Current Topics in Data and AI Group Assignment

Breast Cancer Diagnostic Analysis

Group Members: Paulina Thrasher, Die Tao, Sabina Suleimanova

22/04/2025

Introduction

Breast cancer is a significant health concern for women around the world and is a common cause of female mortality. Early detection and treatment can greatly improve successful outcomes and machine learning advancements have made it possible to develop more accurate models to diagnose patients as soon as possible (Khalid et al., 2023). In this analysis, we aim to build a predictive model to classify breast cancer tumors as benign or malignant, using a dataset containing 30 diagnostic features (Wolberg, 1993). The dataset is publicly available from the University of California Irvine (UCI) Machine Learning Repository. It consists of measurements of cell nuclei taken from breast masses which were sampled using fine-needle aspiration, a common procedure in oncology.

This project is divided into three main stages: data understanding, data pre-processing, and classification. The data understanding phase focuses on understanding the composition of the dataset, understanding the target variables and the features, and planning what data cleaning needs to be completed. The preprocessing steps involve exploratory analysis to ensure the data is ready for modeling. This includes handling missing values, dropping irrelevant columns, transforming categorical variables, normalizing, and assessing correlations.

The dataset is labeled with the diagnosis of malignant or benign, and thus this report employs supervised classification methods. A Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel is used to predict the diagnosis based on the included features. This method is used for classification and works by finding the optimal hyperplane in n-dimensional space to separate datapoints into different classes. It does so by maximizing the margin between the closest points of the two classes, known as support vectors (GeeksForGeeks, n.d.). The RBF kernel allows the model to handle nonlinear relationships by mapping the original features into high-dimensional space, where a linear separation is then possible. This balances the trade-off between misclassification with model complexity (Pedregosa et al., n.d.).

This analysis will also demonstrate the value of visualizations in understanding the results, specifically utilizing Probability of the Alternative Class (PAC) measures, Silhouette Plots, Quasi Residual Plots, Class Maps, and Stacked Mosaic Plots to assess the classification results. Each tool is designed to highlight different aspects of model performance detailed below.

- Probability of the Alternative Class (PAC): measures how close a model's prediction is to the second-best class. Values near 0 indicate a good fit and values near 1 suggest misclassification.
- Silhouette Plot: used to compare classifiers in which a higher average silhouette width indicates a stronger classifier.
- Quasi Residual Plot: plots PAC against one of the features in the dataset. This visual can show how prediction quality changes from input to input.
- Class maps: reflect the probability that a data point belongs to an alternative class (PAC on the y-axis) against a measure of how far it is from its own class (on the x-axis).
- Stacked Mosaic Plots: helps visualize the confusion matrix allowing more clarity into the proportion of correct predictions vs the proportion of incorrect predictions.

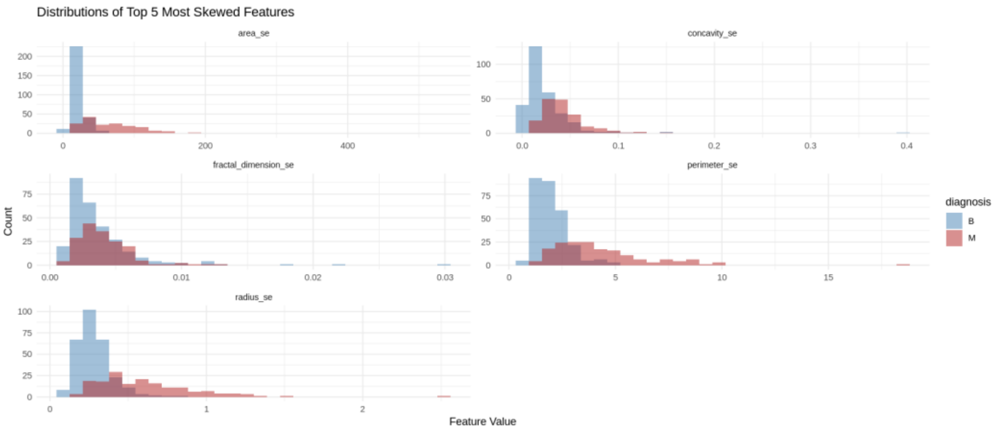
By applying this model and these visual tools to the breast cancer dataset, this report aims to show how these methods can be used to gain a more holistic understanding of model performance in the highly relevant and important context of cancer diagnostics.

