June 30, 2023

Machine Learning Application in cfDNA Analysis to Achieve Tumour Assessment

Yiting Zheng[1*], Zixin Gong[2],Yifeiyang Guo[3], Menghan Huang[4], Rui Li[5], Haiting Gan[6]

[1]School of Life Science and Technology, China Pharmaceutical University, Nanjing 211198, China
[2]College of Life Sciences, Sichuan University, Chengdu 610065, China.
[3]School of Life Sciences, Yunnan University, Kunming 650500, China.
[4]School of Health Economics and Management, Nanjing University of Chinese Medicine, Nanjing 210046, China.
[5]College of Ocean Food and Biological Engineering, Jimei University, Xiamen 361000, China.
[6]College of Ocean Food and Biological Engineering, Jimei University, Xiamen 361000, China.
[*]Correspondence: zhengyt365@163.com; Tel: +86 18964073593

1 Abstract

Breast cancer (BC) is the leading cause of cancer in women and the second leading cause of cancer-related death. Early and accurate screening of BC is a promising way of reducing the proportion of patients with advanced stages of BC. In recent years, the non-invasive test of tumour diagnosis by assessing the level of Plasma cell-free DNA (cfDNA) has become a research hotspot. Here, we demonstrate the use of random forest models to predict BC by evaluating the levels of 26 known breast cancer-related cfDNA methylation molecular markers (model-tested accuracy of 67.88%). Then, we improved the accuracy of the model to 71.52% by parameter optimization. In addition, considering that the diagnosis of BC is closely related to the health of every female, we have extended the project from scientific research to social investigation by carrying out a sample survey of Chinese college students to understand various perspectives on the application of artificial intelligence in the diagnosis of diseases. We found that the response was rather optimistic, while some participants showed concerns about the maturity of the technology and the disclosure of privacy. Therefore, future research should focus on the optimisation of the machine learning model, so as to effectively improve the accuracy of diagnosis and provide better pre-service for the population at risk of cancer.

Keywords:breast cancer,screen,molecular marker,machine learning.

2 Project Results

2.1 Introduction

2.1.1 Current concerns

Malignant tumours are a major cause of illness and death (Siegel, Miller, Fuchs, & Jemal, 2022), and their current treatment can be expensive and ineffective. Furthermore, apart from the complicated procedures and physical pain (Filippiadis, Charalampopoulos, Mazioti, Keramida, & Kelekis, 2018), traditional biopsy methods may have limited sensitivity and may not detect cancer until it has progressed to advanced stages (Mannelli, 2019). Therefore, there is an urgent need for a more efficient and cost-effective solution to improve the diagnosis and treatment of malignant tumours.

2.1.2 Target diagnostic tool

. Circulating free DNA (cfDNA), as a novel non-invasive diagnostic tool, has attracted much attention in recent years. Studies have shown that cancer cells release larger amounts of cfDNA into the bloodstream compared to normal cells, and this DNA often carries mutations and genetic alterations specific to cancer (Mannelli, 2019) (Chabon et al., 2020). By analysing cfDNA, researchers can identify these genetic changes and determine the presence of cancer, its type and stage, and monitor its progression over time (Moss et al., 2020).

2.1.3 Combine machine learning with cfDNA

. According to previous research, the generation of breast cancer (BC) is closely related to the methylation of cfDNA (Liu et al., 2021). Our research group intends to use a database to combine machine learning and BC detection to achieve the purpose of BC diagnosis by identifying characteristic cfDNA fragments. According to analysis, cfDNA methylation profiling may serve as a reliable approach for BC diagnosis (Zhang et al., 2021). With only a few markers, the model can be widely applied to large-scale BC screening at a low cost. The method was first to sift through a batch of data from the database for processing and use it for random forest model (RFM) training, then test it, and after that experiment with new data. Finally, we also conducted social research on AI in healthcare to deepen our understanding of AI itself and its contribution to society.

