Pattern Recognition 155 (2024) 110732

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Conditional advancement of machine learning algorithm via fuzzy neural network

Kevin Bronik [a], Le Zhang [b,*]

[a] Department of Engineering Science, University of Oxford, Oxford, OX1 3PJ, United Kingdom
[b] School of Engineering, College of Engineering and Physical Sciences, University of Birmingham, Birmingham, B15 2FG, United Kingdom

ARTICLE INFO

Keywords:
Machine learning
Convolutional neural network
Fuzzy neural network
Adaptive neuro-fuzzy inference
Segmentation
Validation metrics

ABSTRACT

Improving overall performance is the ultimate goal of any machine learning (ML) algorithm. While it is a trivial task to explore multiple individual validation measurements, evaluating and monitoring overall performance can be complicated due to the highly nonlinear nature of the functions describing the relationships among different validation metrics, such as the Dice Similarity Coefficient (DSC) and Jaccard Index (JI). Therefore, it is naturally desirable to have a reliable validation algorithm or model that can integrate all existing validation metrics into a single value. This consolidated metric would enable straightforward assessment of an ML algorithm's performance and identify areas for improvement. To deal with such a complex nonlinear problem, this study suggests a novel parameterized model named Adaptive Neuro-Fuzzy Inference Systems (ANFIS), which takes any set of input–output precise-imprecise data and uses a neuro-adaptive learning strategy to tune the parameters of the pre-defined membership functions. Our method can be accepted as an elegant and the state-of-the-art method for the nonlinear function approximation, which could be added directly to any convolutional neural networks (CNN) loss functions as the regularization term to generate a constrained-CNN-FUZZY model optimization. To demonstrate the ability of the purposed method and provide a practical explanation of the capability of ANFIS, we use deep CNN as a testing platform to consider the fact that one of the biggest challenges CNN-developers faced today is to reduce the mismatching between the provided input data and the predicted results monitored by different validation metrics. We first create a toy dataset using MNIST and investigate the properties of the proposed model. We then use a medical dataset to demonstrate our method's efficacy on brain lesion segmentation. In both datasets, our method shows reliable validation results to guide researchers towards choosing performance metrics in a problem-aware manner, especially when the results of different validation metrics are too similar among models to determine the best one.

1. Introduction

Machine Learning (ML) algorithms, particularly convolutional neural networks (CNNs), are extensively employed in image analysis tasks such as classification, segmentation, and biomedical data analysis [1]. By leveraging convolution operations, CNNs outperform other ML techniques in these applications. Supervised methods, especially CNNs, have demonstrated superior reliability compared to unsupervised methods and non-CNN-based supervised methods like k-nearest neighbors (k-NN) for tasks such as Multiple Sclerosis (MS) lesion segmentation [2,3]. Furthermore, various enhancement techniques, such as increasing training data, layer freezing, network deepening, and regularization, have been proposed to improve CNN performance [4,5].

The evaluation of ML algorithms presents numerous challenges, particularly in specific domains where relevant performance metrics are crucial [6]. Assuming a linear relationship among these metrics for estimating overall performance is overly simplistic and unrealistic [7]. In reality, the relationships among various validation metrics exhibit complex non-linear behavior. For instance, in MS lesion segmentation, algorithms are typically assessed against expert consensus labels using metrics such as Dice Similarity Coefficient (DSC) [3], Jaccard Index (JI) [2], Sensitivity (SENS) [8], and Specificity (SPEC) [9]. Relying solely on one metric can lead to misleading conclusions about algorithm performance, highlighting the importance of considering multiple metrics simultaneously. The multifaceted approach, which combines multiple metrics, can provide a more comprehensive evaluation and facilitates the identification of underlying causes of inaccuracies. By analyzing multiple metrics together (see Fig. 1), it becomes possible to devise solutions that address deficiencies comprehensively, thereby improving overall algorithm performance.

* Corresponding author.
E-mail address: l.zhang.16@bham.ac.uk (L. Zhang).

https://doi.org/10.1016/j.patcog.2024.110732
Received 3 December 2023; Received in revised form 25 June 2024; Accepted 25 June 2024
Available online 29 June 2024
0031-3203/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

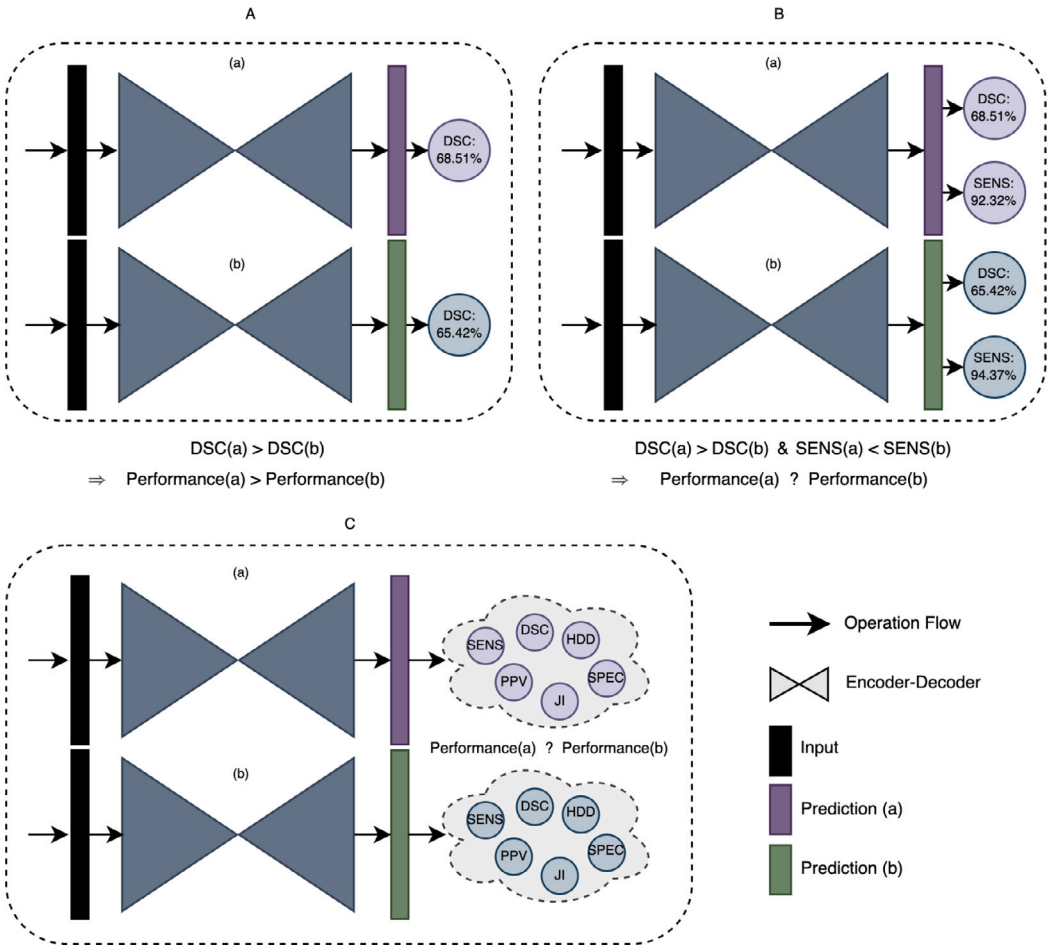

Fig. 1. Different situations for validating the performance of machine learning models. **A**: Only consider DSC; **B**: Consider both DSC and SENS; **C**: Consider as many factors as possible.

