CASE STUDY #5

Cancer Prediction

DA 6813 DATA-ANALYTICS APPLICATION| 28-NOV-2021

KRISCHELLE JOYNER | KATHY KEEVAN | DAYANIRA MENDOZA | CHARLES REYES

Executive Summary

The PCA_NNET provided the highest accuracy when analyzing the models. The following case study shows a few features with more predictive values. As for the observation, the PCA analysis confirmed that the same features are aligned to the main principal component.

Background

According to the National Cancer Institute in the United States, breast cancer affects one in eight women and leads to 6% of cancer-related deaths worldwide. The key to effective treatment is early detection. For those with breast cancer who are diagnosed within five years of the first cancer cell division, survival rates increase from 56% to 86%. Therefore, a precise and reliable system is essential for detecting benign or malignant breast tumors in time. Samples from this disease are typically obtained by surgery, which has the highest recognition accuracy among the methods available but is aggressive, time-consuming, and costly. By using machine learning techniques, Fine Needle Aspiration (FNA) can identify this type of cancer.

Purpose of Study

The Fine Needle Aspiration test can be a simple, inexpensive, noninvasive, and accurate diagnostic tool for detecting breast cancer, but finding appropriate features of the results poses the most difficult diagnostic problem in the early stages.

The analysis will primarily focus on:

What is the best model for predicting whether a patient has benign or malignant cancer?

The predictions can be made using various machine learning models, so we'll focus on accuracy, interpretation, and efficiency, among other factors.

What are the tradeoffs of each model?

We will discuss the advantages and disadvantages of each model in more detail that make it suitable for different types of data.

Review of Literature

Introduction of a New Diagnostic Method for Breast Cancer Based on Fine Needle Aspiration (FNA) Test Data and Combining Intelligent Systems
Fluzy et al., 2012

This is the original paper that is the basis for the assigned case study. The data source is the Wisconsin DataBase Cancer (WDBC) database. From that database, the Iranian researchers pulled the 569 records for FNA tests and employed a combination of several types of algorithms:

- Genetic Algorithm (GA)
- Fuzzy C-Means (FCM)
- Artificial Neural Networks (NN)

A genetic algorithm is a class of evolutionary algorithms. Evolutionary algorithms mimic the characteristics of evolution:

- Random vs deterministic operation (random sampling)
- Population of candidate solutions vs single best solution
- Mutation – random changes in population members that may provide a better solution (or a worse one)
- Crossover – as when two parents contribute DNA to a child, crossover combines elements of existing solutions to create a brand new solution
- Selection – survival of the fittest solutions, and the less fit solutions are eliminated[1]

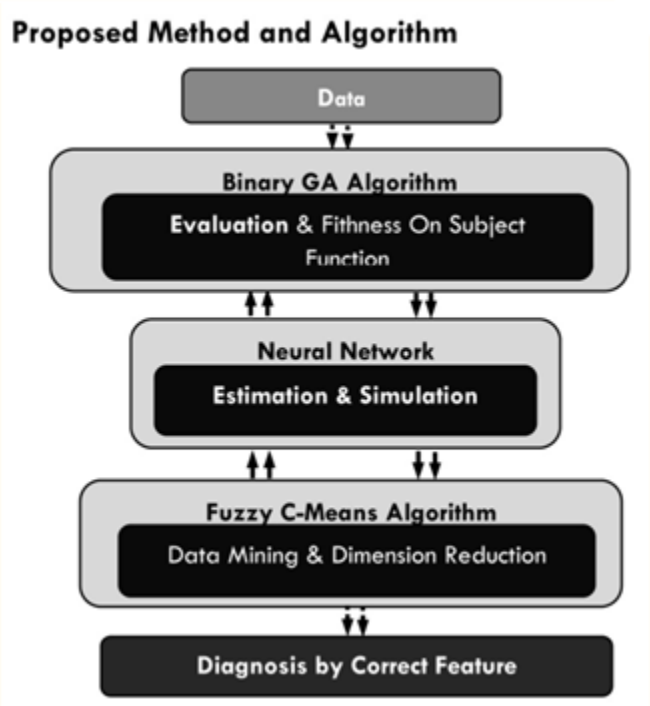
Genetic algorithms use crossover and mutation; evolutionary algorithms don't use crossover.

