CSE 474

PROJECT : Breast Cancer Prediction

Name: Ankita Roy
ID: 20301482
Section: 1

Introduction

In the medical and healthcare fields, breast cancer prediction has long been viewed as a significant research issue. The tissue of the breast is where this cancer grows. Female sex, obesity, a lack of exercise, alcohol consumption, hormone replacement treatment during menopause, ionizing radiation, an early age at first menstruation, having children later in life or not at all, and advanced age are some of the risk factors for breast cancer. Therefore, a system that allowed for early identification and prevention and hence increased the survival rates for breast cancer would be highly helpful.

Several medical prognoses now incorporate a variety of machine learning, deep learning, and bio-inspired computer methodologies. While a numerous modalities have been tested, but none of them can deliver an outcome that is accurate and reliable. The condition was correctly diagnosed using four supervised machine learning techniques. The remainder of the essay is divided into the following sections. The literature review or related work is described in detail, along with all the results, in the next section. The following section provides examples of each machine learning technique's theoretical notion. The parameters for performance measurement are then explained. Before the last portion, the experimental design and result analysis are examined. A conclusion is reached in the last section.

Literature Review

Numerous innovative new technologies for the diagnosis of breast cancer have been developed with the advancement of medical research.The following is a brief summary of the research in this field.

By combining the machine learning methods K-NNs, Naive Bayes (NB), and reduced error pruning (REP) tree with the feature selection algorithm particle swarm optimization (PSO), Sakri et al., 2018 concentrated on improving the accuracy value. Their area of expertise includes the issue of breast cancer in Saudi Arabian women, which is one of the country's key issues, per their study. According to their reports, women over the age of 46 seem to be the main targets of this malicious disease. Authors of Sakri et al., 2018 performed five phase-based data analysis approaches on the WBCD dataset while maintaining this viewpoint. They published a comparison of classifiers that do not use feature selection methods versus classifiers that do use feature selection methods. For NB, RepTree, and K-NNs, they have obtained accuracy of 70%, 76.3%, and 66.3%, respectively. They used the Weka tool to analyze the data. They have discovered four features that work best for this classification assignment when PSO is used. They found accuracy values of 81.3%, 80%, and 75% for NB, RepTree, and K-NNs with PSO, respectively.

On the WBCD and another breast cancer dataset that was downloaded from the UCI library, Juneja and Rana, 2020 adopted a modified decision tree technique they had proposed as a weight improved decision tree. They discovered that they have ranked each characteristic and retained the important features for this classification job using the Chi-square test. Their suggested method achieved an accuracy of about 97% for the WBCD dataset and between 85 - 90% for the breast cancer dataset.

Using the benchmark Wisconsin Breast Cancer Diagnosis (WBCD) dataset,Yue et al., 2018's detailed reviews of SVM, K-NNs, ANNs, and Decision Tree approaches were applied to the prediction of breast cancer. This architecture produced accuracy of 99.68%, whereas the two-step clustering approach used in conjunction with the SVM technique achieved accuracy of 99.10%. They also discussed the ensemble technique, which employed the voting method to implement SVM, Naive Bayes, and J48. Accuracy of the ensemble approach was 97.13%.

Thirumalaikolundusubramanian et al., 2018 highlighted Naive Bayes strategies for breast cancer prediction and described a comparison study on Tree Augmented Naive Bayes (TAN), Boosted Augmented Naive Bayes (BAN), and Bayes Belief Network (BBN). According to their findings, accuracy has been attained for BBN, BAN, and TAN at 91.7%, 91.7%, and 94.11%, respectively, with the aid of gradient boosting. Therefore, according to their research, TAN is the most effective classifier among Naive Bayes approaches for this dataset.