Dataset Description

The UCI Breast Cancer (Diagnostic) Data Set contains diagnostic features computed from digitized images of fine-needle aspirate (FNA) biopsies of breast masses. These measurements describe characteristics of the cell nuclei present in the sample, providing a feature set for modeling and classification. The goal is to classify tumors as either malignant or benign based on these features.

The dataset comprises 568 rows and 33 columns. Each row represents a unique patient case, and the columns include an id, a diagnosis label (our target variable), and 30 numerical features capturing statistics like radius, texture, perimeter, area, smoothness, and symmetry of the cell nuclei. The diagnosis column includes two classes: 356 benign cases and 212 malignant cases. Upon initial inspection of the data, two columns called 'id', and a trailing placeholder column labeled '...33' did not contain predictive information. The id field is simply an identifier and ...33 is an empty column often created by formatting or exporting issues. These two features were removed early in the preprocessing phase. Notably, there were no missing values in the dataset which meant there was no need for imputation, ensuring that the raw structure was preserved.

A summary of the numerical features revealed several variables with highly skewed distributions. Features like area_se, concavity_se, and fractal_dimension_se showed particularly high positive skewness, with area_se having a skewness value above 5. Because of this, normalization was applied after an 80/20 train-test split, using the distribution of the training set to avoid data leakage.



In addition to skewness multicollinearity was assessed by computing a correlation matrix across all numeric features. Several features showed strong correlations above 0.90, including pairs such as perimeter_mean and radius_mean (r = 0.9978), area_worst and radius_worst (r = 0.9835), and concave points_mean and concavity_mean (r = 0.9245). Because the final model uses an SVM which can be sensitive to redundant features, it was decided that it would benefit performance to remove eight highly correlated features from the training and testing data. These features included perimeter_mean, perimeter_worst, area_mean, area_worst, perimeter_se, area_se, concave points_mean, and texture_mean.

This thorough understanding and preparation of the dataset ensures that our classification model is trained on features that are relevant, scaled appropriately, and free of multicollinearity. These preprocessing steps support more stable and interpretable model performance, especially when combined with visualization tools like class maps and quasi residual plots that rely on clean, well-prepared input data.

Analysis

Model selection – SVM with RBF kernel

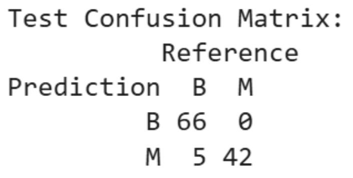
In this task of developing a classification model to predict whether the breast tumor is benign or malignant, the Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel has been selected due to its effectiveness in high-dimensional spaces and its capability in binary classification tasks. Given that the Breast Cancer Wisconsin dataset consists of 30 informative features and a relatively small number of observations (i.e., 568 instances), SVM's ability to find an optimal separating hyperplane that separates the classes with the maximum margin makes it a suitable choice. Moreover, by using a non-linear kernel i.e., RBF, SVM can perfectly handle the boundary between benign and malignant tumors, contributing to a high AUC score and reliable predictive performance.

Compared with other popular classifiers, such as Random Forests which use the power of ensemble learning, SVM has the advantages in this case where the decision boundary between classes is complex and non-linear. While Random Forest is robust and interpretable, SVM often achieves slightly higher performance on clean, high-dimensional datasets like this one.

Modelling and performance evaluation

The training was conducted on a normalized and feature-reduced version of the dataset since SVM assumes all variables are independent and hence is sensitive to multicollinearity. The SVM model, with the following parameters has a strong generalization ability on unseen data with an accuracy of 96% and an AUC of 99%.

- **Preprocessing**: Centering and scaling of predictors
- **Cross-validation**: 5-fold cross-validation
- **Performance metric**: Area Under the ROC Curve (AUC)
- **Tuning Parameters**: Cost (C) = 1, Sigma = 0.1



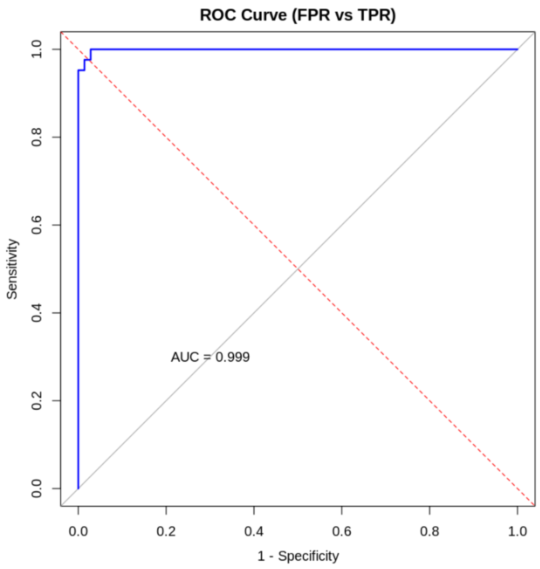
```
Test Confusion Matrix:
          Reference
Prediction  B  M
         B 66  0
         M  5 42
```

