1. Introduction

1.1 Overview

A significant issue in the field of bioinformatics or medical science [1] is the accurate diagnosis of certain important information. The diagnosis of the disease is an energetic and tricky job in medicine domain. There is a huge amount of medical diagnosis data available in many diagnostic centers, hospitals, and research centers as well as on numerous websites. It is hardly necessary to classify them to make the system automated and quick diagnosis of diseases. The disease diagnosis is usually based on the knowledge and skill of the medical planning officer in the medical field. Because of this, there are circumstances of errors, unwanted biases, and also needs a long time for exact diagnosis of disease.

Conferring to the American Cancer Society [2], the ladies are affected by breast cancer in comparison to all other cancers already introduced. Estimation shows that the ladies will be affected with intrusive breast cancer approximately 252,710 and around 63,410 females will be detected within situ breast cancer in the United States in 2017. Men also have a greater chance of breast cancer. An estimation for men is that they will be affected by this cancer approximately 2470 in the United States in 2017. Another estimation shows that about 41,070 persons will die from this cancer in 2017. Recent statistics in the UK reports that 41,000 women are affected by breast cancer every year whereas only 300 men are affected by this disease.

Breast cancer is the leading cancer in females all over the world. Breast cancer is caused due to the abnormal growth of some cells in the breast. Several techniques have been introduced for the correct diagnosis of breast cancer. Breast screening or mammography [3] is a technique to diagnose breast cancer. It is used to check the nipple status of women through X-rays. Generally, it is almost impossible to detect breast cancer at the initial stage due to the small size of the cancer cell seen from outside. It is possible to diagnose cancer atc the early stage through mammography, and this test takes just a few minutes.

Ultrasound [4] is a familiar technique for the diagnosis of breast cancer in which the sound wave is sent inside the body to observe the condition inside. A transducer that emits sound waves is positioned on the skin and the echoes of the tissues of the body are captured with the bounce of sound waves. The echoes are transformed into a gray scale, i.e., a binary value which is represented in a computer. Positron emission tomography (PET) [5] imaging by means of F-fluorodeoxyglucose permits doctors to realize the position of a tumor in the human body. It is constructed on the recognition of radiolabeled cancer-specific tracers. Dynamic MRI [6] has developed the detection procedure for breast distortions. The modality predicts the speed of contrast enhancement by increasing the angiogenesis in cancer. Magnetic reasoning imaging associates with metastasis on contrast enhancement in breast cancer-affected people. Elastography [7] is a newly developed technique based on imaging technology. This technique is applicable when breast cancer tissue is more substantial than the adjacent regular parenchyma. The benign and malignant types are differentiated by a color map of probe compression in this approach.

In very recent years, various machine learning [8,9,10,11], deep learning [12, 13], and bio-inspired computing [14] techniques are used in several medical prognoses. Though a number of modalities have been demonstrated, none of the modalities are able to provide a correct and consistent result. In mammography, the doctors should read a high volume of imaging data which reduces the accuracy. This

procedure is also time-consuming, and in some worse case, detects the disease with the wrong outcome.

1.2 Purpose

An automatic disease detection system aids medical staffs in disease diagnosis and offers reliable, effective, and rapid response as well as decreases the risk of death. Purpose of this project is to create a web application along with machine learning model that predicts the diagnosis for breast cancer whether the cancer id Benign or Malignant. The Wisconsin Breast Cancer dataset is obtained from kaggle. This can be found on a prominent machine learning database named UCI machine learning database.

2. Literature Survey[30]

2.1 Existing Problem

With the evolution of medical research, numerous new systems have been developed for the detection of breast cancer. The research associated with this area is outlined in brief as follows.

Sakri et al. [15] focused on the enhancement of the accuracy value using a feature selection algorithm named as particle swarm optimization (PSO) along with machine learning algorithms K-NNs, Naive Bayes (NB) and reduced error pruning (REP) tree. Their work perspective holds the Saudi Arabian women's breast cancer problem, and according to their report, it is one of the major problems in Saudi Arabia. Their reports suggest that women with age range greater than 46 are the main victim of this malicious disease. Holding this sentiment, authors of [15] implemented five phase-based data analysis techniques on the WBCD dataset. They reported a comparative analysis between classification without feature selection method and classification with a feature selection method. They have acquired 70%, 76.3%, and 66.3% accuracy for NB, RepTree, and K-NNs, respectively. They used Weka tool for their data analysis purpose. With PSO implemented, they have found four features that are best for this classification task. For NB, RepTree, and K-NNs with PSO, they obtained 81.3%, 80%, and 75% accuracy values, respectively. Kapil and Rana [16] proposed a modified decision tree technique as a weight improved decision tree and implemented it on WBCD and another breast cancer dataset which is retrieved from the UCI repository. Using the Chi-square test, they have found that they have ranked each feature and kept the relevant features for this classification task. For the WBCD dataset, their proposed technique acquired approximately 99% accuracy, while for the breast cancer dataset, it acquired approximately 85–90% accuracy.

