Go beyond image-based benign-malignant classification: AI can identify responsible frames better than physicians in breast ultrasound screening videos

Yitao Jiang
Illuminate, LLC; Microport Prophecy
Xiuqin Ye
Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology)
Yan Liu
Qilu Hospital of Shandong University
Chen Cui
Illuminate, LLC; Microport Prophecy
Siyuan Shi
Illuminate, LLC; Microport Prophecy
Huaiyu Wu
Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology)
Hongtian Tian
Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology)
Di Song
Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology)
Jincao Yao
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
Liping Wang
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
Wenhai Sui
Qilu Hospital of Shandong University https://orcid.org/0000-0001-6114-0443
Sijing Huang

Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology)

Keen Yang

Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology)

Jing Chen

Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology)

Jinfeng Xu (xujinfeng@yahoo.com)

Ultrasound Department, First Affiliated Hospital of Southern University of Science and Technology, Second Clinical College of Jinan University,Shenzhen People's Hospital

Dong Xu (xudong@zjcc.org.cn)

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

Fajin Dong (dongfajin@szhospital.com)

Shenzhen People's Hospital (The Second Clinical Medical College，Jinan University，The First Affiliated Hospital, Southern University of Science and Technolog https://orcid.org/0000-0002-4558-4885

Yun Zhang (zhangyun@sdu.edu.cn)

The Key Laboratory of Cardiovascular Remodeling and Function Research, Chinese Ministry of Education and Chinese Ministry of Health, Qilu Hospital of Shandong University

---

Research Article

Keywords: Medical Imaging, Breast Cancer, Artificial Intelligence

Posted Date: February 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1235293/v2

Abstract

Breast Cancer is the most common cancer in the world and the single leading cause of cancer mortality in women. Heavy workload and shortage of ultrasound specialists impede the penetration of breast cancer screening. To reduce the burden of sonographers and empower junior physicians, we propose a novel framework FEBrNet by integrating deep learning architecture with the idea of entropy from Information theory. FEBrNet is capable of auto-selecting responsible frames from ultrasound screening videos based on entropy reduce method and classifying breast nodules using Artificial Intelligence (AI). A combination of 13702 images and 1066 videos from breast ultrasound exams are used to train and test the robustness of the proposed framework. Reader studies show that FEBrNet has equivalent or even superior diagnostic performance to that of ultrasound specialists and that overall physician's performance improves when using FEBrNet's recommended frames and corresponding prediction. Therefore, merging FEBrNet into clinical ultrasound screening workflow might bring actual benefit by helping address the scarcity of sonographers, so as to increase the use of ultrasound screening in cancer prevention.

Introduction

Breast Cancer is the most common cancer in the world[1] and the single leading cause of cancer in women.[2] According to Global Cancer Statistics 2020, the incidence of breast cancer in women has exceeded that of lung cancer, with more than 2.3 million new cases, accounting for 30% of female cancer patients and 11.7% of all cancer patients.[1] Effective, early screening is of great importance for improving 5-year survival rate. Studies have shown it can reduce the mortality rate by 38-48%, as well as local and distant recurrence rates.[3] At present, screening and follow-up mainly rely on two imaging modalities: mammography and ultrasound.[4-6] Mammography has been regularly adopted in breast screening programs in many countries.[7,8] With a higher sensitivity to dense breasts, ultrasound is also widely used worldwide.[9] Meanwhile, for low-income nations or regions, ultrasound can be a practical option for large scopes of screening due to its affordability, portability while without radiation.[10] In addition, ultrasound can potentially detect malignant foci obscured by thick glands as a supplement of mammography.[11]

However, present popularity of breast cancer screening with ultrasonography is insufficient. One primary reason is the shortage of experienced ultrasound specialists.[12] In United States, 40 million breast screening examinations take place, with half of those in women with dense tissue, requiring more than 6 million work hours per year.[13] In other regions with a larger female population of dense breasts, the amount of required ultrasound examinations might be times that of the United States. The burden has beyond physicians' capacity to endure, resulting in serious physical health concerns.[14] Along with the physician shortage, the geographical distribution of experienced physicians is also unequal. The access to fast, reliable ultrasound test in underdeveloped areas is in stark contrast with its counterpart, primarily resulting from the better education, training and opportunities for physicians in developed regions that underdeveloped areas are simply not able to match.[15,16]

The seeking of means to reduce the burden of sonographers and empower junior physicians has been long. Automated Breast Ultrasound(ABUS) is a technology that automates image acquisition, hoping to boost repeatability and decrease operator reliance.[17] It has shown high consistency with handheld ultrasound (HHUS) performed by competent doctors.[18-21] However, the large size and cost of their equipment preclude its use.

Artificial intelligence (AI) algorithm is capable of extracting a large number of quantitative features from breast ultrasound screening, so as to improve clinical detection accuracy. Shen. et al proved that with the help of the AI, false positive rates could be decreased by 37.4% and the number of requested biopsies reduced by 27.8%, while maintaining the same level of sensitivity.[22] Zhang. et al established an artificial intelligence system with the ability to not only identify cancer but also predict its molecular subtypes.[23] Dong. et al demonstrated that the AI system can provide explainable metrics such as diagnosis-based regions in addition to increasing accuracy, sensitivity, and specificity.[24] Further, AI could also be utilized in a variety of ultrasound hardware configurations, including traditional handheld ultrasound, ABUS, and portable handheld ultrasounds to address a variety of clinical needs in a range of scenarios.[25-28]

Most preceding works rely on fixed frames (static images) previously selected by sonographers, and the model's prediction is made using these frames rather than the original video that records the entire screening process. In other words, AI operates on data already filtered by sonographers, potentially resulting in a series of suboptimal issues. To begin with, video may contain information invisible to human eyes, which could be vital for AI to unleash its capabilities. Second, physician-selected frames might not fully represent the screening video, either due to the lack of necessary knowledge by junior physicians, or the lack of scrutiny by busy senior ones. No matter which causes, it could lead to uncertainty in the subsequent diagnosis and obvious bias in AI performance. Therefore, if AI is able to work on the screening video and recommend responsible frames that it deems necessary for diagnosis, it will possibly streamline the process for sonographers, particularly less experienced physicians, and potentially improve diagnostic performance.

