FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 4 No. 1, March, 2020, pp 55 – 61

MACHINE LEARNING MODEL FOR BREAST CANCER DETECTION

[1]Malik Adeiza Rufai, [2]Ahmad Shehu Muhammad, [3]Garba Suleiman and [4]Audu Lovingkindness

[1, 2, 4] Department of Computer Science, Federal University Lokoja Kogi State
[3] Department of Computer Science, FCT College of Education, Zuba Abuja

Corresponding authors email: rufai.malik@fulokoja.edu.ng

ABSTRACT

Cancer is the second most common cause of death in the world, especially in developing countries. Cancer is a serious and dangerous disease that is usually caused by abnormal growth of the cell. This abnormal growth causes lumps (tumors). Breast cancer is ranked the fifth cause of cancer deaths worldwide and the second most common form of cancer amongst females. Machine learning is one of the most trending technologies in computing. It entails the automated conversion of data into useable information that can be related with. This Paper is concerned with the use of a classification algorithm like Support Vector Machine (SVM) to develop a model for the detection of breast cancer. The developed model is to select/extract features from dataset and then classify it as either benign (non-cancerous) or malignant (cancerous). Dataset was collected from the Wisconsin database (in which features are computed from a digitalized image of a Fine Needle Aspirate (FNA) of a breast mass) then trained and tested using the SVM classifier for prediction of breast cancer. In the algorithm, we plot each data item as a point in n-dimensional (here n=9), with the value of each feature being the value of a particular coordinate. Classification is performed by finding the hyper-plane that differentiates the two classes. The model accepts values of attributes as inputs and predicts whether a breast mass is benign or malignant, with an accuracy level of 94.3%. This model is recommended for use in the health sector and also for personal use.

Keywords: Support Vector Machine (SVM), Benign, Malignant, Fine Needle Aspirate (FNA), Machine Learning, Breast Cancer, Medical Images

INTRODUCTION

Cancer is a serious and a very dangerous disease that is usually caused by cells that are not normal and can spread from one part to several parts of the body (Sonawane and Bhutad, 2017). It has eaten deep into humanity and has claimed the lives of millions of people. According to the American cancer society (ACS, 2017) , cancer is the second most common cause of death in the USA and accounts for nearly 1 of every 4 death. Cancer is considered to be one of the leading causes of mortality worldwide (Brook et *al*, 2008; Charan et al, 2018). Some kinds of Cancer include skin cancer, breast cancer, lung cancer, colorectal cancer, cancer of the brain, etc. For the purpose of the project, the focus will be on breast cancer. Breast Cancer has been rated as one of the leading causes of death among women (Fallahi and Jafari, 2011). It is ranked the fifth most cause of cancer deaths worldwide, and the second most common form of cancer amongst females (Lotfy and Salem, 2010). In the past two decades, the incident of breast cancer has shown an upward trend. Breast cancer starts when the cells in the breast suddenly begins to grow out of control. It can start from different parts of the breast. According to (Hamouda, 2017), each year about 182, 000 women are diagnosed with breast cancer in the whole world, and about 43,300 dies. Statistically speaking, one woman out of every eight women, has or will develop breast cancer in her lifetime. The risk of developing breast cancer increases with age. In both developed and developing countries, there has been an increase in the number of individuals suffering from breast cancer. It is really a deadly disease (Levy and Jain, 2016).

The mortality rate of people whose life have been cut short by breast cancer is very high. Every year, more lives are claimed. The detection/diagnosis of cancer related diseases is sometimes very challenging, any may not be accurate if carried out manually (Nahid and Kong, 2018).

To curb this, we embarked on this research, in a bid to come up with a machine learning model that will help reduce the mortality rate and improve the standard of living of infected victims. Many other related works has been carried out before now by several researchers. But, our research has the highest degree of accuracy, compared to previous works.

Nahid and Kong (2018), the MIAS database was used, from which several wavelet entropy features like were extracted and analyzed in digital mammograms for the diagnosis of breast cancer. After the features were extracted, ensemble classification was carried out, using Support vector machine (SVM), Bayes and K-nearest neighbor (KNN) classifiers. SVM gave the highest level of accuracy.

