Comparative Study of SVM and Neural Networks in Classifying B-Lymphocyte Gene Expression for Smoker Detection

1st Destiana Putri
Teknik Industri
Universitas Bina Sarana Informatika
Jakarta, Indonesia
destiana.dtp@bsi.ac.id

2nd Miwan Kurniawan Hidayat
Teknik Industri
Universitas Bina Sarana Informatika
Jakarta, Indonesia
miwan@bsi.ac.id

3rd Eni Irfiani
Sistem Informasi
Universitas Bina Sarana Informatika
Jakarta, Indonesia
eni.enf@bsi.ac.id

4th Sri Watmah
Teknik Elektro
Universitas Bina Sarana Informatika
Jakarta, Indonesia
sriwatmah.wtm@bsi.ac.id

5th Ade Suryanto
Teknik Industri
Universitas Bina Sarana Informatika
Jakarta, Indonesia
ade.ayo@bsi.ac.id

6th Sigit Adi Pratama
Teknik Industri
Universitas Bina Sarana Informatika
Jakarta, Indonesia
sigit.sgp@bsi.ac.id

2025 IEEE International Conference on Advanced Information Scientific Development (ICAISD) | 979-8-3315-7499-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICAISD68166.2025.11385527

***Abstract*—Smoking is a major risk factor that poses serious impacts on global public health. Early detection of smoking status is crucial for supporting prevention efforts and health promotion. This study aims to compare the performance of two classification methods, namely Support Vector Machine (SVM) and Neural Network (NN), in identifying smokers and non-smokers. Model evaluation was conducted using several metrics, including the Area Under the Curve (AUC), accuracy, F1-score, precision, recall, and Matthews Correlation Coefficient (MCC). The experimental results showed that the SVM achieved the best performance, with an AUC of 0.984, accuracy of 0.924, F1-score of 0.924, precision of 0.924, recall of 0.924, and MCC of 0.848. Meanwhile, the NN also demonstrated excellent results, with an AUC of 0.972, accuracy of 0.911, F1-score of 0.911, precision of 0.912, recall of 0.911, and MCC of 0.823. Therefore, it can be concluded that SVM outperforms NN in classifying smoking status, although both methods are capable of providing a very high classification performance. Furthermore, the findings indicate that exposure to cigarette smoke can lead to immune gene dysregulation, including reduced expression of genes associated with cellular defense, increased inflammatory mediators, and alterations in adaptive immune cell functions**

***Index Terms*—Smoking, Gen expression, Machine Learning, SVM, Neural Network**

I. Introduction

The prevalence of tobacco smoking poses a critical challenge to global health [1], causing more than eight million deaths worldwide each year, and continues to be the subject of extensive scientific investigations [2], [3]. In addition to its impact on the respiratory and cardiovascular systems, smoking has widespread consequences for the immune system [4]. In addition to its impact on the respiratory and cardiovascular systems, smoking exerts broad consequences on the immune system [5]. Transcriptomic and epigenetic studies have shown that exposure to cigarette smoke alters gene expression patterns in peripheral blood cells, including B lymphocytes, which play a vital role in adaptive immune responses [6]. These alterations affect various biological pathways, including inflammation, xenobiotic metabolism, apoptosis, and immune cell differentiation [7].

Although several genetic biomarkers have been reported, most studies have focused on univariate analyses or single-gene associations [8]. This approach has limitations in capturing the complexity of multigenic interactions underlying smoking status [9]. In this context, machine learning (ML) has emerged as a relevant approach [10]. Machine learning enables the simultaneous modeling of large-scale gene expression patterns, thereby enhancing the ability to classify smoking status [11]. However, the main challenge lies in selecting an appropriate model for high-dimensional data with a limited number of samples, which is prone to overfitting and poor generalization [12].

The Support Vector Machine (SVM) is well recognized for its strength in handling high-dimensional data owing to its ability to identify an optimal hyperplane that separates classes [11], [13]. On the other hand, neural networks (NN) represent a more flexible nonlinear model capable of capturing complex relationships among variables, although they require a larger number of parameters and data to achieve stable training [14] [14]. To date, there are limited studies that systematically compare the performance of these two approaches in classifying smoking status based on B lymphocyte gene expression [15], [16].