To deepen the usage of both the insights in breast screening data and AI as a powerful feature-extracting tool, we propose a novel framework FEBrNet by integrating deep learning architecture with the entropy method from Information theory.[29] In FEBrNet, all frames in the breast screening video are processed with AI models in parallel, followed by an Entropy Reduce method to auto-select key frames with mutually distinctive features and significant contributions for breast lesion diagnosis. A combination of 13702 images and 1066 videos from breast ultrasound exams are used to train and test the robustness of the proposed framework. Our exhaustive multi-center, multi-reader experiment shows overall physician’s performance is improved when aided by FEBrNet’s recommended frames and corresponding prediction.

Result

Reader studies experiment design

We conducted four reader studies on the same video test set to compare the performance of the AI system and physicians, as well as to assess the benefits of using AI to aid physicians.

Complete AI diagnosis (Complete-AI): Use FEBrNets with a DenseNet or MobileNet backbone to diagnose videos and evaluate their performance when varying the number of responsible frames is employed.

Complete physician diagnosis (Complete-Phy): Six physicians independently read the original video and make diagnoses.

Physicians select frames, then AI diagnoses (Phy-AI): AI diagnoses based on the responsible frames chosen from a video test set by two senior physicians.

AI selects frames, followed by physician diagnosis (AI-Phy): FEBrNet offers physicians the top three responsible frames and predictions for each video. Physicians make diagnosis based on these information and physicians' diagnostic performance is evaluated.

Both 'Complete-Phy' and 'AI-Phy' use the same six physicians. Physicians are classified into three groups (junior, medium-level, and senior, each with two physicians) based on their experience. We conduct 'Complete-Phy' first, followed by 'AI-Phy' one month later, long enough for physicians to forget about their previous diagnosis.

FEBrNet could achieve comparable performance with a limited subset of frames from original video

In this part, we examine the performance of FEBrNet and the effect of changing the number of frames chosen on predication performance. Both the MobileNet and DenseNet121 backbone-based FEBrNet perform well in binary classification. MobileNet has the highest accuracy of 84.25%, while DenseNet121 has the highest accuracy of 84.93%. When the number of responsible frames is no less than 3, MobileNet's AUPR and AUROC values fluctuate at 0.875, whereas DenseNet121 performs slightly better, with AUPR and AUROC around 0.885. Figure 1 is an illustration of how precision, recall, and F1-Score vary for MobileNet and DenseNet121 as the number of responsible frames increases. In Figure 1(a), when less than 3 frames are used, MobileNet performs poorly but picks up with increased number of responsible frames. MobileNet reached its peak performance when only 3 frames is used and plateaued when more frames are added. As seen in Figure 1(b), DenseNet121 performs well in most situations; even when just the top 2 responsible frame is used for prediction, achieving 80.26% recall, 80.14% accuracy, and a 0.808

F1-Score. When more than 15 frames are employed, the effect of adding additional frames is minimal with no increase in F1-Score.

FEBrNet achieves more balanced precision and recall when allowed to directly analyze original videos

We also examine the performance of FEBrNet on the responsible frames chosen by senior doctors from each video in Table 1. MobileNet obtains 82.31% accuracy and DenseNet 81.63%. One observation is the imbalance between precision and recall, possibly implying a preference to reject more true positive cases to obtain greater precision and cause poor F1-Score (0.776 and 0.808 for MobileNet and DenseNet respectively). Such disparity is more noticeable in FEBrNet with MobileNet backbone, with 97.83% accuracy and just 64.29% recall. Meanwhile, FEBrNet with MobileNet backbone achieved 84.25% accuracy with 0.855 F1-Score and FEBrNet with DenseNet121 backbone got 84.93% accuracy with 0.864 F1-Score when AI engages original video file processing. FEBrNet is able to achieve a relatively better F1-Score when they are allowed to select frames on their own and make diagnosis. Overall, in our experiment, it is more stable to utilize AI to directly analyze original videos and forecast malignancy than relying on previously picked frames.

Table 1. Comparation of FEBrNet's performance in 'Physicians select frames, then AI diagnoses' and 'Complete AI diagnosis'

Backbone	AUROC	AUPR	Accuracy	Sensitivity	Specificity	Recall	Precision	F1
FEBrNet's performance in 'Physicians select frames, then AI diagnoses'								
MobileNet	0.883, [0.825,0.941]	0.903, [0.868,0.938]	82.31%	64.29%	98.70%	64.29%	97.83%	0.776
DenseNet121	0.891, [0.840,0.941]	0.897, [0.868,0.927]	80.95%	84.29%	77.92%	84.29%	77.63%	0.808
FEBrNet's performance in 'Diagnose videos completely with AI'								
MobileNet (use 3 frames)	0.865, [0.802,0.929]	0.873, [0.838,0.907]	84.25%	81.93%	87.30%	81.93%	89.47%	0.855
DenseNet121 (use 15 frames)	0.889, [0.834,0.945]	0.890, [0.839,0.941]	84.93%	81.40%	90.00%	81.40%	92.11%	0.864

FEBrNet outperforms physicians and can improve physicians' accuracy of diagnosis

Figure 2 shows ROC, PR curves of FEBrNet in Complete-AI experiments, while dots display results of Complete-Phy and AI-Phy. All of the locations where physicians diagnose themselves are below the curves of FEBrNet with DenseNet121 backbone, suggesting that FEBrNet with DenseNet121 backbone outperforms all of the physicians diagnose alone. Only 2 physicians reached the curves of FEBrNet with

MobileNet backbone, suggesting that even a light convolutional neural network has the potential to achieve equivalent diagnosis capacity as senior physicians.

As a result of FEBrNet's involvement, the performance of all physicians has been improved with the assistance of FEBrNet. Senior-1, senior-2, and medium level-1 were able to beat the FEBrNet with MobileNet backbone, while Senior-1 and senior-2 also reached the performance of FEBrNet with DenseNet backbone.

As indicated in Table 2, when FEBrNet was used, the accuracy and F1-score of all physicians improved. Senior-1, senior-2, and junior-2 improved their precision and recall at the same time, while medium level-1 and junior-1 benefited more from the enhanced recall. Accuracy and F1-Score improvement are greater for physicians with lower baseline performance.

