Prediction for Thyroid Disease using Big Data Analytics

(Preliminary Results by PolyU)

Abstract

In this report, 1479 patient records are collected to analyze the thyroid disease, with 527 healthy samples, 500 thyroid-nodule samples and 452 thyroid-cancer samples respectively. After data pre-processing, we extract 152 features that may influence thyroid cancers. This report firstly shows the density and distribution of each feature with the violin plot, and thereby preliminary select 17 features which are possible to reveal the difference between each classification, with 7 strong-distinctive features and 10 relative-distinctive features. After that, we use four popular supervised learning methods to train the data and predict the thyroid disease classification, which are traditional decision tree method, gradient boosting method, random forest method and bagging method. The results show that random forest method has the highest validation accuracy, which is 76.35% (88.46% in health, 70% in nodule, and 69.57% in cancer).

Data acquisition and processing

The raw data origins from the patient medical records from 2013 to 2018 in Fujian Hospital. The distribution of the samples (patients) is shown in Fig. 1. We first merge the separated data into one file and assign each influence factor a numeric field. For example, 1 represents male, 2 represents female. It helps for further data analysis.

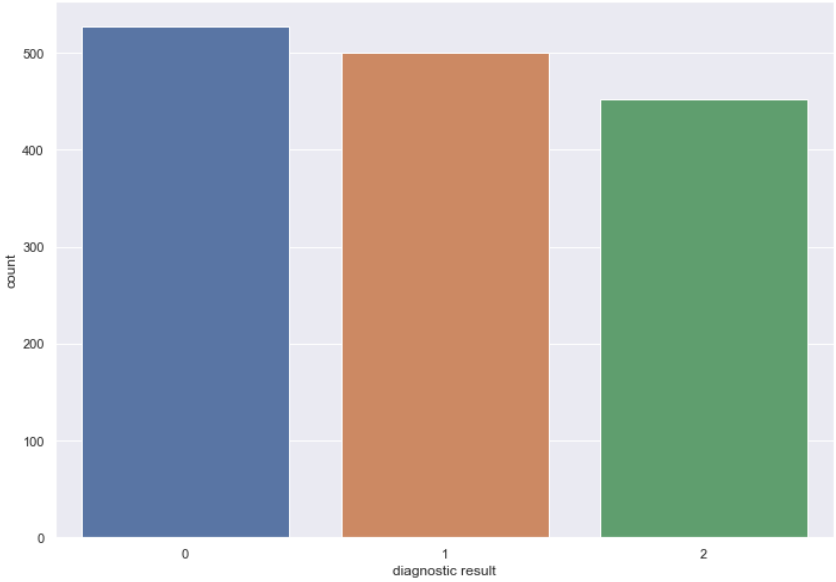

Fig 1. Test samples. Health: 527 samples (2018 year), Nodules: 500 samples (2018 year), Cancer: 452 samples (2013-2018 years).

Exploratory data analysis

Table 1 lists the 152 influence factors we extracted. For the test data, there are mainly two issues:

1) Entire data missing. There is no data at all in some influence factors such as 肌钙蛋白，国际标准化比值 PTINR, 凝血六项 and 甲胎蛋白定量免疫. Therefore, they are excluded from the test data.

2) Partial data missing. Among the remaining valid factors, most of them also have the problem of partial data missing. To solve it, we use the method of **univariate feature imputation** to fill the missing data. The **SimpleImputer** class is used to provide basic strategies for imputing missing values. Missing values can be imputed with a provided constant value, or using the statistics (mean, median or most frequent) of each column in which the missing values are located. This class also allows for different missing values encodings. In this report, we use the mean statistics for imputation.

To show the distribution of each factor value in the three classifications (Health, Nodule, Cancer), A **violin plot** is introduced. It shows both of the density and distribution of a feature in a single plot. Therefore, a total of 152 violin plots are drawn and the one for each factor. Through the analysis, we selected 17 factors which seem to be able to reveal the difference of the three classification. According to the obvious degree, 7 factors have the strong distinction between the classifications, marked with red in Table 1, they are: 甲胎蛋白异质体, 抗 O, 肿瘤相关因子, 血清甲状腺素 T4, TGAB, 甲状腺球蛋白 and 甲状旁腺激素 PTH. The other 10 factors have relatively obvious distinction, marked with yellow in Table 1, they are: 年龄，血红蛋白测定, 血浆粘度 MPAS, 红细胞刚性指数, 最大聚集率 ADP， 白陶土部分凝血活酶时间 AP， 血浆凝血酶时间 TT， AFPL3_AFP，前列腺特异性抗原 and 血清三碘甲状腺原氨酸 T3. The violin plots of all of the above factors are shown in Fig. 2 to Fig. 18.

