Soft Computing in Cancer Prediction

BREAST CANCER PREDICTION

SUBMITTED BY

Gaurab Sedhai (11192612)

Rishita Prakash (11192546)

Sanyam Singla (11192518)

Hemant (11192764)

SUBMITTED TO

Dr. Sandhya Bansal

Abstract

With the rapid growing technologies in the field of Health sector, the popularity of soft computing approach is blooming more than any time in history. Soft computing approaches play a vital role in solving the different kinds of problems and provide promising solutions. This approach has been applied for effectively diagnosing the diseases and obtaining better results in comparison to traditional approaches. Various AI predictive models have been built for prediction of diseases at an early stage. However, the work in developing a decision-suppprt system for healthcare is still in the infancy state. Most of the conventional decision-support systems are based on hard computing which requires exactly state analytic models and does not have any place for approximation and uncertainty. Soft computing, being an approach that imitates the human mind to reason and learns in an environment of uncertainty and impression, helps to provide an optimal solution through its nature of adaptivity and knowledge. The soft computing approach is categorized based on the methodology as genetic algorithm, artificial neural network, fuzzy logic etc. The first objective of this report is to categorize the various soft computing approaches used for diagnosing and predicting the cancer. Second objective of this report is to identify the soft computing approach used for predicting the breast cancer.

Introduction

Artificial intelligence is broadly defined as the programming of machines to enable them to perform tasks like humans. Artificial intelligence simulates human intelligence in learning, perceiving, planning, and decision making.

Breast cancer is the most aggressive type of cancers suffered by women worldwide and becomes the second leading cause of death among women cancer patients. Apparently, the diagnosis and the scrutiny of the breast cancer disease have always been a decisive and critical one in the regard of medical department. Breast cancer is considered as a lump that is formed in the breast cells; when these cells begin to grow irregularly in a human body, it results in flaking and redness of the breast. The cancerous lumps which are also termed as tumors are comprised of two kinds: one is the ***benign*** and the other is ***malignant***. The cancer is still considered as the undiagnosed and untreated disease in various parts of the world. But the early diagnosis of breast cancer can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment of patients.

Diagnosis of breast cancer can be made manually by the physician, but it will take a longer period of time and must be very intricate for the physician to implement the classification. The incompleteness of relevant data can also lead to human errors in diagnosis. Thus, breast cancer detection through an intelligent system is vital in the medical field. Various methods can be applied for classification of breast cancer such as Neural Network, Support Vector Machine, KNN and decision tree.

In this research, fuzzy logic and SVM are used to find out the breast cancer by using different statistical measures such as accuracy, miss rate, false-positive value, false-negative value, likelihood ratio positive, likelihood ratio negative, positive prediction value, and negative prediction values. With the help of these matrices, breast cancer can be found more accurately. Further accurate classification of ***benign*** tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of breast cancer and classification of patients into ***malignant*** or ***benign*** groups is the subject of much research.

Literature Review

S.No.	Paper	Work	Soft Computing Techniques	Advantages
1	1(A)	Soft Computing in skin cancer diagnosis	Artificial Neural Network	ANN is inspired by human brain interactions between synapses and neurons. The ANN method is a good box-based tool for classification of the nonlinear problems with least attempts.
2	1(B)		Convolutional Neural Network	CNN is often employed for image or speech analysis in machine learning. After the application of CNN in image processing, several researchers started to work on using CNN as a tool in medical image processing.
3	2(A)	Soft computing in lung cancer prediction	Convolutional Neural Network	This image processing technique can be used for early detection and treatment to prevent lung cancer. Various features can be extracted from images and therefore, pattern recognition-based approaches are useful in prediction of lung cancer.
4	2(B)		Fuzzy Algorithm	This approach used type-II fuzzy algorithm to improve the quality of raw CT images then, algorithm based on fuzzy c-means clustering is offered in order to achieve another representation of lung regions.
5	3	Soft computing in breast cancer prediction	Fuzzy Logic	It evaluates the features extracted and determines its effects. The Fuzzy-Multi layer SVM (Support Vector Machine) version indicates promising consequences and can offer a density for more sophisticated statistical features based most cancers prognostic models.

Findings

Soft computing methods provide solutions to biologically inspired problem of medical domain like breast cancer. Neural Networks, Fuzzy Logic and Genetic Algorithms contribute novel algorithms to deal with cancer prediction. Breast cancer can be diagnosed using soft computing methods. The effective diagnosis of breast cancer can be achieved by using feature reduction and classification methods.