Higa, (2018) work is concerned with the development of an expert system for the diagnosis and prognosis of breast cancer. Two data mining algorithms, i.e. decision tree and artificial neural network to classify mammograms as either malignant or benign. At the end of the day, the degree of accuracy was high respectively

This paper, prepared by (Khodary*et al*., 2017), aims at developing a MATLAB program, for early detection of breast cancer, based on excellent images produced. This approach works with digital mammogram, from which the tumor regions are separated, and then the images are classified based on edge-sharpness, shape of tumor and feature extraction, after which the system decides whether the mammogram image is normal, and whether it is abnormal. If it is abnormal, it then checks if it is benign or malignant.

In this paper, a model was presented, known as the Adaptive Resonance theory (ART2) Neural Network that could classify breast tumor as either benign or malignant. When this model was compared with other models and tested using the Wisconsin Breast cancer database, the degree of accuracy realized in the result was 82.64%, while the fuzzy system and the digital mammogram breast cancer diagnosis gave a result of 80.6% and 73.7% respectively (Narang *et al.*, 2012).

This research paper is concerned with the development of an expert system for the diagnosis and prognosis of breast cancer. Two data mining algorithms, i.e decision tree and artificial neural network to classify mammograms as either malignant or benign. At the end of the day, the degree of accuracy was high respectively (Higa, 2018).

Wang (2017) went on a survey and reviewed screening and biosensor techniques used for the diagnosis of breast cancer at its early stage. Most of the recent approaches of data mining and biomarkers with biosensors were also reviewed. The micro-wave imaging technique was studied and approved as a very reliable and cost-effective diagnostic tool for the early detection of breast cancer.

This research work is concerned with the use of Bayesian network and data preprocessing for the automatic detection of breast cancer in females. The dimension of the breast cancer database is reduced by an algorithm known as Relief, which is normally followed by data preprocessing, and then Bayesian network is used to classify the images as either malignant or benign. When this approach was compared with others like neural network and neural network combined with association rules, it was discovered that the model performs better (Fallahi and Jafari, 2011).

Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier defined by a separating hyperplane. Alternatively, given labeled training data (*supervised learning),* the algorithm yields an optimal hyperplane. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side (Ade-Ojo, 2018).

SVM is supervised learning technique, mostly used in medical diagnosis for classification. It minimizes the experimental errors and maximizes the geometric margin that is why it is also called Maximum Margin Classifier (Nahid and Kong, 2018). Data represent in the space in the form of points, belonging to either one of 2 classes. SVM creates hyperlane that divide the data according to their belonging classes on the same plane as shown in figure 1.

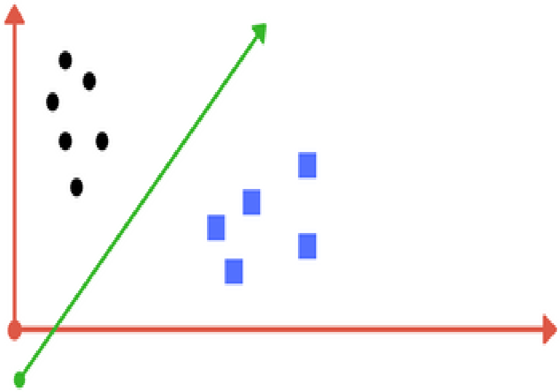


Fig. 1: Support Vector Machine (Ade-Ojo, T. 2018)

Where x is input sample lying on hyperlane, w is the hyperlane orientation and b is the distance of hyperlane from the origin. {+1.-1} are the labels of the class.

The reason to select SVM is:

- It’s a binary classifier and consider best for 2 class problem
- SVM provide unique, optimal and global solution.

MATERIALS AND METHODS

This methodology provides complete measures for the detection of breast cancer. For the purpose of this paper, we used the Cross-Industry Standard Process for data mining (CRISP-DM) methodology. CRISP-DM as shown in figure 2 is one of the best methodologies for data mining, introduced in 1996. The sequence of the phases as shown in figure 2 is not strict and moving back and forth between different phases as it is always required.