Fuzzy C-Means clustering is a "soft clustering" approach. Whereas K-means clustering is a "hard clustering" approach that divides observations into separate clusters, in fuzzy C-means an observation can belong to multiple clusters with a probability of it belonging to each of the clusters.

Neural networks are so named because like neurons in the brain, each neuron can send signals to every other neuron in the adjacent data layer. They are often used in image recognition problems, like this one. They "learn" to recognize images by analyzing the results of pictures translated to numerical data.

[1] ("Genetic Algorithms and Evolutionary Algorithms - Introduction," 2011)

The researchers used the following combination of the above algorithms:



The researchers used a genetic algorithm with a neural network in its cost function for the initial classification. The results were then processed with the FCM model to reduce dimensionality and estimate the accuracy. The data were then run through from the beginning several times until maximum accuracy was achieved. The accuracy of this approach compared to others that were used on the same data was the highest, 96%+:[2]

Table 3

Compare and Result

Patient identify	Best answer	Accuracy test	Method
202	Not Available	95.61%	SVM [7]
202	---	> 95%	RBF+FCM and SVM [10,11,12]
~198	---	> 93%	DT + NN [16]
~201	---	94.4%	GA + DT [17]
~194	---	> 93.6%	Exp Sys [18]
~195	---	> 94%	GA+ AI [19]
~195	---	> 92%	FPRCA [20]
~ 205	96.579%	96 %	Proposed algorithm

[2] (Fiuzy et al., 2012)

Definitions	
SVM	Support Vector Machine
RBF	Radial Basis Function
FCM	Fuzzy Classifier
DT	Decision Tree
NN	Neural Network
GA	Genetic Algorithm
Exp Sys	Expert Systems
AI	Artificial Immune algorithm
FPRCA	Fuzzy Principal Component Algorithm

Methodology and Assumptions

To find the best model, we will be training our models using five different techniques.

1. K Nearest Neighbors (KNN)
2. Support Vector Machine (SVM)
3. Neural Network
4. Neural Network with Principal Component Analysis (PCA)
5. Random Forest with Principal Component Analysis (PCA)

KNN

KNN is a non-parametric algorithm, so there aren't assumptions to be met to implement the model. The KNN algorithm is also easy to understand and implement. It reads through the whole dataset to find K nearest neighbors to classify new data points. We want to see how K nearest neighbors classify benign (noncancerous) or malignant (cancerous) cells.

SVM

We will attempt to apply tuned parameters for the SVM model to yield the best model possible. SVM's are not only suitable for accuracy but are also effective in high-dimensional spaces. We understand that some critical factors to consider are accuracy and interpretability, so we will have to weigh both when selecting our final model.

Neural Network

By analyzing similar events, artificial neural networks learn about the future and how to make decisions according to them. Artificial neural networks can also perform more than one function simultaneously. Backpropagation allows neural networks to achieve better accuracy than several models on the same data set. We will test this on the cancer detection data.

Neural Network with PCA

In addition to the advantages of using a Neural Network to examine cancer data, Principal Component Analysis is great at reducing the number of features in the dataset. This allows better visualizations of our data. The reduction of features also allows better performance and quicker results by eliminating correlated variables that don't contribute to any decision-making.

