Development of Machine Learning Model for Improved Cancer Detection Mechanism Using IBM Machine Learning Cloud

Ibuomo R. Tebepah
Mathematical and Computer Sciences
University of Africa, Toru Orua, Nigeria
tebepahirt@gmail.com

Nathaniel A. Ojekudo
Department of Computer Science
Ignatius Ajuru University of Education, Nigeria
nathojekudo@gmail.com

Abstract

Cancer early detection is vital and has become a major research area for health and computing research. The collaboration between the two fields is essential as more efficient algorithms in machine learning is required. This paper is a demonstration of cancer detection using the IBM machine learning cloud in which different algorithms has been embedded. The Breast Cancer Wisconsin (Diagnostic) Dataset (WDBC) consisting of trained and untrained is used for the analysis. The result shows high rate of accuracy in detecting cancer. The results can also be viewed in different format, thereby making the IBM machine learning cloud an effective means of achieving a comprehensive simulation.

Key Words *machine learning; breast cancer; IBM cloud*

1. Introduction

The level of accuracy in diagnosing cancer in patient has greatly improved. Most pathologist have a 96-98% [1] accuracy rate which is very good considering the large number of persons diagnosed on yearly basis.

However, according to the Oslo University Hospital, the prognoses accuracy is about 60% which complicate or jeopardises the life of the patient. This situation has given rise to applying machine learning in predicting and detecting cancer on patients.

Machine Learning (ML) is one of the core areas in Artificial Intelligence that deals with a system which takes data, manipulate and trains itself using the data, producing an outcome. It means that machine learning enables a computing device to response to an unlabelled data responsibly without a human intervention or preprogramed instructions.

The use of ML in most fields especially in medicine has become essential or necessary because of the high rate of accuracy required in diagnosing and predicting life threatening diseases or infections. In the diagnostic and prognostic analysis of cancer, the application of ML [2] has greatly improved prediction accuracy, early detection, and treatment outcomes. This is because machines generally work faster, repeat series of actions a thousand times without getting exhausted, and high level of accuracy.

2. Literature Review

This section deals with a brief background and review of related works on the current research.

2.1 Background Study

Cancer previously was regarded as an incurable disease. More often people with cancerous growth of any kind dies eventually even after several chemotherapy activities. There are different types or rather causes of cancer, and its location in the human body. Medical research and diagnosis has shown that cancerous growth can be in any part of the body – eye, nose, brain, kidney, liver, lungs just name it. The main cause sometimes is very difficult to determine and medical sciences has not been able to precisely and definitely define the cause. Every stated cause is based on presumed or postulated. For instance cancer of the lungs is akin to heavy smokers, however several heavy smokers do not have lung cancer. Yes life style may contribute, but then how do we define a life style. More recently, it has been discovered that even children are not exempted from the deadly disease.

Cancer is the second cause of death in the world. According to (Tahmooresi et al, 2018) research, 8.8 million patients (people) die due to cancer in 2015. Breast cancer seems to be the leading cause of death among women. This has giving rise to breast cancer campaign awareness from government, social groups, and nongovernmental organisation. Early check-ups and simple self-examination is being advocated.

Abdel-Rahman (2015) work identified that prostate cancer is associated with hereditary. That is patients with family history of prostate cancer are more likely to be diagnosed with the disease.

Peters (2009) research on colorectal cancer (CRC) states it to be the second leading cause of cancer death in the world. According to the research, 35%of CRC is attributed to inherited factors, and identification of associated genetic variants is important to elucidate mechanism underlying this disease.

The above justifies the reason why early detection of cancer is important. Medically, cancer is controllable when the diagnosis is done early, and if discovered, the accurate prognosis will keep the patient alive. The importance of early discovery is so that, the affected cells can be destroyed quickly before it rapidly spreads to other areas.

However, just as it is in the case of most incurable diseases, people are sceptical going for regular check-ups due to cost and not wanting to hear the dreaded news. Some live in denial until it becomes too late to control.

