Published in final edited form as:
IEEE Trans Fuzzy Syst. 2021 January ; 29(1): 34–45. doi:10.1109/TFUZZ.2020.2966163.

A Heuristic Neural Network Structure Relying on Fuzzy Logic for Images Scoring

Cheng Kang,
School of Informatics, the University of Leicester, Leicester, United Kingdom

Xiang Yu [Student Member, IEEE],
School of Informatics, the University of Leicester, Leicester, United Kingdom

Shui-Hua Wang* **[Member, IEEE]**,
School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 454000, P R China & School of Mathematics and Actuarial Science, the University of Leicester, Leicester, LE1 7RH, United Kingdom

David S. Guttery*,
Leicester Cancer Research Centre, University of Leicester, Leicester, United Kingdom

Hari Mohan Pandey*,
Department of Computer Science, Edge Hill University, Lancashire, UK

Yingli Tian* **[Fellow, IEEE]**,
Department of Electrical Engineering, The City College of New York, 10031, USA

Yu-Dong Zhang* **[Senior Member, IEEE]**
Informatics, University of Leicester, Leicester, LE1 7RH, UK & Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Abstract

Traditional deep learning methods are sub-optimal in classifying ambiguity features, which often arise in noisy and hard to predict categories, especially, to distinguish semantic scoring. Semantic scoring, depending on semantic logic to implement evaluation, inevitably contains fuzzy description and misses some concepts, for example, the ambiguous relationship between normal and probably normal always presents unclear boundaries (normal – more likely normal - probably normal). Thus, human error is common when annotating images. Differing from existing methods that focus on modifying kernel structure of neural networks, this study proposes a dominant fuzzy fully connected layer (FFCL) for Breast Imaging Reporting and Data System (BI-RADS) scoring and validates the universality of this proposed structure. This proposed model aims to develop complementary properties of scoring for semantic paradigms, while constructing fuzzy rules based on analyzing human thought patterns, and to particularly reduce the influence of semantic conglutination. Specifically, this semantic-sensitive defuzzier layer projects features occupied by relative categories into semantic space, and a fuzzy decoder modifies probabilities of the last

*Co-correspondence authors.

Europe PMC Funders Author Manuscripts

output layer referring to the global trend. Moreover, the ambiguous semantic space between two relative categories shrinks during the learning phases, as the positive and negative growth trends of one category appearing among its relatives were considered. We first used the Euclidean Distance (ED) to zoom in the distance between the real scores and the predicted scores, and then employed two sample t test method to evidence the advantage of the FFCL architecture. Extensive experimental results performed on the CBIS-DDSM dataset show that our FFCL structure can achieve superior performances for both triple and multiclass classification in BI-RADS scoring, outperforming the state-of-the-art methods.

Index Terms

Fuzzy deep neural networks; transfer learning; fuzzy fully connected layer; medical image scoring

I Introduction

DEEP learning has recently gathered huge interest across a multitude of disciplines [1, 2], which has resulted in researchers applying deep learning to score medical images. However, whether pre-training neural networks by natural images can effectively identify malignant or normal features in medical images has not yet been sufficiently investigated, despite the fundamental features between them being diverse. Further, big datasets may contain high amounts of noise and uncertainties. Ambiguity features, for example, semantic relative processing, impose great challenges on data understanding and classification.

In order to reduce the noise inherent in these systems and improve diagnostic accuracy, fuzzy learning strategies obtain specific inherent logic of humans, and have been established [3, 4], for example, towards image processing [5], image classification [6], and motor control [7]. Researchers have engaged in developing some new neural networks with inherent and embedded common senses to address highly challenge tasks, such as natural language understanding [8], visual question answering [9], and aspect extraction in opinion mining [10]. Fuzzy theory to optimize multi-input and single-output static systems affected by noise has been developed [11], the linear and nonlinear defuzzifiers based on fuzzy rules, compared with conventional deterministic representations, can reduce the uncertainties encountered in these raw data [12], as well as methods to identify nearest-neighbor memeplexes by fuzzy systems [13]. However, this kind of embedded inherent knowledge has not yet referred to deep learning classification regarding to the adjacent overlap of linear scoring. For instance, the Breast Imaging Reporting and Data System (BI-RADS), established by the American College of Radiology, is a scheme for defining mammogram screening into well-defined categories. BI-RADS scoring [14] can evaluate patients' status and provide semantic diagnosis by numerical values, such as probably benign (BI-RADS 3) or benign (BI-RADS 2), and these two categories frequently share similar features, which may increase the difficulty for classifying by using convolutional neural networks (CNNs). This type of semantic or affective diagnosis (denotative and connotative information) [15, 16], contains ambiguous information which causes the partial divergence of neural networks, unlike either auto-categorization or summarization. Therefore, it is natural to ask: regarding

Europe PMC Funders Author Manuscripts

existing CNNs, how can we reduce the relativity of adjacent categories by improving these traditional neural networks with inherent knowledge from human thought?

Our proposed method differs from previous studies since we are assembling priori knowledge derived from the suggestion of experts (mainly about the property of categories), being greedy to lead outputs to the global performance and fitting parameters by modifying back propagation errors. Thus, based on one previous study [17] for self-constructing fuzzy systems, and to verify the question about margins reinforcement and classify ambiguous cases, we designed three experiments in this study, including reinforcement of margins and learning of these reinforced features through six CNNs in the first learning phase, concatenation of the established fuzzy fully connected layers (FFCLs) on the top of the best-performing CNN for triple in the second learning phase and six-class classifications in third learning phase, to gradually improve the inherent structure of traditional neural networks. The influence of margin status is a significant measurement to evaluate breast cancer [18]. We calculated margins by canny and log operators, and designed improved neural networks to learn these important features in this study, because these two operators are recognized as the most generally used edge detectors.

Depending on FFCL, features represented by these pre-trained networks were fused together in this nonlinear layer, and then reserved, deblurred, and adjusted. It can offer traditional neural networks the ability to build cognitive connections among relative categories in the last output layer, and more dependable update of weights and biases thereof. For instance, the forming of the final probabilities for data classification in the output layer can then partially present the distribution of features related to high or low scores. Briefly, after training of several epochs, the neural network without any improvement was able to identify BI-RADS scores with acceptable accuracy. Referring to the probabilities' distribution of the output layer, FFCL can reduce the uncertainties and noise of the original data by updating these output probabilities and back propagation errors. As a result, these updated back propagation errors can influence the presentation of every layer's weights and bias. Overall, FFCL neural networks can be applied to more difficult pattern classification tasks, such as BI-RADS involving data ambiguity and noise. We selected ResNet from seven simple neural networks and supplemented FFCL, as this structure leads to better performance than other state of the art methods in this study.

In this paper, 1) we verified that the enhancement of visual features, such as edges, is not so beneficial to improve the performance of CNNs, which essentially demonstrated that CNNs can extract visual features; 2) we proved that the transfer learning strategy, especially trained by natural categories, can extract medical features, because more and deeper convolutional layers cannot detect new medical information from CBIS-DDSM image dataset; 3) this proposed and introduced FFCL architecture, which essentially focused on fused fuzzy rules deriving from parsing logic representation with traditional convolutional neural networks for semantic BI-RADS scoring, weakens the fusion logic in terms of fuzzy semantic definition, as this type of semantic diagnosis always contains an unstable overlap between two neighbour categories; 4) these extensive experiments demonstrated that the proposed FFCL architecture is effective and outperforms other existing state-of-the-art methods when

scoring BIRADS based on the CBIS-DDSM dataset. Codes and models are available at: https://github.com/ChengKang520/.

II Scoring and fuzzy fully connected layer

A Fuzzy scoring and structure of fuzzy fully connected layer

Let the training set be $T = [(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]$, where x_i is the variable explaining the data and y_i is the corresponding label, for all $i = 1,2,\ldots, n$ where n is the number of training samples. We assumed that the sample was partitioned into m scoring categories, which were defined as real score $S = [S_1, S_2, \ldots, S_m]$. Therefore, for more accurate evaluation, the estimated score $\widetilde{S} = \left[\widetilde{S}_1, \widetilde{S}_2, \ldots, \widetilde{S}_m\right]$ followed by a decimal part.

1) Fuzzy Function—To minimize the influence between two relative categories, this fuzzy function involves directed extensional scores. For CNNs, the probabilities are defined by sigmoid function:

$$P(x_i | y_i = i) = \frac{e^{-E(y_i, x_i)}}{\sum_{y_1}^{y_m} e^{-E(y_i, x_i)}}, \tag{1}$$

and its left and right neighbors:

$$P(x_{i\pm1} | y_{i\pm1} = i \pm 1) = \frac{e^{-E(y_i \pm 1, x_i \pm 1)}}{\sum_{y_1}^{y_m} e^{-E(y_i \pm 1, x_i \pm 1)}}, \tag{2}$$

where $E(y_i, x_i)$ is the expectation that x_i is predicted as y_i, and m is the number of the categories. According to some previous common studies of CNNs, the $P(x_i | y_i) \in (0, 1)$ and this probabilistic distribution has the following affine forms [19]:

$$P(x_i | y_i = i) = \sigma(W_i x_i + b_i), \tag{3}$$

$$P(x_{i\pm1} | y_{i\pm1} = i \pm 1) = \sigma(W_{i\pm1} x_{i\pm1} + b_{i\pm1}), \tag{4}$$

where W_i is the weight in layer i, and b_i is the bias in layer i. To reduce the conglutination between two either neighbors or remote classes, the recursive score was calculated by

$$\widetilde{V_o} = \frac{i \sum_{i-a}^{i+b} P(x_i | y_i)}{\sum_{i-a}^{i+b} P(x_i | y_i)}, \tag{5}$$

where a is the trend of negative growth, while b is the trend of positive increasing, for example, the normal trend to become healthy, and the abnormal trend to become cancer. Therefore, the output value modified by above operators tend to slip forward to the global average position, and we optimized the redistributed probabilities from Equation (1) to

$$\widetilde{P}(x_i \mid y_i = i) = \frac{\left|i - \widetilde{V}_o\right|}{b - a} \times \sum_{j=i-a}^{i+b} \frac{e^{-E(y_j, x_j)}}{\sum_{y1}^{ym} e^{-E(y_j, x_j)}}, \tag{6}$$