Several studies have explored the intricate relationships among various validation metrics. For instance, [10] found that validation metrics with low bias tend to correlate more strongly with manual rankings. Similarly, [11] revealed a profound connection between Receiver Operator Characteristic (ROC) space and Precision–Recall (PR) space, noting that linear interpolation between points in PR space is erroneous and that optimizing the area under the ROC curve does not necessarily optimize the area under the PR curve. This underscores the inherent non-linearity in their correlation, making linear representation challenging, if not impossible. Advanced techniques, such as the neuro-fuzzy system, offer promising avenues for approximating these complex non-linear relationships among validation metrics. The neuro-fuzzy system, a fusion of artificial neural networks and fuzzy logic, has demonstrated its efficacy in various research domains [12, 13]. Leveraging the neuro-fuzzy system, along with neural network approximation methods, provides a robust approach for identifying and modeling non-linear behaviors. Notably, the flexibility of the fuzzy system allows it to model any nonlinear functions with reasonable complexity, accommodating both precise and imprecise input–output data.

In this study, we introduce a novel nonlinear identification method employing Adaptive Neuro-Fuzzy Inference Systems (ANFIS) for evaluating the performance of existing CNN models. We categorize the optimization process in CNN algorithms into intrinsic and extrinsic procedures. Intrinsic optimization addresses overfitting and underfitting to enhance generalizability, while extrinsic optimization involves tuning hyperparameters based on ANFIS analysis of validation metrics. To assess our method, we begin by constructing a toy dataset using MNIST to explore its properties. Subsequently, we validate our approach using brain MRI datasets to demonstrate its efficacy. Our contributions are as follows:

1. We propose the first application of a fuzzy neural network for solving nonlinear optimization problems in medical image analysis, providing a comprehensive evaluation framework for ML algorithms.
2. We emphasize the versatility of fuzzy systems in modeling diverse nonlinear functions, effectively accommodating various types of input–output data.
3. Our proposed parameterized nonlinear function, determined by ANFIS, can be seamlessly integrated into CNN loss functions as a regularization term, facilitating constrained-CNN-FUZZY model optimization.
4. We validate our method across cross-domain datasets, including MNIST digits and MS lesion segmentation datasets, under various experimental setups. Additionally, we demonstrate the potential of ANFIS using synthetic data generated from these real-world datasets. The results on synthetic data intuitively underscore the robustness of our method for evaluating ML algorithms.

2. Hybrid model

We develop a hybrid model that integrates a CNN and a FNN to assess our approach. These networks operate independently during both training and testing phases. However, by exploiting the nonlinear relationships among multiple validation metric outputs, the CNN

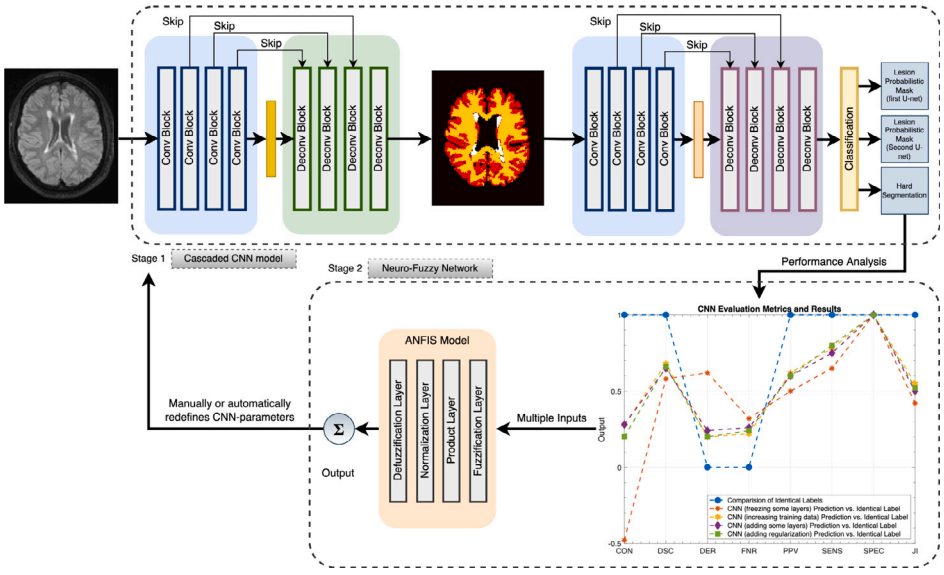

Fig. 2. The architecture of our proposed hybrid model that combines the Convolution Neural Network (CNN) and Fuzzy Neural Network (FNN). Note that the CNN could be replaced by another ML algorithm; the FNN takes the validation outputs of CNN as inputs to generate a value to show the CNN's performance from a more comprehensive approach. The output of FNN could feedback to CNN for manually or automatically fine-tuning CNN's parameters, therefore improve CNN's performance.

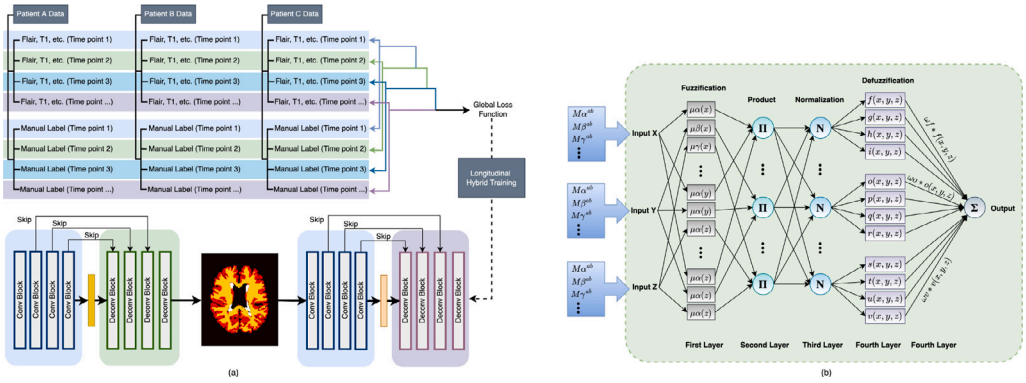

Fig. 3. Left: The architecture of adaptive deep multi-task learning framework for image segmentation; Right: The architecture of adaptive neuro-fuzzy inference system (ANFIS) for modeling high nonlinear function.

undergoes full optimization through a neuro-fuzzy system (see Fig. 2). The interplay between these networks is governed by rules internally controlling the system's behavior while aiming to achieve better agreement between manual and predicted labels. This hybrid architecture, comprising two independent artificial neural networks, can be seamlessly applied to any existing ML algorithms. The FNN plays a central role in achieving optimal performance.

2.1. Adaptive convolutional neural network

CNNs are a category of ML algorithms, representing a progression from traditional ML algorithms to neural network-based approaches. They process information in a hierarchical computational-mathematical manner, aiming to automate the extraction of complex information [14,15]. Recently, many frontier researchers have extensively utilized deep neural networks due to their effectiveness in enhancing information processing accuracy and speed. CNNs address computational model complexity by significantly reducing parameter dimensions through the use of convolutional layers, achieving translation invariance and equivalence properties [16,17]. These properties are typically obtained through data augmentation techniques, which apply various transformations (e.g., rotation, scaling, flipping, translating) to input data. This enhances CNN flexibility, particularly when training data are limited, leading to improved robustness, performance, and consistent results [18,19]. CNNs provide a versatile framework for various research purposes, including video and image analysis, human visual systems, robotics, etc. In our work, we focus on one specific application of CNNs in medical image analysis, particularly in solving segmentation problems [20,21].

To analyze and evaluate our proposed evaluation method, we adapted an advanced U-net model for segmentation tasks in medical images. One such CNN implementation used in this study is publicly available[1] to facilitate reproducibility and enable other researchers to study and test our proposed method easily. It is important to note that any publicly available CNN models, whether related or unrelated to medical engineering, could be utilized for analysis and testing purposes. In this study, the CNN model employed is an adaptive deep multi-task learning framework for image segmentation, supporting longitudinal training and testing strategies (see Fig. 3a). Specifically designed for MS lesion segmentation, the network utilizes a cascaded model of 3D patch-wise U-net. Both U-nets share the same CNN architecture and employ a concatenated objective function based on categorical cross-entropy, Jaccard loss, Symmetric Hausdorff loss, Precision loss, etc., along with

[1] Publicly available at: https://github.com/kbronik2017/.

stochastic optimization techniques such as the Adadelta optimizer during training. The first U-net is trained to detect possible lesion voxels, while the second U-net, based on a threshold (e.g., $l_{threshold} = 0.5$), aims to reduce incorrect predictions made by the first U-net. The segmentation model takes an input image (e.g., MRI image) and outputs two lesion probabilistic masks corresponding to the two cascaded U-net models, along with a final binary segmentation.

