Cancer Detection Using Machine Learning

Team 31
Rikhil Singh, Leong Jing Rou, Nitish Kumar Sahoo, Wang Ruoyu

Abbreviations

DT - Decision Tree, RF - Random Forest, SVM - Support Vector Machine, KNN - K-nearest Neighbours, GB - Gradient Boosting, ANN – Artificial Neural Networks

Introduction

Cancer is a formidable disease that poses great challenges to health globally, being the second leading cause of death worldwide. In Singapore alone, an average of 46 people were diagnosed with cancer and 16 people died of cancer everyday between 2017 and 2021 (SCS, n.d.). The need for effective detection methods is urgent, as early diagnosis is crucial for improving treatment outcomes and reducing mortality rates. This underscores the importance of research into artificial intelligence (AI) and machine learning (ML) methods for cancer detection. With the increasing availability of medical data, AI and ML technologies could offer promising solutions through analysing extensive amounts of data to identify subtle cancer indicators, enabling early detection and intervention before symptoms manifest. Furthermore, ML methods could allow for accurate identifications of different stages in cancer development which aids physicians in prognosis and prescribing the optimal methods of treatment. Therefore, research in this field is pivotal given its potential in improving the diagnosis, treatment, and quality of life of patients, paving the way for a future where cancer could be readily identified and cured.

In this project, we would attempt to develop models to detect cancer using DNA fragments, through the use of a primary and secondary model. The primary model serves to detect cancer and its corresponding stage of development, while the secondary model would serve as a follow-up to differentiate between healthy and screening stage patients.

Related Works

Extensive research has been conducted to identify the most effective and accurate ML models for detecting and predicting various types of cancers. These studies have produced diverse findings regarding the efficacy of different methods. Mokoatle et al. (2023) developed models using raw DNA sequences as input and found that extreme gradient boosting was the best performing classifier for predicting common cancers, with a prediction accuracy of 73%. This is corroborated by Nguyen et al. (2023) who utilised a gradient-boosting tree algorithm to classify liver cancer patients that achieved 81% sensitivity in testing. However, despite good results, the works listed above are limited by their small sample sizes (95 and 110 patient samples used respectively) used for training due to difficulties obtaining clinical data.

On the other hand, Minnoor and Baths (2023) found that RF is the superior method for breast cancer diagnosis using cell attribute data. Su et al. (2022) investigated the diagnosis of colon cancer using gene expression data by first performing feature selection to find characteristic genes associated with cancer. Using the selected genes as predictors, a RF model was able to classify cancer cases according to their stages of development with an average accuracy of 91.5%.

Besides GB and RF, other ML methods commonly used by researchers include DT, SVM, KNN and ANN. As such, we would utilise all the above models in this project, except for ANN due to the lack of large amounts of data for adequate training. We would then evaluate and compare the performances of different types of models through validation and testing.

Dataset

The training data consists of 2193 records of genomic data belonging to patients from one of five different sample classes. These five sample classes of patients are healthy, screening stage, early stage, mid-stage, and late-stage cancer, which make up the set of dependent variables. Each record comprises 350 independent floating values corresponding to the normalised frequencies of DNA fragments with varying lengths. We would conduct a train-validation split of 65-35 for data in the training set for the primary model, and a 75-25 split for the secondary model, which only uses samples classified as healthy and screening stage.

Balancing Data

However, examination of the dataset reveals class imbalances, with only 60 samples of healthy patients compared to 780 samples of early stage cancer patients. This is problematic as it would result in skewed models biased towards the

majority classes. For SVM, the presence of fewer support vectors from the minority class can cause the decision boundary to be less influenced by minority class instances. Underrepresentation of certain classes can also result in poorly defined decision boundaries in DT and RF models.

A variety of methods could be employed to balance the dataset. The first method is under sampling, but this is unfavourable as it further reduces the limited amount of data available. Oversampling methods, like the Synthetic Minority Over-sampling Technique (SMOTE) that creates synthetic samples along line segments between instances of minority classes, could be an option. However, SMOTE is computationally expensive for datasets with large dimensions. Soleymani (2022) has also noted that these methods often fall short in effectiveness compared to the usage of class weights in mitigating class imbalances. Therefore, we choose to utilise class weights to increase the influence of classes with lower representation, such as healthy and late stage cancer categories.

