IT1244 Project: Cancer Detection Dataset

Team 21
Jin Jiarui, Lai Zheyuan, Tan Hui En, Zhuang Yiling

Faculty of Science
National University of Singapore

Introduction

Cancer remains one of the most formidable health challenges of our time, with early detection being crucial for effective treatment and improved survival rates. Traditional methods of cancer screening require substantial medical resources and can often lead to a delay in diagnosis. With the advent of machine learning, there lies potential to revolutionize the way we detect and classify cancerous conditions. This project aims to harness machine learning to accurately distinguish between healthy individuals and those at various stages of cancer using genetic data, thereby optimizing the early detection process.

Dataset

Our project utilizes a comprehensive cancer detection dataset, consisting of approximately 2100 training samples and 1000 test samples. The dataset is structured into 351 columns, with the first 350 columns representing features and the last column indicating the class label. The features are derived from the max normalized frequency of DNA fragment lengths, ranging from `length_51` to `length_400`. These features encapsulate critical genetic information pertinent to the identification of cancerous conditions. The last column of the dataset, `class_label`, is a categorical variable that indicates the health status of the subject: healthy, screening stage, early stage, mid stage, or late stage cancer.

Pre-process the Data

To enhance the quality of data fed into our machine learning models, the following preprocessing measures were adopted:

- **Data Cleaning**: We rigorously cleaned the dataset to handle missing values. This process is pivotal to eliminate potential sources of bias and variance in the model that could stem from noisy or incomplete data.
- **Feature Scaling**: To neutralize the impact of differing ranges of DNA fragment frequencies, we applied feature scaling. This step is critical in ensuring that each feature contributes equally to the analytical process, preventing any single feature from disproportionately influencing the model's predictions due to its scale.

Appropriate Metrics

Since the dataset is **imbalanced** with unequal class distributions (more healthy instances compared to cancer instances), accuracy alone may not be an appropriate metric for evaluating model performance. For imbalanced datasets, it's crucial to consider metrics that account for both positive and negative classes, such as precision, recall, F1-score, and the ROC-AUC score (not covered in IT1244).

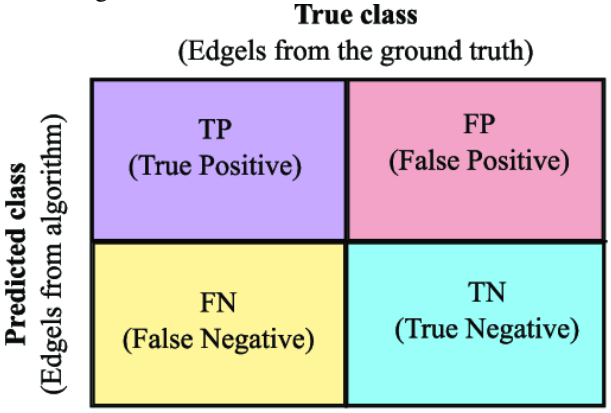
Precision, Recall, and F1-Score are defined as the following:

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F1-Score} = \frac{2 * precision * recall}{precision + recall}$$

For our classification task, TP, FP, TN, and FN are defined according to the **confusion matrix** below:



ROC Curve and ROC-AUC Score

To accurately assess our model's performance in distinguishing between cancer stages, we've adopted the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (ROC-AUC score) as our evaluation metrics. The ROC curve illustrates the trade-off between sensitivity and specificity at various thresholds, while the ROC-AUC score provides a single, aggregate measure of performance across all classification thresh-

olds. These metrics are particularly advantageous for imbalanced datasets, as they reflect a model's ability to classify instances correctly without being misled by the disproportionate class distribution. Although these metrics extend beyond the scope of our course syllabus, they are widely recognized in the field for their effectiveness in evaluating diagnostic models.

Methods

To perform the classification task, we are proposing four methods, including **Logistic Regression**, **Support Vector Machine** (Not covered in IT1244), **Decision Tree** (Not covered in IT1244), and **Artificial Neural Network**.

Logistic Regression

Logistic regression is used to classify instances into two classes: cancer and healthy. Additionally, the code targets specific subgroups within the cancer class, namely screening stage cancer and early stage cancer, to perform more specified classification. The `class_weight='balanced'` parameter is used to address class imbalance during model training. Class imbalance occurs when one class (e.g., cancer) has significantly more instances than another class (e.g., healthy) in the dataset. This can lead to biases in the model's learning process, where the model may prioritize accuracy on the majority class at the expense of the minority class. Therefore, `class_weight='balanced'` is used so the logistic regression algorithm adjusts the weights assigned to each class during training. It assigns higher weights to minority classes and lower weights to majority classes, effectively giving more importance to correctly classifying instances from the minority class.

