CHANDIGARH UNIVERSITY

GHARUAN, PUNJAB

Leukemia Detection using Python and Machine Learning

A STEP TOWARDS THE FUTURE

NAKUL MAGOTRA 17BCS3872

CSE IBM BDA 2

AKWINDER KAUR

8 November 2019

TABLE OF CONTENT

ABSTRACT

Abstract: Leukemia is a cancer of the blood and bone marrow, spongy tissue secretes the bones where blood cells are made. Acute myeloid leukemia (AML) is one of the most common types of leukemia in adults. The signs and symptoms of leukemia are non-specific in nature and are also comparable to symptoms of other interpersonal disorders. Manual microscopic examination of a stained blood smear or bone marrow aspirate is the only method for effective diagnosis of leukemia. But this method is time consuming and less accurate. In this paper, a technique for the automatic detection and classification of AML in blood smears is presented. Logistic regression, Gaussian nave base, decision tree classifier, support vector machine model used for classification.

Keywords: automatic leukemia detection, acute lymphoblastic leukemia, lymphocyte image segmentation, machine learning.

I. INTRODUCTION

Microscopic analysis of peripheral blood smear could be an important step within the detection of malignant neoplastic disease. However, this sort of sunshine microscopic assessment is time overwhelming, inherently subjective, and is ruled by hematopathologists clinical acumen and skill. to bypass such issues, an economical laptop power-assisted methodology for measure of peripheral blood samples is needed to be developed. during this paper, efforts square measure so created to plot methodologies for machine-controlled detection and sub-classification of Acute lymphocytic leukemia (ALL) exploitation image process and machine learning strategies. selection of applicable segmentation theme plays a significant role within the machine-controlled unwellness recognition method. consequently, to phase the conventional mature white cell and malignant lymph cell pictures into constituent morphological regions novel schemes are projected. so as to form the projected schemes viable from a sensible and real–time stand purpose, the segmentation downside is self-addressed within the supervised framework. These projected strategies square measure supported Gaussian Naive Bayes, Logistic regression, Decision Tree Classifier, Random Forest Classifier**,** and SVM field modelling, wherever the segmentation downside is developed as picture element classification, picture element bunch, and picture element labelling downside severally. A comprehensive validation analysis is given to judge the performance of machine-controlled classifier schemes against manual results provided by a panel of hematopathologists. it's discovered that morphological elements of traditional and malignant lymphocytes take issue considerably. To mechanically acknowledge lymphoblasts and observe dead peripheral blood samples, an economical methodology is projected. Morphological, textural and color options square measure extracted from the metameric nucleus and protoplasm regions of the white cell pictures. An ensemble of Random Forest classifiers shows highest classification accuracy of 96.5% as compared to other individual models. These strategies embrace lymph cell image segmentation, nucleus and protoplasm feature extraction, and economical classification. To subtype malignant neoplastic disease blast pictures supported cell lineages, an improved theme is additionally projected and also the results square measure related to thereupon of the flow cytometer. exploitation this theme the origin of blast cells i.e. bodily fluid or myeloid may be determined. An ensemble of call trees is employed to map the extracted options of the leukemic blast pictures into one in all the 2 teams. every model is studied severally and experiments square measure conducted to judge their performances. Performance measures i.e. accuracy effectiveness of the projected machine-controlled systems thereupon of ordinary diagnostic procedures.

A. LEUKEMIA Leukemia is a group of heterogeneous blood-related cancers, differing in its aetiology, pathogenesis, prognosis, and response to treatment (Bain, 2010). Leukemia is considered as a serious issue in modern society, as it affects both children and adults and even sometimes infants under the age of 12 months. In children, leukemia is considered as the most common type of cancer, while, in adults, the World Health Organization

report shows that leukemia is one of the top 15 most common types of cancer (Kampen, 2012). To better understand leukemia, the next sections are dedicated to the discussion of the blood cells lineage, types of leukemia, diagnostic methods currently in use, treatments options as well as prognostic factors.

