Original Research Article

Lungs Cancer Disease Prediction Using Machine Learning

Abstract: *Lung cancer may be a prevalent explanation for death, and it's the sole sort of cancer that's widespread among men and women worldwide. The prime objective of this paper creates a model for predicting lungs cancer using various machine learning classification algorithms like k Nearest Neighbor (kNN), Support Vector Machine (SVM), Logistic Regression (LR), and Gaussian Naive Bayes (NB). Furthermore, assess and compare the performance of the varied classifiers using their accuracy in selecting the best algorithms. The lung cancer dataset is publicly available on Kaggle Machine Learning Repository and therefore the implementation phase dataset is going to be partitioned as 80% for the training phase and 20% for the testing phase then apply the machine learning algorithms. Support vector machine achieved a significant performance in respect of all parameters.*

Keywords: Breast Cancer, Machine Learning, Classification, Accuracy, support vector machine.

I. INTRODUCTION

To put it another way, lung cancer is the leading cause of mortality in both men and women worldwide (Brocken, *et al* (2012). According to other studies, pulmonary cancer accounted for roughly 13% of all cancer diagnoses in the United States in 2015. Lung cancer accounts for approximately 27% of all cancer-related deaths, according to the American Cancer Society Vivekanandan, (2013). As a result, lung nodules in the early stages of development must be properly examined and monitored. Cancer development and progression were investigated by the researchers in this study using the ML and DL methodologies for predicting cancer growth and progression. The prediction models discussed here are built using a variety of supervised

machine learning algorithms as well as various input and data samples. Using the image operator LBP, images can be turned into arrays or images of integer labels, which are referred to as local binary patterns. These labels are used in further image analysis, which is most typically presented in the form of a histogram. As a result of the LBP texture operator's ability to be specific and how easy, it is to use, it has been used in a wide range of applications (Vivekanandan, 2013).

The histogram then makes use of these markers to conduct a more thorough analysis of the image. In the previous three years, cancer mortality from lung disease has remained greater than cancer mortality from prostate or breast cancer in both men and women (D'Cruz, 2016). In large part, this is owing to the sophisticated and systemic character of the prognostic models for prostate and breast cancer that have been developed in recent years. To do this, it is necessary to develop a reliable early-stage lung cancer forecast model as soon as possible (Shen, *et al* 2021). An effective predictor in both linear and nonlinear scenarios, SVM, has found widespread use across many industries, including medicine Abdullah, *et al* (2021). Still, cancer prognostic models are being made even though SVM is a great way to classify things Jenipher, *et al* (2021). Patients' best treatment options are determined by the results of a mutation test (Binson, et al 2021), which has become more important in clinical trials. In addition to screening, direct sequencing can be used to uncover mutations that were missed during the screening process. A genetic mutation in the EGF receptor (EGFR) has been discovered and can be utilized to detect genetic mutations in lung cancer. It has been demonstrated that the artificial neural network (ANN) and support vector machine (SVM) outperform their no ensemble counterparts (Xie, *et al* 2021). Because the majority misjudgment carries a bigger weight than the minority, miss judgment is more likely to occur for the majority than for the minority. Classification algorithms

that rely on traditional methods of doing things do not perform as well as they could (Miller, et al 2021).

II. MACHINE LEARNING ALGORITHMS

Figure 1 shows the bosom breast cancer classification model with machine learning calculations, where the breast cancer dataset is loaded, and features need to be extracted therefore the classification model is often trained and used for the prediction of benign and malignant. Benign cases are considered noncancerous, which is non-perilous. Harmful cancer begins with irregular cell development and may quickly spread or attack close-by tissue altogether that is regularly hazardous.

A. k Nearest Neighbor (kNN)

k Nearest Neighbors algorithm utilizes 'feature similarity' to foresee the estimations of the most recent snippets of data which further methods the new information point will be assigned a value upheld how closely it matches the points inside the training set.

B. Support Vector Machine (SVM)

Support Vector Machine is of the Supervised Machine Learning characterization strategies that are broadly applied inside the field of cancer malignant growth determination and guess. Support Vector Machine works by choosing basic examples from all classes referred to as help vectors and isolating the classes by creating a linear function that partitions them as comprehensively as conceivable utilizing these help vectors. In this way, it is regularly said that planning between an input vector to a high dimensionality space is framed utilizing a Support Vector Machine that intends to search out the preeminent reasonable hyperplane that separates the data set into classes. This linear classifier intends to expand the space between the decision hyperplane and along these lines the closest data, which is named the minimal distance, by finding the most appropriate hyperplane.