Yue et al. [17] mainly demonstrated comprehensive reviews on SVM, K-NNs, ANNs, and Decision Tree techniques in the application of predicting breast cancer on benchmark Wisconsin Breast Cancer Diagnosis (WBCD) dataset. According to the authors, deep belief networks (DBNs) approach with ANN architecture (DBNs-ANNs) has given the more accurate result. This architecture obtained 99.68% accuracy, whereas for the SVM method, the two-step clustering algorithm alongside the SVM technique has achieved 99.10% classification accuracy. They also reviewed the ensemble technique where SVM, Naive Bayes, and J48 were implemented using the voting technique. The ensemble method acquired 97.13% accuracy. Banu and Subramanian [18] have emphasized Naive Bayes techniques on breast cancer prediction and described a comparison study on Tree Augmented Naive Bayes (TAN), Boosted Augmented Naive Bayes (BAN) and Bayes Belief Network (BBN). They used SAS-EM (Statistical Analytical

Software Enterprise Miner) for the implementation of the models. The same popular WBCD dataset is used in their work. According to their findings with the help of gradient boosting 91.7%, 91.7%, and 94.11% accuracy have been achieved for BBN, BAN, and TAN, respectively. Hence, their research suggests that TAN is the best classifier among Naive Bayes techniques for this dataset. Chaurasia et al. [19] implemented Naive Bayes, RBF network, and J48 Decision Tree techniques on WBCD dataset. For their purpose of research, they used the Waikato Environment for Knowledge Analysis (WEKA) version 3.6.9 as a tool of analysis. For Naive Bayes, they obtained 97.36% accuracy which is greater than 96.77% and 93.41% accuracy values resulted from the RBF network and J48 Decision Tree, respectively.

Azar et al. [20] introduced a method for the prediction of breast cancer using the variants of decision tree. The modalities used in this technique are the single decision tree (SDT), boosted decision tree (BDT), and decision tree forest (DTF). The decision is taken by training the data set and after that testing. The outcomes presented that the accuracy obtained by SDT and BDT is 97.07% and 98.83%, respectively, in the training phase which clarifies that BDT performed better than SDT. Decision tree forest obtained an accuracy of 97.51% whereas SDT 95.75% in the testing phase. The dataset was trained by a ten-fold cross-validation fashion. In [21], the authors demonstrated a procedure for the detection of breast cancer. The experiments that have been done for detecting the disease are discussed here using local linear wavelet neural network (LLWNN), and recursive least square (RLS) to enhance the performance of the system. The LLWNN-RLS is providing the maximum values of average Correct Classification Rate (CCR) 0.897 and 0.972 for 2 and 3 predictors, respectively, with a few calculation times. It also provides the lowest value of minimum description length (MDL) and average squared classification error (ASCE) with much lesser time.

Senapati et al. [22] proposed a hybrid system for the detection of breast cancer using KPSO and RLS for RBFNN. The centers, as well as variances of RBFNN, are adjusted using K-particle swarm optimization and adjusted using back-propagation. The classification accuracy achieved by RBFNN-KPSO and RBFNN-extended Kalman filter is 97.85% and 96.4235%, respectively, whereas the coverage time is 8.38 s and 4.27 s, respectively. Hasan et al. [23] developed a mathematical model for the prediction of breast cancer based on the symbolic regression of Multigene Genetic Programming. The ten-fold technique is used to avoid overfitting here. A comparative study is also illustrated. The stopping criteria for the model were generated but the generation level did not reach zero. The highest accuracy obtained by the model is 99.28% with 99.26% precision. A variant of SVM [24] is introduced for the diagnosis of breast cancer. Here six kinds of SVM are explained and used for performance evaluation. The standard SVM results are compared with other types of SVM. Four-fold cross-validation is used for training and testing. The highest accuracy, specificity, and sensitivity achieved by St-SVM are 97.71%, 98.9%, and 97.08%, respectively, in the training phase. The highest accuracy, sensitivity, and specificity obtained by NSVM, LPSVM, SSVM, and LPSVM are 96.5517% 98.2456%, 96.5517%, and 97.1429% individually in the testing phase.