Additionally, we would combine samples of healthy and screening stage patients into the same class. Patients in screening stage are often individuals who are recommended to go through cancer screening as they are facing increased risks of cancer. They may include people of old age, with family histories of cancer or those with abnormal tissue growth that is potentially precancerous. However, the patients at this stage are facing increased risks of cancer but do not necessarily have cancer. Most patients in the class are free from any symptoms, as well as cancerous tissue growth that are severe enough to be classified as cancer at later stages of development (NCI, 2023). As such, we would presume that there would be little differences between the genomic make-up of healthy and screening stage patients and combine these 2 classes to form a '**cancer-free**' class in our primary model.

For the secondary model, due to the extremely large difference in number of samples in healthy (60) and screening stage classes (490), we would use random over-sampling to increase the minority healthy class to 75% the size of the screening stage class.

Feature Selection

Feature selection is often done to extract "informative" features and remove noisy features that are irrelevant and redundant. This serves to simplify the model, which improves its generalisability and reduces model training times (Pudjihartono et al, 2022). This step is especially relevant in this project due to the high dimensionality of data used. The importance of independent variables would be determined using random forest feature ranking. After constructing a RF model, the importance of each feature would be evaluated based on their GINI importance, which refers to the mean decrease in node impurity within each tree. The feature scores from the RF model can be visualised by plotting them against the features. (Fig. 1) More important features can be clearly observed from the graph, as shown through higher corresponding scores.

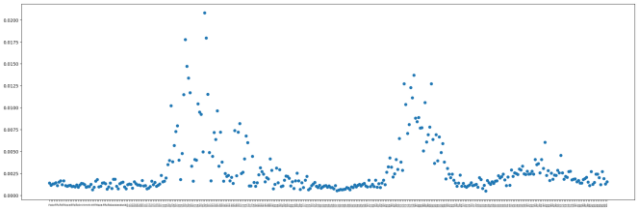

Fig. 1 Feature Scores from RF

Methods

We would model the data using various multiclass classification methods including KNN, SVM, DT, RF and GB classifiers before comparing their performances. Other models such as logistic regression for binary classification problems, and linear regression for continuous predictions would not be appropriate in this project.

Furthermore, despite being a popular model in previous works, we would not utilise ANN due to insufficient data available to train an extensive model. With the amount of data available, the ANN model would be prone to overfitting on the training data and generalise poorly on unseen data. Furthermore, ANN is more difficult to interpret due to the existence of multiple hidden layers, which means that we would not be able to pinpoint the factors behind the algorithm's decision-making process. In a medical context, this is a hindrance as researchers might not be able to analyse what are possible diagnostic indicators that could help identify cancer in a patient.

With regards to the methods used, the Support Vector Machine (SVM) is a supervised learning algorithm that finds the optimal hyperplane in the high-dimensional feature space that best separates the data points into different classes.

Secondly, the Decision Tree (DT) classifier is a supervised machine learning algorithm that recursively splits the dataset into subsets based on the most significant feature at each node. Predictions would be made by traversing the tree from the root node to a leaf node. Each internal node in the tree corresponds to a decision based on a feature, and each leaf node represents a class label.

Next, the Gradient Boosting (GB) classifier is an ensemble learning technique that sequentially combines multiple weak learners, which are typically decision trees to build a strong predictive model. This is done by iteratively fitting new weak learners to the previous model and minimising the errors made by the previous model.

Lastly, the Random Forest (RF) classifier makes use of ensemble learning where it combines the results of multiple decision trees that have been trained on the data and takes

the average of the predictions made across the classifiers. Through this method, specifically known as bagging, the RF Classifier greatly reduces the variance of the decision trees and minimises the effect of the individual bias each tree may have.

