The State-of-Art Machine Learning In Bioinformatics

Mujtaba Shaikh
Ramniranjan Jhunjhunwala College
Department of Data Science and Artificial Intelligence
mujtabashaikh451@gmail.com

ABSTRACT

A branch of artificial intelligence machine learning that provide enormous variety of statistical and probabilistic method that enables the computers to "learn" from historical data. It is widely use in health care domain we use this technology to predict/detect the breast cancer prognosis and detect the cancer from enormous, nosy or complicated bioinformatics dataset. The application of machine learning models bring to us accurate predictions on complex measurements. We use machine learning model to prediction of survival time in breast cancer on the basis of bioinformatics data set. The paper has discussion on the problem on breast cancer with knowledge mixture of bioinformatics and machine learning the decision is made with different attribute like clump thickness, uniformity size ,mitosis etc. The machine learning provide us different set of algorithms we use support vector machine(SVM), Logistic regression and random forest to get the promising result these mode achieve most accurate survival prognosis results.
The paper is based on machine learning algorithms that aim to build a model that accurately differentiate between benign and malignant breast tumours. The cross validation for accuracy of the model demonstrate the best performance on the breast cancer data. A python work flow has been developed and improvement discuss in this paper.

All code of analysis available at http://www.github.com/mujtabashaikh/ML-Bio

I. BACKGROUND

The data is heterogeneous present in different biological terms. However it is complicate to distinguish tumors to even expert with the help of modern technologies such as immunohistochemistry, DNA, or RNA hybridization. Now a days there is an intensive work and vigorously development of new knowledge base diagnostic methods for tumors detection with extended use of tools of bioinformatics, computer science, biostatics, and machine learning [1].

Cytological parameter: To understand breast cancer we must have knowledge of cytology (branch of biology the structure and function of cell) which allow us to understand data and biological terminologies (e.g., clump thickness, bare nuclei, Benign and malignant etc.).How tumors spread in body? Tumors occur when cells divide and grow excessively in the body. Normally, the body controls cell growth and division. New cells are created to replace older ones or to perform new functions. Cells that are damaged or no longer needed die to make room for healthy replacements [4].
The proportion to which epithelial cell is clustered were mono- or multilayered (clump thickness),
coherence of fringe cell of the epithelial cell cultured (marginal adhesion).the diameter of the population of immense epithelial cells relative to erythrocytes, the correspondence of single epithelial nuclei that were vacant by surrounding cytoplasm(bare nuclei),blandness of nuclear

chromatin, normal nucleoli, infrequent mitoses, uniformity of epithelial cell size, and uniformity of cell shape.

Eleven cytological parameter of breast cancer are measured by the class **benign** (the tumor is located only in one particular spot without spreading to correspondence epithelial cell) and

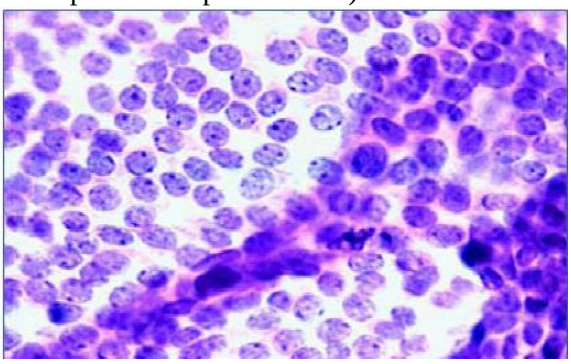

Fig 1. FNA results for benign tumor under the microscope

malignant (tumors are cancerous and can spread to correspondences tissues as they are cancerous). The cancer cell goes under mitosis phase and travel throughout the body to build new tumor

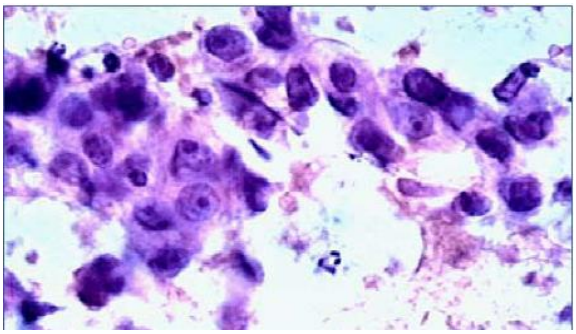

Fig 2. FNA results for malignant tumor under the microscope

Machine Learning: The use machine learning model for classification in breast cancer data deliver more evidence in cancer studies for prediction and prognosis of cancer to be tested more accurately and quickly, in short period of time. About 40% of all Machine learning studies on breast cancer prediction [3]. There are a variation of ML model e.g. K-nearest neighbor, Logistic regression, Random forest and support vector machine and etc. we provide validation with K-fold validation in scope of generality and better accuracy and validation.

