A Prediction Model for Lung Cancer Levels Based on Machine Learning

Huu-Huy Ngo, and Hung Linh Le

***Abstract*—Among cancers, lung cancer is one of the most dreaded conditions, and it is the leading cause of cancer-related deaths worldwide. Early cancer identification and prediction help prevent and treat cancer efficiently, especially the beginning cancer stage. Therefore, this study presents a prediction model for lung cancer level based on machine learning. Machine learning algorithms are applied as primary methods. Firstly, the dataset collection is implemented; then, feature selection algorithms are used to identify essential features. Secondly, the proposed model applies the machine learning algorithms on two datasets (The full dataset and the dataset of essential features). Finally, experimental results demonstrate that this proposed system has an excellent performance, with 100% and 98.7% accuracy on the full dataset and the dataset of the top three essential features, respectively.**

***Keywords*—Classifier model, Feature selection, k-nearest neighbor (kNN), Lung cancer, Machine learning.**

I. Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide [1]. In common cancers worldwide, lung and breast cancers are two of the most dreaded conditions, and each of these cancers contributes 12.3% of the total number of new cases diagnosed in 2018 [2]. Figure 1 shows the global cancer incidence in men and women. Given this increasing global responsibility, one of the most significant challenges in public healthcare is cancer prevention. To prevent cancer, we can work on several ways to archive living healthily, such as changing dietary patterns, reducing alcohol consumption, increasing physical activity, and maintaining a healthy body weight [2].

Accurate risk prediction models for lung cancer provide efficiency to the classification of high-risk people. Lung cancer risk prediction is an attractive research topic of numerous studies. Several risk factors significantly affect lung cancer risk prediction, and these factors are used as predictors of lung cancer risk. The most risk factors for lung cancer are smoking, family history of lung cancer, and respiratory diseases [1, 3].

Given the rapid development of computer science and the explosion of information on the internet, this has provided excellent opportunities for numerous fields, such as production, research and development (R&D), communications, and services. In this process, artificial intelligence (AI) is being actively studied, and machine learning has been widely used to produce useful results efficiently, especially in medicine [4]. The growth in medical data collection and data science offers a new opportunity to improve patient prediction, prevention, and treatment. In the smart healthcare system, computer technologies are applied to improve quality services, and machine learning is becoming an essential solution to support patients' diagnoses. Machine learning algorithms can learn from the provided data, and the accuracy of the model is improved with subsequent training. Machine learning is a useful analytical tool, and it is used to resolve challenging tasks, such as transforming medical records into knowledge, pandemic predictions, and genomic data analysis [5].

Therefore, this study proposes a prediction model for lung cancer levels using machine learning. This model supports early prediction and treatment of lung cancer based on risk factors. First, the dataset collection and pre-processing are implemented; then, the author evaluates feature importance by using feature selection algorithms. Second, the proposed model implements the machine learning algorithms on two datasets (The full dataset and the dataset of essential

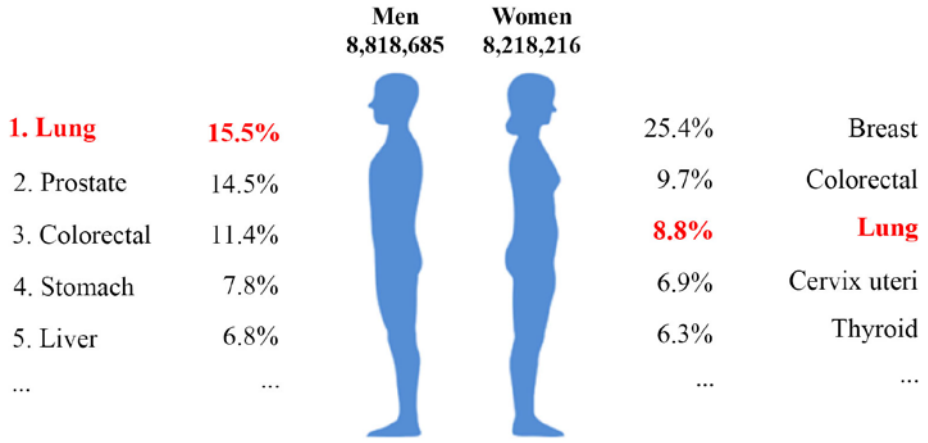

Figure 1. Global cancer incidence in 2018

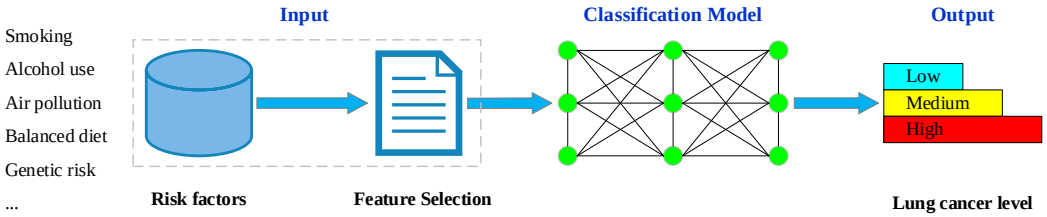

Figure 2. System overview

features). Finally, the output of the model is predicted lung cancer levels, including high, medium, and low levels.

