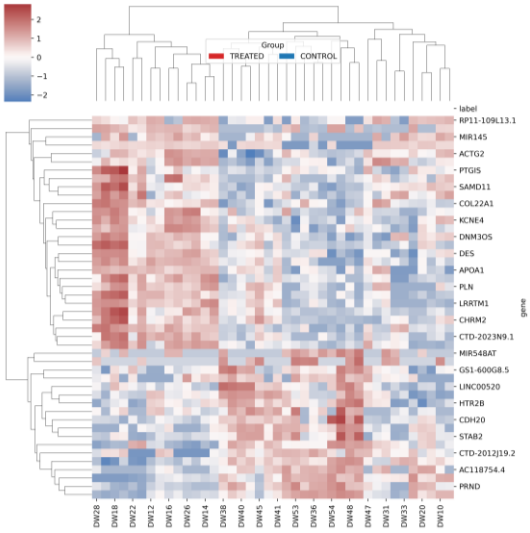

Fig 3. Heatmap of the top 46 DEGs in the GSE274506 dataset

3.2. Pathway Enrichment Analysis

A summary of the significant pathway enrichment results obtained from Gene Ontology (GO), KEGG, and Reactome analyses is presented in Table 2.

Table 2. Summary of significantly enriched pathways identified from GO, KEGG, and Reactome analyses.

Database	Pathway	Adj p-value/Overlap
GO (BP)	GPCR signaling pathway	0.0016 (4/50)
GO (BP)	Calcium ion homeostasis	< 0.05
KEGG	Neuroactive ligand–receptor interaction	5.36×10^{-4} (7/341)
KEGG	cAMP signaling pathway	0.0227 (4/216)
KEGG	Calcium signaling pathway	0.0250 (4/240)
KEGG	Inflammatory mediator regulation of TRP channels	0.0227 (3/98)
Reactome	GPCR ligand binding	< 0.05 (7/458)
Reactome	Class A/1 (rhodopsin-like) receptors	< 0.05 (6/327)
Reactome	*Gα(i)* signaling events	< 0.05 (5/312)

In nicotine-exposed iPSC-derived endothelial cells, functional enrichment analysis of the 46 DEGs (28 upregulated and 18 downregulated) consistently demonstrated activation of the GPCR–cAMP–Ca^{2+} signaling axis. The most highly enriched terms in the Gene Ontology (GO) Biological Process category were G protein-coupled receptor (GPCR) signaling pathway (adjusted p = 0.0016, overlap 4/50), followed by calcium ion homeostasis, negative regulation of inflammatory response, and adenylate cyclase–modulating GPCR signaling. In response to nicotine exposure, these systems show improved signal transduction via second messengers and recalibration of ionic homeostasis.

These results were supported by KEGG pathway analysis, which revealed significant enrichment in the following pathways: cAMP signaling pathway (p = 0.0227, 4/216), calcium signaling pathway (p = 0.0250, 4/240), inflammatory mediator regulation of TRP channels (p = 0.0227, 3/98), and neuroactive ligand–receptor interaction (adjusted $p = 5.36 \times 10^{-4}$, 7/341). These findings show a worldwide trend toward elevated regulation of inflammatory mediators, intracellular calcium dynamics, and receptor activation.

GPCR ligand binding (7/458, FDR < 0.05), Class A/1 (rhodopsin-like) receptors (6/327), GPCR downstream signaling, $G_{\alpha}(i)$ signaling events (5/312) and oxidative stress response modules, such as FOXO-mediated transcription, were among the significant clusters identified in Reactome analysis. This suggests that nicotine exposure activates oxidative stress pathways, regulates vascular contractility, and influences GPCR-mediated signal transduction.