II. INTRODUCTION

Breast cancer (BC) is the malignant tumor that is located in the epithelial cells of breast. As malignant is cancerous cell it spread throughout the all body and clustered a clump (clump thickness).it is commonly occur in women but also occur in men as well. According to research during their life 8% of women are prognosis with Brest cancer. Next to lung cancer Brest cancer is second most common reason of death. Every year one million women are newly diagnosis by with Breast cancer. Discovery of Brest cancer might difficult to detect .Due to absence of symptoms, after some clinical test, the accurate diagnosis should have the ability to differentiate the benign and malignant tumors [5]. A good detection provide low false positive (FP) rate and low false negative (FN) rate [1].

Bioinformatics has leverage to develop methods and software tools for understanding biological data, specifically when the data set is huge and complicated, In other words bioinformatics is the combination of biology, computer science, information engineering and bio-statistics. It has its own detailed "pipeline" that are repeatedly used mainly in the field of genomic (field of biology focusing on the structure, function, evolution, mapping, and editing of genomes), the common use of bioinformatics in identification of genes, nucleotide polymorphism (SNPs), unique adaptation etc.[6]

Machine learning is the technique employ for building the statistical model, I use here the supervised learning (labeled data) method. The relation between bioinformatics and machine learning is not recent, it had been used for decades to classify tumors and other malignancies, predict DNA – sequence, RNA-Sequence and drug-discovery etc. machine learning provide an ability to work rationally with the set of given data.

The aim of this paper to demonstrate the application of machine learning in bioinformatics I use different model for best accuracy to determine which model is best suited for data .I use Wisconsin breast cancer dataset here. To predict the patients have benign or malignant, the model performance will be judge by accuracy of the model and training and testing of the model evaluated in term of K-fold cross validation.

III. RELATED WORK

Cancer becomes most dangerous decease in all over the world. A number of research and studies has been done to diagnosis (chemotherapy) and detection of cancer, several of them used mammography (X-rays to examine the human breast for diagnosis and screening) images and the issue is that images can be miss about 15% of breast cancer [5]. Few techniques are more discrete and used genome or phenotypes to do classification [3, 6, 7]. To predict breast cancer some researcher use ensemble machine learning model [9] and Relevance Vector Machine (RVM)[10].The k-nearest neighbor algorithm most used in machine learning [9].

Different diagnosis techniques were developed for Brest cancer, the accuracy of many of them was evaluated using the dataset taken from Wisconsin breast cancer database [3, 12].for example, in [10] the support relevance machine method performance more than 90% for cancer breast cancer prediction.

IV. ALGORITHM

Algorithm for this approach I use google colab(64-bit) for python programming. Following algorithmic steps, we follow:

1) Import all the modules for feature selection, data splitting, ML models, accuracy score, confusion matrix, classification report and some other required modules.
2) Load the Wisconsin Breas Cancer dataset.
3) Perform Exploratory Data Analysis to preprocessing data.
4) To understand data apply visualization techniques for better understanding the dataset.
5) Divide the datasets as feature and class.
6) Check the significant features for prediction of class using
7) Split the dataset in two training and testing set respectively.
8) Build a various machine learning models using bagging techniques with k-fold (k=3) cross-validation with different estimation trees.
9) Print accuracy and classification reports of different models as comparing the true and predicted class
10) Plot confusion matrices of different models comparing the true and predicted class.

V. DATA DESCRIPTION

Wisconsin Breast Cancer dataset that is use for demonstration in donated by 'University of California, Irvine (UCI)'. There are elven attributes (sample code number, clump thickness, Uniformity of cell size, Uniformity if cell shape, Marginal Adhesion, Single epithelial cell size, Bare Nuclei, Bland chromatin, Normal nucleoli, Mitoses and class (2 for benign and 4 for malignant) and there are 16 missing attribute values represented by 'question mark (?)', The class attribute of benign is 457(65.5%) and malignant is 241(34.5%).

```
Total Number of Sample :698

***count of benign and malignant***
2    457
4    241
Name: class, dtype: int64
```

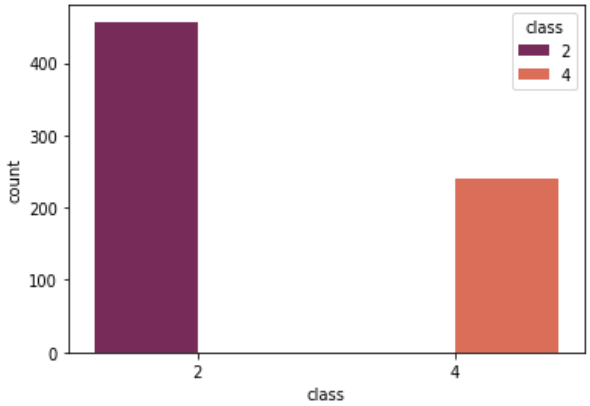

Fig 3. Count plot of Wisconsin Brest Cancer Dataset for Benign tumor (2) and Malignant tumor (4).