2.2 Description of methodologies

2.2.1 Random forest model

(1) Data collection: Due to the current lack of research on cfDNA, there are few credible and available databases. However, our research group found a publicly available dataset for model training. This dataset has been deposited at the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute. After screening and processing of the dataset, 333 methylation scoring data of 26 markers (including a malignant tumor group and a healthy group) were finally collected, which can be used to indicate the degree of methylation of cfDNA fragments.

(2) Random forest model training: Our research group used m features in the dataset to construct a decision tree. First, we iterated through each feature in m, then iterated through each row, selected the optimal feature and feature value through the split-loss function (which can calculate the cost of segmentation), classified according to whether it is greater than this feature value (by dividing into two categories of left and right), and performed the above steps in a loop. This was repeated until it was indivisible or reached a recursive limit (which can prevent overfitting). Then a decision tree was obtained (Figure 1).

(3) Model evaluation: Judging on each row of the test set, the decision tree is a multi-layer dictionary. Each layer is two-class, and each row is according to the classification index in the decision tree, where the inner layer is explored step by step, until the end of exploration when there was no dictionary, and the value obtained was our predicted value.

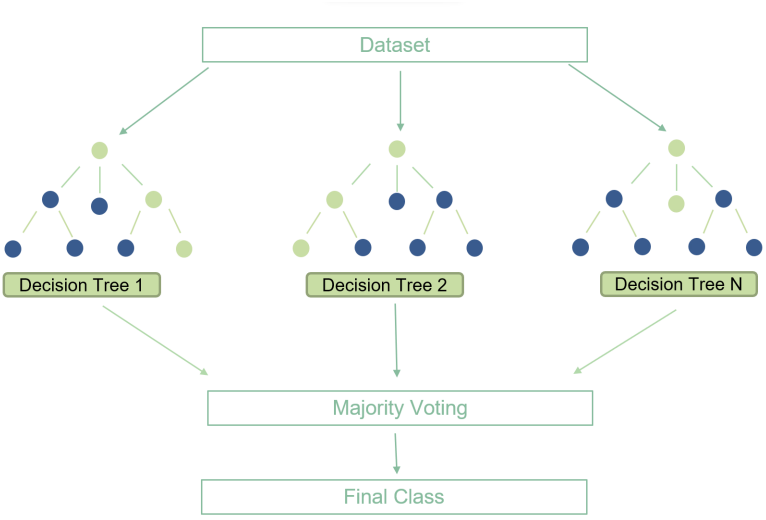

Figure 1: Illustration of the random forest method

2.2.2 Sampling questionnaire survey

(1) Investigation purpose: We aimed to collect the perspectives of Chinese undergraduates on the use of AI in BC detection, and to investigate the respondents' knowledge of BC and their willingness to use AI detection technology.

(2) Questionnaire design: The questionnaire consisted of two parts. The first part aimed to understand the respondents' knowledge of BC, and the second part aimed to investigate the respondents' willingness and relevant opinions on using AI technology to detect BC. In addition, we were also concerned about users' views on the issue of data supervision. Therefore, we set questions about their willingness to disclose their data to open scientific research and their ideal data management institution.

2.3 Results

2.3.1 Differences in methylation levels between breast cancer patients and healthy individuals

Our research group performed a statistical analysis of the obtained methylation score data (Liu et al., 2021) and found that breast cancer patients (BCPs) do have different characteristics from healthy individuals (HIs) at specific methylation marker sites. For instance, aberrant hypomethylation is a common feature of various malignant tumours and usually occurs in cfDNA-rich gene bodies or closed chromatin in intergenic regions (Snyder, Kircher, Hill, Daza, & Shendure, 2016). Four of the 26 methylation markers were selected for box plotting (Fig. 2). It was found that the methylation levels of site cg16304215, cg20072171, and cg21501525 in BCPs were higher than those in HIs. However, the methylation level of BCPs at site cg23035715 was lower than HIs.

2.3.2 Application of RFM based on cfDNA methylation level in BC diagnosis.

Our research group constructed a machine learning model by using random forest to realise a preliminary diagnosis of BC based on cfDNA methylation level. By utilising the methylation score data to train the RFM, the final test results showed that the average accuracy of our model was 67.88% (Figure 3) under the condition of 26 methylation markers, with a tree depth of 15.

