Early Cancer Detection Using Machine Learning Models

IT1244 Team 2: Bryan Kee Tze Ren, Daniel Leong Zhi Kai, Lau Ming Jie, Soh Kai Le

Introduction

A major challenge in cancer diagnostics is achieving the right balance between sensitivity and specificity for early detection. False negatives can lead to untreated cancers, while false positives may result in unnecessary procedures and higher healthcare costs. Traditional methods rely on a combination of variable human judgment and a battery of invasive tests that can be uncomfortable for patients and resource intensive for the healthcare sector.

To tackle this binary classification problem, we propose using machine learning (ML) models to analyze DNA samples from a single blood draw. Our aim is to enhance diagnostic accuracy in distinguishing between those with and without cancer. We will use Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM), all of which are effective for binary classification and can handle imbalanced datasets. SVM performs well in high-dimensional tasks, improving sensitivity and early detection (Liu, 2021). LR, with L1/L2 regularization, reduces overfitting while selecting important biomarkers (Wu, 2018). RF is robust against overfitting and ranks features effectively, making it valuable for identifying cancer biomarkers (Toth, 2019).

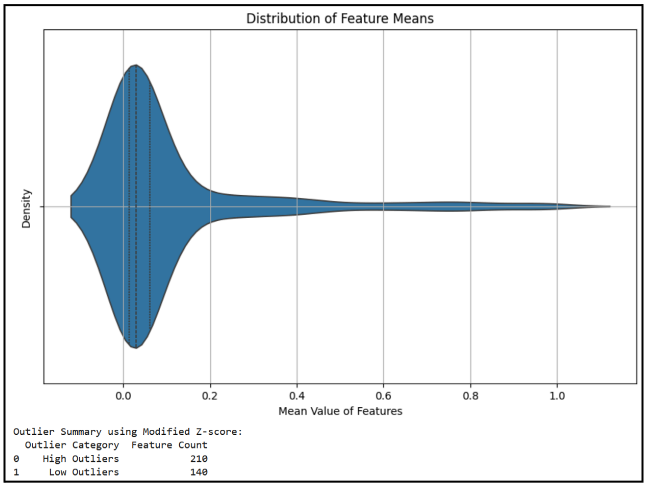
Despite their strengths, these models can struggle with data imbalance and feature selection challenges, especially with many features. To enhance performance, we will apply the Synthetic Minority Oversampling Technique (SMOTE) for class imbalance and use Principal Component Analysis (PCA) with Point Biserial Correlation (r_bp) for efficient feature selection.

Dataset

The dataset comprises 841 training samples and approximately 409 test samples, focusing on classifying healthy versus early-stage cancer samples. It includes 350 continuous features, with each feature representing the maximum normalized frequency of DNA fragment lengths. The last column indicates the class label, categorizing samples as either healthy or cancerous. This dataset presents two significant challenges for accurate classification: firstly, it is notably imbalanced, containing only 80 healthy samples compared to 761 cancer samples; Secondly, the high dimensionality of the feature space can complicate the learning process.

Feature Scaling with Standardization

Feature scaling is essential in our analysis to ensure that all features contribute equally to distance calculations in algorithms like PCA and SVM. The significant variation in feature means, as illustrated by the violin plot of feature distributions (Figure 1), highlights the necessity of feature scaling. We chose standardization due to the considerable presence of outliers in the dataset. Employing the modified Z-score approach, as recommended by Bhargavi and Sireesha (2021), effectively minimizes the influence of extreme values. We defined outliers as values exceeding 3 standard deviations from the median, requiring that they account for more than 1% of the sample size to be deemed significant. Our analysis revealed that approximately 60% of the features displayed a high number of outliers. In summary, the combination of variable magnitudes and a high incidence of outliers led us to standardize the 350 feature inputs.



```
Outlier Summary using Modified Z-score:
  Outlier Category  Feature Count
0    High Outliers            210
1     Low Outliers            140
```

Figure 1: Violin Plot of Feature Means and Outliers

Feature Selection with PCA and r_bp

Using the standardized features, we performed Principal Component Analysis (PCA) to reduce the dimensionality of our dataset. Initially, we selected principal components (PCs) that explained 99.99% of the total variance, corresponding to PCs 1 through 75. We then plotted the explained variance for each PC and applied the elbow method to determine the optimal threshold for selection, which indicated that PCs 1 to 3 were sufficient for our analysis (Figure 2)..

