Cancer Classification Report

Paul Zhang, Julian Tham, Ravindran Dinanjanan, Ong Wei Xuan Titus

Introduction

Cancer is a leading cause of death, with the World Health Organization projecting a 77% increase in cases from 20 million in 2022 to an estimated 35 million by 2050 (World Health Organization, 2022; World Health Organization, 2024). This rise necessitates advancements in detection and diagnosis for effective early treatment (National Cancer Institute, 2024). While cancer is primarily understood as a result of genetic mutations, the existence of over 200 cancer types complicates universal treatment approaches (Biemar & Foti, 2013). Thus, leveraging recent advancements in Artificial Intelligence (AI) and computational power is essential for integrating AI with cancer genomics to identify mutation patterns and potential biomarkers for cancer detection and treatment.

Materials and dataset

In this project, we focus on classifying cells based on their relative normalised frequencies of 350 different DNA fragment lengths using a dataset of 841 training samples and 409 test samples. Both training and test datasets are heavily class-imbalanced where only 7% of the data were labelled as healthy.

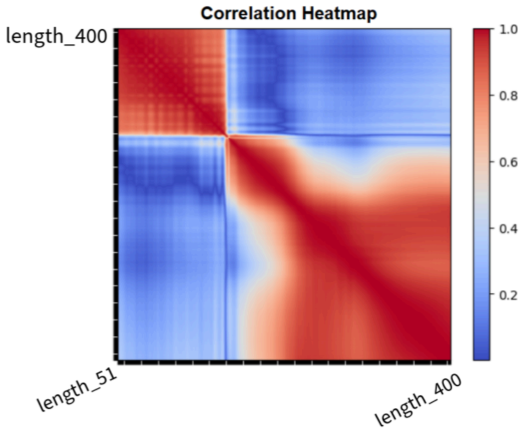


Figure 1: Correlation heatmap between different DNA lengths. Higher correlations are labelled red while lower correlations are labelled blue.

Through our preliminary exploratory data analysis (EDA), we identified many highly correlated feature pairs (refer figure 1). This finding justifies the need for feature extraction to identify key DNA fragments that are relevant for sample classification.

Methods

Multiple machine learning models were used in this study, from Logistic Regression to Neural Networks and SVMs, to classify healthy and early-stage cancer samples in a highly imbalanced dataset, with only 7% healthy cases. We addressed such class imbalances SMOTE for synthetic samples, and combining models in ensembles. Feature reduction with PCA also helped improve performance in some cases. This section will address the various methodologies used and relevant existing literature.

Baseline Reference using Logistic Regression

We first fitted the unprocessed dataset directly into a logistic regression (LR) model to benchmark subsequent models' performance.

Compensating For Class Imbalances

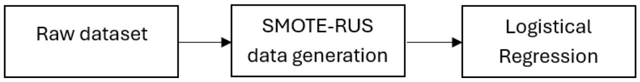


Figure 2: Pipeline for handling class imbalance before logistical regression.

Since our dataset consists of about 7% healthy sample, Synthetic Minority Over-sampling Technique (SMOTE) and Random Under-sampling (RUS) were utilised to create a more balanced dataset. SMOTE generates more data by randomly using parameters on an imaginary straight line between the reference point and its k number of neighbours. By generating more data from the minority class, the subsequent algorithms would better identify samples from the minority class, therefore reducing bias and improving accuracy. Chawla et al. (2011) showed that utilising SMOTE with under-sampling techniques would improve classifier performance. Hence, the combination of SMOTE-RUS was used to generate a new balanced dataset for LR.

Principal Component Analysis with Logistical Regression (PCA-LR)

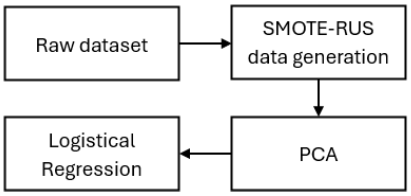


Figure 3: Pipeline for selecting key features using PCA before logistical regression.

To reduce computational load and complexity, dimensionality reduction techniques such as Principal Component Analysis (PCA) was used to extract key features within the dataset. PCA attempts to capture key features by maximising the variance of the data through the calculated eigenvalues and eigenvectors while retaining key information of the dataset. By using the top k eigenvectors, the data would be transformed into a new subspace with a lower dimensionality, thereby focusing on key features that contribute towards sample classification. This model applies PCA towards the SMOTE-RUS data, and PCA-transformed data is fitted using LR.

