Prediction of breast cancer from nuclear features of fine needle aspirate biopsies

Amy Gill

6/2/2019

Contents

Introduction

Background

The purpose of breast tumor biopsy is to differentiate breast cancer from benign masses. Most breast tumors are not cancerous, but it is critical to correctly identify malignant tumors. Because advanced breast cancer is life-threatening, a false negative test result that misses a diagnosis could be a fatal error. It is also important to reduce false positives and correctly identify patients who have benign masses, both to reduce unwarranted psychological and financial distress associated with a cancer diagnosis and to avoid morbidity from unnecessary cancer treatment.

Many breast tumors are assessed by fine needle aspiration (FNA) cytology, the most cost-effective and least invasive biopsy method (Mitra and Dey, 2016). FNA consists of using a thin needle to extract a piece of the tumor, which a pathologist analyzes for physical features that distinguish malignant cells. However, FNA biopsy accuracy changes dramatically with experience of the cytopathologist (Feoli et al., 2008). The false negative rate is variable across studies, with many recent studies showing a 5-10% false negative rate and some centers reporting rates of 15% or greater (Mitra and Dey, 2016). There has been a movement away from use of FNA towards the more invasive and expensive core needle biopsy method, which is widely reputed to be better, although studies show that FNA and core needle biopsy have similar false negative rates of 1.7% when performed by an experienced cytopathologist (Brancato et al., 2012).

Digital analysis of breast biopsies could potentially reduce the disparity in FNA false negative rates. Cytology relies on visual pattern matching of cellular features by the pathologist, and these pattern matching strategies can be mimicked through machine learning. A classifier could supplement the opinion of a cytopathologist, potentially compensating for lack of pathologist experience and providing an accessible and objective second opinion about ambiguous samples. The model could also serve as a teaching tool and highlight which features are most useful for visually distinguishing cancer and normal tissue.

The goal of this study was to train a machine learning classifier to predict whether a breast tumor is malignant (cancer) or benign (not cancer) based on cell nucleus features extracted from digitized images of FNA cytology slides. A useful model should classify tumors better than the worst performing real pathologists (15% false negative rate) in order to be potentially beneficial as a training tool or supplement for this population.

The dataset

The Breast Cancer Wisconsin Diagnostic Dataset consists of 30 features computationally extracted from digital images of FNA biopsy slides of a consecutive series of 569 breast tumors (Street et al., 1993). Features describe properties of the cell nuclei, including attributes related to nucleus size, shape and regularity. For each feature, the average value of the feature across all cells in the image, standard error of the feature across all nuclei (se), and most extreme (worst) value of the feature in the image are reported. Here is a description of the 10 base variables, each of which has a mean, standard error, and worst value:

- `radius`: Nucleus radius (mean of distances from center to points on the perimeter).
- `texture`: Nucleus texture (standard deviation of gray-scale values).
- `perimeter`: Nucleus perimeter.
- `area`: Nucleus area.
- `smoothness`: Nucleus smoothness (local variation in radius lengths).
- `compactness`: Nucleus compactness (perimeter^2/area - 1).
- `concavity`: Nucleus concavity (severity of concave portions of the contour).
- `concave_pts`: Number of concave portions of the nucleus contour.
- `symmetry`: Nucleus symmetry.
- `fractal_dim`: Nucleus fractal dimension ("coastline approximation" - 1).

Approach

The dataset was imported from UCI Machine Learning Repository. Individual features were centered and scaled. Exploratory data analysis revealed clustering of samples based on tumor type and showed that tumor type is the primary source of feature variation between samples. The data were split into an 80% training set and 20% test set, then used to train 7 different machine learning models, which were combined into an ensemble. Model performance was evaluated on the test set using accuracy as a primary measure and false negative rate as a secondary measure.

Performance

The best two models, k-nearest neighbors and linear discriminant analysis, each yielded a 97.4% accuracy and 7% false negative rate. The 7-model ensemble had a slightly lower accuracy and false negative rate (96.5% and 9.3% respectively). Importantly, this project replicates a prior analysis on this dataset that used a multisurface method tree to achieve an accuracy of 97.5% (Wolberg et al., 1995), so the accuracy level is consistent with the literature. The false negative rate is consistent with that of professional cytopathologists and may represent an improvement over some least experienced pathologists.

Methods

Data import and wrangling

The dataset was imported from the UCI Machine Learning Repository as a matrix (samples by features) and an associated vector of tumor type for each sample (malignant or benign). The matrix was scaled by column so that each feature value was reported as a z-score.

Exploratory data analysis

A heatmap of samples was generated by calculating and plotting the Euclidean distance between samples. A heatmap of features was generated by calculating and plotting the correlation coefficient between features. P-values were calculated using two-tailed independent Student's t-tests. Confidence intervals were calculated

using the normal approximation, which was motivated by the apparently normal density of most features when split into benign and malignant groups. Principal component analysis was performed and principal components were investigated for association with tumor type.

Modeling

The data were split into an 80% training set and 20% test set, which had roughly equal proportions of benign and malignant samples. Seven different models were fit to the training data: k-means, logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), loess, k-nearest neighbors (KNN) and random forest. K-means clustering was performed with 2 centers. Logistic regression, LDA, QDA and loess models were trained with default parameters. The KNN model tested odd values of number of neighbors to consider ranging from 3 to 21, and the best value of $k = 9$ was selected using accuracy across 25 bootstrapped samples. The random forest model tested odd values of number of predictors available at each split ranging from 3 to 21 and the best value of `mtry = 7` was selected using accuracy across 25 bootstrapped samples. An ensemble was built using all 7 models that classified samples by majority vote.

