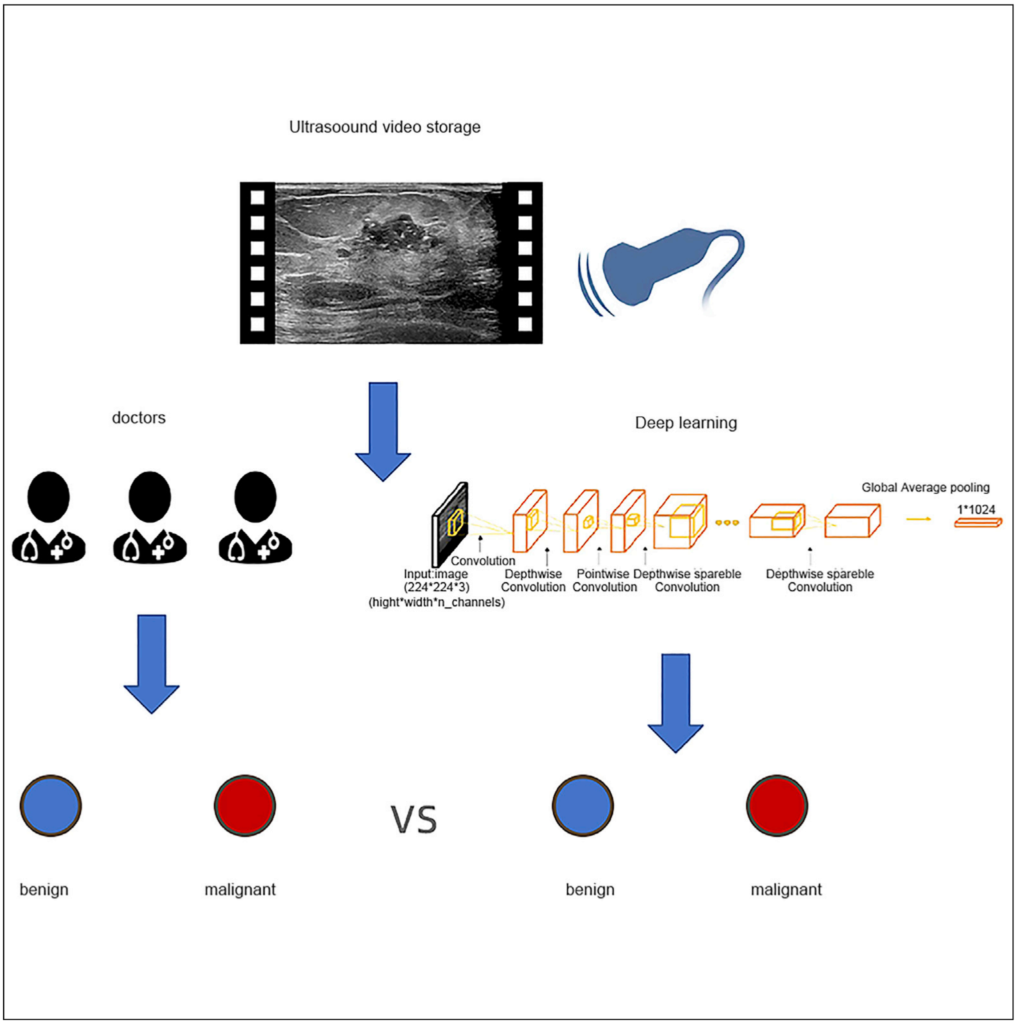
iScience

CellPress
OPEN ACCESS

Article

Feasibility of using AI to auto-catch responsible frames in ultrasound screening for breast cancer diagnosis



Jing Chen, Yitao Jiang, Keen Yang, ..., Jinfeng Xu, Dong Xu, Fajin Dong

xujinfeng@yahoo.com (J.X.)
xudong@zjcc.org.cn (D.X.)
dongfajin@szhospital.com (F.D.)

Highlights

Feature entropy reduction can capture complementary responsibility frames from video

Extracted responsibility frames can be realized by reducing feature entropy

Pre-trained imaged AI model can be screened for the best model to nodule's diagnose

Best AI model can diagnose breast nodules with accuracy up to a senior doctor

Chen et al., iScience *26*, 105692
January 20, 2023 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.105692

iScience

CellPress OPEN ACCESS

Article

Feasibility of using AI to auto-catch responsible frames in ultrasound screening for breast cancer diagnosis

Jing Chen,[1,5] Yitao Jiang,[2,5] Keen Yang,[1,5] Xiuqin Ye,[1] Chen Cui,[3] Siyuan Shi,[3] Huaiyu Wu,[1] Hongtian Tian,[1] Di Song,[1] Jincao Yao,[4] Liping Wang,[4] Sijing Huang,[1] Jinfeng Xu,[1,*] Dong Xu,[4,*] and Fajin Dong[1,6,*]

SUMMARY

The research of AI-assisted breast diagnosis has primarily been based on static images. It is unclear whether it represents the best diagnosis image.To explore the method of capturing complementary responsible frames from breast ultrasound screening by using artificial intelligence. We used feature entropy breast network (FEBrNet) to select responsible frames from breast ultrasound screenings and compared the diagnostic performance of AI models based on FEBrNet-recommended frames, physician-selected frames, 5-frame interval-selected frames, all frames of video, as well as that of ultrasound and mammography specialists. The AUROC of AI model based on FEBrNet-recommended frames outperformed other frame set based AI models, as well as ultrasound and mammography physicians, indicating that FEBrNet can reach level of medical specialists in frame selection. FEBrNet model can extract video responsible frames for breast nodule diagnosis, whose performance is equivalent to the doctors selected responsible frames.

INTRODUCTION

Breast cancer is now the most common malignancy in the world.[1–3] It is the leading cause of death among females.[4,5] Over 2.3 million new cases of breast cancer are diagnosed every year, according to the global cancer burden report for 2021,[3] accounting for 30% of cancers in women and 11.7% of all patients with cancer worldwide, ranking first in cancer and endangering the lives and health of females. Effective screening can detect breast cancer early, reduce the local and long-term recurrence rates, and improve the five-year survival rate.[6] Domestic and international guidelines recommend annual mammography for females aged 40 to 74 years, but the false-positive and false-negative rates for those with dense breasts are relatively high, making miss diagnoses more likely. As a supplementary diagnostic method, ultrasound is not limited by breast glandular tissue types and is especially appropriate for dense breast in Asian women, increasing the overall breast cancer detection rate by 17% and reducing unnecessary biopsies by 40%.[7,8] Thus, breast ultrasound, along with mammography, clinical examination, and needle biopsy, plays an important role in the evaluation of breast disease. However, diagnosis based on breast ultrasound images heavily depends on the sonographer's experience.