The remainder of this paper is organized as follows. Section II reviews relevant studies. The architecture of the proposed system is described in Section III. Section IV demonstrates the experimental results. Finally, Section V summarizes conclusions and future research directions.

II. Related Work

Tammemagi et al. [6] proposed two incidence estimation models for lung cancer. These models used data from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO). Specifically, the first model was used for the general population that includes never-smokers and ever-smokers; meanwhile, the second model was only applied for ever-smokers. Furthermore, these two models combined several other risk factors for lung cancer, such as age, family history of lung cancer, smoking status, and smoking duration. Besides, the second model also included smoking quit-time.

Katki et al. [7] presented two risk models for lung cancer incidence and mortality by using data on ever-smokers from the PLCO control group. They developed Cox hazard ratio models with non-parametric baseline hazards, such as age, sex, race, education, smoking intensity, body mass index, and family history of lung cancer. These models were employed on US ever-smokers aged 50 to 80 years to evaluate the results of risk-based selection for computed tomography (CT) lung screening.

Krishnaiah et al. [8] developed a system based on data mining classification techniques to predict lung cancer. This system was used to extract hidden information from a historical lung cancer disease database. Several classification techniques were used, such as decision trees, Naïve Bayes, and artificial neural networks. The model utilized various lung cancer symptoms, such as age, sex, pain in the chest, shortness of breath, and wheezing, to predict lung cancer patients.

Spitz et al. [9] proposed a risk model to predict lung cancer for never, former, and current smokers. Specifically, this model considers two variables for never smokers, including environmental tobacco smoke and family history of cancer. Meanwhile, this model analyzes various variables for former and current smokers, such as dust exposure, prior respiratory disease, and smoking history. Besides, this model for former smokers also studies two different variables, including the family history of cancer and the age of smoking cessation.

Bach et al. [10] presented a prediction model for smokers to evaluate lung cancer risk. They used data from Carotene and Retinol Efficacy Trial (CARET), a large study for lung cancer prevention. This model used different enrolled person information as input variables, including age, sex, asbestos exposure history, and smoking history.

III. The Proposed Prediction Model

Figure 2 presents an overview of the proposed system. This system contains three primary components, including input, classifier, and output of the model. First, the dataset collection and pre-processing are implemented; then, feature selection algorithms are used to identify the essential features. Second, the proposed model implements the machine learning algorithms on two datasets (The full dataset and the dataset of essential features). Finally, the output of the model includes predicted cancer levels, such as high, medium, and low levels.

A. Input Dataset

1. Dataset Collection

The input of the model is risk factors of lung cancer, such as smoking, passive smoker, alcohol use, air pollution, a balanced diet. The dataset collection has a significant role, and it affects the quality of the model. There are some different resources to gather the lung cancer dataset, such as Data World [11], Kaggle [12], and UCI Machine Learning Repository [13].

In this study, the lung cancer dataset of Data World is used to train and test the proposed model. This dataset covers 1,000 medical records of patients. Except for patient ID and target feature, each record contains 23 risk factors (features) of lung cancer, including age, air pollution, alcohol use, balanced diet, chest pain, chronic lung disease, clubbing of fingernails, coughing of blood, dry cough, dust allergy, fatigue, frequent cold, gender, genetic risk, obesity, occupational hazards, passive smoker, shortness of breath,

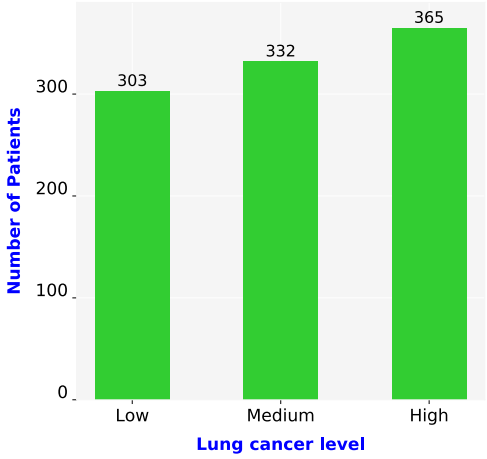

Figure 3. Distributions of the target feature

smoking, snoring, swallowing difficulty, weight loss, and wheezing. The target feature is the level feature, and it includes three classes: low, medium, and high. Figure 3 shows the distributions of the target feature in the collected dataset.

2. Feature Selection

In the dataset, the features have different roles and importance. They directly affect to quality and performance of the model. Therefore, feature selection is a necessary task by evaluating feature importance. Feature selection is a process where essential features are automatically selected in the dataset to improve model performance [14].

Feature selection will have several benefits as follows. Firstly, feature selection reduces overfitting. Due to feature selection is implemented, the model can identify redundant features and remove them. Thus, the model decisions based on noise are reduced. Secondly, feature selection improves accuracy. After the model removes redundant features, the misleading feature is decreased. Therefore, model accuracy is improved. Finally, feature selection reduces training time. Owing to lessening redundant features, the training dataset is reduced. Hence, the training model is implemented faster. There are different approaches to select features from the dataset as follows:

