IEEE Computational Intelligence Society Transactions on Fuzzy Systems

DCNFIS: Deep Convolutional Neuro-Fuzzy Inference System

Journal:	*Transactions on Fuzzy Systems*
Manuscript ID	TFS-2023-0053
Manuscript Type:	Full Paper
Date Submitted by the Author:	28-Feb-2023
Complete List of Authors:	Yeganejou, Mojtaba; University of Alberta, Electrical & Computer Engineering Honari, Kimia; University of Alberta, Electrical & Computer Engineering Kluzinski, Ryan; University of Alberta, Electrical & Computer Engineering Dick, Scott; University of Alberta, Electrical & Computer Engineering Lipsett, Michael; University of Alberta, Electrical & Computer Engineering Miller, James; University of Alberta, Electrical & Computer Engineering
Keywords:	Fuzzy logic, Neuro-fuzzy systems, Deep learning, Explainable artificial intelligence

DCNFIS: Deep Convolutional Neuro-Fuzzy Inference System

Mojtaba Yeganejou, Kimia Honari, Ryan Kluzinski,
Scott Dick, *Member, IEEE,* Michael Lipsett, *Member, IEEE,*, James Miller, *Member, IEEE*

***Abstract*—A key challenge in eXplainable Artificial Intelligence is the well-known tradeoff between the transparency of an algorithm (i.e., how easily a human can directly understand the algorithm, as opposed to receiving a post-hoc explanation), and its accuracy. We report on the design of a new deep network that achieves improved transparency without sacrificing accuracy. We design a hybrid deep learning algorithm based in part upon fuzzy logic, which performs as accurately as existing convolutional neural networks on four well-known datasets. We exploit the transparency of fuzzy logic by deriving explanations, in the form of saliency maps, from the fuzzy rules encoded in the network. We investigate the properties of these explanations in greater depth using the Fashion-MNIST dataset.**

***Index Terms*—Explainable artificial intelligence, Deep learning, Machine learning, Fuzzy logic, Neuro-fuzzy systems.**

I. Introduction

DEEP neural networks are the heart of the "new" Artificial Intelligence (AI) (as exemplified by AlphaGo's defeat of a human champion), and have become the approach of choice to problems including image recognition [1], Natural Language Processing (NLP) [2], speech recognition [3], and many others. However, the Achilles' heel of deep neural networks is that the knowledge they contain is encoded as a distributed pattern of potentially millions of connection weights [4], a format that humans cannot understand. Add to this the fact that some of their decisions seem counterintuitive or even inexplicable to human experts, and it seems likely that humans would be reluctant to trust such a system. Indeed, the literature bears this out [5]–[7]. There is a risk that, far from embracing the AI revolution, human society might reject it if new mechanisms to foster trust in AI are not developed. A classical approach to the interpretability problem is to incorporate an "explanation mechanism" in AI algorithms [5]–[8]. (Such approaches have recently been dubbed eXplainable Artificial Intelligence, XAI.) Explanations in turn enable knowledge discovery [9], algorithm verifiability [10], and even legal compliance (by satisfying the "right to an explanation" that increasingly appears in privacy protection legislation) [11]. A discussion of the history of XAI techniques can be found in [12] and some of the challenges of XAI are discussed in [13]–[15]. Prior research in XAI for shallow learning focused on rule extraction from trained neural networks; see [16] for a discussion. Recent work in this vein for deep networks includes rule extraction (e.g. [17]), but seems to focus more upon visualizations [18], [19]. Alternatively, one could directly design a deep network architecture to be more interpretable, following e.g. the ideas of fuzzy neural networks and neuro-fuzzy systems (e.g. the Adaptive Neuro-Fuzzy Inference System, ANFIS) [20], [21]; interpretability is a key system goal for fuzzy systems and their hybrids [22]. However, there has long been a tradeoff observed between algorithm interpretability and predictive accuracy; interpretable systems tend to be less accurate, while more accurate ones are less transparent, or even opaque to human understanding [23], [24].

