Artificial Intelligence Model Selection for Breast Cancer Risk Screening

Ziwen Fang

Software engineering, Lappeenranta-Lahti University of Technology(LUT), Yliopistonkatu 34,
53850 Lappeenranta, Finland
Email: Ziwen.Fang@student.lut.fi

Abstract. In today's social environment, the risk of breast cancer for women is increasing, and breast cancer has exceeded lung cancer as the most common cancer nowadays. However, if detect breast cancer at an early stage and measures are taken, it can be very effective in improving the chances of survival of breast cancer patients. Meanwhile, with the continuous development of artificial intelligence, it shows a broad prospect in the medical field. In this article experiment try to apply AI to the field of breast cancer risk detection, and help improve the accuracy of breast cancer screening by finding the artificial intelligence model with the highest accuracy rate. This article selected breast cancer data from kaggle, pre-processed the data by Pearson Correlation Coefficient, and then the article compares four of the most common machine learning algorithms namely Random Forest, Logistic Regression, Neural Networks, and Support Vector Machines, using Python. Based on the experimental results the article conclude that Random Forest is highly accurate and shows great affect in the field of breast cancer screening.

Keywords: Artificial Intelligence, Machine Learning, Breast Cancer Risk Screening

1 Introduction

In today's society, breast cancer has become the most common type of cancer among women worldwide. According to the United Nations, in 2020, the number of new cases of breast cancer reached 2.3 million, accounting for 11.7% of new cancer cases worldwide, which is the first time that the number of new cases exceeded that of lung cancer. As the number of breast cancer diagnoses continues to increase, the importance of early diagnosis and treatment of breast cancer is growing. Therefore, it is especially critical to develop technologies that can effectively predict breast cancer in order to detect the risk of the disease in advance, take preventive measures, and provide timely treatment. Meanwhile with the rapid development of big data and computational science, artificial intelligence (AI) has demonstrated great potential in the diagnosis and prediction of breast cancer [1].

Computer-aided diagnosis is becoming a popular area of development in the current field of breast cancer risk monitoring, and attempts have been made to adopt

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_64

computers as well as a number of machine-learning methods to assist doctors in making a diagnosis, but there are no particularly prominent studies showing which machine-learning method is more accurate. In addition, in Nehmat Houssami's study, it was shown that using AI to screen and judge breast images can largely identify breast cancer risk earlier, but there may also be omissions, and it is important to choose AI models that are more accurate [2].

In this study, it aim to find a machine learning based breast cancer prediction model to identify the potential risk of breast cancer through data analytics techniques using patients' medical images and clinical data. The aim of this study is to help doctors understand their patients' cancer risk better and faster by quasi-finding more accurate prediction tools, which will lead to earlier intervention and faster medical intervention for breast cancer.

2 Related Work

The diagnosis of breast cancer is determined by a combination of many factors, including average radius, average texture, average circumference, average area, and average smoothness, and breast cancer is diagnosed when these indicators are outside of a certain range, and the use of these indicators also allows us to determine whether breast cancer is in the early-intermediate stage or advanced stage. As well as research on cancer risk prediction covers a variety of approaches related to AI models, including some machine learning methods such as random forests, logistic regression, Support Vector Machines (SVM) and neural networks, deep learning and some hybrid models.

Recent studies have used deep learning to improve prediction accuracy. Zhang et al. (2019) used a convolutional neural network (CNN) to analyse mammography images to extract subtle patterns that may have been missed by human radiologists [3]. Their results show that convolutional neural networks can significantly improve the detection of early breast cancer. But there is a risk that CNNs trained for a specific set of mammography images may not perform well when processing data from different populations or different types of imaging equipment. Meanwhile the biggest problem that exists with deep learning is the lack of transparency, which can be a major issue in healthcare applications, as clinicians need to understand the basis of model predictions in order to trust and use them effectively in clinical practice.

Also ensemble methods have been applied for breast cancer risk prediction.Lee and Kim (2020) used ensemble models such as Random Forest and Gradient Boosting Machine (GBM) to aggregate the prediction results from multiple models, thereby improving the reliability and robustness of forecasts [4]. These models perform well with a variety of datasets including clinical data, lifestyle factors and patient demographics. However, this approach is computationally complex, particularly as breast cancer detection often relies on breast impact, and the complexity of impact-related data increases prediction risk.

