IT1244 Project Report: Cancer Detection Dataset
Team 28

Oh Rachel
Cho Yu Xin
Tan Xin Yu
Hannah Ng Xiao Han

Introduction

Cancer remains one of the leading causes of death worldwide, and early detection is critical for improving survival rates. One promising approach for early detection is analysing DNA fragment lengths, which vary between healthy and cancerous states due to biological differences in DNA fragmentation. However, manually analysing this high-dimensional data is time-intensive and error-prone. Leveraging machine learning (ML) for this task could significantly reduce processing time and aid in accurate classification, benefiting medical diagnostics. This project aims to classify samples as "healthy" or "early-stage cancer" based on frequency of DNA fragment lengths, utilising ML techniques to create an efficient, accurate predictive model. This project's relevance lies in its potential to automate an essential diagnostic process, aligning with our module's objective of applying and expanding on ML techniques for practical applications.

Literature Review

Recent studies have explored ML-based methods for cancer detection. Key areas include:

- **Automating DNA Data Analysis**: Reducing the manual workload on medical professionals and leveraging ML models to identify patterns in DNA data.
- **Improving Classification Accuracy**: Using genetic data with advanced models to differentiate healthy and cancerous samples.
- **Detecting Cancer Early**: Developing models to classify cancer stages accurately, supporting early intervention.

Despite these advances, existing methods face several challenges, such as:

- **Inaccurate classification** due to class imbalance between healthy and cancer samples.
- The need for improved methods that can **handle noisy or imbalanced data** while still offering high prediction accuracy.

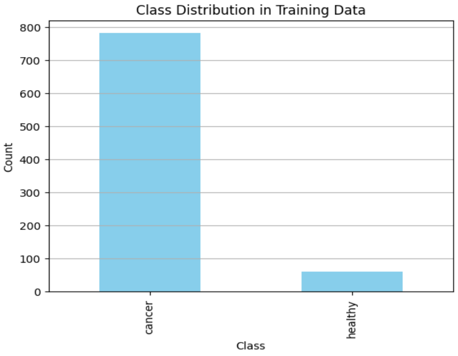
This project aims to address these limitations by implementing ML models that consider class imbalance and high-dimensional feature selection.

Datasets

The dataset comprises maximum normalized frequencies of DNA fragment lengths ranging from 51 to 400 base pairs, with each feature labelled from *length_51* to *length_400* (350 features in total). Each sample is classified as "healthy" or "cancer."

Class Imbalance

A significant class imbalance exists, with a higher number of cancer samples than healthy samples (see Figure 2.1 for distribution). This imbalance may bias models towards the majority class, risking poorer identification of healthy samples. To handle the class imbalance, we apply SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset by generating synthetic samples of the minority class. We also plan to apply class weights, where possible, during training to address this problem. (Li, Adams, and Bellotti 2021)



(Figure 2.1 Class distribution of training data)

Data Preprocessing and Visualization

To prepare the data, we:

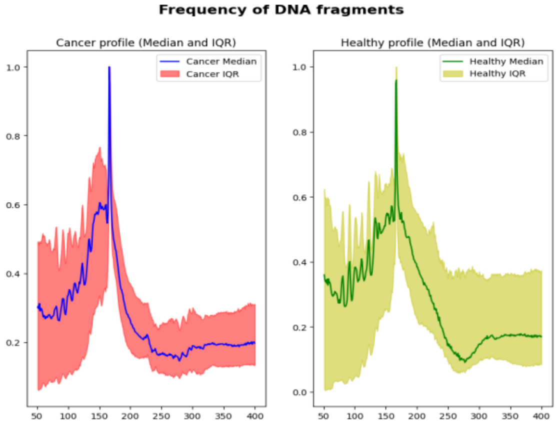
Checked for Missing Values: No missing values were found, ensuring dataset completeness.

Normalized Features: Min-max normalization was applied to scale the feature values, crucial for distance-based algorithms like KNN and its compatible with our other chosen models

Handle Class Imbalance: Using SMOTE to resample the data and save it to a new variable for later use.

Conversion of values: Features converted to float, class labels converted to binary class, where 0 represents healthy and 1 represents cancer.

We visualize our data for both classes independently by computing the median fragment frequency for each class and overlaying the interquartile range of the frequency for each fragment (see Figures 2.2 and 2.3). We observe that the frequencies vary more for some DNA fragments compared to others.