The Cross-Industry Standard Process for data mining (CRISP-DM) methodology was selected for the following reasons:

i. It saves cost and reduces risk of project failure

ii. It supports gradual and timely delivery and testing of system functionalities.

iii. It encourages constant feedback and responses from users.

The Wisconsin dataset, containing 699 data and 11 attributes were used. 70% of the data was used for training and 30% for testing. The selected features includes: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses.

For the purpose of this research, a machine learning algorithm, known as the Support Vector machine (SVM) algorithm was used to test, classify the selected features, and predict whether it's malignant (cancerous) or benign (non-cancerous).

Support vector machine algorithm was used for classification. In this algorithm, we plot each data item as a point in n-dimensional (here n=9), with the value of each feature being the value of a particular coordinate. Classification is performed by finding the hyper-plane that differentiates the two classes (i.e, benign and malignant).

The Django framework was used to design the back end and the front end, which handles the features that were inserted and was designed using HTML, CSS and JavaScript. The features were inserted into columns, and then sent to the back end for processing.

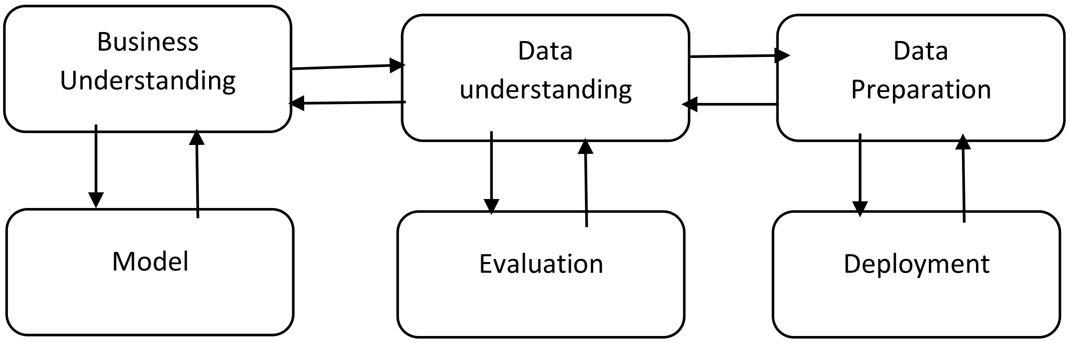


Fig. 2. Illustration of the CRISP-DM Methodology

The development process started with the planning phase where the current system was observed, existing documents were analyzed and the users were met to gather the requirements of the system. It was followed by the analysis phase, where the logical designs of the system were produced. Then to the designing phase where the physical designs of the system were produced. Thereafter, proceeded to the coding and testing phase where the logical and physical designs were translated into programs and the system's outputs were validated with that of a real system. Whenever the output is undesirable, the development process as in figure 3 backtracked to the analysis phase; this is continued until the desired system is achieved.

