Breast Cancer Classification Using The Wisconsin Database Under Different Models

***Abstract*—Breast cancer is now one of the most common cancers among women and the prediction and classification of breast cancer based on data characteristics is extremely important. In this paper, the Wisconsin Breast Cancer Database (WBCD) was selected as the dataset and this paper firstly analyses the data using different methods such as data distribution, statistics, correlation and so on. Secondly, this paper uses different methods for data filtering, feature extraction and dataset partitioning to complete the pre-processing of the dataset. Furthermore, this paper trained five classification models based on WBCD: support vector machines (SVM), random forests (RF), K-Nearest Neighbor (KNN), logistic regression (LG) and Naive Bayes (NB). The classification models are evaluated in terms of accuracy, sensitivity and consistency. Finally, based on the evaluation of the classification models, this paper concludes that the data processing using interpolation of mode, Relief algorithm for feature extraction, and the use of k-folder cross-validation methods to partition the dataset are more suitable for WBCD, and that the RF model achieves a maximum accuracy of 96.35% on WBCD with this data processing method.**

Keywords—Breast Cancer, Data Analysis, Data Processing, Classification

I. Introduction

Breast cancer is now one of the most common cancers and it is important to use appropriate methods to diagnose it early so as to reduce its mortality. In order to differentiate between malignant and benign breast cells, it is now common to use the breast cancer classification to complete its detection.

The analysis and processing of breast cancer data is a necessary process to obtain valid data prior to classifying breast cancer. The breast cancer data used in this paper is the Wisconsin Breast Cancer Database (WBCD), which is currently one of the benchmark databases in the field. In the case of the WBCD, the data analysis allows the relationships and valid patterns of the data to be effectively discovered, thus helping in the subsequent classification [1]. Too many attributes and incorrect data will have a bad impact on the accuracy of the classifier model, so feature selection methods will be effective in filtering out the right features [2].

In breast cancer diagnosis, the main task is to find the appropriate and most accurate classification model, and a variety of classification models and algorithms in machine learning are currently being used in the study of breast cancer data [3]. For example, methods such as support vector machines (SVM), random forests (RF), decision trees (DTs), artificial neural networks (ANN) and Naive Bayes (NB) have been used to make predictions based on the classification of the dataset [4].

This paper uses different data processing methods and analysis models based on the WBCD to classify breast cancer data and to compare the accuracy and stability of the different models.

II. Related Work

Currently, many intelligent methods have been applied to the classification and diagnosis of breast cancer. Based on the WBCD database, a variety of different mathematical methods, traditional algorithms as well as algorithms and data analysis methods such as machine learning have been used to develop unused breast cancer classification models [5].

As in [6], Marcano-Cedeño et al. proposed a neural network (NN) based classifier that used 60% of the data in WBCD as its training set and was trained using a back propagation (BP) method, which ended up with a best accuracy of 99.26% for the classification of the WBCD dataset. Based on the evaluation of the application of ANN with BP in breast cancer classification in the [7], Zaher and Eldeib pre-trained the dataset by combining Levenberg-Marquardt (LM) with ANN, followed immediately by an unsupervised phase, thus constructing a deep belief network neural network (DBN-NN), which resulted in the classification of classification accuracy of 99.68% for WBCD data [8].

Currently, fuzzy logic (FL) is also being used in a variety of medical diagnostic areas. Thani and Kasbe used fuzzy rules to extract nine parameter values from WBCD and used the Mamdani system to construct an expert diagnosis system for breast cancer, resulting in an output for benign or malignant breast cancer [9]. At the same time, the FL-based approach helps to mine association rules. Therefore, an improved fuzzy frequent pattern mining (IFFP) was proposed by Ramesh Dhanaseelan and Jeya Sutha in [10], which analyses the dataset to identify the core factors of breast cancer and then detects whether a person is malignant or benign in WBCD based on the fuzzy association rules formed.

Compared to FL and NN, the construction of classification models for breast cancer data using SVM is relatively simpler and its processing speed is faster. In the [11], Ibrikci, Ustun and Kaya proposed a K-means SVM (K-SVM) algorithm based on conventional SVM applied to WBCD. Based on model performance comparison, the K-SVM algorithm outperformed the conventional algorithm in terms of reliability and accuracy. Ed-daoudy and Maalmi proposed an Association Rules (AR)-based SVM for breast cancer classification, which achieved an accuracy of 98% on WBCD [12].

Although the above methods have provided acceptable models and results, to provide higher accuracy, in [13], Ghiasi and Zendehboudi used RF and Extra Trees (ET) in DTs for WBCD, which were able to achieve 100% accuracy in classifying breast cancer types in WBCD when the RF had 4-10 CARTs.

In summary, the current mainstream classifiers can be classified according to the theoretical principles they use: SVM, DTs, FL, NN, Logistic Regression(LG).

III. Methodology

In this paper, the process of diagnosing breast cancer is divided into two main steps: the analysis and pre-processing of the data and the classification of the data. At each of these different stages, the paper will experiment with different methods so that the performance of different data processing methods and classification models can be compared.