With the integration of machine learning in the diagnostic process, though still expensive in third world countries, it has grossly increased the accuracy rate, giving room for timely intervention, proper medication and follow up.

2.2 Related Works

Agarap (2018) did an extra ordinary work on breast cancer detection using Wisconsin Diagnostic Breast Cancer (WDBC) dataset based on four ML algorithms. Each algorithm accuracy level was high ranging from 94 -99.04%. Although it was clearly stated that the

experimental result may not be applied on real patients, it gives a leading to the possibility of ML, and the extent to which ML can be used to achieve correctness. Also, the work centered on one type of cancer.

Kourou et al (2015) analysed and identified new approaches of ML techniques applied in cancer prognosis and prediction. According to their research, though Artificial Neural Networks (ANNs) and Decision Trees (DTs) has been in existence for decades, the new trend is the use of other supervised learning techniques such as Support Vector Machines (SVMs), and Bayesian Networks (BNs). These has proven to have better accuracy rate in prognosis and predictions. However, from 2015 to date, researchers has also discovered newer models or approaches that give higher accuracy in prediction.

Chen et al (2018) research has integrated deep learning model to a reality microscope that is capable of auto detection of cancer cells in real time. Though the experiment did not use any specific dataset, it demonstrated how fast a system could read through slides containing tissues and examine plus analyse the tissue and present the results within a short period. It is still at the stage of development, but when fully developed, it can easily be utilised as it runs along with a simple computing system.

3. Methodology

3.1 **DataSet** – dataset from the WDBC - ML algorithms were trained to detect different types of cancer using specific but common symptoms amongst the commonest. The images are digitised.

3.2 Machine Learning Algorithms

In other to carry out the supervised experiment, the work used the Multilayer Perceptron (MLP) and Support Vector Machine (SVMs) which in the main related work produced the best accuracy rate even though it looked at only the breast cancer.

3.3 Classification

The data comprises of trained and untrained variables. The algorithm extracts cells from the dataset and classifies them into;

Benign – tumour not spreading. That means patient is safe

Malignant – cancerous. Patient needs urgent attention to stop the spread.

Since the dataset comprises of trained and untrained, certain features are been looked for. Depending on its presence and impression as an image, classification will be made.

Variables:

i. Radius – mean of distance from centre to points on the perimeter
ii. Texture
iii. Area – smoothness
iv. Compactness
v. Concavity
vi. Concave points
vii. Symmetry

viii. Fractal dimensions

3.4 Pseudocode

// an algorithm for detecting different types of cancers from dataset

#import libraries

1. *begin*
2. ***login to IBM Machine Learning cloud***
3. ***create a project***
4. ***create object storage for dataset***
5. ***download and extract dataset from Wisconsin***
6. *Load dataset(data.csv)*
7. *Select variable parameters*
8. *Run simulation*
9. *Display results*
10. *End;*

4. Implementation

In using IBM Watson Machine Learning (Cloud based), it simplified the integration of the algorithm. The machine learning works with Apache Spark, and Object Storage.