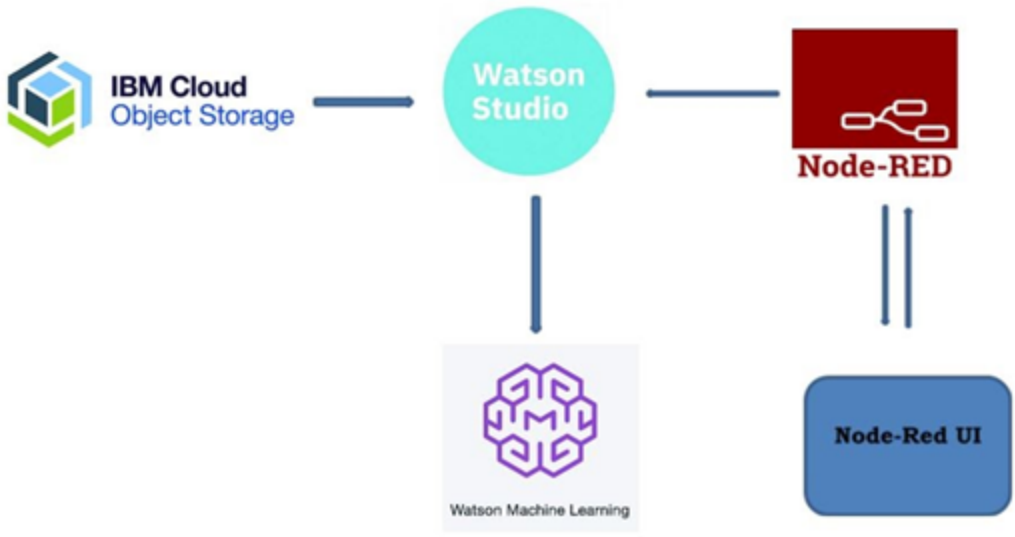
The authors in [25] presented an efficient method for the detection of breast cancer by categorizing the features of breast cancer data utilizing the inductive logic programming technique. A comparison study with a propositional classifier is also drawn. Kappa statistics, F-measure, area under the ROC curve, true-positive rate, etc. are calculated as a performance measure. The system simulated in two platforms named Aleph and WEKA. Jhajharia et al. [26] appraised variants of decision tree algorithms for the diagnosis of breast cancer. The system used the most common decision tree algorithms named CART and C4.5 which are simulated in the WEKA platform using Matlab and Python. The CART implemented in

Python achieved the highest accuracy 97.4% and the highest sensitivity 98.9% is obtained in the CART which is implemented in Matlab, and 95.3% specificity is acquired by CART and C4.5, respectively, which are simulated in WEKA. Some of the smart healthcare systems [27, 28] are developed in the IoT environment for the initial treatment of such types of diseases.

The authors in [30] presented a comparative study of five machine learning techniques for the prediction of breast cancer, namely support vector machine, K-nearest neighbors, random forests, artificial neural networks, and logistic regression. The basic features and working principle of each of the five machine learning techniques were illustrated. The highest accuracy obtained by ANNs is 98.57% whereas the lowest accuracy derived from the RFs and LR is 95.7%.