B. TYPES OF LEUKEMIA Lab tests help the doctor find out the type of leukemia that you have. For each type of leukemia, the treatment plan is different. Acute and Chronic Leukemias Leukemias are named for how quickly the disease develops and gets worse:

Acute: Acute leukemia typically develops quickly. the amount of malignant neoplastic disease cells will increase apace, and these abnormal cells don't do the work of traditional white blood cells. A bone marrow take a look at might show a high level of malignant neoplastic disease cells and low levels of traditional blood cells. individuals with cancer of the blood might feel terribly tired, bruise simply, and obtain infections usually.

Chronic: Chronic leukemia sometimes develops slowly. The malignant neoplastic disease cells work nearly likewise as traditional white blood cells. individuals might not feel sick initially, and therefore the initial sign of ill health is also abnormal results on a routine biopsy. as an example, the biopsy could show a high level of malignant neoplastic disease cells. If not treated, the malignant neoplastic disease cells could later displace traditional blood cells

C. Myeloid and Lymphoid Leukemia's

Leukemia's are also named for the type of white blood cell that is affected:

Myeloid: Leukemia that starts in myeloid cells is termed myeloid, myelogenous, or leukaemia.

Lymphoid: Leukemia that starts in lymphoid cells is termed lymphoid, lymphoblastic, or lymphocytic leukemia. lymphoid leukemia cells might collect within the humour nodes, that become swollen.

D. Four Most Common Types of Leukemia

Acute myeloid Leukemia (AML) affects myeloid cells and grows quickly. Leukemic blast cells collect within the bone marrow and blood. About 15,000 Americans are going to be diagnosed with AML in 2013. Most (about 8,000) are going to be sixty five or older, and concerning 870 kids and teenagers can get this illness.

Acute lymphoblastic Leukemia (ALL) affects body fluid cells and grows quickly. Leukemic blast cells generally collect among the bone marrow and blood. More than half dozen Americans are diagnosed with beat 2013. Most (more than 3,600) are youngsters and youths.

Chronic myeloid leukemia(CML) affects myeloid cells and frequently grows slowly initially. Blood tests show a rise within the variety of white blood cells. There is also atiny low variety of leukemic blast cells within the bone marrow. About 6,000 Americans are going to be diagnosed with CML in 2013. virtually 0.5 (about 2,900) are going to be sixty five or older, and solely regarding a hundred and seventy kids and teenagers can get this malady.

Chronic lymphocytic leukemia(CLL) affects lymphoid cells and typically grows slowly. Blood tests show a rise within the variety of white blood cells. The abnormal cells work nearly similarly because the

traditional blood corpuscle. About 16,000 Americans are diagnosed with CLL in 2013. Most (about 10,700) are sixty five or older. This illness nearly ne'er affects youngsters or teens. Other, less common styles of cancer of the blood can account for over 6,000 new cases in 2013.

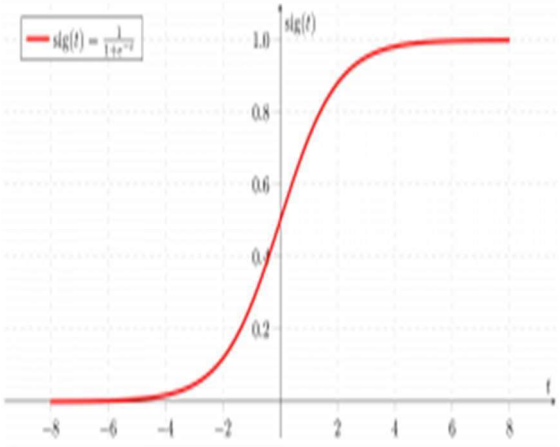
II. EXPERIMENTAL PROPOSED METHOD

A. Logistic Regression

Logistic regression is essentially a supervised classification algorithmic rule. during a classification drawback, the target variable (or output), y, will take solely distinct values for given set of options (or inputs), X.