Our solution (applicable to any CNN) is to first remove the final dense layers of the network, leaving only the network's convolutional base [25]. We then concatenate a modified ANFIS neuro-fuzzy system [21] to it as a new classifier (equivalently, we use the convolutional base as an automated feature extractor for the ANFIS classifier), producing a family of architectures we refer to as Deep Convolutional Neuro-Fuzzy Inferential Systems (DCNFIS) [26]. With a few modifications to the ANFIS algorithm, we are able to perform end-to-end training on DCNFIS; our experiments used the ADAM optimizer [27], but any other optimizer can be used for the training of DCNFIS. Experiments on the MNIST Digits, Fashion-MNIST, CIFAR-10, and CIFAR-100 datasets indicate that DCNFIS is as accurate as the base CNNs they are built from (we have tested our technique on the LeNet, ResNet, and Wide ResNet architectures), unlike the deep fuzzy convolutional networks proposed in [28]–[30]. Those latter papers proposed generating saliency maps only for the medoid elements of a cluster. We further investigate this mechanism in DCNFIS; based on findings from [31], our saliency maps are generated via Guided Backpropagation.

Our contributions in this paper are, firstly, the design and evaluation of a family of novel hybrid deep-learning algorithms, incorporating convolutional neural networks and fuzzy logic. Secondly, we investigate the medoid saliency maps created from DCNFIS using the Fashion-MNIST dataset.

The remainder of this paper is organized as follows. In Section 2 we review essential background on deep learning, neuro-fuzzy systems, and their hybridization. In Section 3

M. Yeganejou, K. Honari, R. Kluzinski, S. Dick, J. Miller are with the Dept. of Electrical & Computer Engineering, M. Lipsett the Dept. of Mechanical Engineering, University of Alberta, Edmonton, AB, Canada, T6G 1H9. Emails: yegaenjo@ualberta.ca, honari@ualberta.ca, rkluzins@ualberta.ca, sdick@ualberta.ca, mike.lipsett@ualberta.ca, jimm@ualberta.ca.

This research was funded in part by the Natural Science and Engineering Research Council of Canada under grant nos. RGPIN 2017-05335 and CRDPJ 543705

we conduct a critical inquiry into the concept of "trusted AI," and how XAI contributes to it. In Section 4 we present our proposed architecture and our methodology for evaluating its performance. Our experimental results characterizing the accuracy of the method are presented in Section 5, and our interpretations are demonstrated in Section 6. We close with a summary and discussion of future work in Section 7.

II. Background Review

A. Neuro-Fuzzy Classification

Fuzzy classifiers [32], [33] assume the boundary between two neighboring classes is a continuous, overlapping area within which an object has partial membership in each class. These classifiers provide a simple and understandable representation of complex models using linguistic if-then rules. The rules are of the basic form:

$$if\ X_1\ is\ A\ and\ X_2\ is\ B\ then\ Z\ is\ C \tag{1}$$

where X_1 and X_2 are the input variables of the classifier; A and B are linguistic terms [34], characterized by fuzzy sets [35], which describe the features of an object; and C is a class label. The firing strength of this rule with respect to a given object represents the degree to which this object belongs to the class C.

B. Convolutional Neural Networks

Modern Convolutional Neural Networks (CNNs) were first described by LeCun et al. in [36]. In that work, the network consisted of stacks of alternating convolution and pooling layers only. More recent CNN architectures have added additional operations or layers to incorporate additional properties. For example, Local Response Normalization (LRN) layers have been added to CNNs in order to implement the concept of lateral inhibition from human vision [37]. Empirically, feature extraction via deep CNNs is currently the leading approach for building accurate neural models [36]–[39].

One recent example is the ResNet architecture [40], which focuses on correcting an observed degradation in training accuracy for CNNs with a large number of layers [1]. The ResNet approach is to change the transfer function of the layer, by adding shortcut links. The network is trained to mimic the mapping F(X)=H(X)-X, where H(X) is the actual optimization target. The original input/output mapping is refactored as F(X) + X. Empirically, this residual appears easier to learn than the original mapping. An updated version of ResNet was recently published in [41]. Wide Residual Networks (WRN) [42] are another developed version of ResNets. ResNet architectures are very deep, and they have the problem of diminishing feature reuse, which makes these networks very slow. WRN architectures are wider (more feature maps per convolutional layer) while they have less depth (fewer convolutional layers), resulting in faster training.