C. Logistic Regression (LR)

Logistic Regression is a key machine-learning classification procedure. It has a place with the gathering of linear classifiers and is fairly practical like polynomial and statistical regression. Logistic regression is quick and similarly simple, and it's helpful for you to decipher the outcomes. Although it's a path for binary classification, it additionally can be applied to multi-class issues. This is frequently not the same as statistical regression, as statistical regression contemplates the forecast of consistent qualities. Logistic regression models the likelihood that a reaction falls into a specific classification. A logistic regression model helps us solve, via the Sigmoid function, situations where the output can take but only two values, 0 or 1.

D. Naive Bayes (NB)

Naive Bayes is a classification method bolstered by Bayes' Theorem with a presumption of independence among predictors. In straightforward terms, a Naive Bayes classifier considers that the nearness of specific features during a class is inconsequential to the nearness of the other element. although these features rely on each other or upon the presence of the contrary features, those properties freely add to the likelihood of a class which is the reason it's referred to as 'Naive'. Naive Bayes (NB) is 'naive' because it makes that features of estimation are free of each other. This is frequently naive because it's (nearly) never evident. The naive Bayes model is easy to make and especially valuable for huge data sets. nearby straightforwardness, Naive Bayes is comprehended to beat even profoundly sophisticated classification methods.

III. ABOUT THE DATASET

This paper is predicated on a dataset that is openly accessible from the US National Lung Screening Trial (NLST). The data contains information about current and former smokers who

were observed for 7 years and were tested for lung cancer each year. No non-smokers were involved in the trial. The dataset has the resulting attributes:

- PID - anonymous identifier of a person
- age - the age of a person at the start of the trial
- gender - Male/Female
- race - the race of a person
- smoker - Former/Current (Former is defined as quitting smoking in the last 15 years)

IV. LITERATURE REVIEW

Benbrahim *et al.* (2020) use classification experimentation to call attention to that the most straightforward accuracy inside the paper was accomplished by the Neural Network calculation, which had, in its best configuration, 96.49% of exactness.

Deepika and Nidhi (2017) use two classification algorithms Naive Bayes and Multi-Layer Perceptron and after analyzing the performance of both algorithms found that Naive Bayes gives more accurate results.

Mariam *et. al.* (2018) use two different classifiers namely Naive Bayes and K Nearest Neighbors for breast cancer classification on comparing accuracy using cross-validation and KNN achieved 97.51% accuracy with the lowest error rate the Naive Bayes Classifier 96.19% accuracy.

Aruna *et al* (2011) uses three different classifiers namely Naive Bayes, Support Vector Machine, and Decision Tree to classify a Wisconsin breast cancer dataset and got the best outcome by utilizing a support vector machine with an accuracy score of 96.99%.

Chaurasia *et al* (2014) looked at the performance of supervised learning classifiers by utilizing a Wisconsin breast cancer growth dataset and Naive Bayes, Support Vector Machine, Neural

Networks, and Decision Tree techniques applied. Reliable with the investigation results, the Support Vector Machine gave the chief the exact outcome with a score of 96.84%.