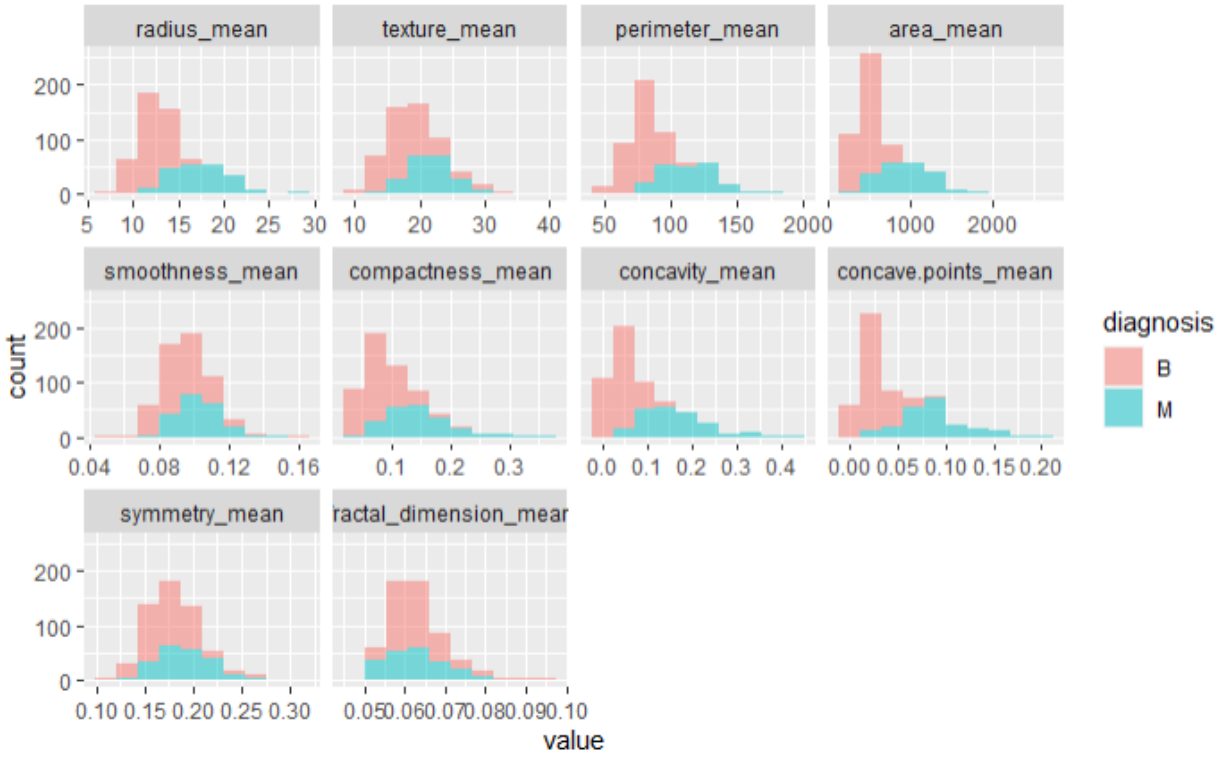
Random Forest with PCA

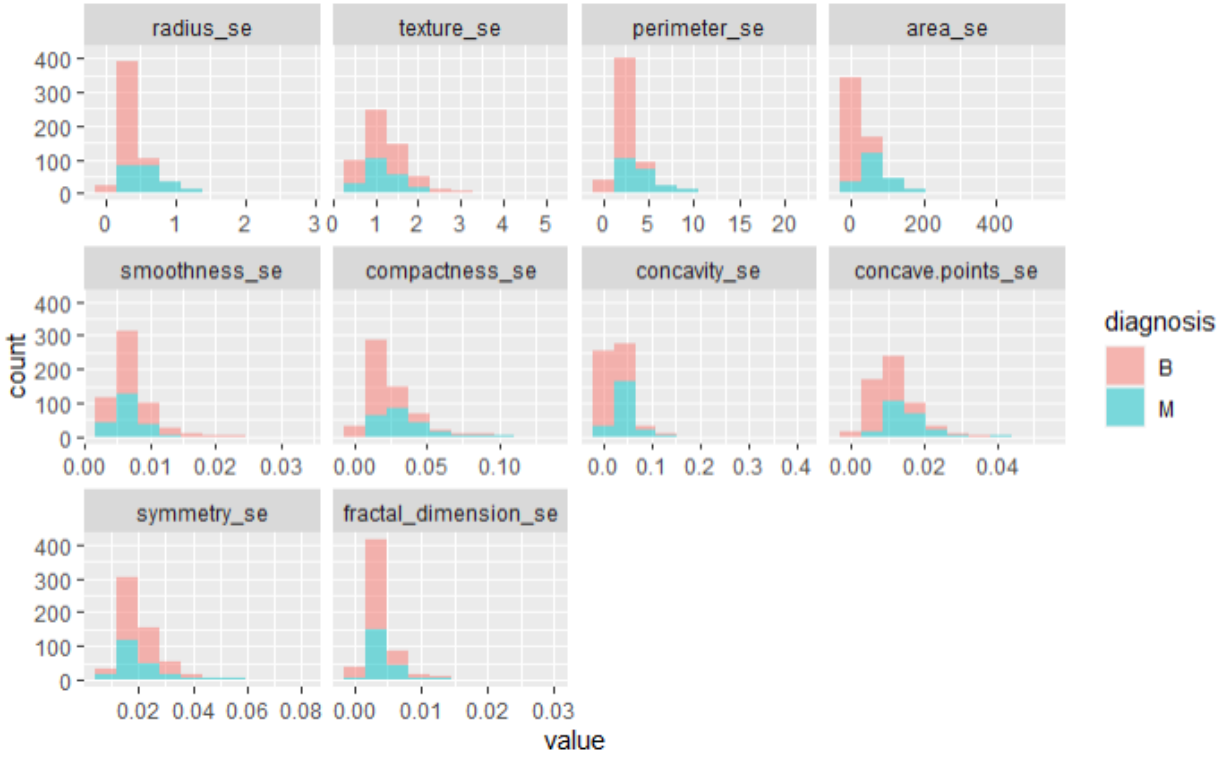
The random forest model is a culmination of many decision trees created using bootstrapping random sampling techniques to predict a response variable. The trees built in the random forest model are uncorrelated and use only a subset of predictors to choose from at each node and tree. They also work very well with data that has outliers, missing values, or is unbalanced. To help address the overfitting of random forests, we used PCA to reduce the number of features.