On the WBCD dataset, Chaurasia et al., 2018 applied Naive Bayes, RBF networks, and J48 Decision Tree algorithms. They employed the Waikato Environment for Knowledge Analysis (WEKA) version 3.6.9 as an analysis tool for their research. They found that Naive Bayes had an accuracy of 97.36%, which is higher than the accuracy values of the RBF network and J48 Decision Tree, which were 96.77% and 93.41%, respectively.

Azar and El-Metwally, 2019 first described a decision tree-based method for predicting breast cancer. The single decision tree (SDT), boosted decision tree (BDT), and decision tree forest are the modalities employed in this method (DTF). The results showed that the accuracy attained by SDT and BDT in the training phase was 97.07% and 98.83%, respectively. This clearly shows that BDT outperformed SDT. In the testing phase, decision tree forest achieved an accuracy of 97.51% versus SDT's 95.75%.

A SVM variation is presented Azar and El-Said, 2020 for the detection of breast cancer. Here, six different SVM types are described and utilized to assess performance. The outcomes of conventional SVM are contrasted with those of other SVM variants. For training and testing, fourfold cross-validation is employed. In the training phase, St-SVM achieves the greatest accuracy, specificity, and sensitivity values of 97.71%, 98.9%, and 97.08%, respectively. In the testing phase, NSVM, LPSVM, SSVM, and LPSVM each achieved the highest accuracy, sensitivity, and specificity scores of 96.5517%, 98.2456%, 96.5517%, and 97.1429%.

The authors of Ferreira et al., 2016 described an effective

strategy for the identification of breast cancer by classifying the characteristics of breast cancer data using the inductive logic programming methodology. Additionally, a comparative study using a propositional classifier is made. As a performance indicator, the following metrics are calculated: kappa statistics, F-measure, area under the ROC curve, true-positive rate, etc. Two platforms called Aleph and WEKA were used to mimic the system.

Methodology

Dataset Description

The analysis makes use of the Wisconsin Diagnostic Breast Cancer (WDBC) data set. There are around 569 instances with 32 attributes in the collection.
The following 10 attributes are employed for the study work: radius (mean distance from center to points on the perimeter), texture, perimeter, area, smoothness, compactness, concavity concave points, symmetry, and fractal dimension. Diagnosis is used as a classification method among the attributes. These characteristics are calculated from a digital picture of a breast mass that was aspirated with a fine needle (FNA). They outline the features of the cell nuclei seen in the picture. So,for the comparison, the key attributes are extracted.

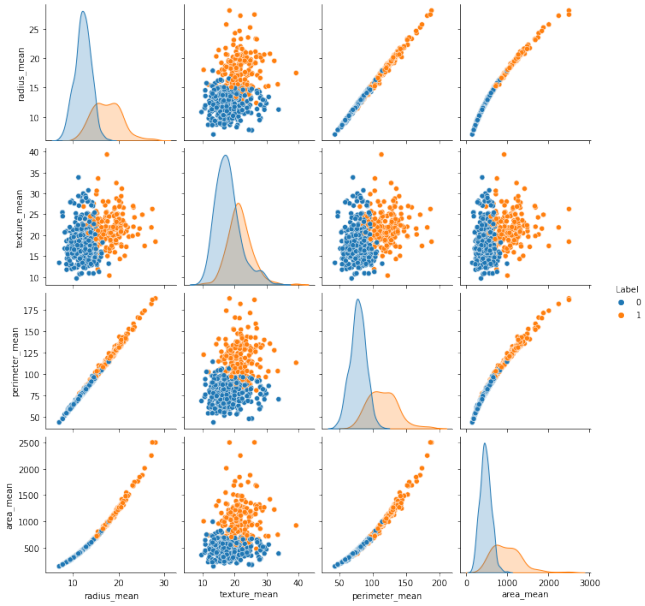

Figure 1: Seeing pairwise relation using pairplot

Data Cleaning and Pre-Processing

There is a particular ”id” that cannot be classified. Our class label is diagnosis. ”Unnamed”:32 feature in NAN. Thus, we do not require it. So, we will drop these superfluous elements. After that we used LabelEncoder to covert the value of M and B to 1 and 0.

