1. Introduction

Lung cancer is a main causes of cancer death worldwide [1]. The different types of lung cancer are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC accounts for 80% to 85% of all lung cancers, while SCLC accounts for 15% to 20%. Smoking is known as a major factor in lung cancer. Only 7% of female patients with lung cancer in Taiwan had a smoking history [2]. Another factor that causes lung cancer in non-smoker women is environmental exposure or heredity. However, there are no molecular mechanisms of NSCLC in women who do not smoke, although several genes, including EML4-ALK, EGFR, TP53.9, and PIK3CA, are linked to lung cancer in smokers [3].

In the case of lung cancer, a chest X-ray and a Computerized Tomography Scan (CT-Scan) are usually used for the diagnosis and prognosis. However, this method can only detect malignant cells in lung cancers in their advanced stages [4][5]. With advances in technology in the field of molecular biology, especially in microarray technology, information about DNA, RNA, and protein will be obtained for early detection of tumor formation [6]. One application of microarray technology is to analyze thousands of DNA, RNA, and protein samples at the same time.

Currently, machine learning on microarrays and bioinformatics analysis are frequently used to identify potential cancer biomarkers, particularly genes related to the prognosis of lung cancer [2]. The use of microarrays produces more complete information about variations in cancer molecules as well as obtains more accurate classification results [7]. In the past decade, machine learning methods have been widely used to medical data analysis [8]. Cancer prognosis is commonly predicted using machine learning techniques, especially Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Bayesian networks [6]. Also microarray data has been implemented to identify a various cases used ensemble method [9].

Identification of lung cancer with microarrays using machine learning methods has been done several times in previous studies. In 2017, Nurul et al. improved cancer classification using the Multi Support Vector Machines (MSVM) method and the Recursive Feature Elimination (RFE) method as feature selection with an accuracy value of 98.6% [10]. In 2017, Wu and Zhao conducted a study that detected SCLC using a novel neural network algorithm with the entropy degradation method (EDM) and the accuracy value reached 77.8% [4]. In 2019, Shrikant and Pawar conducted a web-based study to classify cancer types from microarray of Gene Expression Data (DEGs) using MAS 5.0 and Robust Multi-Array (RMA) as feature selection and SVM method for classification with an accuracy value of 95% [11]. Then in 2019, José et al. conducted research on predicting radiation pneumonitis in advanced stages II-III in NSCLC using machine learning. Using machine learning algorithms, carried out an analysis of contributing factors in the development of radiation pneumonitis to uncover previously unidentified criteria. In multivariate analysis, Random Forest has an accuracy value of 66% [12].

Several other studies have also been carried out previously. In 2020, Yu classified the types of NSCLC using Convolutional Neural Networks with a validated prediction with an accuracy value was 86.4% [13]. In 2020, Pankaj et al. conducted research the hybrid algorithm for the classification of lung cancer using SVM and Neural Network by detecting CT-Scan images with an accuracy value of 98.08% [14]. Reinel and Co conducted research in 2020 comparing machine learning with deep learning to classify cancer types based on gene expression microarray data. The accuracy of the identification result is 90.6% while using Logistic Regression and 94.43% while using Convolutional Neural Network (CNN) [15]. In 2021, Sunil et al. compared the performance of lung cancer classification using the Swarm Intelligence technique with the best results shown when the test was classified with the Decision Tree classifier for one hundred genes, and the highest classification accuracy value was 99.1% [16]. To the best of our knowledge, the studies of NSCLC identification for non-smoker woman is very rare.

In this study, we aimed to predict the potential occurrence of NSCLC in non-smoker women using the GA-SVM methods. This research used a Genetic Algorithm (GA) for feature selection because the evolution operators in a Genetic Algorithm (GA) make this algorithm very effective in global search and this algorithm has high flexibility to be hybridized with other search methods to make it more effective. Also GA has been implemented to develop prediction model on a various cases [17][18]. For the development of the prediction model using the Support Vector Machines (SVM) method because Support Vector Machines (SVM) offer high accuracy results and work well with high dimensional spaces. The Support Vector Machines (SVM) classifier uses a subset of train points so that the result uses very little memory. Several research of SVM has also been done implemented to identify various cases [19][20].