There are also approaches related to machine learning, and one of the main approaches involves traditional machine learning models.Researchers such as Weng, Liu and Hsu (2015) used logistic regression and support vector machines (SVMs) to identify patterns in genetic data and lifestyle factors that can be predictive of breast cancer risk [5]. Their model, while simpler than deep learning methods, provides a strong baseline for understanding feature relationships.Another innovative approach is to combine machine learning with genetic algorithms, which was explored by Tan, Teo and Anderson (2018) [6]. They developed a hybrid model that combines feature selection from genetic algorithms with deep learning models to predict breast cancer risk from high-dimensional data such as gene expression profiles. This approach helps to optimise the model by selecting the most informative features to improve prediction accuracy. All of these approaches require the selection of higher accuracy machine learning models, so in our study we will attempt to identify the most accurate of the four most popular machine learning algorithms currently available to aid in the future development of artificial intelligence detection models for breast cancer risk.

3 Method

In the research will focus on building an artificial intelligence model for breast cancer prediction using machine learning techniques with the collected dataset. It start by performing data relevance assessment, there are many different metrics included in breast impact and by assessing the correlation between different metrics and breast cancer, the features that contribute to the prediction model can be effectively selected. In the study it will use the Pearson correlation coefficient to identify the key physiological indicators that affect the risk of developing breast cancer and generate a heat map to get a clearer understanding of the scope of the effect of different indicators on developing breast cancer, providing the basis for the later work.

After obtaining the relevant metrics, the study processed the large-scale dataset using Random Forest, SVM, Neural Networks, and Logistic Regression, respectively, and constructed a basic model for machine learning to learn and make predictions using the data obtained (Fig.1). The performance of the created models was evaluated using a cross-validation approach to assess their generalization ability on various patient datasets. The evaluation focused on accuracy, recall, F1 score, and ROC curve analysis. This emphasis ensures the models' interpretability and predictive stability in real clinical applications.

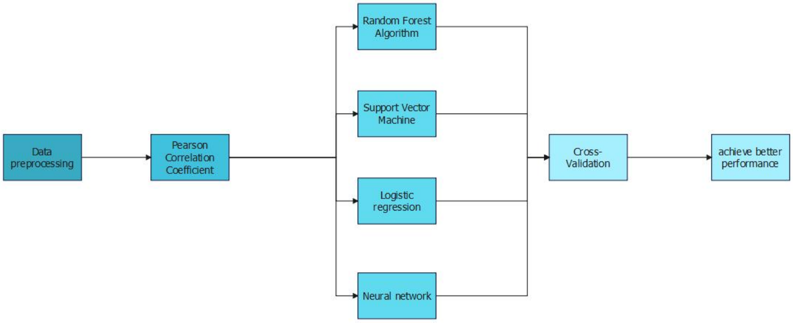

Fig. 1 Flowchart of data processing (Photo/Picture credit :Original)

3.1 Data Preprocessing

The data for the study was selected from Breast Cancer Prediction Dataset on kaggle website and processed to remove erroneous data as well as analysing the correlation of each metric. Meanwhile, in the process of data processing we consider the ethical is Pearson correlation, which as a common method of measuring correlation between data, can be very helpful to detect the correlation between individual indicators and determine the important indicators for the detection of breast cancer [7].

Firstly the study process the data collected for the two variables that need to be checked for correlation, the physiological indicators include: mean_radius, mean_texture, mean_perimeter, mean_area, and mean_smoothness. Ensuring that the data were clean and free of errors, missing values, outliers, and anomalies that could bias the results were dealt with. The mean value of each variable was calculated after correlating the data, where n is the data points number, and x_i, y_i are the data points for each variable. Use these sums to calculate the Pearson correlation coefficient r using the formula.

$$r_{xy} = \frac{(x_i - x)(y_i - y)}{((x_i - x)^2 \quad ((y_i - y)^2} \tag{1}$$

The correlation coefficients are calculated for each pair of variables using the above formula and a matrix of correlation coefficients is formed. This matrix will be a symmetric matrix where the rows and columns of the matrix represent the same variables. The study chose Python to generate heat maps to show whether each indicator is positively or negatively correlated with the risk of cancer and the magnitude of the strength of the correlation for each indicator. As shown in figures 1, 2 and 3 below.

