Predicting Breast Cancer related Malignant Cell Growth using Machine Learning

Olusegun Odumosu [1*], Daniel G. Amoako [1, 2*]

[1] Department of Integrative Biology and Bioinformatics, College of Biological Sciences, University of Guelph, Guelph, Ontario, Canada, N1G 2W1

[2] Department of Pathobiology, Ontario Veterinary College, University of Guelph, Guelph, Ontario, Canada, N1G 2W1

[*] Correspondence: Olusegun Odumosu, Email: odumosu.segun@gmail.com and Daniel Gyamfi Amoako, Email: amoakodg@gmail.com

Running title: Machine learning for Predicting Breast Cancer Status

Abstract

This study aimed to compare the performance of six different machine learning algorithms (Elastic net, CART, SVM, RF, KNN and Bagged CART) for the task of predicting breast cancer status based on a set of features. The evaluation of the models was based on several performance metrics, including accuracy, F1-score, recall, and area under the receiver operating characteristic (ROC) curve. The results showed that the SVM and Elastic net models outperformed the other models, with the highest accuracy, F1-score, and recall. Additionally, the ROC curve achieved a high area under the curve (AUC), indicating a high degree of discrimination power for both SVM and Elastic net models. The LR model also had the best learning curve, achieving high performance with increasing training set size, generalizing well to new data compared to the other models, and are less likely to overfit. Analysis of feature importance revealed that each model was influenced differently by the input features. Overall, the study concluded that SVM and Elastic net models is the recommended model for breast cancer status prediction.

Keywords: Breast Cancer, Status, Classification, Machine learning, Evaluation metrics, Feature importance

Introduction

Breast cancer is a significant public health problem, with an estimated 2.3 million new cases diagnosed in 2020 alone, making it the most commonly diagnosed cancer in women worldwide (Bray et al., 2021). Early detection and diagnosis are crucial for successful treatment and survival, and the use of machine learning algorithms has shown promise in improving breast cancer diagnosis accuracy.

In recent years, researchers have explored the potential of machine learning algorithms in detecting and diagnosing breast cancer. For example, Wang et al. (2020) developed a deep learning algorithm that outperformed radiologists in detecting breast cancer on mammograms. Another study by Esteva et al. (2019) demonstrated the effectiveness of a convolutional neural network (CNN) in identifying skin cancer.

The aim of this assignment is to evaluate the potential of various machine learning models in predicting the cancer status of breast tissue samples as benign or malignant. The dataset provided contains 30 predictor variables and the response variable, cancer status, with two possible values, M (malignant) and B (benign). Six different machine learning models will be considered, including elastic net, KNN, CART, support vector machine, bagged CART, and random forests.

Appropriate measures of accuracy, such as MSE, AUC, sensitivity, and specificity, will be reported for all models, and the most important and influential covariates in the trained model(s) will be highlighted. We also explored the quadratic effects of all covariates on trained models. Finally, the best model for detecting breast cancer will be recommended, and its potential practical usefulness in real-world clinical settings will be discussed.

Methods

Detail of data analysis

In this analysis, six models were evaluated, namely, elastic net, KNN, CART, support vector machine, bagged CART, and random forest. Elastic net is a linear regression method that simultaneously performs variable

selection and regularization (Zou & Hastie, 2005). Lasso and ridge regression, which are incorporated into the elastic net model, are used for selecting important features and shrinking coefficients, respectively (Owen, 2007). KNN is a non-parametric classification algorithm that determines the class of new data based on the k-nearest neighbors in the feature space (Guo et al., 2003). CART, a decision tree-based method, recursively divides the data into homogeneous subsets to predict the outcome variable (Lewis, 2000). SVM is a classification algorithm that seeks to find the best hyperplane that separates the data points into two classes (Noble, 2006). Bagged CART creates multiple decision trees using bootstrap sampling to reduce model variance and prevent overfitting (Bauer & Kohavi, 1999). Random forest is an ensemble learning method that uses multiple decision trees and outputs the mode of classes (Rodriguez-Galiano et al., 2012).