Finally, the back propagation error between the real and the estimation was modified from

$$\varepsilon_i = y_i - P(x_i | y_i = i), \tag{7}$$

to

$$\widetilde{\varepsilon}_i = y_i - \widetilde{P}(x_i | y_i = i) \tag{8}$$

The influence of the modified $\widetilde{\varepsilon}_i$ can be calculated as:

$$\phi = \widetilde{\varepsilon}_i - \varepsilon_i \tag{9}$$

then,

$$\phi = \frac{\left|i - \widetilde{V}_o\right|}{b - a} \times \sum_{j=i-a}^{i+b} \frac{e^{-E(y_j, x_j)}}{\sum_{y1}^{ym} e^{-E(y_j, x_j)}} - \frac{e^{-E(y_i, x_i)}}{\sum_{y1}^{ym} e^{-E(y_i, x_i)}} \tag{10}$$

Because a and b are variables, we can find that the distance from a to b is constant. Thus, the left part of formula should be

$$\frac{\left|i - \widetilde{V}_o\right|}{b + a} < 1\,. \tag{11}$$

If probabilities from category $i - a$ to category $i + b$ are same, we will discover that:

$$\phi = \left(\frac{\left|i - \widetilde{V}_o\right|}{b - a} - 1\right) \times \frac{e^{-E(y_i, x_i)}}{\sum_{y1}^{ym} e^{-E(y_i, x_i)}} \tag{12}$$

Therefore, $\phi \leq 0$. Although probabilities from category $i - a$ to category $i + b$ are not always equal, we define the category i is the highest among the entire categories, and we find that:

$$\sum_{j=i-a}^{i+b} \frac{e^{-E(y_j, x_j)}}{\sum_{y1}^{ym} e^{-E(y_j, x_j)}} \leq \frac{e^{-E(y_i, x_i)}}{\sum_{y1}^{ym} e^{-E(y_i, x_i)}}\,. \tag{13}$$

Followed by above formulas, we can conclude that $\phi \leq 0$ during these two above conditions. That means $\widetilde{\varepsilon}_i$ will bring lesser influence into whole neural networks, when considering the globally optimal strategy in fully connected layer.

2) Gradient Related Optimization—We used the cross-entropy function to calculate the error when implementing back propagation step [20]

$$H(y_i, \widetilde{y}_i) = -\frac{1}{m} \times \sum_i [\widetilde{y}_i \log(y_i) + (1 - \widetilde{y}_i) \log(1 - \widetilde{y}_i)] \tag{14}$$

where $\widetilde{y}_i$ is the probability of an evaluated output $\widetilde{P}(x_i | y_i = i)$.

Based on Equations (1), (5), (6), and (14), the gradients of negative or positive log-probabilities in the last layer then would be presented as:

$$\frac{\partial \varepsilon_i}{\partial \theta_i^{(l)}} = \sum_{layer = l} \frac{\partial \left| y_i - \widetilde{P}(x_i \mid y_i = i) \right|}{\partial o_i^{(l)}} \frac{\partial o_i^{(l)}}{\partial \theta_i^{(l)}} \tag{15}$$

where $\theta_i^{(l)}$ is the parameters in the l_{th} layer for category i, $o_i^{(l)}$ is the output lay according to category i. Therefore, we can get the follow formulas from (3), (4), and (15):

$$\frac{\partial \varepsilon_i}{\partial W_i^{(l)}} = \sum_{layer = l} \frac{\partial \left| y_i - \widetilde{P}(x_i \mid y_i = i) \right|}{\partial o_i^{(l)}} \frac{\partial o_i^{(l)}}{\partial \sigma_i^{(l)}} \frac{\partial \sigma_i^{(l)}}{\partial W_i^{(l),}} \tag{16}$$

$$\frac{\partial \varepsilon_i}{\partial b_i^{(l)}} = \sum_{layer = l} \frac{\partial \left| y_i - \widetilde{P}(x_i \mid y_i = i) \right|}{\partial o_i^{(l)}} \frac{\partial o_i^{(l)}}{\partial \sigma_i^{(l)}} \frac{\partial \sigma_i^{(l)}}{\partial b_i^{(l).}} \tag{17}$$

.

Before assembling FFCL into CNNs, the parameters connected with a specific category updated referring to traditional errors, and this type of error can induce a chain reaction for all layers and the inability to lead parameters to the more properly global optimization. For example, the error ε_1 appearing in Grade 1 point which presents in Figure 1 go through every layer from Grade1 point to input layer. The thick green arrow shows the back propagation of ε_1 before assembling FFCL into CNNs, and referring to the typical ReLu functions which can open or close the connection between previous and current layers, $F_{a\,1}$ connects with L_1 and L_2.

However, after embedding FFCL into CNNs, the influence of ε_1 will switch to the global optimized error. The connections swapped to F_1 with L_1 and L_3, because the hyperparameter $\boldsymbol{W1_{FCL}}$ are more likely to close to dispersive solution for neural network training, when compared with the hyperparameter $\boldsymbol{W1_{FCL}}$ optimized by fuzzy strategy. Sometimes, some redundant functions or blocks will appear in neural networks because of the attribute of neural networks. Thus, the blue $L_{z\text{-}1}$ and L_z are the redundant blocks or functions in this system. Function layers in Figure 1 include the traditional structures, for example, convolutional layers, ReLu layers, pooling layers and so on. After such change, the structure of this proposed neural network will be modified, more especially, these refined

ambiguous features extracted by this type of CNNs can to some extent achieve a high decorrelation.

3) Implementation of FFCL—We set the default accuracy rate at 65% before the beginning of FFCL training tasks. If probabilities of normal trend (score [1]) and possibly normal trend (score [2]) are approximately equal, and if they are obviously greater than that of others, for example, the possibility of the abnormal, this appendix can provide relative and significant assistance to classify these two ambiguous categories. In this study, we defined a proper THRESHOLD. If the maximum of output probabilities located among $score[i-1]$ ($i>0$), $score[i]$ and $score[i+1]$, where i is the score from 0 to m, there are three conditions that should be considered by fuzzy rules (in Figure 2).

We also defined $y_{score[i\nearrow 1]}$ and $y_{score[i\searrow 1]}$ as the ascending or descending trends of normal or abnormal respectively. For example, we define that the trend from the abnormal side to the normal side is negative, which means more normal, and if i=2, then $score[i\nearrow 1]$ will be the $score[1]$, and $score[i\searrow 1]$ will be the $score[3]$. Moreover, this FFCL is a nonlinear function, which can partially verify the error-prone condition. If when the probability of $score[1]$ is 0.32, and that of $score[2]$ is 0.33, it will be frequently identified improperly under this condition, as the difference of these two ratios is not obvious. To widen the gap between these two ratios, the value of $score[3]$ should be considered. Then, the back propagation error will be modified by these fuzzy rules with an adaptive parameter based on the probabilities of these three categories. During the update of weights and bias, the effect of will enhance or restrain the learning process of significant features. The key pseudocode is illustrated in Algorithm 1.

Table 1
Algorithm 1. The algorithm of FFCL

The algorithm of FFCL: Before the start of FFCL training, samples and labels should be trained for several epochs.

Start of iteration:

If the accuracy rate is greater than the value, at 0.8 × average accuracy rate (AAR, we previously trained the neural networks and calculated the average accuracy rate), then we start following iteration: for i = 0,1, …,5

If the maximum of probabilities located among score $i-1$ ($i>0$), score i and score $i+1$:

Rule I: If i = {2, 4},

a) when probabilities of $y_{score[i\searrow 1]}$ and $y_{score[i]}$ are greater than the *THRESHOLD*, but $y_{score[i\nearrow 1]}$ is less than *THRESHOLD*. Modify the output of V_o, by (5), where $n=i-2$, $m=i+1$. And get $P_{score[i\searrow 1]}$ and $P_{score[i]}$ by (6).

b) when probabilities of $y_{score[i\nearrow 1]}$ and $y_{score[i]}$ are greater than the *THRESHOLD*, but $y_{score[i\searrow 1]}$ is less than the *THRESHOLD*. Modify the output of CNNs V_o by (5), where $n=i-1$, $m=i$. And get $P_{score[i\nearrow 1]}$ and $P_{score[i]}$ by (6).

Rule II: If i = {3},

a) when probabilities of $y_{score[i\searrow 1]}$ and $y_{score[i]}$ are greater than the *THRESHOLD*, but $y_{score[i\nearrow 1]}$ is less than the *THRESHOLD*. Modify the output of CNNs V_o by (5), where $n = i - 2$, $m = i + 1$. And get $P_{score[i\searrow 1]}$ and $P_{score[i]}$ by (6).

b) when probabilities of $y_{score[i\nearrow 1]}$ and $y_{score[i]}$ are greater than *THRESHOLD*, but $y_{score[i\searrow 1]}$ is less than the *THRESHOLD*. Modify the output of CNNs V_o by (5), where $n = i - 1$, $m = i + 2$. And get $P_{score[i\nearrow 1]}$ and $P_{score[i]}$ by (6).