2.3.3 Effect of the number of markers on the accuracy of the model.

In order to further improve the accuracy of the model, four of the most representative methylation markers, cg16304215, cg20072171, cg21501525, and cg23035715, were selected to

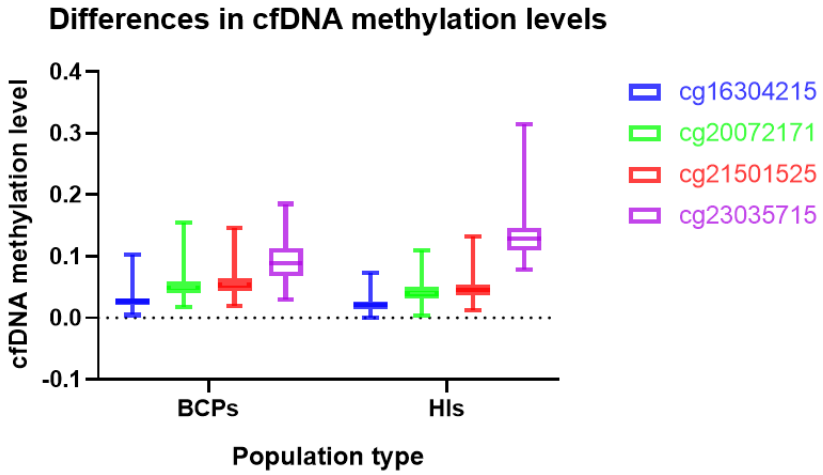

Figure 2: Differences in methylation levels between BCPs and HIs. The population was divided into two groups: breast cancer patients (BCPs) and healthy individuals (HIs). Four characteristic methylation markers were selected, and GraphPad Prism 8.0 was used to draw the box plot.

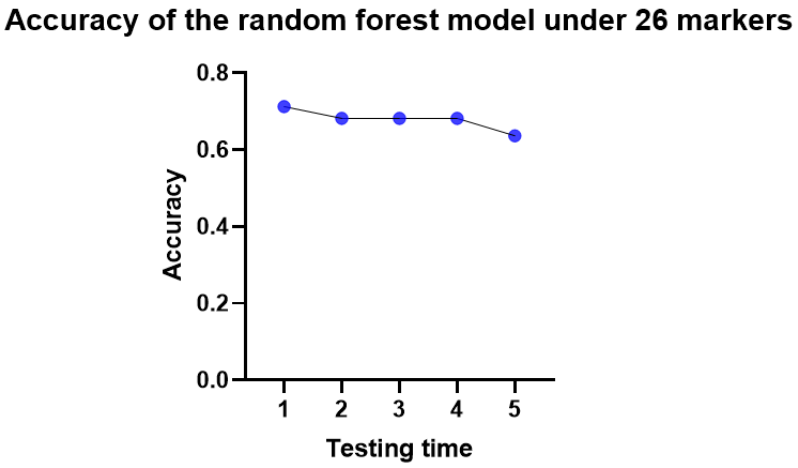

Figure 3: RFM accuracy test results (26 methylation markers, mean score of 67.88%).

re-model and re-run the programme, and the tree depth was 10. We found that the accuracy of the model did show a small improvement, from 67.88% to 71.52% (Fig. 4). This indicates that our model can be further optimised to be more efficient, ideal, and systematic. Moreover, it provided us with the idea of designing a complete breast cancer diagnosing system based on cfDNA methylation level, which will be mentioned in the discussion section.

2.3.4 Social investigation results

The sample validity of 115 questionnaires was screened to achieve validity. The results analysed were as follows (Figure 5):

(1) The social awareness of BC needs to be improved. In the survey, 65.22% of the sample pointed out that the current social awareness of BC was not enough, and 85.22% of the samples had only superficial knowledge of BC.

(2) College students respond well to AI technology for BC screening. The statistics indicate that 95.65% of the sample respondents were willing to use AI technology for BC screening, and a few subjects held a wait-and-see attitude due to concerns about the immature technology and high cost.

