Available online at: http://www.ijmtst.com/vol5issue11.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 05, Issue No: 11, November 2019

Importance of Artificial Intelligence – Machine Learning & Deep Learning Prediction in Cancer Diagnosis using Logistic Regression

Subash Kumar

Bachelor of Engineering specialized in Computer Science, Anna University, India.

To Cite this Article

Subash Kumar, "Importance of Artificial Intelligence – Machine Learning & Deep Learning Prediction in Cancer Diagnosis using Logistic Regression", *International Journal for Modern Trends in Science and Technology*, Vol. 05, Issue 11, November 2019, pp.-17-22.

Article Info

Received on 25-September-2019, Revised on 18-October-2019, Accepted on 25-October-2019, Published on 02-November-2019.

ABSTRACT

There are many ways to define the field of Artificial Intelligence. Here is one way for Artificial Intelligence is "The Study of the computations that make it possible to perceive, reason, act and predict the future possible outcomes". In the field of computer science, the artificial intelligence (AI) are as called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans.As the world grows more complex, we must use our material and human resources more efficiently, and to do that, we need high quality help from computers. Here are few possibilities such as, in Agriculture, Computer Controlled robots should control pests, prune trees, and selectively harvest mixed crops.In Manufacturing, Computer Controlled robots should do the dangerous and boring assembly, inspection, and maintenance jobs.In Medical care, Computers should help practitioners with diagnosis, monitor patients, conditions, manage treatment and make beds.In Household work, computers should give advice on cooking and shopping, clean the floors, mow the lawn, do the laundry and perform the maintenance chores.In Schools, computers should understand why their students make mistakes, not just react to errors. Computers should act as super books, displaying planetary orbits and playing musical scores, thus helping students to understand physics and music([1]). In this world if computers are helpful to the practitioners in medicinal field with diagnosis, monitor patients health conditions and produces directions to multidisciplinary experts, then computers are the greatest gift to the mankind.

KEYWORDS: *Machine Learning, Deep Learning, Artificial Intelligence, Medicine, Lung Cancer, Breast Cancer, Cancer prognosis and prediction, Data Science, Logistic Regression, Regression, Supervised Machine Learning*

I. INTRODUCTION

The International Agency for Research on Cancer (IARC) of the World Health Organization (WHO) released the latest estimates on the global burden of cancer on 12th September 2018 at Geneva, Switzerland. The database is publicly available and it canbe accessed online. The IARC Global Cancer Observatory database wasdeveloped based on the facts about estimates of incidence and mortality

collected from 185 countries for 36 types of cancer from different sources.

The increasing cancer burden is due to several factors such as an increase in population and aging as well as the changing prevalence of certain causes of cancer linked to social and economic development. This is particularly true in fast-growing economies, where a shift is observed from cancers related to poverty and infections to cancers associated with lifestyles more typical of industrialized countries.

The statistics show for men and women combined, nearly half of the new cases and more than half of the cancer deaths worldwide in 2018 are estimated to occur in Asia, in part because the region has nearly 60% of the global population. Breast cancer incidence has gone up by 39.1 percent between 1990 and 2016 and is the most common cancer among women in India, Europe accounts for 23.4% of the global cancer cases and 20.3% of the cancer deaths, although it has only 9.0% of the global population. The Americas have 13.3% of the global population and account for 21.0% of incidence and 14.4% of mortality worldwide. In contrast to other world regions, the proportions of cancer deaths in Asia and in Africa (57.3% and 7.3%, respectively) are higher than the proportions of incident cases (48.4% and 5.8%, respectively), because these regions have a higher frequency of certain cancer types associated with poorer prognosis and higher mortality rates, in addition to limited access to timely diagnosis and treatment in many countries. Cancers of the lung, female breast, and colorectum are the top three cancer types in terms of incidence.*(7)*

Cancer is a killing organism that lives off a host organ, growing by the bio-genetic-molecular mechanism. 9.6 million People died of cancer in 2018 more than from HIV/AIDS, malaria and tuberculosis combined. The incidence of cancer is estimated to double by 2035, with most of these cases expected to occur in low-to-middle income countries (LMICs). 60% of cancer cases occur in LMICs, and 75% of cancer deaths occur in these countries.