Artificial intelligence (AI) has developed rapidly in recent years. AI can provide individualized analysis and assist doctors in clinical decision-making.[9–11] As a branch of AI, deep learning is able to recognize patterns in images without human intervention, facilitating high-throughput correlation between images and clinical data.[12] Various AI techniques have been used to extract useful information from images, including convolutional neural networks (CNNs) and variable autoencoders.[13–15] In the field of medical imaging, AI has made the most significant contributions to imaging medicine in disease severity classification and argan segmentation (or separation). Liver ultrasound images for fatty liver detection,[16,17] ultrasound images for ovarian cancer detection and risk stratification,[18,19] and chaotic ultrasound for stroke risk stratification are examples of disease classification applications.[20]

AI research of breast nodules has mostly relied on static images selected by doctors during scanning. Due to the operator dependence of ultrasound scans, doctors at various expertise levels may have different interpretations of the image and disease judgment. Thus, the responsible frames selected for

[1]Department of Ultrasound, Shenzhen People's Hospital (The Second Clinical School of Medicine, Jinan University; The First Affiliated Hospital of Southern University of Science and Technology), Shenzhen, Guangdong 518020, China

[2]Research and Development Department, Microport Prophecy, Shanghai 201203, China

[3]Research and Development Department, Illuminate, LLC, Shenzhen, Guangdong 518000, China

[4]The Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), Institute of Basic Medicine and Cancer (IBMC), Chinese Academy of Sciences, Hangzhou, Zhejiang 310022, China

[5]These authors contributed equally

[6]Lead contact

*Correspondence: xujinfeng@yahoo.com (J.X.), xudong@zjcc.org.cn (D.X.), dongfajin@szhospital.com (F.D.)

https://doi.org/10.1016/j.isci.2022.105692

decision-making can have inter-observer variation and junior physicians or non-professionals who lack necessary knowledge of breast ultrasound cannot choose a set of responsible frames for diagnosis. Since videos contain complete lesion information, whole videos can be used as inputs for AI without manual frame selection.[21–23] It can broaden the scope of AI applications, especially in assisting less experienced operators that conduct breast diagnosis. However, ultrasound screening is a dynamic process, which records not only the important frames containing nodules but also the frames showing irrelevant tissue and ultrasound artifacts that can mislead AI. AI performance can be improved, if it can self-identify a set of representative frames to describe the whole lesion, just like the frames senior physicians selected.

Currently, video classification models primarily employed the fixed interval time-based method to subsample the frames. Although it reduced the number of frames for analysis, it fails to solve the existence of noisy frames and the problem of repeatedly choosing frames sharing very similar features. To address the issues, we combine greedy algorithm and the idea of information entropy to propose a model called FEBrNet for responsible frame selection. Information entropy (IE) is the average amount of the information contained in each "message" received in information field.[24] Video is considered a collection of images, and each image's information can be projected to multiple feature dimensions. Greedy algorithm is used to select a set of frames, which best summarizes the feature distribution of the whole video. Meanwhile, IE minimizing method is applied to decide when to stop picking a new responsible frame. In this study, we compared the performance of AI based on FEBrNet-recommended frames (R Frames) with AI based on all frames of video (All Frames), AI based on physician-selected images (Phy Images), AI based on fixed interval sampled frames (Fix Frames), ultrasound specialists, and mammography specialists. Area under receiver operating characteristics curve (AUROC), sensitivity, specificity, and accuracy are used as primary evaluating matrices. In 3-fold cross-validation and testing, AI can reach an equivalent performance using FEBrNet-recommended frames as using physician-selected frames and outperforming both ultrasound and mammography specialists.

RESULTS

Comparison of clinical data

A total of 974 patients with breast nodules were included in this study. They were randomly assigned according to the training set and testing set with the ratio of 4:6 since random forest only requires limited training samples. Specifically, the training set contained 387 cases (174 malignant nodules), and the independent testing set contained 587 cases (238 malignant nodules). All patients underwent biopsy or surgery and obtained pathological diagnoses (Figure 1). There was no significant difference in age and nodule size between the patients included ($p > 0.05$).

Comparison of the effectiveness of FEBrNet's responsible frame and others in training set by 3-fold cross-validation

The AUC (0.901 [95% CI: 0.877–0.925]) of FEBrNet's responsible frame method is higher than that of the doctor frame selection method, fixed interval frame selection method, and all frames of video, $p < 0.05$. The cutoff value is 0.402, sensitivity: 84.5%, specificity: 80.6%, and accuracy: 82.3% (Table 1, Figure 2).

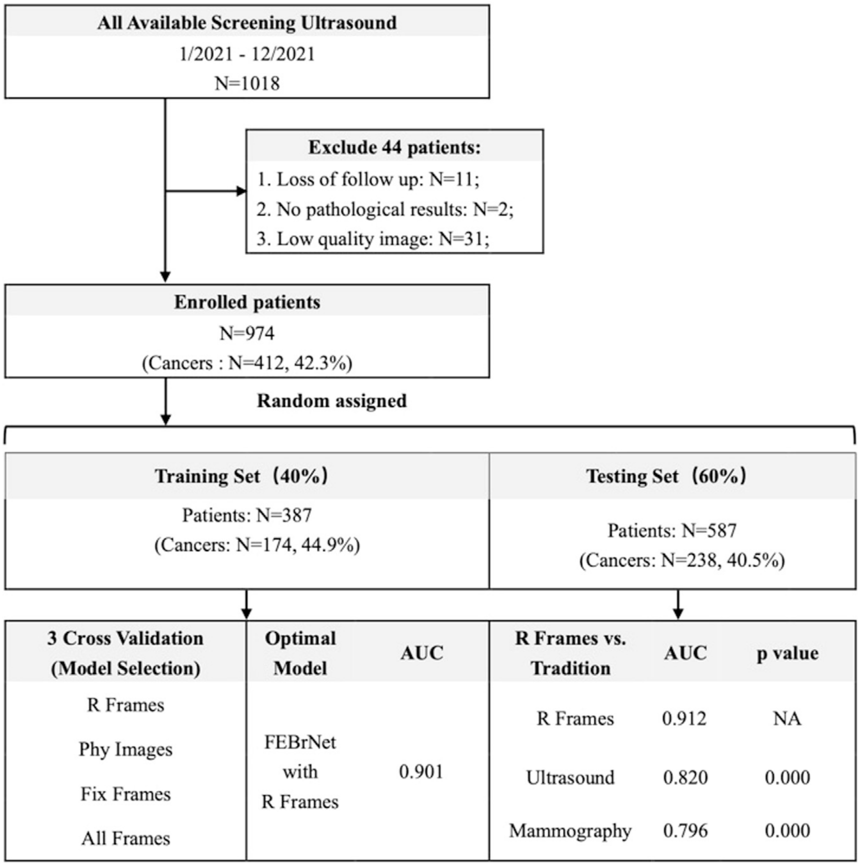
Comparison of the effectiveness of FEBrNet's responsible frame and others AI models in the independent testing set

(1) The AUC (0.912 [95% CI: 0.888–0.936]) of the responsible frame method was higher than that of the fixed interval frame selection and all video frames ($p < 0.05$). The cutoff value was 0.416, sensitivity: 84.4%, specificity: 87.4%, and accuracy: 86.2%.

(2) The AUC of the responsible frame method is slightly higher than that of the model based on the frame selected by doctors (0.909 [95% CI: 0.884–0.935]) ($p = 0.715 > 0.05$), indicating that the frame selection level of AI model can reach the senior experts (Figures 3A–3D).

Comparison of the effectiveness of FEBrNet's responsible frame and ultrasound and mammography specialists in the independent testing set

In this section, we reviewed the 383 cases where both ultrasound and mammography could be found.



3 Cross Validation (Model Selection)	Optimal Model	AUC
R Frames Phy Images Fix Frames All Frames	FEBrNet with R Frames	0.901

R Frames vs. Tradition	AUC	p value
R Frames	0.912	NA
Ultrasound	0.820	0.000
Mammography	0.796	0.000

Figure 1. Flow chart and statistical results

Note: R Frames: responsible frame; Phy Images: doctor frame selection; Fix Frames: fixed interval frame selection; All Frames: all frames of video; Ultrasound: Diagnosis by senior ultrasound doctors; Mammography: Diagnosis by senior mammography doctors; p: R Frames vs. others; AUC: area under the curve; NA: Not Applicable.