The image processing techniques are mostly used for prediction of lung cancer also and for early detection too and to prevent the cancer. To predict lung cancer, various features are extracted from the images therefore, pattern recognition-based approaches are useful to predict the lung cancer.

The significance of image processing in medical applications helps the physicians and radiologists to reduce the complexity and increase the early detection speed for disease diagnosis.

Results

A sample code to predict the type of breast cancer, i.e., ***malignant*** or ***benign*** with values like true positive, true negative, false positive, false negative. Breast Cancer Wisconsin (Diagnostic) Dataset is used for the model and accuracy is determined of the model that is created. Google Colab is used as a platform as it contains all the libraries required to create a model.

- First the libraries like numpy, pandas, seaborn, matplotlib are imported to read the csv (Comma Separated Values) and for plotting the data.

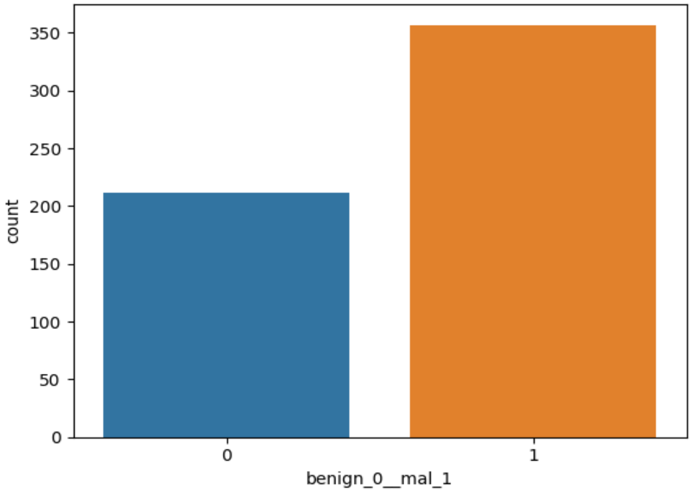
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

- A data frame is created named **df** to load the dataset(cancer_classification.csv) using pandas.

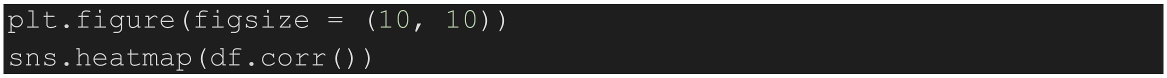
```
df = pd.read_csv('cancer_classification.csv')
```

- A bar graph is plotted representing the type of cancer (***malignant*** as 1 and ***benign*** as 0) with respect to number of patients.