The performance of individual algorithms and the ensemble was evaluated using accuracy as the primary outcome and false negative rate as the secondary outcome. Sensitivity and specificity were also reported as tertiary outcomes. Malignant was treated as the positive class.

Results

Exploratory analysis

Distance between benign and malignant samples

A heatmap of distance between samples shows two main clusters of samples, which correspond to benign and malignant types (Figure 1). Benign samples are most similar to other benign samples, and malignant samples are most similar to other malignant samples. Furthermore, malignant samples are less similar to each other than benign samples, as indicated by a wider variance in distance between malignant samples.

Correlation between features

The heatmap of correlation between features demonstrates that many feature combinations encode distinct information, as shown by the fact that there are several independent clusters of features along the primary diagonal that are relatively uncorrelated with each other (Figure 2). There are also several features that are highly collinear, shown by dark blue squares along the main diagonal. Nuclear size features such as radius, perimeter and area encode redundant information, as indicated by the dark blue box in the upper right. A second cluster in the center relates to nucleus shape, especially concavity and compactness. Mean and worst values of the same variable tend to cluster, and various standard errors tend to cluster, suggesting variance between nuclear features is potentially informative in and of itself.

Distributions of individual features

Most of the 30 nuclear features show a significant difference in distribution between benign and malignant samples (Figure 3). In all cases where the distributions for a feature clearly differ between tumor type, the malignant samples have larger values. Features associated with nucleus size (mean and worst area, perimeter and radius) and irregular nuclear shape (mean and worst concavity and number of concave points) have similar distributions with benign samples tightly clustered at low values and malignant samples right shifted with larger standard errors. This suggests that large nucleus size and more irregular nucleus shape are associated with cancer, which is consistent with cancer being a dysregulation of normal cell architecture and behavior. In general, mean and worst values of features differ more clearly across tumor type than standard errors of features.

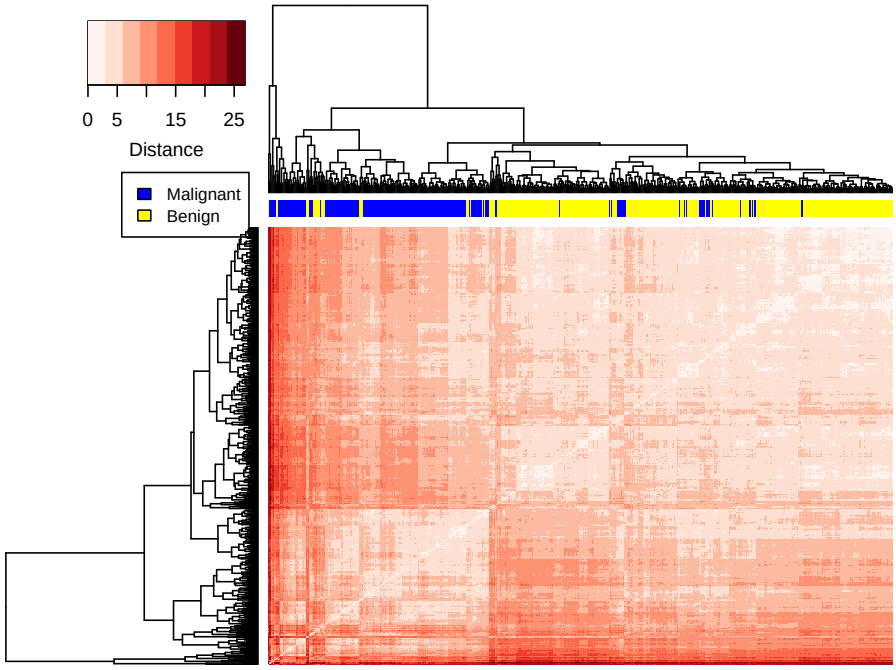

Figure 1: Heatmap of distance between samples.

Feature p-values and confidence intervals

All but 4 of the 30 nuclear features are significantly different between benign and malignant tumors at a p-value of 0.05 using a two-tailed independent Student's t-test (Table 1). The 12 most significantly different variables are all mean and worst features of concave points, concavity, perimeter, area, radius and compactness. The most significantly different attribute was `concave_pts_worst` with a p-value of 1.06×10^{-96}, shortly followed by other features related to nuclear size and irregular shape. Mean and worst values are more significantly different between cancer and normal tissue than standard errors. Importantly, although the means are significantly different for most variables, the 95% confidence intervals all overlap: even the features most highly associated with cancer are not able to unambiguously classify tumor type individually. This suggests models that combine large numbers of features will perform better than models with single features.

Principal component analysis

Principal component analysis shows that the first principal component (PC) accounts for 44.3% of the variance, the second PC accounts for 19.0% of the variance, and every subsequent PC accounts for less than 10% of the variance (Figure 4, Table 2). Seven PCs explain over 90% of the variance and 10 PCs explain over 95% of the variance. Importantly, a boxplot of principal components by tumor type shows that only the first principal component is clearly associated with benign or malignant status. Malignant tumors have significantly higher levels of PC1 than benign tumors, with no overlap between the IQRs for the two groups (Figure 5). A plot of the data using the first 2 PCs shows that PC1 alone is nearly sufficient to separate the samples into benign or malignant classes, though there is some overlap around PC1 values of 0 (Figure 6). Unfortunately, PC1 is not very interpretable: the variable contributing the highest proportion of variance to PC1 (`concave_pts_mean`) only contributes 6.8% of the PC, and 26 of the 30 variables contribute 1% or more to the variance of PC1. There is a rough correspondence between the p-value of a feature's association