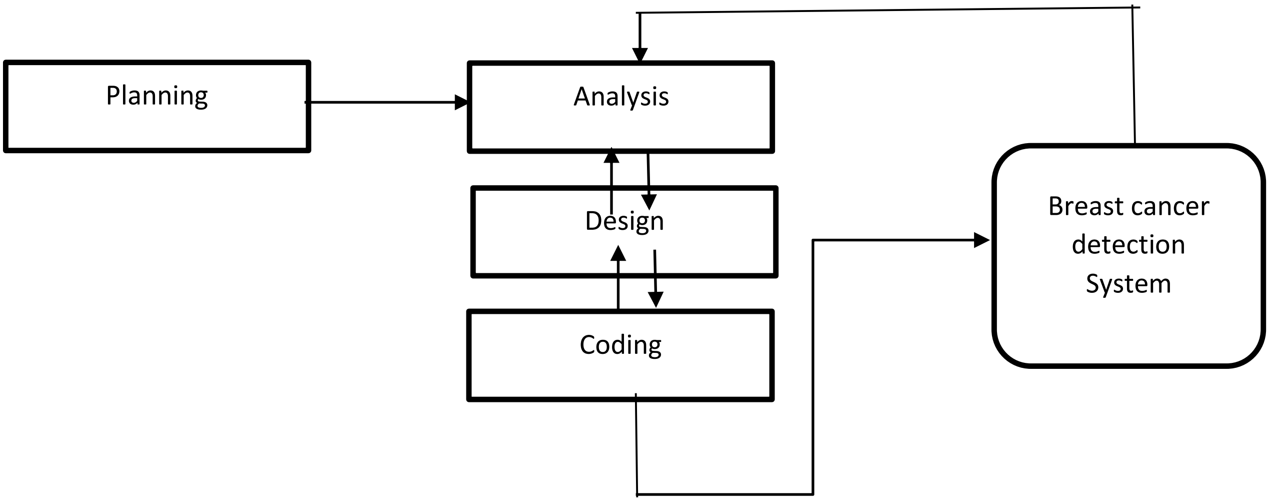
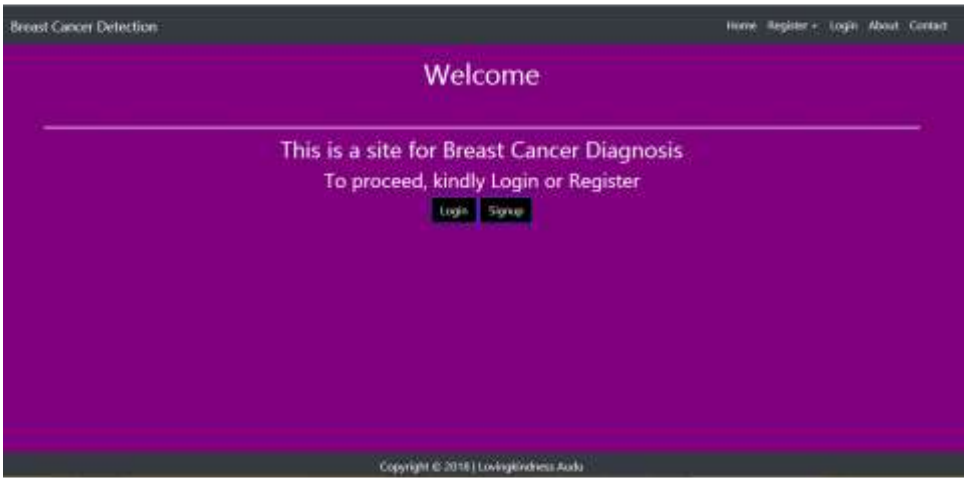


Fig. 3: Development process of the system

RESULTS AND DISCUSSION

As earlier stated, we used Support Vector machine for classification, after which a prediction is made as to whether a set of features falls under malignant or benign. At the end of this research, a model that can be used for the detection of breast cancer in ladies was formulated.

Figure 4 is the screenshot on how to switch to the user registration page by clicking on the **Signup** button.



Welcome

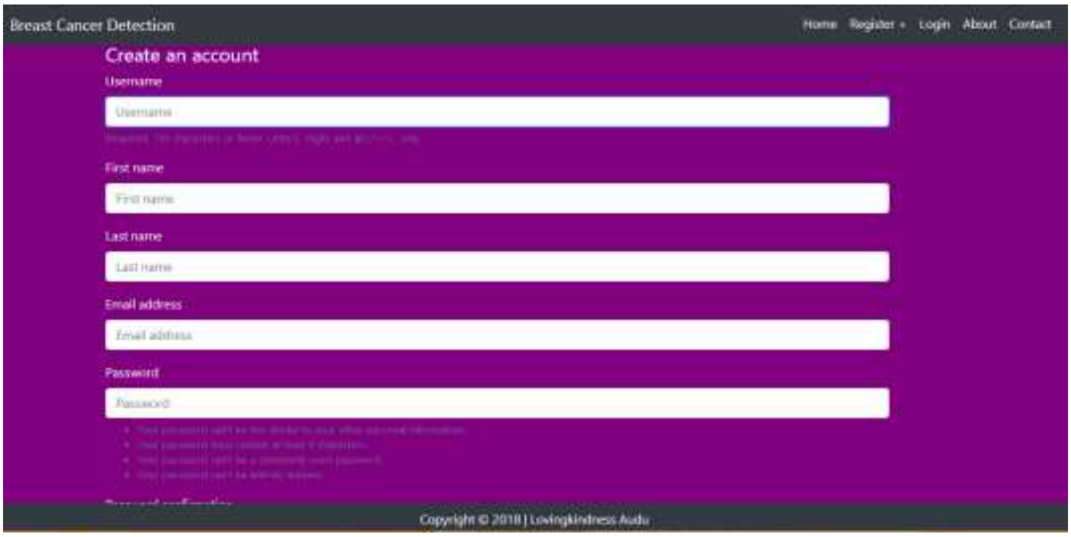
This is a site for Breast Cancer Diagnosis