A. Data Analysis

Before classifying the breast cancer data, the paper first explores the data as well as the data analysis so that the quality of the data can be assessed and the next step of data pre-processing can be facilitated.

In the data analysis phase, this paper firstly completes the description of the data and the analysis of the questions in the WBCD. Secondly, the paper analyses the data in WBCD in two stages and using different statistical analysis methods..

1) Data Quality Analysis: In the data quality analysis phase, the main task is to analyse missing values, outliers of the data.

a) Missing Value: In this section, the causes of missing values, the number of missing values and the rate of missing values are analysed. The processing of missing values will be completed in the data pre-processing section.

b) Outlier Analysis: In outlier analysis, outliers can be detected so that they can be processed during data pre-processing. In this section, the outliers are analysed using box line plots and the Local Outlier Factor (LOF) method respectively.

2) Data Features Analysis: In the characterisation of the data, the paper focuses on the distribution of the data, the statistics of the data and the correlation of the data.

a) Statistical Value Analysis: In this section, trends in the data are analysed. The main statistical values analysed are, for example, the median, the mode, the standard deviation of the data, etc.

b) Distribution Analysis: In this section, the overall distribution of the data is visualised and analysed, giving a holistic view of the data situation. In this section, histograms and density curves are used to analyse and compare data distributions respectively.

c) Correlation Analysis: By analysing the correlation of the data variables, further insight into the correlation of the data can be gained and help with subsequent feature extraction. In this paper, heat maps and correlation and significance fit plots are used to demonstrate this respectively. First, the data is visualised using heatmaps, which indicate correlations between variables by colour. Secondly, the number of features can be reduced by finding attributes with high correlation and selecting only one representative from each group of them.

3) Model Data Analysis: To compare the performance of the models, the prediction data of the models are analysed. In addition to comparing the most important accuracy comparisons, this paper analyses data on consistency, sensitivity, specificity and Area under the ROC curve (AUC) for multiple models using the same dataset.

a) Consistency: In order to compare the consistency of the models, this paper evaluates the models using the Kappa coefficient, where a higher value of this coefficient indicates a higher degree of consistency.

b) Sensitivity: The analysis of this data reflects the ability of the model to detect breast cancer patients in diagnostic experiments, with higher values indicating that the model is good at finding the targets it sets.

c) Specificity: Analysis of specificity data is used to analyse the accuracy of the model's prediction of negative classes in the sample

d) AUC : By calculating and analysing the AUC values of each model, it is possible to visualise the classification ability of the classifier, with values between 0.5 and 1 indicating that the classifier is better than a random classifier and vice versa, and when its value is 1 indicating that the classifier is a perfect classifier [14].

B. Pre-Processing Data

The pre-processing of the data allows the data in the breast cancer database to be transformed into the format required for diagnosis and the data to be standardised to ensure accuracy.

In this paper, the pre-processing of data consists of the following 5 parts:

1) Data cleaning: Observe the data in the WBCD and find any columns of data that are not useful for modelling as well as duplicates and remove them.

2) Missing Data Handling: In WBCD, this database has a lack of data, in order to compare the effect of filling in the data with removing the missing data on the classification model, two different approaches are used to deal with this paper.

a) Removing Missing Data: When dealing with missing data, this paper attempts to analyse the performance of the classification model in this case by removing the very small amount of missing data.

b) Imputation: Instead of removing these missing data, this paper also attempts to repair the dataset by replacing the missing data using the mode.

3) Feature Selection: As part of the data reduction, the extraction of data features reduces the number of unimportant features in the database, allowing for improved accuracy in the subsequent classification of the data by the model. In this section, two different methods, Principal Component Analysis (PCA) and Relief, are used for the extraction of features. Because of the different methods of feature extraction, the features used in the model classification and the number of features differ. After the feature extraction has been completed, the results of the above feature selection are further validated by means of a heat map in the data analysis, as the heat map can show the relationship between the feature variables.

a) PCA: PCA is an unsupervised method that transforms the high latitude of the data into fewer dimensions, and it can be used to obtain the principal components by performing a calculation of the covariance between two features. For example, in this paper, the method can transform the data set into a two-dimensional feature subspace [15].

b) Relief: Relief scores features based on how well samples of features that are close to each other are

differentiated between classes, thus reducing unimportant features [16].

4) Data Division: After the features have been extracted and the data filled in, the data is split into two scaled parts, the training set and the validation set, in order to train and validate the different classification models. At the same time, this paper uses two different methods of partitioning the data to ensure that the partitioning of the dataset is as balanced as possible.

a) K-Folder Cross Valiadation: The method splits the dataset into k groups and each group is used as the validation set, while the remaining other groups are used as the training set at the same time. Thus the same model is performed k times so that the classification accuracy will be averaged. This method, as a form of cross-validation, allows each sample data to be used effectively, avoiding over-fitting and with persuasive classification accuracy [17].

b) Stratified Sampling: Stratified sampling allows the data to be divided into strata by characteristics and samples within each stratum to be sampled, thus reducing internal variation in the training as well as validation datasets and increasing the accuracy of the overall metrics.