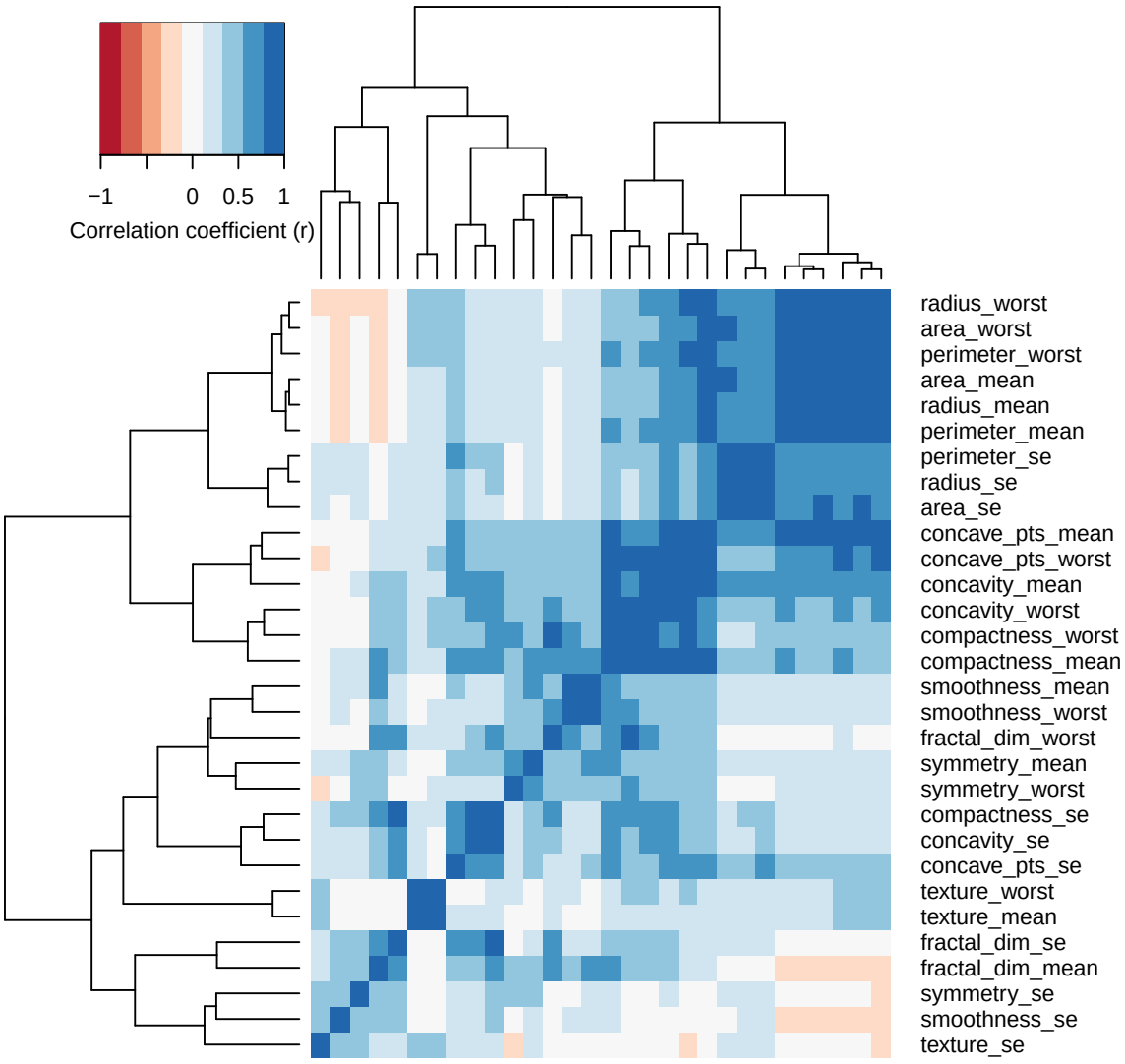

Figure 2: Heatmap of correlation between features.

