International Journal of Scientific Research in Computer Science and Engineering **Research Paper**
Vol.10, Issue.6, pp.01-08, December (2022) **E-ISSN:** 2320-7639

Machine Learning Approaches for Prediction of various Cancer types

Sanjay Garag[1], Anupama S. Nandeppanavar[2*], Medha Kudari[3]

[1,2,3]Dept of MCA, KLE Institute of Technology, Hubballi, India

[*]*Corresponding Author: anupama.nandeppanavar@kleit.ac.in*

Available online at: www.isroset.org | DOI: https://doi.org/10.26438/ijsrcse/v10i6.18

Received: 16/Sept/2022, Accepted: 19/Nov/2022, Online: 31/Dec/2022

Abstract— Cancer is a prevalent disease that affects the people and an early diagnosis will expedite the treatment of this ailment. The Machine Learning is providing enormous contribution to the biomedical field. The main goal of this project is to build a model for predicting cancer using support vector machine classification algorithms. Compare the accuracy of different kernels and apply different parameters to one efficient kernel. Cancer is characterized as a heterogeneous disease of many different subtypes. The Cancer Disease Prediction contains the machine learning models like Random Forest Classifier, Support Vector Machine, K-Nearest Neighbor (KNN), K-Means Clustering, Decision Tree Algorithm and then the collected data is pre-processed using some machine learning techniques. Data divided into the training data and the testing data. Then the Machine Learning Algorithm applied to yield the significant results. The analysis with Decision Tree Algorithm gives the best results for predicting the type of the cancer by considering the symptoms that the patients are bearing. The system is developed to predict that the person is having a cancer or not before going for the lab tests.

Keywords— Random Forest, Support Vector Machine, K-Nearest Neighbor, K-Means Clustering, Decision Tree, Prediction, kernel

I. INTRODUCTION

Living Body is composed of trillions of cells and unique structures. They are the building blocks of the living beings and comprises of millions of functions. The cells contain the body's heredity from the ancestors and they can make the copies of themselves for the upcoming generations. If a person is suffering from any of the cells infected disease in a family then that disease will be transmitted to the later generation of the same family through genetic disorders or hereditary process. A disease is that attacks human body and won't allow them work in healthy and normal conditions. There are different types of disease that affects the human body. But cancer is an infectious and hereditary disease. Today the people are living in tight schedules and won't spend their valuable time for their health issues. If this could be the scenario then it would become even worse to overcome the problems that they face in the future. The infectious disease cancer is one of the major and most hazardous illness that spreads in the human body through cells. There are two types of tumors: Benign and Malignant tumor.

Machine learning has become common tool in health care now a day. It helps in getting untold, logical patterns and relationships in large data set. Statistical models and mathematical models play very important role in early detection of cancer. The numeric prediction provides numeric quantity as outcome but not discrete.

The proposed system "Cancer Disease Prediction Using Machine Learning" helps the cancer patients to identify their cancer disease easily without consulting the doctors directly. This reduces the time and cost of the patients. Using the predictive system, patients will know about their cancer disease condition and information about their disease. They can take the appropriate precautions as soon as possible and they can know the information about the type of cancer they are bearing. The proposed system includes User Interactions. The User provides the different inputs for the different diseases and these inputs are processed by the Machine Learning algorithms. Here the system can be identified as a layered architecture. Then the machine learning algorithm applied to process the user data. The database is maintained for the user.

The main Objective of the System is to predict the disease as per the symptoms. It provides quick analysis of the cancer status of the patients and helps in quick recovery. The System maintains the data that is entered by the patients, who get registered to seek precautions.

Rest of the paper is organized as follows, section One contains the introduction of Cancer Detection Prediction using machine learning algorithms, section Two contain the related work of the same, section Three contains the methodology of proposed along with the working model being presented. Section Four presents the results obtained using the proposed methodology. Section Five describes the discussions based on the results obtained and inferences drawn. Section Six concludes the research work.

II. Related Work

These are the researches related to the machine learning algorithms to predict the type of the cancer. The researches related to this area are discussed briefly.

Melak Akcay et al.[1] made an attempt to predict the survival rate of Stomach Cancer (Gastric Cancer). In the work the Machine Learning methods used are Logistic regression, support vector classification (SVC), XGBoost, Random Forest and Gaussian Naive Bayes (GNB) and multilayer perceptron (MLP). The most accurate algorithms for predicting the survival rate and metastatic stage of the malignancy for gastric cancer were GNB, XGBoost, and Random Forest.

Mou et al.[2] published the research paper to predict the leukaemia cancer by applying some of the different types of machine learning algorithms. They have used algorithms like Random Forest, Decision Tree and Logistic Algorithms to find whether the patient had developed cancer or not. The highest accuracy of 84.15% is provided by the Decision Tree algorithm.

Nasser and Abu-Naser [3] have published research paper to find about lung cancer cells based on human characteristic, symptoms and information. This study discovers lung cancer in the human body with artificial neural network models.

Daqqa et al.[4] proposed the study to predict leukaemia in relation with blood properties. The classification algorithms used for the predicting cancer are mainly SVM, Decision Tree and K-NN. Among all the algorithm results, the Decision tree algorithm provides the best accuracy score by a 77.3% compared with the two different algorithms.

Maria et al.[5] tried to classifying different types of leukaemia or detecting whether a patient has leukaemia. This paper presents a comparative study of different algorithms like SVM, K-NN and DL algorithms.

Ada and Rajneet Kaur [6] the authors focus on early stage detection. The survival rate of patient is also an output. Neural network and feature extraction are used.

Charles Edeki [7] discussed different data mining for breast cancer with different statistical learning techniques for survival analysis. The obtained results were varied regarding which algorithm performed optimal. It showed that the size, high dimensionality of data representation and cleanliness of the dataset affected the performance of each algorithm.

Reeti Yadav [8] presented data mining approaches like classification, clustering and prediction to recognize potential cancer patients. Before going for clinical and lab tests which is cost and time consuming, this research helps in detection of a person's predisposition for cancer.

Arunachalam [9] used K-means Clustering, Marker-Controlled Watershed segmentation, and HSV colour-based segmentation for image segmentation. The SVM classifier is used to further classify leukemia into its different types. This paper identified leukemia and determined its various types.

Nallasivan and Sivaranjani [10] proposed a method for segmentation that initially executes an adequate edge that picks the threshold contingent upon the feeling of the item and setting pixel. This was used in detection of lung cancer from CT scan images.

III. Methodology

The application serves as a user's online resource and support. It provides customer with immediate cancer advice via machine learning model that has been trained to forecast in the online web application. The system will provide cancer prediction with various cancer details. The system also allows users to share health-related issues with their symptoms for cancer prediction. Machine learning models are applied to the user input to provide the most desirable results as a disease that may be associated with patient data. Based on their symptoms the patient gets their prediction results. The patients can select the doctor and also take the appointment from the doctor. The system confirms the appointment for the patients with respective doctor. Rescheduling of appointment is also available. The system will allow the user to view the doctor's details and the hospital details in nearby locality.