In this context, the present study aimed to develop a classification model of smoking status using B lymphocyte gene expression data and to compare the performance of SVM and neural networks using multiple evaluation metrics (AUC, accuracy, F1-score, precision, recall, and MCC). The contribution of this study lies in providing a reproducible methodological

baseline for research on smoking status biomarkers, while also offering comparative insights into the effectiveness of SVM and Neural Network algorithms on high-dimensional gene expression data.

II. Review Of Releted Work

Research on smoking status classification based on gene expression has advanced rapidly in line with the progress of high-throughput omics technologies and machine learning (ML) methods [17]. The main challenge in this field lies in the complexity of high-dimensional transcriptomic data combined with the limited number of samples, making the choice of an An appropriate ML model is crucial to avoid overfitting while ensuring good generalization [18].

The study "Machine Learning Reveals Impacts of Smoking on Gene Profiles of Different Cell Types in Lung" demonstrated that ML algorithms can be employed to identify differences in gene expression patterns between smokers and non-smokers across various lung cell types. This study highlights that cigarette smoke exposure leaves specific molecular signatures that can be modeled using ML, thereby opening opportunities to discover novel biomarkers associated with smoking behavior [7].

In addition, Yang et al. (2018), in their study entitled "Construction of a 26-feature gene Support Vector Machine classifier for smoking and non-smoking lung adenocarcinoma sample classification," developed a classification model based on support vector machine (SVM). They demonstrated that with proper feature selection, SVM is capable of distinguishing between smoker and non-smoker samples in lung adenocarcinoma cases with high accuracy. This finding underscores the strength of SVM in handling high-dimensional data, particularly in the context of gene expression [19].

On the other hand, Zhang et al. (2019), in their article "Machine learning selected smoking-associated DNA methylation signatures that predict HIV prognosis and mortality," highlighted the importance of feature selection strategies in ML. By employing regularization approaches such as LASSO, they successfully identified epigenetic markers associated with smoking behavior and their relevance in predicting HIV prognosis. This indicates that the integration of ML with regularization methods can reduce data complexity and improve the classification performance [20].

Babichev et al. (2021), in their study entitled "A Hybrid Model of Cancer Diseases Diagnosis Based on Gene Expression Data with Joint Use of Data Mining Methods and Machine Learning Techniques," explored various ML algorithms, including SVM and neural networks, for cancer diagnosis based on gene expression data. Their findings revealed that combining data mining approaches with ML techniques can enhance classification accuracy while emphasizing the importance of selecting algorithms that align with the characteristics of the data [17].

Based on this review, it can be concluded that although numerous studies have applied ML in the context of smoking and genetic data, there are limited investigations specifically comparing the performance of different ML methods for smoking status classification based on B lymphocyte gene expression. Therefore, this study seeks to address this gap by developing a classification model and conducting a comparative evaluation of SVM and Neural Network algorithms on high-dimensional gene expression data. The following is a literature review of this study.

A. Support Vector Machine (SVM)

Support Vector Machine (SVM) is well-recognized for its effectiveness in classifying high-dimensional data such as the gene expression. This method operates by determining the optimal hyperplane that maximizes the separation margin between classes, thereby achieving strong generalization even when the number of features (p) is substantially larger than the number of samples (n) [21].

B. Multi-Layer Perceptron (MLP) Neural Network

The Multi-Layer Perceptron (MLP) neural network is capable of learning non-linear representations through hidden layers, resulting in a flexible modeling approach. Nevertheless, its performance is highly dependent on the architecture, the number of neurons and layers, the application of regularization techniques, and the size of the dataset [22].

Several studies have employed machine learning (ML) including Support Vector Machines (SVM) and Neural Networks to detect molecular signatures (biomarkers) of smoking status using PBMC or whole blood data. For example, studies by [11] identified genes such as GPR15 and LRRN3 as strong markers of smoking, which was supported by ML-based analyses.

III. Methode

The subjects were selected from a data source available in Orange, namely Smoking Effect on B Lymphocytes. This dataset contains 79 samples (39 smokers and 40 non-smokers), 3,001 variables (gene expression features randomly selected from the original data), with the target variable being smoker vs. non-smoker.