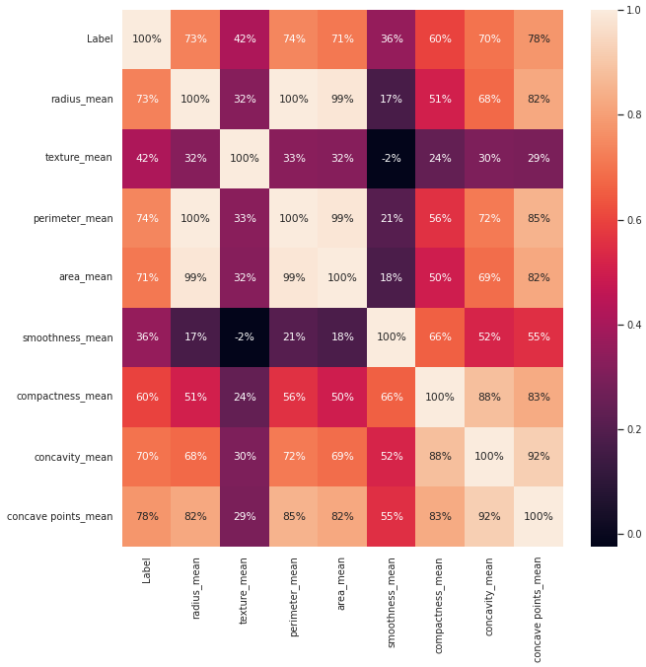

Figure 2: Seeing co-relation between features

From figure2 we can see, radius mean and perimeter mean has the highest corelation which is 100%. Area mean, radius mean (92%) and concavity mean (92%) are also co-realted with each other.

Data Spliting

In the data splitting phase, we divided the data into two datasets: the training dataset and the test dataset. The standard proportion of the splitting process is 70% training and 30% testing. The purpose of splitting the data is to ensure that the model is not overfitted during the model testing with the testing dataset.

Model

Two essential parts that are used in supervised learning are the training data set and the testing data set. To verify the model, test data is frequently randomly selected from the full database. In this study, LR, RF, ANN, and KNN are used for testing.

1. Logistic Regression

Logistic Regression is an analytical modeling technique where the likelihood of a level is associated with a set of explicative variables. It is used for analyzing a dataset in which there are one or more independent variables that decide a result. The result is measured with a binary variable (in which there are only two possible results). It is applied to predict a binary result (True/False, 1/0, Yes/No). [Islam, 2019]

2. Random Forests

Random forest classifier is a powerful supervised classification tool. RF generates a forest of classification trees from a given dataset, rather than a single classification tree. Each of these trees produces a classification for a given set of attributes. [Islam, 2019]

3. K-Nearest Neighbors

K-nearest neighbor algorithm is utilized for grouping and used in pattern recognition. It is widely used in predictive analysis. On the arrival of new data, the K-NN algorithm identifies existing data points that are nearest to it. Any attributes that can differ on a large scale may have sufficient influence on the interval between data points. Given a positive integer k, k -nearest neighbors looks at the k observations closest to a test observation x_0 and estimates the conditional probability that it belongs to class j using the formula

$$Pr(Y = j|X = x_0) = \frac{1}{k} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

Where $mathcalN_0$ is the set of k-nearest observations and I($y_i = j$) is an indicator variable that evaluates to 1 if a given observation in $mathcalN_0$ is a member of class j, and 0 if otherwise.