Contrary to widespread belief, supply regression could be a regression model. The model builds a regression model to predict the likelihood that a given data entry belongs to the class numbered as "1". a bit like linear regression assumes that the information follows a linear function, logistic regression models the data using the sigmoid function.

$$g(z) = \frac{1}{1+e^{-z}}$$



Logistic regression becomes a classification technique only if a call threshold is brought into the picture. The setting of the threshold value could be a vital side of logistic regression and relies on the classification drawback itself.

The decision for {the price | the worth} of the threshold value is majorly suffering from the values of exactness and recall. Ideally, we wish each exactness and recall to be one, however, this rarely is that the case. In case of a Precision-Recall trade-off we use the following arguments to decide upon the threshold: -

1. **Low Precision/High Recall:** In applications wherever we would like to scale back the number of false negatives while not essentially reducing the amount false positives, we decide a call price that includes a low precision of exactitude or high value of Recall. as an example, in a very cancer diagnosing application, we tend to don't need any affected patient to be classified as not affected while not giving a lot of heed to if the patient is being de jure diagnosed with cancer. this can be as a result of, the absence of cancer is detected by additional medical sickness however the presence of the disease can't be detected in an already rejected candidate.

2. **High Precision/Low Recall:** In applications where we tend to would like to chop back the number of false positives whereas not basically reducing the amount of false negatives, we tend to decide a selection a call that encompasses a high worth of precision or low value of Recall. as an example, if we've a bent to classifying customers whether or not they will react absolutely or negatively to a made-to-order advertising, we would like to be totally positive that the shopper will react completely to the promotion as a result of otherwise, a negative reaction can cause a loss of potential sale from the shopper.

B. Gaussian Naive Bayes

Naive Bayes is a statistical classification technique based on Bayes Theorem. It's one in all the only supervised learning algorithms. Naive Thomas Bayes classifier is the fast, correct, and reliable algorithmic rule. Naive Thomas Bayes classifiers have high accuracy and speed on giant data sets. Naive Thomas Bayes classifier assumes that the impact of a specific feature in an exceedingly class is freelance of different options. For instance, a loan applicant is fascinating or not depending on his/her financial gain, previous loan, and group action history, age, and placement. Even though these options are interdependent, these options are still thought of severally. This assumption simplifies computation, and that is why it's considered naive. This assumption is named class conditional independence.

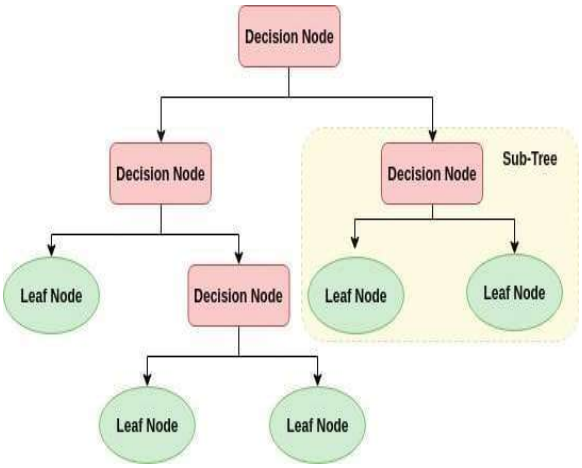
$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

1. ***P(h):*** the probability of hypothesis h being true (regardless of the data). this is often referred to as the previAous probability of h.
2. ***P(D):*** the probability of the data (regardless of the hypothesis). this is often called the prior probability.
3. ***P(h|D):*** the probability of hypothesis h given the data D. this is often called posterior probability.
4. ***P(D|h):*** the probability of knowledge d provided that the hypothesis h was true. this is often called posterior probability.

C. Decision Tree Classifier

A decision tree is a flowchart-like tree structure where an interior node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the result. The top node in AN exceeding call tree is understood is known root node. It learns to partition on the thought of the attribute value. It partitions the tree in an exceedingly algorithmic manner decision call partitioning. This flowchart-like structure helps you in higher cognitive process. It's visual image sort of a flow chart diagram that merely mimics the human-level thinking. that is why decision trees are straightforward to grasp and interpretative reaction will cause a loss

of potential sale from the patron.