Neural Networks for Binary Classification

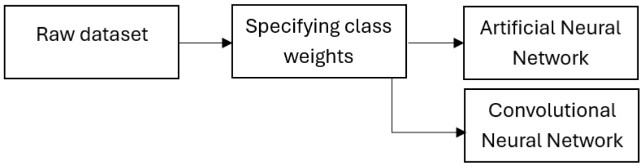


Figure 4: Pipeline for handling class imbalance before implementing neural network for classification

Neural networks (NN) would be suitable for such complex classification problems due to their ability to capture non-linear relationships. Since the DNA length features may interact with each other in non-additive ways, NN's multi-layer structure with relevant activation functions in each layer allows the model to learn and represent intricate feature interactions, which could be crucial for distinguishing cancerous from healthy samples. Furthermore, neural networks with several layers handle high-dimensional data better compared to logistic regression, where the prediction ability is affected by whether the features contribute linearly to the output. NNs are also better equipped to handle class imbalances compared to logistic regression as they have additional techniques such as dropout and batch normalisation. In this project, ANN and CNN were utilised for sample classification.

A simple ANN with three hidden layers was used to learn patterns progressively. The hidden layers were supplemented with dropouts of 0.3, 0.3, and 0.2, with the ReLU activation function to capture non-linear patterns. The model was compiled with binary cross-entropy loss with the adam optimiser for gradient descent. Handling class imbalances called for a different approach with the ANN. Instead of SMOTE, computing class weights and more strongly penalising misclassifying the minority class was found to produce significantly better metrics. Although CNNs' are often used with images, they are powerful when working with spatially structured data in general. In this case, the features consist of DNA fragment lengths in sequential order, making a CNN a viable candidate for this classification task. The CNN implemented consists of two convolution layers with 64 filters and ReLU activation functions. A kernel size of 3 was used with a dropout of 0.5, which was then flattened before being passed to two dense layers of size 64 and 1 with ReLU and sigmoid activation functions respectively. For the CNN, a combination of SMOTE and computing class weights to handle class imbalances resulted in better model performance. The model was trained with early stopping to prevent overfitting, where training is stopped when performance starts to degrade.

Ensembling Learning

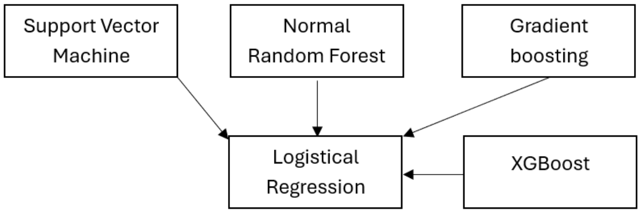


Figure 5: Pipeline for stacking classifiers to predict and classify cancer and healthy samples

Random Forest, Gradient Boosted Trees (GBT) and Extreme Gradient Boosting Trees (XGBoost)

Random Forest is an ensemble learning technique where multiple random decision trees are created. Subsequently, the algorithm combines the result and takes an average or majority vote for either regression or classification tasks respectively. By taking the average or majority vote, this reduces erroneous prediction from the variance occurring from randomness. Hence, random forest could mitigate overfitting from single decision trees while ensuring high accuracy (Belyadi & Haghighat, 2021).

GBT is another robust and flexible ensemble learning technique that utilises sequential decision trees and gradient descent to improve its prediction capabilities, allowing easy implementation and errors to be minimised between each iteration (Belyadi & Haghighat, 2021). Since our datasets consist of many features, GBT would capture complex feature interactions and reduce noise from irrelevant interactions, thereby improving prediction capabilities. However, GBT does not have regularisation and might be prone to overfitting. Hence. XGBoost can be utilised to incorporate regularisation to minimise overfitting. Additionally, XGBoost consists of parallelization and other gradient boosting techniques which improves training speed and further improves robustness and minimising overfitting (Belyadi & Haghighat, 2021).

Support Vector Machine (SVM)

Support Vector Machines (SVM) are supervised learning models effective for both linear and non-linear classification tasks. Unlike linear and logistic regression, SVM creates a hyperplane that separates classes using support vectors, which are identified by maximising the distance from the decision boundary, thereby ensuring optimal class separation (Belyadi & Haghighat, 2021; Pisner & Schnyer, 2020). SVMs excel in high-dimensional spaces and could handle non-linear features, making them suitable for applications like cancer detection, where addressing non-linearity is crucial.