5) Data Standardization: After all data have been converted into a digital dataset, the data is first normalised and then standardised in order to reduce the variation between the data and to minimise data errors.

a) Normalization: Data normalisation causes the values of the data to all be placed between 0 and 1, the formula for which can be expressed by (1).

$$\frac{X_i}{\sqrt{x_i^2 + y_i^2 + z_i^2}} \tag{1}$$

b) StandardScaler: Data normalisation is the process of making the mean of individual features 0 and the standard deviation 1. The formula is shown in (2), where $mean(x)$ represents the average value of x and $stdev(x)$ represents the standard deviation of x.

$$\frac{x_i - mean(x)}{stdev(x)} \tag{2}$$

C. Classifier Models

After completing the data and processing, different classification models were used for the breast cancer diagnosis classification problem and the performance and classification accuracy under WBCD were compared between the different models to obtain a suitable breast cancer diagnosis model.

Based on the exploration of the relevant literature in Part II, the classification models used in this paper are SVM, k-Nearest Neighbourhood (KNN), NB, RF and LG.

1) SVM: In the face of data with features larger than the observed size, SVM is a classifier for data with a large space, so it is suitable for WBCD.

2) KNN: As a supervised algorithm in machine learning, KNN is suitable for classification as well as regression problems, so it is suitable as a classifier for the classification of breast cancer.

3) LG: Logistic regression is a statistical method suitable for classifying multiple types of data, and it can be used to classify biological as well as medical data, so it will be applied as a classification model in this paper.

4) RF: RF is a classifier for big data classification, which can achieve high accuracy and reliability in its application to breast cancer data classification.

5) NB: As one of the statistical classifiers, Naive Bayes requires less data for training to achieve better results in predictive classification, so it is used as a classifier for comparison in this paper.

IV. Experiment and Results

In this stage, this paper firstly introduces WBCD, secondly analyses and processes the WBCD data, and finally based on the processed WBCD this paper implements a variety of classifier models and obtains the accuracy of each classifier.

A. Data Set

A commonly used benchmark dataset for breast cancer classification, WBCD, is available through the UCI Machine Learning Repository [18]. WBCD consists of 699 samples, which contain 16 missing values. In this dataset, excluding the code numbers, each sample contained nine features and one class attribute, which were clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses [19]. For the class attribute, its division into benign and malignant cases indicates that there are two different possibilities for each sample. For the sample features, each feature is assigned a measure, which is an integer value between 1 and 10, where 1 represents its classification closest to benign and 2 represents the closest to malignant [20]. It is the goal of this dataset to process the data and classify the samples by the characteristics and measurements of the samples in the dataset.

B. Data Analysis of WBCD

In the analysis of WBCD data, experiments were first conducted on the quality and characteristics of their data.

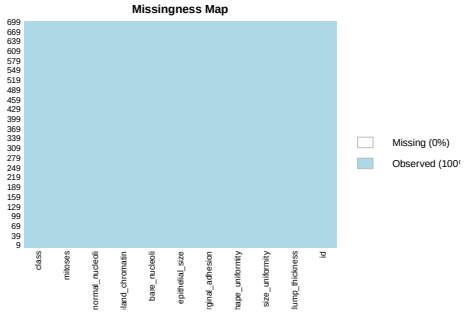

Fig. 1. Missing map of WBCD

	id	bare_nucleoli
24	1057013	?
41	1096800	?
140	1183246	?
146	1184840	?
159	1193683	?
165	1197510	?
236	1241232	?
250	169356	?
276	432809	?
293	563649	?
295	606140	?
298	61634	?
316	704168	?
322	733639	?
412	1238464	?
618	1057067	?

Fig. 2. Missing Value of WBCD

1) Missing Value Detection: In this section, the paper first detects missing data in the data, which can be seen in Fig. 1. As can be seen in Fig. 1, there are no significant missing data and no NULL values in this data.

2) Outlier Detection: As shown in Fig. 2, in the process of outlier detection on the data, this paper further found that there are 16 missing values in the dataset, all of which are in the form of ? in the feature Bare Nuclei. Also, this paper uses box plots and lof for outlier detection (see Fig. 3 and Fig. 4). As can be seen in Fig. 3 and Fig. 4, there are no outliers in class in WBCD, while mitoses has the highest number of outliers.

3) Statistics: In order to better describe the characteristics of the data and to explore the data, a variety of statistics for each variable in WBCD are calculated in this paper. Fig. 5 illustrates the statistics for each variable with its ordinal number, number, mean, standard deviation, median, truncated mean, absolute median, minimum, maximum, value range, skewness, kurtosis and standard error of the mean. In summary, the data can be analysed in a comprehensive manner.

4) Distribution of Data: To explore the data distribution in WBCD more visually, this paper presents it using histograms and density plots respectively, and Fig. 6 shows a comparison of the two.