4. Artificial Neural Networks

The term "Artificial Neural Network" is derived from Biological neural networks that develop the structure of a human brain. Similar to the human brain that has neurons interconnected to one another, artificial neural networks also have neurons that are interconnected to one another in various layers of the networks.3

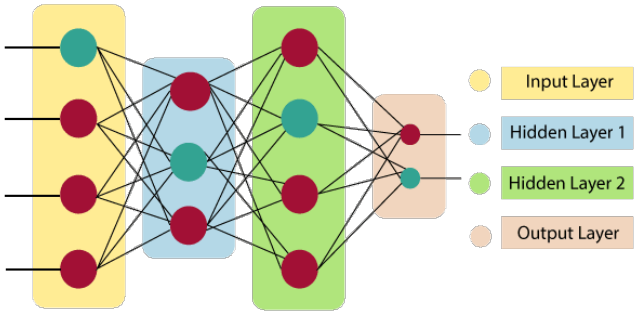

Figure 3: ANN illustration

Input Layer:
As the name suggests, it accepts inputs in several different formats provided by the programmer.

Hidden Layer:
The hidden layer presents in-between input and output layers. It performs all the calculations to find hidden features and patterns.

Output Layer:
The input goes through a series of transformations using the hidden layer, which finally results in output that is conveyed using this layer. The artificial neural network takes input and computes the weighted sum of the inputs and includes a bias. This computation is represented in the form of a transfer function.

$$\sum_{i=1} W_i * X_i + b$$

It determines weighted total is passed as an input to an activation function to produce the output. Activation functions choose whether a node should fire or not. Only those who are fired make it to the output layer. There are distinctive activation functions available that can be applied upon the sort of task we are performing.[JavaPoint, 2020]

PCA

The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D.

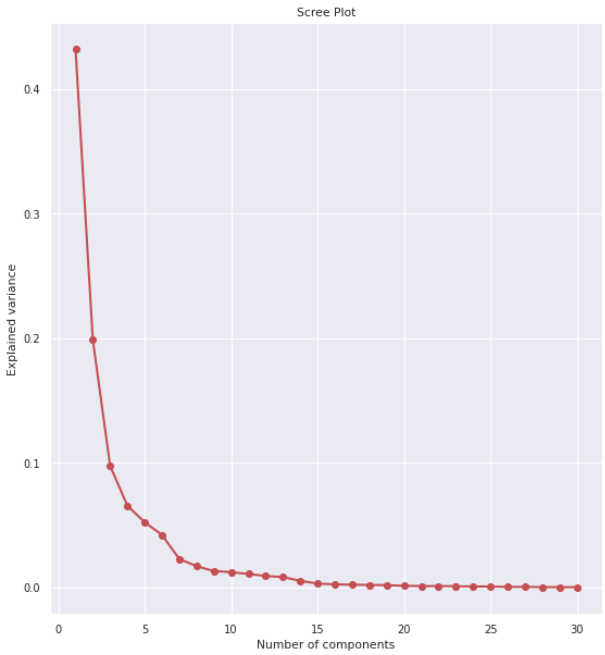

Figure 4: Scree plot

A scree plot is applied 4 in the research to decide the necessary amount of PC, in which a notable drop in the scree plot represents the suitable number of PCs. Based on Figure d thirteen out of thirty PCs are chosen. PCA results was used to build hybrid ANN model.

Result

I have applied four different models in our dataset and one hybrid ANN with PCA. The confusion matrix is calculated for each technique. From the dataset of 569 instances, I used 70% of the total data to train for all four techniques. I used 30% to train both the trained models. The confusion matrix of our best model is illustrated in Fig 5. Where true positive value was 59 and 108 true negative value. Also false positive value was 4 and false negative value was 0.

Figure 5: ANN confusion matrix

Performances of breast cancer prediction system					
	LR	RF	KNN	ANN	ANN with PCA
Accuracy(%)	97.66	97.07	96.49	99.42	96.23
Precision(%)	98.36	98.33	98.30	100	98
Recall(%)	95.23	93.65	92.06	17.46	98
F1-score(%)	96.77	95.93	95.08	29.72	98

From the table above we can see that ANNs outperformed all other machine learning technique with 99.42% accuracy. Where LR has the second highest accuracy of 97.66%. Additionally, the highest precision of 100% was achieved by ANN. Also ANN with PCA has the best and highest recall of all which is 98%. All the techniques have an F1 score of 97% except ANN (29.72%) which is comparatively better.

