# Logistic Regression and Deep Neural Networks in Predicting Lung Cancer, a Comparison

Hayden White
Department of Computer Science
University of South Carolina Upstate
Spartanburg, South Carolina, USA
September 10, 2022
(864) 279 - 0920

mhwhite@email.uscupstate.edu

## Abstract
Cancer is a disease in which normal functioning cells in the body become abnormal and divide uncontrollably. A cluster of these cells is called a tumor, and when a tumor grows and spreads into other tissues of the body the disease can become life-threatening. One of the most common cancers for women is breast cancer and for men, prostate cancer. For both men and women, colon cancer and lung cancer are also common. Lung cancer in particular is the most common type of cancer diagnosis in the world and one of the leading causes of preventable death. This makes it extremely important to confirm a positive diagnosis of lung cancer as early as possible. There are physical symptoms associated with lung cancer such as coughing, wheezing, fatigue, and many more. One method of determining a diagnoses of lung cancer is through machine learning algorithms that input a key list of the physical symptoms associated with the disease that an individual patient might be experiencing, such as the symptoms listed previously, and outputting a likelihood that the culmination of these symptoms in the patient is indicative of a positive case of lung cancer. This practice naturally invites the question of which machine learning algorithm is most accurate and efficient in determining such an intimate and devastating diagnosis as lung cancer. The following research seeks to answer this question by comparing two reputable machine learning models with an established footprint in modern academia: Logistic Regression and Deep Neural Networks.

## Keywords
Passive Smoking, Hemoptysis, Dyspnoea, Mammography, Mesothelioma, Comorbities, Machine Learning, Deep Learning, Neural Network, Deep Neural Network, Logistic Regression, Linear Regression, Backpropagation, Stochastic Gradient Descent, Partial Derivatives, Multivariable Calculus

## 1. Introduction
The epidemiology of Lung Cancer finds its roots in the first half of the twentieth century, sometime around the year 1930. Before this period, Lung Cancer was an extremely rare disease, despite tobacco consumption having been popularized for hundreds of years prior to this point. The prevalence of the disease culminated in the mid-twentieth century when Lung Cancer became the leading cause of cancer death for males. For females it did not take much longer until a sharp rise in lung cancer rates starting in the 1960s to the 2000s propelled lung cancer to become the most frequent cause of female cancer mortality as well [1]. The epidemic of lung cancer among females occurred later and never peaked as high as it did among males because the prevalence of smoking tapered off at substantially higher levels among males around the late eighties to early nineties [1].

Today, we can readily see that past trends in the rise and fall of cigarette smoking has a direct correlation to the fluctuations in the prevalence of positive cases of lung cancer. As previously stated, tobacco had been widely used for hundreds of years, however the modern epidemic of lung cancer directly followed the introduction of mass manufactured cigarettes with addictive properties [1]. An increasing lifespan resulting from the industrial revolution, the sustained exposures of inhaled carcinogens to the lungs, and the addition of new exposures to etiologic agents in cigarettes lead to lung cancer taking the twentieth century by storm. German scientists in the 1930s and early 1940s facilitated some of the earliest research on the connection between lung cancer and smoking and by the early 1950s studies in Britain and the United States corroborated this strong association between cigarettes and the risk of lung cancer. In the year 1964, the United States Surgeon General claimed all evidence was sufficient enough to support the conclusion that cigarette smoking indeed caused lung cancer. Compared to those who have never smoked before, active smokers have an increased risk of developing lung cancer by a factor of twenty. However, while the pattern of lung cancer occurrence is reflective of smoking trends, the rates of occurrence of lung cancer lag behind smoking rates by roughly twenty years [1].

While cigarette smoking is irrefutably the leading cause of lung cancer, other factors that can increase the risk of developing lung cancers also exist. The involuntary inhalation of tobacco smoke, referred to as passive smoking, has been proven to cause lung

cancer. Outdoor air pollution, including but not limited to combustion-generated carcinogens, is considered to increase the risk of lung cancer among urban populations. There is also the concern that in some cases indoor air can carry many respiratory carcinogens such as asbestos and radon. Lastly, there is evidence continuing to emerge that genetic determinate can increase the risk of lung cancer [1]. Despite the numerous other factors that can contribute to the development of lung cancer, active smoking is still responsible for 90% of all lung cancer cases while outdoor air pollution only accounts for about 2% or less of lung cancer cases. One last notable contributor to lung cancer causation is dietary factors, which have been hypothesized to account for approximately 10 to 30% of the lung cancer burden [1].