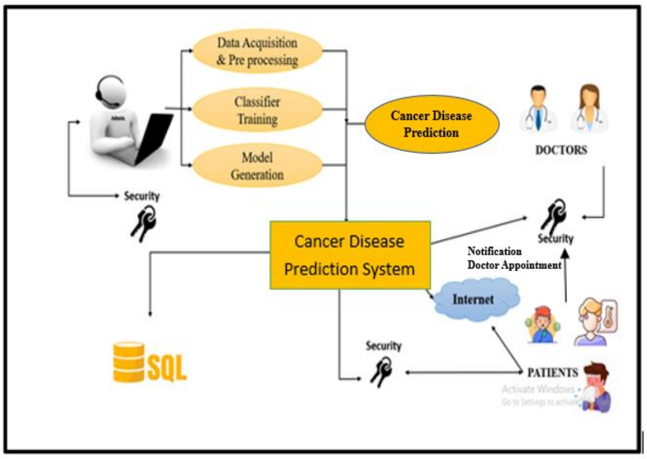


Figure 1. System architecture

In Fig. 1 the system architecture along with different module interactions is shown. The System is divided into different modules like admin, doctors and patients. The user will login into the system and select prediction section in the user module. After selecting prediction, the user will enter the symptoms he is bearing and the prediction result will be displayed to user. The displayed result will contain the type of testing for respective cancer disease. The user can also make an appointment to the doctor and get the appointment notifications to the registered email-id.

The Prediction System “Cancer Disease Prediction” is used for the all the users to know about their health condition and the predict which type of Cancer Disease they are suffering from. Some people avoid taking the cancer scan report due to high cost. This Prediction application will help users save money and time. It provides information about the type of the cancer disease and to get the appointment for consultation of a specialized oncologist. So, it becomes easy to detect the disease in the early stage. The user can take precautions or treatments related to type of Cancer Disease. Users can get prediction results in one click.

A machine learning algorithm is the method by which an Artificial Intelligent (AI) system accomplishes its objective, typically the method by which it predicts the output value from a set of input data. Fig. 2 depicts the steps to build a machine learning model. The first step is to collect the data. The second step is to pre-process the collected data. The pre-processing refers to the cleaning of raw data by filling missing values. The third step is the data encoding, where the categorical values are converted to numerical values so that it can be fitted to a machine learning model. The next step is to split the dataset into training and testing sets. The training set is used to train a model and testing set is used to estimate the performance in the form of accuracy scores. The model is tested using the algorithms: SVM, KNN, Random Forest, decision tree and k-means clustering. Based on the models, the prediction results are obtained, which are presented to the user and the doctor.

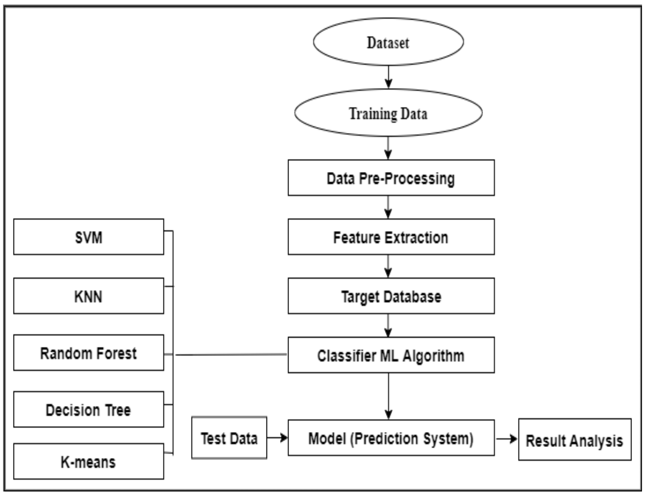
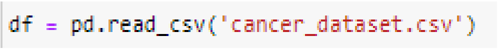


Figure 2. Machine learning model

A. Dataset

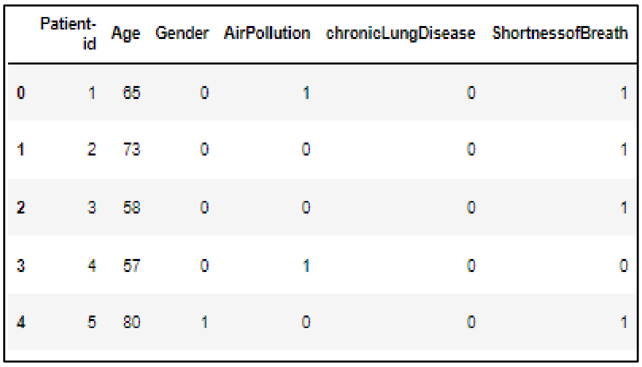
The dataset used for the investigation is the Cancer Disease Prediction dataset. There are 5000 rows in this dataset with 33 columns of data. The attribute Cancer, has a value of 1 or 2 or 3. In the attribute, a value of 1 represents Lung Cancer, 2 indicates Blood Cancer and 3 represents Stomach Cancer. Only 4265 of the rows are the value of 1(Lung Cancer), 614 of the rows are Blood Cancer and 121 of the rows are Stomach Cancer.

The data accessing function to read the dataset (.csv file) into the jupyter notebook is given below:



```
df = pd.read_csv('cancer_dataset.csv')
```

Sample data set is given in Fig. 3 which includes attribute like age, gender, likelihood of air pollution, chances of a person suffering from lung disease and a person suffering from shortness of breath. The attribute values have numeric values of 1 or 0 indicating presence or absence of disease.



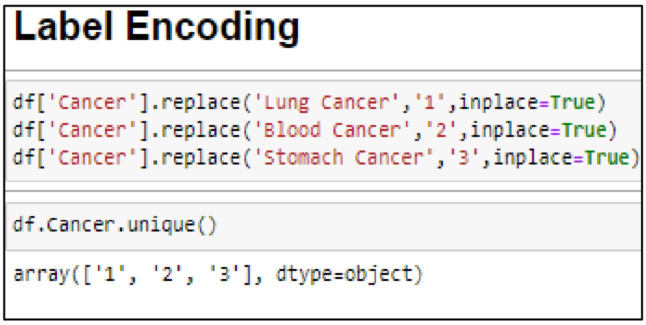
	Patient-id	Age	Gender	AirPollution	chronicLungDisease	ShortnessofBreath
0	1	65	0	1	0	1
1	2	73	0	0	0	1
2	3	58	0	0	0	1
3	4	57	0	1	0	0
4	5	80	1	0	0	1

Figure 3. Dataset

B. Label Encoding

Label encoding is a method of turning textual literals in a dataset into integer values that a computer can understand. Integers must be converted from strings because computers are normally trained on numbers. There has been only one column with string data included in the obtained dataset.