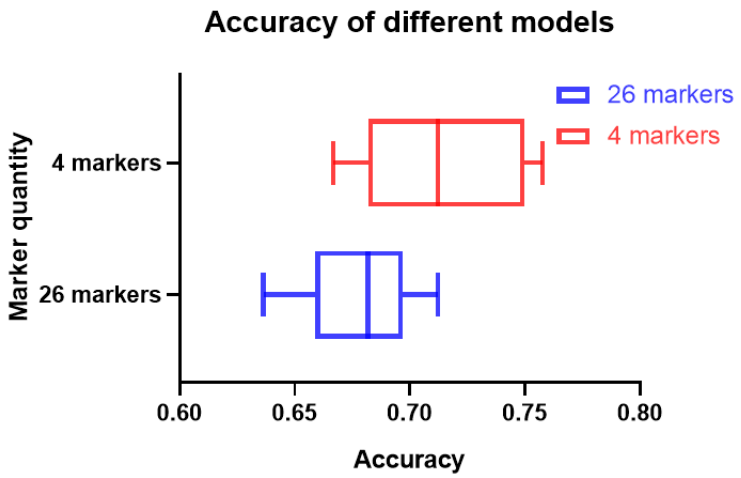

Figure 4: Accuracy comparison of models trained with 4 and 26 methylation markers. Using GraphPad Prism 8.0 to draw the box plot for comparison, the mean score of the 4-marker model was 71.52%, while the mean score of the 26-marker model was 67.88%. .

(3) Data security supervision needs to be updated. Regarding the issue of scientific research data disclosure, 84.35% of the samples were inclined to provide data to the hospital. Minority groups were reluctant to disclose personal data for fear of privacy leaks.

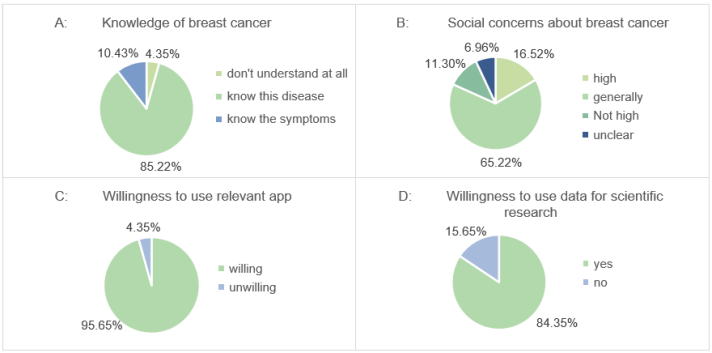

Figure 5: Pie charts of social survey results.

2.4 Discussion

2.4.1 Model optimisation scheme.

Since the model evaluation results we obtained did not meet our expectations, our research group summarised the following optimisation scheme:

(1) Increase data volume. The first and most important thing we need to do is increase the capacity of the database. Our technical limitations resulted in the small amount of available data, which may have contributed to inadequate model training. Expanding the capacity of the database and using more sufficient data for the machine learning model training may greatly improve our model.

(2) Find the best depth for the tree. For RFMs, the mastery of tree depth is of great significance to the entire programme, which will have a considerable impact on the accuracy of the model. Therefore, we need to further find the best depth of the tree to improve the precision of our diagnostic model.

(3) Parameter optimization. We may gain a great chance to improve the precision of our model by majorising our parameters, such as adjusting the number of feature attributes and

recursions, reducing the number of methylation markers involved in the evaluation, or normalising the scoring standard.

2.4.2 Application prospects of diagnostic assays based on cfDNA methylation levels.

At present, BC has developed into one of the most common cancers in the world. To accurately screen BC in time, detection methods such as mammography (MG), magnetic resonance imaging (MRI) and protein hybridization system (PHS) have been developed (He et al., 2020). However, the specificity of cfDNA methylation levels between patients with breast cancer and healthy individuals indicates the potential of incorporating analysis of cfDNA methylation levels into BC diagnostic approaches (Figure 6). On account of the fact that cfDNA testing is non-invasive, its application could lead to a more human health care system.

