Vrije Universiteit Brussel

TAI Project Report

Breast Cancer Prediction

Siyan Luo 0594750

2023-5

Table of content

Introduction

Breast cancer is the most common cancer among females worldwide, which has taken the countless lives. Such that an accurate and swift detection plays an important role in breast cancer treatment.

The X-RAY images obtained from mammography could provide some information on breast cancer diagnosis. However, due to anatomy of the breast, tumors are hard to tell unless they grow big enough. Besides, there are micro-calcifications in the breast, which are tiny white calcium deposits appeared as a natural process of ageing and noticeable on mammogram. The presence of certain groups of micros doesn't indicate breast cancer. But there is a correlation between micros and tumors. In this research, we made the hypothesis that there is a link between individual micros and cancer and classified micros as malignant and benign. Malignant micros present in the neighborhood of a tumor while benign ones do not. The dataset involved in this project extracted properties computed on micros derived from 3D high resolution images [1].

In this research, I processed the breast cancer dataset and constructed 4 machine learning models (Decision Tree, Multi-Layer Perceptron, Bayesian Network, Random Forest) for micro-calcifications prediction and breast cancer prediction separately. The results show that Random Forest has the best performance.

Data and methods

Data Analysis

The dataset consists of 3562 rows and 152 columns with the first column being patient id and last column being patient labels. In other words, there are 150 features in this dataset.

Regarding the balance of the dataset, the number of benign cases is 2020 while the number of malignant cases is 1542. This is a little bit unbalanced, which should be taken into consideration when splitting datasets.

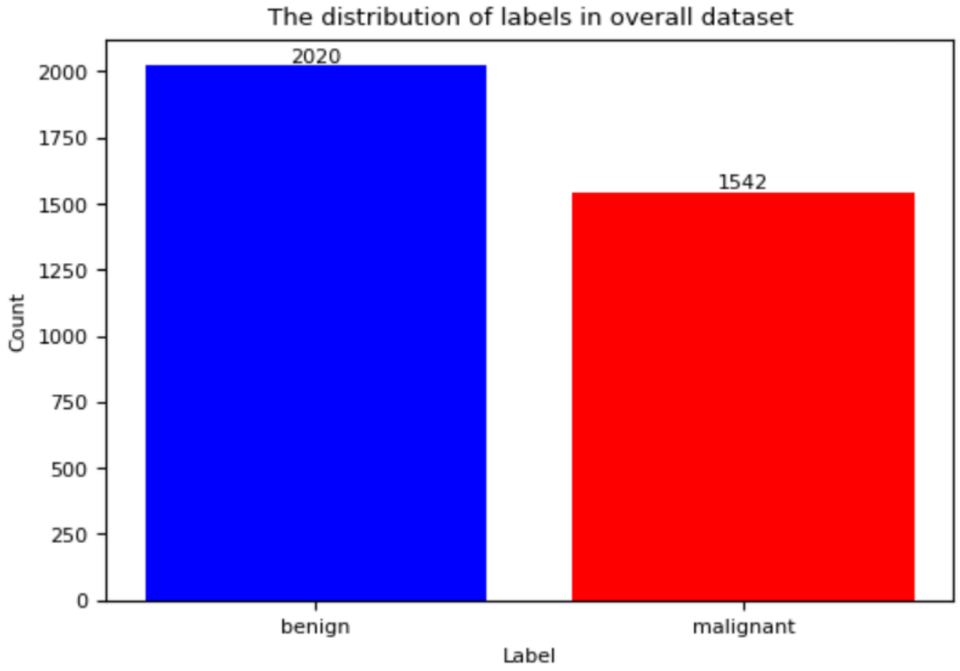


Fig. 1 The number of benign and malignant labels in overall dataset

Every patient in the dataset present multiple micros. There are 96 patients in total. The amounts of micro records per patient varies from 1 to 342. The distribution is showed in fig 2, which is very uneven.