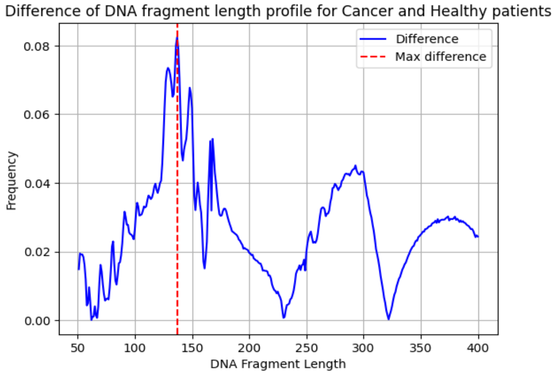
(Figures 2.2 & 2.3 Profile of Cancer and Healthy DNA fragment lengths)

To gain deeper insights, we will compare the two classes by plotting them together on the same graph (refer to appendix for Figure 2.4). This overlay revealed distinct patterns in certain DNA fragment lengths, where cancer and healthy samples diverge significantly.

Methods

Feature Selection:

To reduce the dataset's dimensionality, we calculated the absolute deviation in mean frequencies between healthy and cancer samples for each DNA fragment length (Figure 2.5). This approach allows us to focus on fragment lengths where differences between the two groups are most significant. We selected the top N most deviating features, where N is chosen based on iterating across different feature counts. We want to find N for each model separately to best capture model-specific interactions with the features selected.



(Figure 2.5 Deviation between healthy and cancerous sample)

Classification:

K - Nearest Neighbours

For classification, we applied the k-nearest neighbours (KNN) algorithm. KNN is a simple yet effective classification method that assigns a class label based on the majority class of the nearest neighbours in the feature space. We aim to investigate two parameters: the number of top features to select and the value of k. We employ k-fold cross-validation with varying k values (from 15 to 40) using square root number of samples as a gauge and different feature counts (from 50 to 200) to determine their effects on model performance.

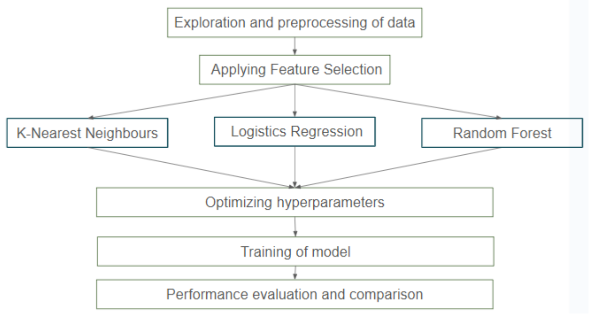
Logistic Regression

Logistic regression is a widely used linear model that estimates the probability of a sample belonging to a particular class by fitting a logistic function to the feature space. We evaluate the model using cross-validation for different feature counts to find the feature count that yields the best results. We use the dataset without SMOTE applied and implement class weights, leveraging the built-in parameter of the sklearn logistic regression function. Finally, we explore how the value of the hyperparameter C affects our model's performance. We use the logarithmic of C since C runs a wide range of values. We employ a 5-fold cross-validation grid search to do this. Additionally, to address the class imbalance in our dataset, we employed the class weighing technique on our original data set, which adjusts the weight of each class based on its frequency. (Li, Adams, and Bellotti 2021). Since this is available, we do not need to use our resampled data for the logistic model.

Random Forest (not covered in IT1244)

Random forest algorithm is an ensemble learning method that combines the predictions of multiple decision trees to improve the accuracy and reduce overfitting. Each tree in the forest contributes to the final prediction based on the majority class output. We aim to investigate the impact of the number of top features selected, and the number of trees in the forest (n_estimators), on model performance using 5-fold cross validation. We choose to evaluate the hyperparameter, n_estimators, as it can affect our model's performance as each tree in a Random Forest is trained to capture patterns in the data to reach a single result. We performed 5-fold cross-validation grid search on different range of values to obtain a suitable range for finding the optimal value for n_estimators.

The following is an overview our methodology:


Exploration and preprocessing of data
Applying Feature Selection
K-Nearest Neighbours
Logistics Regression
Random Forest
Optimizing hyperparameters
Training of model
Performance evaluation and comparison

Results and Discussions

K-Nearest Neighbours

After testing different values of k and different number of top features, we found that the optimal number of neighbours was k=15 and the optimal number of features to use was 80, which resulted in an accuracy of 0.9165 on the train set. Using these parameters, we trained the model and predicted based on the test data to obtain a final accuracy of 0.9169. The confusion matrix (Table 1) shows that the model performs reasonably well in distinguishing between healthy and cancerous samples. However, due to the imbalanced nature of the dataset with more cancer samples, the model tends to have over-inflated scores, with the minority class (negative, 'healthy') being underrepresented.