5) Correlation of Data: The correlation analysis of the data in the WBCD allows for validation of the subsequent feature screening in the WBCD. Fig. 7 shows the correlation analysis using a heat map, while Fig. 8 represents the correlation and significance of the data. The correlation between the uniformity of cell size and the uniformity of cell shape is high, as can be seen in Fig. 7 and Fig. 8.

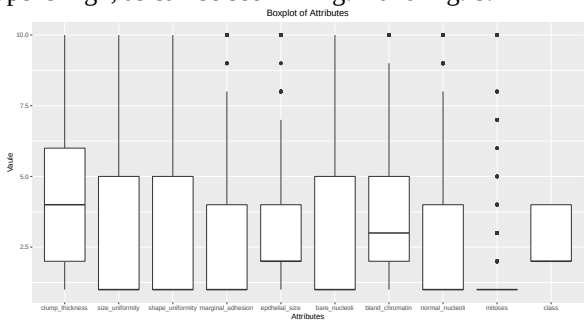

Fig. 3. Box plot of WBCD

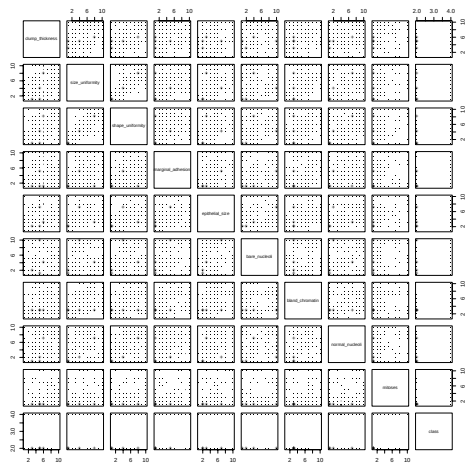

Fig. 4. LOF plot of WBCD

C. Pre-Processing of WBCD

1) Data Cleaning: In this section, firstly, the ID column that have no influence on the data classification model is removed, and secondly, the duplicate data in WBCD are extracted and eliminated. After this step, there are 690 samples left in WBCD, of which 10 variables are left.

2) Missing Value Handling: In this paper, the 16 missing values of Bare Nuclei are treated by interpolating with the mode and removing these samples. In this case, the 16 missing values are each replaced by 1 if the mode method is used. Also, convert all data except class to a numeric type and convert class to a factor type.

3) Feature Extraction: To avoid too many attributes that can mislead the classifier, this paper uses both PCA and Relief feature extraction methods for WBCD. Also, because of the different extraction methods, the final features applied to the classifier and the number of features will be different, allowing this paper to compare the impact of different features on the classifier.

a) PCA: In the PCA analysis of the data, the number of principal components was selected by first visualising the eigenvalues (see Fig. 9). As the first three principal components contain 80% of the contribution, the first three principal components are selected for analysis in this paper. After selecting the number of principal components, the paper analyses the total contribution of each variable to the three principal components, and through this the features are selected. From Fig. 11, it can be seen that mitoses, clump thickness, class, uniformity of cell size, uniformity of cell

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
clump_thickness	1	690	4.43	2.82	4	4.16	2.97	1	10	9	0.59	-0.64	0.11
size_uniformity	2	690	3.13	3.04	1	2.55	0.00	1	10	9	1.23	0.08	0.12
shape_uniformity	3	690	3.20	2.96	1	2.66	0.00	1	10	9	1.16	0.00	0.11
marginal_adhesion	4	690	2.83	2.87	1	2.20	0.00	1	10	9	1.50	0.90	0.11
epithelial_size	5	690	3.21	2.20	2	2.78	0.00	1	10	9	1.71	2.17	0.08
bare_nucleoli	6	690	3.48	3.62	1	2.98	0.00	1	10	9	1.02	-0.73	0.14
bland_chromatin	7	690	3.44	2.44	3	3.10	1.48	1	10	9	1.10	0.17	0.09
normal_nucleoli	8	690	2.89	3.07	1	2.25	0.00	1	10	9	1.40	0.40	0.12
mitoses	9	690	1.59	1.72	1	1.12	0.00	1	10	9	3.53	12.35	0.07
class	10	690	2.69	0.95	2	2.61	0.00	2	4	2	0.65	-1.58	0.04