NPY, MCHR1, KCNE4, PLN, DES, ACTG2, TNFAIP6, GSTM1, and APOA1 are among the elevated genes at the gene level that are associated with neuropeptidergic signaling and calcium-handling activities, suggesting activation of receptor-mediated contraction and extracellular matrix remodeling. In contrast, genes that are downregulated, including VIPR1, HTR2B, IGFBP1, CLEC1B, STAB2, CCL23, and TBX15, indicate that GPCR subfamilies, immunological responses, and insulin-like growth factor signaling pathways are suppressed.

Taken as a whole, these enrichment results paint a clear biological picture of how nicotine exposure triggers a coordinated endothelial response marked by activation of GPCR and TRP-mediated calcium signaling pathways, modulation of oxidative stress and contractility, and reorganization of inflammatory and metabolic circuits. These pathways suggest potential molecular indicators of nicotine-induced endothelial dysfunction.

3.3. Classification Model

The testing dataset was used to assess the effectiveness of Support Vector Machine (SVM) models with linear and Radial Basis Function (RBF) kernels following filtering and logarithmic transformation. Building on these findings, the potential for probabilistic kernel-based models to enhance uncertainty estimates and predictive performance was evaluated using Gaussian Process Classification (GPC). To get objective performance estimates, a nested 5-fold cross-validation approach was used to assess each model. Table 3 presents an overview of the categorization findings.

Table 3. Performance comparison of SVM and GPC models for the GSE274506 dataset.

Model	Acc.	Prec.	Rec.	G-mean	F1	Std. AUC
SVM (Linear)	0.875	0.893	0.900	0.861	0.881	0.088
SVM (RBF)	0.700	0.573	0.700	0.586	0.627	0.293
GPC (RBF)	0.750	0.720	0.700	0.705	0.707	0.085
GPC (Matérn)	0.750	0.720	0.700	0.705	0.707	0.085
GPC (Rational-Quadratic)	0.750	0.720	0.700	0.705	0.707	0.085
GPC (MKL-Sum)	0.800	0.883	0.750	0.774	0.777	0.093
GPC (MKL-Product)	0.750	0.853	0.700	0.705	0.707	0.085
GPC (MKL-Hierarchical)	0.800	0.883	0.750	0.774	0.777	0.085

With an accuracy of 0.875 and an F1-score of 0.881, the linear SVM performed the best and most consistently out of all the SVM variations. This model also achieved the highest geometric mean (G-mean = 0.861), indicating a balanced ability to classify both nicotine-treated and control samples correctly. The SVM-RBF model, on the other hand, performed less consistently and achieved lower accuracy, suggesting it was less resilient on this dataset. RBF, Matern, and Rational Quadratic, the single-kernel GPC models, performed similarly, with F1-scores of 0.707 and accuracies of about 0.750. These models offer probabilistic predictions that enable uncertainty quantification at the sample level, albeit somewhat less effectively than linear SVMs. Compared with single-kernel GPC models, GPC Multiple Kernel Learning (MKL) techniques performed better; the sum and hierarchical strategies achieved accuracies of 0.800 and F1-scores of 0.777. These findings, however, did not outperform the linear SVM, indicating that the gene expression patterns in the GSE274506 dataset are essentially linearly separable. The linear SVM proved to be the most effective overall, providing the best stability and predicted accuracy while keeping the model simple. This makes it an appropriate and effective option for categorizing the response of iPSC-derived endothelial cells to nicotine therapy.

3.4. Feature Importance and Biological Interpretation

The testing dataset was used to assess the effectiveness of Support Vector Machine (SVM) models with linear and Radial Basis Function (RBF) kernels following filtering and logarithmic transformation.

Even though Table 2 shows that the linear SVM achieved the best predictive performance, feature selection in high-dimensional RNA-seq data is highly sensitive to model selection and sampling variability. Instead of relying on a single model, we used a model-agnostic stability technique to ensure robustness by combining feature selection outcomes across all classifiers and folds of the nested CV.