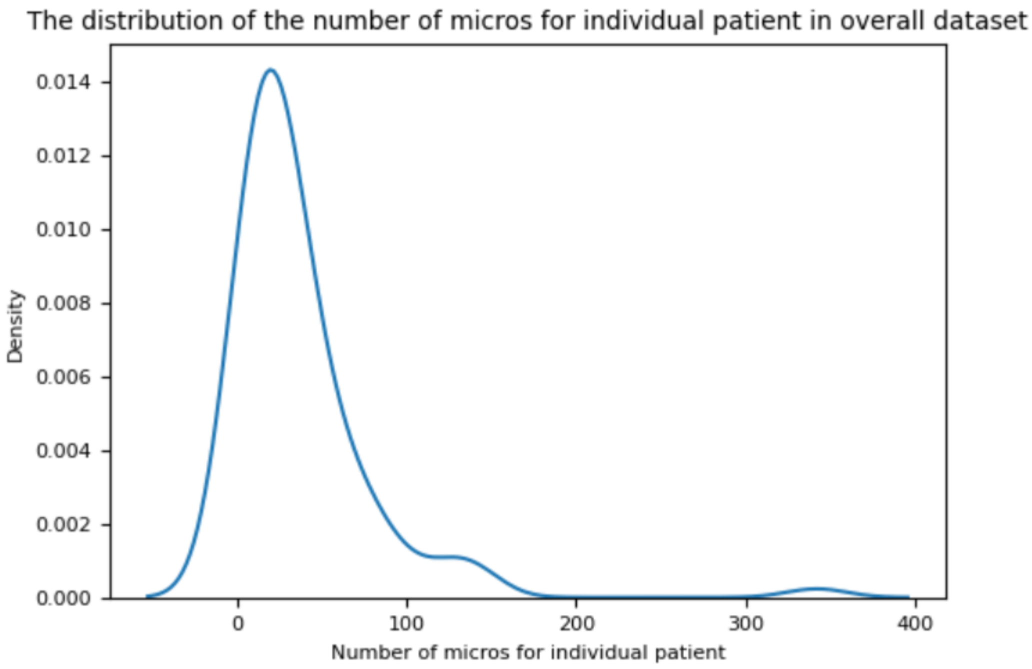


Fig. 2 The distribution of micro amounts per patient in overall dataset

Dataset splitting

To avoid involving the micros of the same patient in both training set and testing set, which results in a biased output, the dataset was split manually by patients id.