To prepare the dataset, covariates were checked for multicollinearity, and variables with a correlation higher than 0.9 were removed. Additionally, a principal component analysis (PCA) was conducted on both the full feature dataset and the dataset with multicollinearity addressed. The robust caret package was used for preprocessing, training, tuning, and prediction (Prabhakaran, 2018). The models were trained using ten-fold cross-validation, repeated three times for validation. Additionally, the features were scaled during training to normalize the range of independent variables (Ozsahin et al., 2022). The quadratic effects of all covariates were explored and the most important/influential ones in our trained model(s)were highlighted.

Results and Discussion

3.1 Data Exploration

a. Class distribution between cancer status

The class distribution in both the training and testing dataset shows an imbalanced distribution of the response variable cancer status, where the majority of the samples are classified as Benign and the remaining samples as Malignant(Figure 1).

Figure 1: Bar plots showing class distribution of cancer status.

b. Correlation between variables (Multicollinearity check)

The correlation plot between variables in the dataset (Figure 2) indicates the strength and direction of the relationships between the variables. It is important to consider the correlations between features, as those that are highly correlated with each other can lead to overfitting and reduce the model's performance (Géron, 2019). In general, features with correlation coefficients greater than or equal to 0.9 are considered to be substantially correlated and may be removed during data pre-processing. In this study, a few variables ("X1", "X2", "X6", "X7", "X12", "X13", "X20", "X22", "X23") were found to have high correlation coefficients and were removed from both datasets accordingly.

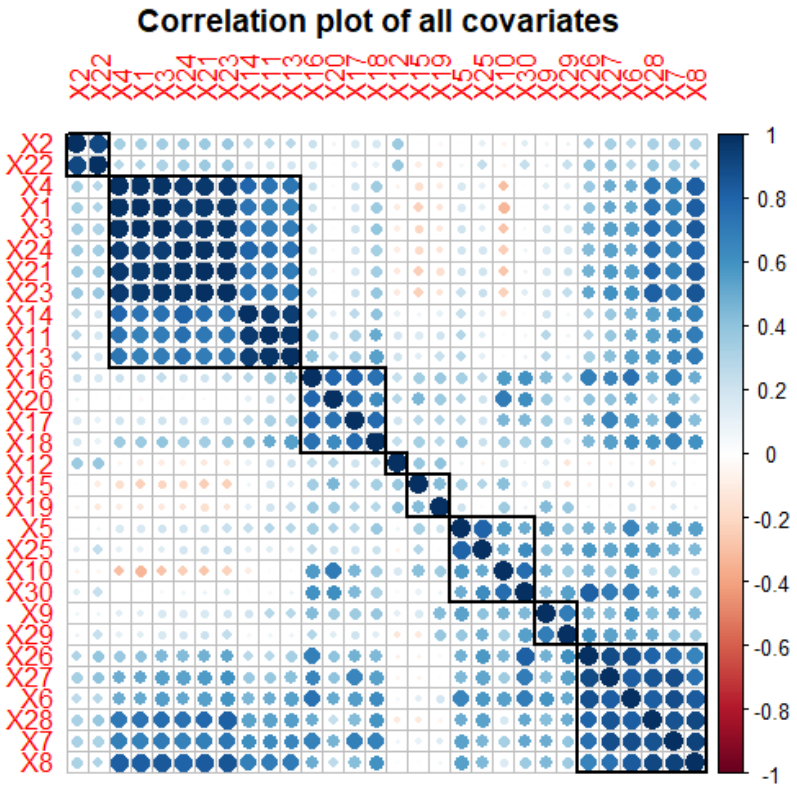

Figure 2: Correlation matrix plot between features.

c. Principal component analysis (PCA)

PCA analysis was performed to reduce the dimensionality of the dataset and identify the most important predictors that explain the variation in the data. The result of the PCA analysis indicates that only 8 principal components were needed to explain 95% of the variance in the data, suggesting that the dataset can be effectively represented with a lower number of dimensions (Figure 3a). The finding that the first 2 principal components (PC1 and PC2) managed to separate the cancer status well compared to the other components suggests that these components contain the most relevant information for predicting the cancer status (Figure b3). A detailed analysis of what variables are the most influential in the first 2 components showed that some variables may be associated with the cancer status prediction (Figure 3b). However, it is important to note that the interpretation of the principal components can be complex and may not always be straightforward.