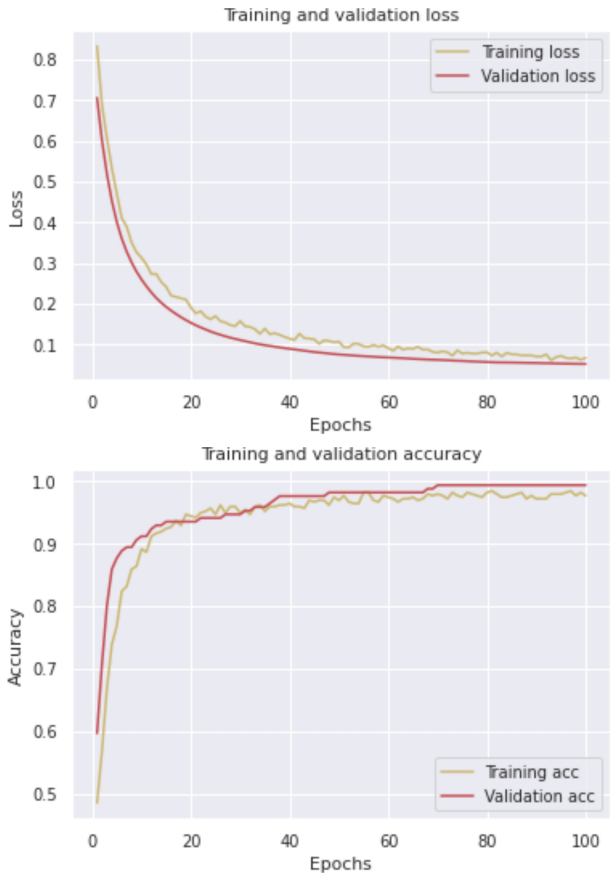

Figure 6: Training and validation loss and accuracy

From Figure 6 the training and validation loss along with training and validation accuracy was plotted. We can see the loss graph in training and validation was almost similar. This is same for the next training and validation accuracy. Though the training accuracy fluctuated a lot but it showed quite impressive performance in the validation accuracy.

Comparison

From Table 1 we can clearly see others authors work on this dataset. The accuracy achieved by the Kernel-based tranform is 98.53%. The author Azar et al. measured the performance of SDT,BDT, DTF for the prediction of breast cancer. The accuracy obtained by the techniques is 95.75%, 97.07% and 97.51%. Local Linear wavelet network(LLWNN) obtained an accuracy of 97.2%. The classification accuracy achieved by RBFNN-KPSO is 97.85%and the RBFNN extended Kalman filter is 96.4235%. The accuracy obtained by LPSVM, LSVM, SSVM, PSVM, NSVM, St-SVM are 97.1429%, 95.4286%, 96.5714%, 96%, 96.5714% and 94.86%, respectively. The proposed system of Kumar et al. combined SVM, Naive Bayes, and J48 using the voting classifier method to achieve accuracy of 97.13% which is better than each of individual classifiers.

Authors	Method	Accuracy
Xu et al.	Kernel-based orthogonal transform	98.53
Azar et al.	SDT	95.75
Azar et al.	BDT	97.07
Azar et al.	DTF	95.51
Senapati et al.	LLWNN	97.20
Senapati et al.	RBFNN-KPSO	97.85
Senapati et al.	RBFNN extended	96.42
Azar et al.	LPSVM	97.14
Azar et al.	LSVM	95.42
Azar et al.	SSVM	96.57
Azar et al.	PSVM	96
Azar et al.	NSVM	96.57
Azar et al.	St-SVM	94.86
Kumar et al.	SVM-Naive Bayes-J48	97.13
Sakri et al.	Naive Bayes	81.3
Sakri et al.	RepTree	80
Sakri et al.	k-NNs	75
Banu and Subramanian	Bayes belief network	91.7
Banu and Subramanian	Boosted augmented Naive Bayes	91.7
Banu and Subramanian	Tree augmented Naive Bayes	94.11
Chaurasia et al.	Naive Bayes	97.36
Chaurasia et al.	RBF network	96.77
Chaurasia et al.	J48	93.41