Fig. 5. Statistics of WBCD

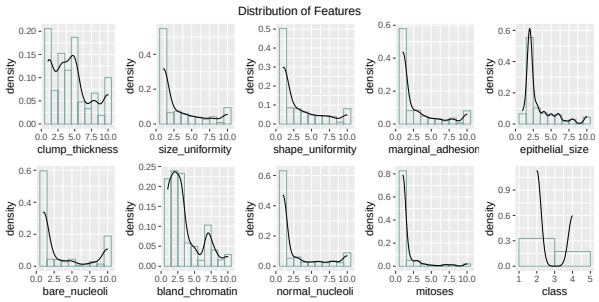

Fig. 6. Distribution of WBCD

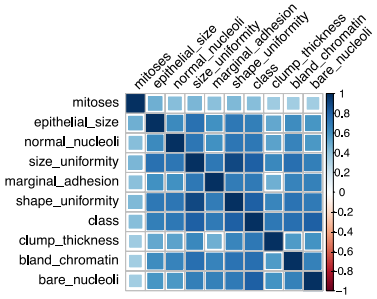

Fig. 7. Heat map of WBCD

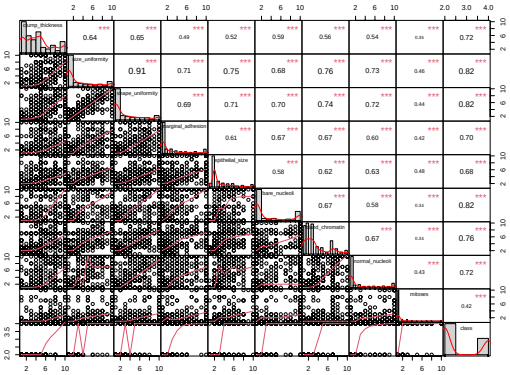

Fig. 8. Correlation plot of WBCD

shape are the five variables with the highest total contribution, so these features were selected to form the new dataset and classifier models were used based on it in this paper.

b) Relief: In the feature extraction using Relief on the data, the contribution values of each variable to the class attribute were obtained. Based on Fig. 11, six features were selected to construct the new dataset, which were bare nuclei, bland chromatin, epithelial cell size, clump thickness, marginal adhesion and class.

4) Data Set Division: In this section, the WBCD dataset is formed into a new dataset after feature filtering, while the k-fold cross-validation and stratified sampling methods are used to divide this dataset into a training set and a test set according to the ratio of 0.7 and 0.3, respectively.

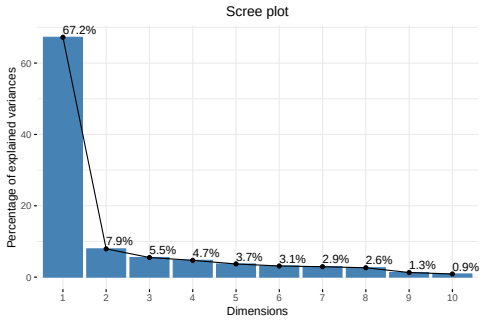

Fig. 9. Scree plot of Features

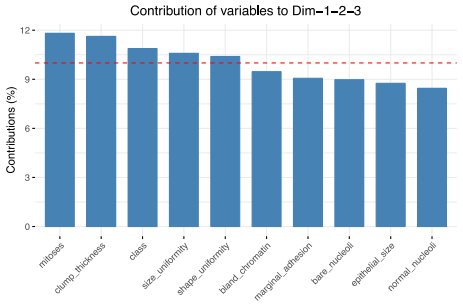

Fig. 10. Contribution of Features

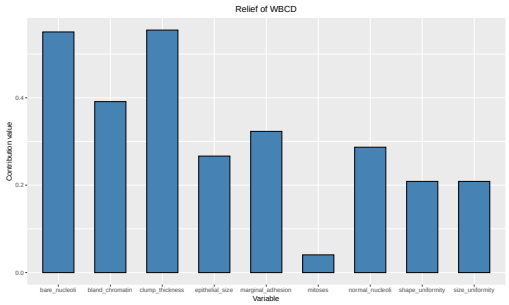

Fig. 11. Relief of Features

D. Classification models

1) KNN: In KNN, a classification model, the choice of k-value is the main influencing factor in the KNN model. Therefore, in this paper, different k-values are tried for model training and k-values are selected based on accuracy. From Fig. 12, it can be seen that this KNN model has the highest accuracy on the training set when the k value is 21. Therefore, this training model is chosen to be validated on the test set in this paper, which yields an accuracy of 95.45% for the KNN model (see Table 2).

2) SVM: In this paper, a non-linear SVM was selected, and in order to obtain a better classification model, the disciplinary factor C and sigma parameters were tuned so that the best parameters for this SVM model could be determined as sigma = 0.05 and C = 0.75 according to Fig. 13. According to table 1, the accuracy of the nonlinear SVM with this parameter on the validation set is 95.91% (see table 2).

3) RF: In the RF, the number of variables in the binomial tree (parameter mtry) is adjusted to select the best RF model, and the accuracy of different mtry values is shown in Figure 14. The specific parameters of the optimal RF model are shown in Table 1, and its accuracy rate on WBCD is 96.36%.