The Research and Development in oncology are supporting the merging of different research fields, such as genetics, genomics, Nanotechnology, Nano medicine Computer Sciences, etc., that are generating new technological pathways for diagnostics and therapeutics. Any hospital in this world will be having a multidisciplinary team of experts involving radiologists, oncologists, surgeons, and other respiratory physicians and consultants. This multidisciplinary team uses a wide variety of data taken from CT Scans, X-Rays, Pet Scans (Positron Emission Tomography) and Biopsies to identify whether a patient has lung cancer. These tests help the team to fully diagnose a patient and the approach would be used by all of them to gather the data. Medical professionals use TNM classification to help characterize lung cancer at the starting stage to the advanced stage of malignant tumors. IA is the earliest stage of cancer which is more likely that it was accidentally discovered, the more difficult it will be to perform a biopsy and ultimately the better the (likelihood to survive). On the other hand, IV is an advanced stage of cancer which means that it is easier to diagnose, including biopsy, highly likely to cause symptoms and the worse the prognosis.

The existing challenge here with medical professionals when dealing with IA classification is that the tumors are so small less than 4 mm that it is very difficult to diagnose them via CT scan images and predict what causes cancer or the likelihood of occurring again. The biopsy needs to be conducted however this process can be very invasive and infeasible to do as IA tumors have around 50/50 chance to be benign/malignant. Malignant being spreading tumor and Benign being a non-spreading tumor. The general medical advice is to rescan in 5-10 weeks to see any evidence for growth which can grow more than double in size.(8)

Having this serious threat to the mankind across the world, "What if the Computer is trained with unlabeled data, the algorithm is trained and left to find the commonalities among its input data and computer might be able to teach new things after it learns patterns from the data about tumors are so small less than 4 mm that it is very difficult to diagnose them via CT scan images that makes these algorithms particularly very helpful to Doctors and Experts, in cases where the human expert doesn't know what to look for in the data . Yes, it is possible with the approaches to Artificial Intelligence Supervised or Unsupervised Machine Learning techniques".(2)

Lung cancer is one of the main diseases in several countries and a leading cause of cancer death of both sexes worldwide. Lungcancer is linked to several risk factors in society, such as smoking, passive smoking, air pollution, etc. Lung cancer can be either small cell lung cancer or no small cell lung cancer, with the latter representing about 80% of the cases. The mortality of lung cancer is very high and the five-year survival rate of

patients is about 2–10%. The detection of the type of lung cancer (e.g., adenocarcinoma, squamous cell carcinoma, etc.), mutations (e.g., EGFR: epidermal growth factor receptor; ALK: anaplastic lymphoma receptor tyrosine kinase rearrangements) and sequential mutations (T790) is a critical process in diagnostics to select the appropriate therapy that can be conventional chemotherapies, target therapies such as gefitinib, erlotinib, etc., or new immune therapies, e.g., nivolumab, atezolizumab. This critical process affects the survival of patients. Classification and mutation of lung cancer type is a critical diagnostic process because of the available treatment options, including conventional chemotherapy and, more recently, targeted therapies, differ for lung adenocarcinoma and lung squamous cell carcinomas. Current technology for an accurate diagnosis and selection of treatment options is based on molecular biomarkers applied on lung biopsies and/or blood testing this approach can diagnose lung cancer type and stage. This current technology has some negative effects. On the one side, patients have to receive an invasive surgical intervention for tissue biopsies or to do liquid biopsy (e.g., blood) and wait a certain time (about a month) before to have the diagnosis of cancer type and correct treatments, and the time in the presence of cancer is a critical variable for survival and, when possible, for healing. On the other side, hospitals have a high cost with these approaches that require appropriate equipment and specialized personnel, which in some regions are scarce or overload of work because of the shrinking public lab and hospital budgets. Let's see how to tackle these issues with Artificial Intelligence Engineering. (2)