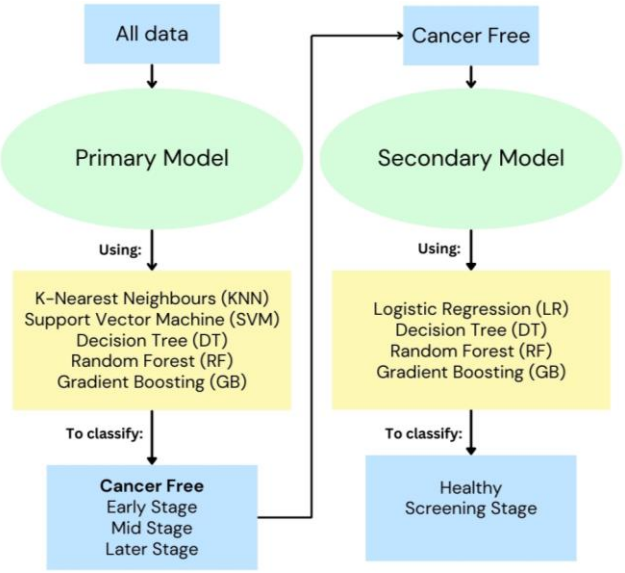

Fig. 2 Structure of model

Our approach to the problem would be divided into 2 segments, as shown in Fig 2 above. Firstly, a primary model with 4 output classes would be developed. The classes would include cancer-free **(**healthy/screening stage), early stage, mid-stage and late stage cancers. A secondary model would then follow, which serves to differentiate healthy samples from screening stage samples. Due to the binary nature of the classes, we would use Logistic Regression (LR) in addition to RF, DT and GB for this part of the analysis. For each model in both segments, we would conduct training on the training set, cross-validation for hyperparameter tuning and performance comparison on the validation set, followed by testing on the test set.

For the primary model, our basis of comparison between different ML methods would be the F1-score generated as a higher F1-score implies low likelihoods of false negatives and false positives in predictions. This is crucial in ensuring the accuracy and reliability of the model in cancer detection. For the secondary model, we would compare models based on precision for more conservative predictions. To illustrate, a false positive test may mean that a patient needs to go through further screening, but a false negative test would suggest that a patient at risk is completely healthy. This would amplify the risk, as the patient may not go for further screening to verify the results which may lead to severe health implications. As such, an ideal model should have high precision and make little false positive predictions.

Results and Discussion

Primary Model

After training of the primary model, each type of model produces F1-scores below when validated against the validation set (Fig. 3). Based on the performance of models shown in Fig. 1, we conclude that RF is the most optimal model as shown by its high F1-scores. The peak of F1-scores for RF at 350 features also indicates that all features should be used for optimal performance.

In general, we observed that tree-based models of RF, DT and GB outperformed KNN and SVM by significant margins. This is due to the presence of large amounts of noise in the high-feature dataset, which resulted in lower accuracies from noise-prone KNN and SVM methods as compared to other noise-resistant tree-based methods (Lehtihet & Åryd, 2021). Both RF and GB performed better than the DT model as they made use of ensemble learning methods to minimise biases that may be present in a single decision tree. By iterating over multiple trees, the models can identify a broader range of patterns in the data, improving their ability to generalise. Finally, we believe that RF performed better than GB due to noise being amplified during gradient boosting. Since GB focuses on incrementally correcting prediction errors by adding up the outcomes of multiple decision trees, it is more prone to overfitting on the noise in the data which reduces its prediction accuracy. RF on the other hand, can handle noisy data effectively through automatic feature selection, which helps to avoid overfitting to noise in the dataset (Lehtihet & Åryd, 2021). Due to these findings, we would only consider the tree-based methods RF, DT and GB in the secondary model.

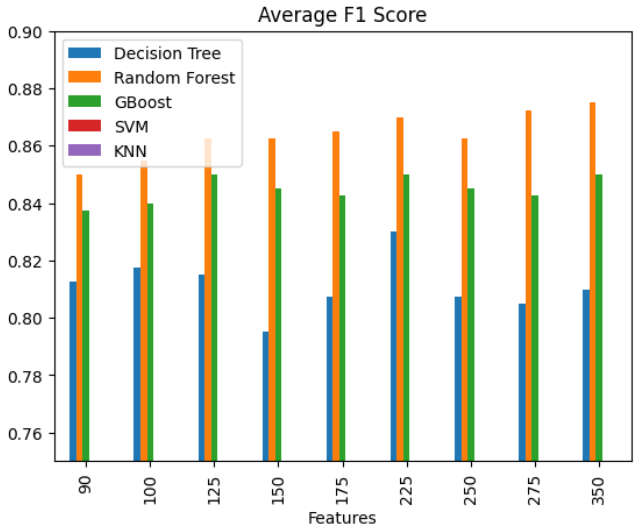

Fig. 3 Average weighted F1-score of models against number of features used. (values for KNN and SVM are too low to be shown)

We then experimented with hyperparameter tuning to further refine the RF model. We first ran GridSearchCV on a set of narrowed down parameters, such as 'n_estimators' and 'criterion', to find their optimal values. K-fold cross validation is then done to assess each hyper-parameterised model's performance on multiple validation sets.