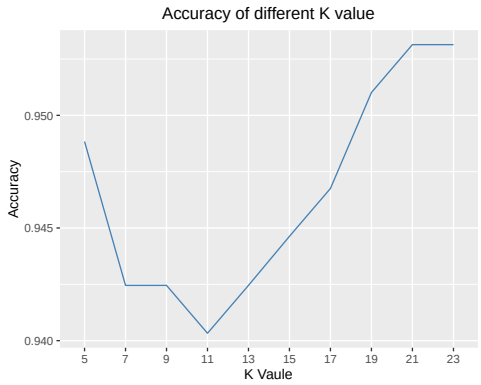

Fig. 12. Accuracy of different K value

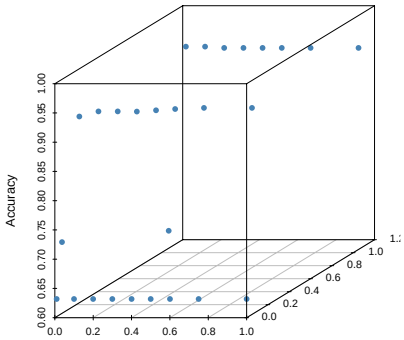

Fig. 13. Accuracy of different K value

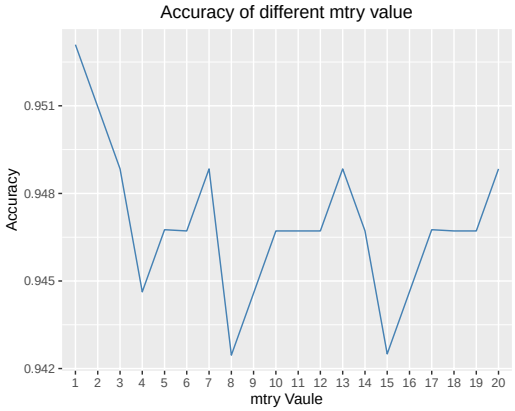

Fig. 14. Accuracy of different Mtry value

4) NB and LR: In NB and LR, the parameters of the two models are not optimized in this paper, and their accuracy on the WBCD-based validation set is 95% and 95.84%, respectively. (see Table 2).

V. Discussion

A. Model Evaluation

In the model evaluation section, the paper focuses on the accuracy, consistency, specificity and sensitivity of each model.

1) Accuracy: By comparing the accuracy of each of the classification models used, it can be seen from Table 2 and Figure 15 that each of the models achieved an accuracy of over 95%, with the RF model achieving the highest accuracy of 96.36%. In contrast, the NB model has the lowest accuracy of 95%, which is due to the fact that the parameters of the NB model were not optimised for this paper.

TABLE I: PARAMETERS OF RF MODEL

RF	**Parameters**			
	mtry	***ntree***	***nodesize***	***maxnodes***
Value	1	800	2	10

TABLE II: ACCURACY OF DIFFERENT CLASSIFICATION MODELS

Accuracy	**Models**				
	KNN	***SVM***	***RF***	***NB***	***LR***
Value	95.45%	95.91%	96.36%	95.00%	95.84%

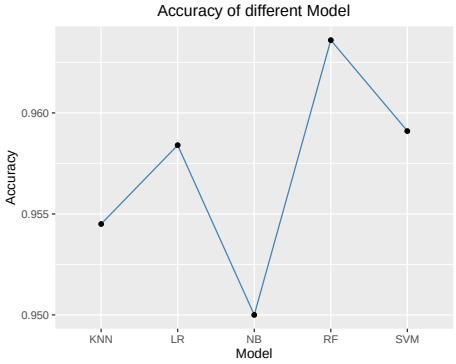

Fig. 15. Accuracy of different Models

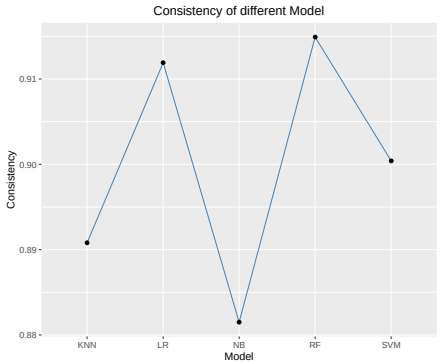

Fig. 16. Kappa value of different Models

As the SVM model with the second highest accuracy, SVM is suitable for datasets with a relatively small number of features. Since the WBCD dataset used in this paper is small and its number of features is also small after feature extraction, the SVM model performs well on this dataset. At the same time, the accuracy of LR is higher than that of NB, and without optimizing the parameters of LR in this paper, the comparison can be concluded that using Relief for feature extraction is appropriate.

2) Comparison of Consistency: This paper uses the k-fold cross-validation method to partition the dataset. Each model has different results when trained using each training set, so this paper uses the kappa coefficient to test the consistency of each model. As can be seen from Fig. 16, the consistency of the RF model is also higher than the other models reaching 0.9149, while on the contrary the consistency of the NB model is only 0.8908. And unlike the accuracy, the consistency of LR is slightly higher than that of SVM. Therefore, RF outperforms the other models not only in terms of accuracy but also in terms of consistency.