In this label encoding method of data pre-processing, the numbering considered for cancer disease are 1 for Lung Cancer, 2 for Blood Cancer and 3 for Stomach Canceras shown in Fig. 4.



Label Encoding

```
df['Cancer'].replace('Lung Cancer','1',inplace=True)
df['Cancer'].replace('Blood Cancer','2',inplace=True)
df['Cancer'].replace('Stomach Cancer','3',inplace=True)
```

```
df.Cancer.unique()
```

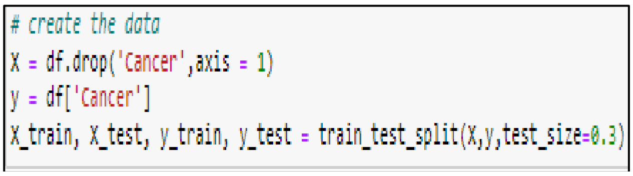
```
array(['1', '2', '3'], dtype=object)
```

Figure 4. Label Encoding

C. Data Splitting

Data splitting step includes splitting data set into two sections namely training and testing set.
Train set: The data is stored as the train set which will be entered into the model. In other words the stored data provides knowledge for the mathematical model.

Test set: The test set consists of data used to validate and test the trained model. It demonstrates how well the mathematical model predicts illogical events and how often it is to do so. A variety of indicators can be used to assess the performance of our model (including precision, recall, and accuracy).



```
# create the data
X = df.drop('Cancer',axis = 1)
y = df['Cancer']
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3)
```

Figure 5. Data splitting

Fig. 5 presents the code snippet for splitting the dataset into two categories like training data and testing data. The training data size will be 70% of the original data set and the testing size of the dataset will be of 30%. This is done using the function called train_test_split.

D. Prediction Page

The Prediction Page screenshot related to application “Cancer Disease Prediction” is shown in Fig. 6. When the Patients logged into the application, they will be redirected to the Patient Home Page where they can use the prediction to predict the type of the cancer disease by selecting the symptoms.

The page has Yes/No options in the symptoms fields followed by clicking on the Predict button. The patient can view their prediction results. After entering the options, the patients will know about their which type of cancer disease the patient are bearing.

IV. Results

The results using different algorithms are discussed below:

A. Confusion Matrix

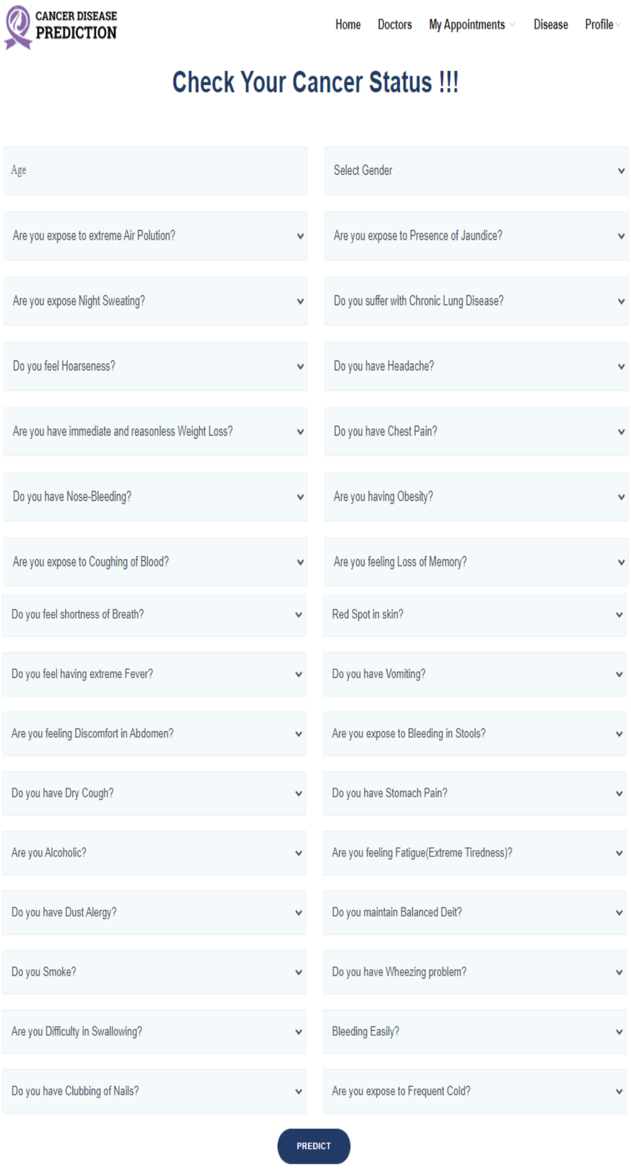
Confusion matrix shows outcome of different class of problems. Values of count help in finding accurate and inaccurate prediction. It contains actual values and predicted values which results in positive and negative classes as shown in Fig. 7.

B. Classification report

Evaluation of classification algorithm is very important part. Accuracy is also much more important. Classification report is the one which does this important work. The report provides per-class examples of the accuracy. Recall and f1-score are the key classification metrics. The metric has measured using positive, true and false negatives.

Classification Rate/Accuracy: Accuracy is considered as one of the parameters for assessing classification models. It is the percentage of predictions. Accuracy is defined as follows:

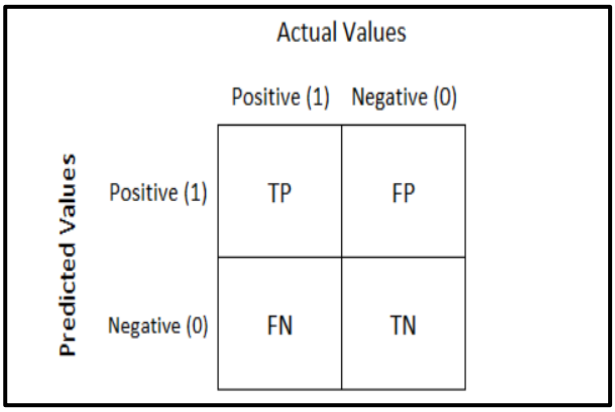
$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$



Check Your Cancer Status !!!

Age
Select Gender
Are you expose to extreme Air Polution?
Are you expose to Presence of Jaundice?
Are you expose Night Sweating?
Do you suffer with Chronic Lung Disease?
Do you feel Hoarseness?
Do you have Headache?
Are you have immediate and reasonless Weight Loss?
Do you have Chest Pain?
Do you have Nose-Bleeding?
Are you having Obesity?
Are you expose to Coughing of Blood?
Are you feeling Loss of Memory?
Do you feel shortness of Breath?
Red Spot in skin?
Do you feel having extreme Fever?
Do you have Vomiting?
Are you feeling Discomfort in Abdomen?
Are you expose to Bleeding in Stools?
Do you have Dry Cough?
Do you have Stomach Pain?
Are you Alcoholic?
Are you feeling Fatigue(Extreme Tiredness)?
Do you have Dust Alergy?
Do you maintain Balanced Deit?
Do you Smoke?
Do you have Wheezing problem?
Are you Difficulty in Swallowing?
Bleeding Easily?
Do you have Clubbing of Nails?
Are you expose to Frequent Cold?