3.2 Random Forest Algorithm

In this article first employed the random forest algorithm to build our model, and setting the appropriate parameters, including the number of trees. Generally, increasing the number of trees enhances the model's performance and stability, but also raises computational demands. Therefore, adjusted this parameter based on the dataset's size. The second is the maximum number of features considered for each decision tree split, a lower value enhances the generalisation of the model and avoids overfitting. And an algorithm is used to control the maximum depth and growth conditions of each tree, limiting the depth there to prevent model overfitting [8].

The article use automatic aggregation to train multiple decision trees using a training dataset, each tree is trained on a different random subsample of the dataset. During the construction of each tree, features are randomly selected for node splitting to increase model diversity and reduce the risk of overfitting.

Finally, study used a trained random forest model to make predictions on the test dataset. In this process, study ensure that each tree makes its prediction independently and the final prediction is determined based on the average or majority voting method derived from the predictions of all trees.

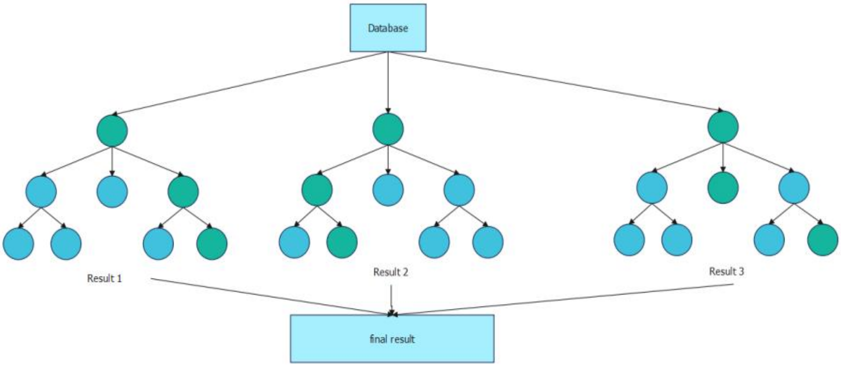

Fig. 2 Diagram of the random forest approach [8]

3.3 SVM

SVM are powerful classification tools as a supervised learning algorithm, and in the experiments and also checked its accuracy with its training algorithm i.e. building a model that can assign new instances to one of the two classes, helping to build a non-probabilistic classifier [9].

Our main goal when constructing the SVM model is to find an optimal hyperplane that divides the samples in the dataset into different classes and maximises the spacing between the two classes.The optimisation problem of SVM can be formulated as finding an optimal hyperplane to minimise the following objective function, where the normal vector of the hyperplane w, the bias term b, a regularisation parameter that controls the degree of punishment for misclassified samples C, ε_i is a slack variable used to allow some samples to be on the wrong side, and the number of samples N.

$$min_{w,b}\frac{1}{2}||w||^2 + C \quad {}_{i=1}^{N}\varepsilon_i \tag{2}$$

Next we use the prediction function of the SVM model for prediction.

$$f(x) = sign(w^T x + b) \tag{3}$$

Similarly, in order to balance the complexity of the model with its generalisation ability, the study determine the optimal C value by means of a cross-validation method. Next, try to train the SVM model using this parameter as well as the training set data, with the aim of obtaining the optimal hyperplane parameters, i.e., the weights w and the bias b.

3.4 Logistic Regression

After data preprocessing is complete, first initialise the model parameters, i.e. the weight parameters (θ). Then, a logistic function is defined which maps a linear combination of input features (denoted as z) to a probability value between 0 and 1 for further computation and analysis.

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{4}$$

Secondly the study will define the cost function which is again the cross-entropy loss function and choose the appropriate cost function to measure the difference between the model predictions and the true labels.

$$J(\theta) =-\frac{1}{m} \ {}_{i=1}^{m}\{y^{(i)}log(h_\theta(x^{(i)})) + (1 - y^{(i)})log(1 - h_\theta(x^{(i)}))\} \tag{5}$$

After calculating the difference, it use gradient descent to update the model parameters (θ) with the aim of minimising the cost function. For each parameter θ, update it according to the rules of gradient descent, and this process will continue until the model satisfies the convergence condition or reaches the preset maximum number of iterations.