V. PROPOSED SYSTEM

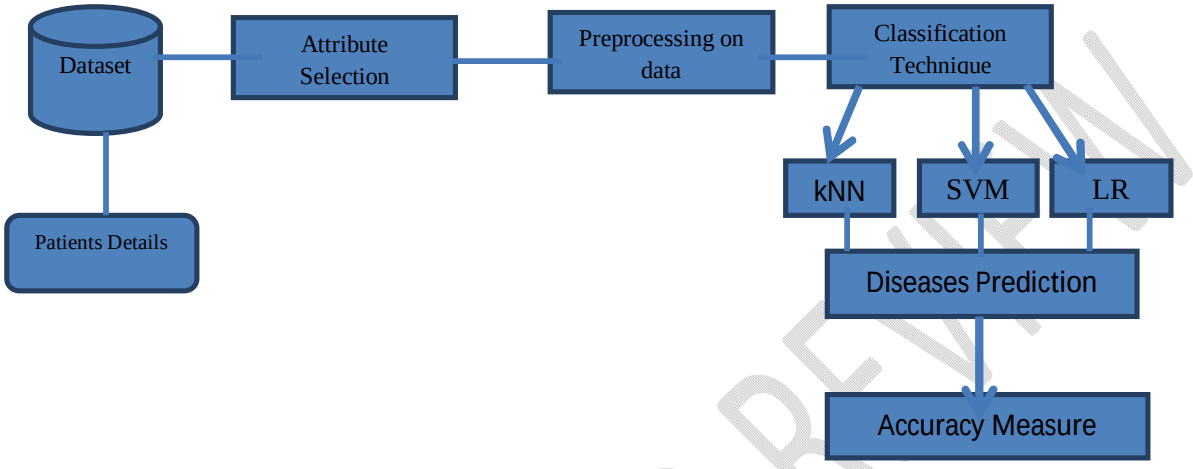


Fig. 1. Bosom breast cancer classification model

VI. SOFTWARE USED

6.1 Python

To collect data a web scraper programmed in Anaconda was used. According to Wikipedia Python's syntax allows programmers to express concepts in fewer lines of code. Python's implementation in December 1989. Python 2.0 was released on October 16th, 2000, and Python 3.0 was released on December 3rd, 2008.

Why use Python for web scraping and not another thing? Python offers a module called 'urllib2', which has suitable functions to open websites and extract information easily. Python is used to program the web scraper that is in charge of collecting the weather data for the model.

6.2 MS Excel

Microsoft Excel is a spreadsheet application developed by Microsoft for Windows and Mac OS X. It features. It features calculation, graphing tools, pivot tables, and a macro-programming language. The first version was released in 1987. Why choose MS Excel versus another similar type of software? MS Excel is a very complete spreadsheet application tool, which supports almost any kind of file extension, and it has a lot of features. Its user-friendly interface helps you most of the time. However, if this doesn't seem enough, I will say that, apart from the typical things a normal user would do in Excel (Charts, Calculation…), it enables you to use the VBA language to create functions to use on the spreadsheets you've created. Excel can also be used as if it were an SQL database as was explained in a previous chapter. Having said this, for me, it is the perfect program.MS Excel is used a lot throughout the project, to visualize the data and perform cleaning tasks on it.

VII. OBSERVATION

The confusion matrix is a table that's frequently wont to depict the performance of a classification model on a gathering of test information that truth values are known.

TABLE 1.
CONFUSION MATRIX

		Predicted Class	
Actual Class		**Class = Yes**	**Class = No**
	Class = Yes	True Positive (TP)	False Negative (FN)
	Class = No	False Positive (FP)	True Negative (TN)

In Table 2 TP and FP are the observations that are accurately predicted and hence shown in blue shading. We might want to decrease false positives and false negatives altogether so that they have appeared in red shading.

VIII. ACCURACY

The classifier exactness is a proportion of how well the classifier can accurately predict cases into their right classification. It's the number of right forecasts separated by the whole number of instances within the data set. It's significant that the accuracy is extremely reliant on the edge picked by the classifier and may, hence, change for different testing sets. Along these lines, it's not the ideal technique to check various classifiers but rather may give a rundown of the classification. Hence, accuracy is often calculated using the following equation:

VIII. RESULT AND DISCUSSION

This project aims to know whether the patients are currents smoker or not (Elkan 1997). The records in the dataset are divided into training sets and test sets. After preprocessing the data. The data classification technique namely support vector. Table 2. Shows the accuracy values for all four machine learning algorithms.

TABLE 2. ACCURACY VALUES

S/N	Algorithms	Accuracy (decimal)	Percentage
1	K Neural Network (kNN)	0.97	97%
2	Support Vector Machine (SVM)	0.98	98%
3	Logistics Regression (LR)	0.89	89%
4	Naïve Bayes (NB)	0.77	77%

Data Source: R-Studio Output

In Table 2 above, the support vector machine accomplishes the critical performance with an accuracy of 98%; the k neural network accomplishes the second performance with an accuracy of 97%. Logistics regression accomplishes the third performance with an accuracy of 89% while naïve-Bayes accomplished the fourth performance with an accuracy of 77%. Consistent with the prediction results, the support vector machine has the very best accuracy for the given dataset. This shows support vector machine is regularly better for the prediction of lung cancer as compared with the Support Vector Machine, Logistic Regression, and Naive Bayes.