Artificial Intelligence (Machine Learning & Deep Learning) detecting the type of lung cancer: There are some studies show that deep learning can be used for the classification of breast, bladder and lung tumors, particularly, the development of new, inexpensive and more powerful artificial intelligence technologies has made possible the training of larger and more complex neural networksThis has resulted in the design of several deep convolutional neural networks (CNNs) that are capable of accomplishing complex visual recognition tasks shows that medicine can benefit from deep learning technology by convolutional neural networks (CNNs) that not only outperforms other methods but also achieves accuracies to classify the type of cancer that is comparable to pathologists for selecting the appropriate therapy. These models maintain their performance when tested on independent datasets of frozen samples as well as on images obtained from biopsies. CNN's idea have been applied to the classification of lung patterns on computerized tomography (CT) scans, with positive results. Google Incorporation in 2014 won the ImageNet Large-Scale Visual Recognition Challenge by using this CNNs Technology. The ImageNet project is a large visual data designed for use in visual object recognition software research. More than 14 million images have been hand annotated by the project to indicate what objects are pictured and in at least one million of the images, this concept is a branch of Unsupervised Machine Learning, bounding boxes are also provided. ImageNet, developing the GoogleNet architecture, increased the robustness to translation and nonlinear learning abilities by using microarchitecture units called inception. Each inception unit includes several nonlinear convolution modules at various resolutions. Inception architecture is particularly useful for processing the data in multiple resolutions, this feature suitable for medicine and in particular for pathology tasks. This complex Convolutional Neural Network (ANN) has already been successfully adapted to specific types of disease classifications like skin cancers and diabetic retinopathy detection and has developed a deep-learning model for the automatic analysis of tumor slides using publicly available whole-slide images available in The Cancer Genome Atlas (TCGA). The scholars trained inception v3 to recognize tumors in lung versus normal tissue. After that, they tested the performance of these methods on the more challenging task of distinguishing lung adenocarcinoma and lung squamous cell carcinoma. They also evaluated the deep-learning model by training and testing the network on a direct three-way classification into the three types of images. Results of this approach are compared to the evaluation of three pathologists (two thoracic pathologists and one anatomic pathologist) that independently classify the whole-slide images in the test set by visual inspection alone, independently of the classification provided by TCGA. Overall, the performance of deep learning models was comparable to that of each pathologist. Hence, new artificial intelligence technology, based on deep-learning models, can assist pathologists in the detection of cancer subtype or gene mutations in any cancer type with a save of time and costs; moreover, poor regions, with these AI technologies,

can also have high benefits by sending the digital images to labs of other countries, generating a reduction of current gap in healthcare between countries. Machines are very helpful not only to identify but also classify, detect or distinguish tumors and other malignancies. In other words, machines have been used primarily as an aid to cancer diagnosis and detection useful for mankind.