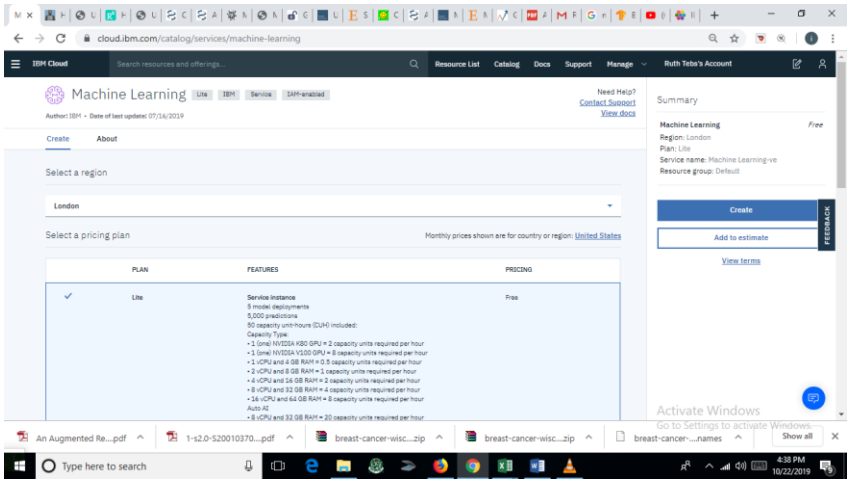


Fig 1: creating machine learning project

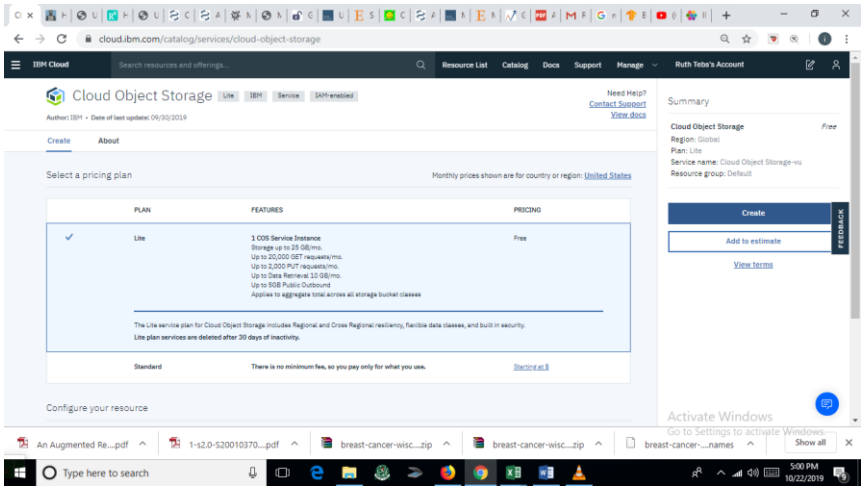


Fig 2: creating object storage for the dataset

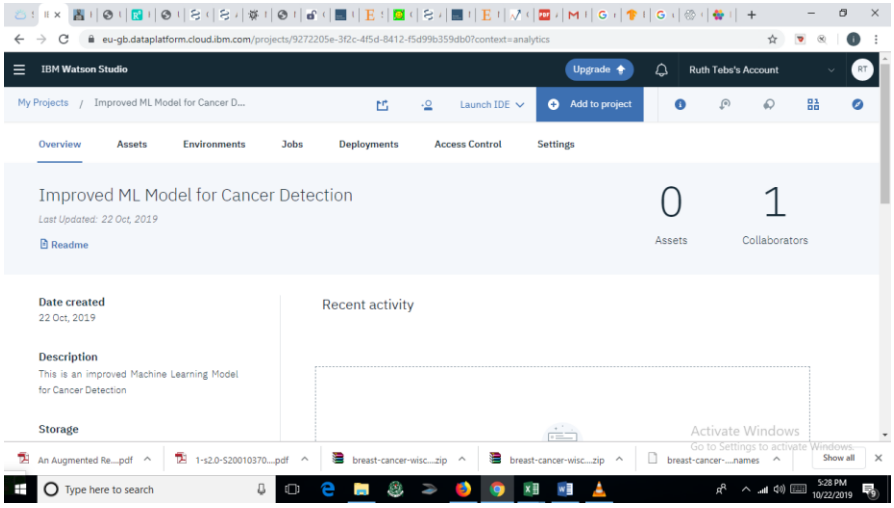


Fig 3: improved ML model for cancer detection

The dataset is uploaded as an asset and published, linking to a catalogue. This makes it accessible for job to be created – the ML model is defined in during the job creation which generate and rates the performance – the accuracy level.