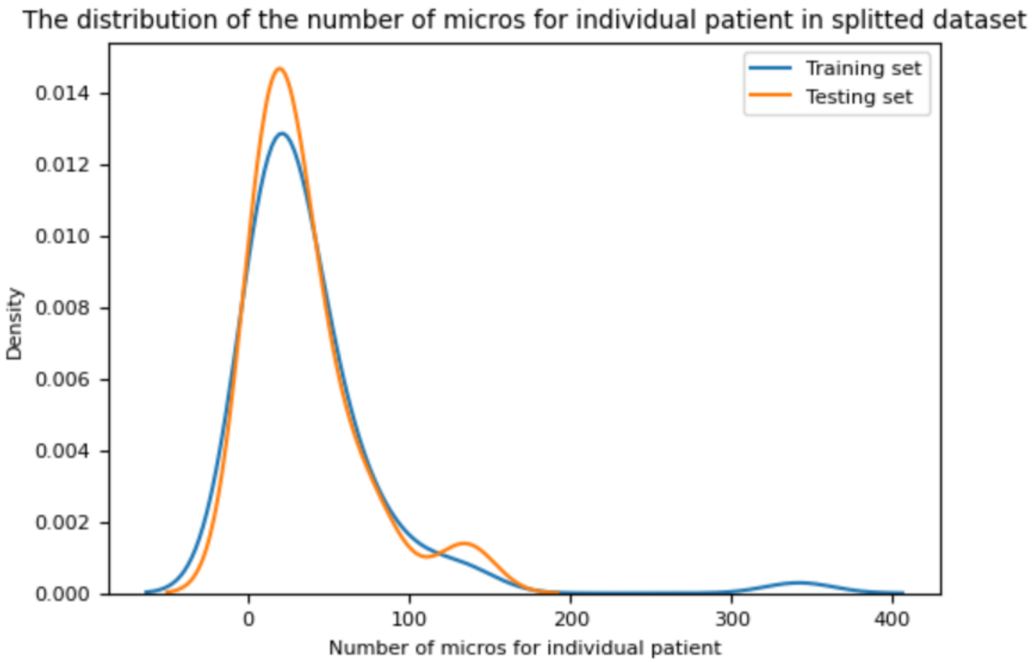


Fig. 3 The distribution of micro amounts for each patient in split datasets

Firstly, patients were sorted by the amounts of micros. For every three patients, the micros of the first and last patient were taken as the training set, the middle one was included in the testing set. In the end, the training set consists of micros records of 2/3 patients and the testing set consists of micros records of 1/3 patients. There are no micro records of the same patient in both sets. The proportion of training set size to testing size is 2434 / 1128. This approach also made the micros distributions in split datasets similar to each other as showed in fig 3.

The analysis of the balance of training dataset showed that the number of benign labels is 1124 and the number of malignant labels is 1310, which makes it very close to a balanced dataset.

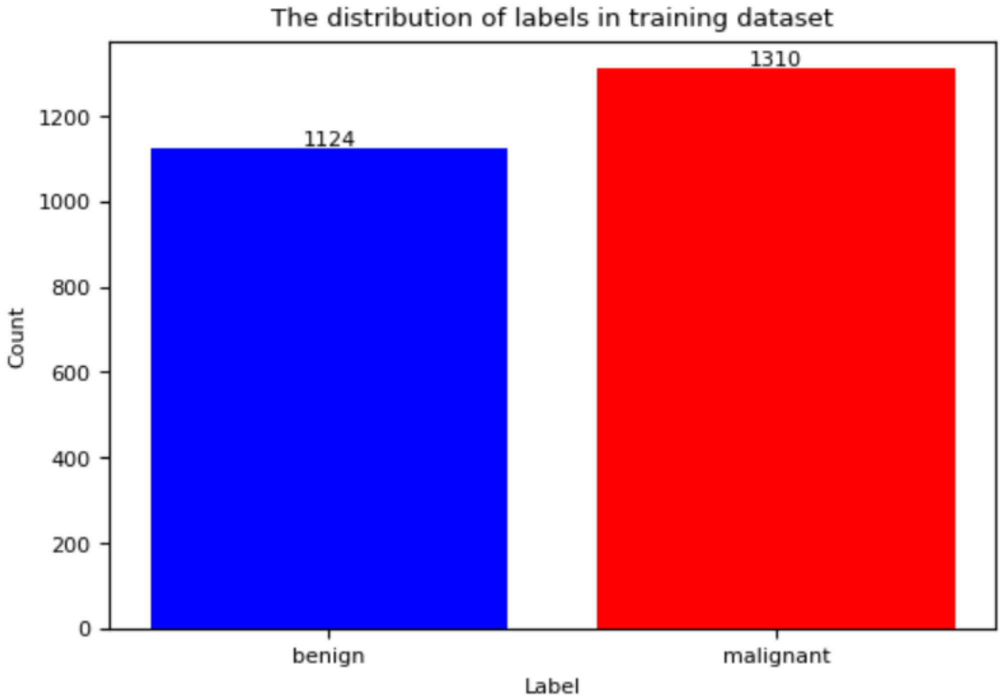


Fig. 4 The number of benign and malignant labels in training dataset

Feature selection

Tor reduce the dimensionality of the datasets, feature selection was applied. By selecting the most informative and discriminative features, we can not only save a lot of time during the model training process but also improve the performance.

Inspired by the similar research done by Redona [2], where she used recursive feature elimination (RFE), recursive feature elimination with cross-validation (RFECV) was implemented here, which is evolved from RFE.

RFE is basically a backward selection of the predictors. This technique begins by building a model on the entire set of predictors and computing an importance score for each predictor. The least important predictors are then removed, the model is re-built, and importance scores are computed again [3]. In practice, RFE is applied with a specified machine learning model and the desired number of features to select.

RFECV adds an additional step of cross-validation to the RFE process. It performs cross-validation on each iteration of feature elimination to evaluate the performance of the model using the selected subset of features. This helps in selecting the optimal number of features that results in the best performance. RFECV provides a ranking of the features based on their elimination order [4].

In this project, RFECV was applied with Random Forest Classifier. It selected 38 features in the end.