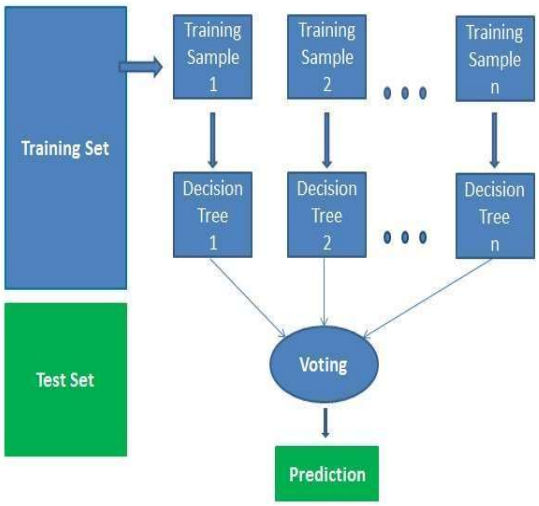
Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy.

D. Random Forest Classifier

Random forests are a supervised learning algorithm. It can be used for both classification and regression. This

algorithm is also the most flexible and easiest to use. A forest is made up of trees. It is said that the more trees there are, the stronger the forests are. Random forests create decision trees on randomly selected data samples, derive a prediction from each tree and select the best solution through voting. It also provides a very good indicator of the importance of convenience.

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning, where you sometimes combine different types of algorithms or the same algorithm to create more powerful predictive models. The random forest algorithm combines multiple algorithms of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "random forest". The random forest algorithm can be used for both regression and classification functions.



E. Support Vector Machines

The so-called support vector machines as SVMs are a supervised learning algorithm that can be used as support vector classification (SVC) and support vector regression (SVR) for classification and regression problems. It is used for small datasets because it takes too long to process. In this set, we will focus on SVC.

The basic intuition to develop here is that the more SV points from the hyperplane, the more likely it is to correctly classify points in its field or classes. SV points are very important in determining the hyperplane because if the position of the vectors changes then the position of the hyperplane changes. Technically, this hyperplane can also be called a margin planning hyperplane.

SVM is based on the idea of finding a hyperplane that separates features into different domains.

- **Linear Kernel**

The linear kernel is used when the data is linearly separated, that is, it can be separated using the same line. It is one of the most common kernels used. It is mostly used when a particular data set has a large number of attributes. One of the examples where there are too many features is text classification, as each alphabet is a new feature. So we mostly use linear kernel in Text Classification.

- **Radial basis function kernel (RBF)/ Gaussian Kernel:**

Gaussian RBF (Radial Basis Function) is another popular kernel method that is used more often in SVM models. The RBF kernel is a function whose value depends on the distance from the origin or some point. Gaussian Kernel is of the following format;

$$K(X_1, X_2) = exponent(-\gamma \|X_1 - X_2\|^2)$$

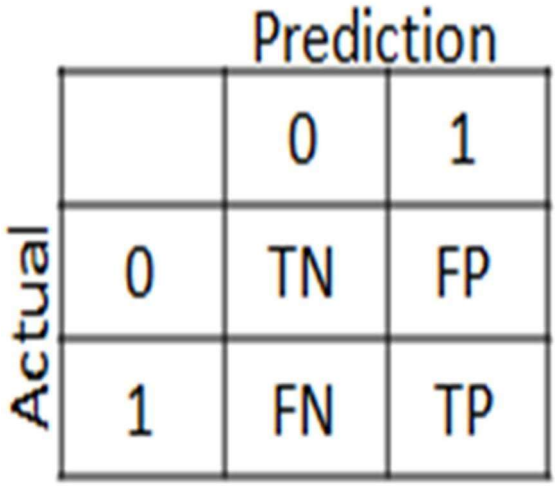
||X1 — X2 || = Euclidean distance between X1 & X2

Using the distance in the original space we calculate the dot product (similarity) of X1 & X2.