Figure 6. Prediction page



	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Figure 7. Confusion matrix

Binary classification and in terms of positives and negatives, the formula below gives accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall: It is the key classification metric. First step is to find all positively classified examples. Then identify proportion of all examples. Here actually examples that were correctly classified are only considered. High Recall means the class has been successfully identified (a small number of FN). It is calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

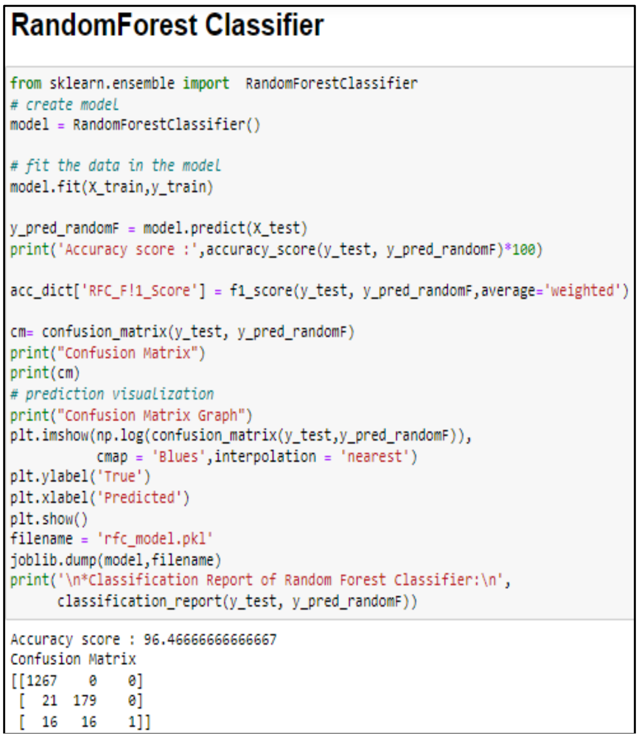
Precision: The precision value is defined by dividing the total number of correctly identified positive cases by the total number of positive examples that were projected to occur. A positive example is one with a high precision rating, hence it must be positive (a small number of FP). It is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

V. Discussion

A. Random Forest

Random Forest classifier is a supervised learning method for classifying and predicting data. It uses the simple regression to build tree structures from a variety of data. Fig. 8 shows the code snippet for Random Forest Classifier. In this algorithm, three performance analyzers are used, namely Accuracy score, Confusion Matrix with the graph and Classification Report. The result obtained when the algorithm is applied on the cancer dataset is shown in Figures 9 and 10. An accuracy score of 96.46% for the Random Forest algorithm is obtained indicating the presence of cancer in the patient.



RandomForest Classifier

```
from sklearn.ensemble import  RandomForestClassifier
# create model
model = RandomForestClassifier()

# fit the data in the model
model.fit(X_train,y_train)

y_pred_randomF = model.predict(X_test)
print('Accuracy score :',accuracy_score(y_test, y_pred_randomF)*100)

acc_dict['RFC_F!1_Score'] = f1_score(y_test, y_pred_randomF,average='weighted')

cm= confusion_matrix(y_test, y_pred_randomF)
print("Confusion Matrix")
print(cm)
# prediction visualization
print("Confusion Matrix Graph")
plt.imshow(np.log(confusion_matrix(y_test,y_pred_randomF)),
           cmap = 'Blues',interpolation = 'nearest')
plt.ylabel('True')
plt.xlabel('Predicted')
plt.show()
filename = 'rfc_model.pkl'
joblib.dump(model,filename)
print('\n*Classification Report of Random Forest Classifier:\n',
      classification_report(y_test, y_pred_randomF))
```

```
Accuracy score : 96.46666666666667
Confusion Matrix
[[1267    0    0]
 [  21  179    0]
 [  16   16    1]]
```

Figure 8. Random forest classifier

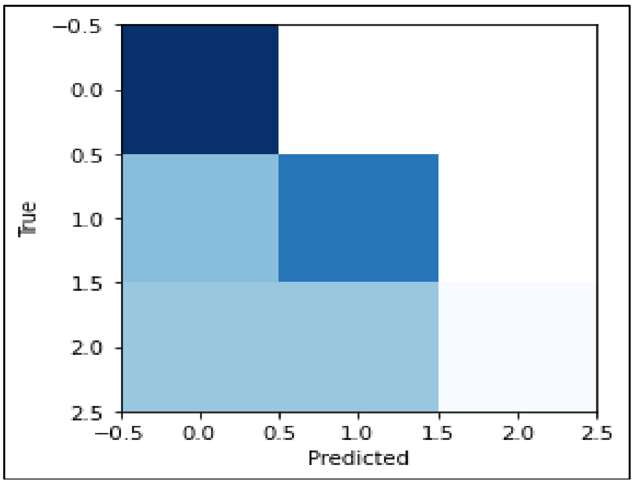
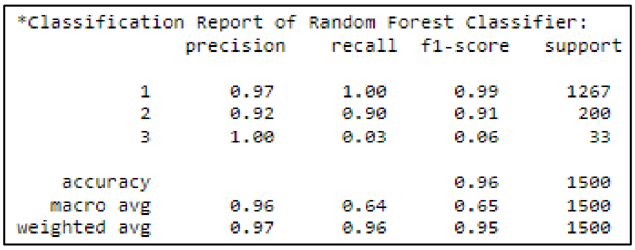


Figure 9. Confusion matrix for random forest classifier

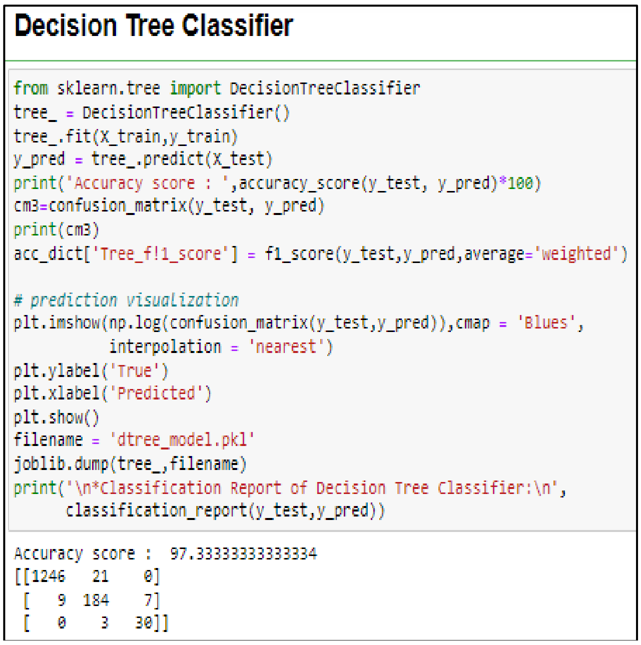


```
*Classification Report of Random Forest Classifier:
              precision    recall  f1-score   support

           1       0.97      1.00      0.99      1267
           2       0.92      0.90      0.91       200
           3       1.00      0.03      0.06        33

    accuracy                           0.96      1500
   macro avg       0.96      0.64      0.65      1500
weighted avg       0.97      0.96      0.95      1500
```