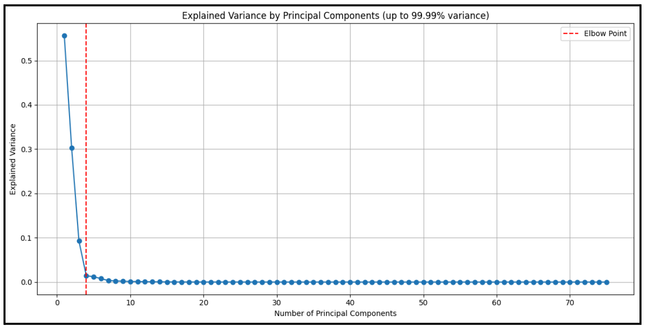


Figure 2: Explained Variance of Principal Components

However, as noted by Sharma and Saroha (2015), selecting principal components (PCs) based solely on explained variance may not adequately capture their relationship with the response variable. To address this limitation, we plotted the point biserial correlation for all selected PCs (Figure 3) to evaluate their correlation with the binary response variable, as discussed by Kornbrot (2014). Using the elbow method on this correlation data, we identified that PCs 6, 17, 4, 18, 45, 19, 30, 67, 42, 13, and 29 exhibited strong correlations with the response variable..

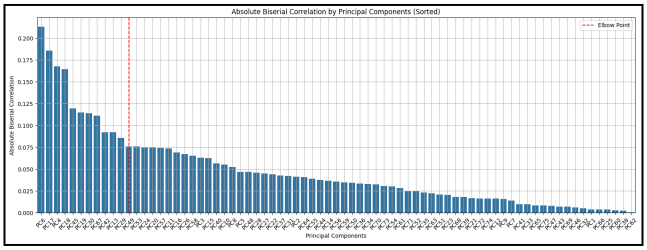


Figure 3: Biserial Correlation of Principal Components

Selected Principal Components and Visualization

In summary, we identified a total of 14 principal components (PCs) that meet the criteria of explained variance and strong correlation with the target variable. To illustrate these relationships, we selected box plots that highlight the differences among the PCs (Figure 4). Notably, while PCs 1 to 3 demonstrate high explained variance, they exhibit lower correlation with the response variable compared to PCs 6, 17, and 45. This comprehensive selection process, considering both variance and correlation, reinforces the effectiveness of our feature selection approach.

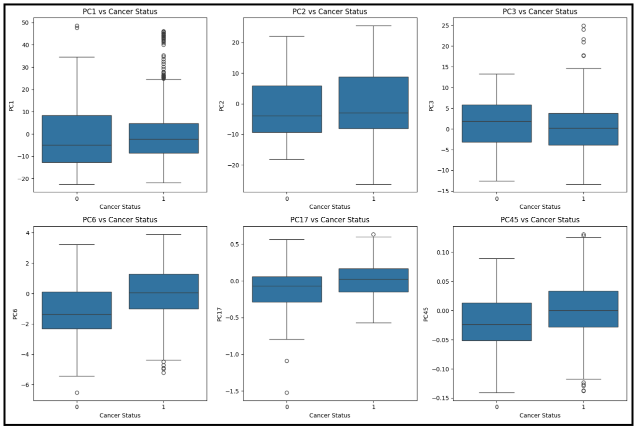


Figure 4: Box Plots of Selected Principal Components

Addressing Class Imbalance

To address the class imbalance in our dataset, where only 10% of the samples belong to the negative class (healthy) and 90% to the positive class (cancer), we recognized that using stratified K-fold cross-validation alone may not sufficiently resolve the imbalance, as the folds would still reflect the same skewed distribution. Therefore, we incorporated the Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes. By applying SMOTE, we generated synthetic samples for the minority class, enhancing the representation of the negative class in the dataset and ensuring that our model is trained on a more balanced distribution of classes (Figure 5). This approach aims to improve model performance and robustness in predicting both classes during training and evaluation.