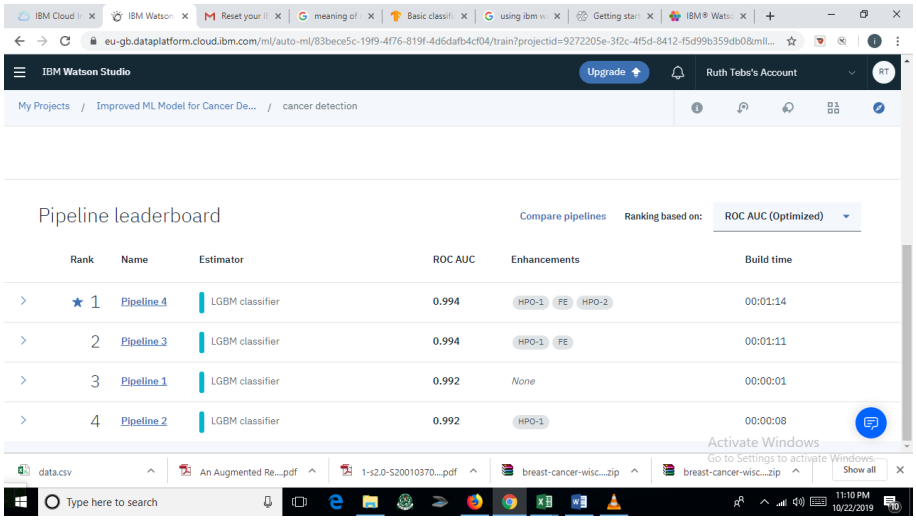


Fig 4: results displaying rate of accuracy

5. Discussion and Results

The experiment showed a high rate of diagnostic accuracy as show in each model use.

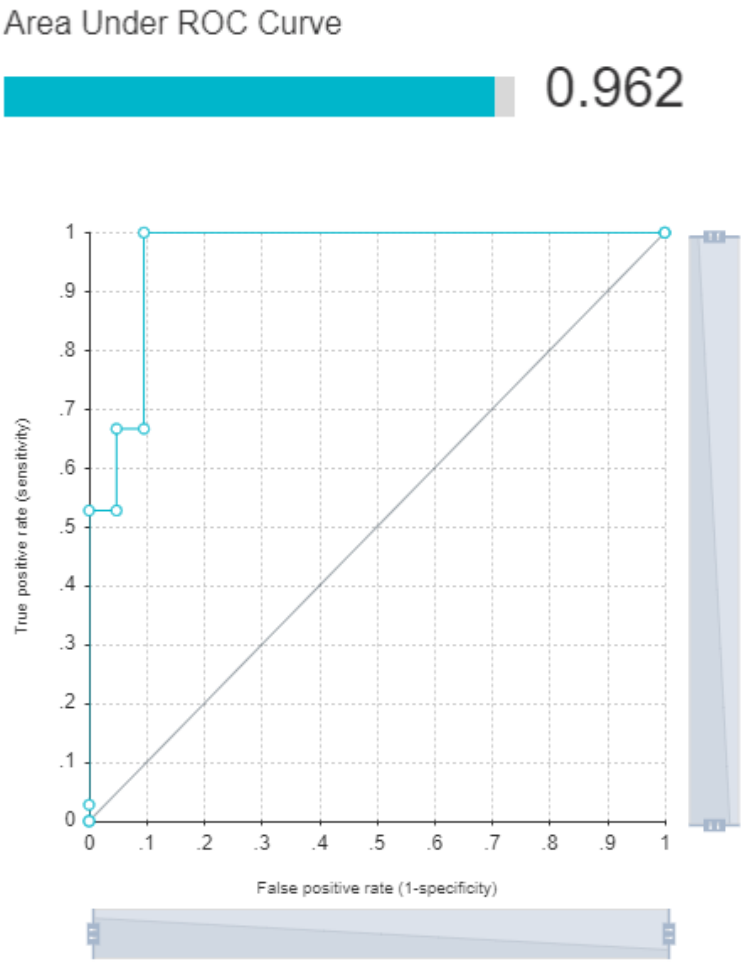
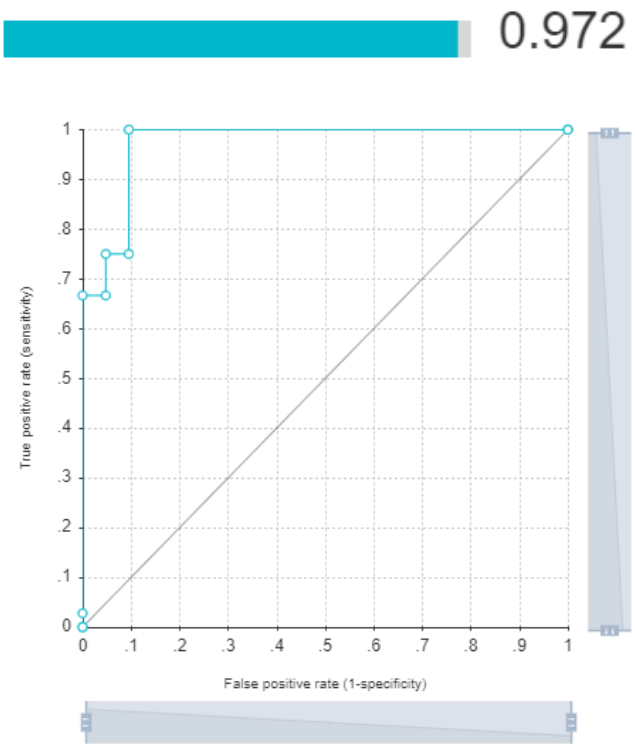


