Comparative Analysis of Recursive Feature Elimination and Feature Importance for Dimension Reduction in Cancer Prediction

1st Muhammad Alif Ramadhan Mappunna
School of Computing
Telkom University
Bandung, Indonesia
alifrm@student.telkomuniversity.ac.id

2nd Mahendra Dwifebri Purbolaksono
School of Computing
Telkom University
Bandung, Indonesia
mahendradp@telkomuniversity.ac.id

3rd Adiwijaya
School of Computing
Telkom University
Bandung, Indonesia
adiwijaya@telkomuniversity.ac.id

***Abstract*—With nearly 10 million deaths in 2020, cancer remains one of the most common causes of death worldwide. In research, microarray techniques have been used to diagnose and predict cancer by simultaneously analyzing gene expression. However, microarray data is characterized by high dimensionality and the presence of many irrelevant genes, necessitating effective feature selection methods. This study explores two major methods—Recursive Feature Elimination (RFE) and Feature Importance (FI)—and evaluates their individual and combined performance through five hybrid strategies: score averaging, intersection, union, RFE→FI, and FI→RFE. Implemented within a Support Vector Machine (SVM) framework, these methods are tested on four publicly available microarray datasets: Lung, Colon, Leukemia, and Ovarian. While hybrid selection has shown advantages over single techniques, the impact of method sequencing has been underexplored. This research finds that the RFE→FI strategy consistently balances relevance and generalizability, leading to superior performance across multiple datasets. To improve generalization and reduce complexity, each dataset was limited to a maximum of 50 selected features. Performance was assessed using accuracy, precision, recall, and F1-score. The results reveal that sequential hybrid approaches—especially RFE→FI—enhance classification performance in complex, high-dimensional genomic data. This study contributes a robust and reproducible hybrid feature selection pipeline, offering practical value for microarray-based cancer prediction systems.**

***Keywords*—Cancer, Microarray, Recursive Feature Elimination, Feature Importance, Support Vector Machine**

I. INTRODUCTION

The human body is composed of networks of cells that grow rapidly. Abnormal proliferation of these cells leads to a disease known as cancer. The general term for a group of diseases that can affect any part of the body is cancer [**?**]. The World Health Organization (WHO) reports that cancer is the leading cause of death worldwide, accounting for nearly 10 million deaths in 2020 almost one in every six deaths and that this figure continues to rise [**?**]. The accumulation of risk factors and the declining effectiveness of cellular repair mechanisms with age cause the incidence of cancer to increase significantly [**?**]. In this context, technologies capable of accurately diagnosing and predicting cancer are urgently needed. One such technology currently in use is microarray analysis.

Numerous studies have been conducted to diagnose and predict cancer, and several of these employ microarray techniques. Microarray analysis is capable of capturing thousands of gene expression profiles from multiple cells simultaneously in a single experiment, thereby generating the data needed to predict and classify the genes active in particular tissues. Consequently, this technique is critical for analyzing cancer and predicting clinical treatment outcomes [**?**]. Typically, microarray datasets are characterized by high dimensionality, small sample sizes, and a large number of irrelevant genes [**?**]. Because there are so many genes in a microarray sample, filtering them down to the optimal subset for effective classification is quite challenging. However, various methods exist to address these issues.

In this study, multiple models are compared. In particular, focusing on the comparison between Recursive Feature Elimination (RFE) and Feature Importance using Support Vector Machine (SVM) can provide a robust approach. RFE and Feature Importance are widely used feature-selection techniques in machine learning for reducing data dimensionality while preserving the most informative features for predictive modeling [**?**] [**?**].

Feature selection in machine learning problems is a crucial component for improving model performance, especially when dealing with high-dimensional data. The greater the number of features or variables used, the higher the risk for the model to experience overfitting and face the curse of dimensionality [**?**]. As a result, the model would require large computational resources but produce low accuracy. Feature selection helps eliminate irrelevant features, allowing the model to focus only on important information. This simplifies the model, reduces the risk of overfitting, and avoids the curse of dimensionality. Models trained with important features tend to be more accurate and are better able to generalize to new data [**?**].

Although several studies, such as those by Abdelwahed et al. proposed a hybrid feature selection approach that significantly improved cancer classification performance. Their study demonstrated that combining feature selection techniques can provide more stable and accurate results compared to single-