To proceed, kindly Login or Register

Copyright © 2018 | Lovingkindness Audu

Fig. 4: screenshot Home page

To register, enter your proposed ‘username’, ‘firstname’, ‘lastname’, ‘email’, ‘password’ and retype your password for verification, then click on **Signup** as shown in figure 5.



Create an account

Username

Username

Required. 150 characters or fewer. Letters, digits and @/./+/-/_ only.

First name

First name

Last name

Last name

Email address

Email address

Password

Password

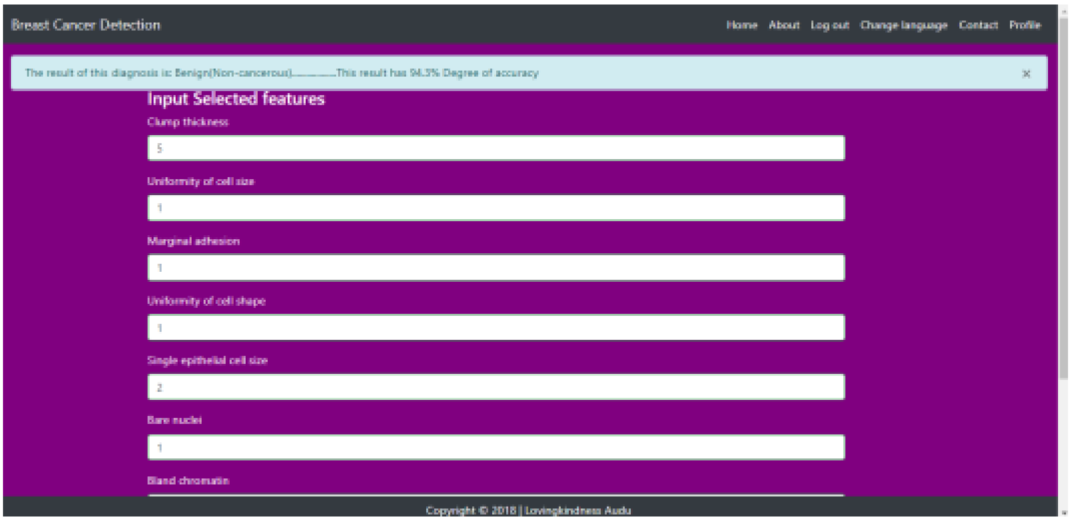
- Your password can't be too similar to your other personal information.
- Your password must contain at least 8 characters.
- Your password can't be a commonly used password.
- Your password can't be entirely numeric.

Password confirmation

Copyright © 2018 | Lovingkindness Audu

Fig. 5: screenshot of Registration page

Figure 6 shows the user interface, where features are inserted. The radiologist gets the mammogram as the patient, carries out several tests on it and then selects the Clump Thickness , Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses, which are then tested to ascertain whether its benign (non-cancerous) or malignant (cancerous), using the model.