Artificial Intelligence (Machine Learning & Deep Learning) detecting the type of Breast cancer: Thousands of our grandmothers, mothers, and daughters fall victim to breast cancer every year. The human body comprises of millions of cells each with its own unique function. When there is unregulated growth of any of these cells it is termed as Cancer. In this, cells divide and grow uncontrollably, forming an abnormal mass of tissue called as tumor. Tumor cells grow and invade digestive, nervous and circulatory systems. Breast cancer is the most frequent cancer worldwide among women. Studies based on advanced countries show that the incidence of breast cancer has been increasing. It has been postulated that breast cancer incidence tends to be higher in more developed countries due to delayed Childbearing, revealed by (WHO), higher usage of hormone replacement therapy, a higher rate of screening, and improved tumor registries. Some studies have also argued that higher-income countries may have higher fat diets, consumption of processed foods, and an increased rate of obesity, both correlated with higher breast cancer incidence rates. In general, researches note that many Western populations have a higher incidence rate of breast cancer than regions in Africa and Asia Breast biopsies, like all types of cancer, are used to diagnose breast cancer type, stage, and mutations. An accurate breast cancer staging is an essential task performed by pathologists to inform clinical management. For instance, evaluation of the extent of cancer spread by histopathological analysis of sentinel axillary lymph nodes is an important part of breast cancer staging. The sensitivity of sentinel axillary lymph nodes evaluation by pathologists, however, is sub-optimal because some studies show that pathology review by experts changed the nodal status in 24% of patients this effect is due to the difficulty of the process of decision making of human behavior. Moreover, sentinel axillary lymph nodes evaluation is time-consuming and has a certain cost. In short, the high cost of exams and a period of more than one month to have breast cancer type, stage, etc. to decide the appropriate therapy, it can affect the results of therapeutic treatments and survival of patients. The application of Artificial Intelligence (AI) machine learning technology with deep learning algorithms to whole-slide pathology images can potentially improve the diagnostic accuracy of breast cancer and metastases. Some scholars have assessed the performance of automated deep learning algorithms at detecting metastases in tissue sections of lymph nodes of women with breast cancer results are accurate than the results compared with pathologists diagnoses, some deep learning algorithms achieved better diagnostic performance than a panel of 11 pathologists participating in a simulation exercise designed to mimic routine pathology workflow; in short, algorithm performance was comparable with an expert pathologist interpreting whole-slide images without time constraints. Hence, this experiment has shown that deep learning algorithms could identify metastases in sentinel axillary lymph nodes slides with 100% sensitivity, whereas 40% of the slides without metastases could be identified as such. This approach can significantly reduce the workload of pathologists and improve the management decisions on whether or not to administer therapy, perform a surgical intervention, etc. Overall, then, this interesting result shows the potential of AI applied in cancer imaging for the detection of lymph node metastases in women with breast cancer to improve processes of diagnosis and therapy of this and other critical cancers.(2)

The continuous development of technology in the medical field will save countless lives and the overall quality of human life continues to improve over time.

II. METHODOLOGY

Regression searches for relationships among variables or features. Regression analysis is one of the most important fields in statistics and machine learning. There are many regression methods available which data scientists work involved. In this paper we will discuss about the Logistic Regression.

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.Binary logistic regression requires the dependent variable to be binary.For a binary

regression, the factor level 1 of the dependent variable should represent the desired outcome.Only the meaningful variables should be included.The independent variables should be independent of each other. That is, the model should have little or no multicollinearity.The independent variables are linearly related to the log odds.Logistic regression requires quite large sample sizes.Logistic regression is useful to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease, Cancer), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.) because Logistic regression deals with Binary Boolean data or Discrete data, for example, Is a person is having cancer or not, if he has cancer then 1 else 0. From the features below mentioned, Patient ID does not gives any meaningful information for cancer prediction, so this particular feature could be eliminated, the last feature Target(Y-axis) is the dependent feature, now we need to build a logistic regression model to regress the independent features(X-axis) against the dependent feature(Y-axis). Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses are considered as Independent features and the last feature, Target is considered as Dependent feature.