B lymphocytes are a type of white blood cellthat play a key role in the immune system, particularly in antibody production. This study aims to classify smoking status (smoker vs. non-smoker) based on B lymphocyte gene expression and to compare the performance of Support Machine (SVM), and Neural Network (MLP). The following are the stages of the research methodology:

Based on the research stages illustrated above, the The following explanations are provided:

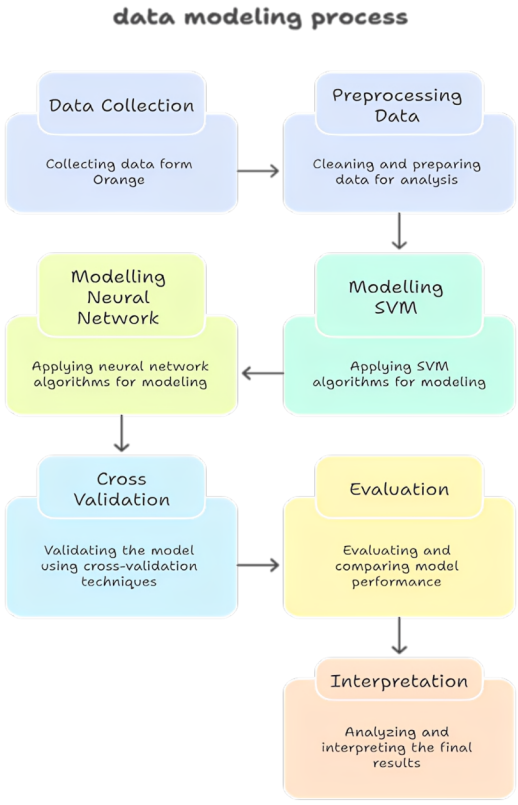

Fig. 1. Research Stages.

A. Data Collection

B lymphocyte gene expression dataset (tabular format; features = genes; label = smoking status). Samples were selected with clearly defined smoking status labels, and adequate expression quality. The following figure presents the data utilized in the classification process in this study:

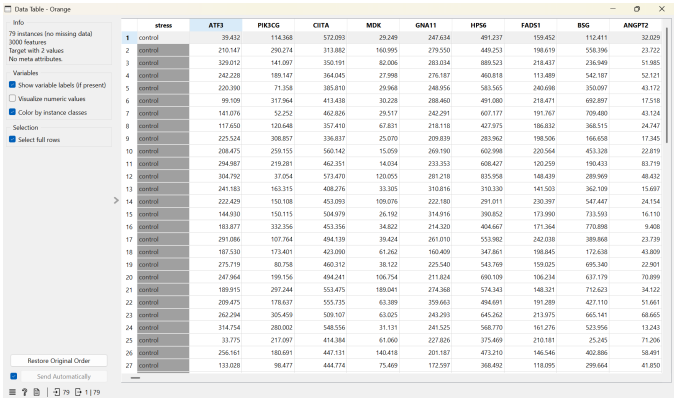

Fig. 2. Smoking effect on B lymphocytes Data

B. Preprocessing

Data cleaning was performed by removing features with entirely missing or constant values. Standardization using z-score normalization per gene (mean = 0, sd = 1) within each fold. Light feature selection: removal of near-constant or low-variance genes to reduce noise.

C. Modelling

Model 1 – SVM (RBF); Effective for cases where p - n with maximum margin optimization, suitable for high-dimensional data. Model 2 – Neural Network (MLP). Capable of capturing non-linear relationships, overfitting mitigated through regularization and early stopping.

D. Cross-Validation

Stratified 5-fold cross-validation was applied to preserving class proportions.

E. Evaluation

Primary metrics include AUC, Accuracy (CA), Precision, Recall, F1-score, and MCC (more informative for imbalanced data), comparing the performance of these two models.

F. Interpretation

Analysis was conducted to determine the best- performing model.