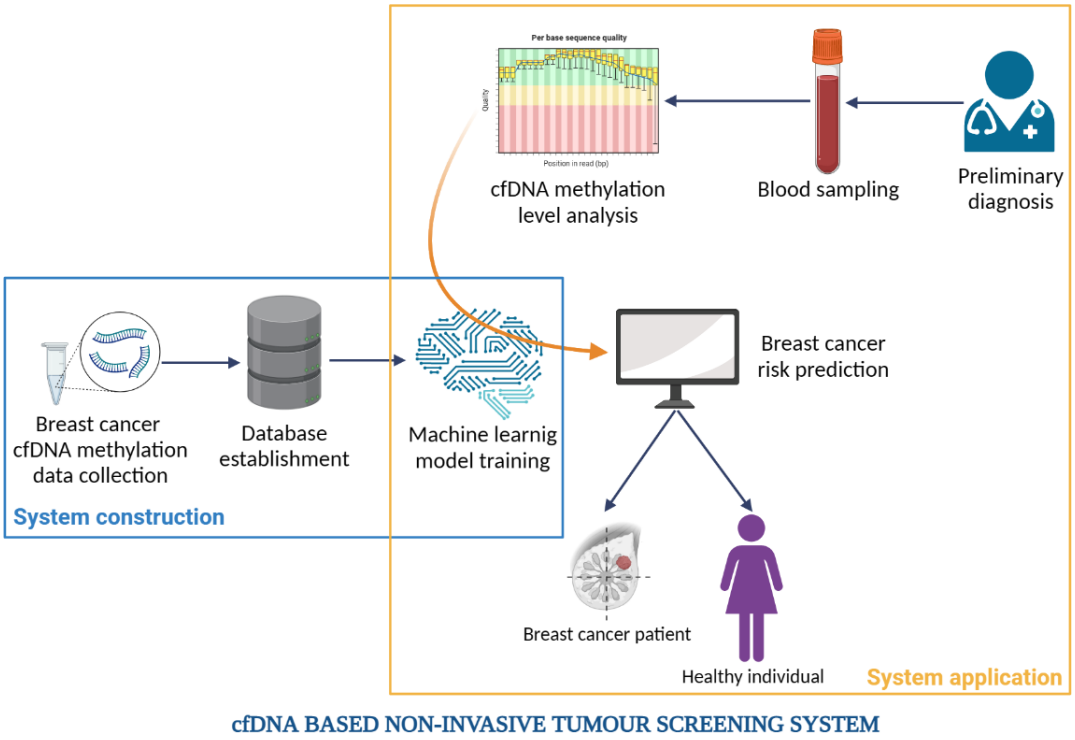

Figure 6: cfDNA based non-invasive tumour screening system. This figure presents a basic procedure of the cfDNA based non-invasive tumour screening system, including system construction and system application. First, a database can be established using the existing data and a feasible machine learning detection system can be constructed. Then, we use the detection system to perform cfDNA analysis. Finally, the system will output a diagnosis result.

2.4.3 The balance between the era of artificial intelligence and data management.

Judging from the collected data of the social investigation, we have reason to believe that using AI technology to realize non-invasive breast cancer screening has certain practical significance and good development prospects. In the age of data diversification, the emphasis on data security and regulatory rationality is conducive to more individuals sharing data for scientific research.

2.5 Limitations

2.5.1 Modelling limitations

As for the modelling section, since it was an initial exploration, there were several limitations.

(1) The data set is relatively small. If there were sufficient valid data, the model should perform better.

(2) In addition, the algorithm is not intuitive and comprehensive enough. Even if random forest analysis is suitable for the model, some other algorithms can be used in conjunction with it, which could possibly make the model become sophisticated and more relevant to the core element of the research.

(3) Apart from that, the framework of the random forest algorithm used in this model is from GitHub. Although some modifications and optimisations have been made to accommodate this study, the algorithm still has a lack of flexibility. A wealth of expertise and deep learning should be applied to the research in the following steps.