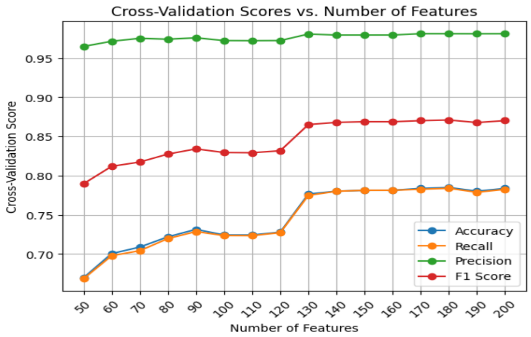
	Predicted Negative	Predicted Positive
Actual Negative	10	31
Actual Positive	3	365

(Table 1: Confusion matrix)

Having applied SMOTE and retrained the model, our model is able to generalize across both classes more effectively. We observed that the selection of the top 100 features and k=15 helped reduce the dimensionality of the problem while retaining the most important information. Without feature selection, the model's performance degraded slightly, likely due to overfitting to noise in the higher-dimensional space.

Logistic Regression

We applied logistic regression to the training data across different top feature counts. We evaluate the model using several metrics: accuracy, precision, recall, and F1-score. Figure 4.1 shows the cross-validation scores for the various number of top features used.



(Figure 4.1: Cross-validation accuracy with varying feature counts)

The greatest increase in all scores occurs when the feature count increases from 120 to 130, indicating that the addition of these 10 features has the most significant impact. Therefore, we selected 130 as the optimal feature count. This is likely due to more relevant features being included in the classification, improving the model's ability to distinguish between classes. Although the scores continue to increase gradually beyond the 130 features, the use of 130 instead of those beyond helps reduce model complexity and prevent overfitting.

The hyperparameter C, of the logistic regression model influences how well the model fits seen data and generalizes to unseen data. We notice that increasing C in our model continues to increase cross-validation accuracy, although theoretically, it should fall after a certain point. For the purpose of this report, we use the default value of C=1.0 for a balance to control overfitting and underfitting.

Finally, we use the top 130 features and C=1.0 to train our logistic regression model.

Random Forest

After testing the different feature counts, we found that using 100 features on the model yielded the best cross-validation accuracy of 0.8938. We also found that the optimal value for the hyperparameter, n_estimators is 103 with a cross-validated accuracy of 0.8957. Based on the classification report (Figure 4.2) for Random Forest, the model struggles to correctly identify healthy samples but is effective in correctly identifying cancer samples even after applying SMOTE which could be due to the model still being slightly biased towards cancer samples.

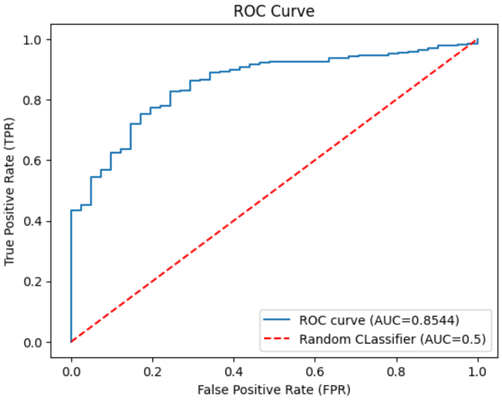
Comparison of results:

Model	Accuracy	Recall	Precision	F1-score
KNN	0.7457	0.7473	0.9615	0.8410
Logistic Regression	0.7873	0.7908	0.9668	0.8700
Random Forest	0.8582	0.7600	0.6900	0.7000

Metric Analysis:
Given the project's focus on cancer detection, we emphasised recall as the most critical measure to capture the maximum number of actual cancer cases. Missing a positive case of cancer can have severe consequences, thus a high recall ensures that most cancer cases are correctly identified, even at the expense of occasional false positives.

Comparing the three models, we note that logistic regression gives a higher recall.

We further analyze the ROC curve and AUC value on the logistic regression model. This analysis acknowledges that while our goal is to maximize true positive rate (recall), we must also ensure that the false positive rate remains at an acceptable level. This is important to avoid subjecting patients who do not have cancer to unnecessary treatments. The figure below shows the results of the ROC curve. (Fig. 4.3)



(Figure 4.3: ROC Curve and AUC value)

Given an AUC value of 0.8544, we can conclude that our model demonstrates a rather low but still acceptable ability to discriminate between true positive rate (TPR) and false positive rate (FPR), as compared to a random classifier. This indicates that the model is reasonably effective in distinguishing between cancerous and non-cancerous cases, achieving a balance.