☐ **Confusion Matrix:**

In the field of machine learning and in particular the problem of statistical classification, a confusion matrix, also known as an error matrix, is a typical table layout that allows visualization of the performance of an algorithm, typically A supervised learning one (in learning unexpectedly) is usually called a matching matrix). Each row of the matrix represents instances in an approximate class while each column represents instances in the actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing the two classes (ie usually confusing each other).

It is a special type of contingency table, which has two dimensions ("real" and "approximate"), and the same set of "classes" in both dimensions (a combination of each class and dimension in the dimension table is a variable).



- ***False positive (FP)*** = a test result that falsely indicates that a particular condition or characteristic exists.
- ***True positive (TP)*** = sensitivity (also called true positive rate, or probability of detection in some areas) measures the proportion of true positivity that is correctly identified.
- ***True Negative (TN)*** = Specificity (also called real negative rate) measures the ratio of true negatives that are correctly identified.
- ***False Negative (FN)*** = a test result that indicates that a condition does not hold, whereas in fact, it does. For example, a test result that indicates that a person does not have cancer when the person actually does it.

- **Steps to work Procedure:**

1. In the first step, we import packages / libraries to make it easier to write programs.
2. Next, I will load the data, and print the first 7 rows of data.
3. Explore the data and count the number of rows and columns in the data set.
4. Count counts by creating plots and encoding hierarchical data.
5. Draw a pair plot and visualize the correlation.
6. First set the data for the model by dividing the data set into a feature data set also known as the independent data set (X), and the target data set is also known as the dependent data set (Y).
7. Redistribute the data, and scale the data.
8. Create a function to hold several different models (such as logistic regression, decision tree classifier, random forest classifier) to perform classification.
9. Create a model that includes all models, and look at the accuracy scores on each model's training data

to classify whether each patient has cancer.

10. Show the confusion matrices and accuracy of models on test data.
11. By accuracy and metric, apply the model that performed best on the test data.

III. EXPERIMENTAL RESULTS AND DISCUSSION

The dataset used in this paper consisted of 569 blood samples. For AML, we accessed the Kaggle to get the sample dataset of blood cancer which consists of 569 and 33 different parameters on which the analysis gets carried out.

- **Load the data**

	id	diagnosis	radius_mean	texture_mean
0	842302	M	17.99	10.38
1	842517	M	20.57	17.77
2	84300903	M	19.69	21.25
3	84348301	M	11.42	20.38
4	84358402	M	20.29	14.34
5	843786	M	12.45	15.70
6	844359	M	18.25	19.98