Results indicate that the SVM model performs exceptionally well on the dataset, achieving high classification accuracy and excellent discriminatory ability, as reflected in the ROC curve below. The confusion matrix shows that there are 66 true negatives, 42 true positives, 5 false positives and 0 false negatives.

So, the following parameters are calculated to further evaluate the model performance:

- **Recall** = TP / (TP+ FN) = 42 / (42+0) = 100%
- **Precision** = TP / (TP + FP) = 42 / (42 + 5) ≈ 89.4%
- **F1 Score** = 2 × (Precision × Recall) / (Precision + Recall) ≈ 2 × (0.894 × 1) / (0.894 + 1) ≈ 94.4%

It achieved a perfect recall (100%) for malignant cases, meaning it did not miss any cancer diagnoses, which is critical in a medical context. The model misclassified 5 benign cases as malignant (false positives), resulting in a precision of ~89.4%. While this may lead to some unnecessary follow-up tests, the high recall ensures no malignant cases were missed, making the model highly suitable for



early detection tasks.

Reflection on 99% AUC

An AUC of 99% can rarely exist in a real-world dataset. K-fold cross-validation and even different random seeds are adopted to examine whether this is an overfitting or data leakage problem.

However, it turns out that high accuracy and high AUC are achievable for this dataset while the model is well generalized without data leakage. This dataset is much simpler than the real-world one since it has 30 numeric features that are already well-engineered (e.g., mean, standard error, worst values of radius, texture, etc.). There is no missing value, and the class separation is very clear. Several features such as “worst radius”, “worst perimeter” have a very strong correlation with the diagnosis, serving as a signal for the positive class; in this case, even simple models perform well.

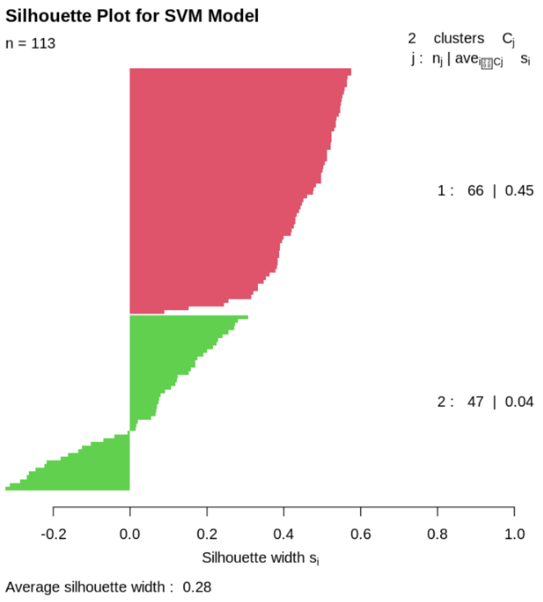
As a benchmark, a community contributor (Buddhini, 2017) on Kaggle achieved 96.5% in accuracy on test dataset by applying all features to the random forest classifier and another author (Ammar, 2025) got 95.6% in accuracy on the test data through Logistic Regression.

The evidence proves the validity and reliability of our SVM model, supporting the strong AUC score of 99% and confirming the high predictive performance instead of overfitting or data leakage issue.

Visualization interpretation

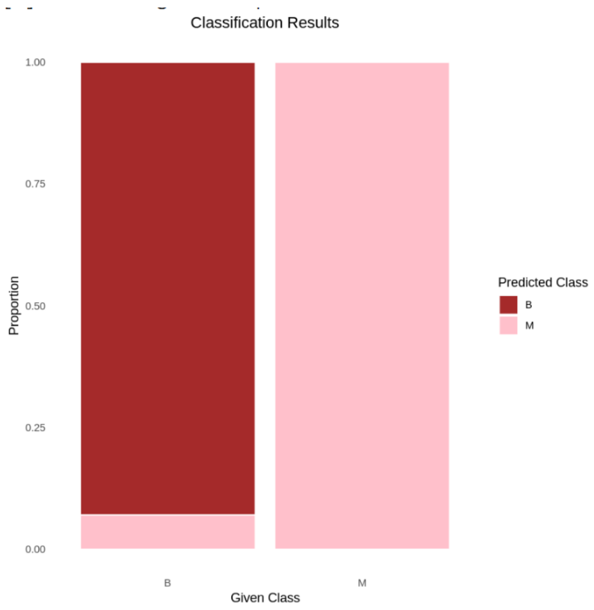
Besides the ROC curve, various visualization techniques are adopted to extract and interpret information from plots.