Rule III: If $i = \{2,3,4\}$,

a) when probabilities of $y_{score[i\searrow 1]}$, $y_{score[i]}$ and $y_{score[i\nearrow 1]}$ are greater than the *THRESHOLD*, and they are approximately equal to each other, modify the output of CNNs V_o by (5) where $n = \mathrm{i} - 1$, $m = i + 1$. And $P_{score[i]}$ will be $max(P_{score[i\searrow 1]}, P_{score[i]}, P_{score[i\nearrow 1]})$, and keep others the same.

End of iteration.

Output: the decimal score and improved CNNs with modified weights and bias.

III Experimental Results

A Dataset and Model Configurations

We used the Curated Breast Imaging Subset of Digital Database of Screening Mammography (CBIS-DDSM) dataset to test our proposed FFCL. The CBIS-DDSM is a large collection of digitized film mammography images, which includes 3,572 images referring to 2689 patient cases. According to BI-RADS, overall BI-RADS assessment from 0 to 5 has been described in this dataset, including BI-RADS score 0 (Incomplete cases), BI-RADS score 1 (Negative cases), BI-RADS score 2 (Benign cases), BI-RADS score 3 (Probably Benign cases), BI-RADS score 4 (Suspicious Abnormal cases) and BI-RADS score 5 (Highly Suspicious Malignant cases), the distribution of which in the CBIS-DDSM dataset is shown in Table 1. Because there are only three normal cases, for triple classification we redistributed three categories, including redefining score 0 as incomplete, combining score 2 with score 3 as benign, and merging scores 4 and score 5 together as malignancy. You can search this type of medical dataset on [21].

As shown in Figure 3, a gray-scale mammogram contains only one gray colour channel, so strategy 1 (S1) used each gray mammogram replicating onto three colour channels. Strategy 2 (S2) applied edge operators to extract margins and stacked them into other two colour channels, while strategy 3 (S3) utilized combination of margins and gray mammogram to submitted other two colour channels. Red lines in Figure 3 edges of mammograms extracted by two basic edge operators, log and canny [22, 23]. Because the ImageNet data has 1000 classes, the last output layer was submitted by a three-class softmax layer, and these three categories consist of incomplete, benign and malignant cases.'

In Figure 4, to simplify the explanation, we defined that the X direction is negative, and its score is 1. The following directions are the same pattern above. There are only two conditions which can be identified with difficulty. The first is that each adjacent category excludes the condition between incomplete cases and negative cases, because incomplete cases approximately have no relationship with other cases. Surfaces XY, yZ, and xy may be difficult to be identified, which means there may be medians between XY, yZ or xy, as their definitions show the high internal relationship. For example, the negative may become the

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

benign in the future, but it actually cannot suddenly transform to high BI-RADS scores, such as probably benign, suspicious abnormality or high suspicious malignancy. Secondly, if there are three probabilities which are approximately equal to each other, such as the probabilities of XYZ, xyZ, we defined that the middle category has the highest probability. Moreover, if these three rectangles seem like that their size on cohorts are not same, that means their probabilities are equal to others.

Many existing CNNs were used in this study (in Table 2), including the 16-layer and 19-layer VGG networks (VGG16 and VGG19) [24], the 18-layer, 50-layer and 101-layer residual networks (ResNet-18, ResNet-50 and ResNet-101) [25], and GoogleNet [26]. Therefore, top layers were designed for whole image classification. In Figure 5, after removal of the 1000-class FCL top layer, six-class FCL or FFCL was stacked behind the top layer in all experiments. However, more convolutional and pooling layers were trained during the second learning phase, and these layers were also added on the top layer. Then during every training task, when the validation rate was reaching the top, the training process was finished and we measured the number of epochs.

B Statistical analysis

Table 2 presents the abbreviation of all plans and the layout of all experiments, for example, S1-ResNet-101-3Conv-FCLbased on FCL. In the plan of S1-ResNet-101- (+3Conv)-FCL, (+3Conv) means that adding the last 3 convolutional layers and training them with FCL together for classifying tasks. NC means the number of classes. Four different learning phases were performed utilizing the CBIS-DDSM dataset in testing the CNN models' recognition capacity for binary, triple and 6-class classifications. RoC curves [27, 28] were generated and aACCs were calculated as a metric of classification accuracy. The confusion matrix, which is a table that can describe the performance of a classification model, was used to test the true values [29]. We used two sample T-test to verify the significance of ACC sequences between two CNNs, and 95% confidence intervals [30] were calculated for ACC values using bootstrapping methods [31]. The deep learning network was implemented using the Matlab platform running on a desktop computer system with the following specifications: Intel Core i7-2670QM CPU@2.20GHZ with 8 GB RAM and a Titan X Pascal Graphics Processing Unit (GPU).

C Networks training strategy

To verify whether the enhancement of visual features is important to improve the performance of CNNs, to validate the advantage of FFCL step-by-step, and to compare with state-of-art, we designed our experiments according to above purposes in this study. Figure 5 explains the structure of training tasks.

Part I – First learning phase—This part determined whether the important visual edge is the significant feature for deep learning improving, and to select the best-performing neural network among these pre-trained CNNs. We stacked two different edges onto two colour channels and then trained these pre-trained neural networks. Depending on pre-trained weights based on the ImageNet database, rather than randomly initialized parameters, these networks were improved by accelerating learning, and more generalizations were

successfully produced to represent features. In this training stage, parameters except the top layer were frozen before training tasks, while simultaneously decreasing the learning rate during training progress. In order to validate whether margin features can be represented or not, S2 and S3 were applied to test these pre-trained VGG-16 and ResNet-101. Table 3 shows that S1-ResNet-101-FCL performed best among residual neural networks, and S1-GoogleNet-FCL was slightly inferior to S1-VGG-16-FCL, which exceeds S1-VGG-19-FCL. After stacking with edges onto the two-colour channels, aACC of S2-ResNet-101-FCL and S2-VGG-16-FCL slightly decreased when compared with these two networks, which only replicated the same mammogram figure. Some researchers' findings supported our result, as they demonstrated that VGG-16 and ResNet-50 have the obvious advantage to classify mammograms [2]. Figure 6 shows ROC curves and confusion matrixes for ResNet-101, VGG-16 and GoogleNet. All categories can be well-distinguished, but the incomplete cases were the most well-defined using both ResNet-101 and VGG-16.

For the confusion matrix of S1-ResNet-101-FCL, of the 46 incomplete cases, this model predicted that 3 cases are benign, and 13 cases are malignant. Of the 2 normal cases, it predicts that all were malignant. And of the 194 benign cases, 4 cases are attached to incomplete, 100 cases are predicted to belong to benign, and the last 90 cases are deemed to be malignant. Of the 462 malignant cases, it predicts that 11 cases are incomplete, 52 cases are benign, and 399 cases are malignant. As the matrix shown in Figure 6, both ResNet-101 and VGG-16 have the disadvantage to distinguish malignancy from benign; but both networks can make obvious distinction between incomplete cases and other types of cases. Among the six CNNs, ResNet-101 performed best, followed by ResNet-50, ResNet-18, VGG-16, GoogleNet and VGG-19 in sequence.

While in the first learning phase, all these CNNs can satisfactorily distinguish each BI-RADS assessment, but only training of the FCL may result in some features that cannot be extracted by pre-trained blocks. Due to there still being some important features that should be represented by CNNs, the larger dataset size or something intrinsic to the characteristics of the DDSM dataset should be represented by our models, therefore, only training FCL is insufficient. According to incomplete cases, which have to be re-examined radiologically, the lack of information for diagnosis can inform CNNs that these kinds of mammograms have insufficient features and should be re-examined. All 6 types of CNNs have an encouraging advantage to identify incomplete and malignant cases, but are less efficient at recognizing malignant from benign cases. Although doctors often disagree on how a particular exam should be classified [32] and less than 1% of the screening population has cancer [32, 33], researchers expect that this problem can be alleviated by using the information about whether a person proceeded to develop breast cancer in the future as an identifier [34]. Even if the mammogram is identified as normal or benign, the incomplete cases may become the mortal potential for patients, therefore, a high rate of incomplete cases' recognition can make diagnosis more reliable. In Table 4, although three different strategies had been utilized, S1 presents the best performance, which provides the evidence that the enhancement of margins in mammograms will result in the graphic degeneration when using ResNet-101 and VGG-16.

To construct a better neural network structure during the following experiments, we subsequently designed the second and the third learning phases through S1 in those following steps. S2 may discard some significant features, and this is the reason why the ACC array of using S1 is significantly greater than that of using S2 ($P < 0.01$). Sometimes, the enhancement of margins for mammograms will result in overfitting, as the ACC array of using S3 is significantly less than that of using S1 ($P < 0.01$). If CNNs cannot efficiently extract margins, there may be no overfitting during this experiment, because these margin features have been reinforced. Thus, this can explain why deep learning can represent features that radiologists may not distinguish.

For traditional computer-aided detection or diagnosis, predefined features are usually used for constructing models, which require pre-emptive determination of which features will contribute to classification tasks [35]. However, in our study, we believe that predefinition of the graphic features is not necessary, and before our training tasks based on CBIS-DDSM, these visible features have already been automatically represented by ImageNet dataset [36]. Obvious features, such as margins, can be recognized by radiologists, and it also can be detected by the learning process, while intrinsic and invisible features which are used for imaging interpretation may not be identified by human beings also can be automatically recognized by CNNs [35].

Many studies have shown the advantage of transfer learning to process limited medical data [37]. We provide deeper insights in developing optimized transfer learning strategies by designing training experiments. However, the incremental transfer learning and the observations made here need to be evaluated by further analyses and comparative studies.