Comparison with Human Performance:
When comparing the performance of our ML models to human diagnostic capabilities, it's essential to note that while humans excel in contextualising clinical data and recognizing subtle patterns, machine learning models can process vast datasets much faster and often with higher consistency. Models like ours can serve as decision-support tools, offering additional insights that may complement a clinician's expertise. However, achieving performance that consistently outperforms human specialists in diverse real-world scenarios remains a challenge. For our project to be deemed useful, the model does not necessarily have to outperform humans but must provide reliable and consistent results, especially in high-stakes situations like cancer detection.

Societal Impacts:
The implementation of ML models for cancer detection carries significant societal implications:

Privacy: The handling of sensitive medical data necessitates privacy protections to prevent unauthorised access or misuse. It is crucial to ensure that data is anonymized and securely stored.

- **Fairness**: Training on diverse datasets and using fairness metrics can help prevent biases that lead to discriminatory healthcare outcomes.
- **Interpretability:** In healthcare, model transparency is essential for clinician trust. Techniques like feature importance visualization can improve understanding of model decisions

- **Impact on Jobs**: The integration of ML tools may change the roles of medical professionals, emphasising the importance of training and adaptation. While these tools can automate certain tasks, they also enable healthcare providers to focus on complex cases, ultimately enhancing patient care.

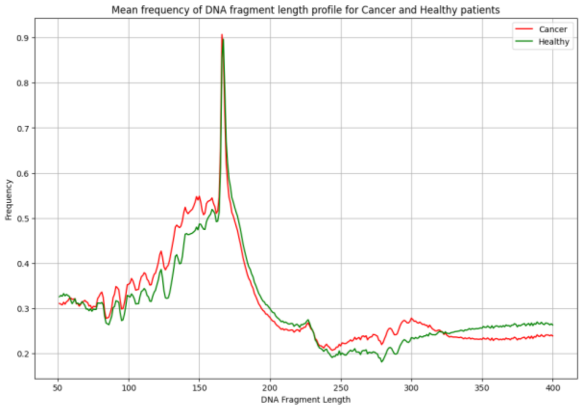
In summary, while our models show promising results, their deployment in real-world clinical settings requires careful consideration of these societal impacts. Striking a balance between technological advancement and ethical considerations is vital to fostering trust and acceptance of AI in healthcare.

References

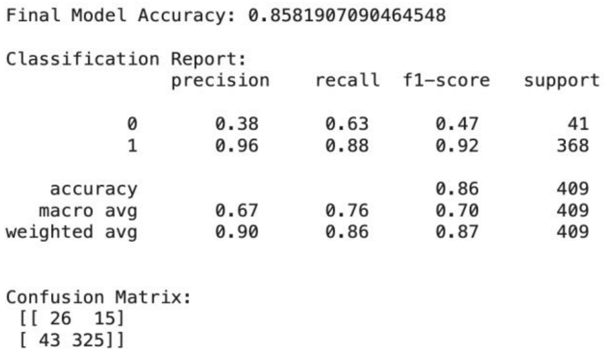
Chhatwal, J., Alagoz, O., Lindstrom, M. J., Kahn, C. E., Jr, Shaffer, K. A., & Burnside, E. S. (2009). A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. *AJR. American journal of roentgenology*, *192*(4), 1117–1127. Retrieved on 26 October 2024, from https://pmc.ncbi.nlm.nih.gov/articles/PMC2661033/

Li, Y., Adams, N., & Bellotti, T. (2021). A Relabeling Approach to Handling the Class Imbalance Problem for Logistic Regression. *Journal of Computational and Graphical Statistics*, *31*(1), 241–253. Retrieved on 26 October 2024, from https://www.tandfonline.com/doi/epdf/10.1080/10618600.2021.1978470?needAccess=true

Appendix



(Figure 2.4 Mean frequency of each fragment)



```
Final Model Accuracy: 0.8581907090464548

Classification Report:
              precision    recall  f1-score   support

           0       0.38      0.63      0.47        41
           1       0.96      0.88      0.92       368

    accuracy                           0.86       409
   macro avg       0.67      0.76      0.70       409
weighted avg       0.90      0.86      0.87       409


Confusion Matrix:
 [[ 26  15]
 [ 43 325]]
```

(Figure 4.2 Classification report for Random Forest)