2.2. Fuzzy neural network

A fuzzy inference system, whether Mamdani-type [22] or Sugeno-type [23], is a logical and human-understandable framework used to determine latent variables based on observations (universe of discourse) within the space of fuzzy rules and membership functions ($\mu_A, \mu_B, \ldots$). Logical operations in fuzzy inference systems follow specifically defined procedures based on so-called T and S norm operators: $T, S := \begin{cases} \mu_{A\cap B} = T(\mu_A, \mu_B) \\ \mu_{A\cup B} = S(\mu_A, \mu_B) \end{cases}$ where $A = \{x, \mu_A(x) \mid x \in X\}$ and $B = \{x, \mu_B(x) \mid x \in X\}$; x is an element of the set of inputs X.

The determination of membership functions is crucial for designing an effective fuzzy inference system to model highly nonlinear functions. In this work, we employ an Adaptive Neuro-Fuzzy Inference System (ANFIS), originally introduced by [24], for evaluating the problem. ANFIS combines fuzzy logic with neural networks, utilizing a neuro-adaptive learning approach similar to neural networks. The parameters of the membership functions in ANFIS are tuned using gradient descent optimization methods to predict optimal performance and minimize prediction errors. The ANFIS network architecture (Fig. 3b) adopted in our work consists of the following layers with associated functionality:

Layer 1: In this layer, an specific membership function $\mu_\alpha(x)$ is associated with any node in this layer whose parameters $(a, b \mid \sigma, c \mid d)$[also called premise parameters] will be tuned during learning process. This can be chosen to be a bell-shaped membership function: $\mu_\alpha(x, a, b, c) = \frac{1}{1+|\frac{x-c}{a}|^{2b}}$, or Gaussian function: $\mu_\alpha(x, \sigma, c) = e^{\frac{-(x-c)^2}{2\sigma^2}}$, or any other membership function such as triangular, trapezoidal etc.

Layer 2: The nodes in this layer apply product $t-norm$:

$T_{prod}(\mu_\alpha(x), \mu_\beta(x)) = \mu_\alpha(x) * \mu_\beta(x),$

to the incoming signals, where the outgoing signals represent rules' firing strengths.

Layer 3: In this layer, the normalization of fuzzy rules' firing strengths will be processed by calculating the ratio of each one rules' firing strength to the sum of all rules' firing strengths.

Layer 4: The nodes in this layers take the incoming normalized rules' firing strengths signals and apply them to the parameterized functions:

$\omega_f * f(x, y, z, ..., p, q, r, ...) = \omega_f * (px + qy + rz + ...),$

where the parameters (also called consequent parameters) will be determined through training process.

Layer 5: This one node layer sums all incoming signals: $\sum_{\forall f} \omega_f * f(x, y, z, ..., p, q, r, ...)$

3. Model implementation and optimization with hybrid structure

To train and optimize the hybrid neural network, we first define the fuzzy inference system using either a Mamdani-type or Sugeno-type fuzzy model. This necessitates prior definition of input and output variables for the inference system. The input variables are the performance validation metric results: M_i (result of comparison between the predicted mask with the expert consensus) and M_i^T (result of comparison of two identical expert consensuses). The output variables can be defined either based on human expertise or a mathematical function that maps input values to a closed interval between 0 and 1 ($l : R^n \rightarrow [0, 1]$, where n is the number of validation metrics in this study). As depicted in Fig. 3b, there is no specific constraint on the symmetry of output values in terms of inputs (it is unnecessary to consider symmetry about the y-axis $x = 1$). Due to changes in parameters β and γ ($\exists \beta, \gamma \mid \beta \neq \gamma$), the following equation is used to generate the ground truth labels (output variables) for ANFIS:

$$label(y; x, \beta, \gamma, \lambda) := \begin{cases} \| \frac{\lambda}{1+e^{\beta(-x+1)}} \|_{L1} & x \leqq 1 \\ \| \frac{\lambda}{1+e^{\gamma(x-1)}} \|_{L1} & x > 1, \end{cases} \quad (1)$$

where $\lambda \in \mathbb{R}$ is a hyperparameter and input x is defined as:

$$x(M_i; W_i, M_i^T) := \frac{1}{n} \sum_{i=1}^{n} \frac{W_i(M_i - M_i^T)}{M_i^T + \varepsilon} + 1, \quad (2)$$

where $\varepsilon := \begin{cases} 1 & M_i^T = 0 \\ 0 & M_i^T \neq 0 \end{cases}$ and the parameters β, γ and the weights $W_{i\in\{1,n\}}$ are determined indirectly through the ANFIS if the labeling is based on human expert, otherwise the parameters are set manually prior to analysis if the expert knowledge is available. Note that, the function at the point $(1, \lambda)$ (see Fig. 4a) is continuous but not differentiable. Taking into the account that the availability of ground truth for training ANFIS, two adaptive neuro-fuzzy inference systems (see Fig. 4b&c) are defined for the independent variables based on the left and right hand sides of the jump discontinuity (point) to model the nonlinear relations of the multiple performance validation metric results. In this work, our defined parameterized function could be added directly to CNN's loss function to generate a constrained-CNN-FUZZY model optimization.

4. Experiments

In the hybrid model, the CNN model is trained through five different experiment settings, including increasing training data (CNN^{+d}), freezing some layers (CNN^{-l}), adding some layers (CNN^{+l}), adding regularization (CNN^{+r}) and early stop (CNN^{-e}), on each dataset to analyze the conditional improvement procedure of the network (see Table 1).

4.1. Datasets

We first define a toy segmentation dataset based on MNIST and study the properties of the proposed method. We then demonstrate the utility of the method on the public medical image dataset. **MNIST dataset:** MNIST was originally constructed to facilitate research in image classification, in the form of recognizing handwritten digits [25], it has found its use in segmentation task, which segment the digital from the background. It can be seen as an image classification task, except that instead of classifying the whole image, we are classifying each pixel individually. MNIST dataset consists of 60,000 training and 10,000 testing examples, all of which are 28×28 gray-scale images of digits from 0 to 9, and we derive the segmentation labels by thresholding the intensity values at 0.5. **ISBI2015 MS Lesion Segmentation Dataset:** The ISBI2015 MS lesion challenge [26] was composed of 5 training and 14 testing subjects with 4 or 5 different image time-points per subject. All of the data were acquired on a 3.0 T MRI scanner (Philips Medical Systems, Best, The Netherlands) with T1-w MPRAGE, FLAIR, T2-w sequences. On the challenge competition, each subject image was evaluated independently, which led to a final training and testing sets composed of 21 and 61 images, respectively. Manual delineations of MS lesions performed by two experts were included

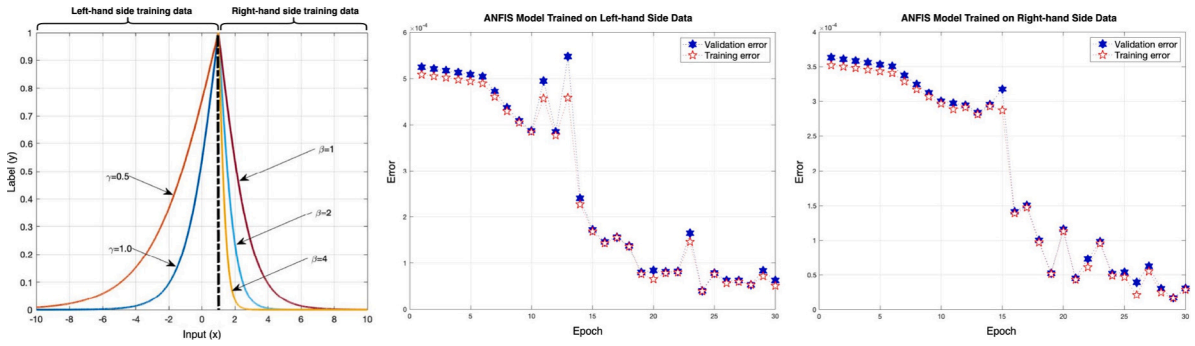

Fig. 4. Left: 2D illustration of $label(y; x, \beta, \gamma, \lambda) := \begin{cases} \| \frac{\lambda}{1+e^{\beta(-x+1)}} \|_{L1} & x \leqq 1 \\ \| \frac{\lambda}{1+e^{\gamma(x-1)}} \|_{L1} & x > 1, \end{cases}$ for different values of β *and* γ and $\lambda = 2$; Middle and Right: Training error and validation error of two ANFIS models ($\beta = 0.6$ *and* $\gamma = 0.7$). For reasons of brevity, only three inputs and one output are shown in the model.