Part II – The second learning phase—The second part aimed to train the last convolutional layers, or add and train the last convolutional layers. After the FCL was removed, parameters from the bottom layers to the final or penultimate VGG and residual blocks were frozen, and the remaining weights and bias were trained and updated in the neural network. By contrast, we also respectively added one or two VGG and residual blocks on the top layer and only trained them to learn features. Because some ambiguous features between two adjacent categories were difficult to identify, we used the FFCL to improve the performance of ResNet-101. In the second learning phase, to carry out which kind of structure will perform best, we selected ResNet-101 and VGG-16 to complete these training tasks. In Table 5, the best single ACC rate of ResNet-101 is 76.82% which was recorded during one best-performed single training task, but significantly greater than that of ACC array in the variance of ACC array in S1-ResNet-101-3Conv-FCL is S1-ResNet-101-FCL ($P < 0.01$). Almost the same performance was found between S1-ResNet-101-3Conv-FCL and S1-ResNet-101-FFCL ($P = 0.243$). On the contrary, training 6 last layers (residual or VGG blocks) made the ACC array dropped significantly ($P < 0.01$).

After adding VGG or residual convolutional blocks onto the top layer, Table 5 shows that the ACC array of S1 -ResNet-101-(+3Conv)-FCL is higher than that of S1-ResNet-101-(+6Conv)-FCL ($P = 0.032$), which indicates that S1-ResNet-101-(+3Conv)-FCL can represent more features. Compared the performance of S1-ResNet-101-FCL and S1-ResNet-101-3Conv-FCL, S1-ResNet-101-FFCL based on FFCL algorithm can significantly

increase ACC array. Moreover, the advantage of S1-ResNet-101-3Conv-FFCL is obvious ($P < 0.01$), therefore, the influence of FFCL can to some extent improve the structures we designed above. To teach and train computers about how to recognize can sometimes achieve extraordinary success. If the algorithm is not able to be strictly recognized as the 'over uncertainties averaged log membership', fuzzy systems can enhance the probability of distinguishing uncertain features [11].

Some traditional machine learning methods, such as Naïve Bayes [38], support vector machine (SVM) [39] and random forest (RF) [40] were utilized to compare each CNNs. Although random forest can reach the same performance of S1-ResNet-101-3Conv-FFCL, but the best single ACC was different. The improved ResNet-101 through updated structures and fuzzy rules has the potential to outperform these traditional machine learning methods.

Part III – The third learning phase—One aim of this part was to check whether the plan about combining score 2 with score 3 as benign and score 4 with score 5 as malignancy will influence the ACC and classification performance. Another aim was to construct an FFCL based on ResNet-101 for 6-class classification.

The first task we designed only trained the FCL. The second task used the FFCL with ResNet-101. The third task applied the structure which performed best in the second learning phase to identify BI-RADS assessment, and the last task was based on the combination of the second and the third tasks.

According to some categories, which are difficult to identify in confusion matrices above, the third learning phase utilized fuzzy rules to improve the neural networks' quality after neural networks can partially identify some classes. Finally, in order to evaluate the distance between our trained models and the convergent globally optimal solution, we used the Euclidean Distance [41] (ED) to measure the effect of classification performance by neural networks:

$$Gd\left(p, q\right) = \sqrt{\frac{\sum_{i=1}^{m} (p_i - q_i)^2}{m}} \tag{18}$$

where p is the value of predicted scores, q is the value of realistic scores, and i is the category. When the output is a decimal, not an integer, it means a mammogram contains uncertain features, and the CNNs will provide radiologists probabilities and decimal scores. The reason why we chose ED to measure the advantage of FFCL is that the measuring distance can be evaluated by t test. The learning rate reduction helped us avoid unlearning important features.

D Comparison with other existing methods

For the CBIS-DDSM medical dataset, after embedding the FFCL into some CNNs, Table 8 shows the significant advantage of this semantic fuzzy layer when comparing with no FFCL before. Although classifying medical images is difficult to implement, as their poor quality, collaboration with relative information can enhance the performance of CNNs, which

indicate that artificial neural networks need basic and inherent knowledge to enrich themselves, and to limit them to overstep the boundary, for example, overfitting.

IV Conclusion

In this study, we verified that the visual enhancement method cannot substantially improve the classification performance, and we provided an evidence that the transfer learning strategy, especially trained by natural categories, can extract medical features. A novel architecture which is based on fuzzy system and embedded in the fully connected layer for scoring images is designed for semantic scoring of medical images.

This proposed optimal structure demonstrates its advantage in CBIS-DDSM dataset for BI-RADS scoring. We firstly proved the mathematical availability of the FFCL and designed three learning phases to gradually develop CNNs based on FFCL. Our proposed framework can also shrink the overlap semantic space explored under an adaptive weight updating environment in this medical dataset.

This FFCL architecture offers the advantage of weakening the influence of equivocal and unclear semantic description for medical diagnosis. Although this architecture can positively deal with the classification tasks which have overlaps between two neighbour classes, it is more likely to weaken the influence of semantic conglutination.

The future work will focus on classifying images annotated by linear categories and relying on another assistant CNN to simulate the cognitive activation of human brain, such as inhibition, disinhibition and maintenance.

Acknowledgments

The work was supported under Royal Society International Exchanges Cost Share Award (RP202G0230), Medical Research Council Confidence in Concept (MRC CIC) Award, Hope Foundation for Cancer Research (RM60G0680), UK.

Biographies

Cheng Kang is a PhD student from the Department of Informatics, University of Leicester, UK. He received his master degree from Shenzhen University. His research interests focus on EEG signal processing, deep learning. Currently, he is working on the early detection of breast cancer by artificial intelligence.

Europe PMC Funders Author Manuscripts

Xiang Yu received his bachelor degree and master degree from Huanggang Normal University and Xiamen University, P.R. China, in 2014 and 2018, respectively. Currently, he is a Ph.D. student in the Department of Informatics, University of Leicester, UK. Also, he was sponsored by CSC and by the University of Leicester as a graduate teaching assistant (GTA). His research interests include medical image segmentation, machine learning and deep learning.

Dr. **Shui-Hua Wang** received her Ph.D. degree from Nanjing University at 2017. She worked as an assistant professor in Nanjing Normal University from 2013-2018. Now she is working as a Research Associate at Loughborough University, and as Research Associate at University of Leicester. She published more than 100 papers, 15 were included as “ESI highly cited paper”. She was rewarded “Highly cited researcher 2019” by Web of Science.

Dr **David Guttery**’s research interests are intertwined with those of Professor Jacqui Shaw (see Professor Jacqui Shaw’s webpage), which focus on the utility of circulating nucleic acids and other circulating biomarkers for early detection and monitoring of cancer. He is currently a co-investigator on an integrated, collaborative programme of clinical and translational research between the University and Leicester and Imperial College funded by Cancer Research UK.

Prof. **Hari Mohan Pandey** received the B.Tech. degree from Uttar Pradesh Technical University, India, the M.Tech. degree from the Narsee Monjee Institute of Management Studies, India, and the Ph.D. degree computer science and engineering from the Amity University, India. He worked as a Postdoctoral Research Fellow with the Middlesex University, London, U.K. He also worked on a European Commission project - Dream4car under H2020. He is a Senior Lecturer with the Department of Computer Science, Edge Hill University, Lancashire, U.K.

Yingli Tian (M'99-SM'01-F'18) received the B.S. and M.S. degrees from Tianjin University, China, in 1987 and 1990, and the Ph.D. degree from Chinese University of Hong Kong, Hong Kong, in 1996. After holding a faculty position at National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, she joined Carnegie Mellon University in 1998, where she was a postdoctoral fellow at the Robotics Institute. She then worked as a research staff member in IBM T. J. Watson Research Center from 2001 to 2008. She is one of the inventors of the IBM Smart Surveillance Solutions. She is currently a CUNY Distinguished Professor in City University of New York. She is a fellow of IEEE.

Prof. **Yu-Dong Zhang** received his PhD degree from Southeast University in 2010. He worked as a postdoc from 2010 to 2012 in Columbia University, USA, and as an assistant research scientist from 2012 to 2013 at Research Foundation of Mental Hygiene (RFMH), USA. He served as a full professor from 2013 to 2017 in Nanjing Normal University, where he was the director and founder of Advanced Medical Image Processing Group in NJNU. Now he serves as Professor in Department of Informatics, University of Leicester, UK.

Europe PMC Funders Author Manuscripts

References

[1]. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28.521:436–444. [PubMed: 26017442]

[2]. Shen L. End-to-end Training for Whole Image Breast Cancer Diagnosis using An All Convolutional Design. arXiv preprint arXiv:1708.09427. 2017

[3]. Buckley JJ, Hayashi Y. Fuzzy Neural Networks - a Survey. Fuzzy Sets and Systems. 1994 Aug 25.66:1–13.

[4]. Lin CT, Yeh CM, Liang SF, Chung JF, Kumar N. Support-vector-based fuzzy neural network for pattern classification. Ieee Transactions on Fuzzy Systems. 2006 Feb.14:31–41.

[5]. Kumar M, Chatterjee S, Zhang W, Yang J, Kolbe LM. Fuzzy theoretic model based analysis of image features. Information Sciences. 2019; 480:34–54.

[6]. Kumar M, Insan A, Stoll N, Thurow K, Stoll R. Stochastic Fuzzy Modeling for Ear Imaging Based Child Identification. Ieee Transactions on Systems Man Cybernetics-Systems. 2016 Sep.46:1265–1278.

[7]. Lin FJ, Lin CH, Shen PH. Self-constructing fuzzy neural network speed controller for permanent-magnet synchronous motor drive. Ieee Transactions on Fuzzy Systems. 2001 Oct.9:751–759.

[8]. Vinyals O, Le Q. A neural conversational model. arXiv preprint arXiv:1506.05869. 2015

[9]. Andreas, J; Rohrbach, M; Darrell, T; Klein, D. Neural module networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. 39–48.