The result of this diagnosis is: Benign(Non-cancerous)...............This result has 94.3% Degree of accuracy

Input Selected features

Clump thickness: 5

Uniformity of cell size: 1

Marginal adhesion: 1

Uniformity of cell shape: 1

Single epithelial cell size: 2

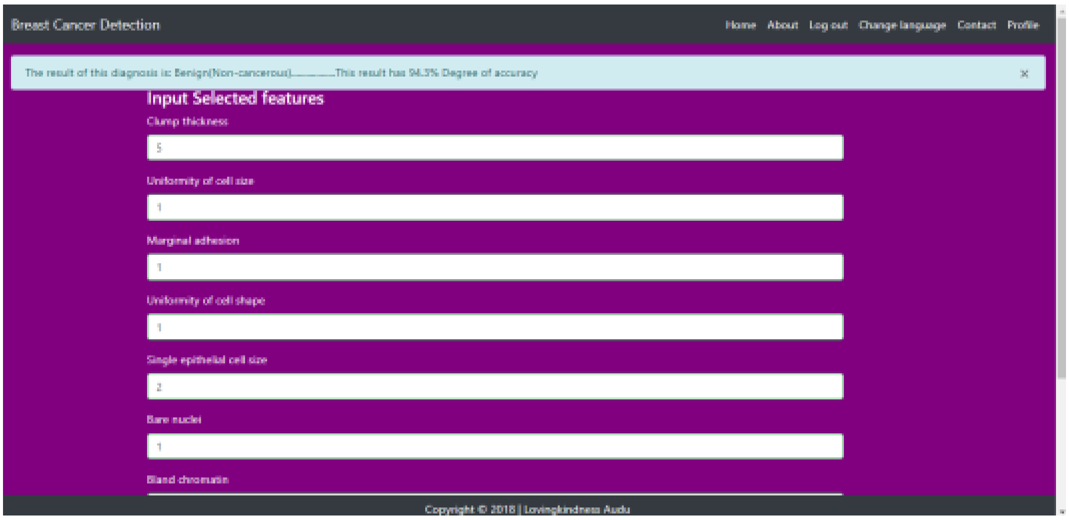
Bare nuclei: 1

Bland chromatin

Copyright © 2018 | Lovingkindness Audu

Fig. 6: User Interface page

As shown in figure 7, the selected features were received as inputs and then processed. The inputs were tested using the machine learning model created, and the result of the diagnosis proved it is benign (Non-cancerous).



The result of this diagnosis is: Benign(Non-cancerous)...............This result has 94.3% Degree of accuracy

Input Selected features

Clump thickness: 5

Uniformity of cell size: 1

Marginal adhesion: 1

Uniformity of cell shape: 1

Single epithelial cell size: 2

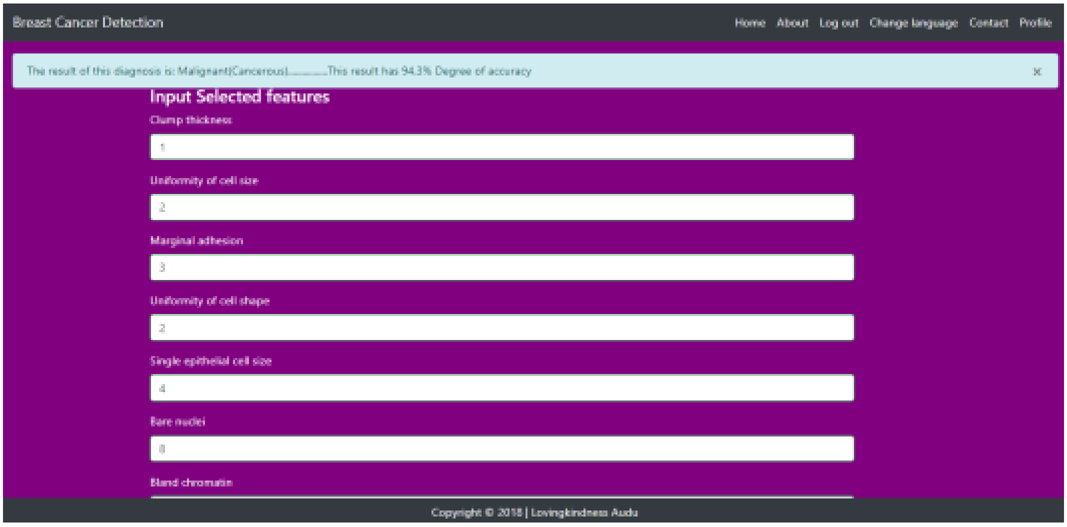
Bare nuclei: 1

Bland chromatin

Copyright © 2018 | Lovingkindness Audu

Fig. 7: Screenshot of Result (Non-cancerous)

From figure 8, the selected features were received as inputs and then processed. The inputs were tested using the machine learning model we created, and the result of the diagnosis proved it is malignant (cancerous).