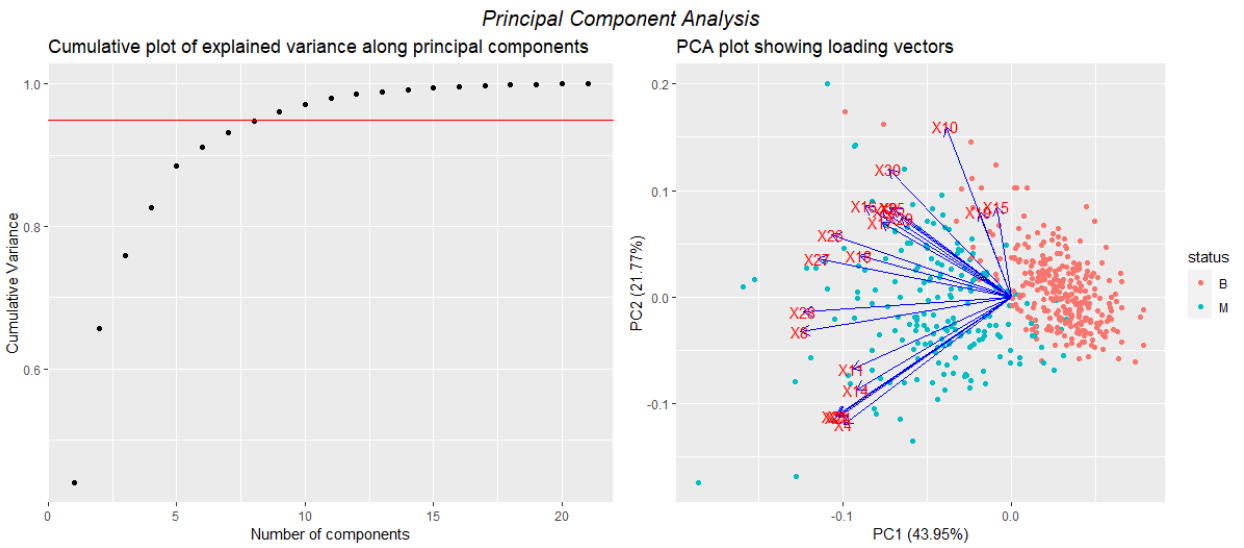

Figure 3: a: Cumulative plot of explained variance along principal components. b: PCA plot showing loading vectors with influential predictors.

3.2 Model Evaluation

a. Confusion matrix

The results of the confusion matrix for the different models provide an insight into how well each model is performing in correctly predicting the cancer status. From the results, we can see that all models have a high number of true positives, indicating that they are able to predict benign cases accurately (Figure 4). However, some models have a higher number of false positives than others. For instance, KNN have a higher number of false positives than the other models, indicating that it may wrongly predict benign cases as malignant. On the other hand, CART, Bagged CART, Random Forest and SVM have a lower number of false positives, indicating that they are better at predicting benign cases correctly.

Overall, the results suggest that Random Forest and SVM may be the better models as they have a high number of true positives and a low number of false positives, indicating that they are better at predicting both malignant and benign cases correctly. However, the final decision should be made by considering other performance metrics, such as accuracy,

precision, recall, and F1 score, in addition to the confusion matrix to make an informed decision about the best model for making prediction.