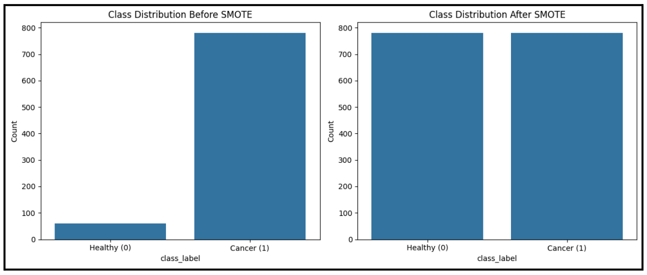


Figure 5: Balancing Class Distribution with SMOTE

Methods

This study tunes three models using 5-Fold cross-validation on SMOTE data to optimize hyperparameters for the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Given the significant class imbalance—approximately 90% cancer cases—we prioritize AUC as our primary evaluation metric. Relying solely on accuracy is misleading, as it can obscure model performance due to the dominance of the majority class. Metrics such as True Positive Rate (TPR) and Precision alone are insufficient, as they may not effectively capture the negative minority class. The ROC offers a comprehensive measure of discrimination between classes across thresholds, optimizing both TPR and False Positive Rate (FPR), which are crucial for sensitivity and specificity in medical contexts. This approach aligns with Hicks et al. (2022) and Mossotto et al., emphasizing the importance of using multiple metrics to gain a balanced understanding of classifier performance in imbalanced datasets.

Logistic Regression

We used logistic regression as the baseline model. While it is effective in handling binary outcomes and providing interpretable results, it struggles with high-dimensional samples. Regularized logistic regression has already seen success in the field of cancer classification, allowing for an improved classification accuracy by shrinking the regression coefficients (Wu et al., 2018). For this study, we undertook a tuning process for this model to optimize its performance. The objective was to identify the best combination of hyperparameters that maximizes the ROC AUC.

The tuning process utilized GridSearchCV, which employs k-fold cross-validation (with stratified splits) to ensure a fair evaluation of model performance across different subsets of the data. By employing a scoring method based on ROC AUC, we aimed to identify the hyperparameters that yield the highest predictive accuracy. In our logistic regression hyperparameter tuning, we selected a few key hyperparameters that directly impact model performance and convergence. We adjusted C, the inverse regularization strength, to control the trade-off between fitting the data well and avoiding overfitting. The penalty type (L1 or L2) was chosen to explore different regularization techniques, each influencing model sparsity and stability. We included solver choices (liblinear and saga) as they support L1 regularization, allowing flexibility with penalty types. max_iter was tuned to ensure adequate convergence in model training, while class_weight was set to balanced to address potential class imbalance.

Random Forest

We also implemented Random Forest as an advanced classification model to enhance our predictive capabilities in cancer classification. This ensemble method is particularly well-suited for handling high-dimensional datasets and mitigating the risk of overfitting, making it an excellent choice for our application.

To optimize the performance of the Random Forest model, we conducted a tuning process aimed at identifying the best combination of hyperparameters that maximizes the ROC AUC score. We initialized the Random Forest Classifier with a fixed random state to ensure reproducibility of results. For tuning our random forest model, we selected a focused set of hyperparameters that significantly influence model complexity, performance, and the handling of class imbalance. We varied n_estimators, the number of trees in the forest, to explore different levels of ensemble strength. max_depth was chosen to control the depth of each tree, helping manage model complexity and overfitting. Adjusting min_samples_split and min_samples_leaf allows us to experiment with tree growth control, influencing the model's ability to generalize. For max_features, we tested log2 and sqrt, optimizing feature selection per split. Finally, the criterion and class_weight parameters were chosen to test alternative split metrics (gini or entropy) and address any class imbalance issues.

Support Vector Machine

The last model implemented was Support Vector Machine as a high-dimensional classification model. It was chosen for its reliability in complex and imbalanced datasets such as the one given. It is effective in distinguishing binary classes with minimal overlap, effective for the high dimensional feature space present in cancer diagnostics. We optimized the model by tuning the parameters to obtain the highest ROC AUC such as done in the 2 earlier models resulting in a clear and optimal separation of healthy and cancerous samples.

