Improving Breast Cancer Detection in Fine-Needle Aspirate Biopsies through Machine Learning

Dhanvee Ivaturi | Silver Creek High School, San Jose, California | Philip Kabranov

1. Purpose

Breast cancer is one of the most common types of cancer and affects millions of women worldwide. The disease is difficult to diagnose in developing countries and is commonly misdiagnosed in developed countries. Thousands of lives are lost every year to undiagnosed and misdiagnosed cases. To combat this, we aim to find the optimal machine learning algorithm for diagnosing breast cancer.

2. Background

Breast cancer can be diagnosed with a fine needle aspirate (FNA) of the tumor in question. This aspirate is then viewed under a microscope to make a diagnosis. The Wisconsin Breast Cancer Database, made by the University of Wisconsin, contains 569 entries with 30 attributes each, and a malignant/benign diagnosis. The 30 attributes describe the cells seen in the FNA, describing their mean radius, texture, and more.

3. Objective

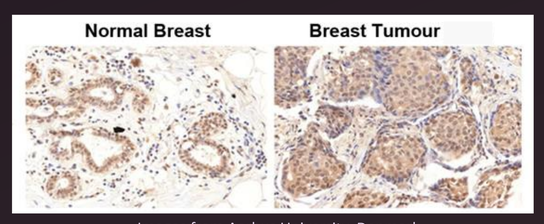
To develop a system that efficiently employs a predetermined machine learning algorithm to accurately diagnose breast cancer using data extracted from an FNA biopsy of a breast tumor, and to investigate the limits of this dataset as a benchmark for future cell image classification research.

4. Engineering Goals

- Test/investigate multiple machine learning algorithms with dataset:
  - Logistic Regression
  - Support Vector Machines
  - Neural Network
- Use data reduction techniques to analyze most impactful features
  - Principal Component Analysis → reducing dimensionality to 5
- Obtain optimal algorithm in terms of accuracy

5. Materials

- Software implemented using Python programming language
- SciKit Learn, Keras – Machine Learning libraries imported
- Pandas, Matplotlib, Numpy – visualizing data, vector computation
- Wisconsin Breast Cancer Dataset [1]
  - features are computed from a digitized image of a fine-needle aspirate from sample of breast tissue



Images from Aarhus University, Denmark

6. Dataset Description

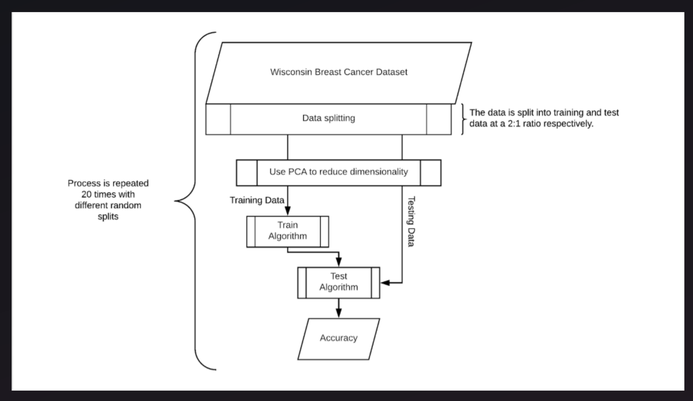
- Class Distribution: 357 benign (62.74%), 212 malignant (37.26%)

Feature Number in Dataset	Feature Description:
1	Sample ID Number
2	Diagnosis (1 = malignant, 0 = benign)
3-32	10 real-valued features are available/calculated for each cell nucleus. 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. 1. **Radius** (mean of distances from center to perimeter) 2. **Texture** (standard deviation of grey-scale values) 3. **Perimeter** 4. **Area** 5. **Smoothness** (local variation in radius lengths) 6. **Compactness** ((perimeter)2/area) 7. **Concavity** (severity of concave portions of the contour) 8. **Concave Points** (# of concave portions of the contour) 9. **Symmetry** 10. **Fractal Dimension** (coastline approximation -1)

7. Methodology

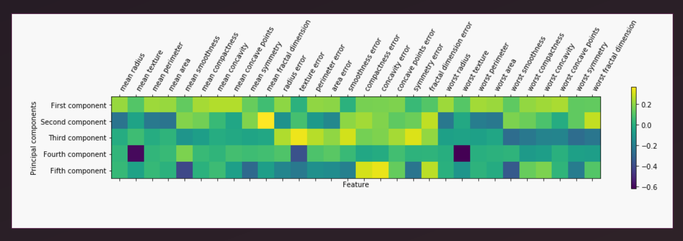
- Split the dataset into training and testing subsets in a 2:1 ratio
- Use PCA to reduce dimensionality (optional)
- Train the algorithm in question (LR, SVM, NN) using the training data obtained in steps 1 & 2.
- Test the trained algorithm using the testing data obtained in step 1; record the accuracy of the program
- Repeat steps 1 - 4 multiple times (20 recommended) to obtain more data for evaluating the algorithms.
- Compare the mean accuracies of the tested algorithms to find the optimal algorithm.