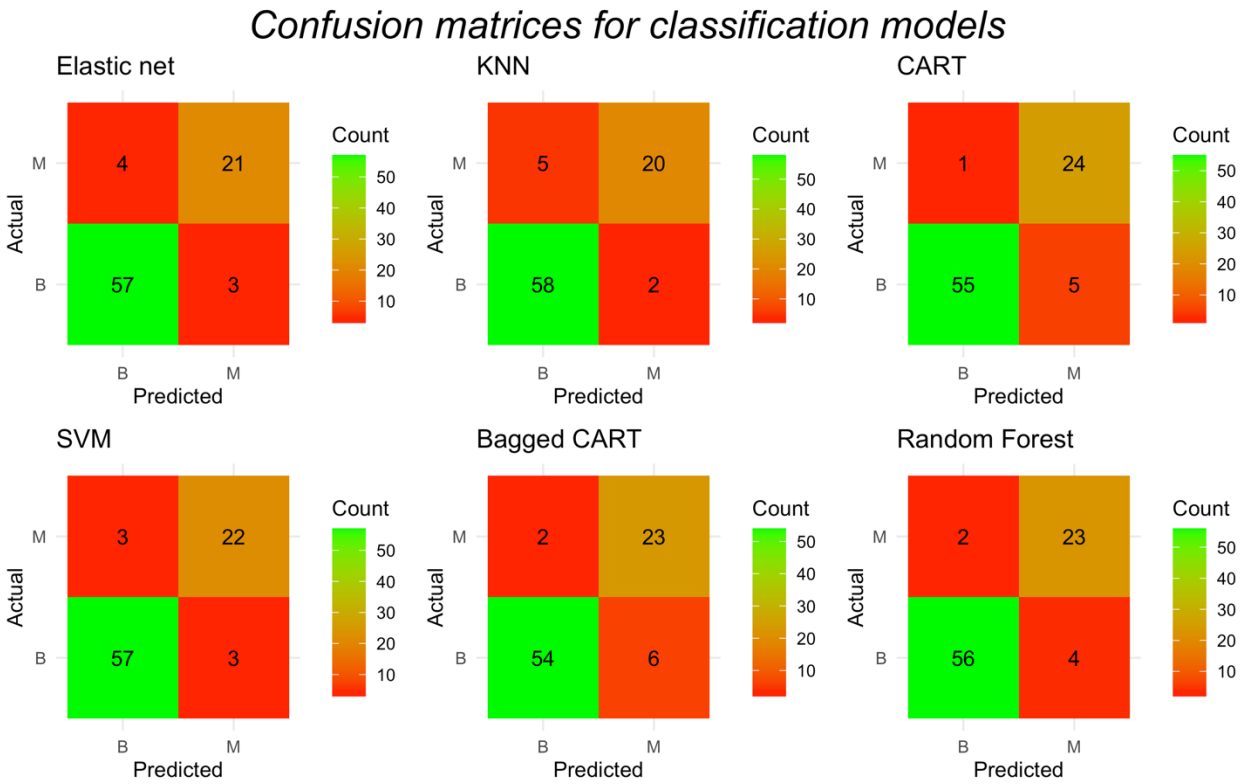

Figure 4: Confusion matrix for Cancer status prediction using all models.

b. Accuracy, Prediction, Recall, F1-score and ROC-AUC

The accuracy, precision, recall, F1-score, and ROC-AUC are important performance metrics in evaluating the effectiveness of classification models (Hicks et al., 2022). In terms of accuracy, most of the models achieved a high level of accuracy, ranging from 0.9058 to 0.9294 (Table 1). The SVM, CART, and Random Forest models had the highest accuracy score of 0.9294. However, it is important to consider other metrics as well, such as precision, recall, and F1-score, as they provide more information on the model's performance in detecting true positives and minimizing false positives and false negatives.

With the exception of Bagged_CART, the precision scores of all the other models are high (ranging from 0.8275 to 0.9090), suggesting that the

remaining models have a low false positive rate, which is important in identifying patients who do not have malignant cancer status accurately (Table 1). The recall score, on the other hand, measures the ability of the models to identify true positives, and the SVM model had a recall score equal to its precision score, indicating that the model has correctly identified all the true positives. However, the KNN model had the lowest recall score (0.8), indicating that it may have missed identifying some patients with benign cancer.

The F1-score is a harmonic mean of the precision and recall scores, and it provides an overall measure of the model's performance (Hicks et al., 2022). The F1-scores of all models are relatively close to each other, ranging from 0.84 to 0.8888, indicating that they are all performing fairly well in terms of both precision and recall (Table 1).