In our SVM hyperparameter tuning, we focused on a few key parameters to optimize the model's balance between complexity and generalization. We varied the regularization parameter C to control the trade-off between maximizing the margin and minimizing classification error, aiming to find the best balance for our data. The kernel parameter was tested with linear, Radial Basis Function, and sigmoid options, allowing us to explore different transformations of the input space and adapt the model to nonlinear patterns. Additionally, we set class_weight to be balanced to handle any potential class imbalance, improving model robustness across diverse class distributions.

Results & Discussions

Model Evaluation

After the tuning process outlined in the methods section, we identified the optimal parameters for each model based on their performance on the training data. This involved systematically adjusting hyperparameters and employing cross-validation to ensure robust evaluation. Once the best parameters were determined, we formed the optimal models using these configurations. Subsequently, each model was tested on unseen testing data to evaluate their performance in a real-world scenario in terms of ROC and Precision Recall (Table 1, Figure 6, Figure 7).

Logistic Regression served as our baseline model, providing a simple yet effective approach to classification. The optimal configuration for this model included an inverse regularization strength of 0.1, balanced class weights, a maximum of 100 iterations, L2 regularization, and the Liblinear solver. Despite its interpretability and ease of implementation, the Logistic Regression model yielded the lowest performance metrics, with a testing ROC AUC of approximately 0.857 and a testing Precision-Recall AUC of about 0.979. These results suggest that while Logistic Regression is suitable for simpler tasks, it may struggle with more complex relationships within the data.

The **Random Forest** model demonstrated powerful capabilities in handling high-dimensional data and interactions among features. The best configuration included balanced class weights, an entropy split criterion, a maximum depth of 15, square root maximum features, a minimum of 5 samples required to split an internal node, a minimum of 5 samples at a leaf node, and a total of 15 decision trees. However, this model exhibited signs of overfitting, as indicated by a significant gap between its training ROC AUC of around 0.991 and its testing ROC AUC of 0.867. Although it achieved a high Precision-Recall AUC of 0.982, the overfitting concern limits its generalizability to unseen data, making it less suitable for practical applications.

In contrast, the **Support Vector Machine (SVM)** model emerged as the most effective solution, achieving the highest testing ROC AUC of approximately 0.882 and a Precision-Recall AUC of 0.982. The optimal configuration for the SVM included a regularization parameter of 5.0, balanced class weights, and a linear kernel type. Notably, the SVM was the only model capable of achieving a True Positive Rate (TPR) above 80% while maintaining a False Positive Rate (FPR) below 20% at a particular threshold. This capability allows the SVM to effectively balance sensitivity and specificity, making it a highly suitable choice for classification tasks.

Machine Learning Models	Training ROC - AUC	Testing ROC - AUC	Testing PR - AUC
Logistic Regression	0.919108	0.856906	0.978940
Random Forest	0.990703	0.867245	0.981946
Support Vector Machine	0.954898	0.881959	0.981956