XIII. CONCLUSION

This paper provides deep insight into machine-learning techniques for the classification of lung cancer due to smoking. The role of the classifier is crucial in the healthcare industry so that the results can be used for predicting the treatment which can be provided to patients. The existing techniques are studied and compared for finding efficient and accurate systems. Machine learning techniques significantly improve the accuracy of lung cancer prediction through which patients can be identified during an early stage of the disease and can be benefitted from preventive treatment. In this paper, we have compared the classification parameters as far as four Machine Learning algorithms, in particular, k Nearest Neighbor, Support Vector Machine, Logistic Regression, and Naive Bayes. The target of this comparative analysis was to search out the foremost accurate machine learning algorithm which will act as a tool for the diagnosis of lung cancer, consistent with the prediction results, k Nearest Neighbor has the very best accuracy for the given dataset. This shows k Nearest Neighbor is regularly better for the prediction of lung cancer as compared with Support Vector Machine, Logistic Regression, and Naive Bayes.

XII. REFERENCE

Abdullah, D. M., Abdulazeez, A. M., & Sallow, A. B. (2021). Lung cancer prediction and classification based on correlation selection method using machine learning techniques. *Qubahan Academic Journal*, *1*(2), 141-149.

Aruna, S., Rajagopalan, S., & Nandakishore, L.(2011). "Knowledge-based analysis of various statistical tools in detecting breast cancer" *Computer Science Information Technology*.

Benbrahim, H., Hachimi, H., & Amine, A. (2020) *Springer, Comparative Study of Machine Learning Algorithms Using the Breast Cancer Dataset*.

Binson, V. A., Subramoniam, M., Sunny, Y., & Mathew, L. (2021). Prediction of pulmonary diseases with electronic nose using SVM and XGBoost. *IEEE Sensors Journal*, *21*(18), 20886-20895.

Brocken, P., Kiers, B. A., & Looijen-Salamon M. G., (2012). "Timeliness of lung cancer diagnosis and treatment in a rapid outpatient diagnostic program with combined ^{18}FDG-PET and contrast enhanced CT scanning," *Lung Cancer*, vol. 75, no. 3, pp. 336–341.

Chaurasia, V., & Pal, S. (2014). "Data mining techniques: To predict and resolve breast cancer survivability" *Int. J.Computer Science*.

D'Cruz, J. Jadhav, A. Dighe, A. Chavan, V. and J. Chaudhari, (2016) "Detection of lung cancer using back propagation neural networks and genetic algorithm," *Computing Technologies and Applications*, vol. 6, pp. 823–827.

Deepika, V., & Nidhi., M. (2017). "Comparative analysis of breast cancer and hypothyroid dataset using data mining classification techniques" IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI).

Elkan C. (1997). "Naive Bayesian Learning, Technical Report CS97-557", Department of Computer Science and Engineering, *University of California, San Diego, USA*.

Jenipher, V. N., & Radhika, S. (2021, February). SVM Kernel Methods with Data Normalization for Lung Cancer Survivability Prediction Application. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 1294-1299). IEEE.

Mariam, A., Saliha, O., Ikram, G., &Tolga, E., (2018). "Breast cancer classification using machine learning" Electric Electronics, Computer Science, *Biomedical Engineerings, Meeting* (EBBT).

Miller, H. A., Yin, X., Smith, S. A., Hu, X., Zhang, X., Yan, J., ... & Frieboes, H. B. (2021). Evaluation of disease staging and chemotherapeutic response in non-small cell lung cancer from patient tumor-derived metabolomic data. *Lung Cancer*, *156*, 20-30.

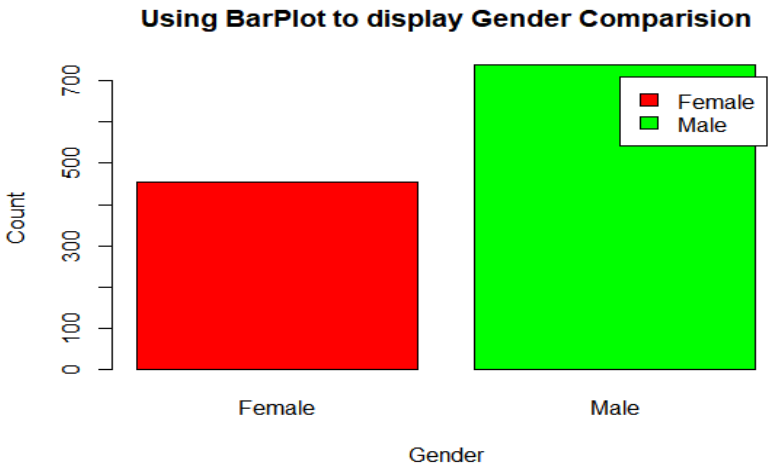
Shen, J., Wu, J., Xu, M., Gan, D., An, B., & Liu, F. (2021). A Hybrid Method to Predict Postoperative Survival of Lung Cancer Using Improved SMOTE and Adaptive SVM. *Computational and mathematical methods in medicine*, *2021*.