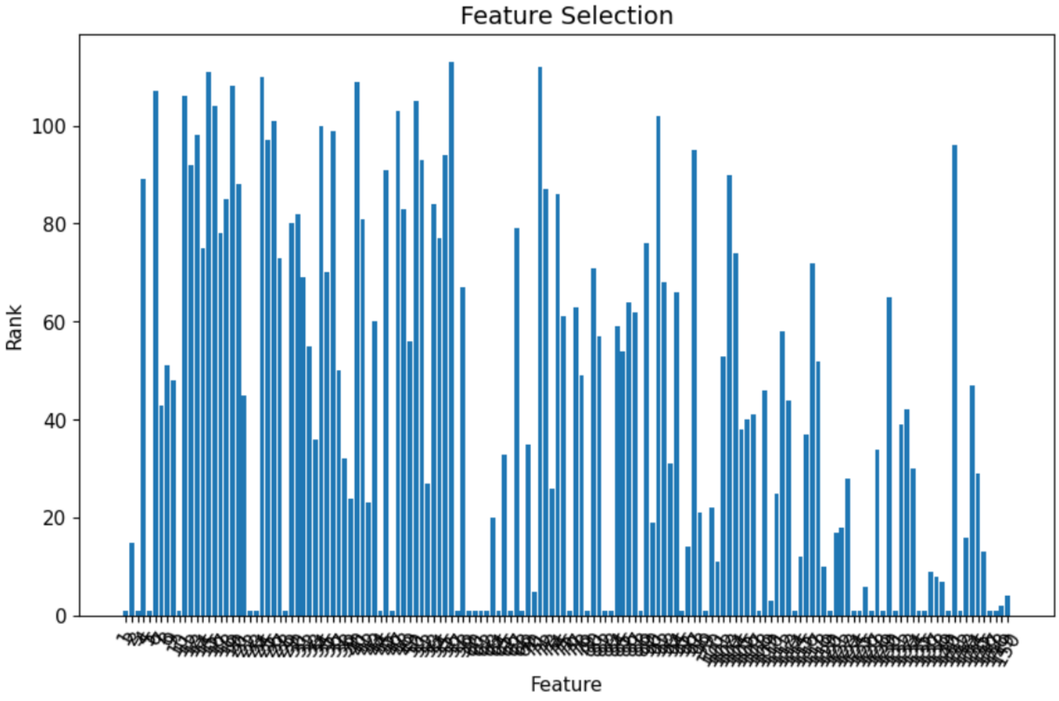


Fig. 5 The ranking of feature selection

Modelling

Task 1: Individual micros classification

With the assumption that all micros per patient have the same label, the patient labels were taken as micro labels. To evaluation different models and make a comparison, 4 binary classifiers as follows were trained: Decision Tree (DT), Bayesian Network (BN), Multi-Layer Perceptron (MLP), and Random Forest (RF).

Since the performance of a model significantly depends on the value of hyper parameters. GridSearchCV was performed in order to determine the optimal values for a given model. By passing predefined hyper parameter values to GridSearchCV function, it tries all the combinations of values and evaluates the model using cross-validation [5].

For model validation, besides evaluating the performance using confusion matrix and ROC curve, I also applied 5-fold cross validation on the model using training set.

Task 2: Patients classification

The result of task 1 showed RF has the best performance. Both training and testing set were input to the trained RF model, predicting the micro labels. Then the predicted micro labels were added as a new feature to the training and testing set for task 2.

The same models and GridSearchCV method were implemented as task 1. After model training, the predictions were just patient labels per micro, which couldn't stand for the patient. To integrate them as one final label for each patient, I calculated the mean value of predicted diseased probabilities of all micros per patient. If the value is large than 0.5, it is reasonable to deduce that the patient can be classified as breast cancer positive and vice versa. In the end, I got the diseased probability, predicted label, and real label for each patient, which were used to construct confusion matrix and ROC curve afterwards for model evaluation.