IV. RESULTS AND DISCUSSION

The gene expression profile of B lymphocytes can classify smoker vs. non-smoker status with very high performance. The results support the use of gene expression–based machine learning as a molecular biomarker of cigarette smoke exposure. Findings indicate significant differences in B lymphocyte gene expression patterns between smokers and non-smokers. Many of the involved genes are associated with adaptive immune functions, inflammation, and oxidative stress responses induced by nicotine exposure in rats. These gene expression changes provide further evidence that smoking weakens immune function and increases Susceptibility to diseases. The following is an overview of data classification processes, including data preprocessing, cross- validation, and modeling using the SVM and Neural Network methods.

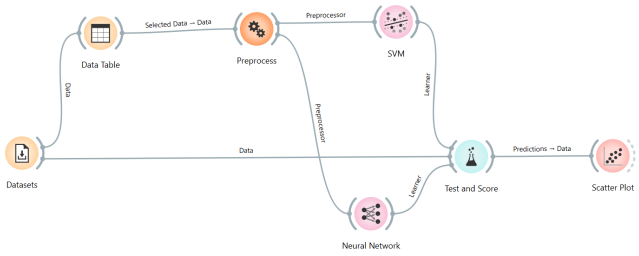

Fig. 3. Preprocessing Settings with Two Models.

The figure above illustrates the application of two modeling approaches: the Support Vector Machine (SVM), a classification algorithm based on the optimal hyperplane, and the Neural Network (Multi-Layer Perceptron, MLP), an artificial neural network algorithm with backpropagation. Both models

were evaluated using the Test and Score procedure with 5-fold stratified cross-validation to maintain class proportions and assess model performance.

The B lymphocyte gene expression data were obtained from transcriptomic databases. Normalization and feature selection were performed to reduce noise and improve the model accuracy. A number of significant genes associated with cigarette smoke exposure were identified, including those related to immune response (e.g., CD19, CD79, etc.). The following section presents the classification The results obtained from both models are as follows:

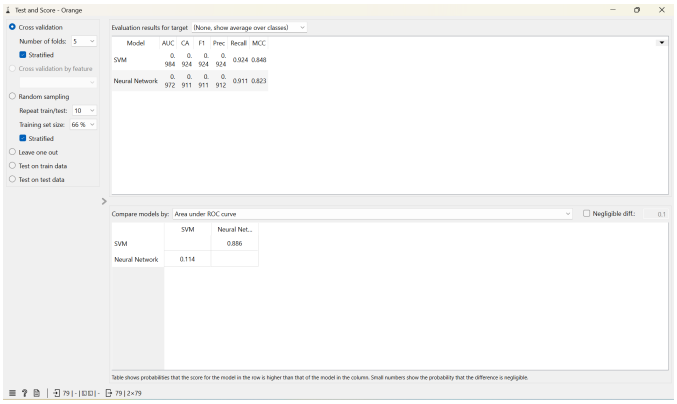

Fig. 4. Classification Results.

The results demonstrate significant differences in the B lymphocyte gene expression profiles between smokers and non-smokers. Many of the differentially expressed genes are associated with adaptive immune function, inflammation, and oxidative stress responses induced by nicotine exposure. These gene expression alterations support the evidence that smoking weakens immune function, and increases susceptibility to diseases.

Support Vector Machine (SVM) proved to be superior in this dataset, consistent with its effectiveness in handling high-dimensional data such as gene expression. The evaluation results show that SVM achieved the best performance on the Smoking Effect on B Lymphocytes dataset. The model obtained an AUC of 0.984, accuracy (CA) of 0.924, F1-score of 0.924, precision of 0.924, recall of 0.924, and MCC of 0.848. These findings indicate that SVM is highly capable of classifying smoking status with excellent accuracy and a well-balanced trade-off between the precision and recall.

Meanwhile, the Neural Network (NN) also demonstrated strong performance, albeit slightly lower, likely due to the limited sample size (only 79 samples). Neural networks generally require larger datasets to avoid overfitting and to better capture complex non- linear patterns. In this study, NN achieved an AUC of 0.972, accuracy (CA) of 0.911, F1-score of 0.911, precision of 0.912, recall of 0.911, and MCC of 0.823. These results confirm that NN provides competitive classification performance, although it does not surpass SVM on this dataset.