Table 1: Classification results of the models.

model name	accuracy	precision	recall	f1_score	auc_roc
Elastic_net	0.9176	0.8461	0.88	0.8627	0.988
Elastic_net_quadratic	0.9058	0.84	0.84	0.84	0.9873
KNN	0.9176	**0.9090**	0.8	0.8510	0.9673
KNN_quadratic	0.9176	0.8461	0.88	0.8627	0.9866
CART	**0.9294**	0.8275	**0.96**	**0.8888**	0.9413
CART_quadratic	0.9058	0.84	0.84	0.84	0.988
SVM	**0.9294**	0.88	0.88	0.88	**0.9933**
SVM_quadratic	0.9058	0.84	0.84	0.84	0.9873
Bagged_CART	0.9058	0.7931	0.92	0.8518	0.9793
Bagged_CART_quadratic	0.9176	0.8461	0.88	0.8627	0.9866
Random_Forest	**0.9294**	0.8518	0.92	0.8846	0.9866
Random_Forest_quadratic	0.9176	0.8461	0.88	0.8627	0.9866

The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification model at different thresholds. The area under the curve (AUC) is a measure of the

performance of the model, where a higher AUC indicates a better performance. The ROC-AUC score measures the ability of the model to distinguish between positive and negative classes. In this study, the SVM model had the highest ROC-AUC score of 0.9933, indicating that it has the best performance in predicting breast cancer status. This was followed by Elastic_net and Random Forest with a score of 0.988 and 0.9866 respectively (Table 1 and Figure 5). However, the CART model had the lowest AUC score of 0.9413, indicating that it may not be the best model for this problem (Table 1 and Figure 5).

Overall, the ROC curve and AUC analysis results support the findings of the previous sections, indicating that the SVM and Elastic net are more suitable for this cancer status classification problem, with SVM showing the best performance. It is worth noting that the ROC curve and AUC analysis provide an important performance evaluation for binary classification models (Gigliarano et al., 2014), and can provide additional insight beyond the traditional performance metrics such as accuracy, precision, recall, and F1-score.

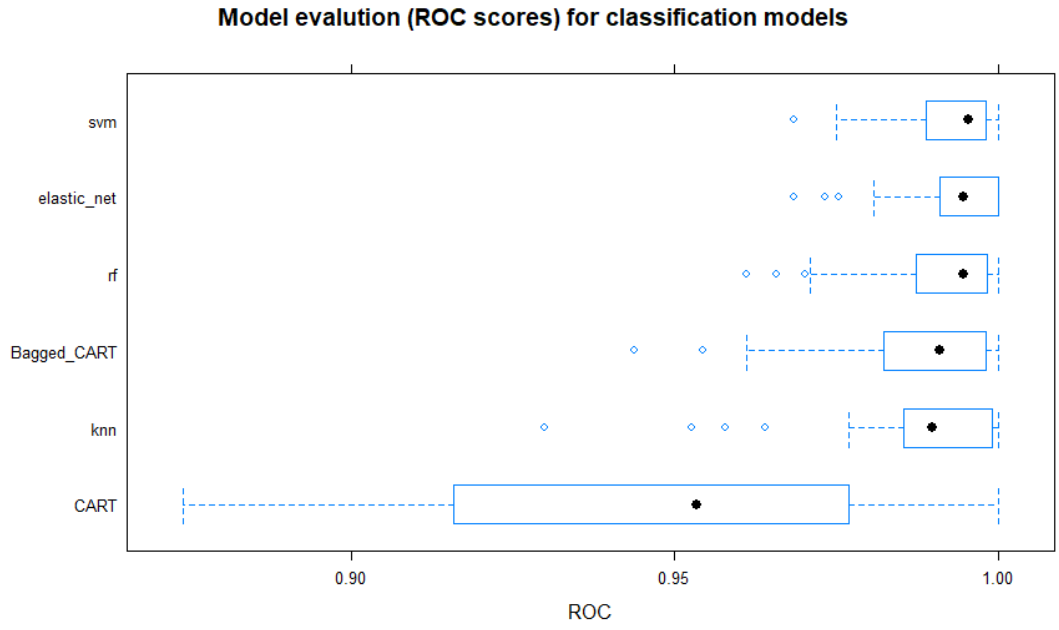

Figure 5: Model performance for all models for Cancer status prediction.