(1) The AUC of responsible frame method was better than that diagnosed by senior ultrasound doctors. The AUC was (0.912 [95% CI: 0.888–0.936] vs. 0.820 [95% CI: 0.788–0.853]) ($p < 0.05$).

(2) The AUC diagnosed by FEBrNet model is better than that diagnosed by senior mammography doctors, and the AUC is (0.912 [95% CI: 0.888–0.936] vs. 0.796 [95% CI: 0.759–0.833]) ($p < 0.05$) (Figures 3E, 3F, and Table 2).

DISCUSSION

Based on our analysis of breast nodule static ultrasound images, a new model was constructed for this study: FEBrNet, which is an AI method for automatically capturing responsible frames in breast ultrasound videos.

Table 1. Statistics of diagnostic efficiency of 3-fold cross-validation in training set

Modality	AUC (95%CI)	Cut-off	Sensitivity (%)	Specificity (%)	Accuracy (%)	p value
R_Frames	0.901 (0.877–0.925)	0.402	84.5	80.6	82.3	NA
Phy_Images	0.857 (0.825–0.889)	0.411	71.8	87.2	80.3	0.009
Fix_Frames	0.815 (0.793–0.837)	0.401	78.7	73.0	75.6	0.000
All_Frames	0.889 (0.864–0.914)	0.521	75.3	89.1	82.9	0.002

Note: AUC: area under the curve; 95% CI: 95% confidence interval; R_Frames: responsible frame; Phy_Images: doctor frame selection; Fix_Frames: fixed interval frame selection; All_Frames: all frames of video; p: R_Frames vs. others; NA: Not Applicable.

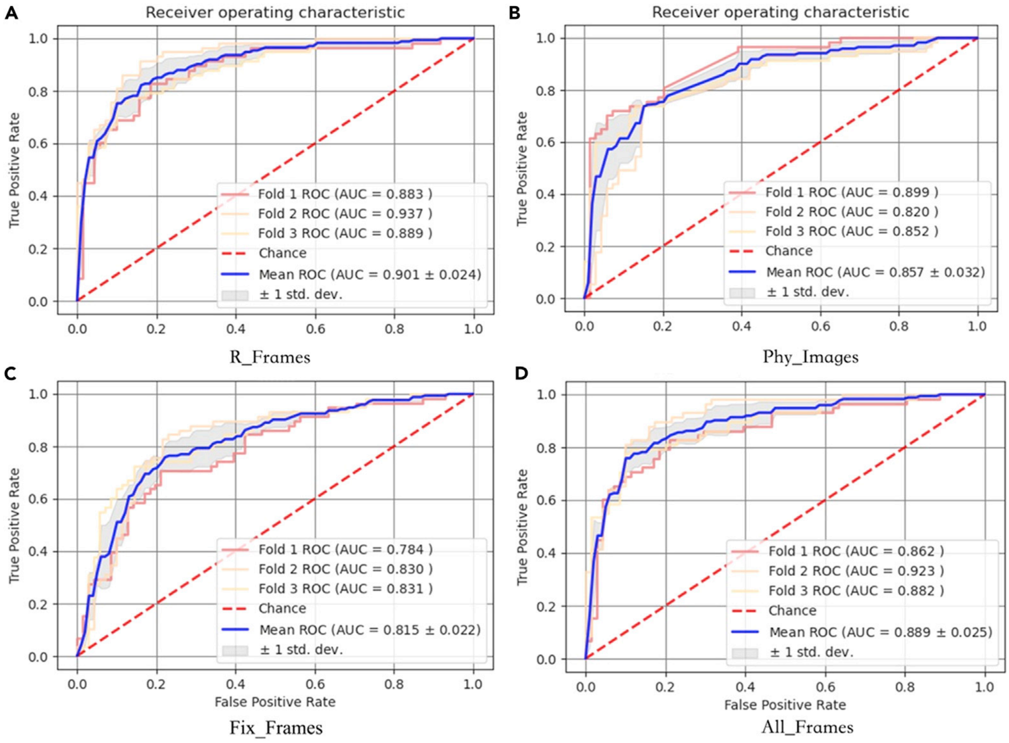


Figure 2. 3-fold cross-validation results of training set

Note: (A) R_Frames: responsible frame; (B) Phy_Images: doctor frame selection; (C) Fix_Frames: fixed interval frame selection; (D) All_Frames: all frames of video.

In this study, FEBrNet was developed to analyze and extract responsible frames from ultrasound video data via defining the feature score (FS) and calculating entropy without adding additional training parameters, self-extracting multiple responsible frames with complementary information for the differential diagnosis of benign and malignant breast nodules. The results show that the AUC (0.912) of the AI model based on FEBrNet's frames is higher than that of the fixed interval frames (0.852) and all video frames (0.881), indicating that the method proposed in this paper is feasible and highly reliable to assist breast diagnosis.

As one major contribution of this work, the FEBrNet addressed the issue of choosing responsible frames while avoiding selecting visually identical frames repetitively. As shown in Figure 4, we use a video taken from a 45-year-old female patient with BIRADS 4c and pathologically confirmed invasive breast cancer as a simple example to demonstrate the capacity of FEBrNet. When frames are sorted by FScore, the top three frames are frame 26, 39, and 27, with FScore of 21.94, 21.29, and 21.03. These frames seem to be relatively similar in Figure 4A and very close in time sequence. After using principal component analysis to compress and display the feature matrices into two dimensions in Figure 4B, it is obvious that the distance in feature dimensions between the three frames is rather close. In Figures 4C and 4D, the same approach is used to assess the top three responsible frames (frame 26, 111, and 96) chosen by FEBrNet. While the FScore of frame 111 is low, it is considerably distant from the first responsible frame (frame 26) in the feature dimensions. The top three frames selected by the FEBrNet are scattered, echoing their various visual attributes in Figure 4D.

Internationally recognized standard BIRADS classification for breast cancer risk is based on the characteristics of malignancy displayed in images.[25,26] In this paper, the diagnostic efficiency of FEBrNet exceeds that of senior ultrasound and mammography doctors, which proves the reliable value of this technology in diagnosing breast cancer. The performance of AI algorithms depends on the training of large, labeled datasets.[27–29] At present, a common problem is that it is difficult to find sufficient training data in the face of specific problems in a certain field. The emergence of transfer learning can alleviate the problem of insufficient data sources.[29,30] Several CNN network models launched in recent years, such as Alexnet,[31] VGG,[32]