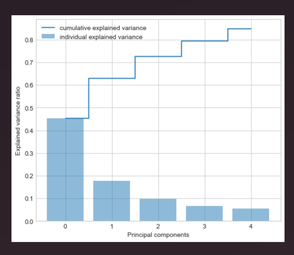
8. Pipeline



9. Relevant vs. Redundant Features

10. Principal Component Analysis (PCA)



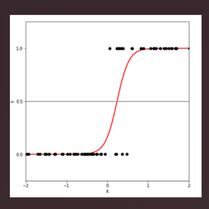


- PCA is used to distinguish relevant/strong patterns in the dataset
- 5 “new” independent variables created from 30 initial features, where each one is a composite of the “old” ones
- We select the “new” variables, preserving 95% of the variance of the “old” ones

11. Machine Learning Techniques

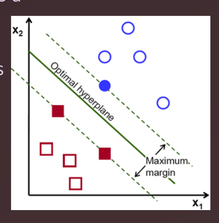
11.1. Logistic Regression (LR)

- Used to estimate probability that an instance belongs to a particular class
- If the estimated probability is greater than 50% than the estimator predicts the instance belongs to “positive” , otherwise it is classified as “negative"
- LR fits the data points as if they were along a continuous function: Logistic function $\sigma(t) = \frac{1}{(1+e^{-t})}$



11.2. Support Vector Machine (SVM)

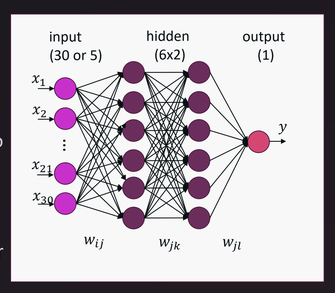
- Also estimates probability that an instance belongs to a particular class
- SVM finds the separating hyperplane that maximizes the distance of the closest points to the margin (the support vectors)
- A “hard margin” SVM will find a hyperplane that separates all the data (if one exists)
- A “Soft” margin will perform better if there is noise in the data (outliers possible)



11.3. Neural Networks (NN)

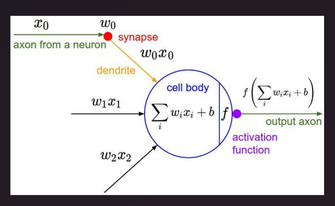
Learning algorithm has 2 phases:

- **1. Forward propagation:** The network computes the outputs and errors
- **2. Backward propagation:** Uses error to adjust weights between neurons through gradient descent algorithm
- The two phases are repeated until error is lower than a certain threshold

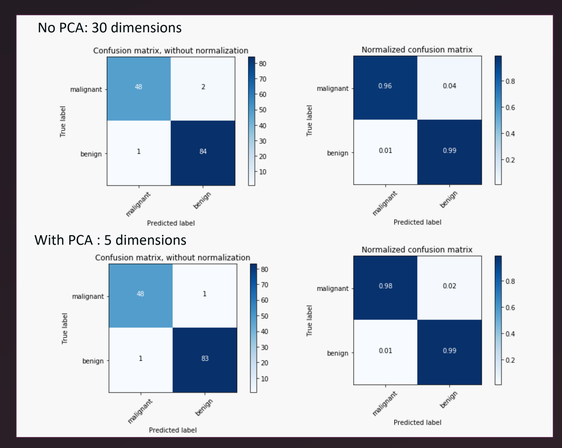


Neuron:

- Input ***X***
- Weight ***W***, bias ***b***
- ***f***: activation function (ReLU: hidden layers, Sigmoid: output layer)

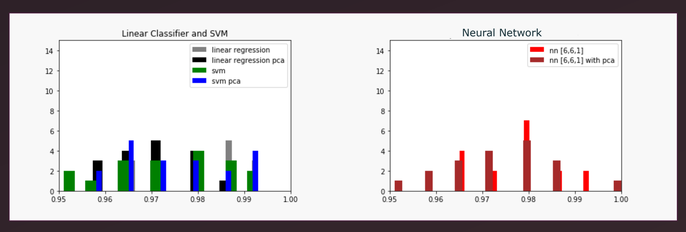


12. Confusion Matrices for Neural Networks



Note: Confusion matrices without normalization were rounded over 20 iterations of training (batch size = 10, # of epochs = 100)

13. Accuracy Distribution & Analysis



- The training is repeated 20 times using original (30 dimensions) and PCA reduced data (5 dimensions) for LR, SVM and Neural Networks
- Neural networks were generally most accurate (~0.5-1% improvement)
- The PCA-reduced data resulted in minimal accuracy deterioration (~0.1%)

14. Conclusions

Classification	Mean Accuracy (%)	Confidence Intervals (%) for α = 0.05
Logistic Regression	97.0924	(96.3534, 97.8314)
Logistic Regression (PCA)	97.2764	(96.6519, 97.9009)
SVM	97.0556	(96.3949, 97.7162)
SVM (PCA)	96.6945	(96.2691, 97.6212)
Neural Network	97.5340	(97.0270, 98.0411)
Neural Network (PCA)	97.3868	(96.8157, 97.9580)

- For the University of Wisconsin Breast Cancer Dataset the LR performs better than SVM, but best results are delivered by Neural Networks.
- The NNs have a narrower confidence interval compared to LR and SVM
- NNs are optimal choice for this particular dataset

15. Further Research

- Work with a university/hospital to obtain newer, more detailed data/images
- Use image recognition recognition/classification and train the networks with image data, without manual extraction of features from each probe.
- Use different NN architectures to compare performance.

Bibliography

[1] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[2] Elmore JG, Longton GM, Carney PA, Geller BM, Onega T, Tosteson ANA, Nelson HD, Pepe MS, Allison KH, Schnitt SJ, O’Malley FP, Weaver DL. Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. JAMA. 2015

[3] Müller, A. C. & Guido, S. *Introduction to Machine Learning with Python, A Guide for Data Scientists*. (O’Reilly Media, Inc., 2017).

Note: All images except those in section 2 were produced by finalists (Dhanvee, Philip).