c. Learning curve

A learning curve is a graphical representation of the model's performance as a function of the training set size or the number of iterations. It is used to diagnose if the model is underfitting (high bias) or overfitting (high variance) and to determine if adding more data or iterations will improve the model's performance. In this study, the performance of six classification models has been evaluated based on their training score, testing score, and resampling score. The training score indicates how well the model fits the training data, while the testing score indicates how well the model performs on unseen data. The resampling score represents the mean cross-validation score of the model, indicating how well the model generalizes to new data.

According to the learning curve, the SVM, Bagged CART, Random Forest, Elastic net, KNN, and CART models all perform relatively well on this classification task (Figures 6 and 7). However, the SVM and Elastic net had the highest resampling scores (Elastic net=0.9675, SVM=0.96) which were not far from the training and testing score indicating that both models generalize well to new data compared to the others (Figures 6 and 7).

Learning curves for SVM, bagged CART and Random Forest

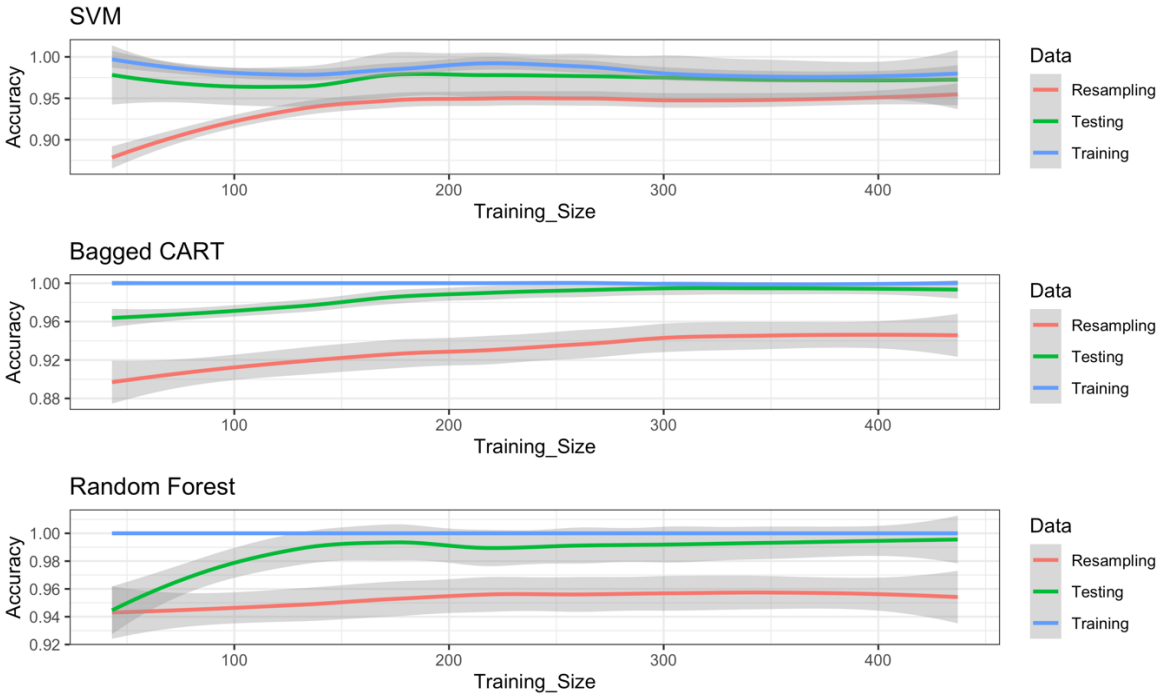

Figure 6: Learning curves for SVM, bagged CART AND Random Forest.