Table 1 The analyzed influence factor list

性别	尿酮体	BUN_CREA	A3 分钟聚集率 ADP
年龄	亚硝酸盐	尿素氮	A5 分钟聚集率 ADP
血压结论	真菌	胱抑素 C	最大聚集率 ADP
舒张压	清晰度	尿酸	血浆纤维蛋白原测定 FIB
体重指数	白细胞	羟丁酸脱氢酶	白陶土部分凝血活酶时间 AP
脉搏	酸碱度	乳酸脱氢酶	血浆凝血酶原时间 PT
身高	尿潜血	肌酸磷酸激酶	抗凝血系统 ATIII
收缩压	尿比重	肌酸激酶同工酶	血浆凝血酶时间 TT
体重	病理性管型	血葡萄糖	D 二聚体定量 DD
血小板压积	滴虫	糖化血红蛋白	AFPL3_AFP
血小板计数	上皮细胞	二氧化碳结合力	癌胚抗原定量
白细胞计数	白细胞计数_1	钙	甲胎蛋白定量
血小板分布宽度	尿颜色	钠	甲胎蛋白异质体
单核细胞计数	葡萄糖	磷	EBNA1IGA
淋巴细胞计数	AFU 血清岩藻糖苷酶	渗透压	EBEAIGA
红细胞体积分布宽度	白蛋白	阴离子间隙	EBRTAIGG
红细胞计数	白球比例	钾	CYFRA211
平均红细胞血红蛋白量	AST_ALT	氯	NSE
粒细胞百分比	球蛋白	镁	糖基抗原 125
平均血小板体积	总胆汁酸	血同型半胱氨酸	糖基抗原 242
淋巴细胞百分比	谷氨酰转肽酶	超敏 C_反应蛋白	糖基抗原 153
红细胞压积	间接胆红素	C 反应蛋白	EBVCAIGA
单核细胞百分比	总胆红素	淀粉酶	FPSA_TPSA
平均红细胞血红蛋白浓度	谷丙转氨酶	PGI	前列腺特异性抗原
粒细胞计数	谷草转氨酶	PGI_PGII	游离前列腺特异性抗原
血红蛋白测定	碱性磷酸酶	PGII	抗 O
平均红细胞体积	腺苷脱氨酶测定	血浆粘度 MPAS	肿瘤相关因子
嗜酸性粒细胞计数	直接胆红素	低切还原粘度	血沉
嗜碱性粒细胞百分比	总蛋白	高切还原粘度	类风湿因子
嗜碱性粒细胞计数	低密度脂蛋白	中切还原粘度	血清甲状腺素 T4
嗜酸性粒细胞百分比	甘油三脂	红细胞变形指数	血清三碘甲状腺原氨酸 T3
尿胆原	载脂蛋白 A1	红细胞压积 HCT	超敏促甲状腺素 STSH
尿蛋白	载脂 A_B	红细胞聚集指数	游离 T4
粘液丝	总胆固醇	全血高切粘度 200_S	游离 T3
透明管型计数	脂蛋白 A	全血低切粘度 3_S	TGAB
结晶	载脂蛋白 B	红细胞刚性指数	TPOAB
红细胞计数_1	高密度脂蛋白	全血中切粘度 30_S	甲状腺球蛋白
胆红素	肌酐	A1 分钟聚集率 ADP	甲状旁腺激素 P T H