Lung cancer prevalence tends to be particularly higher among developed nations in Europe and North America and is less common in the developing nations of South America and Africa. Within those developed countries, the occurrence of lung cancer by race and ethnicity makes the disease relevant for those concerned with the health of minorities. Lung cancer occurs 50% less frequently among white men compared to African-American men [1]. Socioeconomic status is also a major determinate in the risk of lung cancer. In the United States, lung cancer mortality rates among white men is lowest in the Northeast and highest in the Southeast. Low socioeconomic status is associated with an unfavorable profile of several determinants of increased lung cancer risk such exposure to inhaled agents in the workplace, diet, the prevalence of smoking, and the general environment. The Center of Disease Control has documented the deaths caused by smoking-related lung cancer in the United States. For the year 1990, the United States tallied the most deaths in the world with a body count of 127,000. This figure is multiplies when accounting for the rest of the developed world, whose smoking-related fatalities amassed to 457,371 in the year 1990.

Despite a general decreasing trend of lung cancer in the developed world, some countries are experiencing an increase in lung cancer rates. A huge burden of lung cancer cases is predicted for China, which has now become home to one third of the world's smokers. Their smoking-related death count is predicted to amount to several million by the middle of the twenty-first century [1]. Since lung cancer is still the most common diagnosed cancer worldwide, and on the rise in so many countries, the demand for highly accurate models that can predict positive diagnoses of lung cancer are of utmost importance. The earlier lung cancer can be detected, the sooner treatment can begin to combat it, and one of the most cutting-edge facets of medicine today is the development of machine learning algorithms that detect the prevalence of cancer in the human body. In this research, I will be comparing two popular machine learning algorithms: Logistic Regression and Deep Neural Networks. We will investigate the differences in implementing each model, compare the efficiency and complexity of both algorithms, and also the the accuracy of both algorithms in predicting Lung Cancer.

## 2. Literature Review

Surprisingly, despite machine learning algorithms having been used to create predictive models for decades now, applying them towards cancer diagnosis predictions is a relatively new field with few published studies. Luckily, the studies that do exist go into great detail their methodology and results for applying machine learning algorithms to create cancer-related predictive models. A variety of of algorithms  including Decision Trees, Support Vector Machines, Bayesian Networks, and Artificial Neural Networks have been applied in cancer research for the creation of predictive models that have resulted in accurate and effective decision making [3]. There is also a large amount of cancer data that has been made available to the medical research community in which data scientists can find relationships and patterns between complex data to predict future outcomes of a cancer types [3]. The types of machine learning algorithms used effects the performance of a specific model, and each algorithm has their own identifiable advantages and weaknesses. In the last few years, the accuracy of cancer predictions through the use of machine learning models has improved by 15-20%, however these studies will require more adequate validation through larger data samples [3].

A popular machine learning model used in cancer prediction as previously stated, is Artificial Neural Networks. ANNs are supervised learning algorithm, or in other words, they take a labeled set of training data so that they can map or estimate input data to a desired output. The task of a supervised learning algorithm is to categorize data into a set of finite classes. Usually with cancer prediction models we are referring to a binary classification output. For example, an ANN that is modeled to predict the probability that a tumor is or is not malignant (0 for "no," 1 for "yes"). This would also be considered a regression problem, in which a learning function maps input data to a real-value variable [3]. Artificial Neural Networks serve as the gold standard in several classification problems used in cancer prediction, but they also suffer from some drawback. Their layered structure can be time-consuming and lead to poor performance. Also, if an ANN does not work properly in its cancer prediction process, all the hidden layers make it nearly impossible to detect why the neural network did not perform as well as anticipated [3].