- **Univariate feature selection:** The univariate feature selection method uses univariate statistical tests to select the best features. These features are ranked based on different statistical scoring functions. Comparing each feature to the target variable is implemented to identify any statistically significant relationship between them [15]. This method is also called analysis of variance (ANOVA). In evaluating the relationship between one feature and the target variable, other features are ignored. In other words, this method does not consider the dependencies between the features, and features are independent with each other in the dataset. In the univariate feature selection, the Chi-square test and F-test are popular scoring functions, and they are usually used to calculate the test score of the features.
- **Recursive feature elimination (RFE):** RFE is a feature selection method, which works efficiently on small datasets [16, 17]. In various versions of this method, Guyon et al. [18] proposed a support vector machine recursive feature elimination (SVM-RFE) algorithm to identify eliminated features. This is one of the most popular feature selection algorithms. Originally, SVM-RFE was used to classify cancers, and it is applied to solve binary problems. However, this method can easily be extended to deal with multiclass problems. RFE improves performance by removing the least important features, and it identifies features that contribute the most to predicting the target feature. RFE tends to delete redundant and weak features; in contrast, it maintains independent features [16, 19].
- **Principle component analysis (PCA):** PCA is a popular transform method, and it has been extensively explored for various applications, such as computer vision, machine learning, and data mining. In the feature selection, Song et al. [20] applied PCA to select several essential features from all the feature components from the viewpoint of numerical analysis. Each feature has different effects on feature extraction results and exploits the eigenvectors of PCA's covariance matrix. Then, these eigenvectors are used for feature selection.

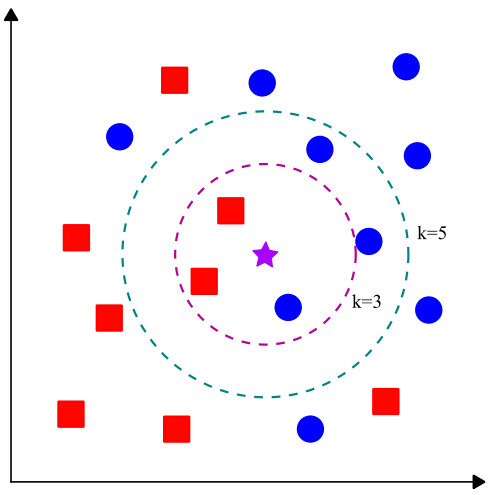

Figure 4. The kNN classification

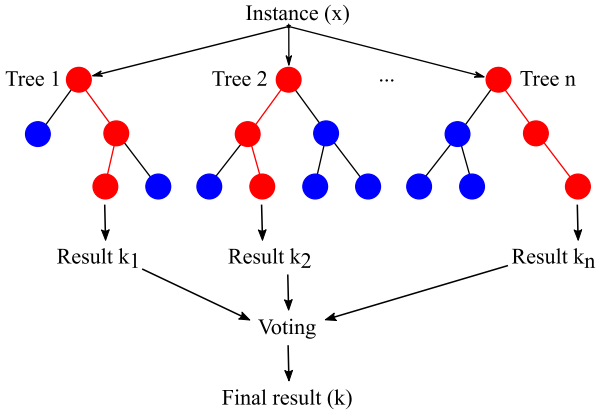

Figure 5. Random forest classification

- **Choosing important features:** A supervised trained classifier is used to evaluate the feature importance and select features. Extra trees, called extremely randomized trees [21], are a popular classifier to estimate the importance of features. During a classifier is trained, each feature is evaluated to create splits, and this measure is used as a feature selector.

B. Input Dataset

This part is the primary model and the most important of the proposed system. Several classification algorithms in machine learning are applied to design the system, such as kNN, random forest, logistic regression, Naïve Bayes, decision tree. Then, the results of these methods are compared, and the most efficient one is chosen to develop the system.

1. k-Nearest Neighbor

The kNN classifier is a conventional non-parametric classifier, and this classification is one of the most basic and straightforward classification methods. In the feature space, each object, called a point, is denoted by a feature vector. The k-NN method classifies a new object by calculating the distances between this new point and different neighbor points. This object will then be allocated to the class most common among its k nearest neighbors, where k is an integer [22, 23]. Figure 4 illustrates the kNN classification method, where the star item denotes a new object. If k equals three, this new point is assigned to the red square class. If k equals five, it is allocated to the blue circle class that is the majority class of the five nearest points.

In kNN classification, numerous distance functions have been utilized to calculate the distance between two points in a feature space, in which the Euclidean distance function is the most popularly used one. Let A and B are represented by feature vectors A = (a_1, a_2, …, a_n) and B = (b_1, b_2, …, b_n), where n is the dimensionality of the feature space. The Euclidean distance between A and B is calculated as in Equation (1). In this study, the kNN classification is established with k equals five; the metric is Euclidean, and the weight is distance.

$$dist(A,B)=\sqrt{\sum_{i=1}^{n}\left(b_i-a_i\right)^2} \tag{1}$$

2. Random Forest

Random forest is an ensemble classifier that uses recursive partitioning to create various trees and then combines the results from different trees [24]. The trees are generated using the Bootstrap sampling method that randomly samples the cases with the database's replacement. Owing to using the Bootstrap sampling method, this approach supports better estimating the original database's distribution and improves accuracy [25]. Figure 5 demonstrates the random forest classification method. The final classification result is the majority voting result from a large number of trees. In this study, the random forest classification is established as follows. The number of generated trees is 10, and the minimum split subsets equal five.