Results Below are the **Testing Classification Reports** for the required tasks:

Table 1: Healthy vs Cancer

Class	Precision	Recall	F1-Score	Support
0	0.14	0.85	0.24	41
1	0.99	0.78	0.87	993
Accuracy: 0.78 (1034)				
Macro Avg: 0.56 Precision, 0.82 Recall, 0.55 F1-Score				
Weighted Avg: 0.96 Precision, 0.78 Recall, 0.85 F1-Score				

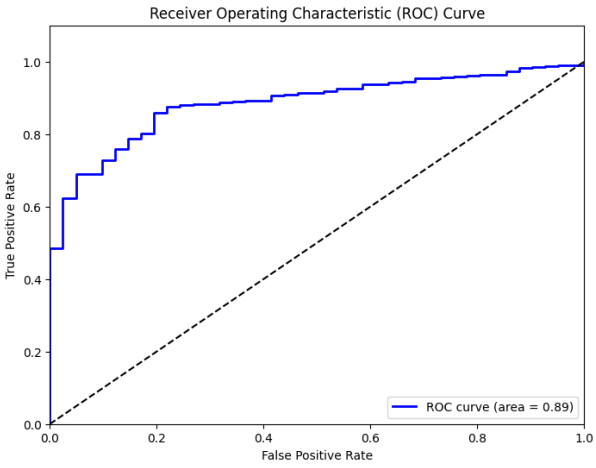
Table 2: Healthy vs Screening Stage Cancer

Class	Precision	Recall	F1-Score	Support
0	0.20	0.59	0.30	41
1	0.89	0.58	0.70	230
Accuracy: 0.58 (271)				
Macro Avg: 0.54 Precision, 0.58 Recall, 0.50 F1-Score				
Weighted Avg: 0.78 Precision, 0.58 Recall, 0.64 F1-Score				

Table 3: Healthy vs Early Stage Cancer

Class	Precision	Recall	F1-Score	Support
0	0.50	0.88	0.64	41
1	0.99	0.90	0.94	368
Accuracy: 0.90 (409)				
Macro Avg: 0.74 Precision, 0.89 Recall, 0.79 F1-Score				
Weighted Avg: 0.94 Precision, 0.90 Recall, 0.91 F1-Score				

Below is the **ROC Curve** (Healthy vs Cancer):



Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful and versatile supervised machine learning algorithm commonly used in classification problems. At its core, SVM seeks to find the optimal hyperplane that best separates the data into its classes. In two-dimensional space, this hyperplane is a line dividing a plane in two parts where in each class lay on either side.

In our case, SVM is utilized to categorize patients based on genetic markers into healthy or various cancer stages. The hyperparameter `kernel` is set to `linear` given our assumption that the data can be linearly separated, a situation where SVM excels. The `class_weight='balanced'` parameter is used to address class imbalance during model training. Furthermore, `probability` setting is enabled, which allows us to generate the ROC Curve and the ROC-AUC Score.

Results Below are the **Testing Classification Reports** for the required tasks:

Table 4: Healthy vs Cancer

Class	Precision	Recall	F1-Score	Support
0	0.13	0.83	0.23	41
1	0.99	0.77	0.87	992
Accuracy: 0.77 (1033)				
Macro Avg: 0.56 Precision, 0.80 Recall, 0.55 F1-Score				
Weighted Avg: 0.96 Precision, 0.77 Recall, 0.84 F1-Score				

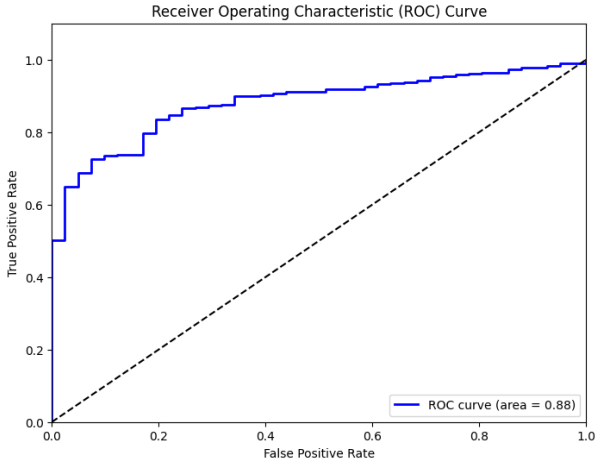
Table 5: Healthy vs Screening Stage Cancer