[10]. Poria S, Cambria E, Gelbukh A. Aspect extraction for opinion mining with a deep convolutional neural network. Knowledge-Based Systems. 2016; 108:42–49.

[11]. Kumar M, Mao YH, Wang YH, Qiu TR, Yang C, Zhang WP. Fuzzy theoretic approach to signals and systems: Static systems. Information Sciences. 2017 Dec.418:668–702.

[12]. Deng Y, Ren ZQ, Kong YY, Bao F, Dai QH. A Hierarchical Fused Fuzzy Deep Neural Network for Data Classification. Ieee Transactions on Fuzzy Systems. 2017 Aug.25:1006–1012.

[13]. Ding W, Lin C-T, Cao Z. Deep Neuro-Cognitive Co-Evolution for Fuzzy Attribute Reduction by Quantum Leaping PSO With Nearest-Neighbor Memeplexes. IEEE Transactions on Cybernetics. 2018

[14]. Obenauer S, Hermann KP, Grabbe E. Applications and literature review of the BI-RADS classification. European Radiology. 2005 May.15:1027–1036. [PubMed: 15856253]

[15]. Cambria E, Olsher D, Rajagopal D. SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. Twenty-eighth AAAI conference on artificial intelligence. 2014

[16]. Cambria E, White B. Jumping NLP curves: A review of natural language processing research. IEEE Computational intelligence magazine. 2014; 9:48–57.

[17]. Prasad M, Lin C-T, Li D-L, Hong C-T, Ding W-P, Chang J-Y. Soft-boosted self-constructing neural fuzzy inference network. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2017; 47:584–588.

[18]. Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER, et al. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. New England Journal of Medicine. 2002 Oct 17.347:1233–1241.

[19]. Chen CP, Zhang C-Y, Chen L, Gan M. Fuzzy restricted Boltzmann machine for the enhancement of deep learning. IEEE Transactions on Fuzzy Systems. 2015; 23:2163–2173.

[20]. Humpert BK. Improving back propagation with a new error function. Neural networks. 1994; 7:1191–1192.

[21]. Smith K, Nolan T. CBIS-DDSM. 2019

[22]. Canny J. A Computational Approach to Edge-Detection. Ieee Transactions on Pattern Analysis and Machine Intelligence. 1986 Nov.8:679–698. [PubMed: 21869365]

[23]. Sotak GE Jr, Boyer KL. The Laplacian-of-Gaussian kernel: a formal analysis and design procedure for fast, accurate convolution and full-frame output. Computer Vision, Graphics, and Image Processing. 1989; 48:147–189.

[24]. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014

[25]. He, KM; Zhang, XY; Ren, SQ; Sun, J. Deep Residual Learning for Image Recognition. 2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr); 2016. 770–778.

[26]. Szegedy, C; Liu, W; Jia, Y; Sermanet, P; Reed, S; Anguelov, D; , et al. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. 1–9.

[27]. Metz CE. ROC methodology in radiologic imaging. Investigative radiology. 1986; 21:720–733. [PubMed: 3095258]

[28]. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006; 27:861–874.

[29]. Stehman SV. Selecting and interpreting measures of thematic classification accuracy. Remote sensing of Environment. 1997; 62:77–89.

[30]. León-Domínguez U, Martín-Rodríguez JF, Leon-Carrion J. Executive n-back tasks for the neuropsychological assessment of working memory. Behavioural brain research. 2015; 292:167–173. [PubMed: 26068585]

[31]. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011; 12:77. [PubMed: 21414208]

[32]. Tabár L, Vitak B, Chen HHT, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. Cancer: Interdisciplinary International Journal of the American Cancer Society. 2001; 91:1724–1731.

[33]. Duffy SW, Tabár L, Chen HH, Holmqvist M, Yen MF, Abdsalah S, et al. The impact of organized mammography service screening on breast carcinoma mortality in seven Swedish counties: a collaborative evaluation. Cancer: Interdisciplinary International Journal of the American Cancer Society. 2002; 95:458–469.

[34]. Geras KJ, Wolfson S, Shen Y, Kim S, Moy L, Cho K. High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv:1703.07047. 2017

[35]. Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep Learning to Distinguish Recalled but Benign Mammography Images in Breast Cancer Screening. Clinical Cancer Research. 2018; 24:5902–5909. [PubMed: 30309858]

[36]. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision. 2015; 115:211–252.

[37]. Zhou, Z; Shin, J; Zhang, L; Gurudu, S; Gotway, M; Liang, J. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. IEEE conference on computer vision and pattern recognition, Hawaii; 2017. 7340–7349.

[38]. Rish, I. An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence; 2001. 41–46.

[39]. Joachims T. Making large-scale SVM learning practical. 1998

[40]. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002; 2:18–22.

[41]. Anton, H. Elementary Linear Algebra, Binder Ready Version. John Wiley & Sons; 2013.

[42]. Geras KJ, Wolfson S, Shen Y, Wu N, Kim S, Kim E, et al. High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv:1703.07047. 2017

[43]. Akselrod-Ballin, A, Karlinsky, L, Alpert, S, Hasoul, S, Ben-Ari, R, Barkan, E. “A region based convolutional network for tumor detection and classification in breast mammography,”Deep Learning and Data Labeling for Medical Applications. Springer; 2016. 19

Europe PMC Funders Author Manuscripts

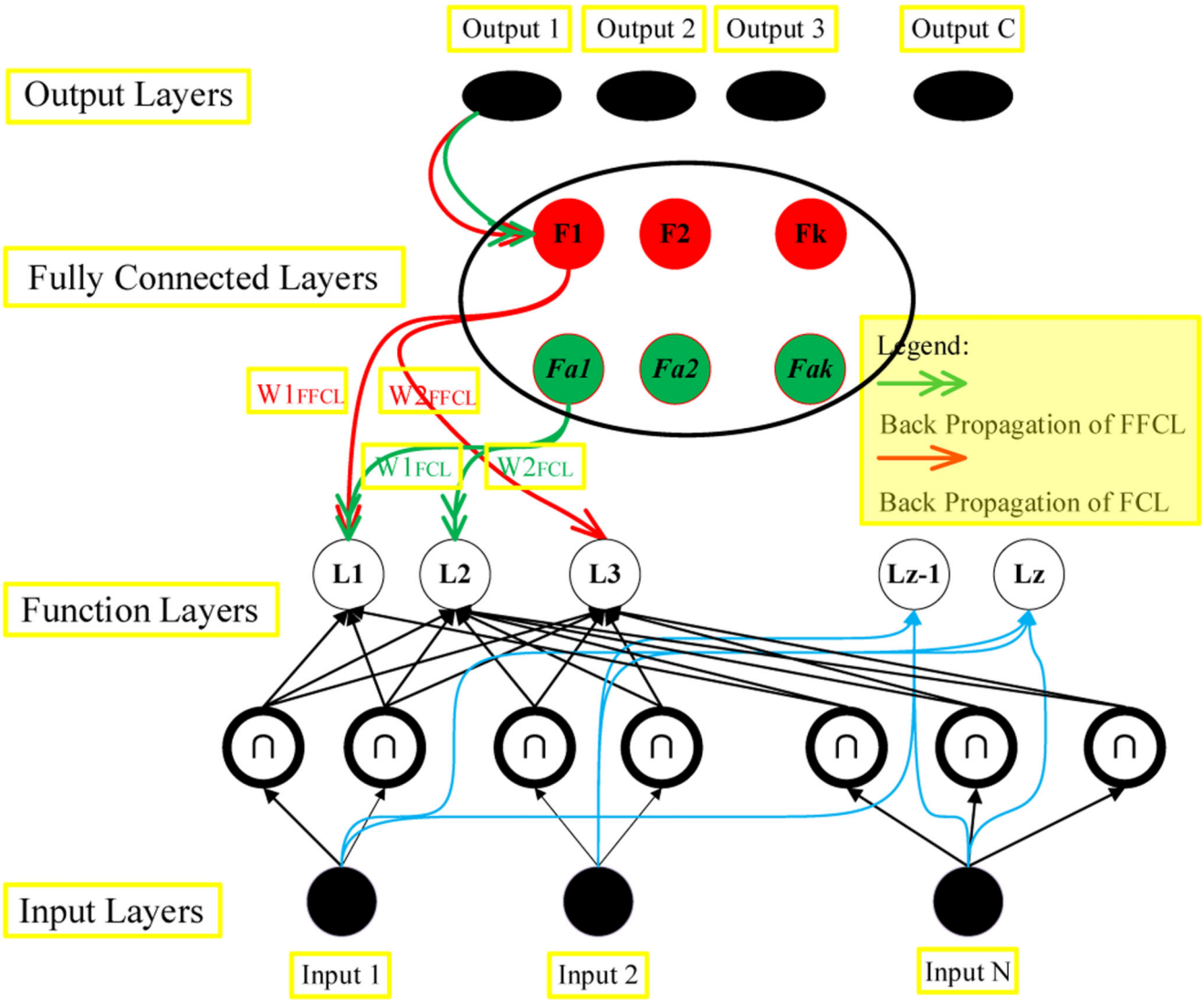

Figure 1.

Conceptual explanation of the FFCL's structure in CNNs. It is composed of four parts, input layer function layers, fuzzy transformation, and output layer.