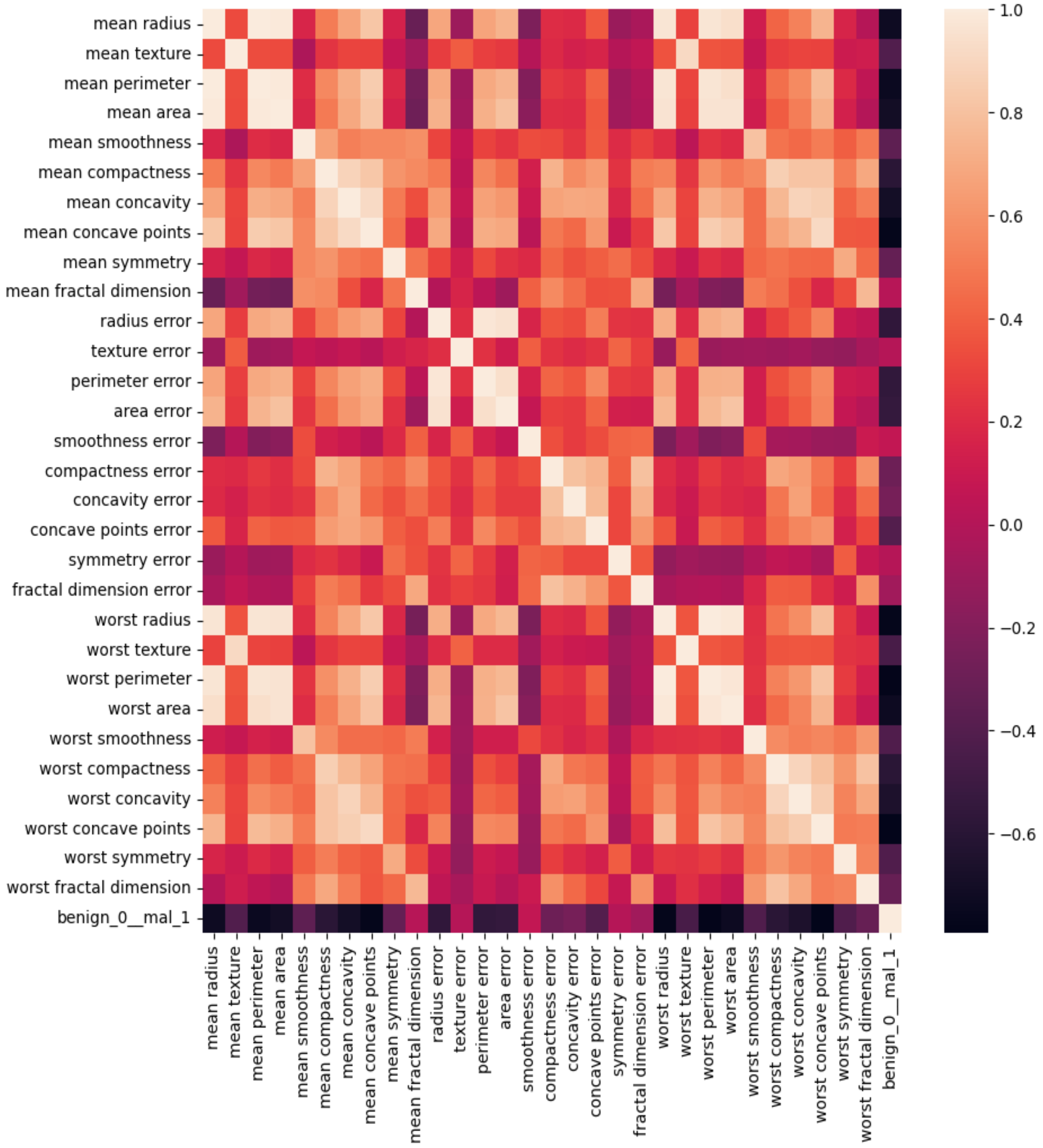
```
sns.countplot(x = 'benign_0__mal_1', data = df)
```



- Heatmap can be used to find the correlation.



```
plt.figure(figsize = (10, 10))
sns.heatmap(df.corr())
```



- The dataset is split into training set and testing set. **Sklearn** library helps us to automatically split the dataset. But first, variables X and y are used to store the values other than benign_0__mal_1 and values of benign_0__mal_1 respectively.

```
from sklearn.model_selection import train_test_split
X = df.drop('benign_0__mal_1', axis = 1).values
y = df['benign_0__mal_1'].values

X_train, X_test, y_train, y_test =  train_test_split(X, y, test_size
= 0.25, random_state = 42)
```

- Feature scaling is important since it normalize the data so feature scaling is applied and sklearn contains the function minmax scaler for that. There should not be outliers in data or should be at minimum count.

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()

scaler.fit(X_train)
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

- Now to train the model, different layers are created and activation functions are used in those layers and last the optimizer is selected.

```
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout, Activation

model = Sequential()
model.add(Dense(units = 30, activation = 'relu'))
model.add(Dense(units = 15, activation = 'relu'))
model.add(Dense(units = 1, activation = 'sigmoid'))

model.compile(loss = 'binary_crossentropy', optimizer = 'adam')
```

- The model is trained by using the training and testing data.

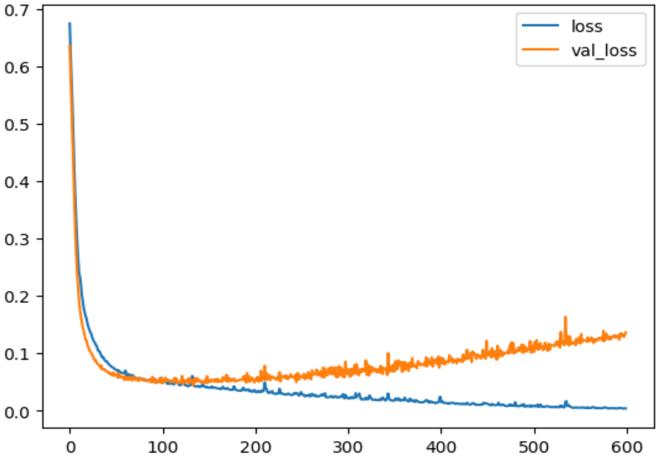
```
model.fit(x = X_train,
          y = y_train,
          epochs = 600,
          validation_data = (X_test, y_test),
          verbose = 1)
```