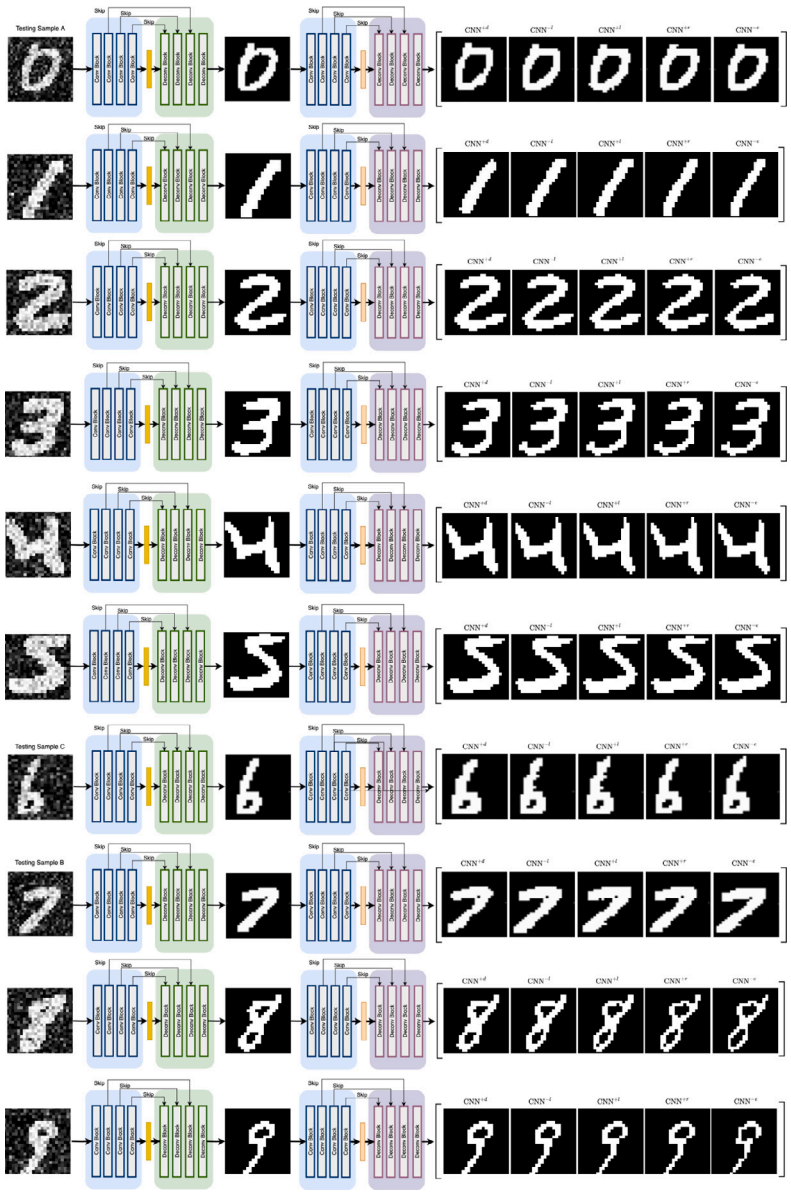

Fig. 5. A graphical display of CNN with five different prediction outputs from different experiment settings on MNIST segmentation dataset.

for each of the 21 training images. In our work, we extract the 2D slices from the 21 training images and use the slices from 3 subjects for training and the slices from another 2 subjects for testing, so the training data and the testing data are not overlapped on same patient. We also applied data augmentation to increase the size of training and testing data. We perform data augmentation on-the-fly at batch time

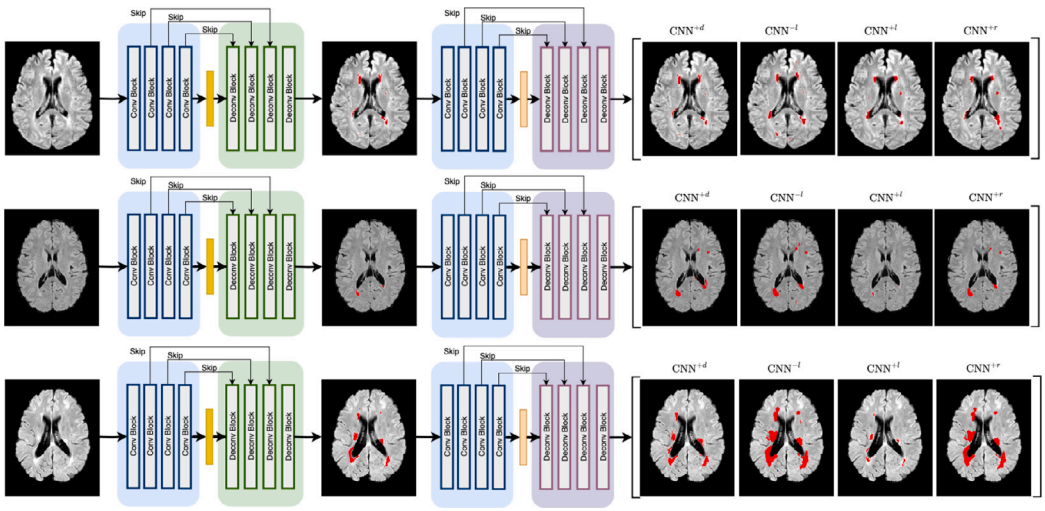

Fig. 6. A graphical display of CNN with four different prediction outputs from different experiment settings on ISBI2015 MS lesion segmentation dataset.

Table 1
Datasets details of MNIST and ISBI2015 MS Lesion Segmentation Challenge in our experiments.

Dataset	Modality	Resolution	Noise Add	Training set	Testing set
MNIST	Single (Grey scale)	28 × 28	Gaussian noise	60,000	10,000
Brain MRI	Multiple (T1, Flair and T2)	0.82 × 0.82 × 2.2 mm^3	N/A	21	61

by multiplying the number of training samples by four following the next transformations [27]: for each mini-batch, all patches are first rotated with 180 degrees in the axial plane. From the original and rotated versions of the patches, new versions are computed by flipping those horizontally. Other rotations than 180 degrees are avoided, in order to roughly maintain the symmetry of the brain and avoid artificial rotations of brain structures (see Table 1).

4.2. Quantitative metrics

Seven complementary metrics are used to evaluate the segmentation results, which are also given as the input of the FNN to generate the ANFIS outcomes. By using these metrics, we are able to thoroughly evaluate the performance of the ML algorithm in segmenting images and provide a detailed and comprehensive assessment of the results.