Europe PMC Funders Author Manuscripts

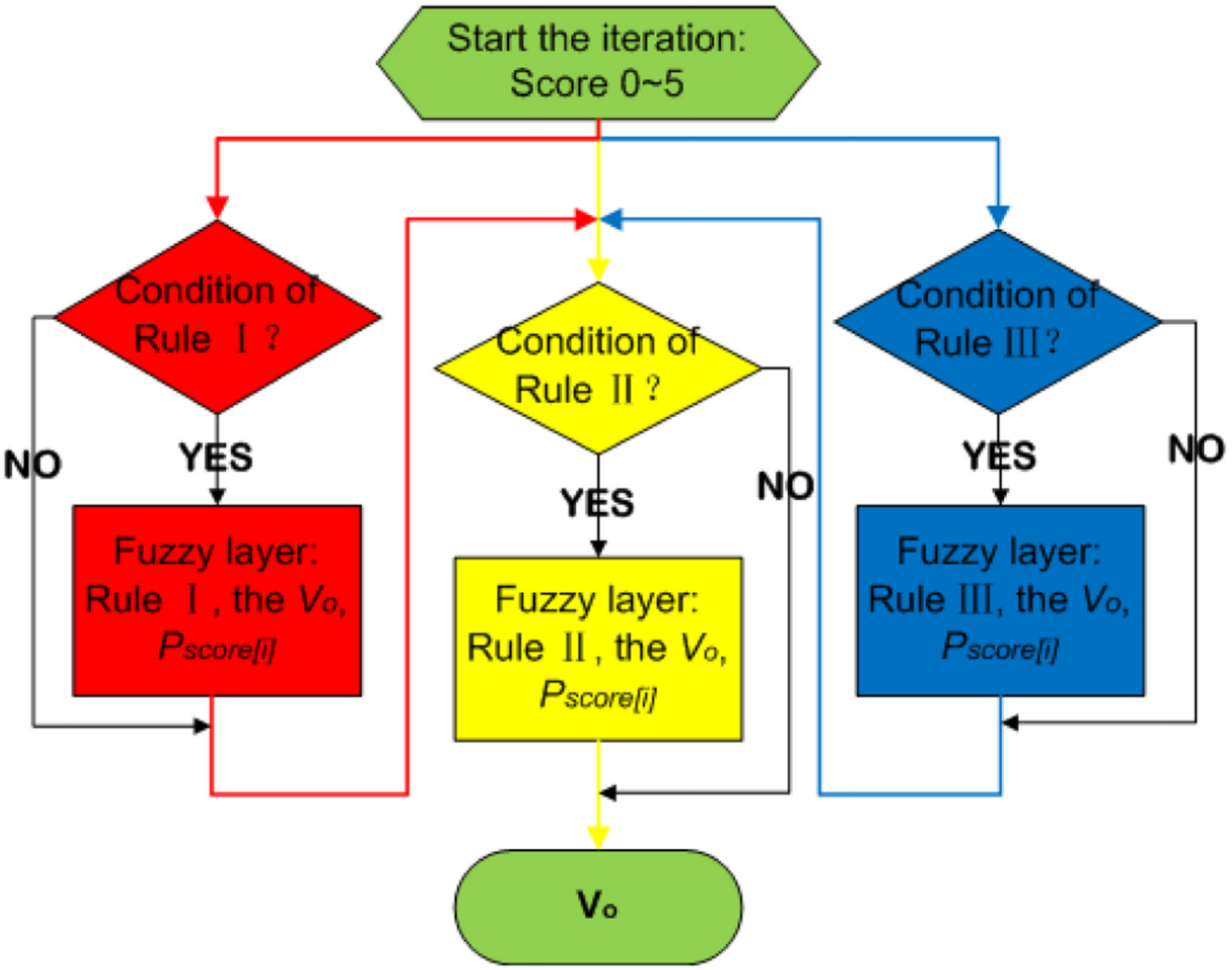

Figure 2.
The core structure of FFCL.

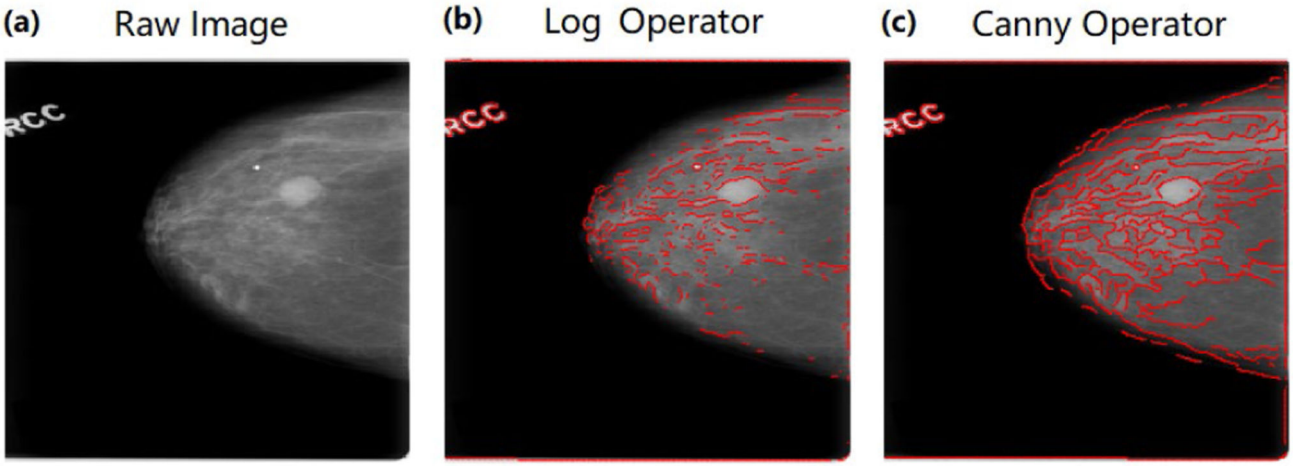

Figure 3.
Mammograms from CBIS-DDSM with the log and canny edges. (a), The gray mammogram. (b), The gray mammogram with log edges. Red lines are log edges. (c), The gray mammogram with canny edges. Red lines are the canny edges.

Europe PMC Funders Author Manuscripts

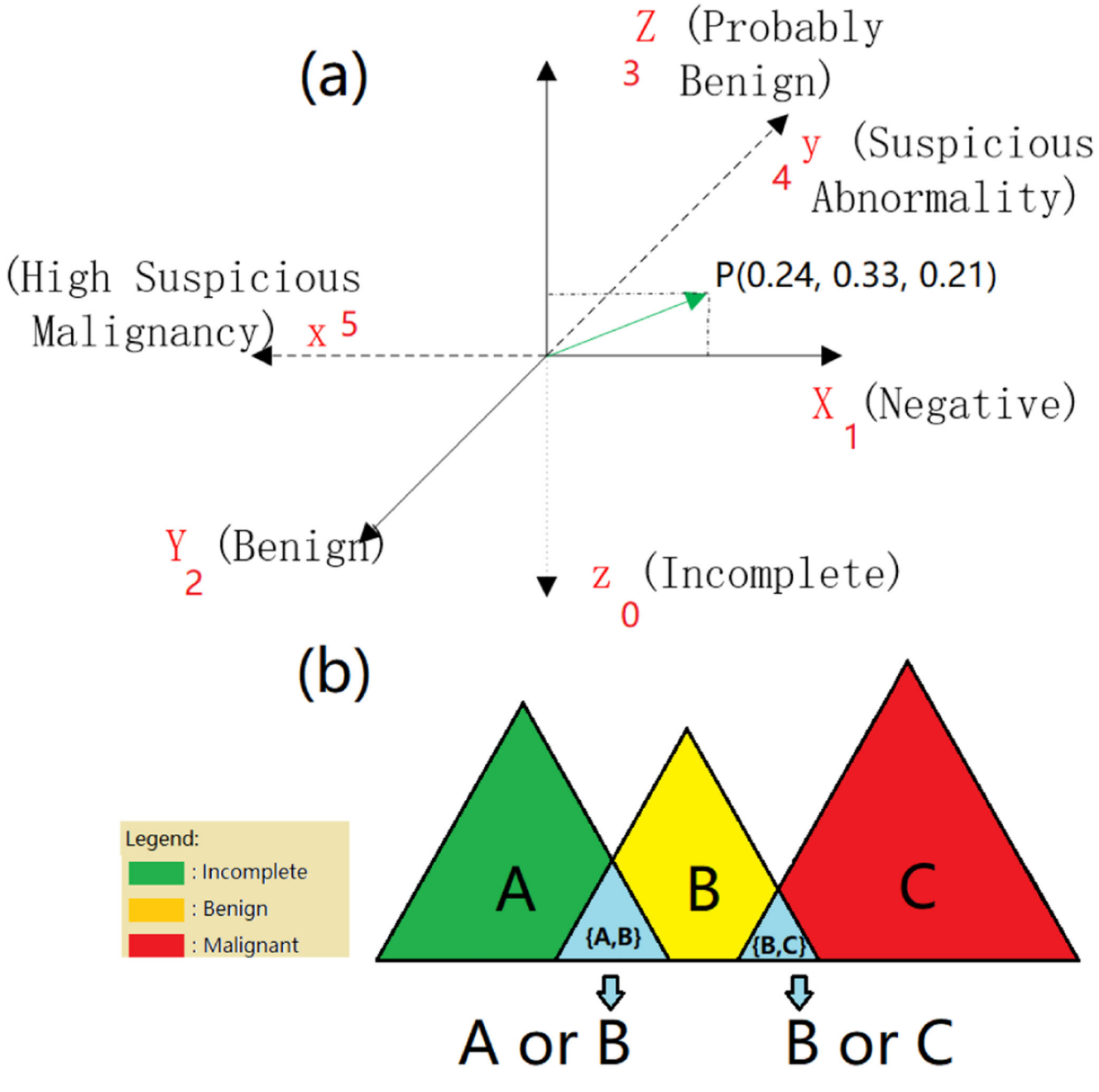

Figure 4.
Clustering arrangement allowing overlap and selecting the scores according to the labels (or classes) attached to them.