Evaluation Metrics

For model evaluation, the following metrics were used: Sensitivity, Specificity, Accuracy, Precision, and F1 Score. The formulas are as followed where TP means True Positive, TN means True Negative, FP means False Positive, and FN means False Negative.

$$Sensitivity = \frac{TN}{TN+FP} \qquad (1\text{-}1)$$

$$Specificity = \frac{TN}{TN+FP} \qquad (1\text{-}2)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1\text{-}3)$$

$$Precision = \frac{TP}{TP+FP} \qquad (1\text{-}4)$$

$$F1\, Score = \frac{2\times Precision\times Sensitivity}{Precision+Sensitivity} \qquad (1\text{-}5)$$

I also plotted ROC curve (Receiver Operating Characteristic Curve) and calculated AUC (Area Under the Curve). The y-axis of ROC curve means specificity while the x-axis means FPR (False Positive Rate), which is also 1 – sensitivity. The higher the specificity and lower the FPR are, the steeper the curve is, which suggests better performance. AUC quantifies the area under the ROC curve. An AUC of 1 indicates perfect distinguish between positive and negative instances and 0.5 means no discrimination [6].

Results

For task 1, RF has the best performance with the AUC at 0.91, accuracy at 0.75, sensitivity at 0.89, specificity at 0.72, and F1 score at 0.6. While MLP has the worst performance with the lowest AUC at 0.7, accuracy at 0.40, specificity at 0.28, and F1 score at 0.38 but it has a high sensitivity reaching 0.88.

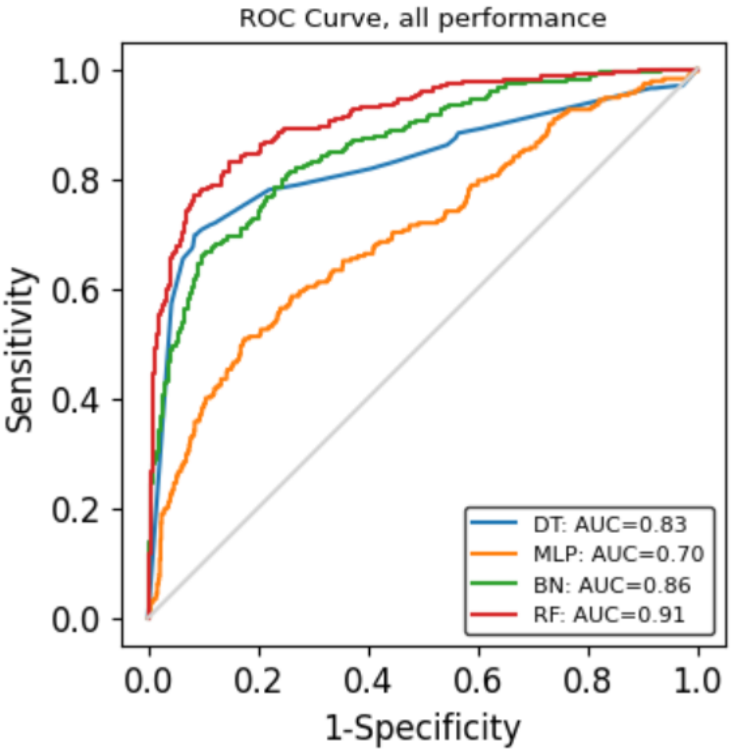


Fig. 6 ROC Curve of all models in Task 1

Table 1 Results of Task 1

	DT	MLP	BN	**RF**
AUC	0.83	0.7	0.86	**0.91**
Accuracy	0.61	0.40	0.66	**0.75**
Sensitivity	0.83	0.88	0.88	**0.89**
Specificity	0.55	0.28	0.61	**0.72**
F1 Score	0.47	0.38	0.52	**0.6**

For task 2, it is also RF that delivered the superior performance with the AUC reaching 0.84, accuracy reaching 0.78, sensitivity reaching 0.82, specificity reaching 0.76, and F1 score reaching 0.72.

However, the other three models also have decent results slightly worse than RF. The partial overlaps in the ROC curves can also indicate that there is similarity in performance between models.