- Dice score coefficient ($DSC = \frac{2|\mu_\alpha \cap \mu_\beta|}{|\mu_\alpha|+|\mu_\beta|}$), a region-based metric, is used to evaluate the region overlap.
- Sensitivity or true positive rate ($SENS = \frac{TP}{TP+FN}$) represents the ability of a model to correctly identify positive instances among all actual positive instances in the segmentation.
- Precision, also known as Positive Predictive Value ($PPV = \frac{TP}{TP+FP}$) is defined as the ratio of true positive predictions to the total number of positive predictions (true positives and false positives).
- Specificity, also known as the true negative rate ($SPEC = \frac{TN}{TN+FP}$) represents the ability of a model to correctly identify negative instances among all actual negative instances in the segmentation.
- Jaccard index ($JI = \frac{|\mu_\alpha \cap \mu_\beta|}{|\mu_\alpha \cup \mu_\beta|}$), also known as the Jaccard similarity coefficient, is a metric used to measure the similarity between two segmentations. It is defined as the size of the intersection of two segmentations divided by the size of their union.
- Hausdorff distance is a mathematical concept used to measure the dissimilarity between two sets of points. In the image segmentation task, Hausdorff distance ($HDD = \max_{a\in A}\left(\min_{b\in B} d(a,b)\right)$) is employed to compare the similarity between two masks.
- Conformity ($CON = (1 - \frac{FP+FN}{TP}) * 100\%$) create a metric that penalizes the model for false predictions (both false positives and false negatives) relative to the true positives.

For all quantitative metrics, we apply the five-folder cross-validation approach to evaluate our method.

4.3. Performance on practical image segmentation task

In Fig. 5 and Fig. 6, we present the CNN testing procedure on cross-domain images. For MNIST dataset, we randomly select three different digital numbers (0, 7 and 9) and show their performance on binary segmentation from the five training procedures. For brain MRI dataset, we present the testing results of one image sample with four from the five training procedures. Meanwhile, the performance analysis graph, shown in Fig. 7, plots the individual metrics results for the brain MRI sample. One can easily verify that neither the binary segmentation masks nor the plotted graphs could lead to an accurate analytical determination of the CNN improvements. In other words, it is difficult to explore whether or not there is an improvement of a CNN by using the five experiment settings. Meanwhile, we present the corresponding quantity results using validation metrics, such as DSC, SENS, PPV, etc, for the two datasets in Tables 2 and 3. Explicitly, in terms of the last column of Table 2 and last two columns of Table 3, which explore the improvement of the CNN using ANFIS, we have a very different result compared to the multiple evaluation metrics, where we only present binary segmentation masks, plotted graph or individual metric result independently. Thus, we can know that ANFIS is a multi-dimensional real-valued high nonlinear function and it computes the closeness to truth or "degrees of truth" for the multiple validation metric results. The ANFIS approach can be used in more sophisticated ways to determine and approve the improvement of CNN, where the perfect case or the maximum available degree of truth is the numerical value 1. Note that the last column of Table 3 shows that how sensible can be the results by weighting the metrics (e.g., DSC), where the quality of ANFIS results can be influenced by the criteria of human expert definition.

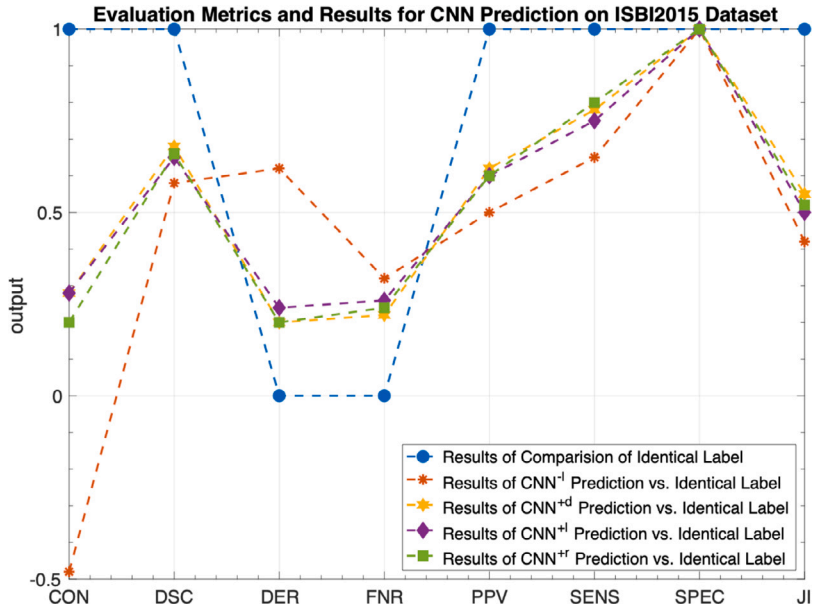

Fig. 7. A graphical display of performance analysis results on ISBI2015 dataset for MS lesion segmentation.

Table 2
The evaluation results of different metrics that have been used in the experiment. Numbers in bold indicate the best method in each table that statistically ($p < 0.01$) better than other methods by computing the p values of paired t-tests on different metrics.

Testing sample A	DSC	SENS	PPV	SPEC	JI	HDD	ANFIS
CNN^{+d}	0.9661	99.1820	94.1747	98.3896	0.9344	0.0572	**0.9927**
CNN^{-l}	0.9224	88.7525	96.0176	99.0338	0.8560	0.1215	**0.9847**
CNN^{+l}	0.9572	96.1145	95.3346	98.7654	0.9179	0.0588	**0.9904**
CNN^{+r}	0.9421	100.0000	89.0710	96.7793	0.8907	0.1080	**0.9883**
CNN^{-e}	0.9720	99.5910	94.9317	98.6044	0.9456	0.0493	**0.9940**
Ground Truth	1.0000	100.0000	100.0000	100.0000	1.0000	0.0000	1.0000
Testing sample B	DSC	SENS	PPV	SPEC	JI	HDD	ANFIS
CNN^{+d}	0.9792	99.5780	96.3265	99.0415	0.9593	0.0409	**0.9958**
CNN^{-l}	0.9220	86.0759	99.2700	99.8402	0.8553	0.1507	**0.9871**
CNN^{+l}	0.9789	97.8902	97.8902	99.4675	0.9586	0.0266	**0.9951**
CNN^{+r}	0.9624	100.0000	92.7592	98.0298	0.9275	0.0770	**0.9927**
CNN^{-e}	0.9853	99.3670	97.7178	99.4142	0.9711	0.0262	0.9969
Ground Truth	1.0000	100.0000	100.0000	100.0000	1.0000	0.0000	1.0000
Testing sample C	DSC	SENS	PPV	SPEC	JI	HDD	ANFIS
CNN^{+d}	0.9636	100.0000	92.9889	99.0952	0.9298	0.0641	**0.9929**
CNN^{-l}	0.9633	99.2063	93.6329	99.1904	0.9293	0.0584	**0.9925**
CNN^{+l}	0.9598	99.6031	92.6199	99.0476	0.9227	0.0682	**0.9920**
CNN^{+r}	0.9298	100.0000	86.8965	98.1904	0.8689	0.1285	**0.9870**
CNN^{-e}	0.9613	98.8095	93.6090	99.1904	0.9256	0.0648	**0.9923**
Ground Truth	1.0000	100.0000	100.0000	100.0000	1.0000	0.0000	1.0000

4.4. Performance on synthetic approaches

In order to determine whether the proposed method meets the defined specification, an appropriate image processing technique based on syntactic method [28] is investigated in the this study. Given X_{input} : $[A(12\times12), B(6\times6), C(3\times3)]$, where A, B, C are three different 2D slices with the image size in corresponding bracket. The following various types of transformations $f_t : X_{input} \rightarrow X'_{output}$ namely: rotation R, translation T and scaling up or down S, are applied to the input images that are the ground truth labels from the brain MRI dataset:

- $f_1 : X_{input} \rightarrow Y_{output} = f_1(X_{input}) := \{R[15, 15, 0; A(12 \times 12)], B(6 \times 6), C(3 \times 3)\}$
- $f_2 : X_{input} \rightarrow Y_{output} = f_2(X_{input}) := \{R[25, 25, 0; A(12 \times 12)], B(6 \times 6), C(3 \times 3)\}$
- $f_3 : X_{input} \rightarrow Y_{output} = f_3(X_{input}) := \{R[25, 25, 0; A(12 \times 12)], R[25, 25, 5; B(6 \times 6)], C(3 \times 3)\}$
- $f_4 : X_{input} \rightarrow Y_{output} = f_4(X_{input}) := \{R[25, 25, 0; A(12 \times 12)], R[25, 25, 5; B(6 \times 6)], R[25, 25, 35; C(3 \times 3)]\}$
- $f_5 : X_{input} \rightarrow Y_{output} = f_5(X_{input}) := \{T[-4, -3, 0; R[25, 25, 0; A(12 \times 12)]], R[25, 25, 5; B(6 \times 6)], R[25, 25, 35; C(3 \times 3)]\}$
- $f_6 : X_{input} \rightarrow Y_{output} = f_6(X_{input}) := \{T[-4, -3, 0; R[25, 25, 0; A(12 \times 12)]], T[-3, 3, 0; R[25, 25, 5; B(6 \times 6)]], R[25, 25, 35; C(3 \times 3)]\}$
- $f_7 : X_{input} \rightarrow Y_{output} = f_7(X_{input}) := \{T[-4, -3, 0; R[25, 25, 0; A(12 \times 12)]], T[-3, 3, 0; R[25, 25, 5; B(6 \times 6)]], T[1, -4, 0; R[25, 25, 35; C(3 \times 3)]]\}$
- $f_8 : X_{input} \rightarrow Y_{output} = f_8(X_{input}) := \{S[104\%, 25\%, 0; T[-4, -3, 0; R[25, 25, 0; A(12 \times 12)]]], T[-3, 3, 0; R[25, 25, 5; B(6 \times 6)]], T[1, -4, 0; R[25, 25, 35; C(3 \times 3)]]\}$
- $f_9 : X_{input} \rightarrow Y_{output} = f_9(X_{input}) := \{S[104\%, 25\%, 0; T[-4, -3, 0; R[25, 25, 0; A(12 \times 12)]]], S[50\%, 117\%, 0; T[-3, 3, 0; R[25, 25, 5; B(6 \times 6)]]], T[1, -4, 0; R[25, 25, 35; C(3 \times 3)]]\}$

Table 3
The evaluation results of different metrics that have been used in the experiment. Numbers in bold indicate the best method in each table that statistically ($p < 0.01$) better than other methods by computing the p values of paired t-tests on different metrics.

Model / Metric	DSC	FNR	DER	SENS	SPEC
CNN^{-l}	0.5794	0.3122	0.6330	68.7701	99.9430
CNN^{+d}	0.7070	0.1653	0.1015	83.4687	99.9553
CNN^{+l}	0.6670	0.1929	0.1212	80.7046	99.9522
CNN^{+r}	0.6855	0.1395	0.0863	86.0495	99.9500
Ground Truth	1.0000	0.0000	0.0000	100.0000	100.0000

PPV	JI	CON	ANFIS /weights=1	ANFIS/weighting Dice Score
50.0617	0.4078	−45.1652	0.5764	**0.5916**
61.3229	0.5468	17.1235	0.5935	**0.6042**
56.8465	0.5004	0.1790	0.5832	**0.5953**
56.9660	0.5215	8.2445	0.5864	**0.5978**
100.0000	1.0000	100.0000	1.0000	1.0000

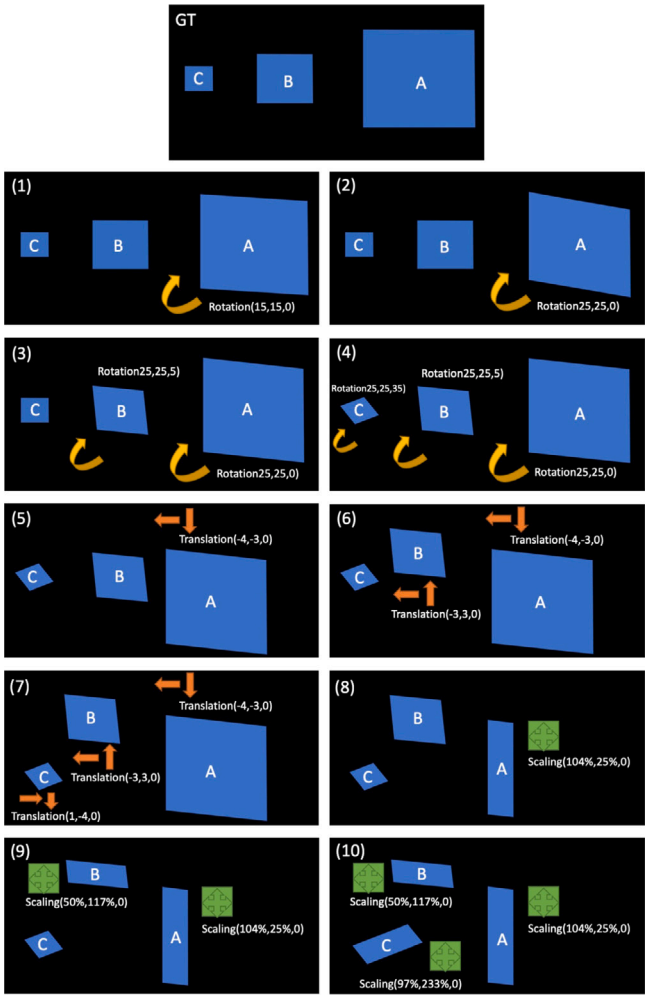

Fig. 8. Applying the transformation ($f_1 \rightarrow f_{10}$) to the Ground Truth image (GT).

- $f_{10} : X_{input} \rightarrow Y_{output} = f_{10}(X_{input}) := \{S[104\%, 25\%, 0; T[-4, -3, 0; R[25, 25, 0; A(12 \times 12)]]], \; S[50\%, 117\%, 0; T[-3, 3, 0; R[25, 25, 5; B(6 \times 6)]]], S[97\%, 233\%, 0; \; T[1, -4, 0; R[25, 25, 35; C(3 \times 3)]]]\}$

We create a 50 synthetic 3D image data to evaluate our pre-trained ANFIS model. Each 3D image data includes five 2D slices. As shown in Fig. 8, we apply the 10 transformation approaches to the five 2D slices in each 3D image. The reason that we generate such 50 image data is to simulate 10 different models with different segmentation abilities and use them as the segmentation results to evaluate ANIFS. It could tell us the truth that which model has the best segmentation performance. From f_1 to f_{10}, transformation will be more and more complex compared to its previous one f_{i-1}. Thus, the 10 models should be with increased segmentation error and consequently decreased in segmentation abilities. Note that the algorithm works under the assumption that at each step only one of the above described transformation is being applied and proceeded on only one slice (A or B or C) while the other objects have been kept unchanged compared to previous step (see the algorithms $f_1 \rightarrow f_{10}$).

The quantity results using six validation metrics for the synthetic data with the sequential transformation are given in Table 4, ANFIS results in the last column are calculated using the first six metric results. In addition, we provided the ANFIS results that are calculated with different input combinations in Tables 5–8. The ANFIS results, which is the application performance of neural calculus algorithm to the nonlinear approximation problem, reflect the decreasing performance

Table 4
Validation Metrics that have been used to evaluate difference between our synthetic labels and the input ground truth labels. ANFIS takes all the six metric results as inputs. Numbers in bold indicate the best method in each table that statistically ($p < 0.01$) better than other methods by computing the p values of paired t-tests on different metrics.

	DSC	FDR	FNR	SPE	PPV	JI	ANFIS
GT	1.0000	0.0000	0.0000	100.0000	100.0000	1.0000	1.0000
f_1	0.9718	0.0040	0.0510	82.8792	57.4271	0.9452	**0.9625**
f_2	0.9207	0.0087	0.1404	84.6239	58.1442	0.8531	**0.9543**
f_3	0.9234	0.0115	0.1335	80.7036	51.5539	0.8577	**0.9444**
f_4	0.9147	0.0160	0.1453	79.9633	50.0396	0.8428	**0.9362**
f_5	0.5707	0.3857	0.4670	79.8985	38.3515	0.3993	**0.4297**
f_6	0.4481	0.5177	0.5815	79.8282	32.7195	0.2887	**0.3066**
f_7	0.4086	0.5602	0.6183	79.7978	30.6822	0.2567	**0.2736**
f_8	0.0644	0.8806	0.9558	90.2391	10.6964	0.0332	**0.1109**
f_9	0.0021	0.9952	0.9985	91.8056	0.4688	0.0011	**0.0786**
f_{10}	0.0021	0.9959	0.9985	90.5212	0.3999	0.0010	**0.0784**

Table 5
Metrics that have been used to calculate ANFIS: FDR, FNR, SPE and PPV. Numbers in bold indicate the best method in each table that statistically ($p < 0.01$) better than other methods by computing the p values of paired t-tests on different metrics.