Each cytological parameter (attribute) of breast cancer FNAs (fine-needle aspirates) were estimated on the scale of 1 to 10, with 1 being the closest to benign and 10 the most anaplastic[4].

VI. PROCEDURE AND METHOD

The preprocessing of data is developed in software module in Python (version 3.8).Pipeline of machine learning for prediction and prognosis of Brest cancer dataset, approach for building ML model is consist of several steps(Figure 4).

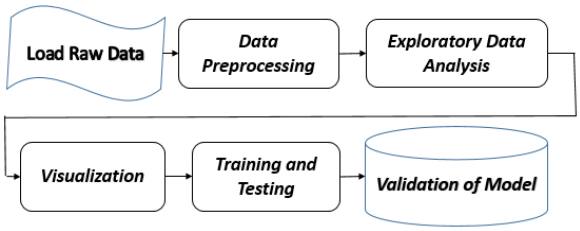

Fig 4.workflo of building ML model

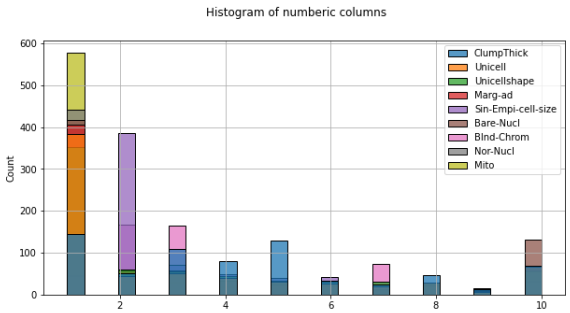

Fig 5.Histogram of cytological attributes

Initially the data is present in raw format, before emplace for testing and training of data we need to clean the data which contained 16 missing we replace the missing value with median of the columns attribute and remove the outliers from Brest cancer dataset to catch the outliers. Here I use visualization technique (box & whisker plot).

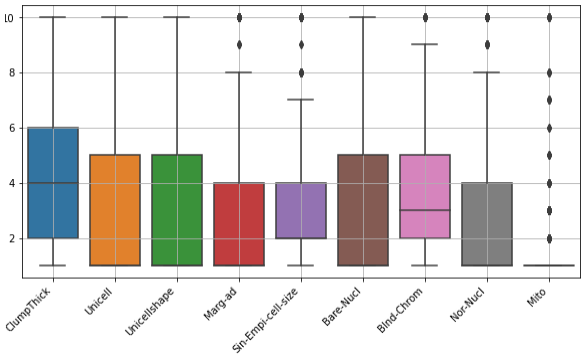

Fig 6.Box & whisker plot of cytological attributes

Plot heat map to display the correlation of cytological attributes.

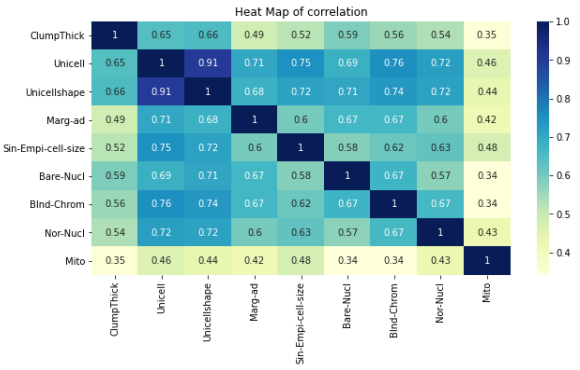

Fig 7.Heatmap of co-relation

After removing the anomaly (unexpected value) and deal with outlier we handover data for training and testing.

VII. EXPERIMENT RESULTS

Logistics Regression: Logistic regression is a statistical machine learning model that uses logistic function to predict the outcome values on and tries to makes a logarithmic line that separates between them. Logistic regression gives 66.39 % accuracy in our approach.

$$p = \frac{1}{1 + e^{-y}}$$

x Where a and b are the model parameters

Support Vector Machine (SVM): Support vector machine (SVM) is a discriminative classifier characterized by an isolating hyper plane. In two dimensional space this hyper plane is a line separating a plane in two sections where in each class lay in either side. SVM gives 97.33% accuracy with linear kernel.

K-Nearest Neighbor (KNN): KNN is very useful for a large dataset do not use mathematical analysis. In the worst scenario, KNN needs more memory to check all data sets. Here we used k=5 the accuracy of KNN is 97.54%

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

x Where x_i, y_i are coordinates of two point.

Random Forest: Random forests is an ensemble machine learning algorithm of decision tree. Random forest build a set of decision trees based on random chosen samples gets expectation from each tree and choose the best prediction results from the voting of each tree. In our approach, it gives 90.91% accuracy of predictions.