#Last line of output of model that is trained.

```
Epoch 600/600
14/14 [==============================] - 0s 6ms/step - loss: 0.0036
- val_loss: 0.1361
```

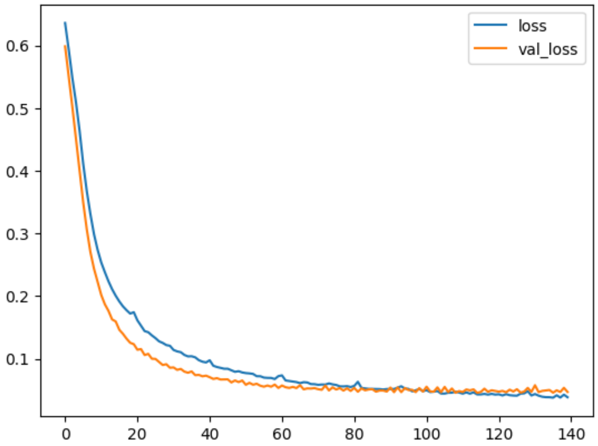
- The model is now trained but we can find the difference in loss, so a graph is plotted to find out the loss since the model can be overfitted without the proper instructions.

```
model_loss = pd.DataFrame(model.history.history)
model_loss.plot()
```

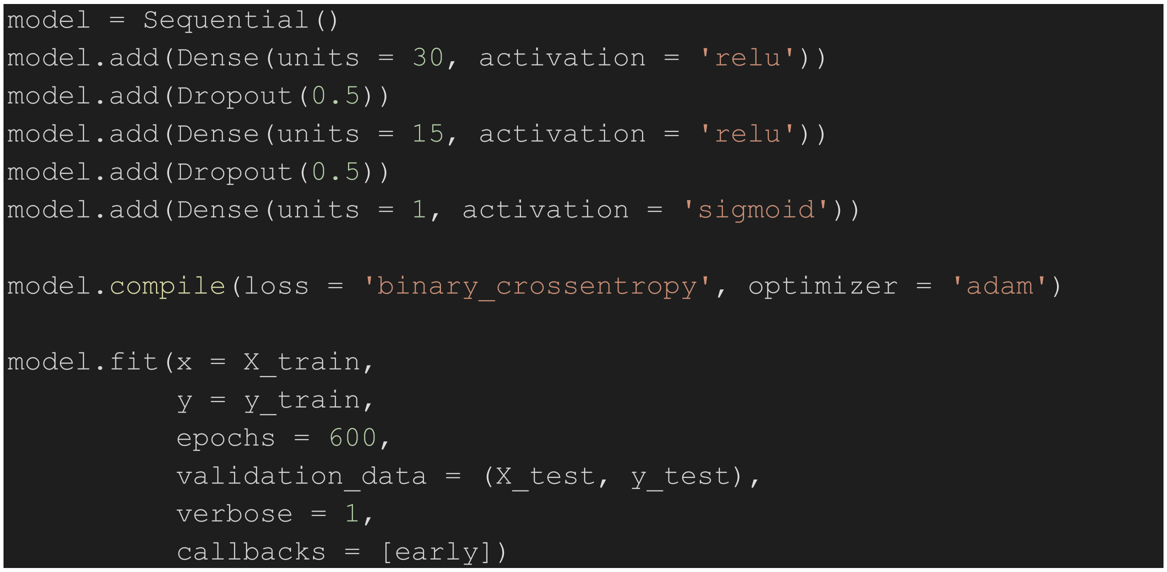


- From the graph, it can be seen that after a certain point the different is huge so EarlyStopping function is used to stop the model from overfitting.