3.5 Neural Network

When using neural network methods for classification tasks, after pre-processing the data, ensure that the features are normalised so that the scales are similar between different features. Secondly, during the dataset division process, the study use the same cut scales as for the random forest above.

When building a neural network model, initially establish its architecture by selecting the number of layers, the quantity of neurons in each layer, and the activation functions to be used. In addition, and also randomly initialised the weights and biases of the neural network [10]. In the article experiments, the neural network was configured as a single hidden layer, which contained 100 neurons, and a linear rectifier function (ReLU) was chosen as the activation function. Also in our code, the initialisation of the weights and biases usually depends on the chosen framework and its default settings; random initialisation's were performed using the MLPClassifier module of scikit-learn to start the training process. In addition, specify the solver for weight optimisation as Adam's algorithm, which serves as an effective adaptive learning rate optimisation algorithm that can handle large and complex datasets very efficiently. With these approaches, configured and optimised the neural network to solve the breast cancer classification problem, while the maximum number of iterations was set to 1000 to allow the model to fully learn and adjust its weights to achieve the best possible performance.

According to the parameters which is set, the output values of each layer are calculated by forward propagation, and for each hidden layer, the outputs of the weighted inputs and activation functions are calculated, and the loss function is calculated, and the difference between the model prediction results and the real labels is measured using an appropriate loss function.

$$J(\theta) =-\frac{1}{m} \ {}_{i=1}^{m}\{y^{(i)}log(y^{(i)}) + (1 - y^{(i)})log(1 - y^{(i)})\} \tag{6}$$

Finally, the gradient of the loss function for each parameter is calculated using the backpropagation algorithm, and the model parameters are updated according to optimisation algorithms such as gradient descent to minimise the loss function, and for each parameter, parameter updates are performed according to the update rule of gradient descent until the convergence condition is reached or the maximum number of iterations is reached.

4 Experimental Setup and Results

4.1 Evaluation Metrics

In the experiments, it use three main evaluation metrics to measure the performance of the model. The first is accuracy, which is an intuitive metric for assessing the overall effectiveness of the model. Next is recall, a metric that evaluates the ratio between actual positive items identified by the model and all actual positive items, which is particularly important in disease diagnosis. Finally, it used the F1 score, which is the reconciled mean of precision and recall, aiming for a balance between the two. In experiments using Python, these metrics were computed via the sklearn metrics library.

4.2 Experimental Setup

In the data processing phase, the data is loaded from CSV files via the load_data function and then split using train_test_split and cross_validation_split to help the model can have a different subsets of data to trained and tested.

In the model training and validation phase, a 5-fold cross-validation was used, which is a method that effectively uses limited data to evaluate the performance of the model. In cross-validation, we divided dataset into small subgroups, each of which alternates as a test set and a training set.

In the performance evaluation phase, the evaluate_algorithm function is used to calculate and print out the mean values of accuracy, recall and F1 score, which are obtained by training and testing on different data folds to ensure the evaluation on fairness and accuracy.

4.3 Experimental Results and Discussion

In the study it use the Pearson correlation coefficient to measure the strength of the linear relationship between two variables. A coefficient of 1 indicates a perfect positive correlation, i.e., as one variable increases, the other variable continues to increase each year. Conversely, a coefficient of -1 indicates a perfect negative correlation, i.e., an increase in one variable is associated with a continued decrease in the other variable each year. A coefficient of 0 indicates that there is no linear correlation between the variables [11].

It also used Python to find the correlation of each metric by reading the data downloaded from Kaggle above, calculating the Pearson correlation coefficients for all columns, and plotting the heatmap.