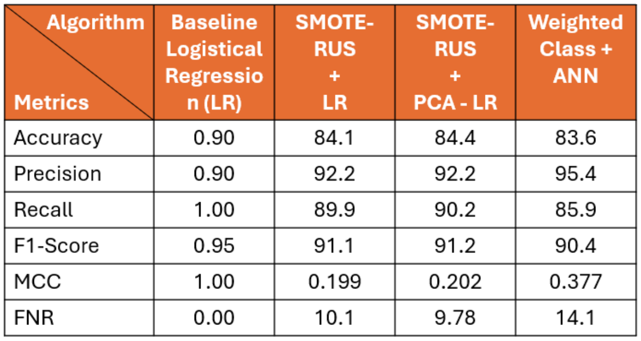
Stacking Classifiers

Stacking ensemble learning utilises various machine learning models and consolidates them into one model. This is done in two tiers - the base classifiers consisting of multiple machine learning models first make predictions based on the training data set, and the chosen meta-classifier takes all the predictions and aggregates them into a single result (Kumar et al., 2022). Several studies have explored the use of ensemble learning with various machine learning methods such as NNs, SVMs and LR (Mohammed et al., 2021; Kumar et al., 2022; Safdar et al., 2022). Notably, Safdar et al. (2022) proposed a model utilising SVM and LR for breast cancer classification. This next section leverages on the findings of this study specifically, which corroborate with the findings of the methods discussed in this report thus far.

The stacking ensemble learning model used in this study firstly leveraged feature engineering by using mean and standard deviation for each feature to capture basic relationships and patterns in the training data (Kiptoon, 2022). Next, PCA (n_components = 50) was applied to reduce the number of features in the dataset. The value of 50 was chosen following trial and error of various values. Next, the principal components were scaled to provide the datasets used for training and testing the model.

Base classifiers were then defined using the models discussed in this report thus far - support vector machine, random forest, gradient boosting and XGBoost. Class imbalances were compensated by adding a balanced class weight parameter to each model. The predictions from these base classifiers were then fed into LR and trained using the training set before being used to make predictions with the test set.

Results and Discussion



Algorithm Metrics	Baseline Logistical Regression (LR)	SMOTE-RUS + LR	SMOTE-RUS + PCA - LR	Weighted Class + ANN
Accuracy	0.90	84.1	84.4	83.6
Precision	0.90	92.2	92.2	95.4
Recall	1.00	89.9	90.2	85.9
F1-Score	0.95	91.1	91.2	90.4
MCC	1.00	0.199	0.202	0.377
FNR	0.00	10.1	9.78	14.1

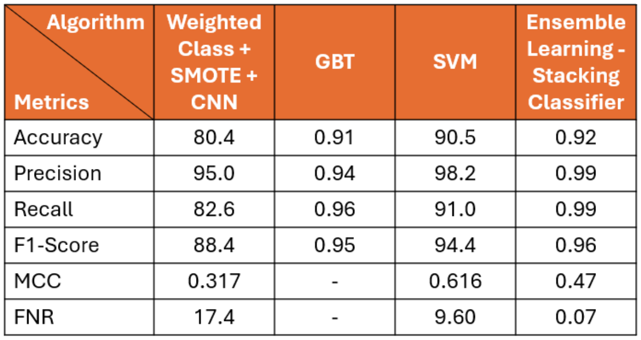
Figure 6: Tabulated results for baseline LR, SMOTE-RUS-LR, SMOTE-RUS-PCA-LR and weighted ANN.

The baseline LR produced an accuracy of 0.90, precision of 0.90, recall of 1.0, F1-score of 0.95, FNR and specificity of 0 and MCC of 1.00. This implies that the model was unable to predict healthy samples due to the class imbalance, resulting in poor training to recognise healthy samples compared to cancer samples. Hence, SMOTE-RUS was needed to improve learning for healthy samples.

Training the LR model with SMOTE-RUS data yielded an accuracy of 84.1%, precision of 92.2%, recall of 89.9%, F1-score of 91.1%, specificity of 31.7%, FNR of 10.1% and MCC of 0.199. Although SMOTE-RUS improved the LR's model to recognise healthy samples, a high FNR and low MCC and specificity implied that LR was still relatively poor in recognising healthy samples. Hence, PCA was utilised to select key features within the dataset to improve prediction capabilities.