This study highlights that machine learning can serve as an important tool in bioinformatics education, public health, and molecular biology, respectively. It also illustrates how computational algorithms can be applied to real biological data to support health-related research in the future.

BIBTEX does not work like magic. It doesn't get the bibliographic data from thin air, but from. bib files. If you use BIBTEX to produce a bibliography you must send the .bib files.

LATEX cannot read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross-referenced as Table IV-B3.

CONCLUSION

Based on the findings obtained from the Smoking Effect on B Lymphocytes dataset, it can be concluded that both Support Vector Machine (SVM) and Neural Network (NN) are capable of classifying smoker and Non-smoker status with high performance. However, SVM demonstrated superior results compared to NN, achieving an AUC of 0.984 and an accuracy of 0.924, which are slightly higher than those obtained by NN (AUC: 0.972, accuracy 0.911).

This indicates that SVM is more effective in handling high-dimensional gene expression data and is able to maintains a balance between precision and recall. Nevertheless, NN remains a relevant alternative classification model with a competitive performance. These findings provide a valuable methodological baseline for developing predictive models of smoking status based on gene expression data and open opportunities for further research through external validation and integration of multi-omics approaches.

This study highlights that the integration of machine learning with biological data can generate promising biomarker prediction tools for detecting smoking exposure.

REFERENCES

[1] N. Baiju, T. M. Sandanger, P. Sætrom, and T. H. Nøst, "Gene expression in blood reflects smoking exposure among cancer-free women in the Norwegian Women and Cancer (NOWAC) postgenome cohort," *Scientific Reports*, vol. 11, no. 1, pp. 1–13, 2021. [Online]. Available: https://doi.org/10.1038/s41598-020-80158-8

[2] S. S. Saharan, P. Nagar, K. T. Creasy, E. O. Stock, J. Feng, M. J. Malloy, and J. P. Kane, "Smoking Classification Using Novel Plasma Cytokines by Implementing Machine Learning and Statistical Methods," *Proceedings - 2023 International Conference on Computational Science and Computational Intelligence, CSCI 2023*, pp. 686–694, 2023.

[3] T. Haase, C. Müller, J. Krause, C. Röthemeier, J. Stenzig, S. Kunze, M. Waldenberger, T. Münzel, N. Pfeiffer, P. S. Wild, M. Michal, F. Marini, M. Karakas, K. J. Lackner, S. Blankenberg, and T. Zeller, "Novel DNA methylation sites influence GPR15 expression in relation to smoking," *Biomolecules*, vol. 8, no. 3, pp. 1–10, 2018.

[4] R. Chen and J. Lin, "Identification of feature risk pathways of smoking-induced lung cancer based on SVM," *PLoS ONE*, vol. 15, no. 6, pp. 1–16, 2020. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0233445

[5] G. Kõks, M. L. Uudelepp, M. Limbach, P. Peterson, E. Reimann, and S. Kõks, "Smoking-induced expression of the GPR15 gene indicates its potential role in chronic inflammatory pathologies," *American Journal of Pathology*, vol. 185, no. 11, pp. 2898–2906, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.ajpath.2015.07.006

[6] X. Wang, M. R. Campbell, H. Y. Cho, G. S. Pittman, S. N. Martos, and D. A. Bell, "Epigenomic profiling of isolated blood cell types reveals highly specific B cell smoking signatures and links to disease risk," *Clinical Epigenetics*, vol. 15, no. 1, pp. 1–20, 2023. [Online]. Available: https://doi.org/10.1186/s13148-023-01507-8