C. Deep learning and fuzzy systems

Hybridizations of fuzzy logic and shallow neural networks have been investigated for over 25 years [43]. Hybridizations of fuzzy logic and deep networks, on the other hand, have only received substantial attention in the last five years. As with the older literature, we can distinguish between fuzzy neural networks - in which the network architecture remains the same, but neuron-level operations are fuzzified in some way – and neuro-fuzzy systems wherein the network architecture is altered to mimic fuzzy inference algorithms [43]. In the former group we find both hybridizations with the basic (type-1) fuzzy logic (e.g. fuzzifying inputs to a deep MLP network [44], and ones involving extensions of type-1 fuzzy sets. Chen et al.'s fuzzy restricted Boltzmann machine [45], for instance, is refactored to use Pythagorean fuzzy sets [46] by Zheng et al. [47], and interval type-2 fuzzy sets by Shukla et al. [48]. They were also stacked to form a fuzzy deep belief network in [49]. Zheng et al. hybridized Pythagorean fuzzy sets and stacked denoising autoencoders in [50]. Echo state networks were modified to include a second reservoir in [51], which uses fuzzy clustering for feature reinforcement (combatting the vanishing gradient problem). Layers of these deep fuzzy echo state networks are then stacked. Stacked fuzzy rulebases, learned via the Wang-Mendel algorithm, are proposed in [52]. An evolving fuzzy neural network for classifying data streams is presented in [53]. A deep fuzzy network for software defect prediction, which incorporates stratification to rectify the class imbalance problem, is proposed in [54]. The algorithm can add or merge layers in response to concept drifts within the stream. Fuzzy clustering is merged with the training of stacked autoencoders by adding constraints the optimize compactness and separation of clusters and within-class affinity to the network's objective function in [55]. A variation on this theme is using fuzzy logic as a preprocessor; in [56], pixels in an image were mapped to triangular fuzzy membership functions, which are three-parameter functions. Feature maps consisting of each of those parameters were then passed to three different NNs, and their outputs fused. Classification errors were reduced on several benchmark datasets. Fuzzy c-means clustering was used to preprocess images for stacked autoencoders in [57]. Fuzzy logic is woven into a ResNet model for segmenting lips within human faces in [58]. The Softmax classification layer is replaced with a fuzzy tree model trained using fuzzy rough set theory in [59].

In the latter group, Aviles et al. [60] engineered a hybrid of ANFIS, recurrent networks and the Long Short-Term Memory for the contact-force problem in remote surgery. Rajurkar and Verma [61] design stacked TSK fuzzy systems, while [62] is a more complex stacked TSK with a focus on interpretability. An adversarial training algorithm for a stacked TSK system is proposed in [63]. Tan et al. suggest using fuzzy compression to prune redundant parameters from a CNN [64]. A specialized neural network for solving polynomial equations with Z-number coefficients is proposed in [65]. A neural network that mimics the fuzzy Choquet integral is proposed in [66]. John et al. [67] employ FCM-based image segmentation to preprocess images before training a CNN, while [68] and [69] instead used the deep network (restricted Boltzmann machines and Resnet, respectively) as a feature extractor prior to clustering. This last is quite similar to the approach in [30] and [29]: the densely-connected layers at the end of a deep network are replaced with an alternative classifier built from a clustering algorithm. The latter two, however, employ the fuzzy C-means and Gustafson-Kessel fuzzy clustering algorithms, respectively. The current paper also follows this basic paradigm.

III. eXplainable AI (XAI)

As often stated, the goal of XAI is to foster user trust in AI algorithms. Anecdotally, that trust is presently very low. Findings in [70] indicate that employees prefer co-workers who are real, trustworthy human beings. AI is only entrusted with low-risk, repetitive work [70], [71]. As one study participant stated, "I don't think that an AI, no matter how good the data input into it, can make a moral decision." [71]

However, there has been relatively little critical inquiry into just what is meant by "trust in AI." In point of fact, trust is a complex, mercurial, situation-dependent concept that defies easy definition. A review in [72] found that various authors define trust as a social structure, a verb, a noun, a belief, a personality trait... the list goes on. It has been studied in the fields of psychology, sociology, organizational science, and information technology (IT); the work in the latter area is perhaps the most congruent to this discussion [73].