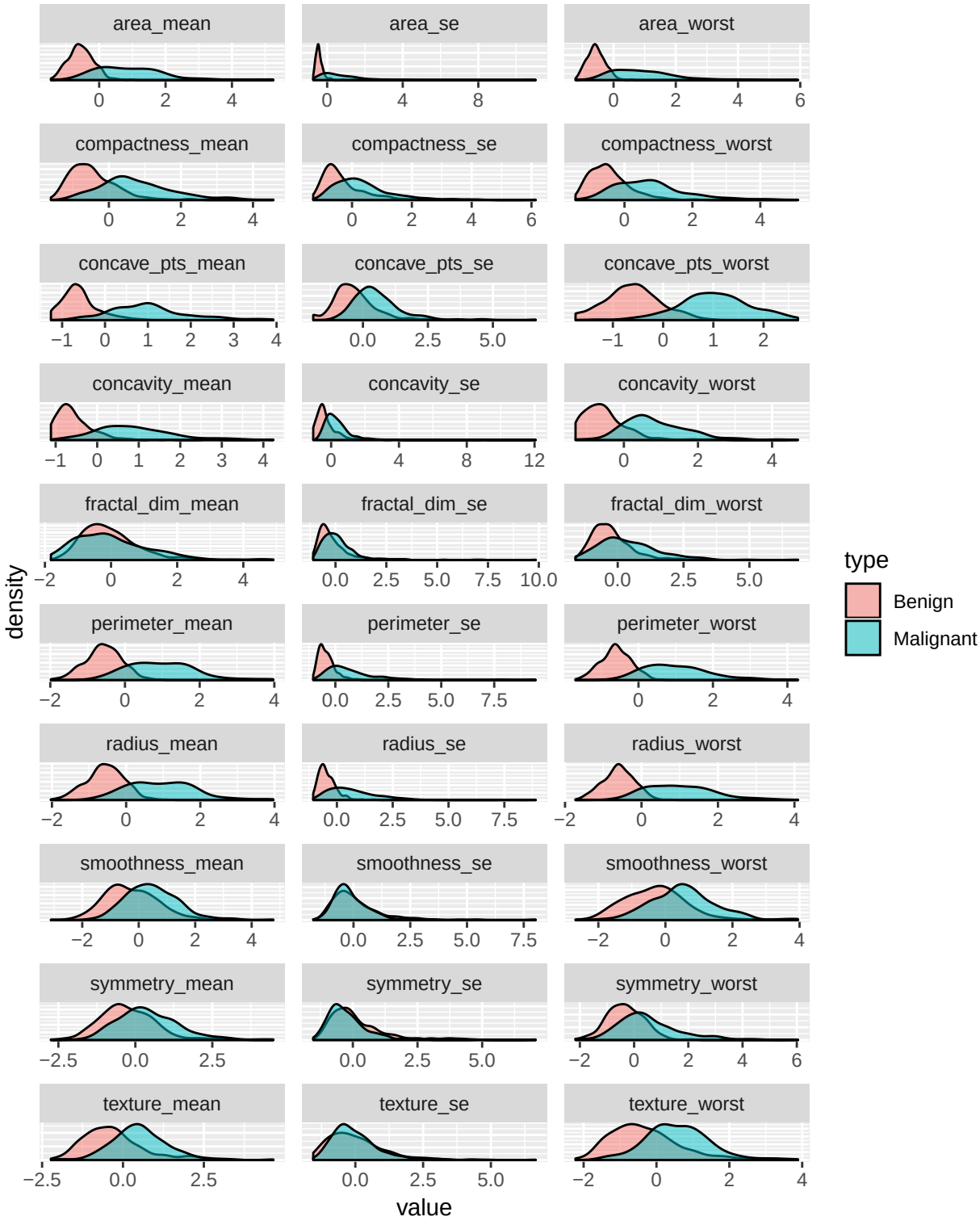

Figure 3: Density plots of nuclear features across tumor type.

Table 1: Feature p-values and confidence intervals by tumor type.

Variable	Benign mean	Benign 95% CI	Malignant mean	Malignant 95% CI	P-value	Significance
concave_pts_worst	-0.611	-1.678 - 0.456	1.029	-0.352 - 2.41	1.06e-96	***
perimeter_worst	-0.603	-1.392 - 0.186	1.015	-0.703 - 2.733	1.03e-72	***
concave_pts_mean	-0.598	-1.402 - 0.206	1.007	-0.729 - 2.743	3.13e-71	***
radius_worst	-0.598	-1.401 - 0.206	1.007	-0.73 - 2.744	3.56e-71	***
perimeter_mean	-0.572	-1.524 - 0.381	0.963	-0.8 - 2.726	1.02e-66	***
radius_mean	-0.562	-1.552 - 0.428	0.947	-0.835 - 2.728	1.68e-64	***
concavity_worst	-0.508	-1.827 - 0.811	0.855	-0.85 - 2.56	9.85e-59	***
concavity_mean	-0.536	-1.604 - 0.532	0.903	-0.942 - 2.747	3.74e-58	***
area_worst	-0.565	-1.128 - -0.002	0.951	-1.107 - 3.01	4.94e-54	***
area_mean	-0.546	-1.294 - 0.202	0.919	-1.13 - 2.968	3.28e-52	***
compactness_mean	-0.459	-1.712 - 0.793	0.773	-1.23 - 2.777	9.61e-42	***
compactness_worst	-0.455	-1.603 - 0.693	0.766	-1.356 - 2.889	1.75e-38	***
radius_se	-0.437	-1.232 - 0.359	0.735	-1.703 - 3.174	1.49e-30	***
texture_worst	-0.352	-2.104 - 1.4	0.592	-1.141 - 2.325	5.2e-30	***
perimeter_se	-0.428	-1.176 - 0.319	0.721	-1.769 - 3.211	6.87e-29	***
area_se	-0.422	-0.803 - -0.041	0.711	-1.933 - 3.354	2.98e-26	***
texture_mean	-0.320	-2.14 - 1.501	0.538	-1.184 - 2.261	3.02e-25	***
smoothness_worst	-0.324	-2.042 - 1.393	0.546	-1.331 - 2.424	3.47e-24	***
concave_pts_se	-0.314	-2.127 - 1.499	0.529	-1.224 - 2.282	4.04e-24	***
smoothness_mean	-0.276	-2.15 - 1.598	0.465	-1.292 - 2.222	5.57e-19	***
symmetry_worst	-0.321	-1.643 - 1.002	0.540	-1.826 - 2.906	6.56e-19	***
symmetry_mean	-0.254	-2.028 - 1.519	0.429	-1.547 - 2.404	5.96e-15	***
fractal_dim_worst	-0.249	-1.747 - 1.249	0.420	-1.919 - 2.759	2.04e-12	***
compactness_se	-0.226	-2.015 - 1.564	0.380	-1.633 - 2.392	6.34e-12	***
concavity_se	-0.195	-2.333 - 1.942	0.329	-1.074 - 1.732	1.27e-11	***
fractal_dim_se	-0.060	-2.236 - 2.116	0.101	-1.411 - 1.613	0.0422	*
smoothness_se	0.052	-1.946 - 2.049	-0.087	-1.974 - 1.8	0.105	
fractal_dim_mean	0.010	-1.863 - 1.883	-0.017	-2.119 - 2.086	0.767	
texture_se	0.006	-2.087 - 2.1	-0.011	-1.727 - 1.706	0.835	
symmetry_se	0.005	-1.654 - 1.664	-0.008	-2.395 - 2.378	0.887	

with tumor type and its contribution to PC1, with highly significant features contributing more to PC1 and unassociated features contributing less than 1% each to PC1 (Table 3).