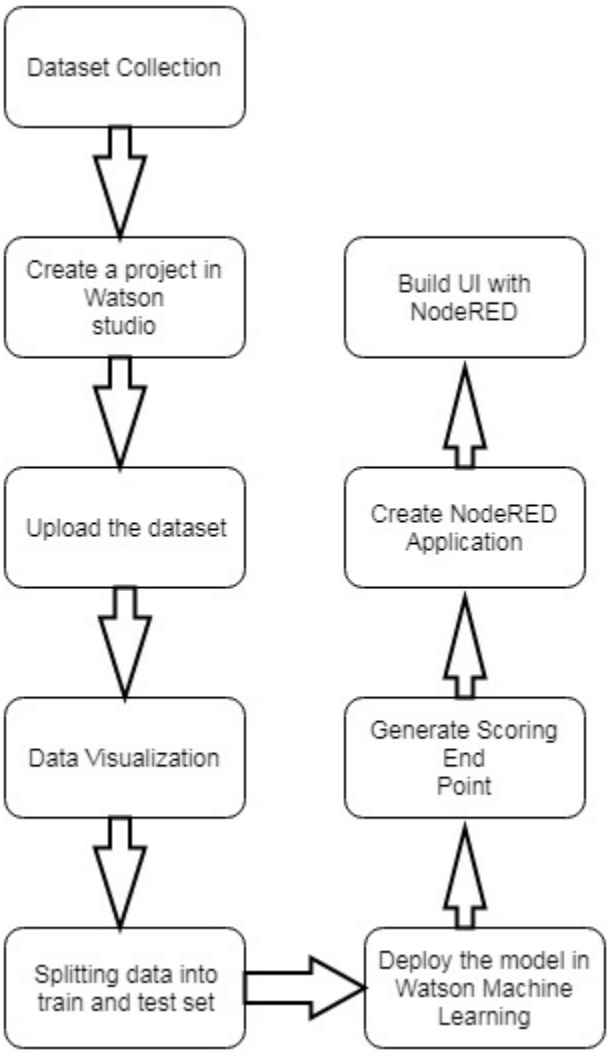
2.2 Proposed Solution

In the proposed work, a model is built that is capable of detecting the Breast Cancer in early stages. The Machine learning model is trained and deployed on IBM Watson Studio and an endpoint is created. The web application is built using IBM Node-Red. Below is the technical architecture.



3. Theoretical Analysis

3.1 Block Diagram



3.2 Hardware/Software designing

- OS: Windows 7+
- Programming Language: Python
- Platform: IBM Watson Studio
- Services: IBM Cloud Object Storage/Watson Machine Learning
- UI Services: NodeRED

4. Experimental Investigations

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

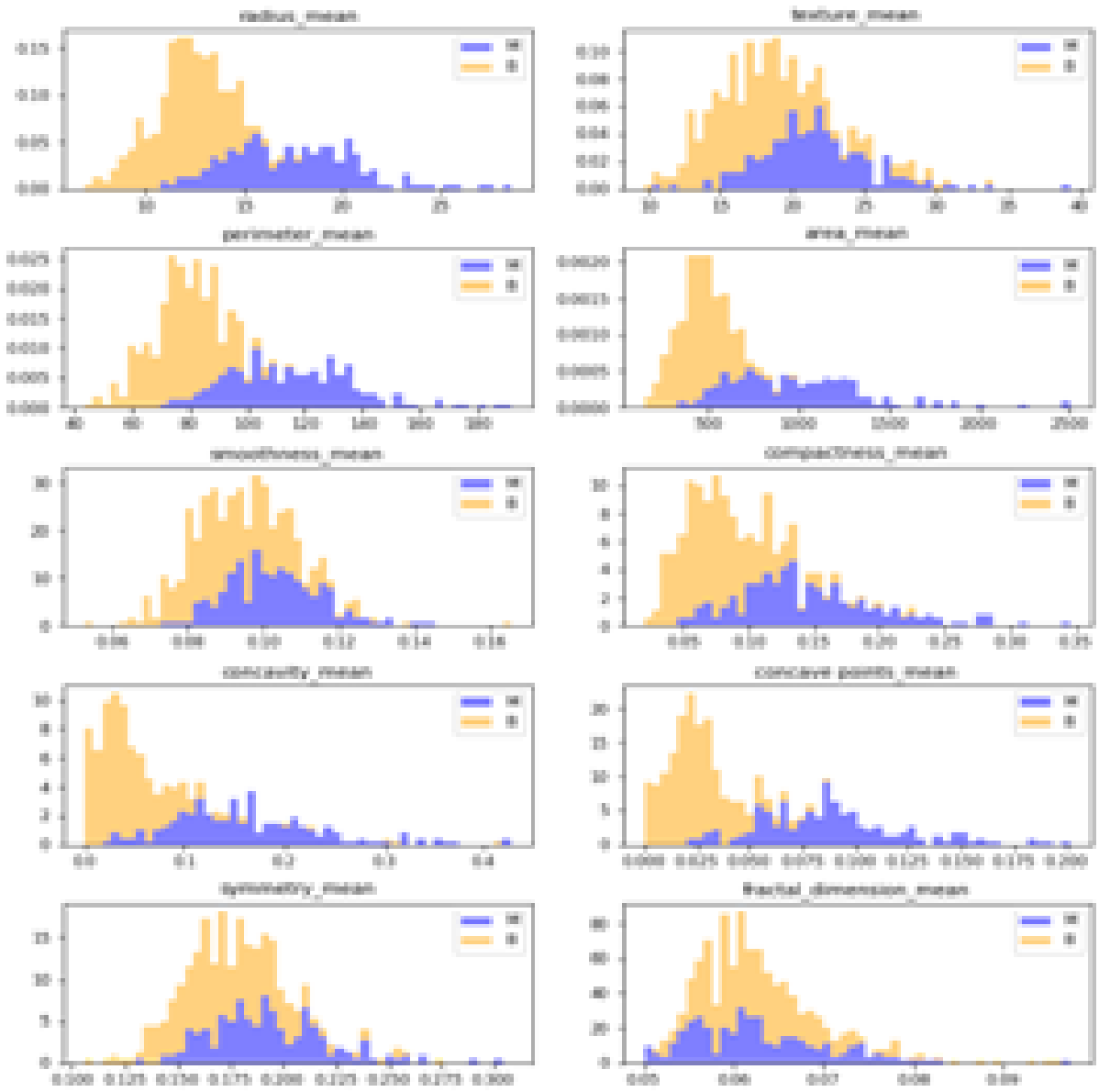
The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