	FDR	FNR	SPE	PPV	ANFIS
GT	0.0000	0.0000	100.0000	100.0000	1.0000
f_1	0.0040	0.0510	82.8792	57.4271	**0.9434**
f_2	0.0087	0.1404	84.6239	58.1442	**0.9432**
f_3	0.0115	0.1335	80.7036	51.5539	**0.9273**
f_4	0.0160	0.1453	79.9633	50.0396	**0.9166**
f_5	0.3857	0.4670	79.8985	38.3515	**0.2962**
f_6	0.5177	0.5815	79.8282	32.7195	**0.1796**
f_7	0.5602	0.6183	79.7978	30.6822	**0.1518**
f_8	0.8806	0.9558	90.2391	10.6964	**0.0413**
f_9	0.9952	0.9985	91.8056	0.4688	**0.0250**
f_{10}	0.9959	0.9985	90.5212	0.3999	**0.0249**

Table 6
Metrics that have been used to calculate ANFIS: DSC,SPE, PPV and JI. Numbers in bold indicate the best method in each table that statistically ($p < 0.01$) better than other methods by computing the p values of paired t-tests on different metrics.

	DSC	SPE	PPV	JI	ANFIS
GT	1.0000	100.0000	100.0000	1.0000	1.0000
f_1	0.9718	82.8792	57.4271	0.9452	**0.8507**
f_2	0.9207	84.6239	58.1442	0.8531	**0.7033**
f_3	0.9234	80.7036	51.5539	0.8577	**0.6878**
f_4	0.9147	79.9633	50.0396	0.8428	**0.6693**
f_5	0.5707	79.8985	38.3515	0.3993	**0.1789**
f_6	0.4481	79.8282	32.7195	0.2887	**0.1214**
f_7	0.4086	79.7978	30.6822	0.2567	**0.1083**
f_8	0.0644	90.2391	10.6964	0.0332	**0.0473**
f_9	0.0021	91.8056	0.4688	0.0011	**0.0413**
f_{10}	0.0021	90.5212	0.3999	0.0010	**0.0412**

Table 7
Metrics that have been used to calculate ANFIS: DSC, FDR and SPE. Numbers in bold indicate the best method in each table that statistically ($p < 0.01$) better than other methods by computing the p values of paired t-tests on different metrics.

	DSC	FDR	SPE	ANFIS
GT	1.0000	0.0000	100.0000	1.0000
f_1	0.9718	0.0040	82.8792	**0.8524**
f_2	0.9207	0.0087	84.6239	**0.7764**
f_3	0.9234	0.0115	80.7036	**0.7596**
f_4	0.9147	0.0160	79.9633	**0.7418**
f_5	0.5707	0.3857	79.8985	**0.3231**
f_6	0.4481	0.5177	79.8282	**0.2289**
f_7	0.4086	0.5602	79.7978	**0.2040**
f_8	0.0644	0.8806	90.2391	**0.0796**
f_9	0.0021	0.9952	91.8056	**0.0674**
f_{10}	0.0021	0.9959	90.5212	**0.0665**

Table 8
Metrics that have been used to calculate ANFIS: FNR, PPV and JI. Numbers in bold indicate the best method in each table that statistically ($p < 0.01$) better than other methods by computing the p values of paired t-tests on different metrics.

	FNR	PPV	JI	ANFIS
GT	0.0000	100.0000	1.0000	1.0000
f_1	0.0510	57.4271	0.9452	**0.7085**
f_2	0.1404	58.1442	0.8531	**0.6396**
f_3	0.1335	51.5539	0.8577	**0.6006**
f_4	0.1453	50.0396	0.8428	**0.5856**
f_5	0.4670	38.3515	0.3993	**0.2595**
f_6	0.5815	32.7195	0.2887	**0.2043**
f_7	0.6183	30.6822	0.2567	**0.1900**
f_8	0.9558	10.6964	0.0332	**0.1093**
f_9	0.9985	0.4688	0.0011	**0.0925**
f_{10}	0.9985	0.3999	0.0010	**0.0924**

of the segmentation representation that is coincided with the visual representation of the decreasing segmentation process from f_1 to f_{10}. This shows our proposed method returns a value that can indicates whether overall performance of a ML algorithm is progressively improved or it becomes progressively worse compared to different experiment settings or different algorithms. Note that, we are using synthetic data in this experiments to validate our hybrid model. In the next step, we will recruit multiple annotators (with experience from beginner to expert) to segment a sequence of real-world images where each image is labeled by different annotators. All images are sorted from the best segmentation to worst one. Then we could apply pre-trained ANFIS model to the sorted the segmentation performance and get the ANFIS value for each image. We can boldly assume that all the images sorted by ANIFS values from highest to lowest stay in agreement with the annotator abilities sorted images from best to worst.

5. Discussion and conclusion

This study presents the development of an adaptive neuro-fuzzy inference system, offering an automated method for comprehensive monitoring of overall ML algorithm performance. We applied our proposed approach to compare different image segmentation results on the MNIST digits dataset and brain MRI lesion segmentation dataset using various optimization methods for CNN. The results demonstrate that such a system can effectively assess the quality of segmentation results obtained by multiple performance analysis metrics using both predicted and manual segmentation masks. To date, there is no existing scientific knowledge or research study, nor a generally accepted paper, that provides a reliable validation algorithm or model capable of combining all existing validation metrics to straightforwardly determine ML algorithm performance. However, we compared our method with almost all existing ML evaluation metrics, including DSC, SENS, PPV, SPEC, JI, HDD, and CON. By utilizing these metrics, we thoroughly evaluated the performance of the ML algorithm in image segmentation and provided a detailed and comprehensive assessment of the results.

To validate the concept of our method, we employed a validation technique to assess whether the proposed method meets the defined specifications. A state-of-the-art image processing technique based on syntactic methods was utilized, representing an appropriate approach for this purpose. Modifications to our proposed method aimed to enhance efficiency, including the development of an adaptive neuro-fuzzy inference system, enabling automated monitoring of overall ML algorithm performance from a comprehensive perspective. This approach facilitates comparison of different image segmentation results and determination of their quality using multiple performance analysis metrics with predicted and manual segmentation masks. Tables 4–8 present the metrics used to calculate the ANFIS results. In our experiments, these metrics include DSC, SPE, PPV, JI, FNR, and others, used to evaluate segmentation algorithm performance by comparing synthetic segmentation outputs (f_1 to f_{10}) with ground truth labels. The ANFIS results in the last column of the tables are calculated using different input combinations and reflect segmentation representation performance. Values in the ANFIS column indicate whether overall segmentation algorithm performance improves or worsens with different sequential transformations. The ANFIS results are independent and unaffected by the number or combinations of metrics used in the analysis. These tables offer a comprehensive overview of evaluation metrics and ANFIS results, providing valuable insights into algorithm performance. Additionally, our method suggests choosing the most relevant validation metrics that reflect and correlate with clinical importance, such as involving highly experienced radiologists for tumor/lesion labeling.