3. Logistic Regression

Logistic regression is one of the primary and popular machine learning algorithms to handle classification problems. The logistic regression model is useful for exploring the relationship between a dichotomous dependent variable and a set of independent explanatory variables [26, 27]. The logistic regression model has the form as in Equation (2). Where y is the dependent variable; x_i is the ith explanatory variable; β_0 is a constant, and β_i is the i^{th} regression coefficient associated with the explanatory variable xi. The probability (P) of the occurrence of y is defined as in Equation (3).

$$logit(y)=\beta_0+\sum_{i=1}^{n}\beta_i x_i \tag{2}$$

$$P=\frac{1}{1+e^{-logit(y)}}=\frac{1}{1+e^{-\left(\beta_0+\sum_{i=1}^{n}\beta_i x_i\right)}} \tag{3}$$

4. Naïve Bayes

A Naive Bayes classifier is a probabilistic machine learning model that is used for classification tasks. This classifier is a supervised multiclass classification algorithm, and it is developed based on applying Bayes' theorem with the "naïve" assumption of conditional independence between every pair of variables [28]. According to Bayes theorem, given a set of independent variables, $X=\{x_1, x_2, \ldots, x_n\}$, the posterior probability is calculated for each possible class, $C=\{c_1, c_2, \ldots, c_m\}$. Naive Bayes classifier computes posterior probability $P(C|X)$ from $P(C)$, $P(X)$, and $P(X|C)$ as in Equation (4). Where $P(C)$ is the prior probability of each class; $P(X)$ is the prior probability of predictor, and $P(X|C)$ is the likelihood.

$$P(C\mid X)=\frac{P(X\mid C)P(C)}{P(X)} \tag{4}$$

5. Decision Tree

The decision tree is a classification method that classifies the labeled trained data into a tree or rules. The decision tree is constructed by recursively partitioning the training data. Figure 6 demonstrates a decision tree classification, in which the decision and leaf nodes represent the attribute name and class label, respectively. Each path of the decision tree describes the decision rule [29, 30].

In this study, the decision tree classification is established as follows. The minimum number of instances in leaves equals two, and the minimum split subsets are five. The maximal depth of the classification tree is 100, and the splitting of the nodes stops when the majority reaches 95%.

IV. Experimental Results

A. Analysis of Feature Selection

The choosing important features method is used for feature selection in our dataset. Extra trees are used to estimate the importance of features. We get the result of feature importance, as shown in Figure 7. The obesity feature has the highest importance value of 0.0992. Next, the coughing of blood and passive smoker features have an importance value corresponding to 0.0783 and 0.0653. The top 5 essential features include obesity, coughing of blood, passive smoker, fatigue, and wheezing. These features are used to evaluate the proposed system. By contrast, the three least essential features cover patient ID, gender, and age. Especially, the patient ID feature has an importance value of 0 because this feature is only used to identify a record of the patient, and it does not affect the classification results.

B. The Model Evaluation

1. The Model Evaluation on the Full Dataset

We compared and evaluated the proposed model using other classification algorithms: kNN, random forest, logistic regression, Naïve Bayes, and decision tree. The input data includes 24 features (full dataset). The classification results are demonstrated in Figs. 8 and 9. As presented in Figure 8, we observe that three methods have an accuracy value of 100%, which is the maximum value, such as kNN, random forest, and decision tree. By contrast, the lowest value of the accuracy is 97.7%, the Naïve Bayes method; however, it is still outstanding. Figure 9 illustrates the confusion matrix of five classifiers on the full dataset. The medium class of the target feature has the highest false negative rate (FNR). FNR of logistic regression and Naïve Bayes methods are 4.8% and 4.6%, respectively.