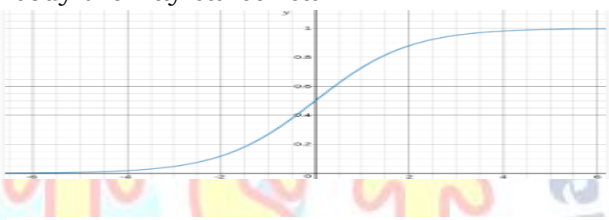
Here we developed a Logistic Regression model using Anaconda Python 3.6 with the Wisconsin data set available in public domain and we found the accuracy of the model as 96.5%

Features Explanation:

1. Patientid id number
2. Clump Thickness 1 – 10
3. Uniformity of Cell Size 1 – 10
4. Uniformity of Cell Shape 1 – 10
5. Marginal Adhesion 1 – 10
6. Single Epithelial Cell Size 1 – 10
7. Bare Nuclei 1 - 10
8. Bland Chromatin 1 – 10
9. Normal Nucleoli 1 - 10
10. Mitoses 1 - 10
11. Target: (0 for Benign, 1 for Malignant)

Now we need to Identify for every Patient ID, how every Independent features (X-axis) are closely correlating to the Output or Responses or Dependent Variable (Y-axis), what should be the Diagnosis for every patient ID falls under either. Diagnosis (M being malignant, B being benign). Which are considered as Y – axis or Target variable.

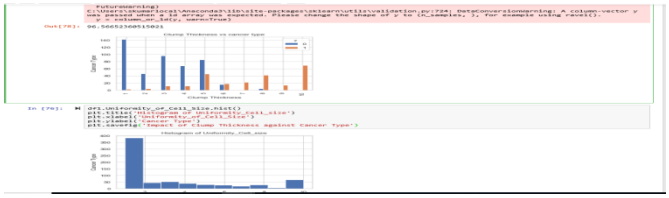
Malignant Tumor: Malignant tumors are cancerous growths. They are often resistant to treatment, may spread to other parts of the body and they sometimes recur after they were removed. Benign tumor is not a malignant tumor, which is cancer. It does not invade nearby tissue orspread to other parts of the body the way cancer can



When implementing regression of some dependent variable Y - axis on the set of independent variables X = (X1, X2, X3 …., Xr), where r is the number of predictors, Y = 1 is Malignant Tumor and 0 is Benign Tumor, finally after developing the correlation of Independent features against the Target variable, we can see the machine scores on predicting more accurate for the cancer patients who has Malignant Tumor and right treatment will be provided to those patients.What do we infer from applying the Statistics in Breast cancer scenario, A patient who has Malignant tumor should not be diagnosed with Benign tumor treatment, which is actually misdiagnosed if it treated with Benign tumor medicine, Machine Learning is really helpful to see how closely the individual features are correlating with dependent features

Algorithm:

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
sns.set(style="white")
sns.set(style="whitegrid", color_codes=True)
df = pd.read_excel('C:\\Users\\skumarlocal\\Desktop\\AI(Machine Learning)\\Cancer_Wisconsin_Datas
df1 = df[['Clump_Thickness','Uniformity_of_Cell_Size','Uniformity_of_Cell_Shape','Marginal_Adhesi
df1['y'].value_counts()
# 0 being Benign tumor
# 1 being Malignant tumor
df1.groupby('y').median()
df1.groupby('Clump_Thickness').median()
df1.groupby('Uniformity_of_Cell_Size').median()
df1.groupby('Uniformity_of_Cell_Shape').median()
df1.groupby('Marginal_Adhesion').median()
df1.groupby('Single_Epithelial_Cell_Size').median()
df1.groupby('Bare_Nuclei').median()
df1.groupby('Bland_Chromatin').median()
df1.groupby('Normal_Nucleoli').median()
df1.groupby('Mitoses').median()
%matplotlib inline
pd.crosstab(df1.Clump_Thickness,df1.y).plot(kind='bar')
plt.title('Clump Thickness vs cancer type')
plt.xlabel('Clump Thickness')
plt.ylabel('Cancer Type')
plt.savefig('Impact of Clump Thickness against Cancer Type')
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(Features, Target)
y_pred = classifier.predict(Features)
predict_excel = y_pred.tolist()
df2 = pd.DataFrame(predict_excel)
df_test_final = pd.concat([X,df2],axis=1,ignore_index=False)
writer2 = pd.ExcelWriter('Cancer_Probability_Percentage_prediction.xlsx',engine='xlsxwriter' )
df_test_final.to_excel(writer2, sheet_name = 'output')
writer2.save()
classifier.score(Features,Target)*100
```



True negatives (TN): We predicted no, and they don't have the disease.

False positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

False negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

True positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.

Confusion Matrix: A confusion matrix is a summary of prediction results on a classification problem against the actual results.

III. RESULTS

The breast cancer data with 10 attributes was analyzed to identify the error rates and accuracy of 649 patients. The accuracy of the model is 96.56%, which shows 96.5% is correct predictions.

	Instances	Probability Percentage
Correctly Classified Instances	611	95.56%
Wrongly Classified Instances	38	4.44%

IV. DISCUSSION

The Logistic regression used here is to predict a patient is diagnosed with Malignant or Benign tumor. In today's world, Breast Cancer is the huge Killer among women worldwide, around 500,000 women died in 2018, which make us feel sad. There is another statistics tells that 90,000 of 500,000 breast cancer patients are misdiagnosed, Women dying every day in this world, which is the huge problem which needs to be prevented using these advanced techniques. I feel that the current method of diagnosing breast cancer is not efficient, so there is a way to improve this using Machine Learning.Few Recommended screening guidelines detecting the Breast Cancer

Mammography: The most important screening test for breast cancer is a mammogram. A mammogram is an X-ray of breasts. It has the ability to detect breast cancer up to two years before you or the doctor would ever notice.Women age 40–45 or older that have an average risk of getting breast cancer should get a mammogram once a year.

Women at high risk should have yearly mammograms along with an MRI starting at age 30.

Factors for Breast Cancer: These are many risk factors for breast cancer. We cannot come to a conclusion that a particular factor is makes breast cancer.

Age: As the age increases the risk of getting breast cancer increases. Almost 80% of breast cancer patients are women over the age of 50.

Having a personal history of breast cancer**:** A woman who had breast cancer in one breast in the past has a higher risk of developing cancer in her other breast.

Having a family history of breast cancer: If a patient mother, sister daughter or other blood-related females have had breast cancer, especially at a young age (before 40) increases the risk of them getting cancer.

Genetic factors: Some women has certain genetic mutations. For example, changes to the BRCA1 and BRCA2 genes. These genetic mutations increase the risk of developing breast cancer. Other changes in a patient's genes may raise the risk of breast cancer as well.

Menstrual history: older a woman is when she has her first child, the greater her risk of breast cancer is.Women who menstruate for the first time before 12. Women who go through menopause after the age of 55. Women who haven't had children. (5)

REFERENCES

[1] Artificial Intelligence Third Edition Patrick Henry Winston

[2] Artificial Intelligence Technology in Oncology: A New Technological Paradigm Author: Mario Coccia

[3] World Health Organization 2019. International Agency for research on cancer, May, 2019

[4] ScienceDirect 2019. Advanced Research, https://www.sciencedirect.com/search/advanced?qs=artificial%20intelligence%20and%20retinopathy&show=25&sortBy=relevance&lastSelectedFacet=years&years=1998 (accessed May 2019)

[5] https://realpython.com/linear-regression-in-python/,https://medium.com/@nyla.pirani/how-to-build-a-basic-breast-cancer-model-machine-learning-9bae53c40f18

[6] Applications of Machine Learning in Cancer Prediction and Prognosis Joseph A. Cruz, David S. Wishart Departments of Biological Science and Prognosis Joseph A. Cruz, David S. Wishart

[7] https://www.google.com/search?q=world+health+organization+cancer&rlz=1C1CHBD_enUS865US866&oq=world+health+organization+cancer&aqs=chrome..69i57.6326j0j1&sourceid=chrome&ie=UTF-8

[8] Lung Cancer Detection using Deep Convolutional NetworksJelo Salomon