A sample of the first 7 rows of data

- **Visualize the count**

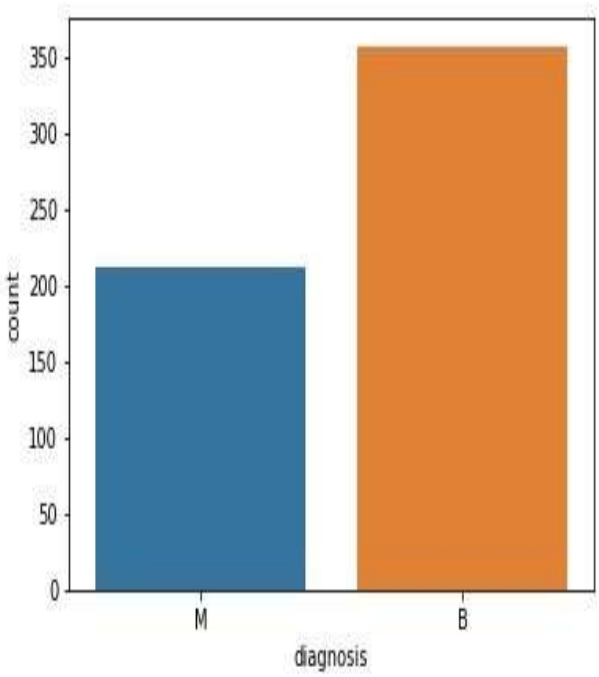
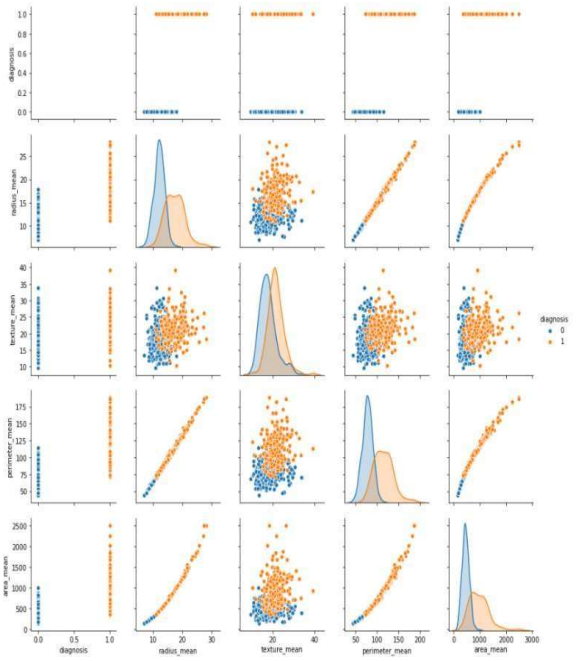
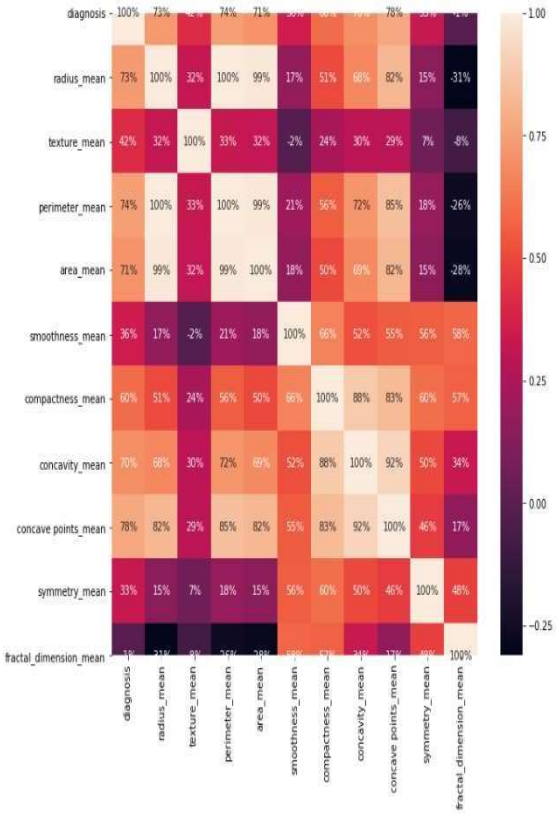


Chart displaying Malignant (cancerous) & Benign(non-cancerous) diagnosis

- **Create a pair plot**.



- **Visualize the correlation by creating a heat map**




```
[[86  4]
 [ 4 49]]
Model[0] Testing Accuracy = "0.9440559440559441!"

[[89  1]
 [ 5 48]]
Model[1] Testing Accuracy = "0.958041958041958!"

[[87  3]
 [ 2 51]]
Model[2] Testing Accuracy = "0.965034965034965!"

[[88  2]
 [ 3 50]]
Model[3] Testing Accuracy = "0.965034965034965!"

[[85  5]
 [ 6 47]]
Model[4] Testing Accuracy = "0.9230769230769231!"

[[84  6]
 [ 1 52]]
Model[5] Testing Accuracy = "0.951048951048951!"

[[87  3]
 [ 2 51]]
Model[6] Testing Accuracy = "0.965034965034965!"
```

- **Show the confusion matrix and accuracy for all of the models on the test data**