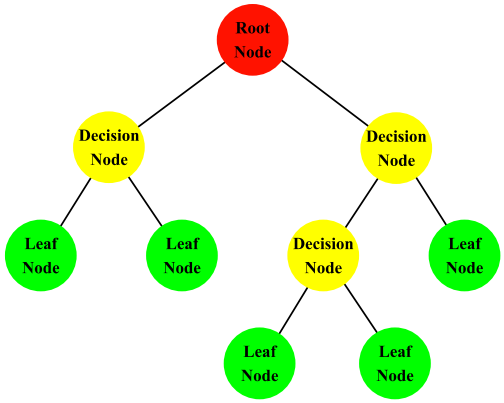

Figure 6. Decision tree classification

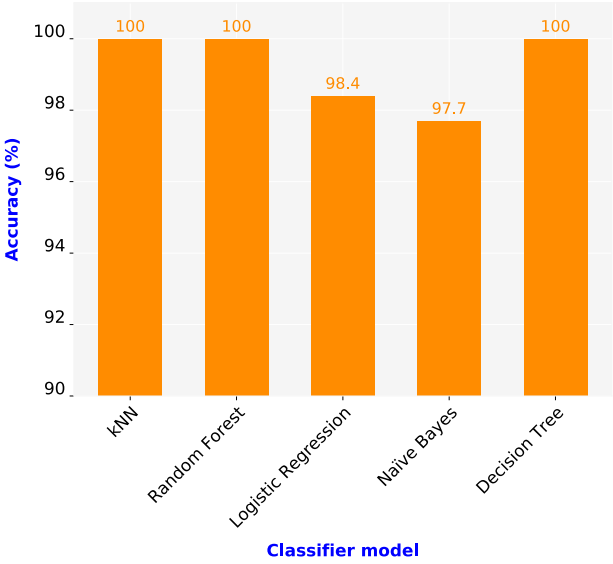

Figure 8. The accuracy of the models on the full dataset

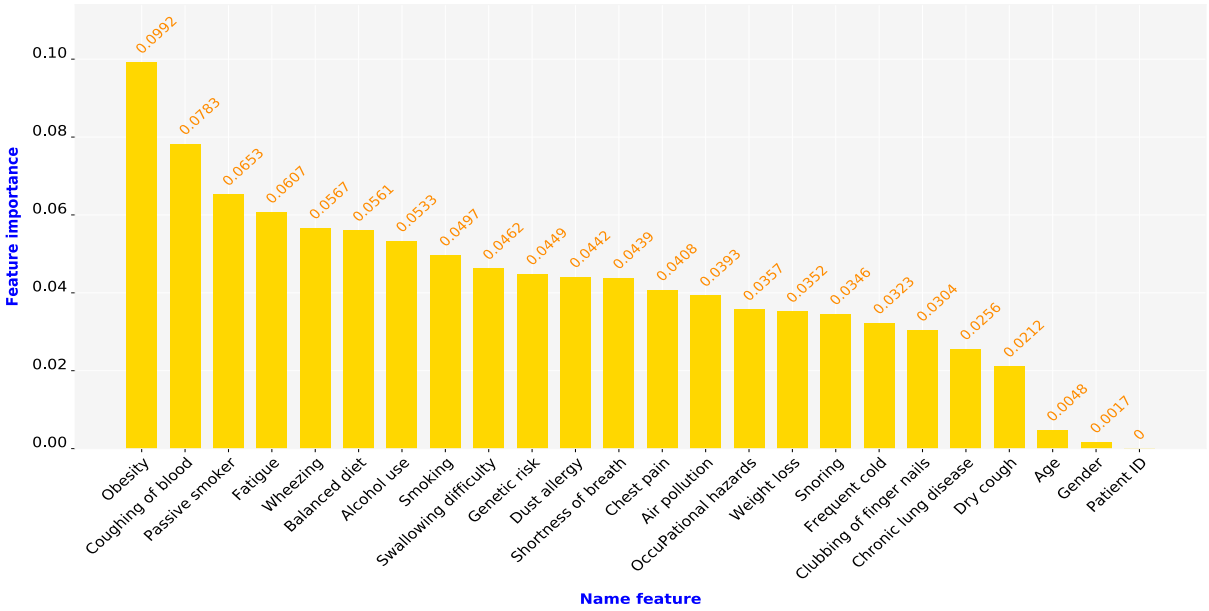

Figure 7. The result of feature importance

2. The Model Evaluation on the Dataset of the Top Five Essential Features

The features have different roles and importance. After identifying the importance of features, the proposed model is implemented on the dataset of the top five essential features, including obesity, coughing of blood, passive smoker, fatigue, and wheezing. Figure 10 shows the classification results of five different methods on the dataset of the top five essential features. The accuracy value of methods remained consistently very high, especially kNN, random forest, and decision tree methods with an accuracy value of 100%. Meanwhile, the logistic regression method has the lowest accuracy value of 92.2%. Figure 11 demonstrates the confusion matrix of five classifiers on the dataset of the top five essential features. FNR of the medium class is still highest in classes of the target feature. In other words, the medium class is usually misclassified into other classes of the target feature.