Table 2. Performances of physicians in 'Complete physician diagnosis' and 'AI selects frames, followed by physician diagnosis'

	Accuracy	Sensitivity	Specificity	Recall	Precision	F1
Senior-1	79.45%	82.86%	76.32%	82.86%	76.32%	0.795
Senior-1 + AI	80.82%	83.33%	78.38%	83.33%	78.95%	0.811
Senior-2	79.45%	83.82%	75.64%	83.82%	75.00%	0.792
Senior-2 + AI	80.82%	88.57%	73.68%	88.57%	75.61%	0.816
Medium level-1	76.71%	81.43%	72.37%	81.43%	73.08%	0.774
Medium level-1 + AI	80.14%	78.57%	81.58%	78.57%	79.71%	0.791
Medium level-2	73.29%	88.57%	59.21%	88.57%	66.67%	0.761
Medium level-2+ AI	76.03%	76.62%	75.36%	76.62%	77.63%	0.771
Junior-1	69.86%	85.71%	55.26%	85.71%	63.83%	0.732
Junior-1 + AI	76.03%	77.14%	75.00%	77.14%	73.97%	0.755
Junior-2	71.92%	80.00%	64.47%	80.00%	67.47%	0.732
Junior-2 + AI	74.66%	81.43%	68.42%	81.43%	70.37%	0.755

Discussion

The conflict of limited medical resources and an increasing amount of patient population with breast cancer has long been a dilemma in public health, especially for underdeveloped regions.[33-36] To tackle this problem, efforts have been made to improve convenience and efficiency of breast screening. Artificial intelligence has potential to tackle this issue, and its value has been proposed by prior studies.[22-24] Beyond using AI in static breast images, we take a step further in processing ultrasound screening videos and auto-select key responsible frames in this research. To the best of our knowledge, it is the first time to combine deep learning models with entropy method of information theory in processing each frame of ultrasound screening videos. According to the result of our multi-reader studies, FEBrNet has shown its ability to diagnose breast cancer by outperforming physicians as well as improving physicians' accuracy of diagnosis, particularly for ones with limited expertise and baseline performance.

Responsible frames are vital in clinical workflows for locating key diagnostic information and predicting malignancy. As one major contribution of this work, the Entropy Reduce method is novel in addressing

the issue of choosing responsible frames while avoiding selecting visually identical frames repetitively. As shown in figure 3, we use a video taken from a 45-year-old female patient with BI-RADS 4c and pathologically confirmed invasive breast cancer as a simple example to demonstrate the capacity of entropy reduce method. When frames are sorted by *FScore*，a variable to evaluate the apparent degree of malignancy on image level, detailed definition could be found in Method part，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 3(a) and very close in time sequence. After using Principal Component Analysis (PCA)[37] to compress and display the feature matrices into two dimensions in figure 3(b), it is obvious that the distance in feature dimensions between the three frames is rather close. In Figures 3(c) and 3(d), the same approach is used to assess the top three responsible frames(frame 26, 111 and 96) chosen by Entropy Reduce method. 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 Entropy Reduce method are scattered, echoing their various visual attributes in Figure 3(d).

We also notice that FEBrNet can identify the features easy to be neglected by physicians. FEBrNet proposes two malignant indicative frames with architectural distortion in figure 4(a) that were overlooked by physicians during their earlier diagnosis. Figure 4(b) illustrates a more typical situation in which clinicians have difficulty in determining the lesion's malignancy due to a combination of benign (parallel orientation) and malignant (not circumscribed margin) features from physician chosen frames. FEBrNet contributes to the certainty of malignant likelihood by supplying frames with an additional malignant feature (calcification). With the ability to choose appropriate responsible frames, FEBrNet could alleviate sonographers' daily burden and enable physicians with less expertise to perform ultrasound breast screening. Hence, FEBrNet can potentially help address the scarcity of sonographers and contribute to the widespread use of ultrasound screening to diagnose diseases early.

Although FEBrNet has solved the problem of selecting responsible frames for sonographers in part, there are still areas that need further investigation. The first task is to find out how to get high-quality ultrasound screening videos, which serves as the foundation for all subsequent work. AI has shown a great potential for assistance in the acquisition of high-quality ultrasound data[26,38] and several studies have revealed the advantages of applying AI to assist ultrasound imaging[39,40] One next endeavor is to develop a navigational AI for breast ultrasound screening. Furthermore, after we have obtained the responsible frames, we will need a tool to analyze the features of images. There are explicable aspects of FEBrNet worth examining, including currently established diagnostic features (e.g. margin, calcification and other features in BI-RADS)[41] and the features have not yet been discovered. In the long term, it can be anticipated to construct an AI-based ultrasound screening system streamlining guided ultrasound imaging, video data processing, diagnosis, and explanation generation.

FEBrNet has further exploration possibilities, and our current work has certain limits. To begin, the quantity of breast ultrasound data included in this research is limited, and more multicenter breast ultrasound data from a broader patient group should be incorporated. Second, our study focuses only on breast nodule malignancy classification, although its applicability might be expanded to include subtype

disease categorization, molecular phenotype prediction, and other tasks. Finally, ultrasonic screening for a variety of disorders may be possible using the FEBrNet architecture, and more indications should be evaluated to ensure its robustness.

Materials And Methods

Ethical Approval and Informed Consent

This study obtained ethical approval from the Institutional Review Board of the Shenzhen People's Hospital. The approval included the collection of data on implied consent. We only used retrospective data and the patients were not actively involved in the study. The requirement of written informed consent was waived by the Institutional Review Board.

Data sources and entry criteria

This retrospective study was conducted in accordance with the procedures specified by the hospitals that participated. The Ethics Committees of the Cancer Hospital of The Chinese Academy of Sciences and Shenzhen People's Hospital authorized this research. To ensure the quality of the data, we based the experiment's inclusion and exclusion criteria on clinical guidelines.

The following criteria apply to data inclusion:

(1) Ultrasound detection of breast nodules;

(2) Nodule diameter must be between 5.0-30.0mm;

(3) Breast tissue surrounding the nodule must be at least 3.0mm thick;

(4) Nodules must be BIRADS 0, 2, 3, 4a, 4b, 4c, or 5;

(5) No intervention or surgery on the nodule has been performed before the ultrasound test;

(6) Patients must undergo surgery or biopsy within one week of the ultrasound data collection and obtain pathological results.