Conclusions from exploratory analysis

There are significant differences between benign and malignant samples, suggesting it should be possible to find an actionable prediction algorithm. The principal component analysis demonstrates that the most significant source of variance in sample features is whether they are malignant or benign. Furthermore, the first PC informs whether a sample is cancerous, while other PCs differ little between benign and malignant tumors. Features most strongly associated with the first PC tend to be more significantly different between malignant and benign samples. The features most significantly associated with cancer status relate to large nucleus size (radius, perimeter, area) and irregular nucleus shape (concavity, number of concave points, compactness), with mean and worst values of these parameters more strongly associated with tumor type than the standard error of these parameters. Importantly, no single variable is sufficient to distinguish malignant and benign samples, suggesting that the best performing models will incorporate numerous complementary variables.

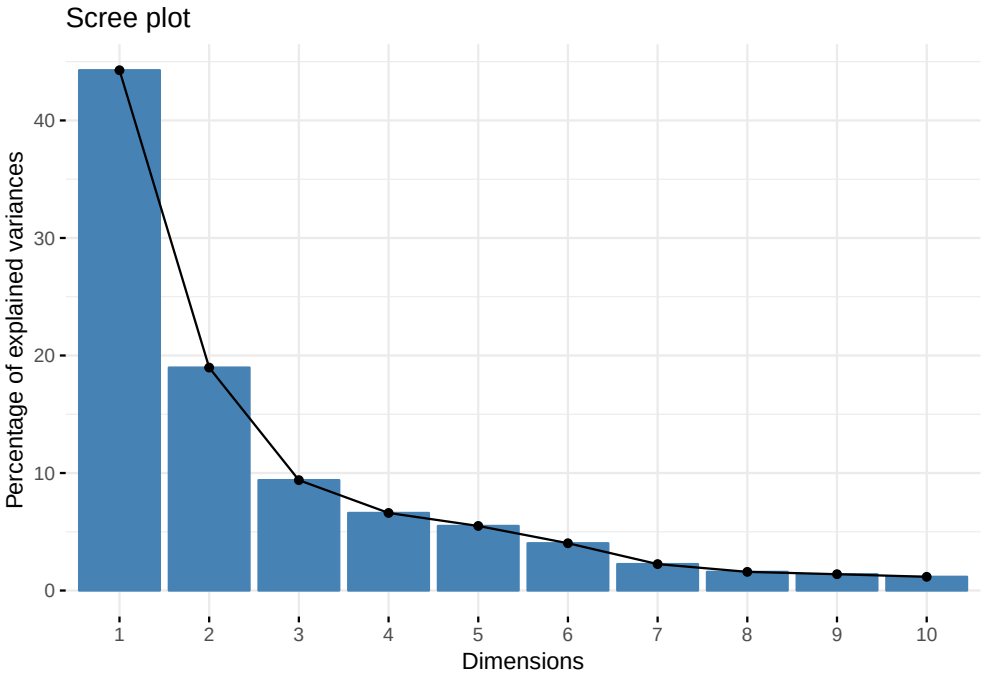

Figure 4: Scree plot of variance explained by principal components.

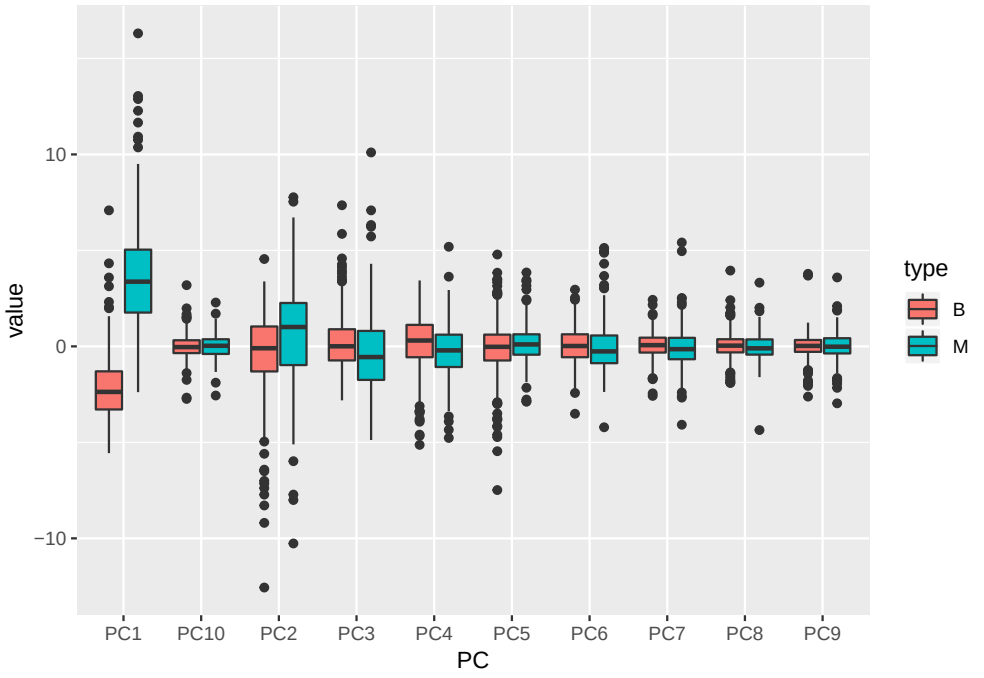

Figure 5: Principal components by tumor type. The first PC encodes most of the difference between benign and malignant samples.