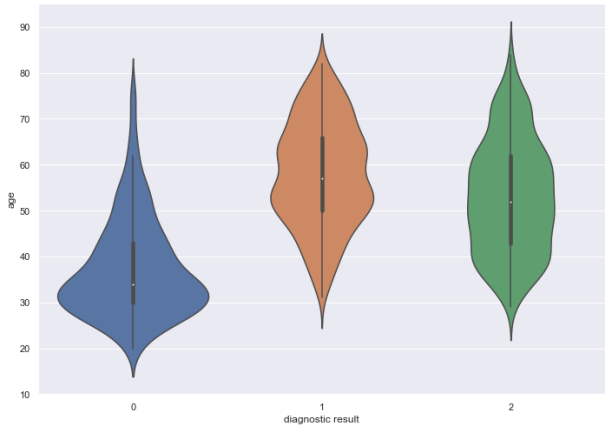

Fig. 2 Influence of age

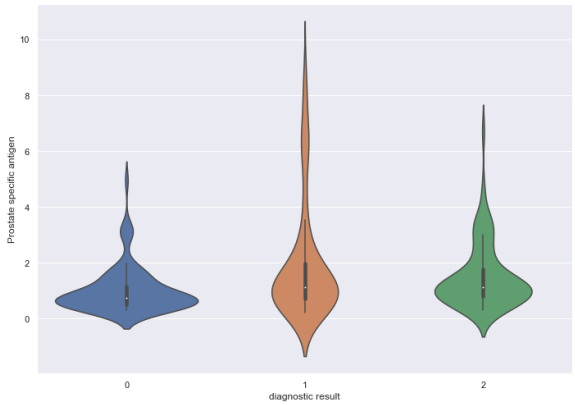

Fig. 3 Influence of 前列腺特异性抗原

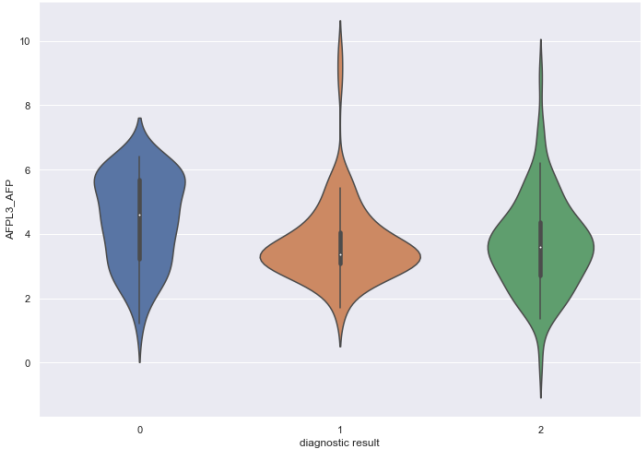

Fig. 4 Influence of AFPL3_AFP

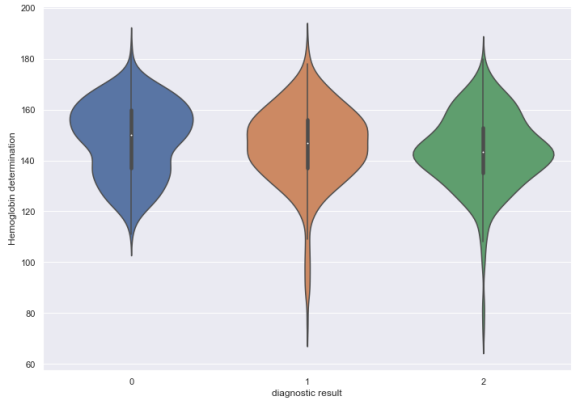

Fig. 5 Influence of 血红蛋白测定

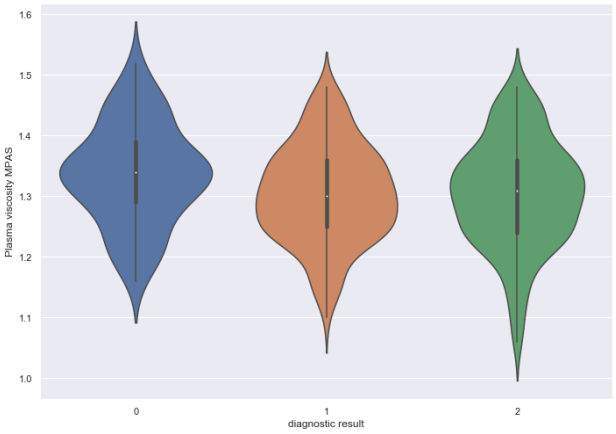

Fig. 6 Influence of 血浆粘度 MPAS

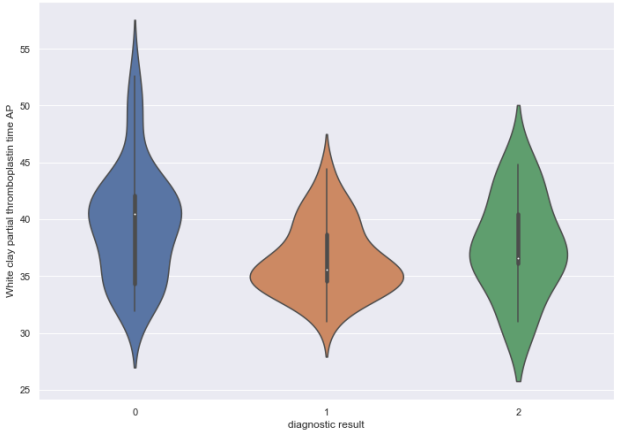

Fig. 7 Influence of 白陶土部份凝血活酶时间 AP

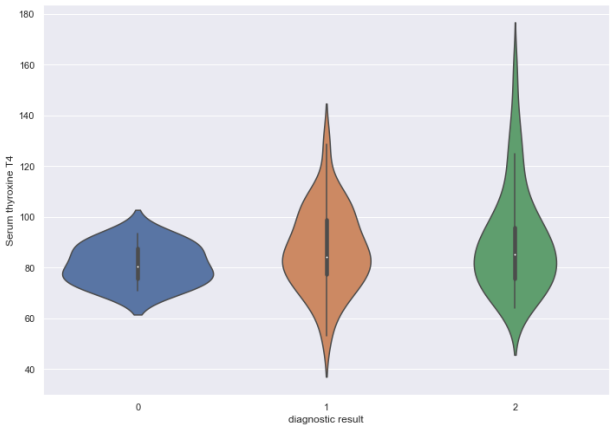

Fig. 8 Influence of 血清甲状腺素 T4

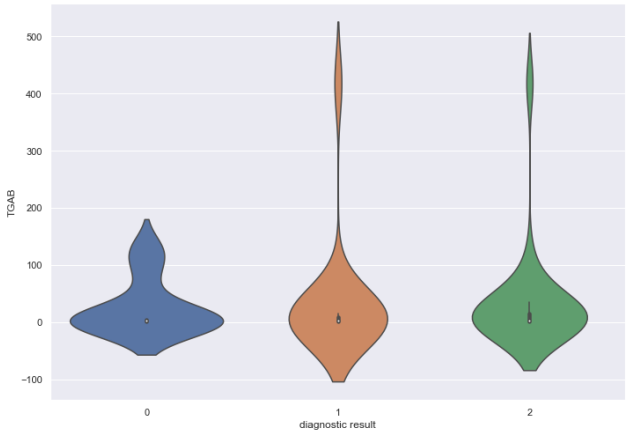

Fig. 9 Influence of TGAB

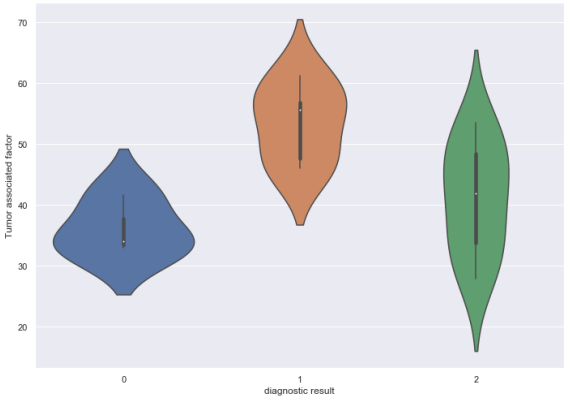

Fig. 10 Influence of 肿瘤相关因子

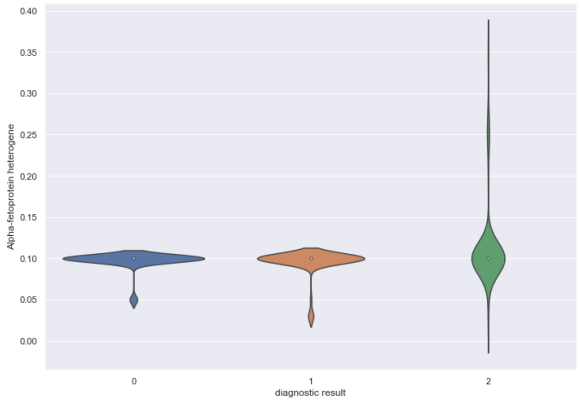