The following criteria are used to exclude data:

(1) normal breasts (BIRADS category 1);

(2) a history of breast surgery or interventional therapies;

(3) image quality is poor;

(4) clinical data for the case are insufficient, and pathological outcomes are untraceable.

Study population and data distribution of image set

The study comprised 13702 2D ultrasound breast nodule images with pathology results acquired from 3448 female patients between 2020.10 and 2021.10 (9177 images from 2457 patients with benign pathology, 4545 images from 991 patients with malignant pathology), as stated in Table 3.

All pictures utilized are grayscale ultrasound images from each of one which a region of interest (ROI) is extracted. All non-object regions in the ultrasound image are eliminated. The image dataset is utilized to build the CNN image classifier in the first step of FEBrNet, which is then transferred to a video classifier.

Table 3. The study population and images in training set, validation set, and test set. Bn: benign. Mal: malignant.

	Malignancy	Bn	Mal
Train	Number of patients	2004 (72.14%)	774 (27.86%)
	Total number of patients	2778	
	Number of images	7324 (67.78%)	3482 (32.22%)
	Total number of images	10806	
Validation	Number of patients	254 (72.99%)	94 (27.01%)
	Total number of patients	348	
	Number of images	889 (68.75%)	404 (31.25%)
	Total number of images	1293	
Test	Number of patients	199 (61.80%)	123 (38.20%)
	Total number of patients	322	
	Number of images	946 (59.68%)	639 (40.32%)
	Total number of images	1585	

Study population and data distribution of video set

As shown in Table 4, the study includes 1066 ultrasound breast nodule films with pathology results from 440 female patients between 2020.10 and 2021.10. (546 videos from 237 patients with benign pathology and 520 videos from 203 patients with malignant pathology). Additionally, we gathered the physician chosen responsible frames for each video in the dataset, which are the frames that two senior physicians

confirm include significant characteristics indicative of malignancy (random number of responsible frames for each video, including raw frames and annotated frames).

The video dataset is used to train the random forest feature classifier, which processes the pretrained CNN image classifier's features. To prevent information leakage during model training, we ensure that the patients in the video data set do not overlap with the patients in the image data set.

Table 4. The distribution of videos in training set, validation set, and test set. Bn: benign. Mal: malignant.

Malignancy	Train set		Test set	
	Bn	Mal	Bn	Mal
Number of patients	161	133	76	70
(Percentage of malignancy)	54.8%	45.2%	52.1%	47.9%
Total number of patients	294		146	
Number of videos	367	332	179	188
(Percentage of malignancy)	52.5%	47.5%	48.8%	51.2%
Total number of videos	699		367	
Number of responsible images selected by senior doctors	538	436	1052	896
(Percentage of malignancy)	55.3%	44.7%	54.0%	45.0%
Total number of responsible images selected by senior doctors	974		1948	

Statistical evaluation

In this paper, we ran 3 trials to evaluate FEBrNet:

1. Results of eleven different numbers of responsible frames of FEBrNet are compared, using AUROC, AUPR, Accuarcy, Sensitivity, Specificity, Recall, Precision, and F1-Score, to assess the impact of number of responsible frames on FEBrNet's performance.
2. 'Physicians select frames, then AI diagnoses' and 'Complete AI diagnosis' compare, using AUROC, AUPR, Accuarcy, Sensitivity, Specificity, Recall, Precision, and F1-Score, to benchmark FEBrNet when utilizing raw videos without human judgment.
3. 'Complete physician diagnosis', 'Complete AI diagnosis' and 'AI selects frames, followed by physician diagnosis' compare, using scatter plot of FEBrNet and physician performances, AUROC, AUPR, Accuarcy, Sensitivity, Specificity, Recall, Precision, and F1-Score, to evaluate how FEBrNet compares physicians when diagnosing alone and whether physicians benefit from the assistance of FEBrNet.

Model section: Philosophy and DataFlow of FEBrNet

In this part, we illustrate the philosophy of FEBrNet, its computational operations and data flow step by step. In short, there are four steps in the workflow of FEBrNet: 1) feature distillation, 2) entropy matrix generation, 3) responsible frame recommendation, and 4) binary classification. Figure 5 depicts the data flow of FEBrNet. In the first step, state-of-the-art deep learning models are trained to acquire the knowledge from static breast ultrasound images. The backbone of the model is transferred to the second step for parallel feature extraction from breast ultrasound videos, which are independent of the static images. Together with key weights from the first step, frame-by-frame feature vectors are concatenated into a new feature matrix (feature entropy matrix). Third, we design a new entropy reduce method to select a subset of all frames to represent the entire ultrasound video for this particular event of breast lesion diagnosis. Meanwhile, binary classification of benign or malignancy is also conducted based on the feature entropy matrix to assist physicians in the diagnosis.

Model section: Feature distillation

Two major merits stand out in feature distillation: 1) Pre-accumulated physician-selected images contain a plethora of breast lesion features, in particular malignant ones. Therefore, to create a model capable of extracting essential malignant characteristics, Convolutional Neural Networks (CNN) can be pretrained on a relatively large ultrasound image dataset. 2) To a large extent, the backbone of the pretrained model can accelerate the process of video feature extraction compared to training parallel model from scratch. Here, most standard CNN models, including DenseNet, ResNet, and MobileNet, could be used. DenseNet and MobileNet are used in this experiment for comparing light-weight model (3,230,914 parameters for MobileNet) and sophisticated model (7,039,554 parameters for DenseNet).

We split the entire image dataset containing 13072 2D images into three portions in our research (train : valid : test = 8:1:1), with images of the same patient appearing in only one of the subsets. For data augmentation, random stiff transformations on the original image can be used to mimic the image displacement, zoom, and flip that might occur as physicians scan for nodules in real-world clinical procedures. Some of the specialized methods including rotation, zoom, translation, and flip, as well as grayscale adjustment are also used. Binary Cross-entropy Loss is utilized to calculate classification loss and adjust network weights.