Figure 10. Performance of random forest classifier

B. Decision Tree

The Decision Tree is the most well-known and widely used divide and prediction tool. Each node represents a feature test, so every branch reflects the test's outcome and each node in a decision tree (terminal node) includes the class label. Fig. 11 shows the Decision Tree Classifier code snippet. The accuracy of Decision Tree algorithm is found to be 97.3% as shown in Fig. 12 and 13. This algorithm gives the Highest Accuracy Score compared to all the four algorithms.



Decision Tree Classifier

```
from sklearn.tree import DecisionTreeClassifier
tree_ = DecisionTreeClassifier()
tree_.fit(X_train,y_train)
y_pred = tree_.predict(X_test)
print('Accuracy score : ',accuracy_score(y_test, y_pred)*100)
cm3=confusion_matrix(y_test, y_pred)
print(cm3)
acc_dict['Tree_f!1_score'] = f1_score(y_test,y_pred,average='weighted')

# prediction visualization
plt.imshow(np.log(confusion_matrix(y_test,y_pred)),cmap = 'Blues',
            interpolation = 'nearest')
plt.ylabel('True')
plt.xlabel('Predicted')
plt.show()
filename = 'dtree_model.pkl'
joblib.dump(tree_,filename)
print('\n*Classification Report of Decision Tree Classifier:\n',
      classification_report(y_test,y_pred))
```

```
Accuracy score :  97.33333333333334
[[1246   21    0]
 [   9  184    7]
 [   0    3   30]]
```

Figure 11. Decision tree classifier

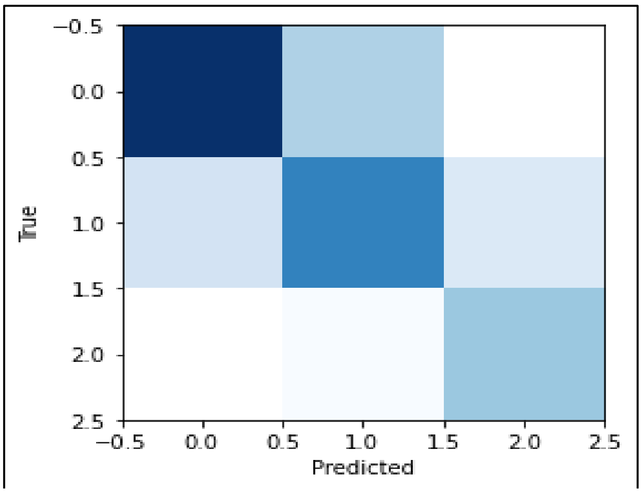
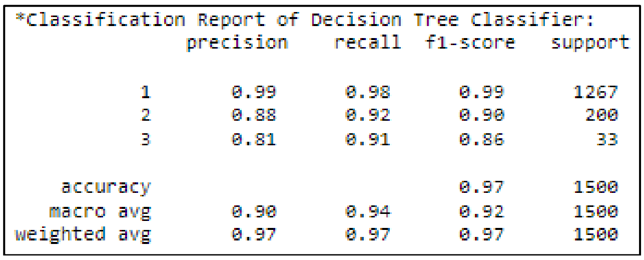


Figure 12. Confusion matrix for decision tree classifier

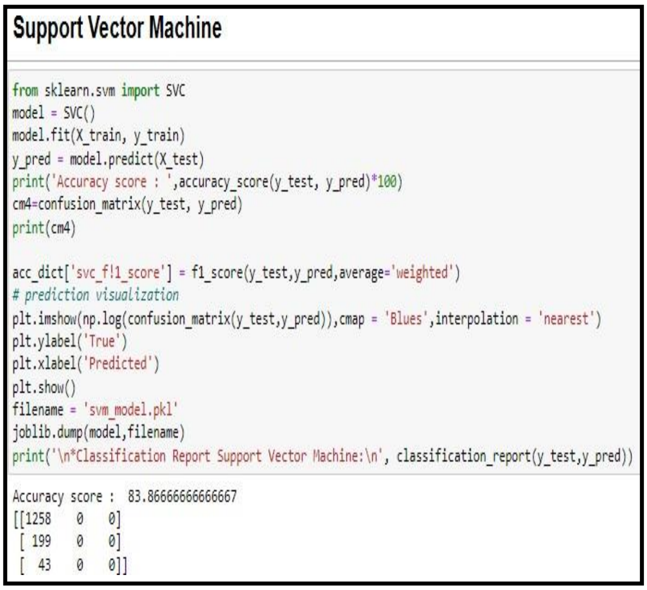


```
*Classification Report of Decision Tree Classifier:
              precision    recall  f1-score   support

           1       0.99      0.98      0.99      1267
           2       0.88      0.92      0.90       200
           3       0.81      0.91      0.86        33

    accuracy                           0.97      1500
   macro avg       0.90      0.94      0.92      1500
weighted avg       0.97      0.97      0.97      1500
```

Figure 13. Performance of decision tree classifier

C. *Support Vector Machine*

To determine the classification and regression, SVM uses supervised machine learning. Identify data points in N-dimensional space. SVM actually peer these hyper plane. The hyper plane's size is number of features. Two input features identify it as line. Chances of collapsing hyper plane occurs if there are more than three input features. Figure 14 show the implementation of Support Vector Machine. An accuracy score of 84.46% for the support vector machine algorithm is obtained as in Figures 15 and 16 indicating the presence of cancer in the patient.