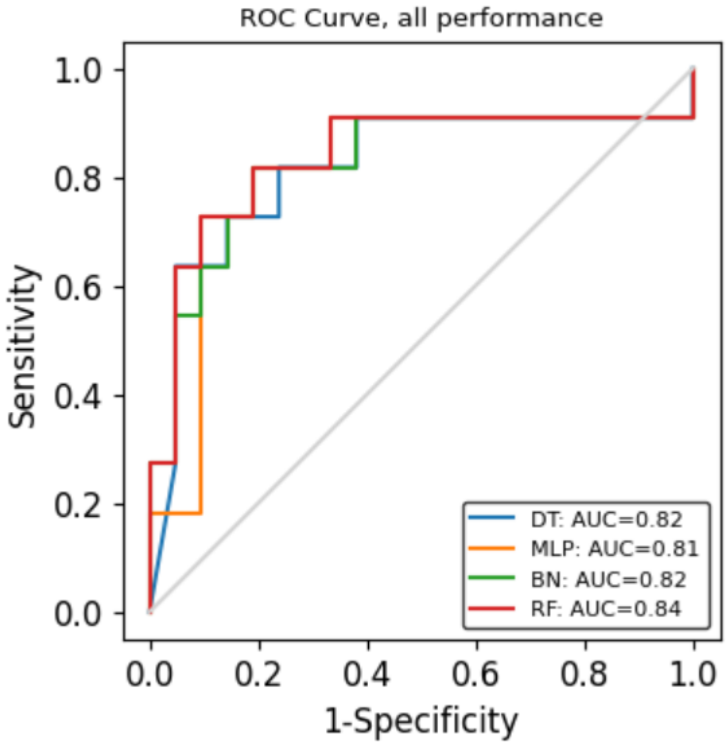


Fig. 7 ROC Curve of all models in Task 2

Table 2 Results of Task 2

	DT	MLP	BN	**RF**
AUC	0.82	0.81	0.82	**0.84**
Accuracy	0.78	0.72	0.78	**0.78**
Sensitivity	0.82	0.82	0.82	**0.82**
Specificity	0.76	0.67	0.76	**0.76**
F1 Score	0.72	0.67	0.72	**0.72**

As it is noticeable that for all models in both tasks, sensitivity is higher than specificity, which is the most obvious in MLP. It means for this dataset, these models are more likely to correctly identify positive cases but might also misclassify some negative instances as positive.

Since RF is an ensemble of multiple decision trees, usually it provides robust and accurate predictions compared to DT. The results verified that. However, in task 2, RF didn't outperform DT much, it might because that DT itself is already able to capture the relevant patterns in the data.

Discussion

Although the final results are decent, there are still limitations and improvement that could be performed.

In feature selection, it's better to apply the same model to RFECV as the following modelling used. But the implementation of REFCV requires specific attributes which are missing in MLP and BN. Both DT and RF were applied to RFECV, it turned out using RF for feature selection has a better performance, so I stuck to this one.

In hyper parameter tuning, since this process is quite time-consuming, the values I set are limited. More optional hyper parameter values could be passed to GridSearchCV for better optimal values searching.

For micro classification modelling, I used 5-fold cross validation to evaluate the performance, which would split the training set into 5 subsets. To have a more accurate outcome, the dataset splitting in cross validation should also be done manually to guarantee non-overlap of the same patient data in different subsets and similar micro distributions.

In task 2, the method to deduce the labels per patient by grouping the labels per micro could be performed differently. What I used is taking the mean predicted diseased probabilities. What could also be tried is that if there is one diseased label in all micros then the patient is disease positive. Another idea is calculating the amounts of positive and negative records per patient, setting a threshold, if the proportion of positive records reaches the threshold, it's reasonable to deduce the label for the patient is positive.

I only used 4 models in this project, to gain more insights, more binary classifiers could be applied. Besides, to narrow down random errors, repeating model training and taking the mean value as the result could give more accurate figures.

Conclusion

When every subject in the dataset presents multiple records, other than the balance of the dataset, we also have to consider the distribution of records amounts per subject. Besides, it is necessary to split the datasets manually because including the data of the same patient in both training set and testing set results in biased outcome.

In this project, for individual micros classification, Random Forest classifier displayed the best performance, and Multi-layer Perceptron is the worst at predicting the pathogenicity of micros.

For patient classification, it is also Random Forest classifier that delivered the superior results. However, Decision Tree, Multi-layer Perceptron, and Bayesian Network can also serve as alternatives since their performance are just slightly worse than Random Forest.