Three complementary signals were used to quantify feature importance: (i) selection frequency across all models (SVM-linear, SVM-RBF, GPC-RBF, GPC-Matérn, GPC-Rational Quadratic, and MKL variants) and across nested CV; (ii) differential expression statistics (FDR, ROC-AUC, and $\log_2$FC); and (iii) univariate discriminative power. To rank stable and biologically significant genes, they were merged into a composite *BIOMARKER_SCORE*. The top 10 genes with the highest composite scores are shown in Table 4.

Table 4. Performance top 10 candidate biomarkers ranked by the composite *BIOMARKER_SCORE*.

Gene	Score	Stab.	Freq.	AUC	$\log_2$FC	Direction
NPY	0.719	1.0	25	0.109	−1.91	Down in the untreated
AC078942.1	0.641	0.6	15	0.835	1.29	Up in the untreated
MIR145	0.609	1.0	25	0.178	−1.40	Down in the untreated
CDH20	0.594	0.6	15	0.798	1.17	Up in the untreated
PRND	0.587	0.6	15	0.788	1.11	Up in the untreated
IGFBP1	0.516	0.2	5	0.913	0.93	Up in the untreated
TNFAIP6	0.505	0.8	20	0.140	−1.24	Down in the untreated
MIR548AT	0.501	0.4	10	0.723	1.58	Up in the untreated
GSTM1	0.480	0.6	15	0.228	−1.59	Down in the untreated
CTD-3032J10.4	0.475	0.4	10	0.800	1.01	Up in the untreated

High-scoring genes generally show robust differential expression, considerable discriminative power, and consistent selection across models and folds. Recurrent selection and coherent regulatory patterns of genes, including NPY, MIR145, TNFAIP6, and GSTM1, suggest their roles in endothelial responses associated with nicotine. Several genes, including CDH20, PRND, and IGFBP1, were more highly expressed in controls, indicating pathways that were inhibited following nicotine use. These genes are relevant as potential biomarkers of nicotine-induced endothelial dysfunction because several of them are associated with enriched pathways in inflammation, GPCR signaling, calcium signaling, and oxidative stress.

3.5. Discussion

Using a human iPSC-derived endothelial cell (iPSC-EC) model, this study combines differential expression analysis, pathway enrichment, and kernel-based machine learning to give an integrative assessment of the endothelial transcriptional response to nicotine. 46 genes were found to be differentially expressed after nicotine exposure using RNA-seq data from genetically diverse iPSC-ECs. These genes showed coordinated regulation of oxidative stress and inflammatory pathways, disruption of calcium homeostasis, and activation of GPCR-mediated signaling. These results confirm that iPSC-ECs are a suitable in vitro technology for capturing nicotine-induced endothelial dysfunction at the transcriptome level.

Pathway enrichment analysis further indicated that nicotine produces a system-level reconfiguration of endothelial signaling in iPSC-ECs, highlighted by coordinated activation of the GPCR-cAMP-Ca2+ axis, TRP channel-mediated inflammatory pathways, and oxidative stress-responsive programs. The convergence of these pathways indicates that nicotine affects endothelial function through closely related receptor-mediated and second-messenger signaling pathways, which are consistently observed in endothelial populations derived from genetically heterogeneous iPSCs.

Machine learning analyses validated these biological findings. The linear SVM outperformed the non-linear SVM and Gaussian Process classifiers in terms of accuracy and stability across all assessed models. This data demonstrates that nicotine-associated transcriptional alterations in iPSC-ECs are largely linear and driven by broad, coherent shifts across gene expression profiles, which are efficiently captured by linear decision limits.

More complicated non-linear models did not generate incremental performance gains, suggesting minimal higher-order interaction effects in the present iPSC-EC transcriptomic dataset.

Feature importance analysis identified numerous substantial discriminative markers, including NPY, MIR145, TNFAIP6, and GSTM1, all of which have well-established roles in vascular regulation, inflammatory responses, and oxidative stress. The uniform selection of these genes across models highlights their potential importance as transcriptome indicators of nicotine-induced endothelial dysfunction in human iPSC-ECs, with possible implications for cardiovascular risk assessment.