Table 1: The comparison of our study with the state of the art

Disussion

While working on this project many problems arised. Such as while implementing decision tree our model was overfitted and providing 100% accuracy. PCA was surely significant while choosing features that are important for our project. As from the result we can see PCA increased the precision, recall and F1 score.

Conclusion and Future Work

This study compared the effectiveness of the machine learning techniques logistic regression, K-nearest neighbors, random forests, and artificial neural networks for predicting breast cancer. Each of the four machine learning approaches' fundamental characteristics and operation were demonstrated. The lowest accuracy came from ANNs using PCA, whilst the best accuracy was achieved by ANNs at 99.42%. In the medical field, the diagnosing process is highly time- and money-consuming. The system suggested that machine learning techniques might serve as a clinical aid for the detection of breast cancer and would be highly beneficial for newly qualified medical professionals or doctors in the event of a misdiagnosis. The constructed model via ANNs is more reliable than any other method mentioned, and it could be able to transform the field of breast cancer prediction. We may infer from the findings that machine learning approaches can accurately and automatically diagnose the illness. Future work will concentrate on investigating additional dataset values and producing more intriguing results. By lowering total cost, time, and mortality rate, this study can aid in establishing more accurate and reliable illness prediction and diagnostic systems, which will assist to develop a better healthcare system.

References

Azar, A. T., & El-Metwally, S. M. (2019). Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, *23*(7), 2387–2403.

Azar, A. T., & El-Said, S. A. (2020). Performance analysis of support vector machines classifiers in breast cancer mammography recognition. *Neural Computing and Applications*, *24*(5), 1163–1177.

Chaurasia, V., Pal, S., & Tiwari, B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*, *12*(2), 119–126.

Ferreira, P., Dutra, I., Salvini, R., & Burnside, E. (2016). Interpretable models to predict breast cancer. *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1507–1511.

Hasan, M. K., Islam, M. M., & Hashem, M. (2016). Mathematical model development to detect breast cancer using multigene genetic programming. *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, 574–579.

Https://colab.research.google.com/drivewezxybgoxq z8lmp_r6zq6ohydvwaz2df?uspsharingscrolltoȲhqev7jmc3z2. (n.d.).

Https://github.com/codeforlife200/breast_cancer_prediction. (n.d.).

Islam, M. M. (2019). Breast cancer prediction: A comparative study using machine learning techniques. *1*(1).

JavaPoint. (2020). Artificial neural network tutorial. *1*(1).

Juneja, K., & Rana, C. (2020). An improved weighted decision tree approach for breast cancer prediction. *International Journal of Information Technology*, *12*(3), 797–804.

Kaggle datasets download d uciml/breastcancerwisconsin-data. (n.d.).

LUGAT, V. (2019). Https://www.kaggle.com/code/ vincentlugat/breast-canceranalysisand-prediction. *1*(1).

Sakri, S. B., Rashid, N. B. A., & Zain, Z. M. (2018). Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access*, *6*, 29637–29647.

Senapati, M. R., Panda, G., & Dash, P. K. (2019). Hybrid approach using kpso and rls for rbfnn design for breast cancer detection. *Neural Computing and Applications*, *24*(3), 745–753.

Thirumalaikolundusubramanian, P., et al. (2018). Comparison of bayes classifiers for breast cancer classification. *Asian Pacific journal of cancer prevention: APJCP*, *19*(10), 2917.

Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, *2*(2), 13.