When developing an ANN model, or any machine learning algorithm for that matter, it is important to keep data-related issues in mind that can occur when processing a large amount of complex data. This can include issues such as duplicate or missing data, outliers, noise, or data that is considered biased-unrepresentative. There are a few techniques for dealing with these issues in a stage referred to as "data preprocessing" that include dimensionality reduction, feature selection, and feature extraction. Machine learning algorithms learn better when the dimensionality is lower, making dimensionality reduction an important step in reducing noise and providing a more robust model. Feature selection coincides with dimensionality reduction as it is the process of selecting a new feature set that is a subset of the old feature set. Finally, feature extraction creates a new set of features from the initial set that captures all the significant information in a dataset [3].

A successful disease prognosis is dependent on the quality of a medical diagnosis, but a prognostic prediction should ideally take into account more than a simple diagnostic decision. Three predictive tasks should be of concern when reckoning with cancer prediction: the prediction of cancer survival, the prediction of cancer recurrence, and the prediction of cancer susceptibility. For the first case, the main objective is predicting a survival outcome that is disease specific or overall survival after cancer treatment begins. For the last two cases, the objective is finding the likelihood of developing a type of cancer and the likelihood of redeveloping that type of cancer after partial or complete remission [3]. Machine learning algorithms such as ANNs and Decision Trees have been used in these types of prediction models

for nearly three decades, and a growing trend in the last ten years has been the use of supervised learning algorithms specifically towards cancer prediction and prognosis. Physicians conclude the integration of features such as weight, diet, age, family history, high-risk habits, and exposure to environmental carcinogens play an important role in predicting the development of cancer. The expression of certain genes, cellular parameters, and molecular biomarkers are also proven to be informative cancer prediction indicators [3].

One study looks into developing decision making tools that can discriminate between malign and benign breast cancer tumors. When developing these prediction models, risk stratification is important and other existing studies use computer models using techniques such as ANNs to asses the risk of breast cancer patients. These neural networks are distinguishing between benign and malignant tumors through mammography findings, or the use of x-rays to diagnose and locate tumors in the breast. These ANN models use a large amount of hidden layers which work better than neural networks with a smaller number of hidden nodes. For this study in particular, the data collected and used in these models consisted of 48.774 mammography findings as well as tumor characteristics and demographic risk factors. Note, radiologists reviewed all of the mammography records [3]. This data was then inputted in the ANN model. Performance was estimated by ten-fold cross validation. The calculated Area Under the Curve of this model gave a  0.965 accuracy following its training and testing, which the authors claim signifies this model can accurately estimate the risk assessment of breast cancer patients. They also concluded the most important factors used to train the neural network were the mammography findings with tumor registry outcomes [3].

The second study we will analyze identifies a combination of early predictive symptoms and sensations indicative of primary Lung Cancer. An e-questionnaire comprised of descriptors of these symptoms and sensations were administered for patients suspected of having Lung Cancer, each descriptor requesting a binary response of “yes” or “no.” This data was then inputted into a machine learning multivariate regression algorithm called OPLS, or Orthogonal Projection to Latent Structures [4]. The goal of this study was to create a model that helps identify early risk symptoms and sensations of Lung Cancer that can flag individuals for screening and early detection. The early identification of these lung cancer symptoms will greatly impact overall lung cancer mortality resulting from greater survival in early-identified stages [4].

Data was collected from a total of 1200 patients, however 530 of those patients were excluded due to not meeting inclusion criteria, declining to participate, or for other reasons not specified. This resulted in a full sample of 670 patients, which was then filtered again by excluding patients had another or earlier cancer diagnosis, mesothelioma, or other factors. This resulted in a total dataset of 506 patients. All data was anonymized to protect the privacy of the study participants [4]. The e-questionnaire consisted of many symptoms starting with Background (smoking habits, comorbidities, sociodemographic characteristics, etc.), Breathing Difficulties, Phlegm/Expectorates, Pain/Aches/ Discomfort, Fatigue, Voice Changes, Appetite/Eating/Taste Changes, Olfactory Changes, Fever/Chills/Sweating, and Other Changes adding up to 342 total potential items and 285 descriptors, however, after several tests with the OPLS model, these descriptors were narrowed down to 70 variables in total (63 descriptors and 7 background variables) [4].