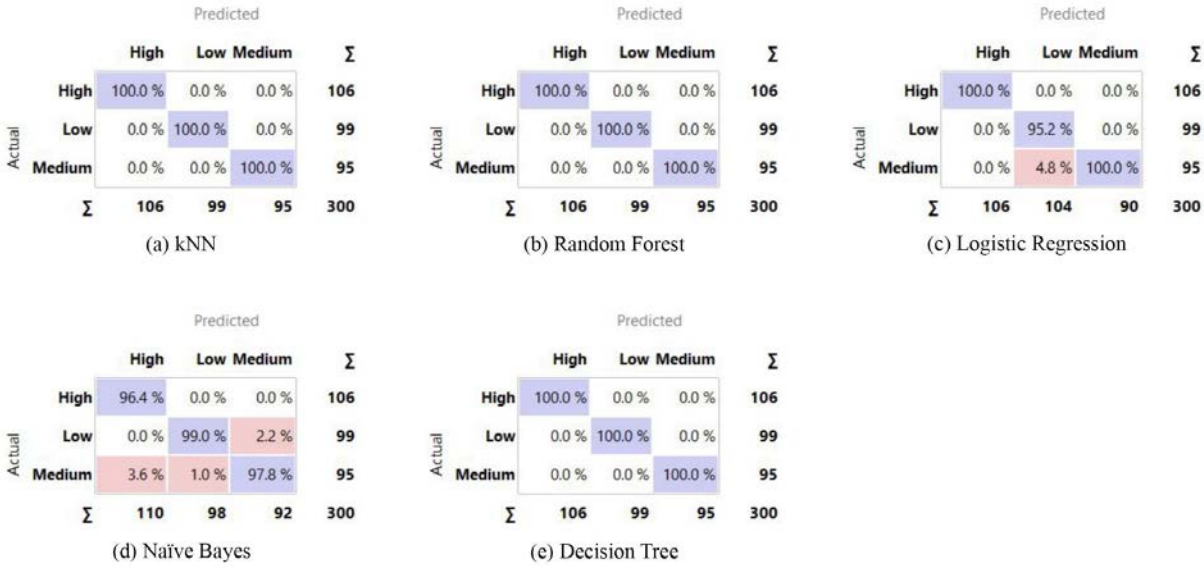

Figure 9. The confusion matrix of five classifiers on the full dataset

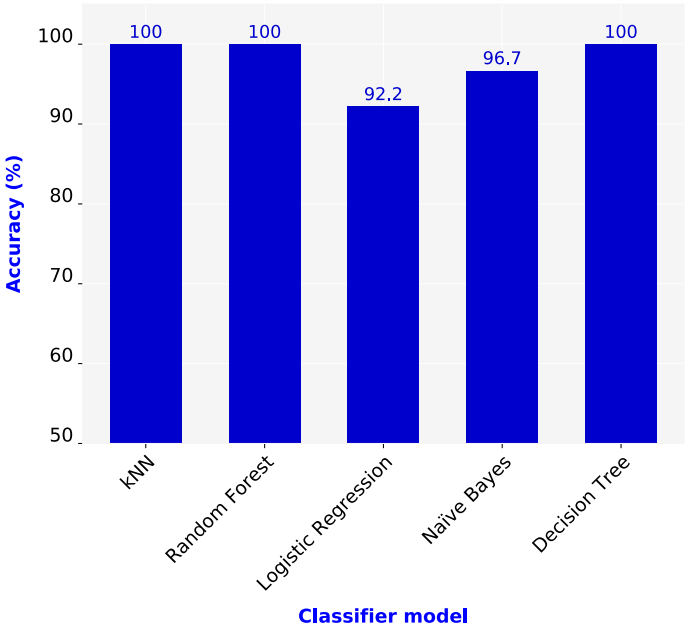

Figure 10. The accuracy of the models on the dataset of the top five essential features

Figure 11. The confusion matrix of five classifiers on the dataset of the top five essential features

3. The Model Evaluation on the Dataset of the Top Three Essential Features

In this section, we consider the performance of the proposed model on the narrow dataset that only includes the three most essential features: obesity, coughing of blood, and passive smoker. The classification results on the dataset of the top three essential features are illustrated in Figure 12. The kNN, random forest, and decision tree methods remained consistently very high accuracy values of 98.7%. Next, the accuracy of Naïve Bayes and logistic regression methods are 91.0% and 75.9%, respectively. These results demonstrated that the proposed model has a good performance by selected a suitable classified method. Figure 13 shows the confusion matrix of five classifiers on the dataset of the top three essential features. We observe that the medium class of the target feature is regularly misclassified, and the false positive of this class occurred in five classification methods.

V. CONCLUSIONS

This study examined a prediction model for lung cancer level based on machine learning. This model early predicts lung cancer level from its risk factors to help prevent and treat it efficiently. Machine learning algorithms are applied as primary methods, such as kNN, random forest, logistic regression, Naïve Bayes, and decision tree. Besides, feature selection algorithms are used to identify essential features. The results show that the proposed model has an excellent performance using three classification methods: kNN, random forest, and decision tree. These methods have an accuracy value of 100% and 98.7% on the full dataset and the dataset of the top three essential features, respectively. In the future, we will improve this model by collecting more datasets. In addition, we also intend to extend system implementation in medical facilities specializing in lung cancer.

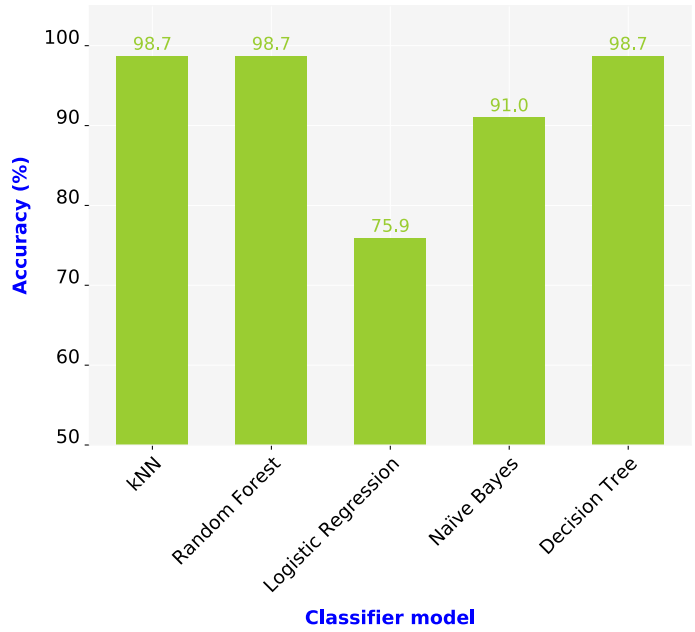

Figure 12. The accuracy of the models on the dataset of the top three essential features