Table 1: Model Performance for Various Metrics

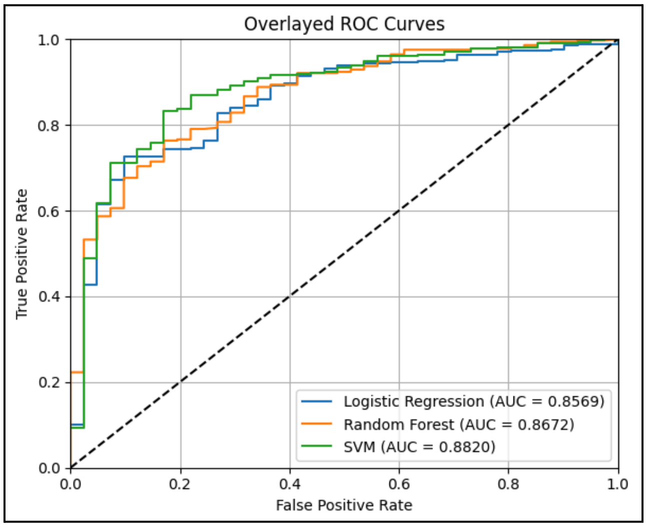


Figure 6: Receive Operating Characteristic Curves

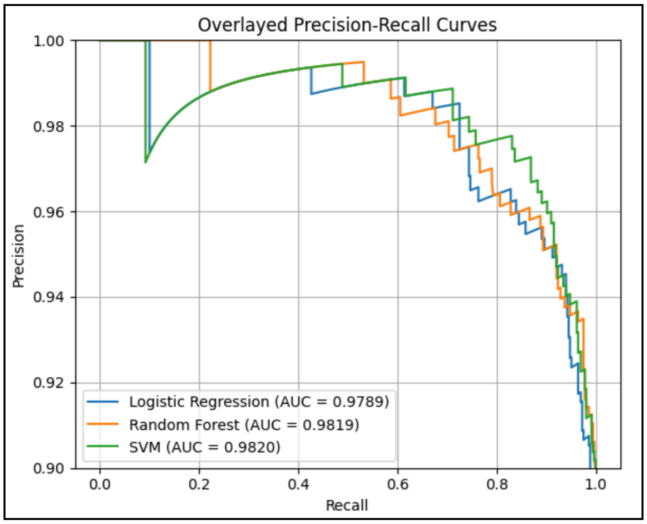


Figure 7: Precision Recall Curves

Conclusion

In conclusion, the Support Vector Machine (SVM) emerged as the best model for our classification task, achieving a testing ROC AUC of approximately 0.882 and a Precision-Recall AUC of 0.982. This performance was enhanced by employing Principal Component Analysis (PCA) for feature selection and utilizing SMOTE to address class imbalance, highlighting the effectiveness of these preprocessing techniques in improving model accuracy.

The societal impacts of this project are multifaceted. One major concern is **interpretability**; the use of Principal Component Analysis (PCA) means that the model relies on linear combinations of DNA fragment lengths, which may lack meaning for genomics experts trying to identify specific genes. This gap in understanding can hinder the model's acceptance and application in real-world scenarios. Additionally, the **impact on jobs** can be a double-edged sword. While machine learning techniques could replace some existing diagnostic tools and processes, they also hold the potential to complement the work of medical professionals. By providing data-driven insights, such models can enhance decision-making and improve patient outcomes, allowing professionals to focus on more complex cases that require human judgment.

References

Bhargavi, M. V., and Sireesha, V. 2021. A comparative study for statistical outlier detection using colon cancer data. *Advances and Applications in Statistics* 72: 41–54. doi.org/10.17654/0972361722003.

Davis, J. J., and Goadrich, M. H. 2006. The relationship between precision-recall and ROC curves. In *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, 233–240. doi.org/10.1145/1143844.1143874.

Hicks, S. A.; Strümke, I.; Thambawita, V.; Hammou, M.; Riegler, M. A.; Halvorsen, P.; and Parasa, S. 2022. On evaluation metrics for medical applications of artificial intelligence. *Nature* 12: 5979. doi.org/10.1038/s41598-022-09954-8.

Kornbrot, D. 2014. Point biserial correlation. In *Wiley StatsRef: Statistics Reference Online*. 22 April 2014. doi.org/10.1002/9781118445112.stat06227.

Liu, L.; Chen, X.; and Wong, K. 2021. Early cancer detection from genome-wide cell-free DNA fragmentation via shuffled frog leaping algorithm and support vector machine. *Bioinformatics* 37(19): 3099–3105. doi.org/10.1093/bioinformatics/btab236.

Rajpoot, C. S.; Sharma, G.; Gupta, P.; Dadheech, P.; Yahya, U.; and Aneja, N. 2024. Feature selection-based machine learning comparative analysis for predicting breast cancer. *Applied Artificial Intelligence: An International Journal* 38(1). doi.org/10.1080/08839514.2024.2340386.

Sharma, N.; and Saroha, K. 2015. A novel dimensionality reduction method for cancer dataset using PCA and feature ranking. In *Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 10–13 August 2015. IEEE. doi.org/10.1109/ICACCI.2015.7275954.

Toth, R.; Schiffmann, H.; Hube-Magg, C.; et al. 2019. Random forest-based modeling to detect biomarkers for prostate cancer progression. *Clinical Epigenetics* 11: 148. doi.org/10.1186/s13148-019-0736-8.

Wu, S.; Jiang, H.; Shen, H.; and Yang, Z. 2018. Gene selection in cancer classification using sparse logistic regression with L1/2 regularization. *Applied Sciences* 8(9): 1569. doi.org/10.3390/app8091569.