The 70 variable model resulted in the most accurate model performance with an Area Under the Curve of 0.767, a sensitivity of 84.8%, and a specificity of 55.6%. The author claims this was the first study to their knowledge to utilize an e-questionnaire given to individual patients suspected for Lung Cancer that asks for self-evaluation of pre-diagnostic descriptors of symptoms for associated with Lung Cancer. This questionnaire was combined with a cutting-edge multivariable machine learning analysis of multi-dimensional data to find how different combinations of variables perform in predicting Lung Cancer [4]. In addition to active smoking being the most recognized risk factor, chest pain, cough, dyspnoea, and hemoptysis were also important contributors. The study does recognize, however, that patient recall bias is a potential inhibitor to the accuracy of the prediction model. They also recognize that a larger sample would aid in discovering the importance the selected descriptors, and that their model needs to be tested against the general population in order to properly evaluate its validity as a tool in determining patients to flag as at-risk for Lung Cancer [4].

Now let's discuss Logistic Regression, one of the models I will be comparing in this research. Logistic Regression is a learning algorithm that estimates the association of one or multiple predictor variables with a binary output variable. We call the output variable “binary” because it can only take the form of two states (on or off, yes or no, 1 or 0, etc.). More directly, Logistic Regression is used to estimate the probability of an outcome happening depending on the value of the input variables [5]. This probability has a sigmoidal relationship with the input variables, conveniently constraining the output between 0 and 1. Logistic Regression can easily accommodate for more than one independent variable, allowing us to analyze the relationship each variable has on the binary outcome. This is one of the reasons Logistic Regression is one of the most commonly used predictive models for dichotomous outcomes in medicine [6]. For our research, this means we can analyze the effect each individual symptom associated with Lung Cancer has on the likelihood of the patient being diagnosed with Lung Cancer.

One of the major advantages of Logistic Regression is that the exponentiated logistic regression slope coefficient can be easily interpreted as an odds ratio, indicating how much the odds of an outcome can change versus a reference category (when dealing with categorical input variables, specifically) [5]. The second model we will be comparing, a Neural Network, also has advantages. They require less formal training statistically, they have the ability to implicitly detect complex nonlinear relationships between input and output variables, the ability to detect all possible interactions between predictor variables, and also the availability of several training algorithms [6]. Neural networks also have disadvantages, one of them being the “black-box” nature of them that makes it hard to pinpoint exact locations in the model that are problematic. Other disadvantages are the computational burden, proneness to overfitting, and the empirical nature of model development [6].

## 7. References


1. Alberg, A. J., & Samet, J. M. (2003). Epidemiology of Lung Cancer. *Chest*, *123*(1), 21s–49s.

2. Islami, F., Torre, L. A., & Jemal, A. (2015). Global trends of lung cancer mortality and smoking prevalence. *Translational lung cancer research*, *4*(4), 327–338. https://doi.org/10.3978/j.issn.2218-6751.2015.08.04

3. Kourou, Konstantina, et al. "Machine Learning Applications in Cancer Prognosis and Prediction." *Computational and Structural Biotechnology Journal*, vol. 13, Nov. 2014, pp. 8–17. *ScienceDirect*, www.sciencedirect.com/science/article/pii/S2001037014000464.

4. Levitsky, A., Pernemalm, M., Bernhardson, BM. *et al.* Early symptoms and sensations as predictors of lung cancer: a machine learning multivariate model. *Sci Rep* **9**, 16504 (2019). https://doi.org/10.1038/s41598-019-52915-x

5. Schober, P., & Vetter, T. R. (2021). Logistic Regression in Medical Research. *Anesthesia and analgesia*, *132*(2), 365–366. https://doi.org/10.1213/ANE.0000000000005247

6. Tu, Jack. "Advantages and Disadvantages of Using Artificial Neural Networks Versus Logistic Regression for Predicting Medical Outcomes." *Journal of Clinical Epidemiology*, vol. 49, no. 11, Sept. 1995, pp. 1225–31. www.sciencedirect.com/science/article/abs/pii/S0895435696000029.

7. Michael A. Nielson, "Neural Networks and Deep Learning", Determination Press, 2015