Class	Precision	Recall	F1-Score	Support
0	0.20	0.56	0.29	41
1	0.88	0.59	0.71	230
Accuracy: 0.59 (271)				
Macro Avg: 0.54 Precision, 0.58 Recall, 0.50 F1-Score				
Weighted Avg: 0.78 Precision, 0.59 Recall, 0.65 F1-Score				

Table 6: Healthy vs Early Stage Cancer

Class	Precision	Recall	F1-Score	Support
0	0.51	0.85	0.64	41
1	0.98	0.91	0.94	368
Accuracy: 0.90 (409)				
Macro Avg: 0.75 Precision, 0.88 Recall, 0.79 F1-Score				
Weighted Avg: 0.94 Precision, 0.90 Recall, 0.91 F1-Score				

Below is the **ROC Curves** (Healthy vs Cancer) for SVM:



Decision Tree

Decision Tree is a widely used supervised machine learning algorithm that utilizes a tree data structure to achieve classification. Decision Tree aims at finding a series of conditions that maximally separates the data into classes. Every condition in the tree is called a **decision node**. At each decision node, one of the features is considered and served as the benchmark for decision making and classification at a time. When a subset of data points is not subdivided further, the node that contains the subset is called a **leaf node** (also known as an **end node**). When a leaf node contains only data points of the same class, that leaf node is said to be *pure*. Data points are classified according to the majority vote.

In our Decision Tree model, the evaluation function chosen is `'gini'` (which stands for Gini impurity), which helps to choose the optimal condition at each decision node by computing the likelihood of misclassification for every possible conditions. We set `max_depth=None` to guarantee all leaf nodes are pure. We also set `random_state=0`. This yields a deterministic algorithm that returns consistent output every time. By restricting the stochastic nature of Decision Tree, it is easier for us to tune our model. The `class_weight` parameter is set to `'balanced'` as with the Logistics Regression model and SVM mentioned previously.

Results Below are the **Testing Classification Reports** for the required tasks:

Table 7: Healthy vs Cancer (Decision Tree)

Class	Precision	Recall	F1-Score	Support
0	0.20	0.24	0.22	41
1	0.97	0.96	0.96	993
Accuracy: 0.93 (1034)				
Macro Avg: 0.58 Precision, 0.60 Recall, 0.59 F1-Score				
Weighted Avg: 0.94 Precision, 0.93 Recall, 0.93 F1-Score				

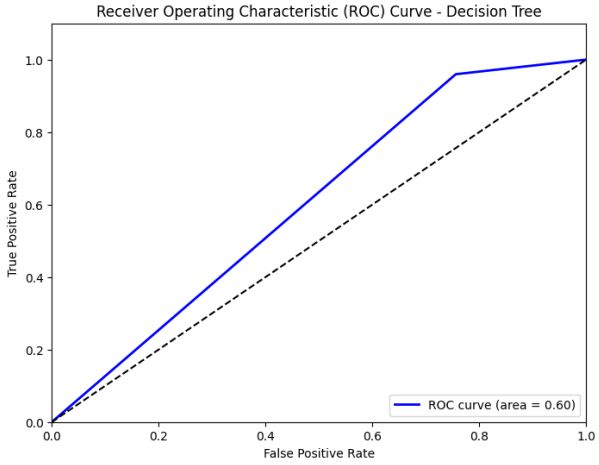
Table 8: Healthy vs Screening Stage Cancer (Decision Tree)

Class	Precision	Recall	F1-Score	Support
0	0.28	0.41	0.33	41
1	0.89	0.81	0.85	230
Accuracy: 0.75 (271)				
Macro Avg: 0.58 Precision, 0.61 Recall, 0.59 F1-Score				
Weighted Avg: 0.79 Precision, 0.75 Recall, 0.77 F1-Score				

Table 9: Healthy vs Early Stage Cancer (Decision Tree)

Class	Precision	Recall	F1-Score	Support
0	0.37	0.54	0.44	41
1	0.95	0.90	0.92	368
Accuracy: 0.86 (409)				
Macro Avg: 0.66 Precision, 0.72 Recall, 0.68 F1-Score				
Weighted Avg: 0.89 Precision, 0.86 Recall, 0.87 F1-Score				

Below is the **ROC Curves** (Healthy vs Cancer) for Decision Tree model:



Artificial Neural Network (ANN)

Artificial neural network (ANN) models are part of supervised machine learning and are capable of solving both classification and regression problems. ANN is able to learn any non-linear function, allowing it to map a complex relationship between input and output.