The result of this diagnosis is: Malignant(Cancerous)..............This result has 94.3% Degree of accuracy

Input Selected features

Clump thickness: 1

Uniformity of cell size: 2

Marginal adhesion: 3

Uniformity of cell shape: 2

Single epithelial cell size: 4

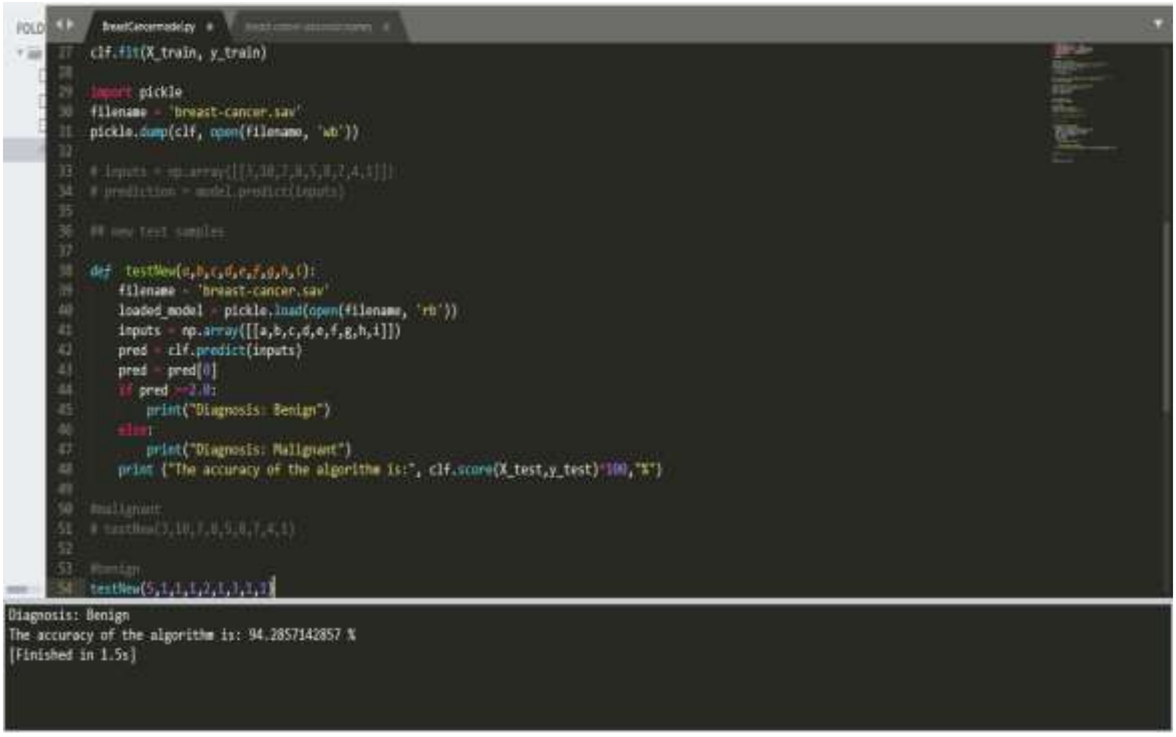
Bare nuclei: 8

Bland chromatin

Copyright © 2018 | Lovingkindness Audu

Fig. 8: Screenshot of Result (Cancerous)