2.5.2 Social surveying limitations

In regard to the questionnaires section, the survey was conducted mainly among university students from four cities and six universities in China, which does not give a large representation of views. Besides, the questionnaire is pre-designed with a fixed range of responses, so that answers of respondents are restricted, and detailed information may be missed out.

2.6 Conclusion

In conclusion, a primary diagnostic prediction model for BC based on a combination of machine learning and cfDNA methylation was performed by our group. A preliminary model was built and tested through Python, and a number of data were collected to train it. Ultimately, the accuracy of the model was obtained after running the code, which was advanced by optimising the parameters afterwards. Furthermore, a variety of university students in China were surveyed to help us learn about the views and attitudes of AI applications in non-invasive BC assessment. Based on the results, the prospects of the project can be initially glimpsed.

3 Project skills

3.1 Individual project contribution

On the basis of a literature search, the project content and scope were determined jointly by all the team members (Figure 7). We studied the mechanism of cfDNA affecting cancer and the combination of cfDNA methylation and algorithm to realize tumor diagnosis and prediction.

3.2 Project management approach

We adopted the method of stage management, from the topic identification to the whole process of the completion of the project. According to the characteristics of the project, the whole project was divided into several small stages, and the content completed in each stage was discussed and summarised regularly. In the process of the project, we encountered problems such as database suitability selection, algorithm and code operation, Gantt chart production, and Overleaf use. Faced with these problems, we learned from different websites such as CSDN and GitHub, actively sought suggestions from students and professors in related majors, and solved the problems we encountered properly.

3.3 Build an efficient team

In order to build an efficient team, it is beneficial to formulate good cooperation guidelines, arrange various tasks reasonably based on mutual trust, combine the advantages of each member, and conduct timely and effective internal communication and feedback. During the project learning process, team members increased their understanding of the project by sharing relevant knowledge, useful skills, and vital information. In the process of project promotion, team members were able to take the initiative to work with each other and solve problems together. In the process of project implementation, during the progress of the

project, various plans needed to be adjusted when necessary. Team members needed to work together to relentlessly pursue the desired goals and results.

Task	Content	Responsible person
Article searching	Find background literature to support the project	All
cfDNA mechanism exploring	Learn about cfDNA mechanism to design the project framework	Gong Zixin, Guo Yifeiyang, Huang Menghan, Zheng Yiting
Data collecting	Get available data sets	Guo Yifeiyang
Algorithm adjusting	Modify the algorithm to meet our expectations	Gong Zixin, Huang Menghan, Zheng Yiting
Results analysing	Conduct analysis and comparison based on model run results	Gong Zixin, Zheng Yiting
Questionnaire designing	Design questionnaires to help understand the application prospects of the projects	Gan Haiting, Li Rui
Questionnaire distributing	Distribute the questionnaires online	All
Results collecting and analysing	Collate the results of the questionnaire and analysing them in relation to our research	Gan Haiting, Li Rui
Project summarising and report writing	Conclude our project and formulate a research report	All

Figure 7: Allocation of tasks

Author contribution: Conceptualisation: Z.Y.T., G.Z.X., G.Y.F.Y., H.M.H., G.H.T. and L.R.; Article Searching: All; cfDNA Mechanism Exploring: Z.Y.T., G.Z.X., G.Y.F.Y. and H.M.H.; Data Collecting: G.Y.F.Y.; Algorithm Adjusting: Z.Y.T., G.Z.X. and H.M.H.; Results Analysing: Z.Y.T. and G.Z.X.;Questionnaire Designing: G.H.T. and L.R.; Questionnaire Distributing: All; Questionnaire Results Collecting and Analysing: G.H.T. and L.R.; Project Summarising and Report Writing: All.All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study has followed the research guidelines of Biotechnology Engineering & Healthcare Technology, Cambridge Programme 2023. Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability: please contact the corresponding author(s) for all reasonable requests for access to the data.

Acknowledgments: We would like to give our thanks to our faculty professors and teachers at Cambridge University for their guidance.

Conflicts of Interest: The authors declare no conflict of interest.