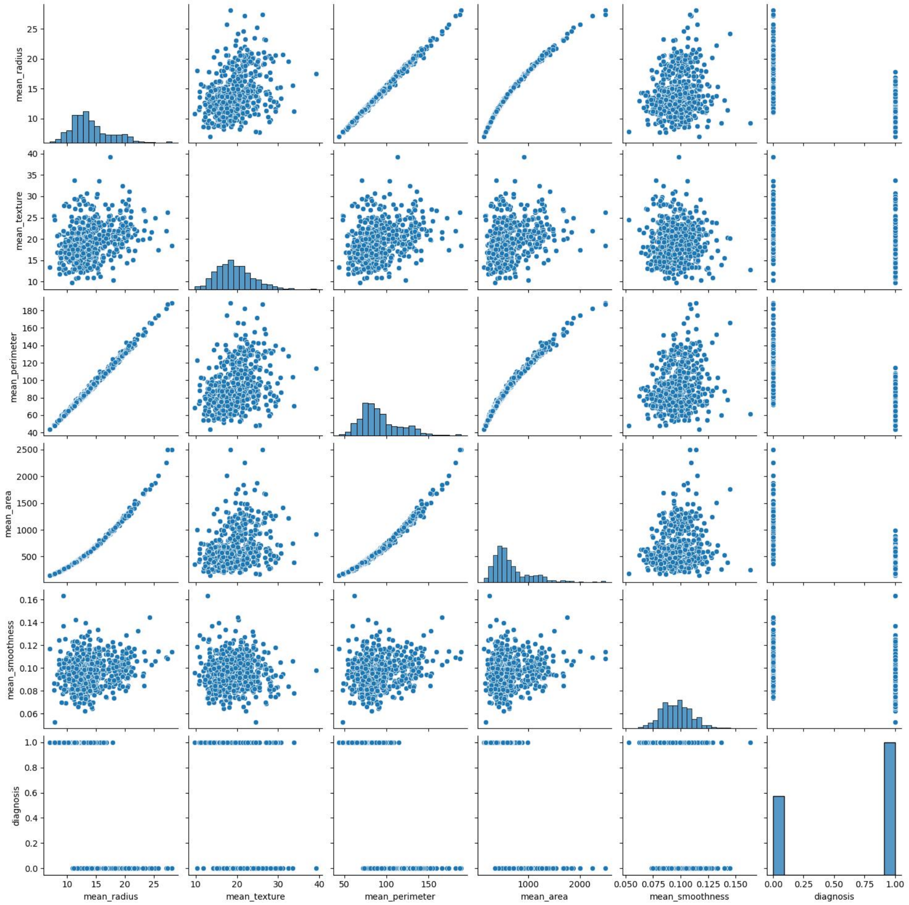

Fig. 3 Indicator correlation scatter plot (Photo/Picture credit :Original)

As can be seen from the experimental results (Fig.3 and Fig.4), the features mean_radius, mean_perimeter, and mean_area show strong correlation with diagnosis. Whereas mean_texture and mean_smoothness showed a weaker correlation with cancer diagnosis and may not have a direct correlation for tumour size or having cancer. However, we can see that these features have some influence on cancer diagnosis, and are all indicators that we need to consider when performing AI model training.

In the scatterplot mean_radius, mean_perimeter, mean_area show similar distributions, all of them are positively skewed, i.e. most of the data are clustered in a smaller range of values, but there are some larger values lengthening the tails. mean_texture and mean_smoothness have a more uniform distribution, but also slightly positively skewed. Diagnosis as a binary variable is represented as a histogram of two peaks in the plot. The scatterplot shows a very strong positive correlation between mean_radius, mean_perimeter, and mean_area, that this consistent with the findings from the correlation heatmap. These features are strongly related to each other because they measure the physical size of the tumour. The relationship with diagnosis is also clearly categorical, especially in mean_radius, mean_perimeter, and mean_area, where two distinct clusters of points can be seen, suggesting that these features may be very effective in distinguishing cancer status. mean_texture and mean_smoothness are also very strong in the correlation heat map. smoothness also have some correlation and need to be considered when we screen for cancer.