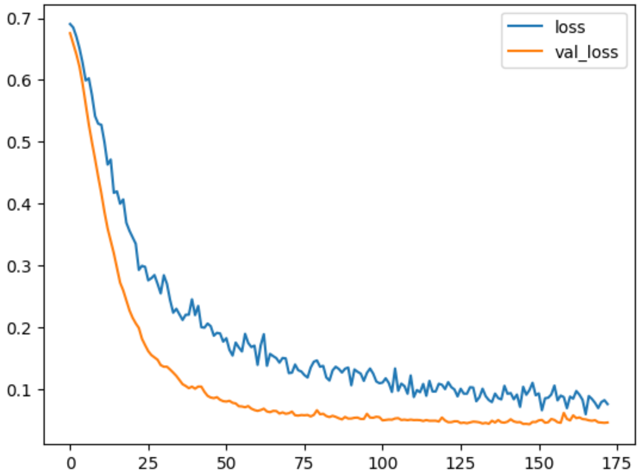
```
model = Sequential()
model.add(Dense(units = 30, activation = 'relu'))
model.add(Dense(units = 15, activation = 'relu'))
model.add(Dense(units = 1, activation = 'sigmoid'))

model.compile(loss = 'binary_crossentropy', optimizer = 'adam')

from tensorflow.keras.callbacks import EarlyStopping

early = EarlyStopping(monitor = 'val_loss', mode = 'min', verbose =
1, patience = 25)
model.fit(x = X_train,
          y = y_train,
          epochs = 600,
          validation_data = (X_test, y_test),
          verbose = 1, callbacks = [early])
```



- Now the model looks better but there is still slight crossing between the loss. To further optimize the model, dropout can be used.



```
model = Sequential()
model.add(Dense(units = 30, activation = 'relu'))
model.add(Dropout(0.5))
model.add(Dense(units = 15, activation = 'relu'))
model.add(Dropout(0.5))
model.add(Dense(units = 1, activation = 'sigmoid'))

model.compile(loss = 'binary_crossentropy', optimizer = 'adam')

model.fit(x = X_train,
          y = y_train,
          epochs = 600,
          validation_data = (X_test, y_test),
          verbose = 1,
          callbacks = [early])
```



- Now the model is tested for the prediction and values such as accuracy, confusion matrix is calculated and classification report is calculated too.

```
predictions = (model.predict(X_test) > 0.5).astype("int32")

from sklearn.metrics import confusion_matrix, classification_report,
accuracy_score

print(accuracy_score(y_test, predictions) * 100)
print(confusion_matrix(y_test, predictions))
print(classification_report(y_test, predictions))
```

- The following output is obtained.

```
97.9020979020979
```

The model gave the accuracy of 97.90%.

```
[[53  1]
 [ 2 87]]
```

In confusion matrix the values are divided into following parameters:

- True Positive: 53
- True Negative: 87
- False Positive: 1
- False Negative: 2

```
              precision    recall  f1-score   support
           0       0.96      0.98      0.97        54
           1       0.99      0.98      0.98        89
    accuracy                           0.98       143
   macro avg       0.98      0.98      0.98       143

weighted avg       0.98      0.98      0.98       143
```

And classification report can also be observed.

Conclusion

Fuzzy algorithm is reliable and managed to generate good performances in the classification of breast cancer data. The gained result is presented in page no. 10. There were total of 569 cases during the evaluation process. Then for the implementation of fuzzy algorithm, they were divided into training and testing sets, where testing set size was 25% of total data. Then the algorithm was applied and the model was trained with 75% of training set of data. Compiling and comparing the values of training and testing set, the accuracy of 97.90% was achieved, close to 98%. And from the confusion matrix we could find the following values which are the result of the 25% testing set of data.

True positive, i.e., Cytological and suspicious positive diagnosis, which are positive in pathological tests = 53

False positive, i.e., Cytological and suspicious positive diagnosis, which are negative in pathological tests = 1

True negative, i.e., Cytological and suspicious negative diagnosis, which are negative in pathological tests = 87

False negative i.e., Cytological and suspicious negative diagnosis, which are positive in pathological tests = 2

Simulation results show that using data mining process can increase accuracy and efficiency of the network training. The main goal of this research is to achieve 100% accuracy, so to guarantee those artificial intelligence systems and furnish them with practical aspects, we need adequate knowledge, and satisfactory control over intelligent systems, data mining, and breast cancer.

References

1. Role of Soft Computing Approaches in HealthCare Domain: A Mini Review.
2. Soft Computing Techniques for Medical Diagnosis, Prognosis and Treatment.
3. Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis.
4. Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning.
5. Early Detection of Breast Cancer using Soft Computing.
6. Fuzzy Multi-Layer SVM classification of Breast Cancer Mammogram images.
7. Computer-aided diagnosis of skin Cancer based on soft computing techniques.
8. Early detection of cancer using soft computing.
9. Dataset used during implementation.