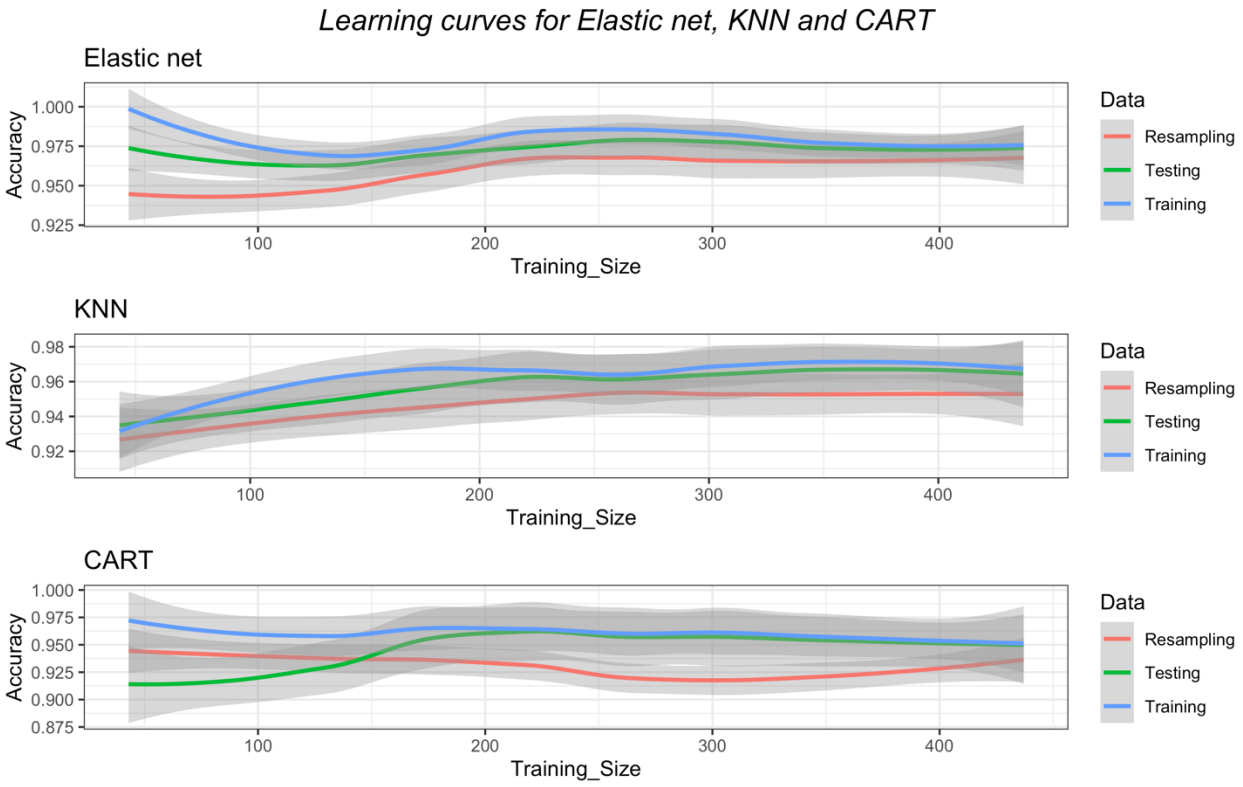

Figure 7: Learning curves for Elastic net, KNN and CART.

3.3. Feature importance and quadratic effects of covariates

Analyzing feature importance is crucial in comprehending the factors that contribute to the models' predictions (Chowdhury et al., 2020). However, the performance of the quadratic versions of the models did not differ significantly from the non-quadratic versions (Table 1), implying that the quadratic effects of the covariates did not have a significant impact on the overall model performance. Nonetheless, the results show that the models exhibit varying levels of sensitivity to different input features, making it essential in selecting the best model for a specific use case.