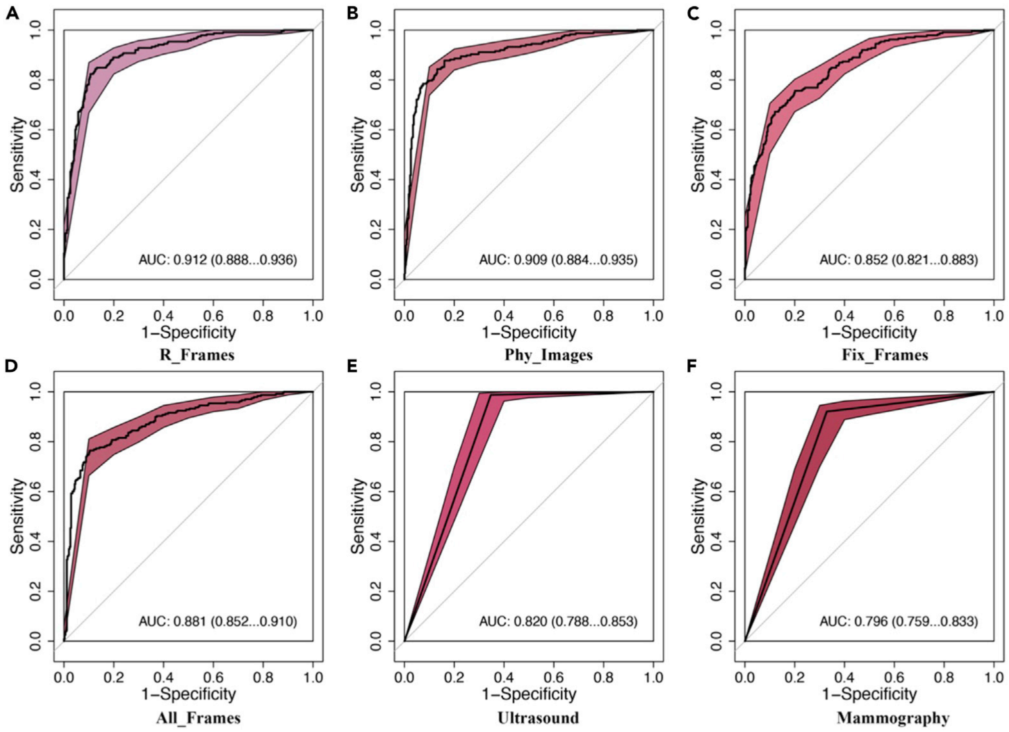


Figure 3. Comparison of the effectiveness of FEBrNet's responsible frame and others in the independent testing set

Note: AUC:area under the curve, 95% CI: 95% confidence Interval; (A) R_Frames: responsible frame; (B) Phy_Images: doctor frame selection; (C) Fix_Frames: fixed interval frame selection; (D) All_Frames: all frames of video; (E)Ultrasound: Diagnosis by senior ultrasound specialists; (F) Mammography: Diagnosis by senior mammography specialists.

GoogLeNet,[27] ResNet,[33] and Inception,[34,35] can solve the visual classification task in natural images. This study used the static image pre-training and screening model and then inherits the weight of its backbone and full connection layer through the method of transfer learning (Figure 5). FEBrNet model was constructed to process video data and automatically extract responsible frames from video data and make the diagnosis (Figures 6 and 7); the results show that the model directly predicts the video responsible frame, which is very close to the AUC of the doctor's frame selection method (0.909 [95% CI: 0.884–0.935]) ($p = 0.715 > 0.05$), which shows that AI model can reach the level of senior experts in frame selection. Mammography screening can detect early breast cancer and reduce breast cancer mortality by more than 30% in women over 40 years of age.[36] Although mammography is the gold standard for breast cancer screening[37] and has been proven to reduce mortality, the accuracy of results suffers among women with dense breasts.[38] In the independent testing set, the AUC of the FEBrNet model responsible frame method was better than that of senior ultrasound and mammography doctors ($p < 0.05$).

Table 2. Statistics of diagnosis efficiency of FEBrNet-recommended frames and other frame sets

Modality	AUC (95%CI)	Cut-off	Sensitivity (%)	Specificity (%)	Accuracy (%)	p value
R_Frames	0.912 (0.888–0.936)	0.416	84.4	87.4	86.2	NA
Phy_Images	0.909 (0.884–0.935)	0.401	87.8	84.0	85.3	0.715
Fix_Frames	0.852 (0.821–0.883)	0.490	75.6	79.9	78.2	0.000
All_Frames	0.881 (0.852–0.910)	0.436	76.4	89.7	84.3	0.002
Ultrasound	0.820 (0.788–0.853)	NA	98.7	65.3	79.6	0.000
Mammography	0.796 (0.759–0.833)	NA	92.1	67.1	77.8	0.000

Note: AUC: area under the curve; 95% CI: 95% confidence interval; R_Frames: responsible frame; Phy_Images: doctor frame selection; Fix_Frames: fixed interval frame selection; All_Frames: all frames of video; Ultrasound: Diagnosis by senior ultrasound doctors; Mammography: Diagnosis by senior mammography doctors; p: R Frames vs. others; NA: Not Applicable.

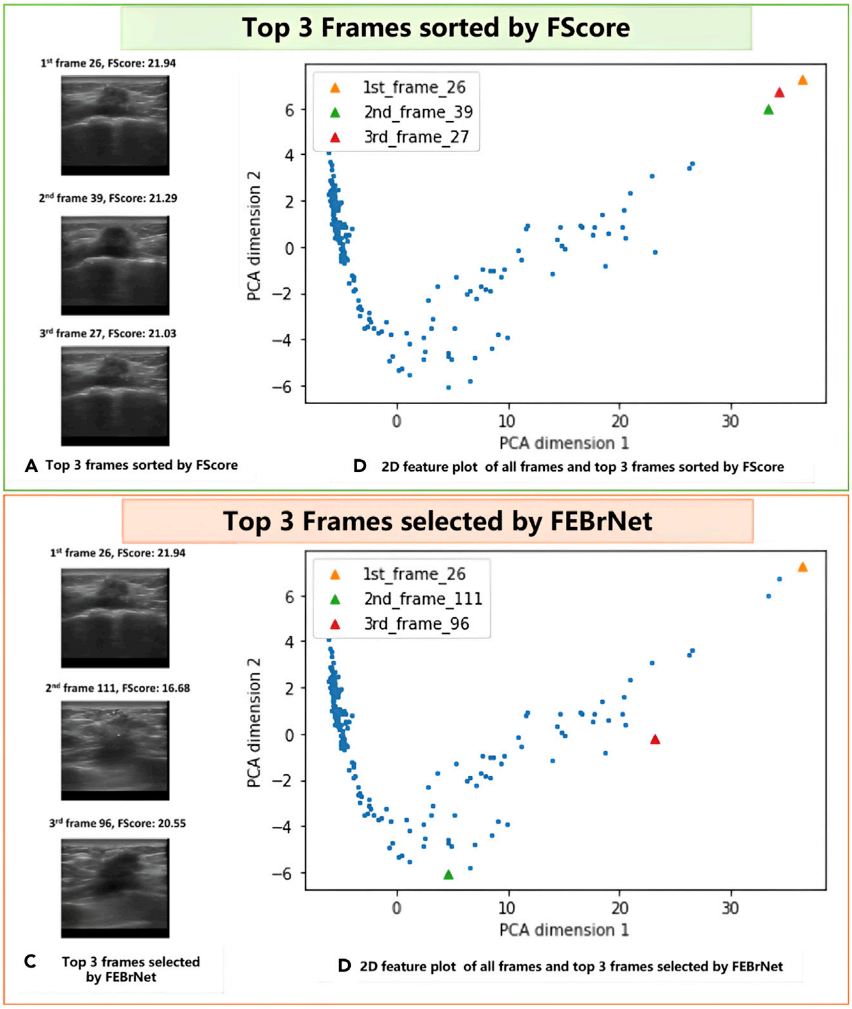


Figure 4. A case study of responsible frames selected by FEBrNet

Note: (A) and (B) The top 3 frames sorted by FScore are relatively similar in time sequence, visually identical, and contribute comparable characteristics; (C) and (D) The top 3 frames chosen by Entropy Reduce method show more diverse image characteristics and scattered on 2D feature plot.