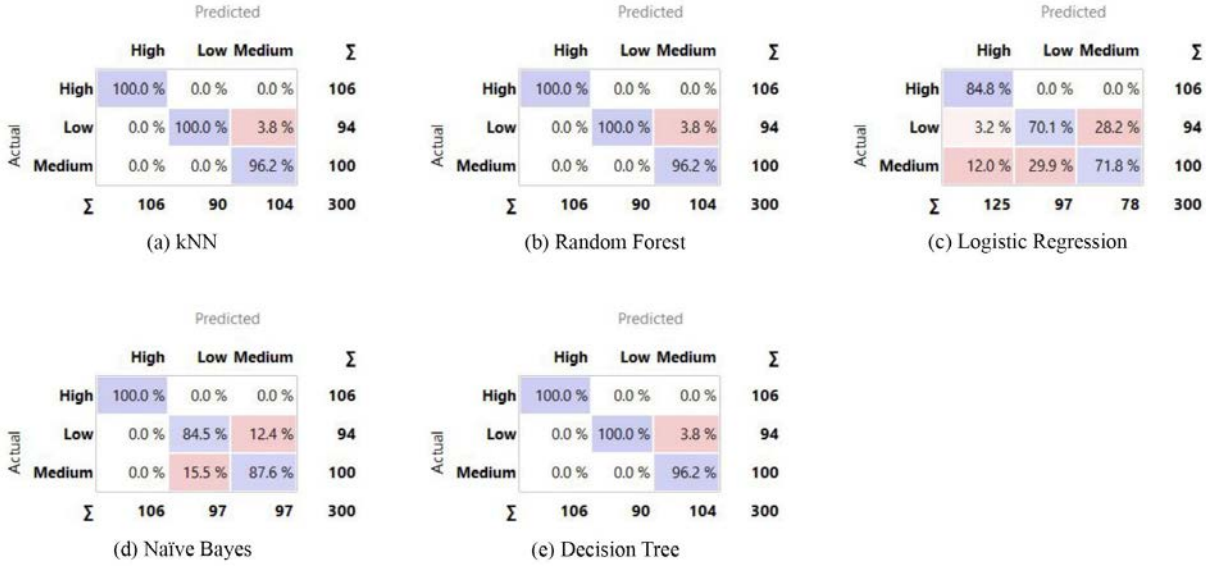

Figure 13. The confusion matrix of five classifiers on the dataset of the top three essential features

REFERENCES

[1] I. Toumazis, M. Bastani, S. S. Han and S. K. Plevritis, "Risk-Based Lung Cancer Screening: A Systematic Review," *Lung Cancer*, vol. 147, pp. 154–186, Sep. 2020.

[2] Worldwide Cancer Data, World Cancer Research Fund International, 2018, *https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data*, [09-Aug-2020].

[3] Z. Lyu, N. Li, S. Chen, G. Wang, F. Tan, X. Feng, X. Li, Y. Wen, Z. Yang, Y. Wang, J. Li, H. Chen, C. Lin, J. Ren, J. Shi, *et al.*, "Risk Prediction Model for Lung Cancer Incorporating Metabolic Markers: Development and Internal Validation in a Chinese Population," *Cancer Medicine*, vol. 9, no. 11, pp. 3983–3994, 2020.

[4] H. Liu, "Feature Selection," in *Encyclopedia of Machine Learning*, Boston, MA, USA: Springer, pp. 402–406, 2010.

[5] A. K. Gárate-Escamila, A. Hajjam El Hassani and E. Andrès, "Classification Models for Heart Disease Prediction Using Feature Selection and PCA," *Informatics in Medicine Unlocked*, vol. 19, pp. 1–11, Jan. 2020.

[6] C. M. Tammemagi, P. F. Pinsky, N. E. Caporaso, P. A. Kvale, W. G. Hocking, T. R. Church, T. L. Riley, J. Commins, M. M. Oken, C. D. Berg and P. C. Prorok, "Lung Cancer Risk Prediction: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Models and Validation," *JNCI: Journal of the National Cancer Institute*, vol. 103, no. 13, pp. 1058–1068, Jul. 2011.

[7] H. A. Katki, S. A. Kovalchik, C. D. Berg, L. C. Cheung and A. K. Chaturvedi, "Development and Validation of Risk Models to Select Ever-Smokers for CT Lung Cancer Screening," *JAMA*, vol. 315, no. 21, pp. 2300–2311, Jun. 2016.

[8] V. Krishnaiah, D. G. Narsimha and D. N. S. Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques," *International Journal of Computer Science and Information Technologies*, vol. 4, no. 1, pp. 39–45, 2013.

[9] M. R. Spitz, W. K. Hong, C. I. Amos, X. Wu, M. B. Schabath, Q. Dong, S. Shete and C. J. Etzel, "A Risk Model for Prediction of Lung Cancer," *JNCI: Journal of the National Cancer Institute*, vol. 99, no. 9, pp. 715–726, May 2007.

[10] P. B. Bach, M. W. Kattan, M. D. Thornquist, M. G. Kris, R. C. Tate, M. J. Barnett, L. J. Hsieh and C. B. Begg, "Variations in Lung Cancer Risk Among Smokers," *JNCI: Journal of the National Cancer Institute*, vol. 95, no. 6, pp. 470–478, Mar. 2003.

[11] Lung Cancer Data, Data World, 2017, *https://data.world/cancerdatahp/lung-cancer-data*, [15-Sep-2020].

[12] Lung Cancer Dataset, Kaggle, 2018, *https://www.kaggle.com/yusufdede/lung-cancer-dataset*, [15-Sep-2020].