Model section: Entropy matrix generation

Various nodular features are scattered over the temporal sequence of frames. The principle of entropy matrix generation is mapping all features of all frames to a high-dimensional space, where all features of a single frame can be presented as one vector. Transferred from step 1, backbone of pretrained CNN model serves as a feature extractor to distill essential features from each frame of video in parallel and creates feature matrices. Feature matrices reveals the feature intensity of each frame in diverse feature

dimension, where the number of dimensions is determined by the backbone model (1024 dimensions in the DenseNet121 backbone). By incorporating the weights of the final layers in step 1 as they represent the indicative information of malignancy, we obtained feature entropy matrices.

Model section: Responsible frame recommendation

Here, feature entropy matrices are used to rank the contribution of frames for breast nodule diagnosis, and a key variable we define is called *FScore*, which is the sum of values of feature entropy matrices from all feature dimensions at each frame. With a higher *FScore*, the frame contributes more characteristics indicating the possibility of malignant. *FScore* assists in locating the frame that contributes the most to the possibility of malignancy, which we defined as the first responsible frame of the model's prediction. However, since adjacent frames usually share similar image signatures and have very close *FScore*, which means the frame with the largest *FScore* and the second largest *FScore* may looks almost identical.

To choose a comprehensive set of responsible frames with varied features, we first extend the concept of *FScore* from a single image to a collection of frames, because video in essence, is a collection of frames. Second, we propose a novel entropy reduce method to select a minimal set of all frames with a lowest sum of entropy values. This set of frames is considered to have the highest likelihood to represent the entire video. The essential philosophy of the entropy method is a greedy mechanism, where the next frame is repeatedly searched to reduce the sum of entropy of all selected frames until the entropy sum can not be reduced. More mathematical illustrations and examples can be found in Supplement.

Model section: Binary classification

For classification task, we re-organize (MaxPool in deep learning) the feature entropy matrix of all frames or selected responsible frames (option 1 or 2 in figure 1) to compress and shape the information into a consistent shape of a vector (1*k). The matrix shows the maximum contribution of video in each feature dimension and is indicative to classify video as benign or malignancy. We refer to the compressed feature entropy matrix as the video feature entropy matrix.

In FEBrNet, we employ a classic machine learning model to analyze video feature entropy matrices and produce final benign-malignant predictions. Here, random forest is adopted to train on our video training set,[30] with 1024 feature estimators and the maximum depth below 10. The classification is made based on entropy matrices instead of original videos for two reasons: 1) the feature entropy matrices already encapsulate key information from previous steps for classification; 2) undesirable noisy frames during the continual screening process might also undermine the classifying accuracy.

Conclusion

To summarize, we developed a system called FEBrNet, which is capable of auto-selecting responsible frames from ultrasound screening videos and classifying breast nodules using AI. Results of reader studies prove that FEBrNet has equivalent or even superior diagnostic performance to that of ultrasound specialists. It has the potential to empower physicians with less expertise to conduct better breast ultrasound screening and achieve better accuracy. Therefore, merging FEBrNet into clinical ultrasound screening workflow might bring actual benefit by helping address the scarcity of sonographers, so as to increase the use of ultrasound screening in cancer prevention.

Declarations

Conflict of Interest

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.

Data Availability

The dataset used in this study contain protected health information (PHI) and cannot be published for reasons of data protection.

Code Availability

The source code of this publication may be obtained from the corresponding author upon reasonable request.

References

1 Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, doi:10.3322/caac.21660 (2021).

2 Chen, W. *et al.* Cancer incidence and mortality in China, 2013. *Cancer Lett* **401**, 63-71, doi:10.1016/j.canlet.2017.04.024 (2017).

3 Cedolini, C. *et al.* Type of breast cancer diagnosis, screening, and survival. *Clin Breast Cancer* **14**, 235-240, doi:10.1016/j.clbc.2014.02.004 (2014).

4 Slomski, A. Long-term Mortality Outcomes in Early Breast Cancer Screening Trial. *Jama* **324**, 2020, doi:10.1001/jama.2020.22569 (2020).

5 Bell, K. J. L. *et al.* Mammography screening for breast cancer-the UK Age trial. *Lancet Oncol* **21**, e504, doi:10.1016/s1470-2045(20)30528-3 (2020).

6 Lotter, W. *et al.* Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med* **27**, 244-249, doi:10.1038/s41591-020-01174-9 (2021).

7 Schünemann, H. J. *et al.* Breast cancer screening and diagnosis: a synopsis of the European Breast Guidelines. *Annals of internal medicine* **172**, 46-56 (2020).

8 Narayan, A. K., Lee, C. I. & Lehman, C. D. J. M. C. Screening for breast cancer. *Medical Clinics* **104**, 1007-1021 (2020).

9 Rajaram, N. *et al.* Differences in mammographic density between Asian and Caucasian populations: a comparative analysis. *Breast cancer research* **161**, 353-362 (2017).

10 Mascara, M. & Constantinou, C. J. C. O. R. Global Perceptions of Women on Breast Cancer and Barriers to Screening. *Current Oncology Reports* **23**, 1-9 (2021).

11 Wang, J. *et al.* Is Ultrasound an Accurate Alternative for Mammography in Breast Cancer Screening in an Asian Population? A Meta-Analysis. *Diagnostics* **10**, 985 (2020).

12 Waring, L., Miller, P. K., Sloane, C. & Bolton, G. J. U. Charting the practical dimensions of understaffing from a managerial perspective: The everyday shape of the UK's sonographer shortage. *Ultrasound Med Biol* **26**, 206-213 (2018).

13 Brem, R. F., Lenihan, M. J., Lieberman, J. & Torrente, J. J. A. J. o. R. Screening breast ultrasound: past, present, and future. *American Journal of Roentgenology* **204**, 234-240 (2015).

14 Zhang, D. *et al.* Evaluation of work-related musculoskeletal disorders among sonographers in general hospitals in Guangdong province, China. *International Journal of Occupational Safety* **26**, 802-810 (2020).

15 Lili, X. *et al.* Analysis of breast cancer cases according to county-level poverty status in 3.5 million rural women who participated in a breast cancer screening program of Hunan province, China from 2016 to 2018. *Medicine* **99** (2020).