Furthermore, the variable importance trends in models with normal covariates differ from those with quadratic covariates. For instance, in the Elastic net models, the most crucial variables are X11, X21, X24, X28, X14, X8, and X27. However, the importance of these variables varies

when quadratic covariates are used. In the quadratic Elastic net model, X11, X21, X24, X14, and X27 still have high importance scores, while other variables like X28_sq, X29_sq, X25_sq, and X11_sq are also ranked high. Similarly, other models exhibit differences in variable importance rankings. Thus, we can conclude that there are general variations between the normal models and quadratic models regarding feature importance trends. It is recommended that future studies explore other possible transformations or interactions between variables to improve model performance. Polynomial transformations may be useful in capturing more complex relationships between variables that were not captured by the quadratic effects examined in this study. Furthermore, it is recommended to investigate the impact of including interaction terms between variables in the models. Interactions between variables may play a significant role in predicting outcomes, particularly when analyzing complex systems with numerous interacting components.

4.0. Conclusion

In this study, six machine learning models (Elastic net, CART, SVM, RF, KNN and Bagged_CART) were compared for their ability to predict breast cancer status. The evaluation metrics were used to assess the performance of these models. The SVM and Elastic net models demonstrated the highest evaluation metrics, including accuracy, F1-score, and recall, as well as a good balance between true and false positives. The ROC curves of the SVM and Elastic models achieved highest AUC, indicating a high degree of discrimination power. Furthermore, SVM and Elastic net had the best learning curves, achieving high performance with increasing training set size, generalizing well to new data compared to the other models, and are less likely to overfit. Based on these results, it can be concluded that the SVM and Elastic models are the best models for predicting breast cancer status using the provided features. Therefore, they are recommended for the breast cancer status prediction problem. Overall, the study provides valuable insights into the performance and feature importance of different machine learning algorithms in predicting breast cancer status. However, it is important to consider other factors such as computational efficiency, interpretability, and feasibility in real-

world applications, and to validate the results on larger datasets before applying the models in a clinical setting.

Acknowledgements

We would like to express our gratitude to Professor Dirk Steinke and Maysa Niazy for their invaluable guidance throughout this course. We also extend our appreciation to the Department of Integrative Biology and Bioinformatics, College of Biological Sciences, University of Guelph, Guelph, for providing us with the necessary resources to complete our research. We are proud to be part of such a reputable institution.

Authors' contributions

All authors improved and contributed to the editing of the manuscript. All authors read and approved the final version of the manuscript. All authors contributed equally to the study.

Disclosures

The authors have no competing interests to declare.

Funding information

No funding received

References

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, *36*, 105-139.

Bray, F., Jemal, A., Grey, N., Ferlay, J., & Forman, D. (2021). Global cancer transitions according to the Human Development Index (2008–2030): a population-based study. The Lancet Oncology, 22(2), 205-215.

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family medicine and community health*, *8*(1).

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Gigliarano, C., Figini, S., & Muliere, P. (2014). Making classifier performance comparisons when ROC curves intersect. *Computational Statistics & Data Analysis*, 77, 300-312.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. *Lecture Notes in Computer Science*, *2888*, 986–996

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2019). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12(1), 5979.

Lewis, R. J. (2000). An introduction to classification and regression tree (CART) analysis. *Annual meeting of the society for academic emergency medicine in San Francisco, California* (Vol. 14). California: Department of Emergency Medicine Harbor-UCLA Medical Center Torrance.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567.

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, *443*(7), 59–72.

Ozsahin, D. U., Mustapha, M. T., Mubarak, A. S., Ameen, Z. S., & Uzun, B. (2022). Impact of feature scaling on machine learning models for the diagnosis of diabetes. *2022 International Conference on Artificial Intelligence in Everything (AIE)* (pp. 87-94). IEEE.

Prabhakaran, S. (2018). *Caret Package – A Practical Guide to Machine Learning in R*. Machinelearningplus. https://www.machinelearningplus.com/machine-learning/caret-package/

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS journal of photogrammetry and remote sensing*, *67*, 93-104.

Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., Zhou, J., & Cheng, L. (2020). Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. Nature Communications, 11(1), 1-9.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005). *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *67*, 768–768.

List of Figures

List of Tables