Fig 5: model 1 pipeline

Area Under ROC Curve



Model Evaluation Measures

	Holdout Score	**Cross Validation Score**
Accuracy	0.965	0.951
Area Under ROC Curve	0.972	0.994
Precision	1.000	0.928
Recall	0.905	0.948

Model Evaluation Measures

	Holdout Score	Cross Validation Score
F_1 Measure	0.950	0.936
Average Precision	0.968	0.992
Log Loss	0.287	0.132

Fig: 6 Pipeline

Conclusion

In carrying out this project, the relevance of machine learning cannot be over looked especially if any human based activities requires a high rate of accuracy in prediction and detection. Machines are gradually becoming more efficient as more efficient algorithms are developed. The overall expectation is that, accuracy rate should be less than minus 1 (-1) at every point when it has to do with life threatening diseases.

References

Tahmooresi, M., Rad, B.B., Afshar, A., Nowshath, K.B., and Bamiah, M.A (2018), Early Detection of Breast Cancer using Machine Learning Techniques, 10, JTECE, 21-27

Abdel-Rahman, O (2015), Prostate Cancer Incidence and Mortality In Relationship to Family History of Prostate Cancer: Findings from the PLCO Trial, Clinical Genitourinary Cancer, 17(4) e837-e844

Peters, U (2009), Screen for Rare Alleles by Deep Resequencing of Colorectal Cancer Cases, CDA

Ippolito, P.P., https://towardsdatascience.com/understanding-cancer-using-machine-learning-84087258ee18 accessed on 21st October, 2019.

Agarap, A.F.M., (2018), On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset, ICMLSC, arxiv.org/pdf/1711.07831.pdf accessed on 21st October, 2019

Kourou, K, Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., and Fotiadis, D.I (2015) Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, 13, 8-17

Huyghe et al (2019), Discovery of Common and Rare Genetic Risk Variants for Colorectal Cancer, 51 (1), Nature Genetics, 51, 76-87

Salaria, S (2019), Breast Cancer Detection Using Machine Learning, https://medium.com/datadriveninvestor/breast-cancer-detection-using-machine-learning-475d3b63e18e, accessed on 30th October, 2019

Chen et al (2018), An Augmented Reality Microscope for Real-time Automated Detection of Cancer, https://drive.google.com/file/d/1WRBCqJItaGly-9PDSMlwQ5Ldhc8lB0lf/view accessed on 22nd October, 2019

Danaee, P., Ghaeini, R., and Hendrix, D.A (2017), A Deep Learning Approach For Cancer Detection And Relevant Gene Identification, 219-229 https://pdfs.semanticscholar.org/9cf6/62e3bacaa81dcb1ec3867c730fe6be2650e5.pdf?_ga=2.87191256.1947193168.1576638617-1419913009.1571136569 Viewed on 26th October, 2019