There are of course some commonalities in how trust is conceptualized across the various domains. Perhaps the most common of these is that the act of extending trust necessarily renders the person extending it (the trustor) vulnerable to the person being trusted (the trustee) [74]–[76]. Extending trust is thus not a trivial act, and a number of factors influence a person's decision to trust another or not. The Theory of Reasoned Action (TRA) groups these together as cognitive trust (trust based on rational assessment) and emotive trust (an emotional drive to extend trust, or not) [76]. In addition, there are two forms of trust: relational trust is specific to a particular individual trustee, while generalized trust is a trustor's belief about the trustworthiness of people in general. Evidence from behavioral psychology indicates that the two are distinct [73], [74]

Focusing on the IT literature specifically, a meta-study in [73] found that "online trust" (a user's trust in the persons/vendor operating a website) is influenced by at least 18 different factors. Perceived risk, for instance, is an obvious and potent influence on the decision to trust. Actions that reduce risk tend to improve cognitive trust. In contrast, predictability (a website's conformance to expectations, both directly concerning that website and pages similar to it) seems to impact online trust by engendering a more positive attitude towards the website (i.e., improving emotive trust). Cognitive enjoyment also positively impacts attitude, as does ease of use. Considered more widely, the findings in [73] indicate that users search for signals that an online vendor is trustworthy. What we thus find is that online trust proceeds from generalized trust; there is no individual trust relationship formed.

What, then, are the implications for XAI? Let us first consider how trust is defined. Several definitions appear in the XAI literature. There is of course the classic meaning of trust as a willingness to be vulnerable to another entity [77]; as pointed out in [78], there is a correlation between how vulnerable the user is to the entity, (i.e. the level of risk) and how reliant they are on it [79], [80]. There also seems to be a tendency for the most automated systems to pose the greatest risk [81].

In a definition more directly aimed at autonomous systems, Lee and See define trust as an "attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [71], [82]; this seems to be the most commonly adopted definition in the XAI literature. Israelsen and Ahmed define it as "a psychological state in which an agent willingly and securely becomes vulnerable, or depends on, a trustee (e.g., another person, institution, or an autonomous intelligent agent), having taken into consideration the characteristics (e.g., benevolence, integrity, competence) of the trustee" [83]. Others take a more behaviorist slant; trust was "the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid" in [84].

Explanations, then, are a means of persuading users to be vulnerable to an AI, in exchange for the AI's services. As Miller points out, explanation is a social act; an explainer communicates knowledge about an explanandum (the object of the explanation) to the explainee; this knowledge is the explanans [85]. In XAI, a decision or result obtained from an AI model is the explanandum, the explainee is the user who needs more information, the explainer is the Explanation Interface (EI) of the AI, and the explanans is the information communicated to the user [86].

The XAI field also employs some of the theoretical frameworks for conceptualizing trust used in IT. These each posit several factors that influence trust (some overlapping, some not as found in [73]). XAI researchers then hypothesize that an explanans influences one or more of these factors, thus influencing user trust in the AI. One of the most common frameworks postulates Ability, Benevolence, and Integrity (the so-called ABI model) as major factors influencing user trust in AI [77], [87](the ABI+ variant adds the fourth factor of Predictability [88]). The constructs of the Technology Adoption Model (TAM) [89], particularly Ease of Use and Usefulness, are also frequently used to frame XAI research [90]. Trust is also related to Usefulness of the Explanation and User Satisfaction in [90]. A few other models have been put forward; for instance, Ashoori and Weisz propose seven factors influencing trust in an AI; trust itself was a multidimensional construct, measured by five psychometric scales [71].

There is, however, a crucial difference between trust in IT and in AI in the literature: the basic question of "whom are we trusting?" The locus of trust (LoT), the person being trusted, is key to the trusting relationship. A website, for instance, is created and completely controlled by humans. They are clearly the LoT [91]–[93]. A large deep network, on the other hand, cannot be said to be fully under human control; it has almost certainly discovered and exploited patterns in its input data that were not perceptible to the human analyst. Certainly, the designers have influence over it, but typically not even they grasp the entirety of the induced model. Are we instead trusting the AI itself?