Limitations of the study

(1) The amount of breast ultrasound data in this study is limited. In the future, multi-center clinical research and video data of various instruments should be carried out to improve the effectiveness and robustness of training.

(2) This study only focuses on the classification of benign and malignant breast nodules. In the future, it should be expanded to include image tasks such as infiltration range prediction, subtype classification, molecular phenotype prediction, and distant metastasis prediction.

(3) The FEBrNet architecture can be used for ultrasound screening of various diseases, and more diseases should be evaluated in the future.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY

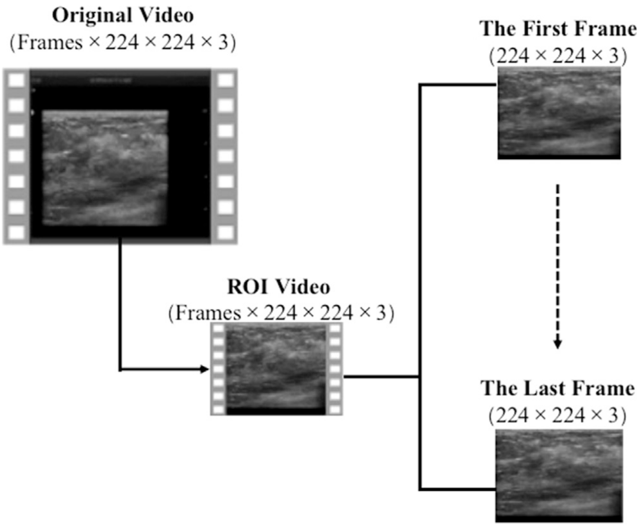


Figure 5. Ultrasound video preprocessing

- ○ Lead contact
- ○ Materials availability
- ○ Data and code availability
- ● EXPERIMENTAL MODEL AND SUBJECT DETAILS
- ○ Ethic statement
- ○ Inclusion criteria

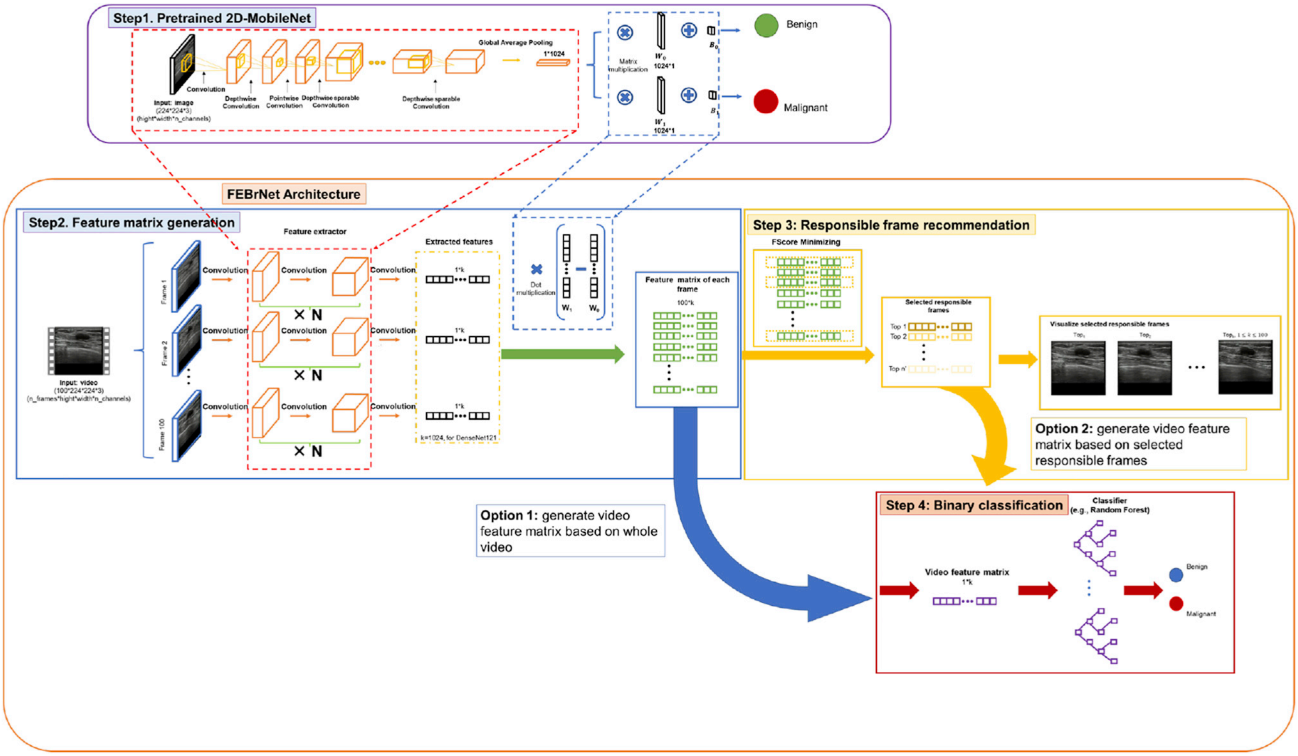


Figure 6. FEBrNet model structure

Note: The backbone model is a feature extractor and weights of fully connected layer from pre-trained ultrasound image dataset filtered MobileNet_224. FEBrNet uses a feature extractor and weights from a pre-trained fully connected layer to generate a feature matrix that identifies responsible frames and makes diagnostic predictions.