Despite these revelations, there are several limitations to this study. The analysis lacked external validation across separate cohorts and was limited to a single iPSC-EC RNA-seq dataset. In addition, functional validation at the protein or cellular level was not performed, and focusing solely on transcriptome data may not adequately capture post-transcriptional regulation. Future studies incorporating bigger iPSC-EC cohorts, experimental validation, and multiomics data will be necessary to increase the translational significance of the identified biomarkers and pathways.

4 Conclusion

This study merges transcriptome profiling with kernel-based machine learning to characterize the molecular response of human iPSC-derived endothelial cells to nicotine exposure. A total of 46 important genes were identified through differential expression analysis, and their combined roles indicate activation of GPCR-mediated signaling, altered calcium homeostasis, inflammatory regulation, and oxidative stress responses. GO, KEGG, and Reactome enrichment analyses consistently highlighted the GPCR-cAMP-Ca^{2+} axis and TRP-related pathways as essential components of the endothelial response to nicotine.

The Support Vector Machine with a linear kernel outperformed SVM-RBF, single-kernel GPC models, and multiple-kernel GPC variations, achieving the highest and most consistent performance among the classification models studied (accuracy = 0.875, F1-score = 0.881). This finding suggests that transcriptomic divergence between nicotine-treated and untreated samples is largely linear, favoring simpler kernel architectures over more complex non-linear models.

Feature-importance analysis further identified several strong candidate biomarkers, including NPY, MIR145, TNFAIP6, and GSTM1, which demonstrated substantial differential expression, stable model selection, and clear biological relevance. Overall, these results highlight the value of integrating differential expression analysis, pathway enrichment, and machine learning to uncover mechanistic insights, potential biomarkers of nicotine-induced endothelial dysfunction, and key molecular pathways disrupted by nicotine exposure.

References

[1] W. Qin *et al.*, “Endothelial to mesenchymal transition contributes to nicotine-induced atherosclerosis,” *Theranostics*, vol. 10, no. 12, pp. 5276–5289, 2020, doi: 10.7150/thno.42470.

[2] S. M. Braß *et al.*, “Nicotine Potentially Alters Endothelial Inflammation and Cell Adhesion via LGALS9,” *J. Cardiovasc. Dev. Dis.*, vol. 11, no. 1, p. 6, Dec. 2023, doi: 10.3390/jcdd11010006.

[3] N. Jimenez-Tellez, D. Williams, Y. Liu, M. Wang, M. Chandy, and J. C. Wu, “Transcriptomic analysis of nicotine on the cardiovascular system using a diverse population of human induced pluripotent stem cell-derived endothelial cells,” *J. Mol. Cell. Cardiol.*, vol. 198, pp. 21–23, Jan. 2025, doi: 10.1016/j.yjmcc.2024.11.001.

[4] A. H. P. Chan, C. Hu, G. C. F. Chiang, C. Ekweume, and N. F. Huang, “Chronic nicotine impairs the angiogenic capacity of human induced pluripotent stem cell-derived endothelial cells in a murine model of peripheral arterial disease,” *JVS-Vasc. Sci.*, vol. 4, p. 100115, 2023, doi: 10.1016/j.jvssci.2023.100115.

[5] A. Khaleel, B. Alkhawaja, T. S. Al-Qaisi, L. Alshalabi, and A. H. Tarkhan, “Pathway analysis of smoking-induced changes in buccal mucosal gene expression,” *Egypt. J. Med. Hum. Genet.*, vol. 23, no. 1, p. 69, Dec. 2022, doi: 10.1186/s43042-022-00268-y.

[6] X. Cao, L. Xing, E. Majd, H. He, J. Gu, and X. Zhang, “A Systematic Evaluation of Supervised Machine Learning Algorithms for Cell Phenotype Classification Using Single-Cell RNA Sequencing Data,” *Front. Genet.*, vol. 13, p. 836798, Feb. 2022, doi: 10.3389/fgene.2022.836798.