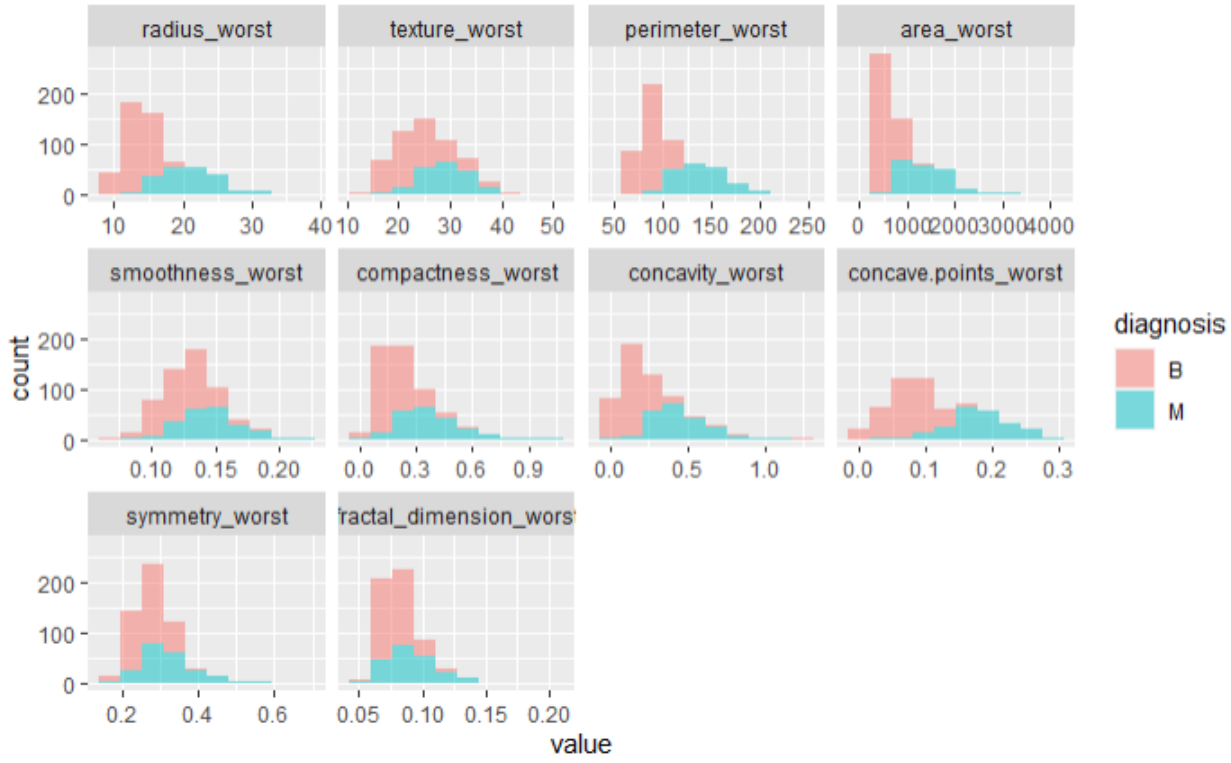
Data

The features of the data set were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass for breast cancer patients. It described the characteristics of the cell nuclei present in the image. The data was relatively clean and consisted of 569 observations, each with 32 variables. We removed the ID column as it did not contribute any predictive value. We also changed the diagnosis predictor to a factor instead of a character value as it should be a binary variable. The data did not contain any missing values, so we moved on to the dataset's attributes.

These histograms show that our data are more or less normally distributed. The mean, standard error, and "worst" variables were grouped by the data.


radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave.points_mean
symmetry_mean
ractal_dimension_mear
count
value
diagnosis
B
M


radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave.points_se
symmetry_se
fractal_dimension_se
count
value
diagnosis
B
M



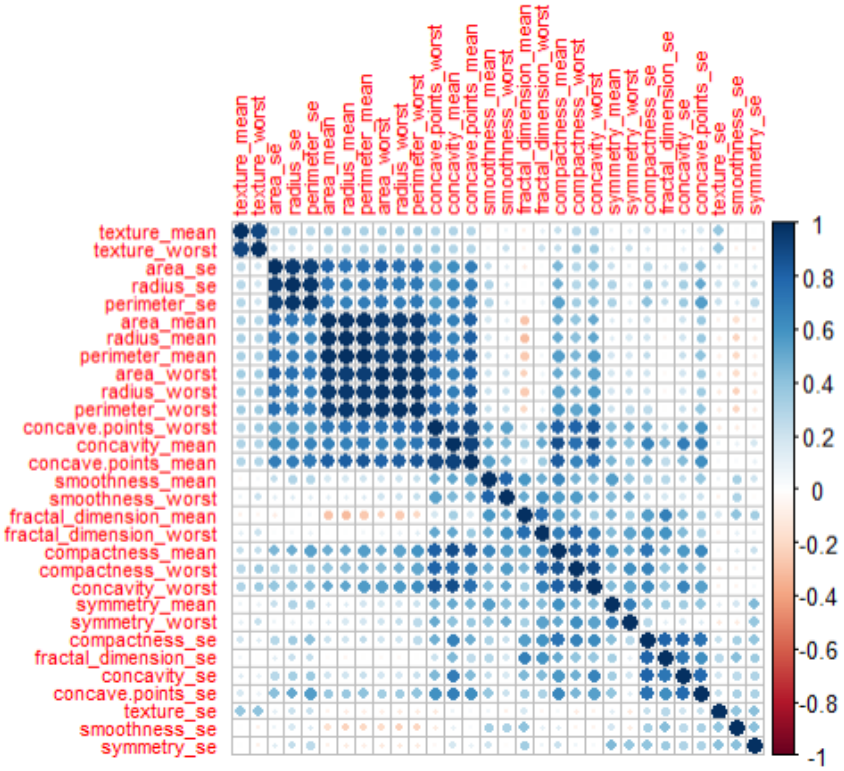
The visualization data allowed us to identify which features were predictive of malignant or benign cancers and general trends that may help choose models and hyperparameters. We analyzed the characteristics, determined which ones had higher predictive values, and formulated a model to determine whether the cells were cancerous. Regarding frequency, 357 of the observations represent 62.7% of the non-cancer cells, while 212 observations represent 37.3% of the observations that are cancer cells. This could affect the results of classification models like KNN by biasing the results of benign growth.