- **Show other ways to get the classification accuracy & other metrics**

```
Model  1
              precision    recall  f1-score   support

           0       0.95      0.99      0.97        90
           1       0.98      0.91      0.94        53

    accuracy                           0.96       143
   macro avg       0.96      0.95      0.95       143
weighted avg       0.96      0.96      0.96       143

0.958041958041958

Model  2
              precision    recall  f1-score   support

           0       0.98      0.97      0.97        90
           1       0.94      0.96      0.95        53

    accuracy                           0.97       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.97      0.97      0.97       143

0.965034965034965

Model  3
              precision    recall  f1-score   support

           0       0.97      0.98      0.97        90
           1       0.96      0.94      0.95        53

    accuracy                           0.97       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.96      0.97      0.96       143

0.965034965034965

Model  4
              precision    recall  f1-score   support

           0       0.93      0.94      0.94        90
           1       0.90      0.89      0.90        53

    accuracy                           0.92       143
   macro avg       0.92      0.92      0.92       143
weighted avg       0.92      0.92      0.92       143

0.9230769230769231

Model  5
              precision    recall  f1-score   support

           0       0.99      0.93      0.96        90
           1       0.90      0.98      0.94        53

    accuracy                           0.95       143
   macro avg       0.94      0.96      0.95       143
weighted avg       0.95      0.95      0.95       143

0.951048951048951

Model  6
              precision    recall  f1-score   support

           0       0.98      0.97      0.97        90
           1       0.94      0.96      0.95        53

    accuracy                           0.97       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.97      0.97      0.97       143

0.965034965034965
```

- **Print Prediction of Random Forest Classifier model and Print the actual values**

```
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 1 0 0 1 0 0
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]

[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 1 0 0 1 0 1
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]
```

Table 1: Accuracy Table

Method	Accuracy Percentage
Logistic Regression	94.4%
K Nearest Neighbour	95.8%
Support Vector Machine (Linear Classifier)	96.5%
Support Vector Machine (RBF Classifier)	96.5%
Gaussian Naive Bayes	92.3%
Decision Tree Classifier	95.1%
Random Forest Classifier	96.5%

IV. CONCLUSION

Make the prediction/classification on the test data and show both the Random Forest Classifier model classification/prediction and the actual values of the patient that shows rather or not they have cancer.

I notice the model, misdiagnosed a few patients as having cancer when they didn't and its misdiagnosed patients that did have cancer as not having cancer. Although this model is good, when dealing with the lives of others I want this model to be better and get its accuracy as close to 100% as possible or at least as good as if not better than doctors. So, a little more tuning of each of the models is necessary.

V. REFERENCES

[1] Cristianini, N., and J. Shawe-Taylor. "An Introduction to support vector machines and other kernel-based learning methods" New York: Cambridge University Press, 2000.

[2] Vapnik, V. N. "The Ature of Statistical Learning Theory" New York: Springer, 1995.

[3] A. Madabhushi, "Digital pathology image analysis: opportunities and challenges," Imaging in Medicine, vol. 1, no. 1, pp. 7– 10, 2009.

[4] A. N. Esgiar, R. N. G. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray, "Fractal analysis in the detection of colonic cancer images," IEEE Transactions on Information Technology in Biomedicine, vol. 6, no. 1, pp. 54–58, 2002.

[5] L. Yang, O. Tuzel, P. Meer, and D. J. Foran, "Automatic image analysis of histopathology specimens using concave vertex graph," in Medical Image Computing and Computer-Assisted Intervention—MICCAI 2008, pp. 833–841, Springer, Berlin, Germany, 2008.

[6] R. C. Gonzalez, Digital Image Processing, Pearson Education India, 2009.

[7] S. Liao, M. W. K. Law, and A. C. S. Chung, "Dominant local binary patterns for texture classification," IEEE Transactions on Image Processing, vol. 18, no. 5, pp. 1107–1118, 2009.

[8] J. C. Caicedo, A. Cruz, and F. A. Gonzalez, "Histopathology image classification using a bag of features and kernel functions," in Artificial Intelligence in Medicine, vol. 5651 of Lecture Notes in Computer Science, pp. 126–135, Springer, Berlin, Germany, 2009.

[9] H. S. Wu, J. Barba, and J. Gil, "Iterative thresholding for segmentation of cells from noisy images" Journal of Microscopy, vol. 197, no. 3, pp. 296–304, 2000.