In all trained models, the sensitivity is higher than the specificity, which indicates these models are more biased towards classifying micros as positive more often than negative or the training data is biased. More models and sampling methods could be performed to dive in the underlying reasons. Nevertheless, in the terms of disease prediction, usually a highest possible sensitivity is required to avoid missing out any possible positive cases. Such that it’s acceptable when the specificity is within tolerance.

References

[1] Course Slides: Project Techniques of AI 2022-2023.pdf

[2] Brahimetaj, Redona, et al. "Improved automated early detection of breast cancer based on high resolution 3D micro-CT microcalcification images." *BMC cancer* 22.1 (2022): 1-13.

[3] Max Kuhn, and Kjell Johnson. “Feature Engineering and Selection: A Practical Approach for Predictive Models.” Chapman & Hall/CRC Data Science Series, 2022.

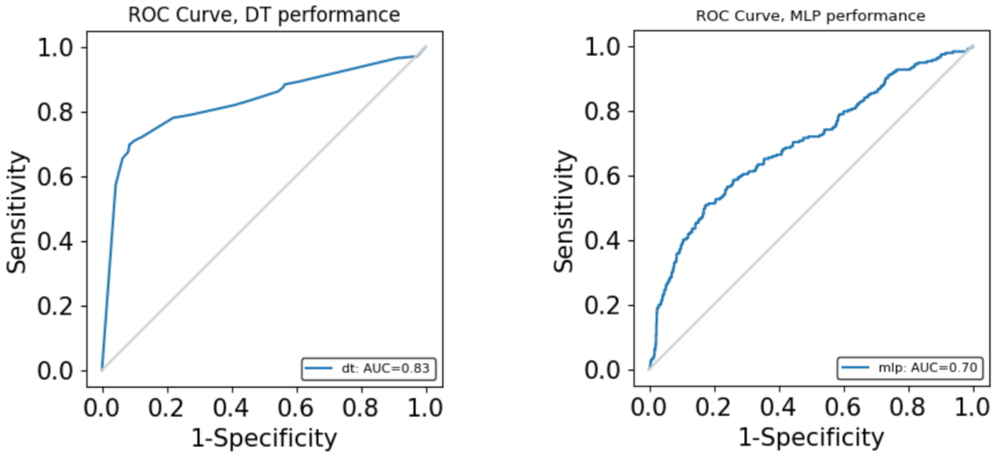
[4] Feature Engineering — Recursive Feature Elimination With Cross-Validation

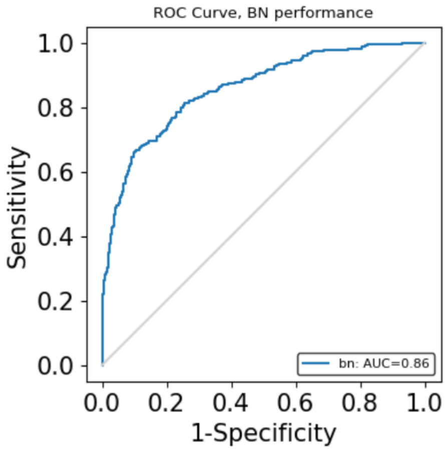
[5] Hyperparameter Tuning with GridSearchCV