Our experiments demonstrate that a neuro-fuzzy network can effectively reduce the complexity of comparing individual metric results with each other and solve the approximation problem for high-dimensional, real-valued, highly nonlinear functions defining complex relations between all applicable validation metrics. Our proposed method is fully adaptable to a variety of validation metrics and is suitable for evaluating any segmentation models. Furthermore, our method can suggest selecting the most relevant validation metrics that reflect and correlate with clinical importance, such as involving highly experienced radiologists for lesion labeling. In our brain MRI segmentation experiments, we utilized 200 sets of metric values to train ANFIS, with a computation time of approximately 10 min on an NVIDIA RTX 2080 Ti GPU. Parallel profiling or utilizing Multi-core and Multiprocessor Nodes can significantly improve performance and efficiency, showing superior scalability compared to non-parallel runs (achieving around a 50% increase in time efficiency).

In addition to the discussed applications of our developed hybrid model, this study suggests that the proposed parameterized nonlinear function, critically determined by ANFIS, can be directly incorporated into CNN loss functions as a regularization term to optimize constrained-CNN-FUZZY models. However, the accuracy of the ANFIS model is influenced by several factors, including the selection of input, the number of iterations, the number and type of membership functions. Apart from these factors, the results of the proposed technique may directly face challenges. One such challenge is dealing with noisy data, where the collected dataset is contaminated by unknown noise. In such cases, the training data may fail to capture all pertinent features necessary for effective modeling within the ANFIS framework. Another potential issue is overfitting, particularly concerning the generalization of the trained ANFIS across diverse data modalities, such as data sourced from various centers. This concern arises from the fixed structure of the ANFIS model, which includes a considerable number of parameters. Consequently, there is a risk of the model excessively fitting the training data, especially when subjected to numerous training epochs.

CRediT authorship contribution statement

Kevin Bronik: Writing – Original draft, Validation, Software, Methodology, Formal analysis. **Le Zhang:** Writing – Methodology, Validation, Formal analysis, Review & Editing, Supervision.

Declaration of competing interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Data availability

Data will be made available on request.

References

[1] S. Asgari Taghanaki, K. Abhishek, J.P. Cohen, J. Cohen-Adad, G. Hamarneh, Deep semantic segmentation of natural and medical images: A review, Artif. Intell. Rev. 54 (1) (2021) 137–178.

[2] R.A. Kamraoui, V.T. Ta, T. Tourdias, B. Mansencal, J.V. Manjon, P. Coupé, DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation, Med. Image Anal. 76 (2022) 102312.

[3] S. Valverde, M. Salem, M. Cabezas, D. Pareto, J.C. Vilanova, L. Ramió-Torrentà, À. Rovira, J. Salvi, A. Oliver, X. Lladó, One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks, NeuroImage: Clin. 21 (2019) 101638.

[4] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, H. Jégou, Going deeper with image transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 32–42.

[5] M. Yazdanpanah, A.A. Rahman, M. Chaudhary, C. Desrosiers, M. Havaei, E. Belilovsky, S.E. Kahou, Revisiting learnable affines for batch norm in few-shot transfer learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9109–9118.

[6] B. Cheng, R. Girshick, P. Dollár, A.C. Berg, A. Kirillov, Boundary IoU: Improving object-centric image segmentation evaluation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15334–15342.

[7] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (1) (2020) 1–13.

[8] L. Zhang, R. Tanno, M.-C. Xu, C. Jin, J. Jacob, O. Cicarrelli, F. Barkhof, D. Alexander, Disentangling human error from ground truth in segmentation of medical images, Adv. Neural Inf. Process. Syst. 33 (2020) 15750–15762.

[9] L. Zhang, R. Tanno, K. Bronik, C. Jin, P. Nachev, F. Barkhof, O. Ciccarelli, D.C. Alexander, Learning to segment when experts disagree, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 179–190.

[10] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, M.B. Blaschko, Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index, IEEE Trans. Med. Imaging 39 (11) (2020) 3679–3690.

[11] A.R. Rachakonda, A. Bhatnagar, Aratio: Extending area under the ROC curve for probabilistic labels, Pattern Recognit. Lett. 150 (2021) 265–271.

[12] D. Nauck, R. Kruse, Neuro-fuzzy systems for function approximation, Fuzzy Sets and Systems 101 (2) (1999) 261–271.

[13] L.J. Herrera, H. Pomares, I. Rojas, J. González, O. Valenzuela, Function approximation through fuzzy systems using taylor series expansion-based rules: interpretability and parameter tuning, in: Mexican International Conference on Artificial Intelligence, Springer, 2004, pp. 508–516.

[14] M.A. Khan, T. Akram, Y.D. Zhang, M. Sharif, Attributes based skin lesion detection and recognition: A mask RCNN and transfer learning-based deep learning framework, Pattern Recognit. Lett. 143 (2021) 58–66.

[15] H. Yu, L.T. Yang, Q. Zhang, D. Armstrong, M.J. Deen, Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives, Neurocomputing 444 (2021) 92–110.

[16] O.S. Kayhan, J.C.v. Gemert, On translation invariance in cnns: Convolutional layers can exploit absolute spatial location, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14274–14285.

[17] V. Biscione, J.S. Bowers, Convolutional neural networks are not invariant to translation, but they can learn to be, J. Mach. Learn. Res. 22 (229) (2021) 1–28.

[18] M. Xu, S. Yoon, A. Fuentes, D.S. Park, A comprehensive survey of image augmentation techniques for deep learning, Pattern Recognit. 137 (2023) 109347.

[19] H. Chen, Z. Qin, Y. Ding, L. Tian, Z. Qin, Brain tumor segmentation with deep convolutional symmetric neural network, Neurocomputing 392 (2020) 305–313.

[20] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, X. Chen, CPFNet: Context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imaging 39 (10) (2020) 3008–3018.

[21] N. Wang, S. Lin, X. Li, K. Li, Y. Shen, Y. Gao, L. Ma, MISSU: 3D medical image segmentation via self-distilling TransUNet, IEEE Trans. Med. Imaging (2023).

[22] E.H. Mamdani, S. Assilian, An experiment in linguistic synthesis with a fuzzy logic controller, Int. J. Man-Mach. Stud. 7 (1) (1975) 1–13.

[23] M. Figueiredo, F. Gomide, Design of fuzzy systems using neurofuzzy networks, IEEE Trans. Neural Netw. 10 (4) (1999) 815–827.

[24] J.S.R. Jang, et al., Fuzzy modeling using generalized neural networks and Kalman filter algorithm, in: AAAI, vol. 91, 1991, pp. 762–767.

[25] K. Faber, D. Zurek, M. Pietron, N. Japkowicz, A. Vergari, R. Corizzo, From MNIST to ImageNet and back: benchmarking continual curriculum learning, Mach. Learn. (2024) 1–28.

[26] L. Zhang, R. Tanno, M. Xu, Y. Huang, K. Bronik, C. Jin, J. Jacob, Y. Zheng, L. Shao, O. Ciccarelli, et al., Learning from multiple annotators for medical image segmentation, Pattern Recognit. 138 (2023) 109400.

[27] J. Krüger, R. Opfer, N. Gessert, A.-C. Ostwaldt, P. Manogaran, H.H. Kitzler, A. Schlaefer, S. Schippling, Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks, NeuroImage: Clin. 28 (2020) 102445.

[28] T.R. Dean, J.R. Cordy, A syntactic theory of software architecture, IEEE Trans. Softw. Eng. 21 (4) (1995) 302–313.

Dr. Kevin Bronik is a postdoc researcher at the University of Oxford. Before that, he was a research fellow at the Queen Square Multiple Sclerosis (MS) Centre, Institute of Neurology, University College London. He received his Ph.D. degree from University of Cardiff, United Kingdom. His research lies on data analysis and machine learning.

Dr. Le Zhang is an Assistant Professor in School of Engineering, College of Engineering and Physical Sciences at the University of Birmingham, affiliated with Digital Healthcare and Medical Imaging Research Group. Dr. Zhang was a Researcher at the University of Oxford since 2022. Before that, he was a Research Fellow at University College London since 2019 working with Prof. Daniel Alexander. Under the supervision of Prof. Alejandro F Frangi, he obtained his Ph.D. in Medical Image Computing from the University of Sheffield in 2019.