Observations while data exploring:

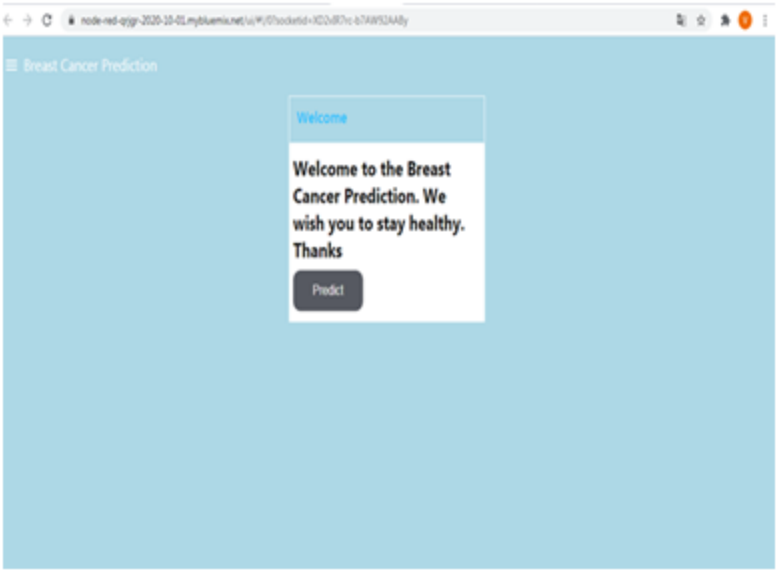
1. mean values of cell radius, perimeter, area, compactness, concavity and concave points can be used in classification of the cancer. Larger values of these parameters tends to show a correlation with malignant tumors.
2. mean values of texture, smoothness, symmetry or fractual dimension does not show a particular preference of one diagnosis over the other. In any of the histograms there are no noticeable large outliers that warrants further cleanup.



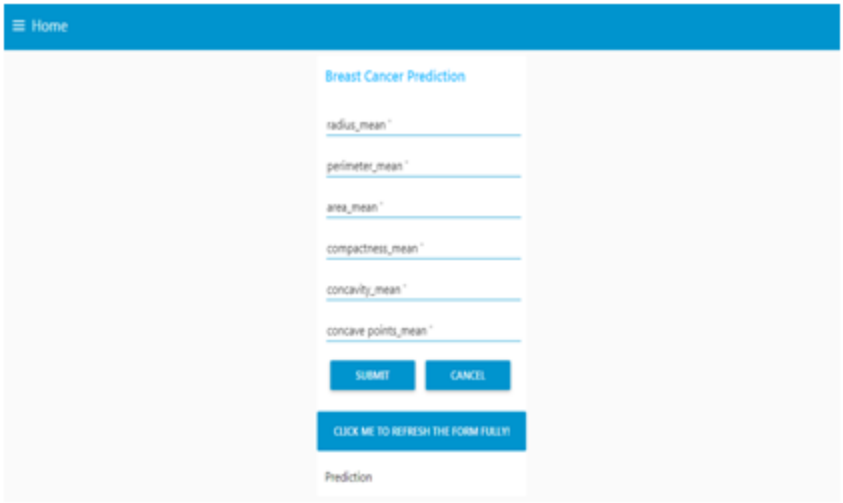
Breast cancer project created in Watson studio using auto AI experiment. After exploring it was found that only six attributes are sufficient for prediction as discussed above and dataset was trained on those six attributes with 90% training data and 10% holdout data. 3 folds were used for validation. Total 8 pipelines were generated. The performance of the project is measured with respect to accuracy. All the algorithms suggested by Watson are taken into consideration from which LGBM classifier stood top by giving 92.8% accuracy. After deploying the model, user interface was built using nodeRED service that consists of two pages: Welcome page and prediction page. In the next section screenshots are attached when running the project.