Europe PMC Funders Author Manuscripts

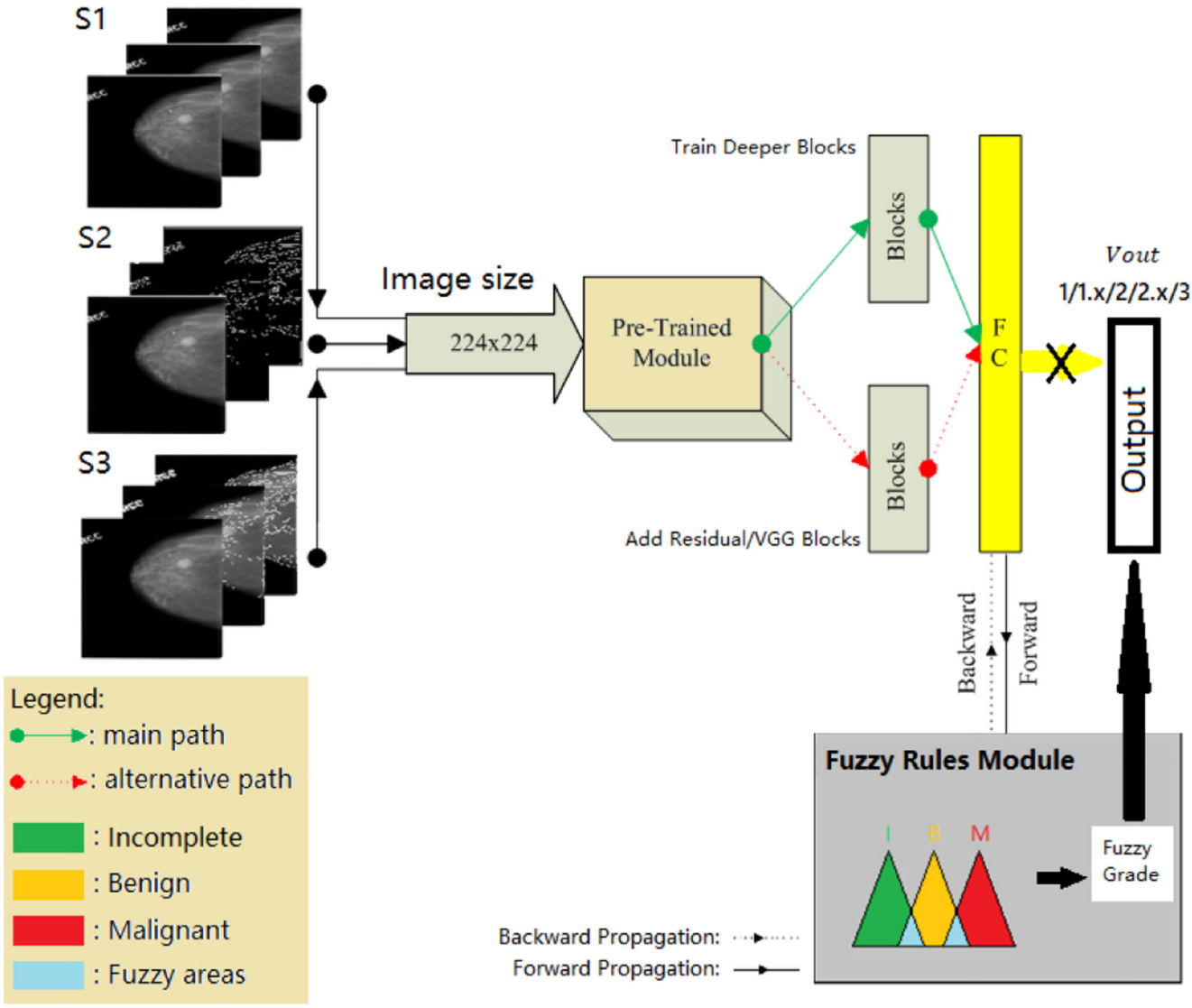

Figure 5.

The deep learning structure for improving CNNs by training deeper last convolutional layers, adding and training the last residual convolutional layers, and establishing fuzzy rules block on top.

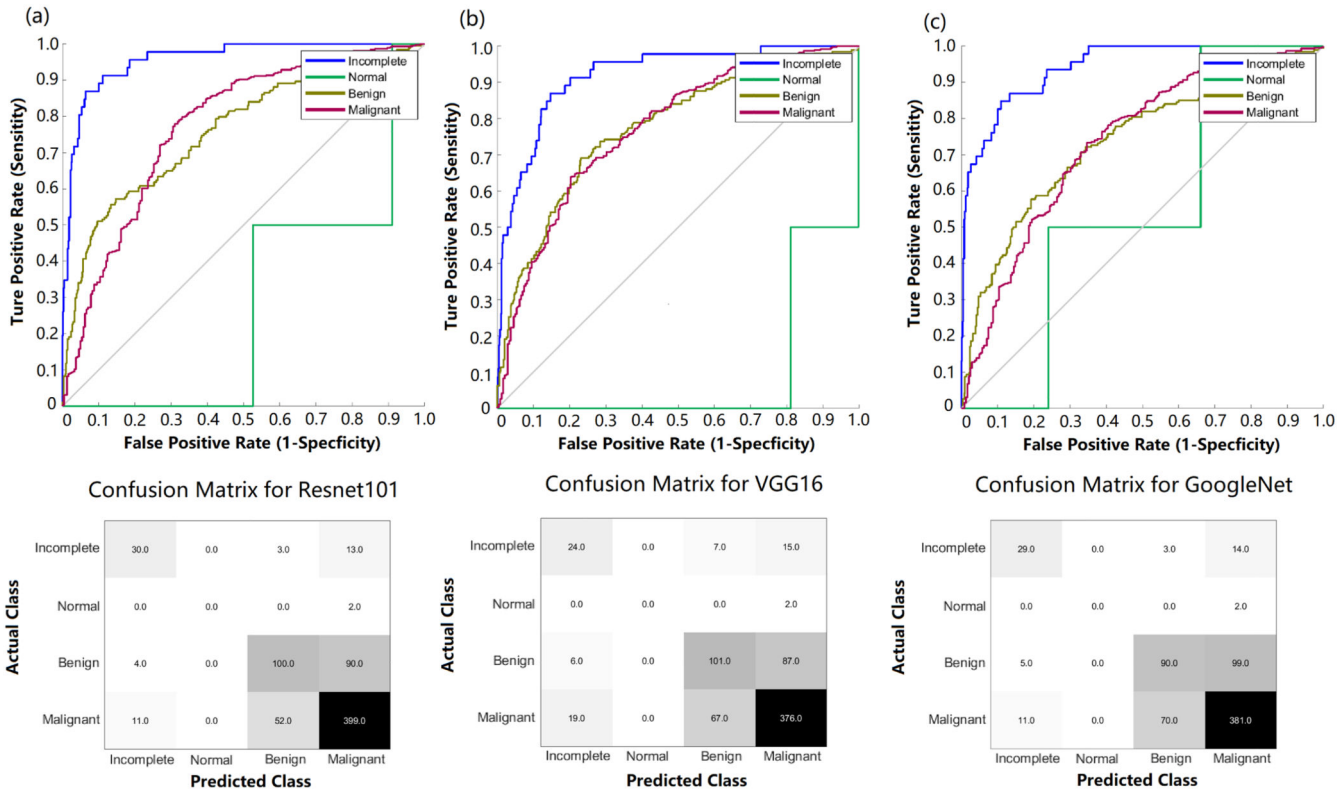

Figure 6.

The ACC curve and confusion matrixes. From left to right successively (a) S1-ResNet-101-FCL, (b) S1-VGG-16-FCL and (c) S1-GoogleNet-FCL in the first learning phase for the triple classification scenarios based on CBIS-DDSM dataset.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