Support Vector Machine

```
from sklearn.svm import SVC
model = SVC()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print('Accuracy score : ',accuracy_score(y_test, y_pred)*100)
cm4=confusion_matrix(y_test, y_pred)
print(cm4)

acc_dict['svc_f!1_score'] = f1_score(y_test,y_pred,average='weighted')
# prediction visualization
plt.imshow(np.log(confusion_matrix(y_test,y_pred)),cmap = 'Blues',interpolation = 'nearest')
plt.ylabel('True')
plt.xlabel('Predicted')
plt.show()
filename = 'svm_model.pkl'
joblib.dump(model,filename)
print('\n*Classification Report Support Vector Machine:\n', classification_report(y_test,y_pred))
```

```
Accuracy score :  83.86666666666667
[[1258    0    0]
 [ 199    0    0]
 [  43    0    0]]
```

Figure 14. Support vector machine

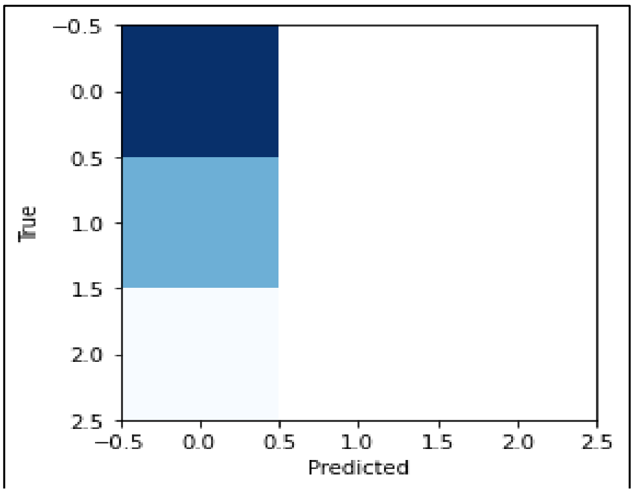
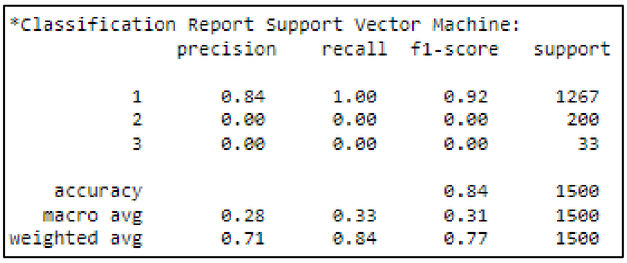


Figure 15. Confusion matrix for support vector machine

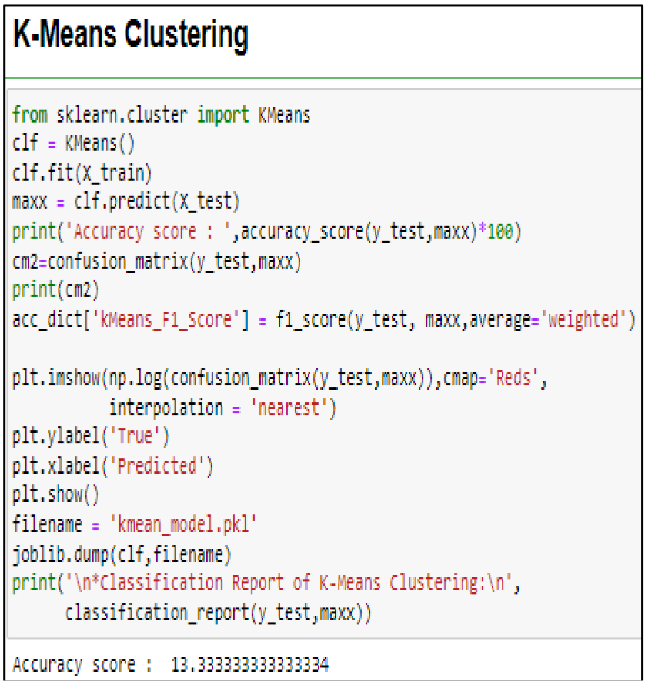


```
*Classification Report Support Vector Machine:
              precision    recall  f1-score   support

           1       0.84      1.00      0.92      1267
           2       0.00      0.00      0.00       200
           3       0.00      0.00      0.00        33

    accuracy                           0.84      1500
   macro avg       0.28      0.33      0.31      1500
weighted avg       0.71      0.84      0.77      1500
```

Figure 16. Performance of support vector machine

D. *K-means clustering*

The unsupervised learning algorithm K-Means Clustering is used to solve a clustering problem where the dataset is divided into a number of k clusters. Fig. 17 shows the code snippet of K-Means Clustering. An accuracy score of 13.33% for the K-Means Clustering is obtained indicating the presence of cancer in the patient as in Figures 18 and 19.



K-Means Clustering

```
from sklearn.cluster import KMeans
clf = KMeans()
clf.fit(X_train)
maxx = clf.predict(X_test)
print('Accuracy score : ',accuracy_score(y_test,maxx)*100)
cm2=confusion_matrix(y_test,maxx)
print(cm2)
acc_dict['KMeans_F1_Score'] = f1_score(y_test, maxx,average='weighted')

plt.imshow(np.log(confusion_matrix(y_test,maxx)),cmap='Reds',
           interpolation = 'nearest')
plt.ylabel('True')
plt.xlabel('Predicted')
plt.show()
filename = 'kmean_model.pkl'
joblib.dump(clf,filename)
print('\n*Classification Report of K-Means Clustering:\n',
      classification_report(y_test,maxx))
```

```
Accuracy score :  13.333333333333334
```

Figure 17. K-means clustering

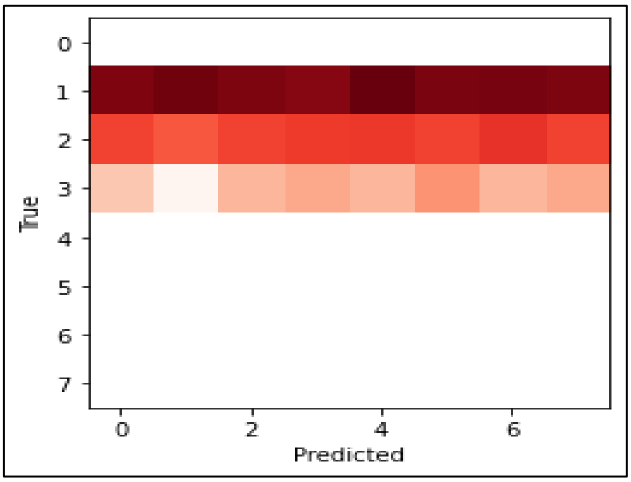
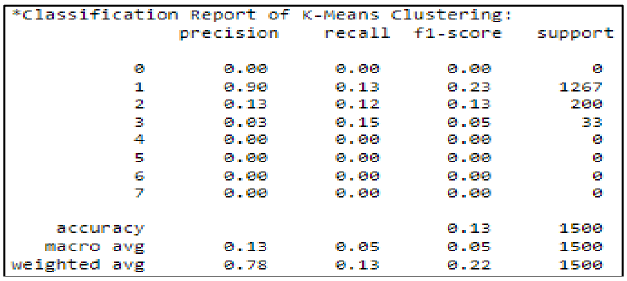