5. Results

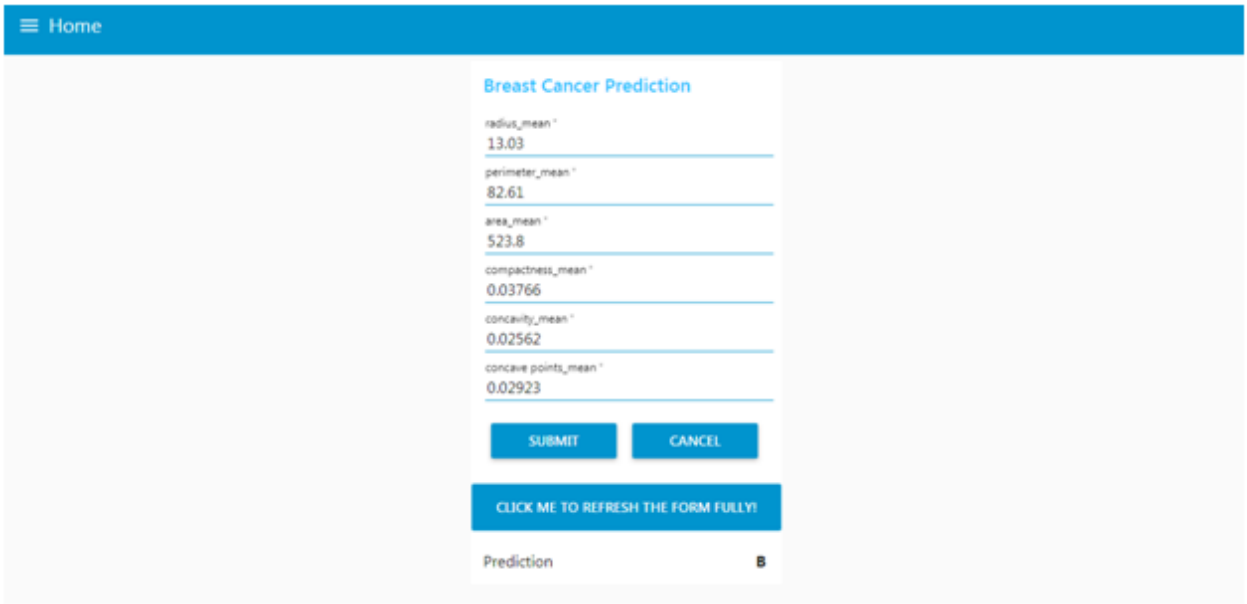
Welcome Page



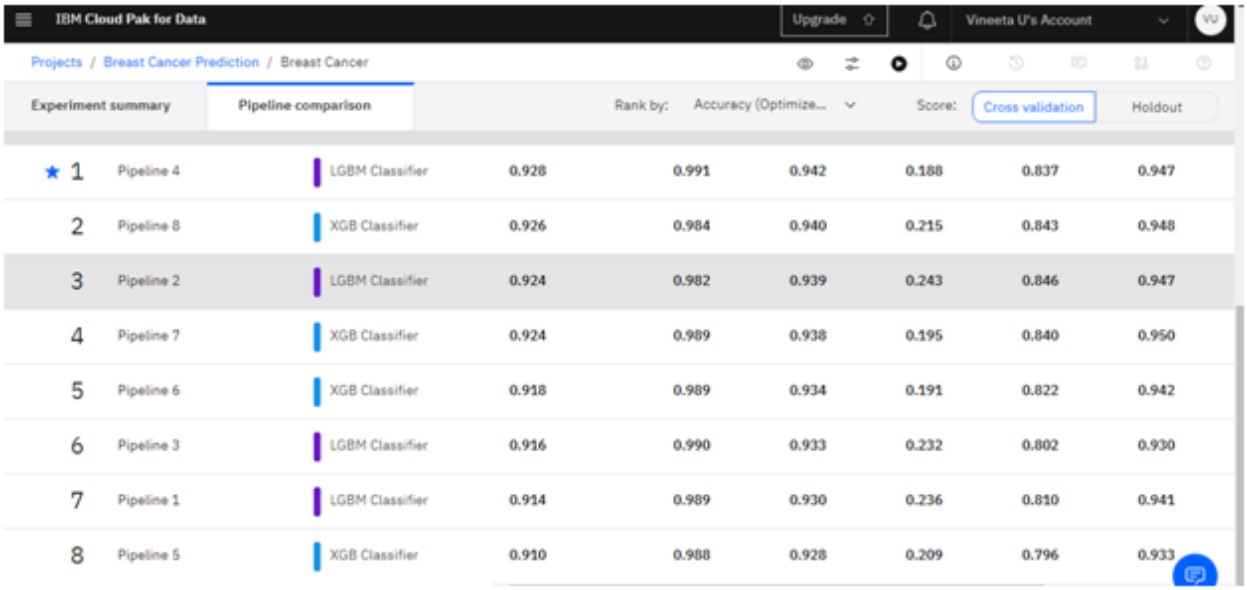
Prediction Page



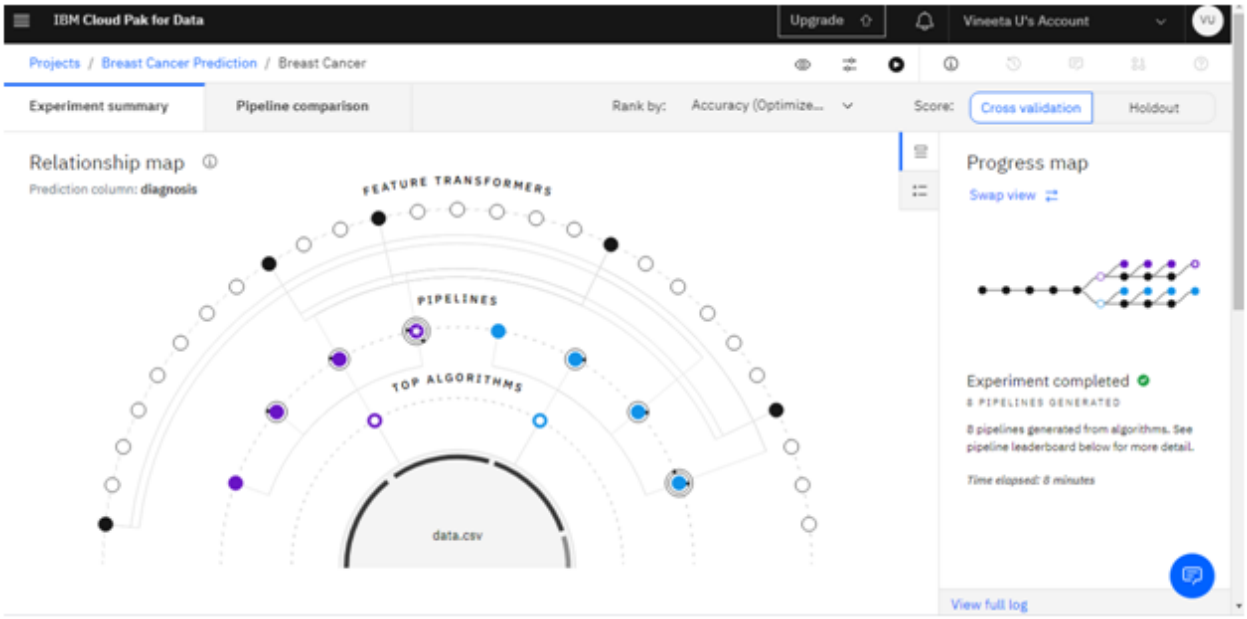
Results after Prediction



Pipeline comparison in cloud



Experiment Summary



6. Applications

Breast cancer is one of the main causes of cancer death worldwide. Early diagnostics significantly increases the chances of correct treatment and survival, but this process is tedious and often leads to a disagreement between pathologists. Computer-aided diagnosis systems showed the potential for improving diagnostic accuracy. But early detection and prevention can significantly reduce the chances of death. It is important to detect breast cancer as early as possible. The dashboard created can be used by healthcare professionals for early detection of breast cancer after getting lab reports.

7. Conclusion

Early detection of disease has become a crucial problem due to rapid population growth in medical research in recent times. With the rapid population growth, the risk of death incurred by breast cancer is rising exponentially. Breast cancer is the second most severe cancer among all of the cancers already unveiled. An automatic disease detection system aids medical staffs in disease diagnosis and offers reliable, effective, and rapid response as well as decreases the risk of death [30]. In this project, a machine learning model is developed that detects the breast cancer and predicts whether it is Benign or Malignant. Along with that user interface is also built to interact with the model.