Table 2: Variance explained by the first 10 principal components.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	3.644	2.386	1.679	1.407	1.284	1.099	0.822	0.690	0.646	0.592
Proportion of Variance	0.443	0.190	0.094	0.066	0.055	0.040	0.023	0.016	0.014	0.012
Cumulative Proportion	0.443	0.632	0.726	0.792	0.847	0.888	0.910	0.926	0.940	0.952

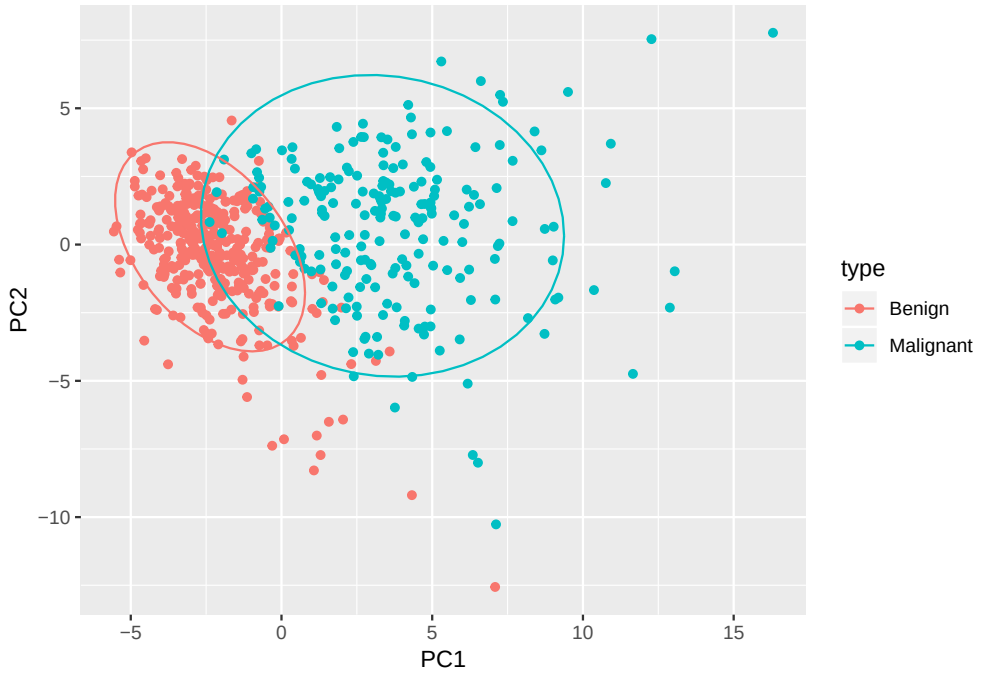

Figure 6: The first principal component separates samples into benign and malignant clusters.

Modeling

Testing and training sets

The data were split into an 80% training set and 20% test set. The training and test sets had equal proportions of cancer and normal tissue, with 37.2% and 37.4% malignant samples respectively (Table 4).

Individual models

Seven models were trained and their performance was assessed on the test set (Table 5). The best performing models are LDA and KNN with an accuracy of 0.974. These models have identical sensitivities for malignant samples of 0.930 and both have a perfect specificity of 1, correctly identifying all benign samples. These two models, along with logistic regression and loess, are tied for the lowest false negative rate of 7%. The best-performing models, LDA and KNN, are not easily interpretable. To get a sense of which variables drive classification, variable importance was extracted from the random forest model, which had an accuracy of 0.948 and was the best performing model for which importance can be calculated (Figure 7). Variable importance was measured in mean decrease in Gini index. 5 variables account for most of the importance: the worst number of concave points in the nucleus outline, worst perimeter, worst area, worst radius and mean number of concave points in the nucleus outline.

Ensemble models

Table 3: Feature contribution to the variance of PC1. PC1 is not well determined by a small number of variables.

	Percent of PC1
concave_pts_mean	6.804
concavity_mean	6.677
concave_pts_worst	6.294
compactness_mean	5.726
perimeter_worst	5.600
concavity_worst	5.233
radius_worst	5.198
perimeter_mean	5.177
area_worst	5.057
area_mean	4.884
radius_mean	4.792
perimeter_se	4.466
compactness_worst	4.414
radius_se	4.243
area_se	4.116
concave_pts_se	3.364
compactness_se	2.903
concavity_se	2.359
smoothness_mean	2.033
symmetry_mean	1.909
fractal_dim_worst	1.737
smoothness_worst	1.637
symmetry_worst	1.511
texture_worst	1.091
texture_mean	1.076
fractal_dim_se	1.052
fractal_dim_mean	0.414
symmetry_se	0.181
texture_se	0.030
smoothness_se	0.021

A heatmap of model predictions shows a core set of benign and malignant tumors that are correctly classified by all 7 models, a small number of benign samples that are incorrectly classified by loess and logistic regression, 2 small clusters of malignant samples that are only sometimes correctly classified, and a small cluster of malignant samples that is not identified by any algorithm (Figure 8). The two top algorithms, KNN and LDA, have identical accuracies of 0.974 and perfect classification of benign samples (specificity = 1), but they identify different subsets of malignant samples.