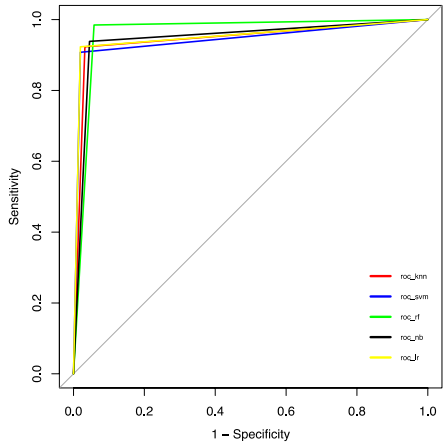

Fig. 17. ROC of different Models

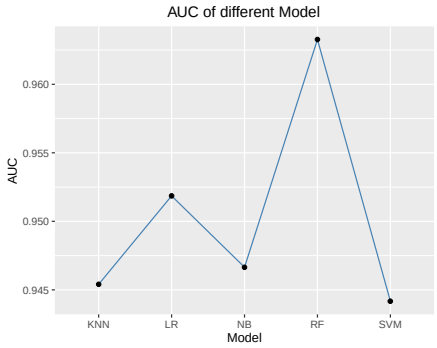

Fig. 18. AUC of different Models

3) ROC and AUC: The ROC curve is a graph that reflects both the specificity and sensitivity of a model, so this paper uses the ROC curve to evaluate these two indicators for the 5 models (see Fig. 17). Although the ROC curve can reflect the performance of each classification model in terms of specificity and sensitivity, it sometimes does not provide a visual indication of which classifier is superior, so this paper also examines the effectiveness of the classifier by the AUC value. According to Fig.17, Fig.18, it can be seen that the AUC values of all the models used in this paper exceed 0.94. According to the classification of AUC values, the classification models developed in this paper outperformed the random models in terms of WBCD, so all these models have predictive value. The AUC value of the RF model are close to 1, so although it is not perfect classifiers it can have high predictive power in the WBCD classification scenario.

B. Comparison of Methods and Results

In this section, the paper compares the five classification models used in the paper with the models proposed in the literature based on WBCD in Part II in terms of accuracy, sensitivity and specificity, and explores the differences in model performance brought by different methods.

As shown in Table III, the best performance in WBCD classification is the RF model with four Classification and regression tree (CARTs) and the EF model with three CARTs proposed by [12], which not only reach 100 in accuracy but also reach 100 in specificity and sensitivity, and It can be considered as a perfect predictor on the WBCD dataset. Since this paper only modifies some of the parameters of the RF and does not incorporate CARTs, it is not possible to process the classification and regression at the same time, which makes the RF model in this paper not achieve the performance of the model in [12].

TABLE III: THE CLASSIFICATION PERFORMANCE OF USED MODELS CAMPARED WITH LITERATURE MODELS

Model	**Performance**		
	Accuracy	***Sensitivity***	***Specificity***
BDG-NN [5]	84.00%	None	None
LM-NN [6]	99.30%	None	None
DBN-NN [7]	99.60%	1	0.99
IFFP [9]	97.20%	0.97	0.95
K-SVM [10]	98.30%	0.99	0.94
AR-SVM [11]	97.10%	0.98	0.98
RF (4 CARTs) [12]	100.00%	1	1
EF (3 CARTs) [12]	100.00%	1	1
KNN	95.45%	0.92	0.97
SVM	95.91%	0.91	0.98
RF	96.36%	0.99	0.94
NB	95.00%	0.94	0.96
LR	95.54%	0.92	0.98

Among SVM classifiers, both the K-SVM in [10] and the AR-SVM in [11] outperformed the non-linear SVM used in this paper, so the SVM classifier can be combined with other data processing methods through parameter optimisation, as well as combining different kernel functions to further improve the accuracy on WBCD.

At the same time, based on the effectiveness of [5], [6] and [7] in terms of NN models for WBCD classification, classifiers on NN can also be applied to this paper, thus enabling further comparison of different types of classifiers.

VI. CONCLUSION AND FUTURE WORK

Now, It is now particularly important for the diagnosis of breast cancer to achieve a prediction of malignancy or benignity of breast cancer cells based on data from breast cancer.

In this paper, different classifier models are developed to classify the data in WBCD. Firstly, the data from WBCD is analysed using different methods. In the data cleaning, we analyse and process the missing values, outliers and useless values of the data to complete the data cleaning. Secondly, this paper explores the WBCD data by analysing the statistics, data distribution and correlation between the data. In order to construct a suitable dataset for classification, PCA and Relief methods are used to extract features from the data set, thus effectively reducing the variables that have little impact on the classification. In the process of dividing the processed WBCD dataset into training and validation sets, the K-fold cross-validation method and stratified sampling were used to train the model more effectively and stably. For the classification of WBCD, five classifiers, KNN, SVM, RF, NB and LR, are used to further compare the performance of the different classifiers while completing the classification. Finally, in order to analyse the performance of the models used in this

paper and the models proposed in the literature, the accuracy of each model is compared, followed by the consistency analysis of the models using kappa coefficients, and the performance, specificity and sensitivity of the models are compared using ROC curve and AUC values. In summary, this paper concludes that the RF model is the most suitable classification model for WBCD among the five models.