- **Silhouette plot**



Silhouette Plot is used here to assess the structure of the test data after classification and the silhouette score shows how similar an object is to its own cluster compared to other clusters. From the silhouette plot, there are benign (green) and malignant (red) classes with a test data size of 113. The malignant class has an average silhouette width of 0.45 which indicates it is well-clustered, while the figure is 0.04 for the benign class, resulting from possible mislabeling of the class. This can be proved by the negative part in the benign class that certain data points are misclassified into malignant class. This aligns with the result from the confusion matrix that there are 5 false positive cases.

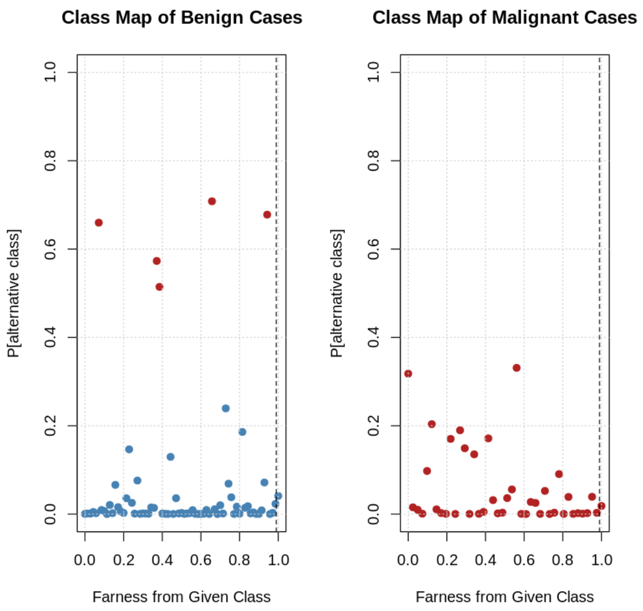
- **Stacked Mosaic Plots**



Stacked mosaic plot is a way of visualizing the confusion matrix by comparing the predicted vs. actual class labels. For the actual benign class, almost everything has been correctly predicted with a small portion of data classified into the malignant category; for the actual malignant class, nearly everything is predicted correctly as M class. It clearly shows the proportion of each class in a clean, layered way, which might help more for classification tasks with over 2 classes.

- **Class map**

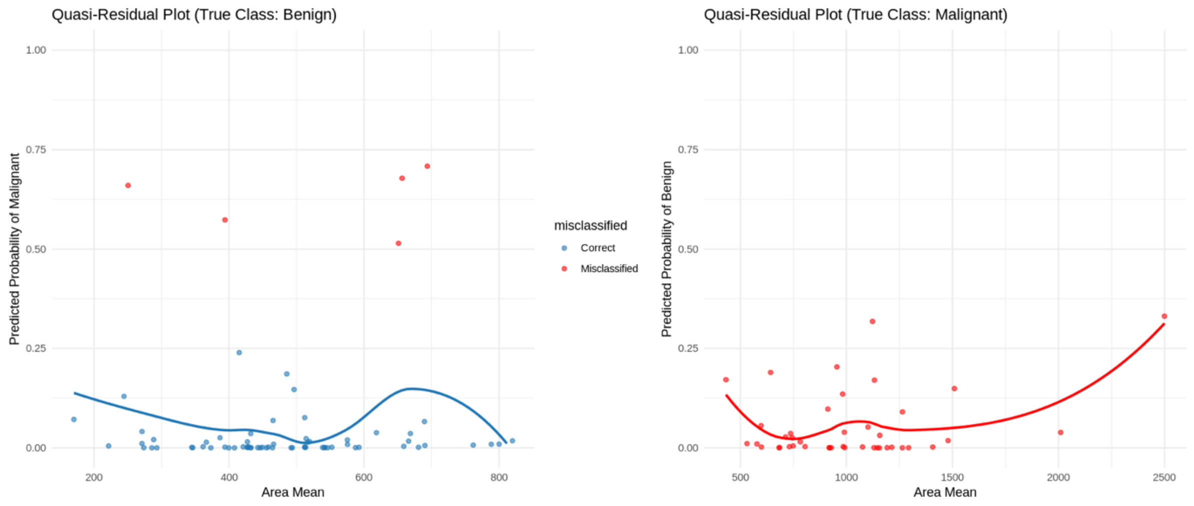
The class maps visualize the predicted probability of the alternative class against the farness value. The predicted probability of each class by the classification model is plotted.