The hyperparameter model performed marginally better compared to the baseline model with default parameters, producing a weighted average F1-score of 0.89 on validation set and 0.70 on the test set compared to 0.88 and 0.70 of the baseline model.

Secondary Model

	LR	RF	DT	GB
Healthy	0.13	0.10	0.18	0.40
Screening Stage	0.91	0.89	0.90	0.90
Weighted Avg.	0.82	0.80	0.82	0.85

Fig 4. Precision metrics for secondary models in validation

As shown in Fig. 4, the performance of all models are relatively consistent, with GB having slight better precision for healthy patients. Therefore, we conclude that GB to be the most optimal secondary model. These differences could be attributed to regularisation embedded in the GB algorithm which improves generalisability and reduce overfitting to noisy data.

Similarly, hyperparameter tuning is done on the selected GB model to find the optimal parameter values. The tuned model is able to produce a higher precision of 0.83 compared to 0.79 of the baseline model.

Final Model

The final model is a combination of the primary and secondary models, which are used together to classify cancer growth within a patient using the sample of their DNA fragments.

As shown in Fig. 5, the model is able to achieve a weighted average F1-score of 0.66 across all classes when tested on the test set. The model does well in predicting late stage cancer but remains relatively limited in identifying healthy samples.

	Precision	Recall	F1-score
Healthy	0.47	0.17	0.25
Screening Stage	0.64	0.51	0.57
Early Stage	0.64	0.72	0.68
Mid Stage	0.62	0.76	0.68
Late Stage	0.89	0.78	0.83
Weighted Avg.	0.67	0.67	0.66

Fig 5. Metrics of final model in testing

While the model had achieved some success, it is far from an idea model which is able to perfectly discriminate between patients with and without cancer. The small sample used for training might have caused models to overfit to noise and fail to capture the important trends present in the data. As such, this method necessitates additional research and validation using more extensive datasets. Further splitting of classification tasks, such as through using multiple layers of models to differentiate similar stages of cancer, may further improve the model. Finally, although such a ML model may offer some insights into cancer detection, it must be used in tandem with other methods of diagnosis to obtain reliable results.

References

In: Common types of cancer. https://www.singaporecancersociety.org.sg/learn-about-cancer/cancer-basics/common-types-of-cancer-in-singapore.html.

Mokoatle M, Marivate V, Mapiye D, et al. 2023. A review and comparative study of Cancer Detection Using Machine Learning: SBERT and Simcse Application. *BMC Bioinformatics*. doi: 10.1186/s12859-023-05235-x

Minnoor M, Baths V. 2023. Diagnosis of breast cancer using random forests. *Procedia Computer Science* 218:429–437. doi: 10.1016/j.procs.2023.01.025

Nguyen V-C, Nguyen TH, Phan TH, et al. 2023. Fragment length profiles of cancer mutations enhance detection of circulating tumor DNA in patients with early-stage hepatocellular carcinoma. *BMC Cancer*. doi: 10.1186/s12885-023-10681-0

Su Y, Tian X, Gao R, et al. 2022. Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. *Computers in Biology and Medicine* 145:105409. doi: 10.1016/j.compbiomed.2022.105409

Soleymani A. 2022. Stop using smote to treat class imbalance. take this intuitive approach instead. In: Medium. https://towardsdatascience.com/stop-using-smote-to-treat-class-imbalance-take-this-intuitive-approach-instead-9cb822b8dc45.

Cancer screening overview. In: National Cancer Institute. https://www.cancer.gov/about-cancer/screening/patient-screening-overview-pdq#:~:text=person%20live%20longer.-,Cancer%20screening%20is%20looking%20for%20cancer%20before%20a%20person%20has,may%20have%20grown%20and%20spread.

Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. 2022. A review of feature selection methods for machine learning-based disease risk prediction. Frontiers in Bioinformatics. doi: 10.3389/fbinf.2022.927312

Schwalbe Lehtihet O, Åryd V. 2021. A Comparison of Performance and Noise Resistance of Different Machine Learning Classifiers on Gaussian Clusters. Thesis, School of Electrical Engineering and Computer Science, KTH, Stockholm, Denmark.