I. INTRODUCTION

Cancer is a symptom of abnormal cell growth caused by uncontrolled changes in gene expression. Cancer can cause significant morbidity and can cause death if not treated quickly [1]. One of the cancers that occur in the organs of the body is the lung which can also be called lung cancer.

In 2020, WHO stated that lung cancer is cancer with the highest number of deaths reaching 1.8 million people, even reaching twice the number of deaths second-highest cancer, i.e. colon and rectal cancer which reached 935 thousand people [2]. Symptoms commonly experienced include relentless coughing, hoarseness, constant chest pain, frequent lung infections such as bronchitis, pneumonia, and coughing up blood [3].

The most common cause of lung cancer is smoking. Cigarettes cause 90% of lung cancer cases. Tobacco in cigarettes contains many chemicals that cause lung cancer. Even former smokers still have a risk of developing lung cancer, although the risk is smaller. Passive smokers also have the same risk [4]. Lung cancer can be prevented by 30% to 50% by doing early detection of cancer and doing the right treatment [2]. Cancer has a high chance of being cured if it can be diagnosed early. Early diagnosis can be done if there are symptoms of lung cancer.

One of the efforts in the early detection of lung cancer is screening. The types of screening used are low-dose CT scan (LDCT) and chest X-Ray. The drawback of this physical technology is that it can only detect malignant cancer cells in late-stage cancer [5]. Furthermore, screening is only effective for some cancers, and cancer cells are generally much more complex, requiring resources as well as specialized equipment and medical personnel. Screening methods also need to be pre-differentiated based on age and symptoms so as not to generate too many false positives [2].

With the advancement of DNA microarray technology, the gene expression level of thousands of genes or cells in a particular tissue can be measured. Microarray technology can search systematically and can perform analysis for the classification of cancer. Then it predicts results in the form of various tumors or cancers. Thus, the microarray is a very important tool for studying the transcriptome or transcription process in cancer cells [6].

Several studies have used machine learning methods to classify or identify and detect lung cancer. Most of them used gene expression data either by using the GSE4115 dataset from GEO or another dataset. Also, some of them used the Ensemble Methods model specifically. The GSE4115 dataset was used because it was specific that all samples were smokers. Some studies used the GSE4115 dataset with various methods. In 2019, Wu, et al. used Sparse Logistic Regression with L1/2 Regularization with 0.83 accuracy [7]. In 2020, Yin and Chen used Deep Forest and Semi-Supervised with Self-Training (DSST) with 0.73 accuracy [8]. In 2020, Wang, et al. used Random Forest with Self-Paced Learning (RFSPL) with 0.82 accuracy [9].

For other dataset, in 2017 Wu and Zhao used a neural-network-based algorithm, i.e. the Entropy Degradation Method (EDM) with an accuracy of 0.77 [10]. In 2019 Nasser and Abu-Naser used an Artificial Neural Network (ANN) with an accuracy of 0.96 [11]. In 2016, Podolsky, et al. used K-Nearest Neighbor (KNN) K=5 with AUC and MCC values of 0.96 and 0.77 respectively [12]. In 2019 Pati used Multilayer Perceptron (MLP) with 0.86 accuracy [13]. For Ensemble Methods specifically, in 2018 Faisal, et al. used the Majority Voting Ensemble Methods on three methods, i.e. Multilayer Perceptron (MLP), Gradient Boosted Tree (GBT) and Support Vector Machine (SVM) with 0.88 accuracy [14]. In 2017, Wang, et al. used the Random Forest and AdaBoost Ensemble Methods with AUC values of 0.916 and 0.914 respectively [15]. Unfortunately, the study of lung cancer detection for the case for smoker person is still very rare. Also, the study with GSE4115 dataset only had a small variant of Ensemble Methods and this method is not often used in this dataset.

Hence, in this study, we aimed to build a prediction model for the identification of lung cancer in smokers using the Ensemble Methods based on gene expression data (microarray). We choose Ensemble Methods for this study because it is a machine learning technique that combines several other learning algorithms to solve the same problem. This method was chosen because it has good performance and is more profound in making predictions than using individual models [16]. Ensemble Methods used in this study are Random Forest and AdaBoost. Furthermore, we also aimed to get the optimal feature selection for classification and get the accuracy of the Ensemble Methods in making predictions.