The class map of the benign cases demonstrates that most blue points are located around PAC = 0. The model correctly classified these cases as benign with confidence and assigns a very low probability of belonging to the malignant class. The 5 red points are misclassified with a PAC > 0.5, representing label noise. These were classified as malignant while actually belonging to benign, which corresponds to the 5 false positives in the confusion matrix. A few points have a farness value above > 1, which entails that these cases are classified correctly but are far from its class.

In the malignant class map, most red points are also located around PAC = 0, meaning the SVM correctly predicted the malignant cases with confidence. There are no blue cases present, which means that there are no misclassifications, also corresponding to the 0 false negatives in the confusion matrix.

- **Quasi residual plot**



The quasi residual plots demonstrate the predicted probability of the alternative class against the relevant variable, area mean. This variable was appointed as there is a strong correlation between malignancy and the size of the tumor.

In the benign class plot, the blue points are scattered around zero on the y-axis, indicating that the model predicts these cases with high confidence. The majority of the benign cases have a low probability of being classified as malignant, however a few points with a PAC > 0.5 are misclassified and actually belong to the benign class, aligning with the confusion matrix. The smoothed line demonstrates a relatively consistent prediction across the area mean values, between a PAC of 0 and 0.25. The benign cases are located between an area mean of 200 and 800, which is along expectations as benign tumors tend to be smaller.

The residual plot of the malignant cases demonstrates similar findings where most red points are also located around PAC = 0, indicating that the malignant classes are correctly classified and have a low probability of belonging to the benign class. The PAC values range between 0 and 0.3, entailing that there are no misclassifications as aligned with the confusion matrix. The area mean values range between 500 and 2500, which is expected as malignant cases have a larger tumor size. The smoothed line rises as the area mean increases, demonstrating a slightly higher PAC for those cases yet still below the threshold for misclassification.

Conclusion

For the Wisconsin Breast Cancer dataset, the SVM model with RBF kernel was selected as classification technique for its effectiveness on high-dimensional data. The classification model achieved an AUC of 0.99 with an accuracy of 96%, demonstrating the model's ability to correctly classify benign and malignant cases. The strong recall score of 100% and the precision score of 89.4% further support the model's optimal performance.

The confusion matrix consists of 66 true negatives, 42 true positives, 5 false positives and 0 false negatives. This demonstrates the model's reliability in medical diagnosis, especially given the absence of false negatives, which could be very dangerous if an actual malignant case was classified as benign. The 5 false positives classify benign as malignant which is less severe but can still have consequences due to additional tests and procedures having to be performed.

Among the various visualizations used, the class map provided the most valuable insights for evaluating the performance of the SVM model. The class map provides information on misclassifications, the prediction confidence (PAC) and the farness value. With separate plots for the benign and malignant cases, the visualization allows for class specific analysis that aligns with the confusion matrix. Each class map demonstrates the correctly and incorrectly classified points and how confident the model is in its prediction. Additionally, it provides insights on whether cases are misclassified due to label noise or feature noise. Furthermore, the plots also inform on the farness value of the cases and how far they are located from their class. While all visualizations methods provide valuable information, the class map offers the richest insight and interpretation for the SVM model evaluation.

Citations

Ammar L., 2025. Breast Cancer Diagnostic - Logistic Regression. https://www.kaggle.com/code/ammarlouah/breast-cancer-diagnostic-logistic-regression

Buddhini W., 2017. Breast cancer prediction. https://www.kaggle.com/code/buddhiniw/breast-cancer-prediction#Conclusion

Khalid, A., Mehmood, A., Alabrah, A., Alkhamees, B. F., Amin, F., AlSalman, H., & Choi, G. S. (2023). Breast cancer detection and prevention using machine learning. *Diagnostics*, *13*(19), 3113. https://doi.org/10.3390/diagnostics13193113

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (n.d.). *1.4. Support vector machines*. Scikit-learn. https://scikit-learn.org/stable/modules/svm.html

William Wolberg, O. M. (1993). *Breast cancer wisconsin (Diagnostic)* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B