As we know, the dataset is imbalanced, hence we need to do some pre-processing of data. There are a few methods of handling imbalanced dataset. In our case, we decided to employ synthetic minority over-sampling (SMOTE). In our case, SMOTE has clear advantages over other methods such as the problems of overfitting in the case of over-sampling and reduction in size of dataset in under-sampling. Since we have a small dataset, it is a good choice to use SMOTE over other methods to avoid such issues.

We compared the difference between two methods of feature scaling in the case of ANN, and concluded that min-max scaling performs slightly better than standard scaling, hence our choice of scaling for ANN. We used Adam (Adaptive Moment Estimation) optimizer with binary cross entropy as the preferred loss function for binary classification.

Results Below are the **Testing Classification Reports** for the required tasks and the **ROC Curve** (Healthy vs Cancer):

Table 10: Healthy vs Cancer

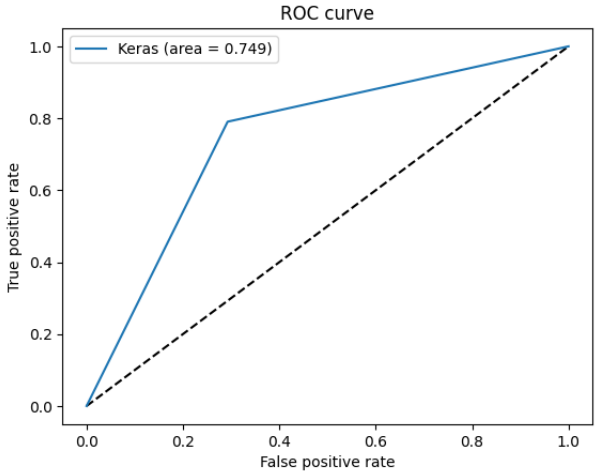
Class	Precision	Recall	F1-Score	Support
0	0.12	0.71	0.21	41
1	0.98	0.79	0.88	993
Accuracy: 0.78 (1034)				
Macro Avg: 0.55 Precision, 0.75 Recall, 0.54 F1-Score				
Weighted Avg: 0.95 Precision, 0.79 Recall, 0.85 F1-Score				

Table 11: Healthy vs Screening Stage Cancer

Class	Precision	Recall	F1-Score	Support
0	0.17	0.68	0.27	41
1	0.88	0.41	0.56	230
Accuracy: 0.45 (271)				
Macro Avg: 0.52 Precision, 0.55 Recall, 0.42 F1-Score				
Weighted Avg: 0.77 Precision, 0.45 Recall, 0.51 F1-Score				

Table 12: Healthy vs Early Stage Cancer

Class	Precision	Recall	F1-Score	Support
0	0.35	0.71	0.47	41
1	0.96	0.85	0.90	368
Accuracy: 0.84 (409)				
Macro Avg: 0.66 Precision, 0.78 Recall, 0.69 F1-Score				
Weighted Avg: 0.90 Precision, 0.84 Recall, 0.86 F1-Score				



Results & Discussions

With careful evaluation, we decided to use Decision Tree (DT) as our final model as it yields the best overall performance for all three classification tasks. The relatively high accuracies attained by our DT model is the primary reason why it is chosen as the final model. Such high accuracies could be explained by the fact that our DT model ensures all leaf nodes are pure.

Nevertheless, our model possesses a series of limitations. From Table 7, we observed that the performance of our model is notably skewed towards class 1. With the assumption that a confirmed cancer diagnose (class 1) as tested **positive** and the opposite (class 0) as tested **negative**, we obtained from our model a much higher sensitivity measure (*0.96*) compared to specificity measure (*0.24*). This means that our model is biased in terms of classification capability, where it is much more accurate in classifying 'cancer' datapoints than 'healthy' datapoints. This is also reflected by the ROC curve, where the small AUC value indicates that DT cannot satisfactorily differentiate healthy and cancerous data entries.

Furthermore, we noticed an anomaly that our DT model yielded a slightly lower accuracy for the Healthy vs Early Stage Cancer classification than the Logistic Regression model and the SVM model (Table 3, 6 and 9). The difference of 0.04 (accuracy score) could be due to random chance or other implicit factors that were not considered in this study. As a possible future area of improvement and extension, it is worth investigating the cause behind such observation.

References

1. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Chapman et Hall/CRC.

2. Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine in machine learning. *Academic Press*, 101-121.

3. Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3), 221–234.

4. Ting, K.M. (2011). Confusion Matrix. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. *Springer*, Boston, MA.