In comparison with the models proposed in the literature, the models used in this paper are only partially optimised for the parameters and do not incorporate specific methods suitable for the respective models, thus making the performance of the classifier models not as good as in the literature. In the analysis of the WBCD data, more methods are needed to improve the analysis of the data and to make the training and validation sets used for the classifier more efficient. Therefore, in future work, this paper can combine the current methods in deep learning and data mining to further explore the analysis and feature extraction part of the data in more depth, and also try to combine the classifier with regression analysis and deep learning algorithms, so as to not only improve the accuracy of the model but also increase the stability and generalisation ability of the model.

REFERENCES

[1] S. G. Jacob, "Evolving efficient clustering and classification patterns in lymphography data through data mining techniques," *International Journal on Soft Computing*, vol. 3, no. 3, pp. 119–132, 2012.

[2] H.-Y. Liao, W.-W. Zhang, J.-Y. Sun, F.-Y. Li, Z.-Y. He, and S.-G. Wu, "The clinicopathological features and survival outcomes of different histological subtypes in triple-negative breast cancer," *Journal of Cancer*, vol. 9, no. 2, pp. 296–303, 2018.

[3] S. Singh, J. Prasad, S. Prasad, and S. Naik, "Breast cancer prediction using supervised machine learning techniques," *2021 IEEE 18th India Council International Conference (INDICON)*, 2021.

[4] V. Venkatesh, M. M. Anishin Raj, K. Mohamed Sajith, R. Anushiadevi, and T. Suriya Praba, "A precision-based diagnostic model adobe-accurate detection of breast cancer using logistic regression approach," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 6, pp. 8419–8426, 2020.

[5] N. Modi and K. Ghanchi, "A comparative analysis of feature selection methods and associated machine learning algorithms on Wisconsin Breast Cancer Dataset (WBCD)," *Advances in Intelligent Systems and Computing*, pp. 215–224, 2016.

[6] A. Marcano-Cedeño, J. Quintanilla-Domínguez, and D. Andina, "WBCD breast cancer database classification applying Artificial Metaplasticity Neural Network," *Expert Systems with Applications*, vol. 38, no. 8, pp. 9573–9579, 2011.

[7] A. Bhattacherjee, S. Roy, S. Paul, P. Roy, N. Kausar, and N. Dey, "Classification approach for breast cancer detection using back Propagation Neural Network," *Deep Learning and Neural Networks*, pp. 1410–1421, 2020.

[8] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using Deep Belief Networks," *Expert Systems with Applications*, vol. 46, pp. 139–144, 2016.

[9] I. Thani and T. Kasbe, "Expert system based on fuzzy rules for diagnosing breast cancer," *Health and Technology*, vol. 12, no. 2, pp. 473–489, 2022.

[10] F. Ramesh Dhanaseelan and M. Jeya Sutha, "Detection of breast cancer based on fuzzy frequent itemsets mining," *IRBM*, vol. 42, no. 3, pp. 198–206, 2021.

[11] T. Ibrikci, D. Ustun, and I. E. Kaya, "Diagnosis of several diseases by using combined kernels with support vector machine," *Journal of Medical Systems*, vol. 36, no. 3, pp. 1831–1840, 2011.

[12] A. Ed-daoudy and K. Maalmi, "Breast cancer classification with reduced feature set using association rules and Support Vector Machine," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, 2020.

[13] M. M. Ghiasi and S. Zendehboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," *Computers in Biology and Medicine*, vol. 128, p. 104089, 2021.

[14] A. M. Carrington, D. G. Manuel, P. Fieguth, T. O. Ramsay, V. Osmani, B. Wernly, C. Bennett, S. Hawken, O. Magwood, Y. Sheikh, M. Mcinnes, and A. Holzinger, "Deep Roc analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022.

[15] Z. Mushtaq, A. Yaqub, A. Hassan, and S. F. Su, "Performance analysis of supervised classifiers using PCA based techniques on breast cancer," *2019 International Conference on Engineering and Emerging Technologies (ICEET)*, 2019.

[16] H.-C. Lu, E.-W. Loh, and S.-C. Huang, "The classification of mammogram using convolutional neural network with specific image preprocessing for breast cancer detection," *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 2019.

[17] Z. Nematzadeh, R. Ibrahim, and A. Selamat, "Comparative studies on breast cancer classifications with K-fold cross validations using Machine Learning Techniques," *2015 10th Asian Control Conference (ASCC)*, 2015.

[18] *UCI Machine Learning Repository: Breast Cancer wisconsin (original) data set*. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29. [Accessed: 18-Apr-2022].

[19] A. reddy, "Support vector machine classifier for prediction of breast malignancy using wisconsin breast cancer dataset," *Journal of Artificial Intelligence, Machine Learning and Neural Network*, no. 21, pp. 1–8, 2022.

[20] M. Hosni, I. Abnane, A. Idri, J. M. Carrillo de Gea, and J. L. Fernández Alemán, "Reviewing Ensemble Classification Methods in Breast Cancer," *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 89–112, 2019.