Accuracy: It is an percentage of correct predictionsfor the test data.It can be calculated easily by deviding the number of correct predictions by the number of total prediction

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{all predictions}}$$

Fig 8.Formula for calculating accuracy

Machine Learning Models	*Accuracy*
Logistic Regression	66.39%
Supper Vector Machine	97.33%
K-Nearest Neighbor	97.54%
Random Forest	99.79 %

Table1. Accuracy of Machine Learning Models

Confusion Matrix A confusion matrix is an outline of prediction. The number of accurate and inaccurate predictions are précised with count values and broken down by each class. The confusion matrix is the methods in which ML model is confused when it is predicted. It gives us intuition not only about the inaccuracies of classifier but also tells about the types of errors in which class.

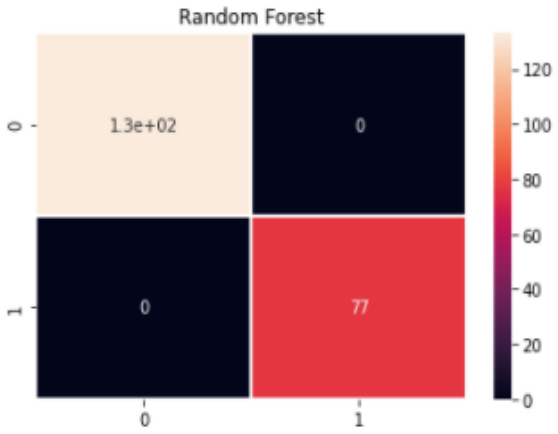

Fig 8.Heatmap of random forest confusion matrix

Cross Validation: Cross validation is a re-sampling method used to evaluate machine learning models on a constrained data samples. There is number of validation technique I use k-fold were k is chosen as 10-fold. The general strategies follow in crossover validation is: x Mix the dataset arbitrarily. Loading Breast Cancer Dataset with Classes (698) Feature Importance with Extra Tree Classifier Normalization Stage: Standard Scaling Classification of Breast Cancer using Ensemble Machine Learning Approach Cross Validation with 10-folds Summarize results of each group to utilizing the model accuracy.

VIII. CONCLUSION AND FUTURE WORK

After compare different machine learning algorithms, which are like support vector machine, k-nearest neighbor random for etc. On Wisconsin Breast Cancer dataset the best suited algorithm is random forest which is precise and reliable, The accuracy of random forest is a high efficiency(99.79%) , however I also notice K-nearest neighbor has pretty good accuracy(97.54%) as well, if the data set is larger the KNN time for detection increase.

References

1. Mihaylov, Iliyan; Nisheva, Maria; Vassilev, Dimitar. 2019. "Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies" *Information* 10, no. 3: 93. https://doi.org/10.3390/info10030093

2. Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc Natl Acad Sci U S A*. 1990;87(23):9193-9196. doi:10.1073/pnas.87.23.9193

3. M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, 2018, pp. 1-4, doi: 10.1109/EBBT.2018.8391453.

4. Joyce AP, Zhang C, Bradley P, Havranek JJ. Structure-based modeling of protein: DNA specificity. Brief Funct Genomics. 2015 Jan;14(1):39-49. doi: 10.1093/bfgp/elu044. Epub 2014 Nov 19. PMID: 25414269; PMCID: PMC4366589.

5. P. Baldi, S.R.B., Bioinformatics: The machinelearning approach. 2 ed, ed. S.r.B. Pierre Baldi, 2001.

6. N. Bhatia, "Survey of Nearest Neighbor Techniques", International Journal of Computer Science and Information Security, Vol. 8, No. 2, 2010.

7. A. Francillon, P.R., "Smart Card Research and Advanced Applications": 12th International Conference, CARDIS 2013, Berlin, Germany, 2013.

8. A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," 2010 5th International Symposium on Health Informatics and Bioinformatics, Ankara, Turkey, 2010, pp. 114-120, doi: 10.1109/HIBIT.2010.5478895.

9. Naveen, R. K. Sharma and A. Ramachandran Nair, "Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models," 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), Bangalore, India, 2019, pp. 100-104, doi: 10.1109/RTEICT46194.2019.9016968.

10. B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5, doi: 10.1109/ICCIC.2016.7919576.

11. S.K. Prabhakar, H. Rajaguru, "Performance Analysis of Breast Cancer Classification with Softmax Discriminant Classifier and Linear Discriminant Analysis", In: Maglaveras N.,Chouvarda I., de Carvalho P. (eds) Precision Medicine Powered by pHealth and Connected Health. IFMBE Proceedings, vol 66. Springer, Singapore, 2018.

12. . Z. Zhou, Y.J., Y. Yang, S.F. Chen, "Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles Artificial Intelligence", Medicine Elsevier, Vol. 24, pp. 25-36, 2002.