Fig. 11 Influence of 甲胎蛋白异质体

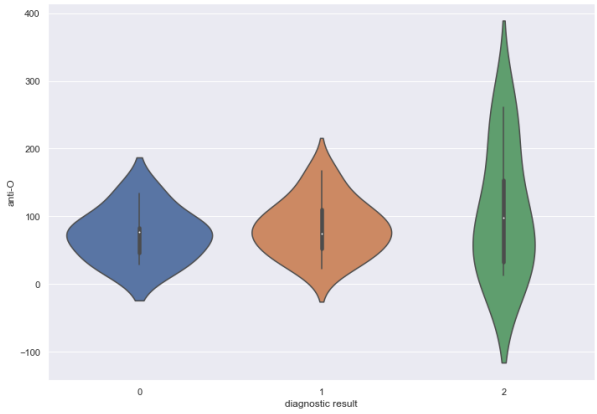

Fig. 12 Influence of 抗 O

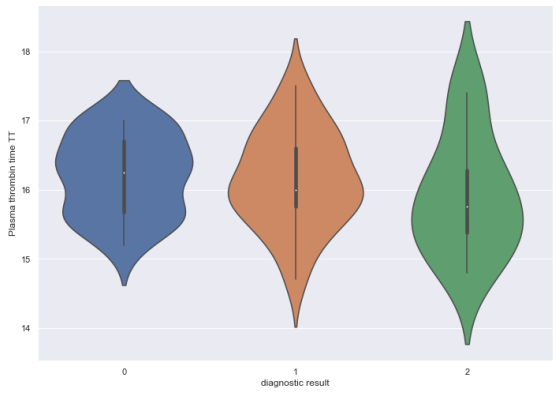

Fig. 13 Influence of 血浆凝血酶时间 TT

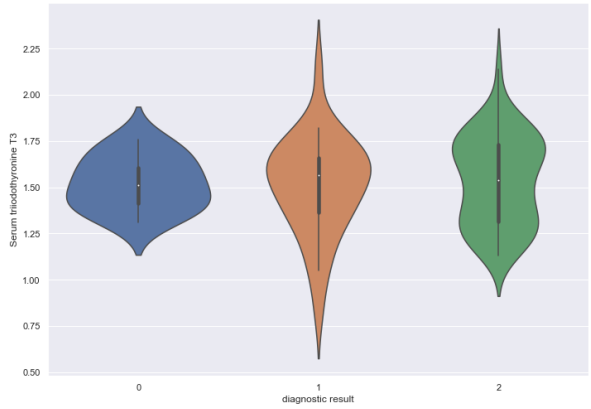

Fig. 14 Influence of 血清三碘甲状腺原氨酸 T3

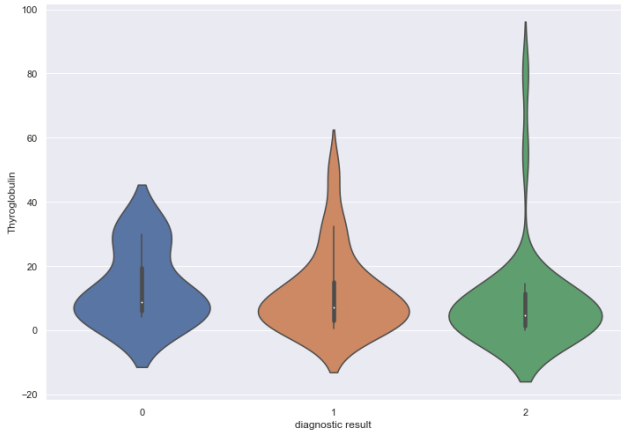

Fig. 15 Influence of 甲状腺球蛋白

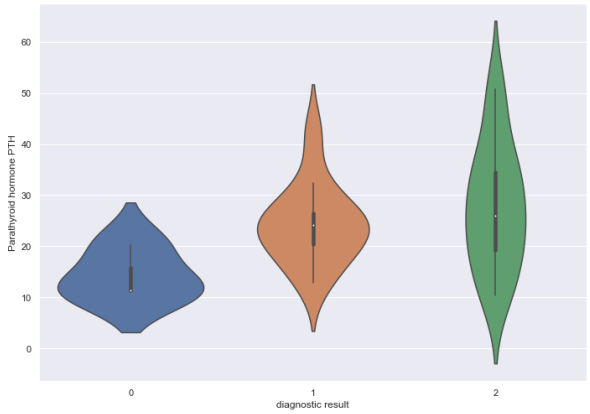

Fig. 16 Influence of 甲状旁腺激素 P T H

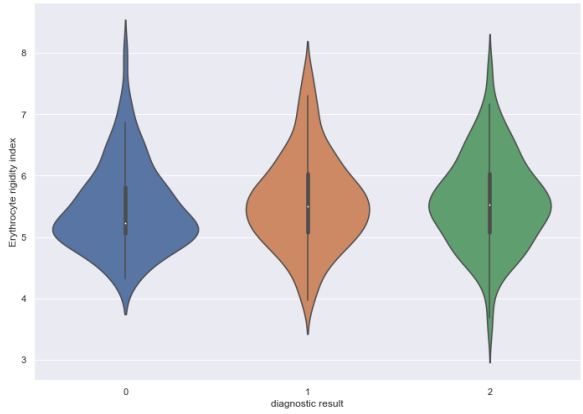

Fig. 17 Influence of 红细胞刚性指数

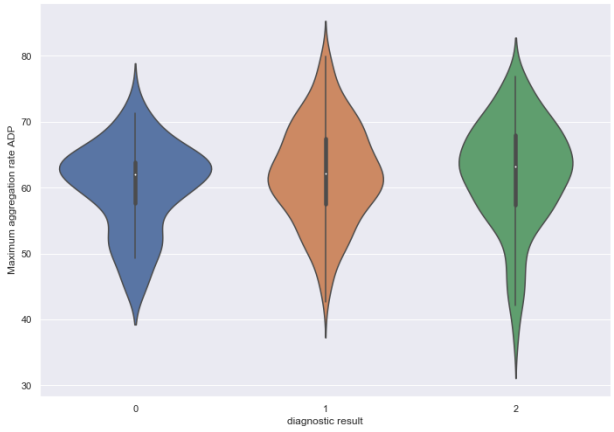

Fig. 18 Influence of 最大聚集率 ADP

Prediction model

There have been many methods for the data classification, the **decision tree** is a popular and traditional decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Ensemble learning, in general, is a model that makes predictions based on a number of different models. By combining individual models, the ensemble model tends to be more flexible (less bias) and less data-sensitive (less variance). **Bagging** method is to training a bunch of individual models in a parallel way. Each model is trained by a random subset of the data. **Random forest** is an ensemble model using bagging as the ensemble method and decision tree as the individual model. **Gradient boosting** is one of the boosting models. Boosting model's key is learning from the previous mistakes and gradient Boosting learns from the mistake, called the residual error directly, rather

than update the weights of data points. In this report, we try to train the data and predict the model separately using the above four methods.

We split the dataset into a training and test set. Before writing our models, we need to define the appropriate error metric. In this case, since it is a classification problem, we could use a **confusion matrix** and use the classification error.

Lastly, **scikit-learn** tool is used to implement the four mentioned methods. Confusion matrix is generated as show in Fig. 19. The results show that **random forest method shows the highest classification accuracy**, with 76.35% (88.46% in health, 70% in nodule, and 69.57% in cancer).