The seven models have mostly consistent results on benign samples and are more variable on malignant samples (Table 6). Of the benign samples in the test set, all models correctly classified 88.9% (64/72) and no samples were incorrectly classified by more than half of the models. In contrast, only 76.7% (33/43) of malignant samples in the test set were correctly classified by all models, and 9.3% (4/43) of cancers were incorrectly classified by over half of the models, including 4.7% (2/43) that no model called correctly.

In an attempt to improve performance beyond that of individual models, an ensemble model was built that classified samples by majority vote from all 7 models (Table 7). This ensemble performed slightly worse than the two top individual models, KNN and LDA (accuracy 0.965 versus 0.974, false negative rate 0.093 versus 0.07), although it correctly identified all benign samples (specificity = 1).

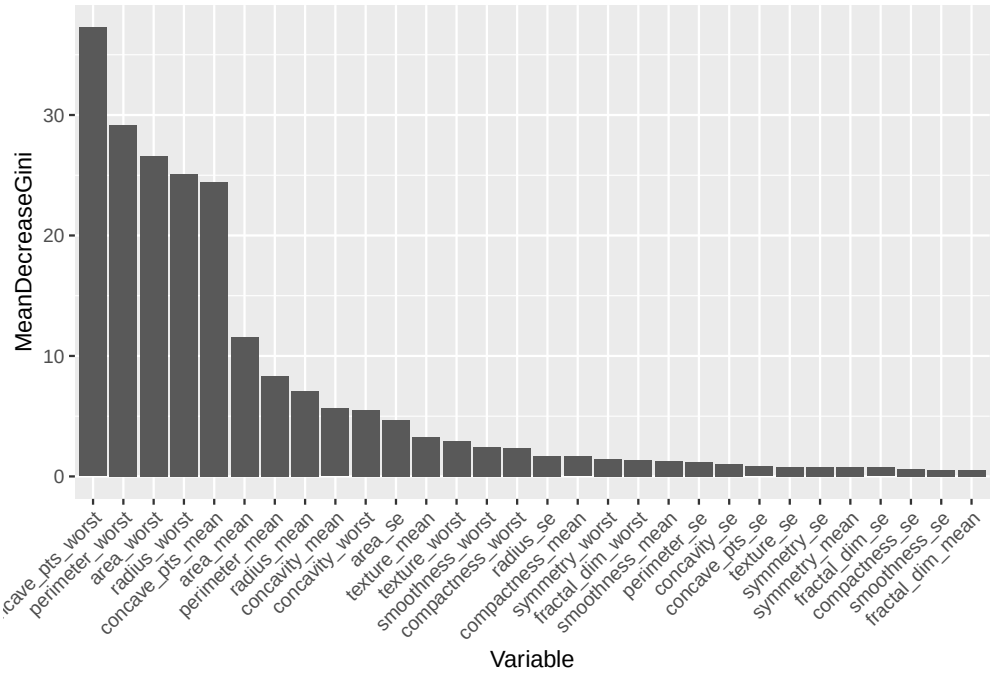

Figure 7: Importance of random forest features.

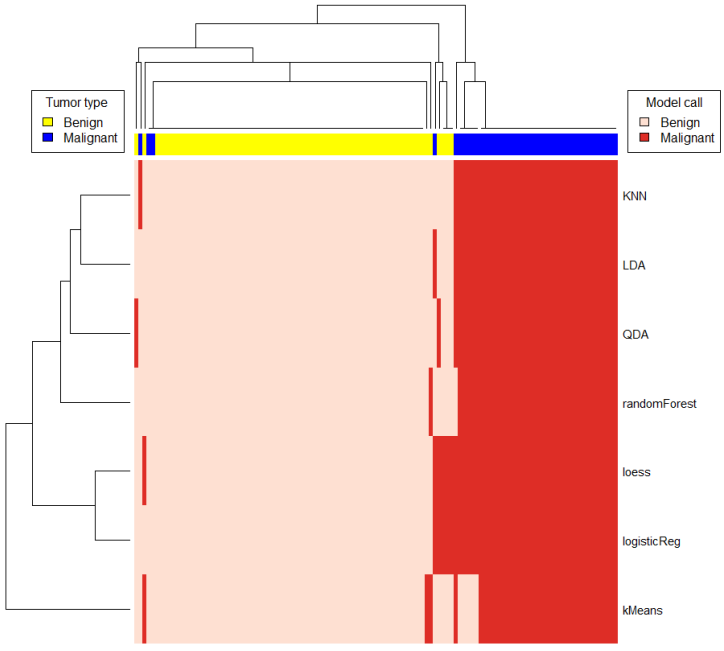

Figure 8: Predictions of tumor type across seven models.

Table 4: Proportions of benign and malignant tumors in training and test sets.

Dataset	Benign	Malignant
Train	0.628	0.372
Test	0.626	0.374

Table 5: Performance metrics for machine learning classifiers.