Figure 18. Confusion matrix for K-means clustering

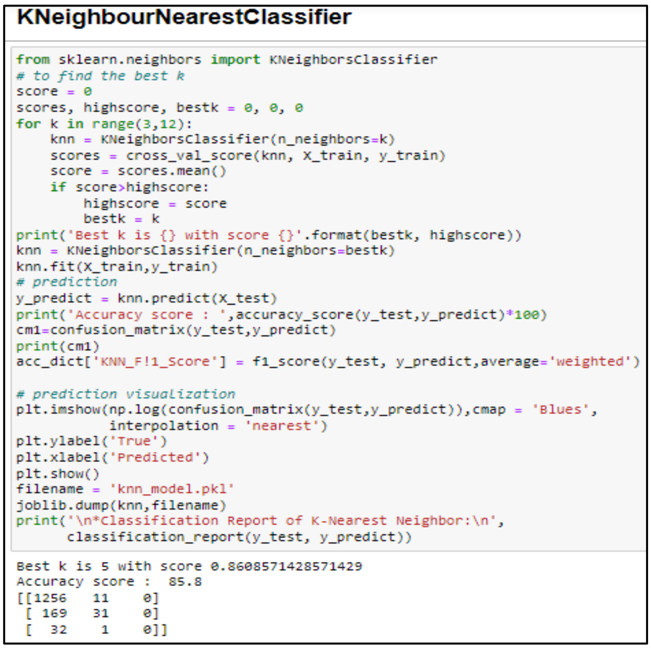


```
*Classification Report of K-Means Clustering:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         0
           1       0.90      0.13      0.23      1267
           2       0.13      0.12      0.13       200
           3       0.03      0.15      0.05        33
           4       0.00      0.00      0.00         0
           5       0.00      0.00      0.00         0
           6       0.00      0.00      0.00         0
           7       0.00      0.00      0.00         0

    accuracy                           0.13      1500
   macro avg       0.13      0.05      0.05      1500
weighted avg       0.78      0.13      0.22      1500
```

Figure 19. Performance of K-means clustering

E. K-nearest neighbours

K-Nearest Neighbour (KNN) is the supervised learning model, used both for classification and regression problems. In this classification, the output will be known as class membership. ‘k’ is a positive integer, which is small. The object is assigned to the single nearest neighbor of the class when k=1 which is shown in Fig. 20. An accuracy score of 85.80% for the K-Means Clustering is obtained indicating the presence of cancer in the patient as in Figures 21 and 22.



KNeighbourNearestClassifier

```
from sklearn.neighbors import KNeighborsClassifier
# to find the best k
score = 0
scores, highscore, bestk = 0, 0, 0
for k in range(3,12):
    knn = KNeighborsClassifier(n_neighbors=k)
    scores = cross_val_score(knn, X_train, y_train)
    score = scores.mean()
    if score>highscore:
        highscore = score
        bestk = k
print('Best k is {} with score {}'.format(bestk, highscore))
knn = KNeighborsClassifier(n_neighbors=bestk)
knn.fit(X_train,y_train)
# prediction
y_predict = knn.predict(X_test)
print('Accuracy score : ',accuracy_score(y_test,y_predict)*100)
cm1=confusion_matrix(y_test,y_predict)
print(cm1)
acc_dict['KNN_F!1_Score'] = f1_score(y_test, y_predict,average='weighted')

# prediction visualization
plt.imshow(np.log(confusion_matrix(y_test,y_predict)),cmap = 'Blues',
           interpolation = 'nearest')
plt.ylabel('True')
plt.xlabel('Predicted')
plt.show()
filename = 'knn_model.pkl'
joblib.dump(knn,filename)
print('\n*Classification Report of K-Nearest Neighbor:\n',
      classification_report(y_test, y_predict))
```

```
Best k is 5 with score 0.8608571428571429
Accuracy score :  85.8
[[1256   11    0]
 [ 169   31    0]
 [  32    1    0]]
```

Figure 20. K-nearest neighbour

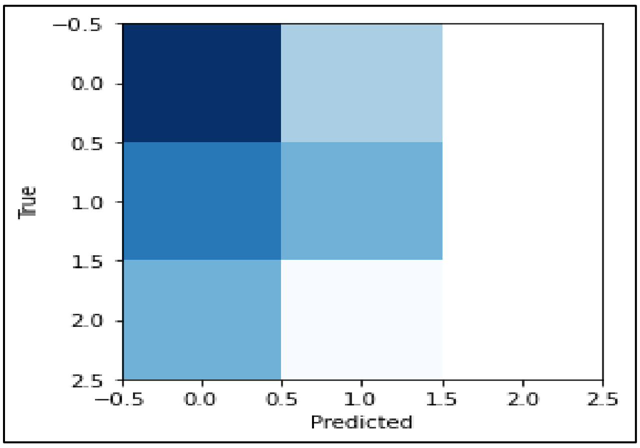
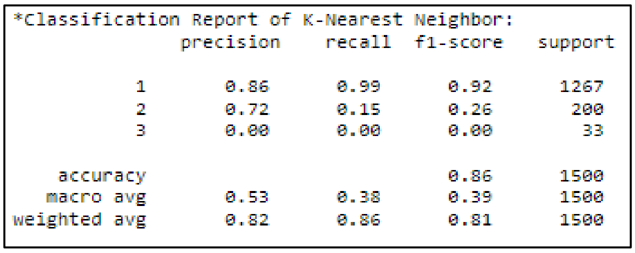


Figure 21. Confusion matrix for K-nearest neighbor

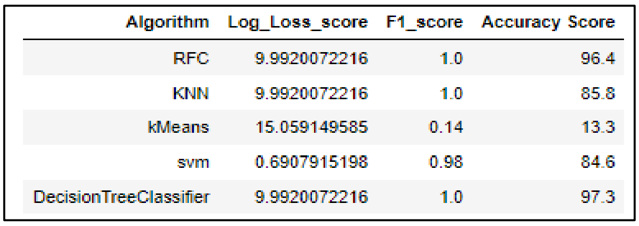


```
*Classification Report of K-Nearest Neighbor:
              precision    recall  f1-score   support

           1       0.86      0.99      0.92      1267
           2       0.72      0.15      0.26       200
           3       0.00      0.00      0.00        33

    accuracy                           0.86      1500
   macro avg       0.53      0.38      0.39      1500
weighted avg       0.82      0.86      0.81      1500
```

Figure 22. Performance of K-nearest neighbor

F. Performance of Algorithms

Fig. 23 shows the performance of all the algorithms that are used in this analysis. All the algorithms performed well for the given cancer dataset, with Random Forest Algorithm giving the accuracy of 96.8%, K-Nearest Neighbor algorithm having 85.0% accuracy, K-Means Clustering algorithm providing an accuracy of 12.8%, Support Vector Machine algorithm obtained 84.6% accuracy and lastly Decision Tree algorithm giving the highest accuracy score of 97.3% among all five algorithms.