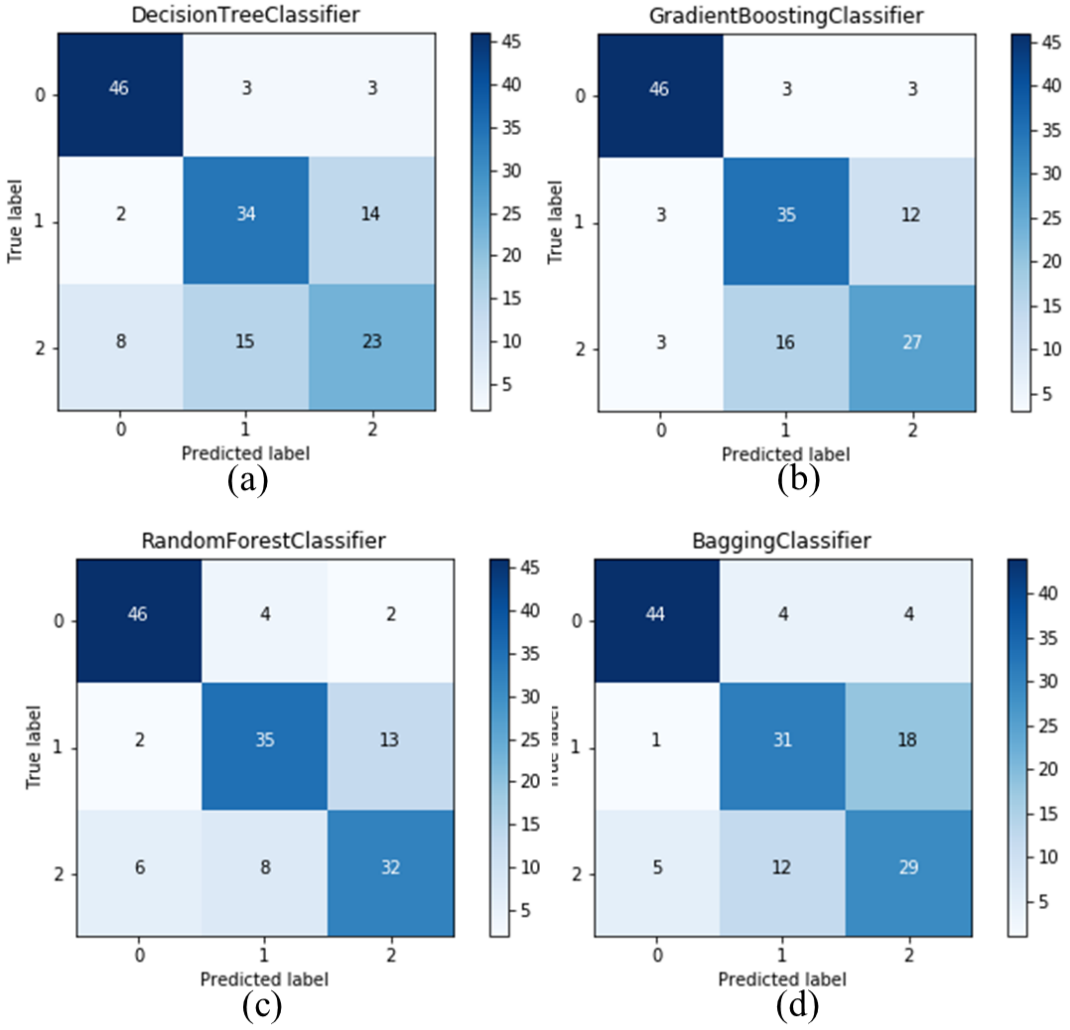

Fig. 19 Comparison results of confusion matrices using different methods (a) Decision Tree Classifier (b) Gradient Boosting Classifier (c)Random Forest Classifier (d) Bagging Classifier. 0 represents health, 1 represents thyroid nodules and 2 represents thyroid cancer.

Results and future work discussion

The current prediction accuracy is about 70%, it is promising to be improved when the training data becomes larger and the more targeted algorithm is developed in the future. Hence, the future work will mainly focus on the development of algorithms and data processing. However, from the view of the raw data source, some suggestions should be addressed here:

1) All the results in this report are based on the current insufficient data especially in some factors. For example, from the violin plot results, although we have found there are some factors have the prediction value. But the data of these factors are extremely little, approximately only 50 valid sample data among 1479 samples. Therefore, the current results are lack of convince. More completed raw data need to be provided.
2) Similar to suggestion 1), the problems of the subjective data seem to be more serious. And some data seems to be contradictory. Therefore, this report only analyzes the objective medical examination data.
3) From the convenience of data pre-processing, it would be better to record the data with numeric field instead of words. For example, 1 represents male, 2 represents female.
4) According to the current literature review, the more popular solutions to predict and analyze the thyroid disease are using the thyroid ultrasound images. The project can be more meaningful if we can combine the images and the current data together into the big database.