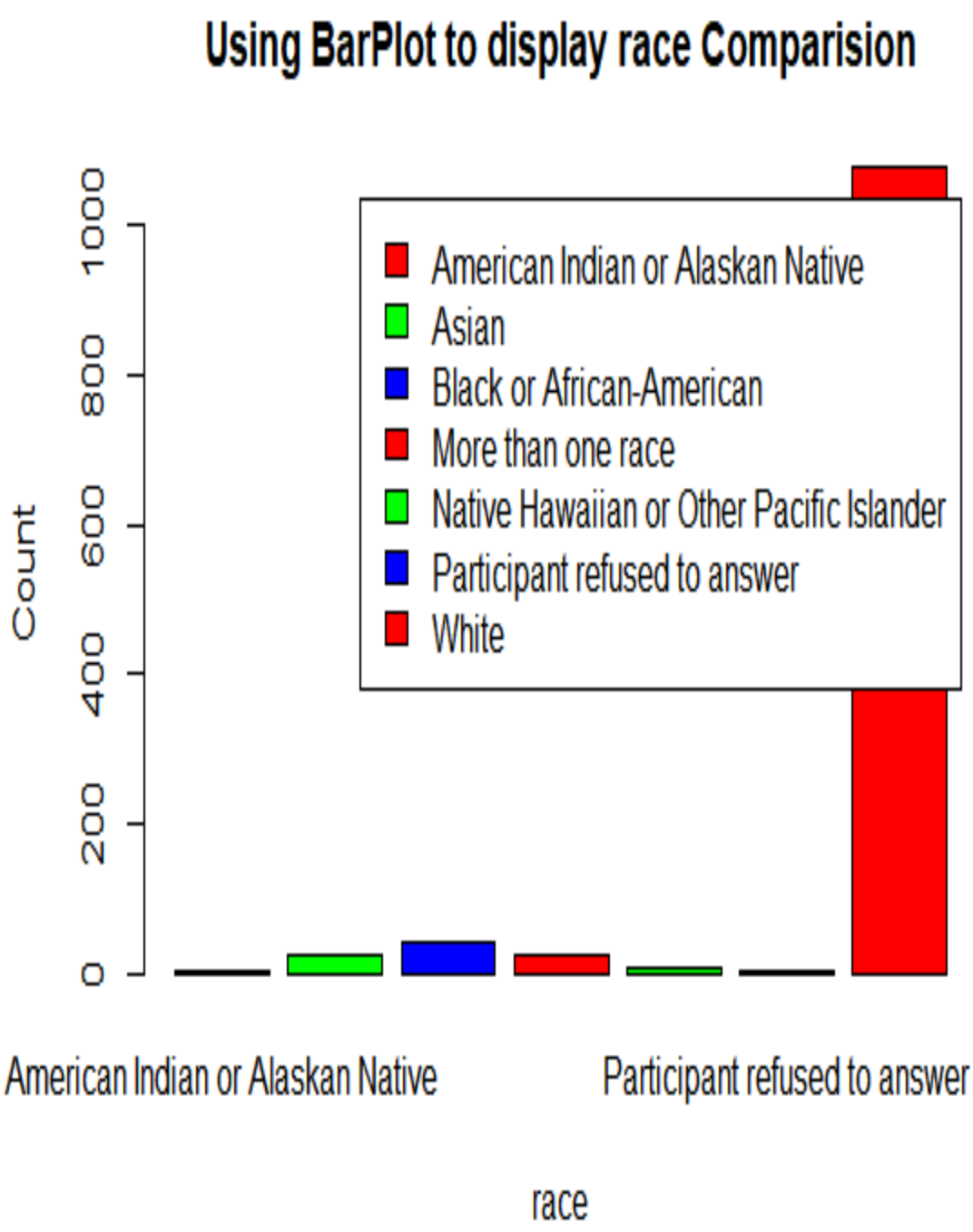
Vivekanandan, P. (2013). "An efficient SVM based tumor classification with symmetry non-negative matrix factorization using gene expression data," in *In 2013 International conference on information communication and embedded systems (Icices)*, pp. 761–768, IEEE, USA.