[13] Lung Cancer Dataset, UCI Machine Learning Repository, 1992, *https://archive.ics.uci.edu/ml/datasets/Lung+Cancer*, [15-Sep-2020].
[14] J. Brownlee, "Feature Selection in Python with Scikit-Learn," Machine Learning Mastery, 2014.
[15] Md. R. H. Subho, Md. R. Chowdhury, D. Chaki, S. Islam and Md. M. Rahman, "A Univariate Feature Selection Approach for Finding Key Factors of Restaurant Business," in Proceedings of *IEEE Region 10 Symposium*, Kolkata, India, pp. 605–610, Jun. 2019.
[16] X. Chen and J. C. Jeong, "Enhanced Recursive Feature Elimination," in Proceedings of *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, Cincinnati, Ohio, USA, pp. 429–435, Dec. 2007.
[17] P. M. Granitto, C. Furlanello, F. Biasioli and F. Gasperi, "Recursive Feature Elimination with Random Forest for PTR-MS Analysis of Agroindustrial Products," *Chemometrics and Intelligent Laboratory Systems*, vol. 83, no. 2, pp. 83–90, Sep. 2006.
[18] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, Jan. 2002.
[19] K. Yan and D. Zhang, "Feature Selection and Analysis on Correlated Gas Sensor Data With Recursive Feature Elimination," *Sensors and Actuators B: Chemical*, vol. 212, pp. 353–363, Jun. 2015.
[20] F. Song, Z. Guo and D. Mei, "Feature Selection Using Principal Component Analysis," in Proceedings of *International Conference on System Science, Engineering Design and Manufacturing Informatization*, Yichang, China, pp. 27–30, Nov. 2010.
[21] P. Geurts, D. Ernst and L. Wehenkel, "Extremely Randomized Trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, Apr. 2006.
[22] C.-R. Dow, W.-K. Wang, H.-H. Ngo and S.-F. Hwang, "An Advising System for Parking Using Canny and k-NN Techniques," *Computer Science & Information Technology (CS & IT)*, vol. 9, no. 6, pp. 27–34, May 2019.
[23] L.-Y. Hu, M.-W. Huang, S.-W. Ke and C.-F. Tsai, "The Distance Function Effect on K-Nearest Neighbor Classification for Medical Datasets," *SpringerPlus*, vol. 5, no. 1, pp. 1–9, Aug. 2016.
[24] K. Moorthy and M. S. Mohamad, "Random Forest for Gene Selection and Microarray Data Classification," in Proceedings of *Knowledge Technology*, Kajang, Malaysia, pp. 174–183, Jul. 2011.
[25] C. Liu, F. Tang and C. Leth Bak, "An Accurate Online Dynamic Security Assessment Scheme Based on Random Forest," *Energies*, vol. 11, no. 7, pp. 1–17, Jul. 2018.
[26] M. E. A. Budimir, P. M. Atkinson and H. G. Lewis, "A Systematic Review of Landslide Probability Mapping Using Logistic Regression," *Landslides*, vol. 12, no. 3, pp. 419–436, Jun. 2015.
[27] T. K. Hembram, G. C. Paul and S. Saha, "Spatial Prediction of Susceptibility to Gully Erosion in Jainti River Basin, Eastern India: A Comparison of Information Value and Logistic Regression Models," *Modeling Earth Systems and Environment*, vol. 5, no. 2, pp. 689–708, Jun. 2019.
[28] P. Valdiviezo-Diaz, F. Ortega, E. Cobos and R. Lara-Cabrera, "A Collaborative Filtering Approach Based on Naïve Bayes Classifier," *IEEE Access*, vol. 7, pp. 108581–108592, Aug. 2019.
[29] D. Lavanya and K. U. Rani, "Ensemble Decision Tree Classifier for Breast Cancer Data," *International Journal of Information Technology Convergence and Services*, vol. 2, no. 1, pp. 17–24, Feb. 2012.
[30] B. Thangaparvathi, D. Anandhavalli and S. M. Shalinie, "A High Speed Decision Tree Classifier Algorithm for Huge Dataset," in Proceedings of *International Conference on Recent Trends in Information Technology (ICRTIT)*, Chennai, Tamil Nadu, India, pp. 695–700, Jun. 2011.

Huu-Huy Ngo received his B.S. and M.S. degrees from Thai Nguyen University of Information and Communication Technology, Vietnam, in 2010 and 2012, respectively, and Ph.D. degrees in Information Engineering and Computer Science from Feng Chia University, Taiwan, in 2021. Currently, he is a lecturer at the Thai Nguyen University of Information and Communication Technology, Vietnam. His research interests include computer vision, deep learning, embedded system, neural networks, and object detection. Email: nhhuy@ictu.edu.vn.

Hung Linh Le received his B.S. and M.S. degrees from VNU University of Engineering and Technology, Vietnam, in 2003 and 2007, respectively, and Ph.D. degrees in Control Engineering and Automation from Vietnam Academy of Science and Technology, Vietnam, in 2016. Currently, he is a lecturer at the Thai Nguyen University of Information and Communication Technology, Vietnam. His research interests include measurement system design, control system, embedded system, neural networks, and multimedia applications. Email: lhlinh@ictu.edu.vn.