Intellectual Property: The authors attest that copyright belongs to them, the article has not been published elsewhere, and there is no infringement of any intellectual property rights as far as they are aware.

References

Chabon, J. J., Hamilton, E. G., Kurtz, D. M., Esfahani, M. S., Moding, E. J., Stehr, H., & Diehn, M. (2020). Integrating genomic features for non-invasive early lung cancer detection. *Nature*, *580*(7802), 245-251.

Filippiadis, D. K., Charalampopoulos, G., Mazioti, A., Keramida, K., & Kelekis, A. (2018). Bone and soft-tissue biopsies: What you need to know. *Seminars in interventional radiology*, *35*(4), 215-220.

Liu, J., Zhao, H., Huang, Y., Xu, S., Zhou, Y., Zhang, W., ... Su, J. (2021). Genome-wide cell-free dna methylation analyses improve accuracy of non-invasive diagnostic imaging for early-stage breast cancer. *Molecular cancer*, *20*(1), 36.

Mannelli, C. (2019). Tissue vs liquid biopsies for cancer detection: Ethical issues. *Journal of Bioethical Inquiry*, *16*(4), 551-557.

Moss, J., Zick, A., Grinshpun, A., Carmon, E., Maoz, M., Ochana, B. L., & Dor, Y. (2020). Circulating breast-derived dna allows universal detection and monitoring of localized breast cancer. *Annals of oncology : official journal of the European Society for Medical Oncology*, *31*(3), 395-403.

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Ca: a cancer journal for clinicians. *Cancer statistics*, *71*(1), 7-33.

Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., & Shendure, J. (2016). Cell-free dna comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, *164*(1-2), 57-68.

Zhang, X., Zhao, D., Yin, Y., Yang, T., You, Z., Li, D., ... Pang, D. (2021). Circulating cell-free dna-based methylation patterns for breast cancer diagnosis. *NPJ breast cancer*, *7*(1), 106.

A Questionnaire

Questionnaire

Thank you very much for your participation in this survey. We hope to discover your knowledge of breast cancer and your willingness to use AI for breast cancer screening by filling in the questionnaire. Please fill in the questionnaire according to your actual situation. This questionnaire is conducted anonymously, and the contents you fill in will be kept strictly confidential. Please be sure to fill in your true information. Thank you sincerely for your support and cooperation!

1. Gender:
A. Male
B. Female
2. Grade:
A. Freshman
B. Sophomore
C. Junior
D. Senior
E. Postgraduate.
3. Do you think there is a high degree of concern about breast cancer in society at present?
A. High
B. Common
C. Not high
D. Unclear
4. What do you know about breast cancer?
A. I don't know about it at all
B. I have only heard about it
C. I know the specific symptoms
5. Through which channels do you know about breast cancer? (multiple choice):
A. Relatives and friends
B. TV broadcasting
C. Health education
D. Books
E. Hospitals
F. Others
6. What methods do you know about breast cancer screening? (multiple choice):
A. Mammography
B. Breast color Doppler ultrasound
C. Breast magnetic resonance examination
D. CT biopsy
E. Cell puncture
F. I don't know
7. If there is an AI technology that can realize breast cancer screening through simple blood collection, are you willing to participate in (or recommend family and friends to participate in) this project?
A. Yes
B. No
8. (Depending on the second option in question 7) What is the reason why you don't want to participate in it?
A. The technology is immature
B. The accuracy of the results is doubtful
C. The cost is too high
D. Privacy may be revealed
9. Do you think the accuracy of screening breast cancer with AI technology is high?
A. High
B. Common
C. Not high
D. Unclear
10. Do you prefer invasive testing (such as puncture) or non-invasive testing (such as

blood sampling)?

A. Invasive

B. Non-invasive

11. Would you like your inspection data to be used for scientific research?

A. Yes

B. No

12. (Depending on the first option of question 11) Which research institution would you prefer to hand over the data to?

A. Hospitals

B. Testing companies

13. (Depending on the second option of question 11) Reasons for unwillingness (multiple choices):

A. Unreliable supervision

B. Data involving privacy

C. Others