We assessed the correlations between the 30 predictors using a correlation plot. This allowed us to identify ten variables that were highly correlated and therefore removed from our analyses:

```
```{r}
highlyCor
```
```

```
[1] "compactness_mean"   "concavity_mean"     "texture_worst"      "fractal_dimension_se" "texture_mean"
[6] "perimeter_worst"    "diagnosis"          "texture_se"         "perimeter_se"         "radius_mean"
```

Our final dataset contained 21 variables.



We removed the following ten predictors due to high correlation:

- compactness_mean
- texture_se
- concavity_mean
- perimeter_se
- texture_worst
- radius _mean
- fractal_dimension_se
- texture_mean
- perimeter_worst
- diagnosis

Results

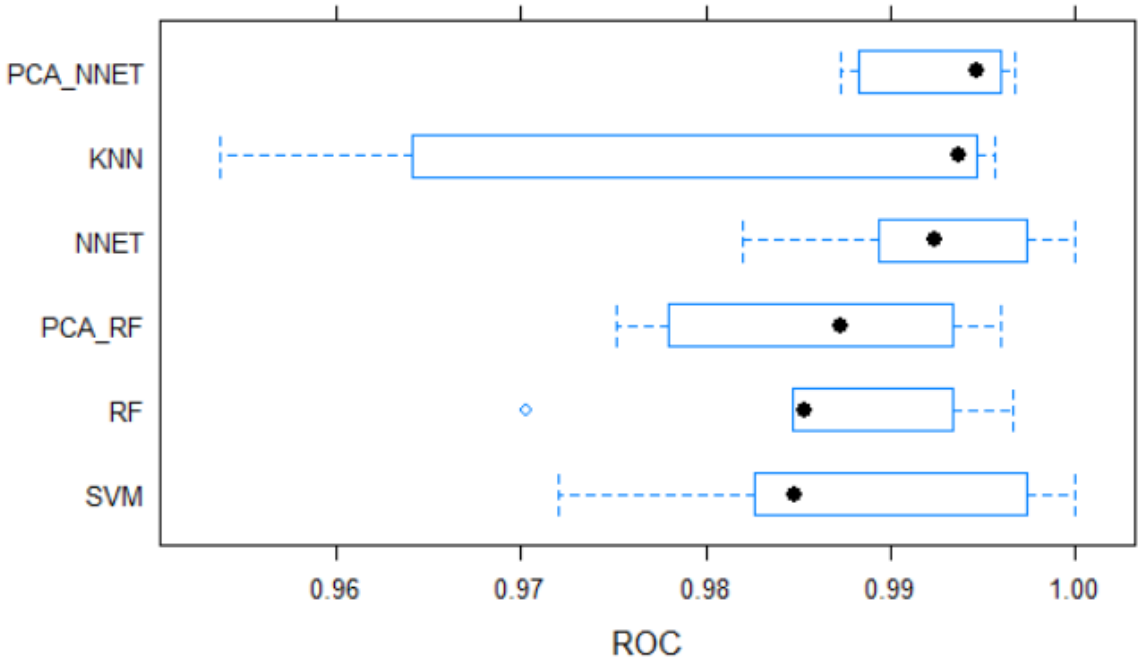
Our analysis showed that the random forest model had an accuracy rate of 94.1%, Random forest with PCA accuracy of 93.53%, KNN 92.9%, NNET 95.88%, NNET with PCA 97.65%, SVM 92.9%.