SMOTE-RUS-PCA-LR had a slight improvement with an accuracy of 84.4%, precision of 92.2%, recall of 90.2%, F1-score of 91.2%, specificity of 31.7%, FNR of 9.78% and MCC of 0.202. This highlights that LR was still limited in detecting healthy samples despite Zhou et al. (2004) showing that LR was effective in similarly high-dimensional biomedical applications to capture relationships between gene expression levels and disease status. SMOTE-RUS-PCA-LR performance was better optimised using L1 regularisation than L2 regularisation, indicating selecting key features was more effective for learning and predicting. The mean cross-validation score of 78.9% indicates that while the model performs well on average, the significant variability in its performance on different subsets of the data demonstrates potential sensitivity to certain distributions in the dataset. Hence, any changes towards the input's distribution would yield varying results from SMOTE-RUS-PCA-LR.

The ANN with tuned hyperparameters produced good results with an accuracy of 83.6%, precision of 95.4% and a specificity of 63.4%. For cancer classification, it yielded a recall of 85.9% and f1 score of 90.4%. The model had a 14.1% FNR and scored 0.377 for MCC. It was noted that feature selection degraded the performance of the neural network NNs are able to understand more complicated patterns in the features, and the drops in the evaluation metrics under Random Forest feature selection supports this. Research by Bergstra and Bengio (2012) showed that, in many cases, only a few hyperparameters have a significant impact on model performance, while the rest may only need to be within a reasonable range. Hyperparameters for the ANN were tuned using random search, which tries sample combinations randomly. This enables it to test a wider variety of combinations without needing to evaluate each one exhaustively, saving on computational resources and time. The optimal hyperparameters: {'neurons_layer3': 64, 'neurons_layer2': 128, 'neurons_layer1': 128, 'learning_rate': 0.0001, 'dropout_rate2': 0.2, 'dropout_rate1': 0.3, ‘epochs’: 50, ‘batch_size’: 64} were obtained.



Algorithm Metrics	Weighted Class + SMOTE + CNN	GBT	SVM	Ensemble Learning - Stacking Classifier
Accuracy	80.4	0.91	90.5	0.92
Precision	95.0	0.94	98.2	0.99
Recall	82.6	0.96	91.0	0.99
F1-Score	88.4	0.95	94.4	0.96
MCC	0.317	-	0.616	0.47
FNR	17.4	-	9.60	0.07

Figure 7: Tabulated results for weighted CNN with SMOTE, GBT, SVM and stacking classifier.

For the CNN model utilising SMOTE and class weights, an accuracy of 80.4%, a precision of 95% and a specificity of 61.0% was obtained. For cancer prediction, the model scored 82.6% for recall and 88.4% for the F1 score. It also had a FNR of 17.4% and a MCC of 0.317. The performance of the CNN was mostly similar to the ANN, except for the specificity metric which quantifies how correctly the model identifies healthy patients.

GBT showed strong performance in detecting cancer cases (class 1), achieving an accuracy of 0.91, precision of 0.94, recall of 0.96, and an F1-score of 0.95. However, its performance on the healthy class was weaker, with a precision of 0.59 and recall of 0.46, resulting in an F1-score of 0.52. This discrepancy indicates a significant limitation, as the model struggles to accurately classify healthy samples, which is crucial for medical diagnostics. While GBT excels in cancer detection, addressing the class imbalance is essential to enhance overall performance and reliability across both classes.

The SVM performed well for this cancer classification task. The model had an accuracy of 90.5%, precision of 98.2%. The low FNR of 8.9% indicates a good classification ability. This corroborates with the recall score of 91.0% and f1 score of 94.4%. The high specificity of 85.4% and MCC of 61.6% represents an effective balance between sensitivity to cancer samples and specificity for healthy cases. This performance likely results from SVM’s capacity to handle high-dimensional data and class separability, making it a strong candidate for further application in similar biomedical contexts. The robustness of SVM in this study points to its potential as a primary model for such imbalanced classification tasks.