16 Raza, S. *et al.* Training Nonradiologist Clinicians in Diagnostic Breast Ultrasound in Rural Rwanda: Impact on Knowledge and Skills. *Journal of the American College of Radiology* **18**, 121-127 (2021).

17 Zanotel, M. *et al.* Automated breast ultrasound: basic principles and emerging clinical applications. *La radiologia medica* **123**, 1-12 (2018).

18 Vourtsis, A. & Kachulis, A. J. E. r. The performance of 3D ABUS versus HHUS in the visualisation and BI-RADS characterisation of breast lesions in a large cohort of 1,886 women. *European radiology* **28**, 592-601 (2018).

19 Wilczek, B., Wilczek, H. E., Rasouliyan, L. & Leifland, K. J. E. j. o. r. Adding 3D automated breast ultrasound to mammography screening in women with heterogeneously and extremely dense breasts: report from a hospital-based, high-volume, single-center breast cancer screening program. *European journal of radiology* **85**, 1554-1563 (2016).

20 Giger, M. L. *et al.* Automated breast ultrasound in breast cancer screening of women with dense breasts: reader study of mammography-negative and mammography-positive cancers. *American Journal of roentgenology* **206**, 1341-1350 (2016).

21 Kim, S. H., Kim, H. H. & Moon, W. K. J. K. j. o. r. Automated breast ultrasound screening for dense breasts. *Korean journal of radiology* **21**, 15-24 (2020).

22 Shen, Y. *et al.* Artificial Intelligence System Reduces False-Positive Findings in the Interpretation of Breast Ultrasound Exams. *medRxiv* (2021).

23 Zhang, X. *et al.* Evaluating the Accuracy of Breast Cancer and Molecular Subtype Diagnosis by Ultrasound Image Deep Learning Model. *Frontiers in oncology* **11**, 606 (2021).

24 Dong, F. *et al.* One step further into the blackbox: a pilot study of how to build more confidence around an AI-based decision system of breast nodule assessment in 2D ultrasound. *European Radiology*, 1-10 (2021).

25 Akkus, Z. *et al.* Artificial Intelligence (AI)-Empowered Echocardiography Interpretation: A State-of-the-Art Review. *Journal of Clinical Medicine* **10**, 1391 (2021).

26 Drukker, L., Noble, J., Papageorghiou, A. J. U. i. O. & Gynecology. Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. *Ultrasound in Obstetrics Gynecology* **56**, 498-505 (2020).

27 Zhou, Y. *et al.* Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. *Medical Image Analysis* **70**, 101918 (2021).

28 Akkus, Z. *et al.* A survey of deep-learning applications in ultrasound: Artificial intelligence–powered ultrasound for improving clinical workflow. *Journal of the American College of Radiology* **16**, 1318-1328 (2019).

29 Shannon, C. E. J. T. B. s. t. j. A mathematical theory of communication. *The Bell system technical journal* **27**, 379-423 (1948).

30 Breiman, L. Random forests. *Machine learning* **45**, 5-32 (2001).

31 Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 4700-4708.

32 Howard, A. G. *et al.* Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint* (2017).

33 Sharma, R. J. J. o. P. H. Breast cancer burden in Africa: evidence from GLOBOCAN 2018. *Journal of Public Health* (2020).

34 Sharma, R. J. B. C. Breast cancer incidence, mortality and mortality-to-incidence ratio (MIR) are associated with human development, 1990-2016: evidence from Global Burden of Disease Study 2016. *Breast Cancer* **26**, 428-445 (2019).

35 Li, N. *et al.* Global burden of breast cancer and attributable risk factors in 195 countries and territories, from 1990 to 2017: results from the Global Burden of Disease Study 2017. *Journal of hematology oncology* **12**, 1-12 (2019).

36 Mubarik, S. *et al.* Trends and projections in breast cancer mortality among four Asian countries (1990-2017): evidence from five stochastic mortality models. *Scientific reports* **10**, 1-12 (2020).

37 Jolliffe, I. J. E. o. s. i. b. s. Principal component analysis. (2005).

38 Muse, E. D. & Topol, E. J. J. T. L. Guiding ultrasound image capture with artificial intelligence. *The Lancet* **396**, 749 (2020).

39 Cheema, B. S., Walter, J., Narang, A. & Thomas, J. D. J. C. R. Artificial intelligence-enabled POCUS in the COVID-19 ICU: a new spin on cardiac ultrasound. *Case Reports* **3**, 258-263 (2021).

40 Schneider, M. *et al.* A machine learning algorithm supports ultrasound-nagtive novices in the acquisition of diagnostic echocardiography loops and provides accurate estimation of LVEF. *International Journal of Cardiovascular Imaging* **37**, 577-586 (2021).

Figures

Figure 1

Recall, Precision and F1-Score of FEBrNet when using different numbers of responsible frames

(a) MobileNet's recall falls and precision increases as the number of responsible frames used grows when less than three frames used.

(b) DenseNet121 works well in most circumstances. More than 15 responsible frames had little influence on precision and recall.