[7] Q. Ma, Y. Shen, W. Guo, K. Feng, T. Huang, and Y. Cai, "Machine Learning Reveals Impacts of Smoking on Gene Profiles of Different Cell Types in Lung," *Life*, vol. 14, no. 4, 2024.
[8] M. A. Hossain, M. Z. Rahman, T. Bhuiyan, and M. A. Moni, "Identification of Biomarkers and Molecular Pathways Implicated in Smoking and COVID-19 Associated Lung Cancer Using Bioinformatics and Machine Learning Approaches," *International Journal of Environmental Research and Public Health*, vol. 21, no. 11, 2024.
[9] T. Liu, Y. Huang, Q. Hui, and Y. V. Sun, "Epigenetic prediction of smoking status using machine-learning methods," *s*, pp. 1–23, 2020.
[10] S. Bollepalli, T. Korhonen, J. Kaprio, S. Anders, and M. Ollikainen, "Epismoker: A Robust Classifier to Determine Smoking Status from DNA Methylation Data," in *Epigenomics*, 2016. [Online]. Available: http://tandfonline.com/doi/full/10.2217/epi-2019-0206
[11] F. Huang, Q. Ma, J. Ren, J. Li, F. Wang, T. Huang, and Y. D. Cai, "Identification of Smoking-Associated Transcriptome Aberration in Blood with Machine Learning Methods," *BioMed Research International*, vol. 2023, 2023.
[12] Y. Wang, D. J. Miller, and R. Clarke, "Approaches to working in high-dimensional data spaces: Gene expression microarrays," *British Journal of Cancer*, vol. 98, no. 6, pp. 1023–1028, 2008.
[13] Victo Sudha George and Cyril Raj, "Review On Feature Selection Techniques And The Impact Of Svm For Cancer Classification Using Gene Expression Profile," *International Journal of Computer Science Engineering Survey*, vol. 2, no. 3, pp. 16–27, 2011.
[14] S. S. Saharan, P. Nagar, K. T. Creasy, E. O. Stock, J. Feng, M. J. Malloy, and J. P. Kane, "Optimization of Smoking Classification by Applying Neural Network with Variable Importance Using Cytokine Biomarkers," *Proceedings - 2023 International Conference on Computational Science and Computational Intelligence, CSCI 2023*, pp. 661–670, 2023.
[15] J. K. Nowak, E. Dybska, A. T. Adams, and J. Walkowiak, "Immune cell-specific smoking-related expression characteristics are revealed by re-analysis of transcriptomes from the CEDAR cohort," *Central European Journal of Immunology*, vol. 47, no. 3, pp. 246–259, 2022.
[16] P. Beineke, K. Fitch, H. Tao, M. R. Elashoff, S. Rosenberg, W. E. Kraus, and J. A. Wingrove, "A whole blood gene expression-based signature for smoking status," *BMC Medical Genomics*, vol. 5, no. 1, p. 1, 2012. [Online]. Available: BMCMedicalGenomics
[17] Y.-D. L. L. I. Babichev, Sergii, "A Hybrid Model of Cancer Diseases Diagnosis Based on Gene Expression Data with Joint Use of Data Mining Methods and Machine Learning Techniques," *Applied Sciences (Switzerland)*, vol. 13, 2023.
[18] B. H. Chen, "Minimum standards for evaluating machine-learned models of high-dimensional data," *Frontiers in Aging*, vol. 3, no. September, pp. 1–6, 2022.
[19] L. Yang, L. Sun, W. Wang, H. Xu, Y. Li, J. Y. Zhao, D. Z. Liu, F. Wang, and L. Y. Zhang, "Construction of a 26-feature gene support vector machine classifier for smoking and non-smokinglung adenocarcinoma sample classification," *Molecular Medicine Reports*, vol. 17, no. 2, pp. 3005–3013, 2018.
[20] X. Zhang, Y. Hu, B. E. Aouizerat, G. Peng, V. C. Marconi, M. J. Corley, T. Hulgan, K. J. Bryant, H. Zhao, J. H. Krystal, A. C. Justice, and K. Xu, "Machine learning selected smoking-associated DNA methylation signatures that predict HIV prognosis and mortality," *Clinical Epigenetics*, vol. 10, no. 1, pp. 1–15, 2018.
[21] E. P. Chou and T.-W. Ko, "Dimension Reduction of High-Dimensional Datasets Based on Stepwise SVM," *Journal*, no. 64, 2017. [Online]. Available: http://arxiv.org/abs/1711.03346
[22] H. Yu, D. C. Samuels, Y. yong Zhao, and Y. Guo, "Architectures and accuracy of artificial neural network for disease classification from omics data," *BMC Genomics*, vol. 20, no. 1, pp. 1–12, 2019.