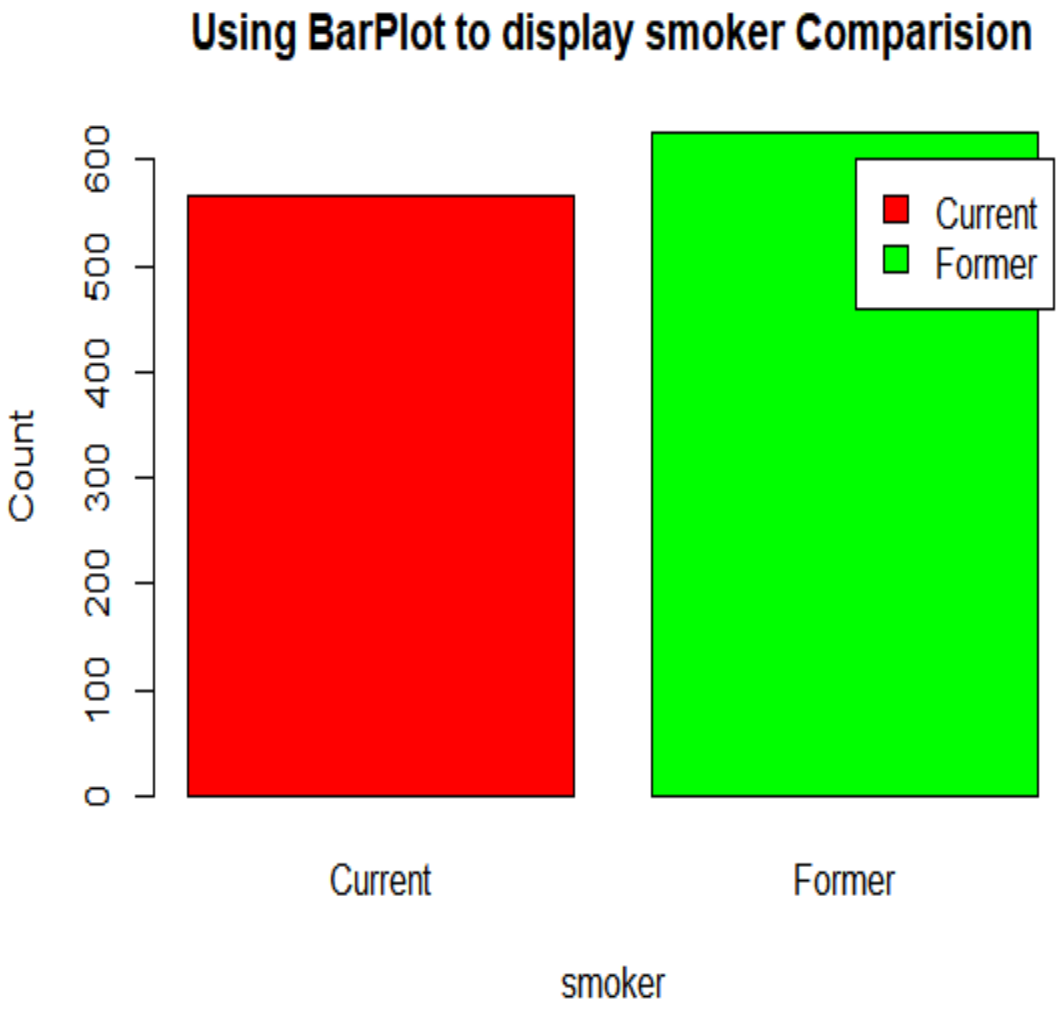
Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., ... & Leung, E. L. H. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational oncology*, *14*(1), 100907.

Appendix

```
> head(cancer)
    pid age gender  race  smoker
1 100001  70   Male White Current
2 100002  66   Male White Current
3 100003  64   Male White Current
4 100004  60   Male White  Former
5 100005  64   Male White  Former
6 100006  56 Female White Current
> tail(cancer)
       pid age gender                race  smoker
1187 101195  71 Female               White  Former
1188 101196  61 Female               White Current
1189 101197  70   Male More than one race  Former
1190 101198  60   Male               White Current
1191 101199  55 Female               White Current
1192 101200  69 Female Black or African-American Current
```




Using BarPlot to display race Comparision
Count
0
200
400
600
800
1000
American Indian or Alaskan Native
Asian
Black or African-American
More than one race
Native Hawaiian or Other Pacific Islander
Participant refused to answer
White
American Indian or Alaskan Native
Participant refused to answer
race


Using BarPlot to display smoker Comparision
Count
0
100
200
300
400
500
600
Current
Former
smoker
Current
Former