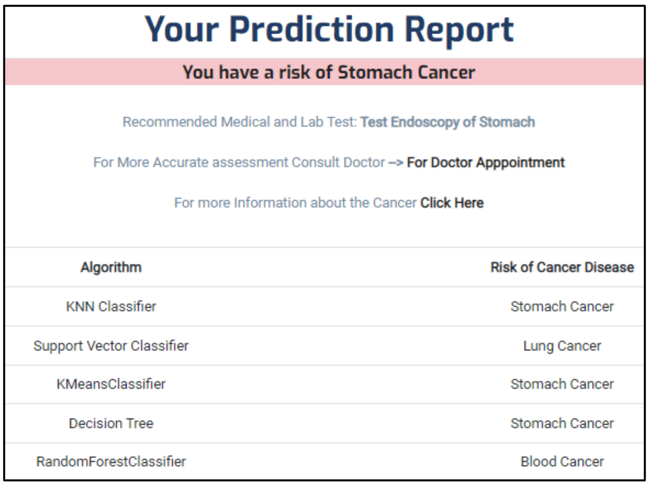


Algorithm	Log_Loss_score	F1_score	Accuracy Score
RFC	9.9920072216	1.0	96.4
KNN	9.9920072216	1.0	85.8
kMeans	15.059149585	0.14	13.3
svm	0.6907915198	0.98	84.6
DecisionTreeClassifier	9.9920072216	1.0	97.3

Figure 23. Performance analysis on cancer dataset

G. Inference

The prediction page of the application displays the type of cancer disease the patient is having and the algorithms applied are shown in Fig. 24. The text message displayed is “You have risk of Stomach Cancer” along with the related medical tests. It describes the algorithms applied and the results obtained from the different algorithms applied. In the example considered above, the KNN Classifier, K-Means Clustering and Decision Tree algorithms are predicting Stomach Cancer whereas SVM and random forest algorithms are predicting other two types of Cancer i.e., Lung cancer and Blood cancer. Since majority of the algorithms are predicting stomach cancer, hence the conclusion is that the patient is having a high risk of suffering from stomach cancer.



Your Prediction Report

You have a risk of Stomach Cancer

Recommended Medical and Lab Test: Test Endoscopy of Stomach

For More Accurate assessment Consult Doctor –> For Doctor Appointment

For more Information about the Cancer Click Here

Algorithm	Risk of Cancer Disease
KNN Classifier	Stomach Cancer
Support Vector Classifier	Lung Cancer
KMeansClassifier	Stomach Cancer
Decision Tree	Stomach Cancer
RandomForestClassifier	Blood Cancer

Figure 24. Prediction page

VI. Conclusion

The project “Cancer Disease Prediction” is used for all the users to know about their health condition and to find out which type of Cancer Disease is the patient suffering from. The Prediction application helps users save money and time. The patients can view the information about the type of the cancer disease and consult a specialized Doctor listed by the application. With early detection of cancer, the user can take precautions or treatments related to type of Cancer Disease. The users can get Prediction results online.

The Decision Tree algorithm gives the highest accuracy compared to the four different algorithms, namely KNN, K-means clustering, SVM and random forest. The Predicting system considers the majority for the Final Prediction Report. Additional features for the Patients include recommendation of related medical tests.

The application “Cancer Disease Prediction” is a web application. This application is used by all users like Patients, Doctors and Admin. One of the most effective ways to reduce the number of deaths by cancer is to detect the type of cancer as soon as possible. A person with minimum knowledge will be able to utilize the application. The user has to enter the symptoms based on their health condition to predict the type of cancer. After getting the prediction report, patient can consult the doctor for the treatment.

References

[1] Akcay, M., Etiz, D., & Celik, O. “Prediction of survival and recurrence patterns by machine learning in gastric cancer cases undergoing radiation therapy and chemotherapy”, *Advances in Radiation Oncology*, vol. **5,** Issue **6**, pp. **1179-1187**, **2020**.

[2] A. D. Mou, M. W. Hasan, P. K. Saha, N. A. R. Priom and A. Saha, "Prediction and Rule Generation for Leukemia using Decision Tree and Association Rule Mining," *11th International Conference on Electrical and Computer Engineering (ICECE)*, pp.**133-136**, **2020.**

[3] Nasser, I. M., & Abu-Naser, S. S., “Lung cancer detection using artificial neural network”, *International Journal of Engineering and Information Systems (IJEAIS)*, vol. **3**, issue **3**, pp.**17-23**, **2019**.

[4] K. A. S. A. Daqqa, A. Y. A. Maghari and W. F. M. A. Sarraj, "Prediction and diagnosis of leukemia using classification algorithms," *8th International Conference on Information Technology (ICIT)*, pp. **638-643**, **2017**.

[5] Maria, I. J., Devi, T., & Ravi, D. “Machine learning algorithms for diagnosis of leukemia” *International Journal of Science and Technology Research*, vol. **9**, issue **1**, pp.**267-270**, **2020.**

[6] Ada and Rajneet Kaur “Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient” *International Journal of Computer Science and Mobile Computing, IJCSMC*, Vol. **2**, Issue. **4**, pp.**1 – 6**, **April 2013.**

[7] Charles Edeki “Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability”, *Mediterranean journal of Social Science,* Vol. **3** issue. **14**, **November 2012**.

[8] Reeti Yadav “Chemotheraphy Prediction of Cancer Patient by Using Data Mining Techniques”, *International Journal of Computer Applications,* vol. **76**-No.**10**, **August 2013.**

[9] Arunachalam, Siddhika. "Applications of Machine learning and Image processing techniques in the detection of leukemia", *International Journal of Scientific Research in Computer Science and Engineering*, vol. **8**, issue **2**, pp. **77-82**, **April 2020**.

[10] G. Nallasivan, M. Sivaranjani "Lung Detection and Segmentation for Cancer Diagnosis in Machine Learning Approach", *International Journal of Scientific Research in Biological Sciences*, vol. **8**, issue **1**, pp. **49-54**, **2021**.

AUTHORS PROFILE

Mr. Sanjay Garag has pursued his Master in Computer Applications degree from KLE Institute of Technology, Hubballi. His interests include analysis of machine learning algorithms and Web Development.

Mrs. Anupama S Nandeppanavar pursued her Bachelor of Engineering in Information Science Engineering from Visvesvaraya Technological Univesity in 2005 and Master of Technology in Computer Science from Visvesvaraya Technological Univesity in 2010. She is currently working as Assistant Professor in Department of MCA, KLE Institute of Technology, Hubballi since 2013. She is a member of IFERP and ICTD. She has published papers in reputed international journals. Her main research work focuses on Cyber Security, Machine Learning and Web Technologies. She has 15 years of teaching experience.

Dr Medha Kudari pursued Bachelor of Science and Master of Science from Mangalore University in year 2005. She has been awarded Ph.D. from Karnataka University, Dharwad in the year 2019 and currently working as Assistant Professor in Department of MCA, KLE Institute of Technology since 2019. She has published research papers in reputed international journals and presented at international conferences. Her main research work focuses on Image Processing and Pattern Recognition along with Data Mining, IoT and Machine Learning. She has 12 years of teaching experience and 4 years of Research Experience.