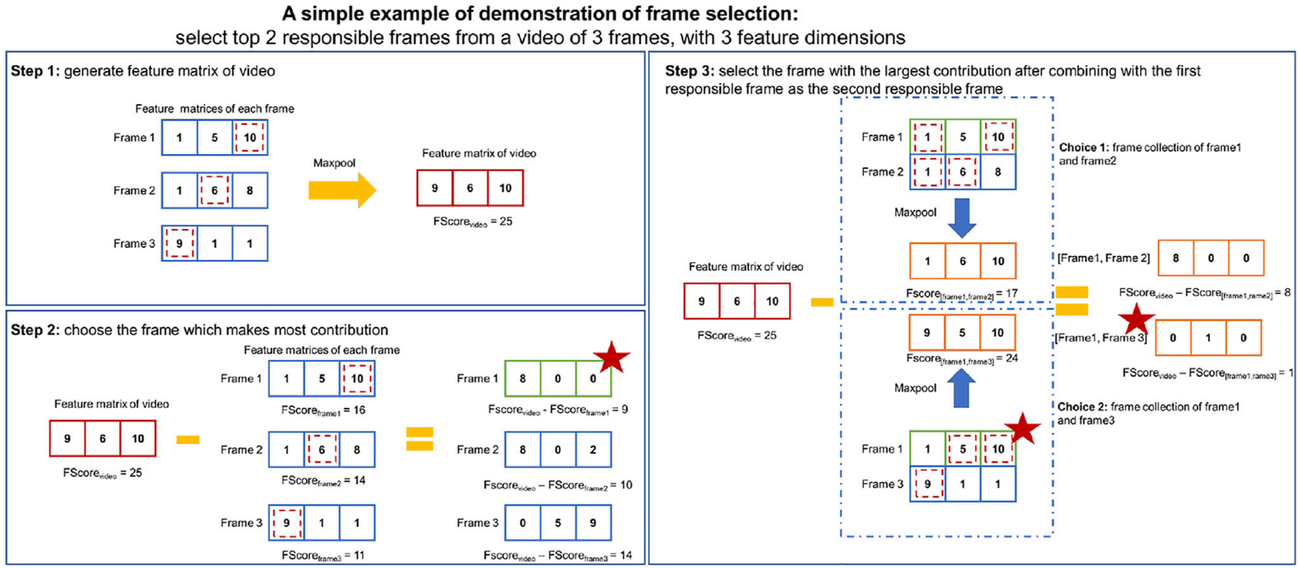

A simple example of demonstration of frame selection:
select top 2 responsible frames from a video of 3 frames, with 3 feature dimensions
Step 1: generate feature matrix of video
Feature matrices of each frame
Frame 1
1 5 10
Frame 2
1 6 8
Frame 3
9 1 1
Maxpool
Feature matrix of video
9 6 10
FScore$_{video}$ = 25
Step 2: choose the frame which makes most contribution
Feature matrix of video
9 6 10
FScore$_{video}$ = 25
Feature matrices of each frame
Frame 1
1 5 10
FScore$_{frame1}$ = 16
Frame 2
1 6 8
FScore$_{frame2}$ = 14
Frame 3
9 1 1
FScore$_{frame3}$ = 11
Frame 1
8 0 0
Fscore$_{video}$ - FScore$_{frame1}$ = 9
Frame 2
8 0 2
Fscore$_{video}$ – FScore$_{frame2}$ = 10
Frame 3
0 5 9
Fscore$_{video}$ – FScore$_{frame3}$ = 14
Step 3: select the frame with the largest contribution after combining with the first responsible frame as the second responsible frame
Feature matrix of video
9 6 10
FScore$_{video}$ = 25
Frame 1
1 5 10
Frame 2
1 6 8
Maxpool
1 6 10
Fscore$_{[frame1,frame2]}$ = 17
Choice 1: frame collection of frame1 and frame2
[Frame1, Frame 2]
8 0 0
FScore$_{video}$ – FScore$_{[frame1,rame2]}$ = 8
9 5 10
FScore$_{[frame1,frame3]}$ = 24
Maxpool
Frame 1
1 5 10
Frame 3
9 1 1
[Frame1, Frame 3]
0 1 0
FScore$_{video}$ – FScore$_{[framo1,ramo3]}$ = 1
Choice 2: frame collection of frame1 and frame3

Figure 7. An example of selecting responsible frames

Step 1: MaxPool the three frame feature matrices and yield the video feature matrix. Step 2: choose frame1 as the first responsible frame for it minimizing the difference between *FScore$_{video}$* and *FScore$_{framei}$*. Step 3: add another frame to the responsible frame collection, whereas step 2 already chose one. Since the *FScore* difference between the responsible frame collections of frame1 and frame3 is the smallest, frame3 is picked as the second responsible frame. Example Figure 1 Schematic diagram of automatic stop of responsible frame selection. Note: RF_1: The first responsibility frame, and so on; When the model selects the video responsibility frame, it gradually rises from the initial malignant prediction value to RF_9 (0.93), from RF_10 starts to decline gradually, so the model is selected to RF_9 stop.

ACKNOWLEDGMENTS

This project was supported by Commission of Science and Technology of Shenzhen (GJHZ20200731095401004). Our D.team provided all the data for the experiments. Thanks for our D.team of Ultrasound Department of Shenzhen People's Hospital.

AUTHOR CONTRIBUTIONS

J.X.: Conceptualization, Methodology, Investigation, Funding acquisition. D.X.: Conceptualization, Methodology, Investigation, Supervision. F.D.: Conceptualization, Methodology, Investigation, Analyzed the data, Supervision. J.C.: Methodology, Writing – original draft, Writing – review & editing, Analyzed the data. Y.J.: Software, Writing – review & editing, Analyzed the data, Methodology. K.Y.: Resources, Analyzed the data, Methodology. X.Y.: Analyzed the data, Resources, Analyzed the data. S.S.: Software, Supervision, Analyzed the data. C.C.: Software, Supervision, Analyzed the data, Funding acquisition. H.W., H.T., D.S., J.Y., and S.H.: Analyzed the data, Resources.

DECLARATION OF INTERESTS

The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: August 8, 2022
Revised: October 31, 2022
Accepted: November 26, 2022
Published: January 20, 2023

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA. Cancer J. Clin. *71*, 209–249.

2. Siegel, R.L., Miller, K.D., Fuchs, H.E., and Jemal, A. (2021). Cancer statistics, 2021. CA. Cancer J. Clin. *71*, 7–33.

3. Siegel, R.L., Miller, K.D., Fuchs, H.E., and Jemal, A. (2022). Cancer statistics, 2022. CA. Cancer J. Clin. *72*, 7–33.

4. Maliszewska, M., Maciążek-Jurczyk, M., Pożycka, J., Szkudlarek, A., Chudzik, M., and Sułkowska, A. (2016). Fluorometric investigation on the binding of letrozole and resveratrol with serum albumin. Protein Pept. Lett. *23*, 867–877.

5. Chen, W., Zheng, R., Zhang, S., Zeng, H., Xia, C., Zuo, T., Yang, Z., Zou, X., and He, J. (2017). Cancer incidence and mortality in China, 2013. Cancer Lett. *401*, 63–71.

6. Cedolini, C., Bertozzi, S., Londero, A.P., Bernardi, S., Seriau, L., Concina, S., Cattin, F., and Risaliti, A. (2014). Type of breast cancer diagnosis, screening, and survival. Clin. Breast Cancer *14*, 235–240.

7. Niell, B.L., Freer, P.E., Weinfurtner, R.J., Arleo, E.K., and Drukteinis, J.S. (2017). Screening for breast cancer. Radiol. Clin. North Am. *55*, 1145–1162.

8. Osako, T., Takahashi, K., Iwase, T., Iijima, K., Miyagi, Y., Nishimura, S., Tada, K., Makita, M., Akiyama, F., Sakamoto, G., and Kasumi, F. (2007). Diagnostic ultrasonography and mammography for invasive and noninvasive breast cancer in women aged 30 to 39 years. Breast Cancer *14*, 229–233.

9. Tonekaboni, S., Joshi, S., McCradden, M.D., and Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. Preprint at axRiv. https://doi.org/10.48550/arXiv.1905.05134.

10. Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digit. Med. *1*, 18.

11. Suzuki, K. (2017). Overview of deep learning in medical imaging. Radiol. Phys. Technol. *10*, 257–273.

12. Yasaka, K., Akai, H., Kunimatsu, A., Kiryu, S., and Abe, O. (2018). Deep learning with convolutional neural network in radiology. Jpn. J. Radiol. *36*, 257–272.

13. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., and Aerts, H.J.W.L. (2018). Artificial intelligence in radiology. Nat. Rev. Cancer *18*, 500–510.

14. Mazurowski, M.A., Buda, M., Saha, A., and Bashir, M.R. (2019). Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. J. Magn. Reson. Imaging. *49*, 939–954.

15. He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X., and Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. Nat. Med. *25*, 30–36.

16. Acharya, U.R., Sree, S.V., Ribeiro, R., Krishnamurthi, G., Marinho, R.T., Sanches, J., and Suri, J.S. (2012). Data mining framework for fatty liver disease classification in ultrasound: a hybrid feature extraction paradigm. Med. Phys. *39*, 4255–4264.

17. Saba, L., Dey, N., Ashour, A.S., Samanta, S., Nath, S.S., Chakraborty, S., Sanches, J., Kumar, D., Marinho, R., and Suri, J.S. (2016). Automated stratification of liver disease in ultrasound: an online accurate feature classification paradigm. Comput. Methods Programs Biomed. *130*, 118–134.