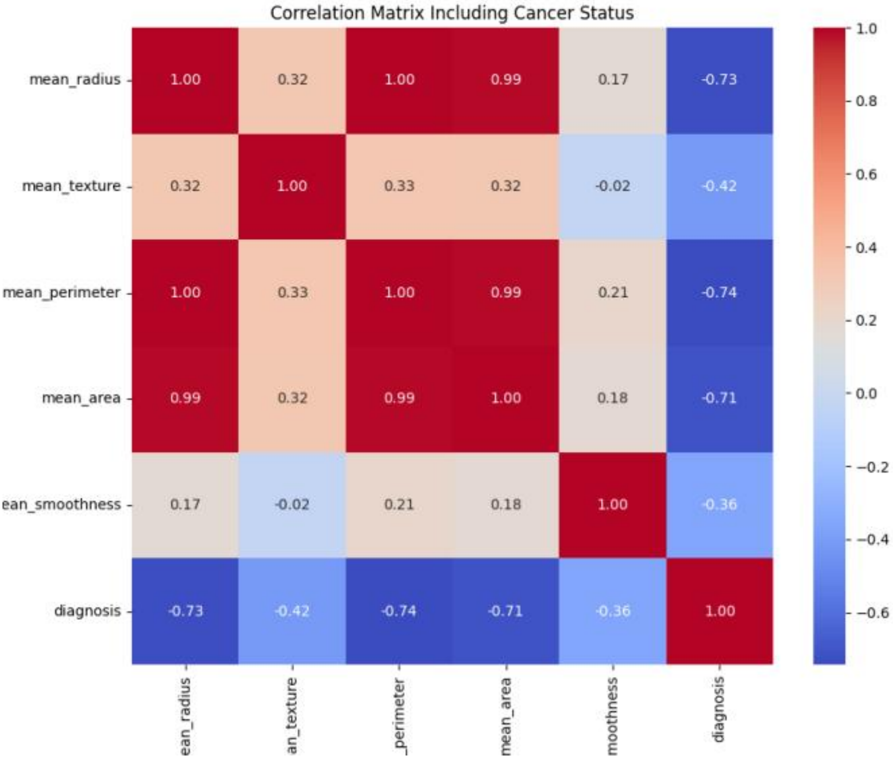

Fig. 4 Pearson's correlation coefficient heat map (Photo/Picture credit :Original)

At the same time study used Pearson correlation coefficient to calculate the size of the correlation between each indicator and developing breast cancer to get the following results (Fig.5).

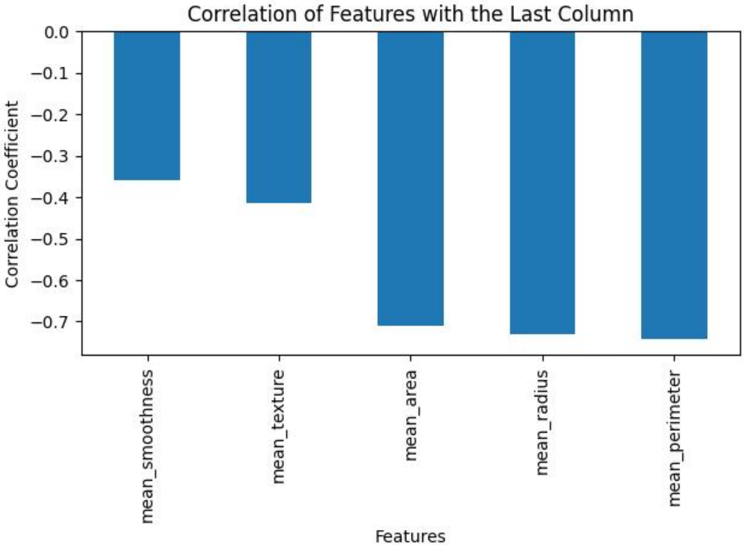

Fig. 5 Indicator correlation coefficients graph (Photo/Picture credit :Original)

From the feature coefficients it can also see that all of the selected features have a strong correlation with having breast cancer. mean_perimeter, mean_radius, and mean_area have the strongest correlation with having breast cancer, while the correlation for mean_texture and mean_smoothness is weaker but still significant, suggesting that these features related to the texture and smoothness of the tumour surface, while having some impact on cancer diagnosis, have less of an impact compared to the physical size of the tumour, which is also consistent with our experimental results above.

After obtaining the correlation used Python to test and train the four methods, Random Forest, SVM, Logistic Regression and Neural Networks, and performed cross-validation and parameter tuning to arrive at the following results with high accuracy (Table 1).