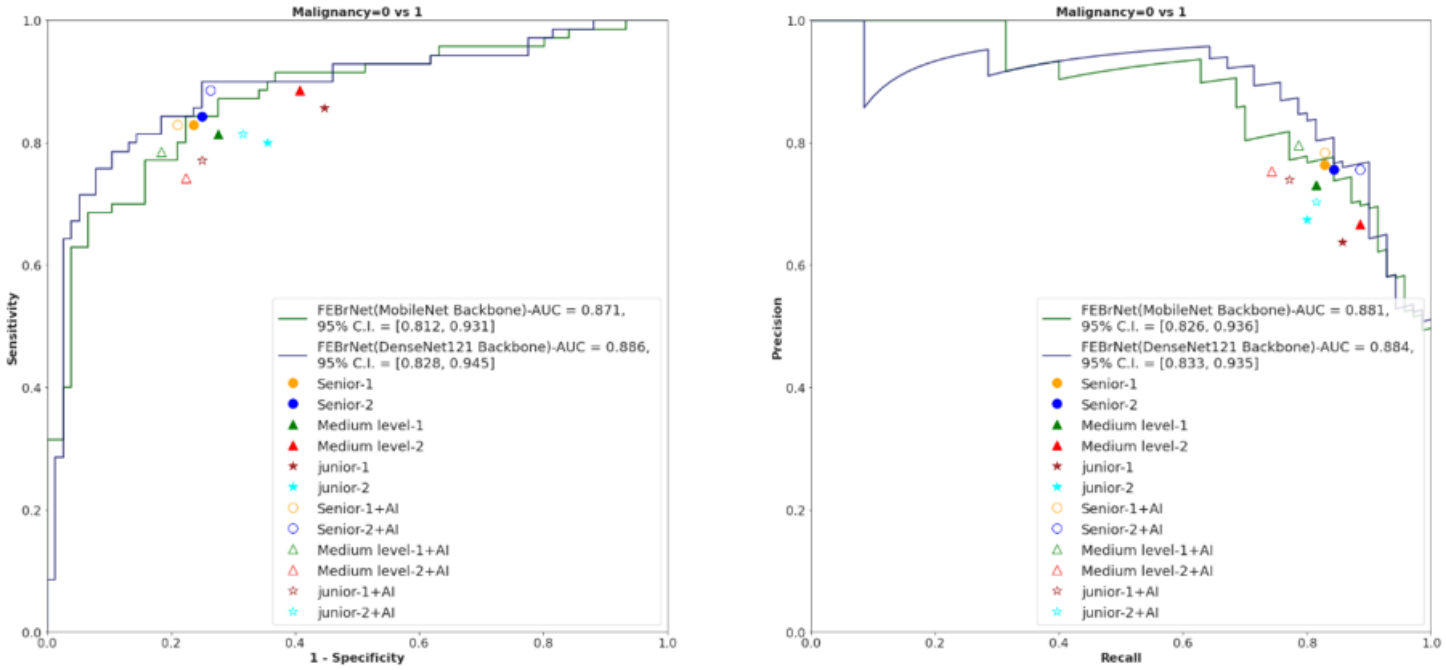


Figure 2

Comparison of 'Complete physician diagnosis', 'Complete AI diagnosis' and 'AI selects frames, followed by physician diagnosis'.

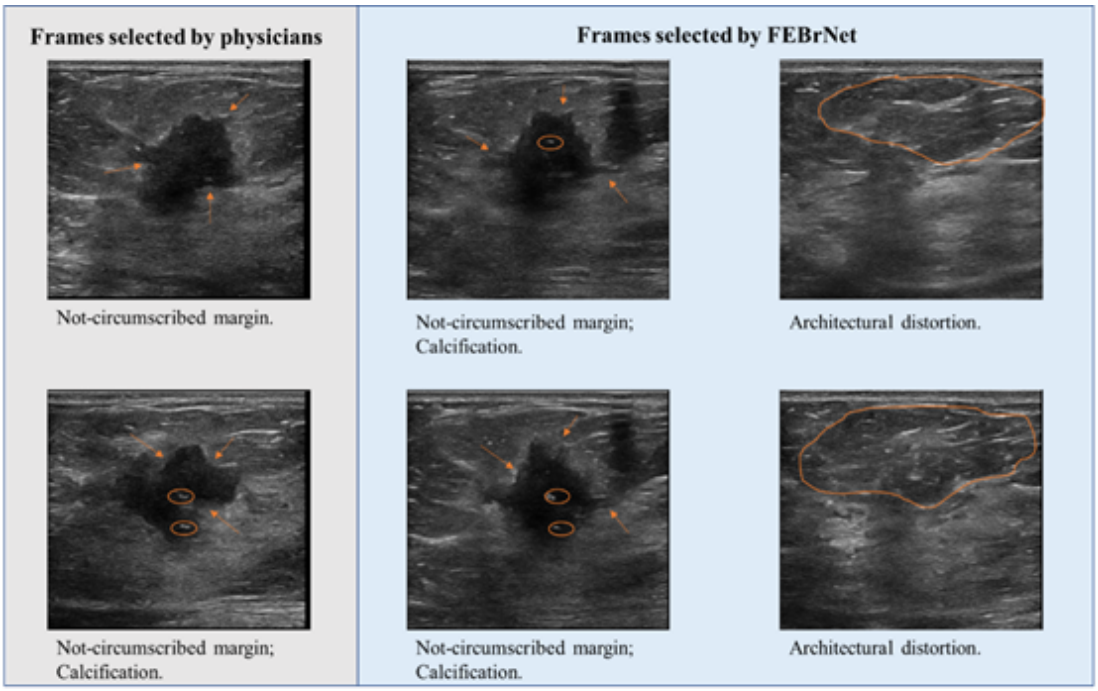
When physicians diagnose alone, FEBrNet with DenseNet backbone outperforms all of them. With the assistance of FEBrNet, all physicians' performances have been enhanced, and physicians 1 and 2 exhibits comparable diagnostic capability to AI.