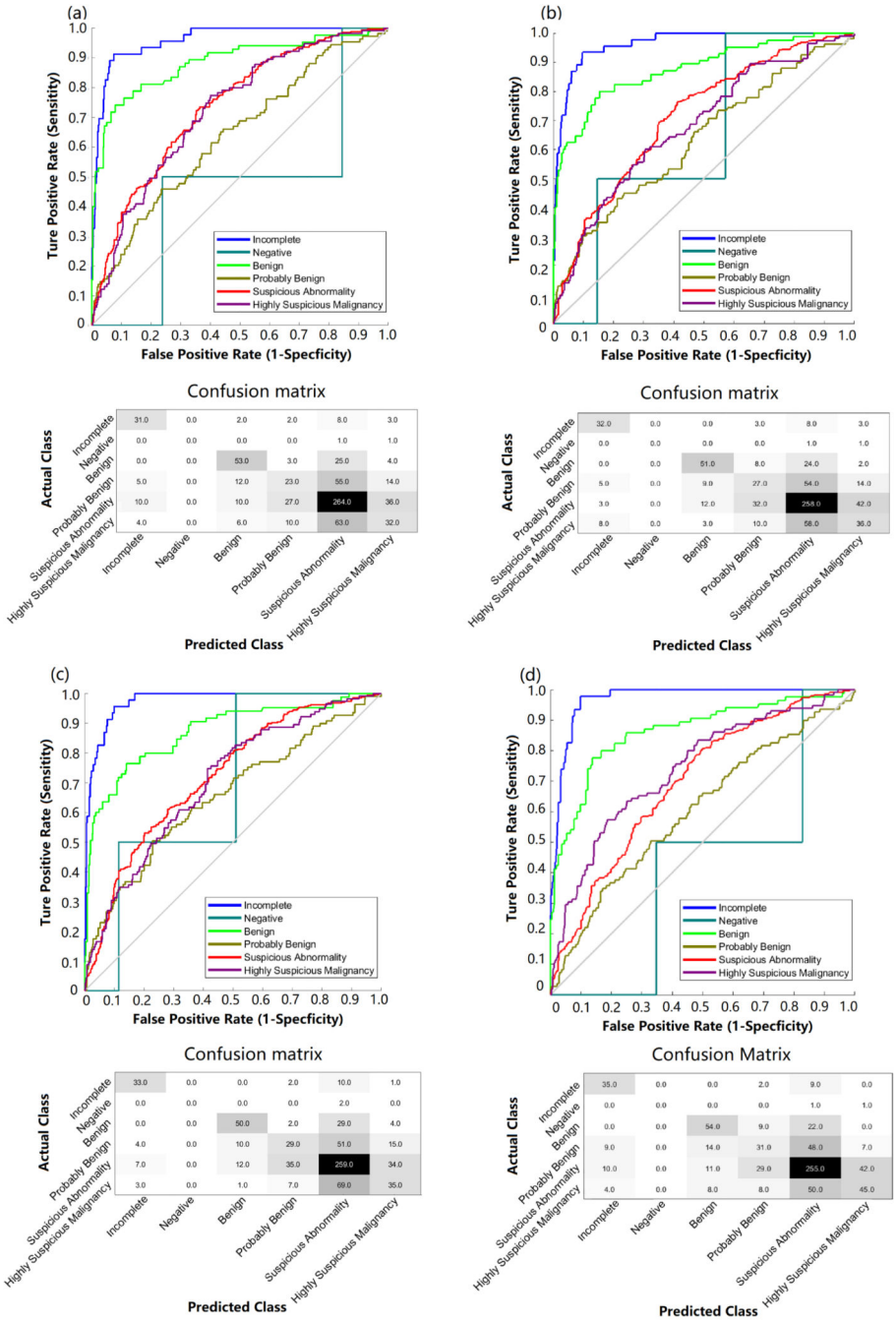

Figure 7.

ROC curves and confusion matrixes. From left to right successively (a) S1-ResNet-101-FCL, (b) S1-ResNet-101-FFCL and (c) S1-ResNet-101-3Conv-FCL and (d) S1-ResNet-101-3Conv-FFCL for 6-class classification (Incomplete, Negative, Benign, Probably Benign, Suspicious Abnormality and Highly Suspicious Malignancy) on the CBIS-DDSM dataset in the third learning phase.

Europe PMC Funders Author Manuscripts

Table 1

The distribution of samples of CBIS-DDSM dataset based on BI-RADS assessment

	0	1	2	3	4	5
Mass + Calcification Training Set	192 (129+63)	1	559 (77+482	368 (279+89)	1286 (533+753)	458 (299+159)
Mass + Calcification Testing Set	46 (33+13)	2	85 (14+71)	109 (85+24)	347 (169+178)	115 (75+40)

Table 2

Experiments and the structure of neural networks in this study.

Abbreviation	NC	Strategy	Base Network	Retrained layers
First Learning Phase				
S1-ResNet-18-FCL	4	S1	ResNet-18	FCL
S1-ResNet-50-FCL	4	S1	ResNet-50	FCL
S1-ResNet-101-FCL	**4**	**S1**	**ResNet-101**	**FCL**
S1-VGG-16-FCL	**4**	**S1**	**VGG-16**	**FCL**
S1-VGG-19-FCL	4	S1	VGG-19	FCL
S1-GoogleNet-FCL	4	S1	GoogleNet	FCL
S1-ResNet-101-FCL	**4**	**S1**	**ResNet-101**	**FCL**
S1-VGG-16-FCL	4	S1	VGG-16	FCL
S2-ResNet-101-FCL	4	S2	ResNet-101	FCL
S2-VGG-16-FCL	4	S2	VGG-16	FCL
S3-ResNet-101-FCL	4	S3	ResNet-101	FCL
S3-VGG-16-FCL	4	S3	VGG-16	FCL
Second Learning Phase				
S1-ResNet-101-3Conv-FCL	**4**	**S1**	**ResNet-101**	**3Conv-FCL**
S1 -ResNet-101 -3Conv-FCL	4	S1	ResNet-101	6Conv-FCL
S1 -ResNet-101 -(+3Conv)-F CL	4	S1	ResNet-101	(+3Conv)-FCL
S1-ResNet-101-(+6Conv)-FCL	4	S1	ResNet-101	(+6Conv)-FCL
S1-Naive Bayes	4	S1	NO	NO
S1-SVM	4	S1	NO	NO
S1-Random Forest	4	S1	NO	NO
Third Learning Phase				
S1-ResNet-101-FFCL	4	S1	ResNet-101	FFCL
S1-ResNet-101-3Conv-FFCL	**4**	**S1**	**ResNet-101**	**3Conv-FFCL**
S1-ResNet-101-FCL	6	S1	ResNet-101	FCL
S1-ResNet-101-FFCL	6	S1	ResNet-101	FFCL
S1 -ResNet-101 -3Conv-FCL	6	S1	ResNet-101	3Conv-FCL

Europe PMC Funders Author Manuscripts

Abbreviation	NC	Strategy	Base Network	Retrained layers
S1-ResNet-101-3Conv-FFCL	6	S1	ResNet-101	3Conv-FFCL

Europe PMC Funders Author Manuscripts

Table 3

The learning phase used pre-trained neural networks.

Model	ACC (Best)	Epochs	aACC / (ACC range)
S1-ResNet-18-FCL	73.30%	27	72.62%/ (72.16%~73.30%)
S1-ResNet-50-FCL	74.15%	27	73.62%/ (72.73%~74.15%)
S1-ResNet-101-FCL	**75.71%**/ 71.42%	**28** 15	**74.76%/(73.72%~75.71%)**
S1-VGG-16-FCL	**72.02%**	**25**	**70.95%/ (69.89%~72.02%)**
S1-VGG-19-FCL	70.31%	29	69.44%/ (68.32%~70.31%)
S1-GoogleNet-FCL	71.73%	28	70.82%/ (70.03%~71.73%)

Europe PMC Funders Author Manuscripts

Table 4

Considering to reinforce the visible features, we only trained the FCL based on the pre-trained ResNet-101 and pre-trained VGG-16 in this task.

Model	ACC (Best)	Epochs	aACC / (ACC range)
S1-ResNet-101-FCL	**75.71%**	**28**	**74.76%/ (73.72%~75.71%)**
S1-VGG-16-FCL	72.02%	25	70.95%/ (69.89%~72.02%)
S2-ResNet-101-FCL	72.73%	25	71.51%/ (70.60%~72.73%)
S2-VGG-16-FCL	69.32%	28	68.41%/ (67.76%~69.32%)
S3-ResNet-101-FCL	70.03%	26	69.13%/ (68.18%~70.03%)
S3-VGG-16-FCL	71.16%	28	70.11%/ (69.32%~71.16%)

Europe PMC Funders Author Manuscripts

Table 5

The performance of adding and retraining three or six last layers.

Model	ACC (Best)	Epochs	aACC / (ACC range)
S1-ResNet-101-3Con v-FCL	**76.70%**	**16**	**74.12%/ (70.88%~76.70%)**
S1-VGG-16-3Conv-FCL	64.55%	5	
S1-ResNet-101-6Conv-FCL	74.29%	16	71.46%/(69.32%~74.29%)
S1-ResNet-101-(+3Conv)-FCL	74.72%	18	72.78%/(69.74%~74.72%)
S1-VGG-16-(+3Conv)-FCL	62.43%	6	
S1-	74.57%	16	72.12%/
ResNet-101-(+6Conv)>-FCL			(69.89%~74.57%)
S1-ResNet-101-FFCL	75.92%	25	75.15%/(74.58%~75.92%)
S1-ResNet-101-3Conv-FFCL	**76.82%**	**15**	**74.15%/(70.68%~76.82%)**
S1-Naïve Bayes	41.19%		
S1-SVM	61.93%		
S1-Random Forest	**75.99%**		

Europe PMC Funders Author Manuscripts

Table 6

The performance of ResNet-101 for 6-class classification.

Model	ACC (Best)	Epochs	aACC / (ACC range)
S1-ResNet-101-FCL	57.53%	16	56.34%/(55.54%~57.53%)
S1-ResNet-101-FFCL	57.88%	16	56.99%/(55.97%~57.88%)
S1 -ResNet-101 -3Conv-FCL	59.09%	17	56.86%/(54.55%~59.09%)
S1-ResNet-101-3Conv-FFCL	**59.94%**	**17**	**57.40%/(55.11%~59.94%)**

Europe PMC Funders Author Manuscripts

Table 7

The Euclidean Distance among S1-ResNet-101-FCL, S1-ResNet-101-FFCL and S1 - ResNet-101 -3Conv-FFCL.

Models	Ed (average and variance)	P-value
S1-ResNet-101-FCL < S1-ResNet-101-FFCL	1.466+0.0104~ 1.482+0.0120	0.0045
S1-ResNet-101-FCL < S1-ResNet-101-3Conv-FFCL	1.466+0.0104~ 1.455+0.0464	0.4380
S1-ResNet-101-FFCL < S1-ResNet-101-3Conv-FFCL	1.482+0.0120~ 1.455+0.0464	0.0791

Europe PMC Funders Author Manuscripts

Table 8

Comparison with existing methods on DDSM in terms of ACC.

	DDSM (Scenario)	
Methods	**CNN+FCL**	**CNN+FFCL (proposed)**
Geras [42]	68.8% (BI-RADS 0/1/2)	70.1% (BI-RADS 0/1/2)
Akselrod-Ballin [43]	60.0%(BI-RADS 2/(3-4-5))	62.3%(BI-RADS 2/(3-4-5))
Ours	72.0%(BI-RADS 0/(2-3)/(4-5))	74.1%(BI-RADS 0/(2-3)/(4-5)))
	56.34%±1.4% (BI-RADS 0/1/2/3/4/5)	57.40%±1.7% (BI-RADS 0/1/2/3/4/5)