We observe that radius_worst, concave.points_mean, area_worst, area_mean, concave.points_worst, perimeter_mean, area_se, and concavity_worst are the most important features. Most of them are also in the list of features with higher dimensions in the leading Principal Components plane or aligned with the leading Principal Component, PC1, as well as aligning with the random forest results.

Tradeoffs of Models

We then tested the data on the models. After testing them, we used the ROC metric to measure the AUC of the ROC curve of each model. The following metric is independent of any threshold. We see

here that some models have great variability. The model PCA NNET achieved a significant AUC with some variability, meaning that it is the best result for detecting breast cancer. The feature analysis showed features with more predictive values for diagnosing. The PCA analysis confirmed the observations showing that the same features were aligned. The boxplots of the ROC statistic for resampling the models we used show the accuracy of PCA NNET consistently high, with the highest mean ROC of all the models.



Full metrics for all of the models show that PCA NNET has the second-highest sensitivity (true positive rate) and the second-highest specificity (true negative rate). For cancer diagnosis, these measures are especially vital. Missing a positive diagnosis can have devastating consequences. Telling a patient she has cancer when this is not true can cause unnecessary worry and perhaps lead to unwarranted treatment. Another metric of note is balanced accuracy (the average of sensitivity and specificity), which is helpful in evaluating unbalanced classes like the cancer data, which is about ⅔ benign and ⅓ cancerous.

Metric	RF	PCA_RF	PCA_NNET	KNN	SVM
Sensitivity	95.2%	77.8%	90.5%	85.7%	90.5%
Specificity	98.1%	100.0%	99.1%	99.1%	98.1%
Pos Pred Value	96.8%	100.0%	98.3%	98.2%	96.6%
Neg Pred Value	97.2%	88.4%	94.6%	92.2%	94.6%
Precision	96.8%	100.0%	98.3%	98.2%	96.6%
Recall	95.2%	77.8%	90.5%	85.7%	90.5%
F1	96.0%	87.5%	94.2%	91.5%	93.4%
Prevalence	37.1%	37.1%	37.1%	37.1%	37.1%
Detection Rate	35.3%	28.8%	33.5%	31.8%	33.5%
Detection Prevalence	36.5%	28.8%	34.1%	32.4%	34.7%
Balanced Accuracy	96.7%	88.9%	94.8%	92.4%	94.3%

Conclusion and Recommendations

In conclusion, the feature analysis shows a few more predictive values for the diagnosis. The PCA analysis confirmed the observations, indicating that the same features align with the leading in the principal component. We have found a model based on neural network and PCA preprocessed data with good results over the test set. This model has a sensitivity of 0.905 with a F1 score of 0.942.

Bibliography

Fiuzy, M., Haddadnia, J., Mollania, N., Hashemian, M., Hassanpour, K., 2012. Introduction of a New Diagnostic Method for Breast Cancer Based on Fine Needle Aspiration (FNA) Test Data and Combining Intelligent Systems. Iran J Cancer Prev 5, 169–177.

Genetic Algorithms and Evolutionary Algorithms - Introduction [WWW Document], 2011. . solver. URL https://www.solver.com/genetic-evolutionary-introduction (accessed 11.24.21).

Appendix

The data frame below shows 12 variables and 32 features. 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 containing cell nucleus information. The additional two features are the patient id number and diagnosis for a total of 32 features.

patient ID number	Patient identification number
diagnosis	M = malignant (cancerous) or B = benign (noncancerous)
	Ten real-valued features are computed for each cell nucleus:
radius	Mean of distances from center to points on the perimeter
texture	Standard deviation of gray-scale values
perimeter	Perimeter of nucleus
area	Area of nucleus
smoothness	Local variation in radius lengths
compactness	perimeter^2 / area - 1.0
concavity	Severity of concave portions of the contour
concave point	Number of concave portions of the contour
symmetry	Nuclei symmetry
fractal dimension	Coastline approximation - 1