Model	Accuracy	Sensitivity	Specificity	False Negative Rate
K means	0.896	0.791	0.958	0.209
Logistic regression	0.939	0.930	0.944	0.070
LDA	0.974	0.930	1.000	0.070
QDA	0.948	0.907	0.972	0.093
Loess	0.930	0.930	0.931	0.070
K nearest neighbors	0.974	0.930	1.000	0.070
Random forest	0.948	0.884	0.986	0.116

Discussion

The purpose of this study was to develop a machine learning model to classify breast tumor biopsy samples as malignant or benign using features of cell nuclei extracted fine needle aspirate cytology images. The model accuracy needed to be comparable to the original peer-reviewed analysis of this dataset (Wolberg et al., 1995), which trained a multisurface model tree algorithm that achieved an accuracy of 97.5%. The two best individual models, KNN and LDA, both achieved accuracies of 97.4% on the test set. The equivalent accuracies suggest the current analysis was performed correctly and that these models reproducibly identify features that can classify almost all breast biopsies as benign or malignant.

Furthermore, in order to be potentially useful in the clinic, this model needed to perform better than the worst reported false negative rate of 15%. The KNN and LDA models both had false negative rates of 7%, better than the performance of the least experienced cytopathologists. This model could provide decision-making support for pathologists diagnosing breast cancer, especially those with less experience. However, it is important to note that FNA biopsy error rates depend both the pathologist's diagnostic skill and the pathologist's competence performing the FNA biopsy procedure (Willems et al., 2012). The models designed here can only compensate for deficiencies in image analysis, not aspirator performance.

For most samples, all seven models agree on whether the tumor is benign or malignant. These algorithms or ones like them could be used in the clinic to flag cytology samples as clearly benign, clearly malignant or ambiguous. This could identify samples that require more careful follow-up or additional analysis, and increase confidence in diagnoses when the pathologist and algorithm agree. When the pathologist thinks a sample is benign but most models suggest the tumor is malignant, then further follow-up could be save a life.

A majority vote ensemble of all 7 models had an accuracy of 96.5% and false negative rate of 9.3%, which is worse than the individual KNN and LDA models. This is because most models fail to recognize the same set of malignant samples. Creating an ensemble of models that make the same errors will not improve the performance of the ensemble. However, individual models are more vulnerable to overfitting than an ensemble. Importantly, the KNN and LDA models had higher accuracy on the test set (97.4%) than on the training set (95.3% for KNN and 94.9% for LDA), suggesting that the accuracy metrics on the test set may be overly optimistic and that the ensemble model may be more appropriate despite its slight decrease in accuracy. A future analysis could include cross-validation of these results on several training and test set partitions to determine whether these models are robust to sampling error.

Another future direction involves applying more machine learning algorithms to the current test, including boosted algorithms that would increase the odds of correctly classifying samples that are routinely assigned incorrectly in other algorithms. Boosting involves an iterative modeling process that progressively prioritizes

Table 6: Number of models classifying a sample as malignant by tumor type.

	Benign	Malignant
0	64	2
1	2	1
2	5	0
3	1	1
6	0	6
7	0	33

Table 7: Performance metrics for ensemble model. The individual KNN and LDA models perform slightly better.

Model	Accuracy	Sensitivity	Specificity	False Negative Rate
7-model ensemble	0.965	0.907	1	0.093

identifying samples that were incorrect in previous modeling cycles. Addition of more sophisticated models could improve the performance of the ensemble.

In addition, this analysis used only features related to the appearance of the cell nuclei. Cancers have additional morphological features aside from nuclear irregularity that could be informative. Some examples include evidence of mitosis, nucleus-cytoplasm ratio, cellular pleomorphism, properties of cell aggregates, features of nucleoli, lymphocytic infiltration, uniformity of cell shape and size, presence of bare nuclei and cytoplasmic staining features (Garud et al., 2012). Malignant tumors that were incorrectly classified as benign could have normal appearing nuclei but have other distinguishing features. Inclusion of additional digitally extracted features from this list in a diagnostic algorithm could further improve accuracy and reduce false negatives. Pairing of computer vision for extracting such features and machine learning classification could further improve the clinical utility of FNA breast biopsies, reducing the incidence of incorrect diagnosis of breast disease.

References

B. Brancato, E. Crocetti, S, Bianchi, S. Catarzi, G.G. Risso, P. Bulgaresi, F. Piscioli, M. Scialpi, S. Ciatto, and N. Houssami. Accuracy of needle biopsy of breast lesions visible on ultrasound: audit of fine needle versus core needle biopsy in 3233 consecutive samplings with ascertained outcomes. Breast, 21(4):449-454.

F. Feoli, M. Paesmans, and P. Van Eeckhout. Fine Needle Aspiration Cytology of the Breast: Impact of Experience on Accuracy, Using Standardized Cytologic Criteria. Acta Cytologica, 52:145-151, 2008.

H.T. Garud, D. Sheet, M. Mahadevappa, J. Chatterjee, A.K. Ray and A. Ghosh. Breast fine needle aspiration cytology practices and commonly perceived diagnostic significance of cytological features: A pan-India survey. Journal of Cytology, 29(3):183-189, 2012.

S. Mitra and P. Dey. Fine-needle aspiration and core biopsy in the diagnosis of breast lesions: A comparison and review of the literature. CytoJournal, 13:18, 2016.

I.A. Park and E.K. Ham. Fine needle aspiration cytology of palpable breast lesions: histologic subtype in false negative cases. Acta Cytologica, 41(4):1131–1138, 1997.

W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, 1905:861-870, San Jose, CA, 1993.

S.M. Willems, C.H.M. van Deurzen, P.J. van Diest. Diagnosis of breast lesions: fine-needle aspiration cytology or core needle biopsy? Journal of Clinical Pathology, 65(4):287-292, 2012.

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, 26:792-796, 1995.