8. Bibliography

1. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiol Soc N Am. 2018;286(3):800–9.
2. Breast Cancer: Statistics, Approved by the Cancer.Net Editorial Board, 04/2017. [Online]. Available: http://www.cancer.net/cancer-types/breast-cancer/statistics. Accessed 26 Aug 2018.
3. Mori M, Akashi-Tanaka S, Suzuki S, Daniels MI, Watanabe C, Hirose M, Nakamura S. Diagnostic accuracy of contrast-enhanced spectral mammography in comparison to conventional full-field digital mammography in a population of women with dense breasts. Springer. 2016;24(1):104–10.
4. Kurihara H, Shimizu C, Miyakita Y, Yoshida M, Hamada A, Kanayama Y, Tamura K. Molecular imaging using PET for breast cancer. Springer. 2015;23(1):24–32.
5. Azar AT, El-Said SA. Probabilistic neural network for breast cancer classification. Neural Comput Appl. 2013;23(6):1737–51.
6. Nagashima T, Suzuki M, Yagata H, Hashimoto H, Shishikura T, Imanaka N, Miyazaki M. Dynamic-enhanced MRI predicts metastatic potential of invasive ductal breast cancer. Springer. 2002;9(3):226–30.
7. Park CS, Kim SH, Jung NY, Choi JJ, Kang BJ, Jung HS. Interobserver variability of ultrasound elastography and the ultrasound BI-RADS lexicon of breast lesions. Springer. 2013;22(2):153–60.
8. Ayon SI, Islam MM, Hossain MR. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE J Res. 2020;.
https://doi.org/10.1080/03772063.2020.1713916.
9. Muhammad LJ, Islam MM, Usman SS, Ayon SI. Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery. SN Comput Sci. 2020;1(4):206.
10. Islam MM, Iqbal H, Haque MR, Hasan MK. Prediction of breast cancer using support vector machine and K-Nearest neighbors. In: Proc. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 226–229.
11. Haque MR, Islam MM, Iqbal H, Reza MS, Hasan MK. Performance evaluation of random forests and artificial neural networks for the classification of liver disorder. In: Proc. International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, 2018, pp. 1–5.
12. Ayon SI, Islam MM. Diabetes prediction: a deep learning approach. Int J Inf Eng Electron Bus (IJIEEB). 2019;11(2):21–7.
13. Islam MZ, Islam MM, Asraf A. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, 2020. pp. 1–20.
14. Hasan MK, Islam MM, Hashem MMA. Mathematical model development to detect breast cancer using multigene genetic programming. In: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 574–579, 2016.
15. Sakri SB, Rashid NBA, Zain ZM. Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access. 2018;6:29637–47.
16. Juneja K, Rana C. An improved weighted decision tree approach for breast cancer prediction. In: International Journal of Information Technology, 2018.
17. Yue W, et al. Machine learning with applications in breast cancer diagnosis and prognosis. Designs.

2018;2(2):13.
18. Banu AB, Subramanian PT. Comparison of Bayes classifiers for breast cancer classification. Asian Pac J Cancer Prev (APJCP). 2018;19(10):2917–20.
19. Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. J Algorithms Comput Technol. 2018;12(2):119–26.
20. Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. Neural Comput Appl. 2012;23(7–8):2387–403.
21. Senapati MR, Mohanty AK, Dash S, Dash PK. Local linear wavelet neural network for breast cancer recognition. Neural Comput Appl. 2013;22(1):125–31.
22. Senapati MR, Panda G, Dash PK. Hybrid approach using KPSO and RLS for RBFNN design for breast cancer detection. Neural Comput Appl. 2014;24(3–4):745–53.
23. Hasan MK, Islam MM, Hashem MMA (2016) Mathematical model development to detect breast cancer using multigene genetic programming. In: Proc. 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, 2016, pp. 574–579.
24. Azar AT, El-Said SA. Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Comput Appl. 2013;24(5):1163–77.
25. Ferreira P, Dutra I, Salvini R, Burnside E. Interpretable models to predict Breast Cancer. In: Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, 2016, pp. 1507–1511.
26. Jhajharia S, Verma S, Kumar R. A cross-platform evaluation of various decision tree algorithms for prognostic analysis of breast cancer data. In: Proc. International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1–7.
27. Islam MM, Rahaman A, Islam MR. Development of smart healthcare monitoring system in IoT environment. SN Comput Sci. 2020;1(3):185.
28. Rahaman A, Islam M, Islam M, Sadi M, Nooruddin S. Developing IoT based smart health monitoring systems: a review. Rev d'Intell Artif. 2019;33(6):435–40.
29. Breast Cancer Wisconsin (Original) Data Set, [Online]. https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data.
30. Islam, M.M., Haque, M.R., Iqbal, H. et al. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN COMPUT. SCI. 1, 290 (2020). https://doi.org/10.1007/s42979-020-00305-w