Figure 3

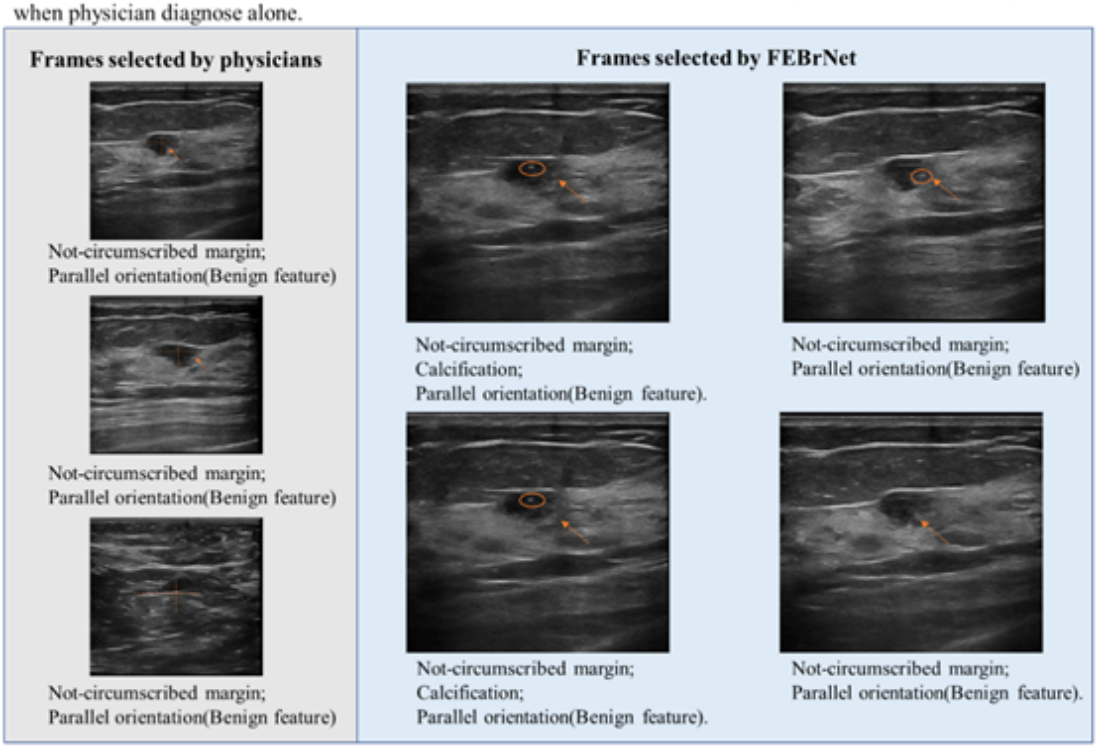
A case study of responsible frames selected by Entropy Reduce method

(a) The top 3 frames sorted by *FScore* are relatively similar in time sequence and visually identical;

(b) The top 3 frames chosen by Entropy Reduce method show more diverse image characteristics and are scattered on 2D feature plot.



(a) Case study 1: FEBrNet detected architectural distortion(confirmed by two senior-physicians) which is neglected when physician diagnose alone.



(b) Case study 2: FEBrNet detected calcification(confirmed by two senior-physicians) which is neglected when physician diagnose alone.

Figure 4

Cases of FEBrNet discovered physicians overlooked malignant features.

Comparison of frames selected by physicians and FEBrNet reveals that FEBrNet has the capacity of identifying features easy to be overlooked by physicians.

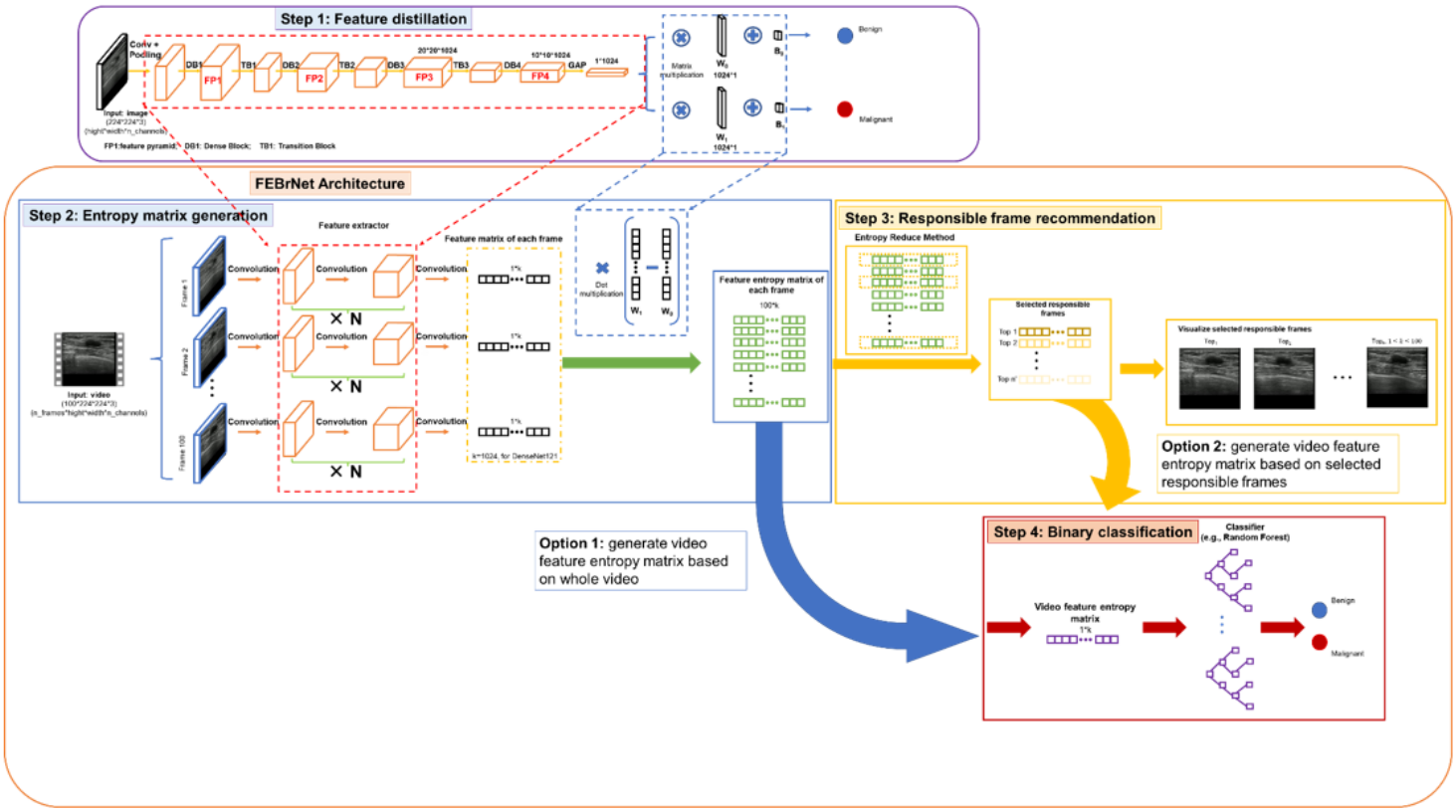


Figure 5

Architecture of FEBrNet

The backbone model is a CNN trained on our ultrasound dataset, which supplies the feature extractor and weights of fully connected layer. FEBrNets generate feature entropy matrices using the feature extractor and weights from CNN's fully connected layer. These matrices are used to identify responsible frames and produce predictions lat

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplement.docx