[7] Q. Sun *et al.*, "Characterizing hub biomarkers for metabolic-induced endothelial dysfunction and unveiling their regulatory roles in EndMT through RNA sequencing and machine learning approaches," *Front. Cardiovasc. Med.*, vol. 12, p. 1585030, May 2025, doi: 10.3389/fcvm.2025.1585030.

[8] 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.

[9] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.

[10] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR : a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010, doi: 10.1093/bioinformatics/btp616.

[11] The Gene Ontology Consortium *et al.*, "The Gene Ontology knowledgebase in 2023," *GENETICS*, vol. 224, no. 1, p. iyad031, May 2023, doi: 10.1093/genetics/iyad031.

[12] T. Wu *et al.*, "clusterProfiler 4.0: A universal enrichment tool for interpreting omics data," *The Innovation*, vol. 2, no. 3, p. 100141, Aug. 2021, doi: 10.1016/j.xinn.2021.100141.

[13] K. Wijesooriya, S. A. Jadaan, K. L. Perera, T. Kaur, and M. Ziemann, "Urgent need for consistent standards in functional enrichment analysis," *PLOS Comput. Biol.*, vol. 18, no. 3, p. e1009935, Mar. 2022, doi: 10.1371/journal.pcbi.1009935.

[14] F. Huang *et al.*, "Analysis and prediction of protein stability based on interaction network, gene ontology, and KEGG pathway enrichment scores," *Biochim. Biophys. Acta Proteins Proteomics*, vol. 1871, no. 3, p. 140889, May 2023, doi: 10.1016/j.bbapap.2023.140889.

[15] B. Jassal *et al.*, "The reactome pathway knowledgebase," *Nucleic Acids Res.*, p. gkz1031, Nov. 2019, doi: 10.1093/nar/gkz1031.

[16] A. Fabregat *et al.*, "The Reactome Pathway Knowledgebase," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D649–D655, Jan. 2018, doi: 10.1093/nar/gkx1132.

[17] M. Milacic *et al.*, "The Reactome Pathway Knowledgebase 2024," *Nucleic Acids Res.*, vol. 52, no. D1, pp. D672–D678, Jan. 2024, doi: 10.1093/nar/gkad1025.

[18] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. in Springer Texts in Statistics. New York, NY: Springer US, 2021. doi: 10.1007/978-1-0716-1418-1.

[19] 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.

[20] 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.

[21] R. D. Leone, F. Maggioni, and A. Spinelli, "A Robust Twin Parametric Margin Support Vector Machine for Multiclass Classification," Jun. 24, 2025, *arXiv*: arXiv:2306.06213. doi: 10.48550/arXiv.2306.06213.

[22] B. Guo *et al.*, "Identification of potential biomarkers in cardiovascular calcification based on bioinformatics combined with single-cell RNA-seq and multiple machine learning analysis," *Cell. Signal.*, vol. 131, p. 111705, Jul. 2025, doi: 10.1016/j.cellsig.2025.111705.

[23] 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.

[24] A. Banerjee, D. Dunson, and S. Tokdar, "Efficient Gaussian Process Regression for Large Data Sets," Jun. 28, 2011, *arXiv*: arXiv:1106.5779. doi: 10.48550/arXiv.1106.5779.

[25] A. Muyskens, B. W. Priest, I. R. Goumiri, and M. D. Schneider, "Identifiability and Sensitivity Analysis of Kriging Weights for the Matern Kernel," Oct. 10, 2024, *arXiv*: arXiv:2410.08310. doi: 10.48550/arXiv.2410.08310.

[26] C. Cortes, P. Haffner, and M. Mohri, "Rational Kernels: Theory and Algorithms".

[27] M. Briscik, G. Tazza, L. Vidács, M.-A. Dillies, and S. Déjean, “Supervised multiple kernel learning approaches for multi-omics data integration,” *BioData Min.*, vol. 17, no. 1, p. 53, Nov. 2024, doi: 10.1186/s13040-024-00406-9.

[28] 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.