18. Acharya, U.R., Sree, S.V., Krishnan, M.M.R., Saba, L., Molinari, F., Guerriero, S., and Suri, J.S. (2012). Ovarian tumor characterization using 3D ultrasound. Technol. Cancer Res. Treat. *11*, 543–552.

19. Acharya, U.R., Sree, S.V., Kulshreshtha, S., Molinari, F., En Wei Koh, J., Saba, L., and Suri, J.S. (2014). GyneScan: an improved online paradigm for screening of ovarian cancer via tissue characterization. Technol. Cancer Res. Treat. *13*, 529–539.

20. Acharya, R.U., Faust, O., Alvin, A.P.C., Sree, S.V., Molinari, F., Saba, L., Nicolaides, A., and Suri, J.S. (2012). Symptomatic vs. asymptomatic plaque classification in carotid ultrasound. J. Med. Syst. *36*, 1861–1871.

21. Furht, B., Smoliar, S.W., and Zhang, H. (2012). Video and Image Processing in Multimedia Systems (Springer Science & Business Media), p. 326.

22. Schmidhuber, J. (2015). Deep learning in neural networks: an overview. Neural Netw. *61*, 85–117.

23. Golden, J.A. (2017). Deep learning algorithms for detection of lymph node metastases from breast cancer: helping artificial intelligence be seen. JAMA *318*, 2184–2186.

24. Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. *27*, 379–423.

25. Drukker, L., Noble, J.A., Papageorghiou, A.T., et al. (2020). Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. Ultrasound Obstet. Gynecol. *56*, 498–505.

26. Muse, E.D., and Topol, E.J. (2020). Guiding ultrasound image capture with artificial intelligence. Lancet *396*, 749.

27. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Boston, USA: IEEE).

28. Wu, S., Zhong, S., and Liu, Y. (2018). Deep residual learning for image steganalysis. Multimed. Tools Appl. *77*, 10437–10453.

29. Aneja, N., and Aneja, S. (2019). Transfer Learning Using CNN for Handwritten Devanagari Character Recognition, 2019 1st International Conference on Advances in Information Technology (ICAIT) (IEEE), pp. 293–296.

30. Torrey, L., and Shavlik, J. (2010). Transfer learning, Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI global, 242–264.

31. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Commun. ACM *60*, 84–90.

32. Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.1409.1556.

33. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.1512.03385.

34. Xia, X., Xu, C., and Nan, B. (2017). Inception-v3 for Flower Classification, 2017 2nd International Conference on Image, Vision and Computing (ICIVC) (IEEE), pp. 783–787.

35. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In

Proceedings of the IEEE conference on computer vision and pattern recognition (IEEE), pp. 2818–2826.

36. Tohno, E., Umemoto, T., Sasaki, K., Morishima, I., and Ueno, E. (2013). Effect of adding screening ultrasonography to screening mammography on patient recall and cancer detection rates: a retrospective study in Japan. Eur. J. Radiol. *82*, 1227–1230.

37. Le-Petross, H.T., and Shetty, M.K. (2011). Magnetic resonance imaging and breast ultrasonography as an adjunct to mammographic screening in high-risk patients. Semin. Ultrasound CT MRI *32*, 266–272.

38. Ohnuki, K., Tohno, E., Tsunoda, H., Uematsu, T., and Nakajima, Y. (2021). Overall assessment system of combined mammography and ultrasound for breast cancer screening in Japan. Breast Cancer *28*, 254–262.

39. van Rossum, G. (1995). Python Reference Manual (Department of Computer Science [CS]).

40. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., and Kudlur, M. (2016). {TensorFlow}: a system for {Large-Scale} machine learning. In 12th USENIX symposium on operating systems design and implementation. OSDI *16*, 265–283.

41. Hunter. (2007). Matplotlib: A 2D Graphics Environment. In Computing in Science and Engineering, *9* (IEEE), pp. 90–95.

42. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362.

43. Jiang, Y. (2022). Code for FEBrNet (Mendeley Data). https://doi.org/10.17632/wyjy6pr445.1.

44. Wu, H., Ye, X., Jiang, Y., Tian, H., Yang, K., Cui, C., Shi, S., Liu, Y., Huang, S., Chen, J., et al. (2022). A comparative study of multiple deep learning models based on multi-input resolution for breast ultrasound images. Front. Oncol. *12*, 869421.

45. Muhammad, W., Aramvith, S., and Onoye, T. (2021). Multi-scale Xception based depthwise separable convolution for single image super-resolution. PLoS One *16*, e0249278.

46. Wang, W., Hu, Y., Zou, T., Liu, H., Wang, J., and Wang, X. (2020). A new image classification approach via improved MobileNet models with local receptive field expansion in shallow layers. Comput. Intell. Neurosci. *2020*, 8817849.

47. Zhou, J., Zhang, Y., Chang, K.T., Lee, K.E., Wang, O., Li, J., Lin, Y., Pan, Z., Chang, P., Chow, D., et al. (2020). Diagnosis of benign and malignant breast lesions on DCE-MRI by using radiomics and deep learning with consideration of peritumor tissue. J. Magn. Reson. Imaging. *51*, 798–809.

48. Ezzat, D., Hassanien, A.E., and Ella, H.A. (2021). An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization. Appl. Soft Comput. *98*, 106742.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python	van Rossum,[39]	https://www.python.org/
Tensorflow	Abadi et al.[40]	https://www.tensorflow.org/
OpenCV	OpenCV team	https://opencv.org/
Matplotlib	Hunter,[41]	https://matplotlib.org/
NumPy	Harris et al.[42]	https://numpy.org/
Code for this paper	Jiang,[43]	Code for FEBrNet - Mendeley Data

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and should be directed to and will be fulfilled by the lead contact, FaJin Dong(dongfajin@szhospital.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Data is not publicly shared but available upon reasonable request from the lead contact.

Code of FEBrNet is available on Code for FEBrNet - Mendeley Data. Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Ethic statement

This study retrospectively analyzed the female patients between the ages of 24 and 75 who underwent breast ultrasound screening at Shenzhen People's Hospital from July 2015 to December 2021. The study was approved by the Ethics Committee of Shenzhen People's Hospital, and the requirement for informed consent was waived because of its retrospective design.

Inclusion criteria

(1)Breast nodules were detected by ultrasound, which were classified as 0, 3, 4a, 4b, 4c or 5 according to BIRADS; (2) At least 3.0 mm breast tissue can be displayed around the nodule; (3) No intervention or operation was performed on the nodule to be evaluated before ultrasonic examination; (4) Patients underwent surgery or biopsy within one week of ultrasonic data collection and obtained pathological results.

Exclusion criteria

(1) BIRADS 1 and 2; (2) Have a history of breast surgery or intervention; (3) Poor image quality; (4) The clinical data of cases are incomplete, and the pathological results are not tracked.

Patient anonymize and video preprocessing

Firstly, we cut the original video, remove the equipment information and patient sensitive information around the ultrasound image, and only keep the image window. Zoom video to 224 × 224 pixels, and then expand them frame by frame. These scaled frames will be used as the input data for the responsibility frame selection and comparison test of the new model (Figure 5).