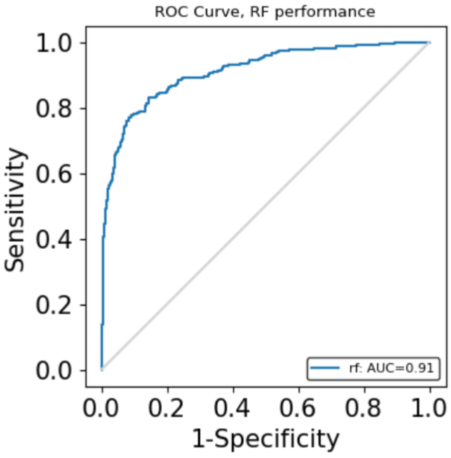
[6] Understanding AUC - ROC Curve

Appendix

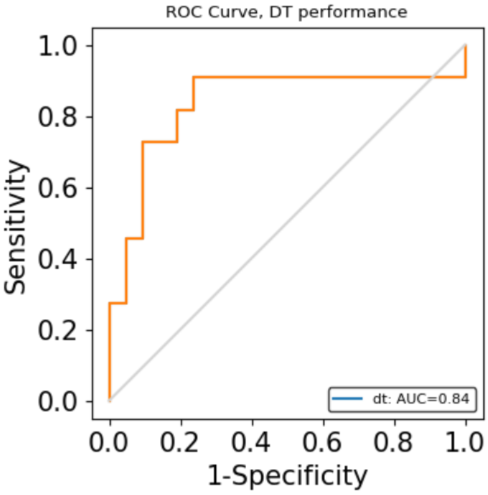
The ROC curve plots of Task 1

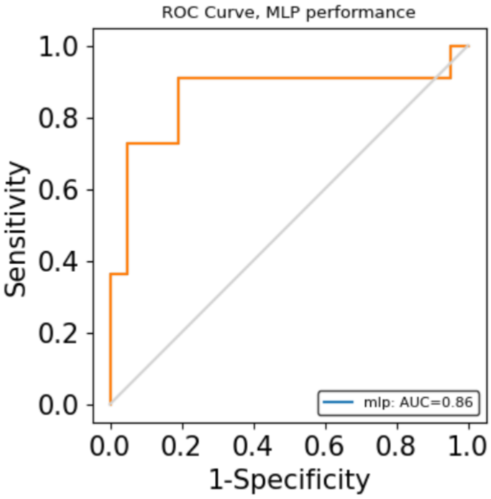


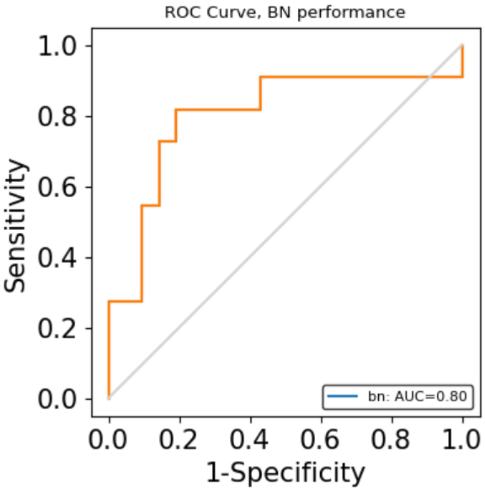


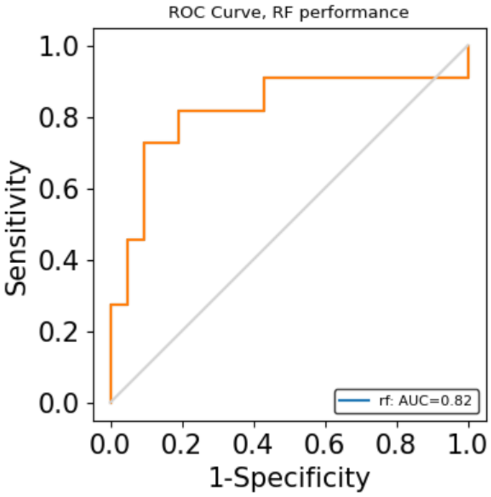


The ROC curve plots of Task 2









Cross validation result of Task 1

	DT	MLP	BN	RF
5-fold cv AUC	0.74	0.68	Null	0.83
5-fold cv Accuracy	0.68	0.64	0.71	0.76
5-fold cv Sensitivity	0.63	0.71	0.69	0.71
5-fold cv Precision	0.74	0.67	0.75	0.81
5-fold cv F1 Score	0.68	0.67	0.72	0.76

Hyper parameter tuning result

Model	Hyper parameter	Task 1	Task 2
DT	class_weight	None	{0:2, 1:1, 2:2}
	max_depth	6	8
	max_features	8	8
MLP	alpha	0.0001	0.0001
	hidden_layer_sizes	50	50
	learning_rate	invscaling	invscaling
BN	learningMethod	TAN	GHC
	prior	Smoothing	Smoothing
	scoringType	AIC	AIC
RF	max_depth	10	6
	max_features	log2	sqrt
	n_estimators	500	200