The stacking ensemble learning method produced the best performance for this cancer classification task, with an accuracy of 92.4%, a precision of 99.5%. For cancer detection, it produced a f1-score of 95.9% and recall of 92.6%. It produced a high specificity of 85.7% but a moderate MCC of 0.474. False negative rates were also very low at 7%. This is likely due to the ensemble learning method utilising the strengths of the methods discussed in this study, integrating them to produce the best performing result across all models. While the ensemble approach is effective in identifying cancer cases, improving the classification of the healthy class should be a focus for future enhancements.

Conclusion

The primary challenge in this study was to accurately classify an imbalanced dataset of cancerous and healthy samples, where the detection of healthy samples presented significant difficulties due to their minority status in the dataset. Through systematic experimentation with Logistic Regression (LR), Gradient Boosting Trees (GBT), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and an ensembling approach using a stacking classifier, we aimed to improve recognition of the healthy class while maintaining high accuracy for cancer detection. Our best performing model in terms of accuracy - the stacking classifier ensemble learning model, explores a slightly different model from existing literature where various ensemble learning methods are stacked together. This method does not perform as well as the study conducted by Sadfar et al. (2022), and future research could explore how such ensemble learning methods can be further optimized to yield a higher accuracy.

The machine learning models explored in this project show potential in classifying cancer samples but still fall short of human diagnostic capabilities. The considerable presence of false positives and false negatives in the models could result in serious consequences such as a false diagnosis leading to inappropriate treatment methods, which can result in death. Instead of seeking to surpass human expertise, the role of these models should merely be supplementary decision aids to healthcare professionals during the diagnostic process. As AI systems enable increased efficiency in diagnosis and treatment, growing concerns about job displacements among medical professionals become increasingly valid. This calls for more attention in balancing technological processes with workforce stability.

References

Belyadi, H., & Haghighat, A. (2021). Chapter 5 — Supervised learning. In H. Belyadi & A. Haghighat (Eds.), *Machine Learning Guide for Oil and Gas Using Python (pp. 169–295)*. Gulf Professional Publishing. https://doi.org/10.1016/B978-0-12-821929-4.00004-4

Bergstra, J., Ca, J., & Ca, Y. (2012). Random Search for Hyper-Parameter Optimization Yoshua Bengio. *Journal of Machine Learning Research, 13*, 281–305. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf?ref=broutonlab.com

Biemar, F., & Foti, M. (2013). Global progress against cancer—Challenges and opportunities. *Cancer Biology & Medicine, 10*(4), 183. https://doi.org/10.7497/j.issn.2095-3941.2013.04.001

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*(16), 321–357. https://doi.org/10.1613/jair.953

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. *Ecology, 88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Kiptoon, D. (2023, August 18). Understanding Feature Engineering in Machine Learning. *Medium.* https://medium.com/@jdkiptoon/understanding-feature-engineering-in-machine-learning-59fc343a29c9

Kumar, M., Singhal, S., Shekhar, S., Sharma, B., & Srivastava, G. (2022). Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning. *Sustainability, 14* (21). https://doi.org/10.3390/su142113998

Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., & Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Computers in Biology and Medicine, 121, 103761. https://doi.org/10.1016/j.compbiomed.2020.103761

Mohammed, M., Mwambi, H., Mboya, I. B., Elbashir, M. K., & Omolo, B. (2021). A stacking ensemble deep learning approach to cancer type classification based on TCGA data, *11*. https://doi.org/10.1038/s41598-021-95128-x

National Cancer Institute. (2024, May 30). AI and Cancer—NCI (nciglobal,ncienterprise) [cgvArticle]. https://www.cancer.gov/research/infrastructure/artificial-intelligence

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. Machine Learning, 101–121. https://doi.org/10.1016/b978-0-12-815739-8.00006-7

Sadfar, S., Rizwan, M., Gadekallu, T. R., Javed, A. R., Rahmani, M. K. I., Jawad, K., & Bhatia, S. Bio-Imaging-Based Machine Learning Algorithm for Breast Cancer Detection. *Diagnostics (Basel), 12*(5). https://doi.org/10.3390/diagnostics12051134

World Health Organization. (2022, February 3). *Cancer.* https://www.who.int/news-room/fact-sheets/detail/cancer

World Health Organization. (2024, February 1). *Global cancer burden growing, amidst mounting need for services.* https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services

Zhou, X., Liu, K.-Y., & Wong, S. T. C. (2004). Cancer classification and prediction using logistic regression with Bayesian gene selection. Journal of Biomedical Informatics, 37(4), 249–259. https://doi.org/10.1016/j.jbi.2004.08.002