Prior work

In our previous work,[44] we had completed training and testing four AI models (Xception,[45] MobileNet,[46] ResNet50,[47] and DenseNet121[48]) with three input image resolutions (224 × 224, 320 × 320, and 448 × 448 pixels). According to the inclusion and exclusion criteria, 3447 females breast nodules were finally included. Finally, we selected MobileNet in 224 × 224 pixels is the optimal model and image resolution, AUC: 0.893 (95% CI: 0.864-0.922), sensitivity: 81.6%, specificity: 82.1%, accuracy: 81.8%.

METHOD DETAILS

FEBrNet algorithm overview

In this study, we established the feature entropy breast network (FEBrNet), inheriting the backbone of the pretrained optimal models as feature extractor and adding a feature calculation layer at the last layer of models. The feature extractor processes each frame image and generates the corresponding feature matrix. We performed maxpooling on the feature matrix of all frames to get the reference feature matrix of entire video. A variable called *FScore* is defined by summing the values of malignant feature dimensions, indicating the total information of malignancy in feature matrix. The frame sharing most similar *FScore* to the reference feature matrix is selected as the first responsible frame and added to the responsible frame set. When adding a new frame into the responsible frame set, we use greedy algorithm to search all possible frames, which can minimize the differences in *FScore* of responsible frame set and the reference feature matrix. Meanwhile, for each added responsible frame, we calculate the malignant probability and the entropy of probability; once the entropy increases at a certain frame, the model's certainty of prediction decreases, which is the stop signal for frame selection.

Identifying responsible frames using FScore

In our method, we focus on minimizing the difference between the *FScore* of the video and the *FScore* of the responsible frame collection, with a smaller difference suggesting that the information of the responsible frame collection is closer to the whole video. By gradually increasing the number of frames in the responsible frame collection from 1 to n, and in each step selecting the frame with the smallest difference and adding it to the responsible frame collection, we eventually obtain the local optimal responsible frame collection, with each frame contributing various features (Figure 6).

The figure below shows a basic example of picking the top two responsible frames from a video with three frames (Figure 7). The video feature matrix is constructed in step 1 by MaxPooling the three frame feature matrices, and the $FScore_{video}$ is 25. In step 2, we specify the number of frames in responsible frame collection as one, and there are three options with *FScores* of 16, 15, and 11. Since the difference between $FScore_{video}$ and $FScore_{frame1}$ is the lowest when selecting frame1, we choose frame1 as the first responsible frame. Step 3 increases the number of frames in the responsible frame collection to two and the first responsible frame (frame1) has already been chosen. $FScore_{[frame1,frame2]}$ is 17 and $FScore_{[frame1,frame3]}$ is 24, thus frame3 should be selected as the second responsible frame. Despite the fact that $FScore_{frame2}$ is larger, we will not choose frame2 as the second responsible frame for the benefit of adding frame2 to the responsible frame collection are few. Frame2 provides almost the same features as the already selected frame1, implying that they may also look similar.

In math, we defined *FScore* to quantify the contribution of a frame in certain feature dimension. We revised the equation of final prediction in neural network. The output of the CNN in the basic CNN architecture could be written as $Y_{pred} = Softmax([W_0, W_1]^T * X + B) = [Y_0, Y_1]$, where Y_0 denotes the benign probability and Y_1 is the malignant probability. The number of $(w_1^i - w_0^i) \times x_i$ in the preceding equation represents the malignancy contribution of a single feature dimension, it could be separated into two parts: $(w_1^i - w_0^i)$ representing the amount of information that feature dimention i could contribute to final possibility of malignant and x_i representing the intensity of the frame in feature dimention i. Meanwhile we hope to concentrate on the features indicating malignancy, therefore we use the equation $FE_i^j = max(0, (w_1^j - w_0^j)) \times x^j$ to describe the contribution of framei in jth feature dimension, and the whole fearure matrix of framei can be represented as $[FE]_i = [FE_i^1, FE_i^2, FE_i^3 \cdots FE_i^k]$. To concentrate on the most representative characteristics of video, we used MaxPooling to alter the feature matrices of all frames and constructed the video feature matrix, where

$$[FE]_{video} = MaxPool([FE]_1, [FE]_2 \cdots [FE]_n) = \Big[max(FE_1^1, FE_2^1 \cdots FE_n^1), \\ max(FE_1^2, FE_2^2 \cdots FE_n^2) \cdots max\left(FE_1^k, FE_2^k \cdots FE_n^k\right)\Big]$$

Mathematically, we define the *FScore* of frame$_i$ as $FScore_i = sum([FE]_i) = \sum_{j=1}^{k} FE_i^j$. When it comes to a collection of frames A (A = [frame$_a$, frame$_b$, … frame$_c$]), we use $FScore_A = sum[FE]_{A\ max} = \sum_{j=1}^{k} max(FE_a^j, FE_b^j \cdots FE_c^j)$ to express its *FScore*.

The process of finding the ith responsible frame can be described as the following equation:

$$i^{th} frame = argmin\left(FScore_{video} - FScore_{[top1, top2, top3 \cdots topi-1, i]}\right) = argmin \\ \times \sum_{j=0}^{k}\left(max\left(FE_1^j, FE_2^j, \cdots FE_n^j\right) - max\left(FE_{top1}^j, FE_{top2}^j, \cdots FE_{topi-1}^j, FE_i^j\right)\right)$$

Automatically stop frame selection using entropy

To set the auto-stop strategy for frame selection, we borrow the idea of entropy in information theory which is proposed by Shannon in 1948.[24] In information theory, entropy is a measure of the uncertainty of random variables, and mathematically described as $H(X) = \sum_{x \in X} (-log_2\, \rho(x))\rho(x)$, where $\rho(x)$ is the probability density of random variable X, and $-log_2\, \rho(x)$ measure the amount of information which is also called self-information amount. We update the feature matrix every time the responsibility frame is added and calculate the entropy using malignant prediction value. With the increase of the number of responsibility frames, the prediction changes. When the certainty of malignant prediction value decreases, the entropy will increase, which is the stop signal (example Figure 1).

Experiment design

To evaluate whether FEBrNet-selected frames can be used as diagnostic frames, we trained AI models based on FEBrNet-recommended frames, all frames, fixed interval frames and physician-selected frames. Detailly, AI models based on different inputs have the same architecture, where pre-trained feature extractor process the inputted set of frames to generate feature matrix. Then, each model uses a random forest with same parameters (1024 estimators, max depth equals 10) to analyze the feature matrix of the frames set and generate predictions. Since the feature extractor is inherited, it requires no retraining process and only the random forests need training. Videos of all enrolled patients are divided into training and testing sets, where pathological results of breast nodules are used as ground truth. To ensure no crossover, the same patient can only appear in the same set. The results of models are compared, and an inference is that the model based on better-inputted frame set can achieve higher performance of AUROC.

First, we conducted three-fold cross-validation to evaluate models' performances on the training. Then we compare the diagnostic performances of all models on testing set and the performance of real-world ultrasound and mammography doctors. The real-world ultrasound and mammography doctors' performances are obtained by tracing back the recordings of patient's final examination before biopsy and re-evaluated by senior physicians.

QUANTIFICATION AND STATISTICAL ANALYSIS

Continuous variable data are expressed as mean ± standard deviation. Categorical variable data is expressed as a percentage. The paired sample t-test was used to compare the differences within the group. R 3.6.3 was used for the statistical analysis. Firstly, draw the receiver operating characteristic curve (ROC), calculate and output the area under the curve (AUC), and 95% confidence interval (95% CI), $p < 0.05$ is statistically significant. Output the optimal cut-off value, specificity, sensitivity, and accuracy.