Figure 9 shows the codes and model accuracy result of the algorithm which gives 94.28%



```
clf.fit(X_train, y_train)

import pickle
filename = 'breast-cancer.sav'
pickle.dump(clf, open(filename, 'wb'))

# inputs = np.array([[3,10,7,8,5,8,7,4,1]])
# prediction = model.predict(inputs)

## new test samples

def  testNew(a,b,c,d,e,f,g,h,i):
    filename = 'breast-cancer.sav'
    loaded_model = pickle.load(open(filename, 'rb'))
    inputs = np.array([[a,b,c,d,e,f,g,h,i]])
    pred = clf.predict(inputs)
    pred = pred[0]
    if pred ==2.0:
        print("Diagnosis: Benign")
    else:
        print("Diagnosis: Malignant")
    print ("The accuracy of the algorithm is:", clf.score(X_test,y_test)*100,"%")

#malignant
# testNew(3,10,7,8,5,8,7,4,1)

#benign
testNew(5,1,1,1,2,1,3,1,1)
```

```
Diagnosis: Benign
The accuracy of the algorithm is: 94.2857142857 %
[Finished in 1.5s]
```

Fig. 9: screenshot of the model (Result)

CONCLUSION

This research work is concerned with the development of a machine learning model for the extraction of features and the detection of breast cancer, and also a Django web application to demonstrate the applicability of the model. The web application authenticates a user and then allows the user to input nine features obtained from a test. These features include: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses. The system then predicts based on the inputs of the user, if the breast mass is benign (Cancerous) or malignant (non-cancerous). This research will be beneficial to the government most especially the health sector in their planning to minimize the spread of cancer related issues. At the end of this research, a machine learning model and a web application was developed which can be used for the detection of breast cancer with an accuracy level of 94.3%.

Research can still be carried out in this area (via machine learning) in other to come up with a more efficient model that will have a higher level of accuracy. Especially for breast cancer detection from mammograms (x-ray image of the breast), algorithms like the Convolutional Neural Network can be used to bring about a better model with a very reliable result.

REFERENCES

Ade-Ojo, Toluwani (2018). Development of an intelligent decision support system for prompt diagnosis of Ebola and Lassa fever disease (Doctoral dissertation, Federal University Oye- Ekiti).

ACS. (2017). Breast Cancer Basics. *About Breast Cancer*, 6-8.

Brook, A., El-Yaniv, R., Isler, E., Kimmel, R., Meir, R., and Peleg, D. (2008). Breast cancer diagnosis from biopsy images using generic features and SVMs. *Technion - Computer Science Department - Technical Report CS-2008-07 - 2008*, 1-16.

Charan, S., Khan, M. J., and Khurshid, K. (2018). Breast cancer detection in mammogram using convolutional neural network. *Department of Electrical Engineering Institute of Space Technology*, 1-6.

Fallahi, A., & Jafari, S. (2011). An expert system for detection of breast cancer using data preprocessing and bayesian network. *International Journal of Advanced Science and Technology Vol. 34*, 1-6.

Hamouda S, (2017). Enhancement accuracy of breast tumor diagnosis in digital mammograms. *Journal of Biomedical Sciences*, 1-8.

Higa, A. (2018). Diagnosis of breast cancer using decision tree and artificial neural network algorithms. *International Journal of Computer Applications Technology and Research*,7(1):1-6.

Khodary, S., El-Ezz, R. H., and Wahed, M. E. (2017). Enhancement accuracy of breast tumor diagnosis in digital mammograms. *Journal of Biomedical Science,* 6(4): 1-8.

Levy, D. and Jain, A. (2016). Breast mass classification from mammograms using deep convolutional neural networks. 1-6.

Lotfy, E. A. and Salem, A.-B. M. (2010). A breast cancer classifier based on a combination of case-based reasonong and ontology approach. *International Multiconference on Computer Science and Information Technology*, 1-8.

Nahid, A.A. and Kong, Y. (2018). Histopathological breast-image classification using local and frequency domains by convolutional neural network. *School of Engineering, Macquarie University, Sydney*, 1-26.

Narang, S., Verma, H. K., & Sachdev, U. (2012). Breast cancer detection using art2 model of neural networks. *International Journal of Computer Applications (0975 – 8887)*, 1-5.

Wang, L. (2017). Early diagnosis of breast cancer. *Sensors*, 1-20.