Table 1 Statistics on the results of the four methods

Model	Mean Accuracy	Mean Recall	Mean F1 Score
Random Forest	91.327%	0.959	0.932
Support Vector Machine	90.795%	0.940	0.927
Neural Network	87.611%	0.944	0.0.904
Logic regression	90.973%	0.960	0.0.930

In the final experimental results, which can see that Random Forest performs best with high accuracy, recall and F1 score. By its very nature, Random Forest can efficiently handle datasets with a large number of features and mitigate the effects of overfitting through the 'integration' method of integrated learning. This makes

Random Forests generally have more stable and reliable performance on multi-feature datasets.

In addition logistic regression and support vector machines also show good performance, especially in terms of recall and F1 scores, but not as good as random forests. Logistic regression is a relatively simple model that relies on the linear divisibility of the data, whereas we are breast cancer related data is sometimes not linear, which may lead to their lack of performance in the face of complex non-linear relationships. SVMs, despite being able to deal with non-linear data, may require appropriate kernel functions and parameter tuning to achieve optimal performance when the number of features is high.

Neural networks, while performing well in terms of recall, have relatively low overall accuracy and F1 scores. The performance of neural networks is affected by several factors, especially the need for a large dataset to learn complex function mappings, and our dataset is very limited, which leads to neural networks that may not be adequately tuned to the characteristics of the data in the network architecture, with inappropriate number of layers or neurons, which affects the model's ability to learn. To improve the performance of neural networks we need larger datasets to conduct experiments to develop and adjust the network architecture or optimise the parameters.

5 Conclusion

In order to find the machine learning algorithm with the highest accuracy, we kept changing. The results show that the accuracy rate of random forest is above 90%, which is the highest accuracy rate among the four models. And Random Forest has high interpretability, which is very suitable for medical applications. However, there are still shortcomings in our experiments, especially the limitation of the dataset. In our future work we have to collect more data to validate our experiment. Meanwhile, in our future work, we can try to use an integrated approach to apply random forest while combining the advantages of different models. In the future we also need to keep conducting comparative studies to compare the Random Forest model with the state-of-the-art models for similar tasks or to explore the performance of the model in different datasets or domains. The application of random forest model is not limited to breast cancer detection; it can be widely used in the detection of other diseases. We should continue to explore and expand the potential of the Random Forest Model in the field of disease detection in order to enhance its functionality and provide more support for future disease diagnosis.

References

1. Chan, H. P., Samala, R. K., & Hadjiiski, L. M. CAD and AI for breast cancer—recent development and challenges. The British journal of radiology, 93(1108), 20190580, (2019).

2. Houssami, N., Lee, C. I., Buist, D. S., & Tao, D. Artificial intelligence for breast cancer screening: opportunity or hype?. The Breast, 36, 31-33, (2017).
3. Zhang, Q., Zhao, L., Luo, X., & Chen, H. Enhancing Breast Cancer Detection via Deep Learning Analysis of Mammographic Images. Clinical Radiology, 74(6), 337-345, (2019).
4. Lee, J.H., & Kim, M.Y. Breast Cancer Risk Prediction Using an Ensemble of Machine Learning Methods. Health Informatics Journal, 26(2), 1234-1245, (2020).
5. Weng, C.Y., Liu, Y.F., & Hsu, W.L. Predictive Modeling of Genetic Risk Factors for Breast Cancer Using Logistic Regression and SVM. Journal of Medical Informatics and Decision Making, 15(3), 17-29, (2015).
6. Tan, S.M., Teo, J.P., & Anderson, R. A Hybrid Model of Deep Learning and Genetic Algorithms for Breast Cancer Prediction. Computational Biology and Chemistry, 42, 53-62, (2018).
7. Schober, P., Boer, C., & Schwarte, L. A. Correlation coefficients: appropriate use and interpretation. Anesthesia & analgesia, 126(5), 1763-1768, (2018).
8. Biau, G., & Scornet, E. A random forest guided tour. Test, 25, 197-227, (2016).
9. Wikipedia contributors. Support Vector Machine. In Wikipedia, The Free Encyclopedia. Retrieved 5:32, June 10, (2024), from https://en.wikipedia.org/wiki/Support_Vector_Machine
10. Goldberg, Y. A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, 57, 345-420,(2016).
11. Wikipedia contributors. Pearson correlation coefficient. In Wikipedia, The Free Encyclopedia. Retrieved 12:37, June 10, (2024), from https://en.wikipedia.org/wiki/Pearson_correlation_coefficient