A number of researchers do treat the AI itself as the LoT, e.g. [78], [83], [84], [94]. More significantly, evidence indicates that users themselves react to AI agents as persons, rather than programmed constructs. Participants in user studies treated autonomous systems as intentional beings in e.g. [91]. People were found to respond to a robot's social presence in [95].

Alternatively, we can view the LoT as being the designers of an AI. It is the designers, after all, who determine what biases and moral values are or are not embedded in AI code or in the training of a machine learner. It is the designers who determine how fairness, ethics, transparency, and safety are realized within the AI. It is the designers who will monitor and correct the learning process for the AI, and will guide its ongoing adaptation to real-world usage. Thus, [96] and others argue that the human designers should be the LoT, and held responsible for the AI's behavior. Similarly, accountability for the decisions of an AI is another aspect or trustworthiness [97], [98]; one that must necessarily be borne by humans.

IV. Methodology

A. Datasets

We have chosen 4 commonly-used benchmark datasets for deep learning: MNIST [99], Fashion MNIST [100], CIFAR-10 [37], and CIFAR-100 [101]. MNIST is a collection of grayscale images, each representing one handwritten digit and labeled with the correct value. The samples were gathered from roughly 250 individuals; one half are U.S. Census Bureau employees, while the other half are high school students. MNIST is pre-partitioned into 60,000 training samples and 10,000 testing samples; the writers for the two groups are disjoint. Each MNIST image is a 28x28 pixel square, centered on the center of mass of the digit pixels. These images are constructed from a set of 20x20 pixel square images containing the actual digits, size-normalized to this bounding box. While the original samples were binary images, the anti-aliasing technique used in the normalization introduced gray levels. See [99] for more details. Fashion MNIST shares the same image size and structure of training and testing splits with MNIST. Each Fashion MNIST sample is a 28x28 grayscale image, associated with a label from one of ten classes (t-shirts, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots) [100]. CIFAR-10 [37] is an RGB image dataset, consisting of 60,000 images of size 32x32 in ten mutually exclusive classes, with 6000 images per class, again pre-partitioned into training and testing sets. There are 50,000 training images and 10,000 testing images. CIFAR-100 [101] is similar to CIFAR-10, except it has 100 classes. This dataset also includes 50,000 training images and 10,000 testing images.

B. Proposed Architecture

Our fuzzy classifier is based on the Adaptive Neuro-Fuzzy Inference System (ANFIS) architecture [21], [102]. ANFIS is known to be a universal approximator [21], and is thus theoretically equal to the fully-connected layer(s) it replaces

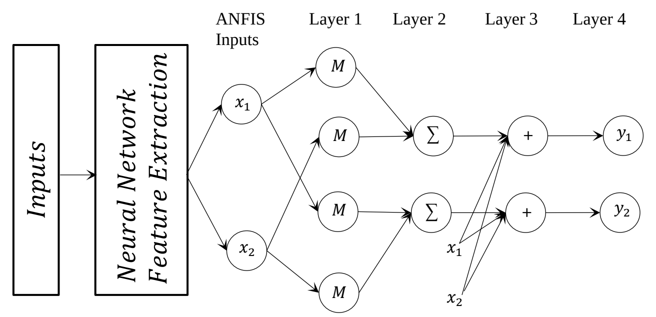

Fig. 1. Simplified Block Diagram of our Deep Convolutional Neuro-Fuzzy Classifier

in the ResNet or LeNet architectures. ANFIS is a layered architecture where the first layer computes the membership of an input in a fuzzy set (in this case, defined by a Gaussian membership function) using (6), for fuzzy rule i and input j. There is one membership per input for each rule; Fig. 1 shows a system with 2 inputs and 2 fuzzy rules, resulting in 4 membership functions. We compute the natural logarithm of the membership values, for numerical stability in the tails.

$$\begin{gathered} M_{ij} = \exp(-\beta_{ij}(x_j - \mu_{ij})^2) \\ \log M_{ij} = -\beta_{ij}(x_j - \mu_{ij})^2 \end{gathered} \tag{2}$$

Layer 2 computes the firing strength of each fuzzy rule. This is usually the product of all linked membership functions; however, as we are using logarithms, we sum the log memberships

$$\begin{gathered} \omega_i = \prod_j M_{ij} \\ \log \omega_i = \sum_i \log M_{ij} \end{gathered} \tag{3}$$

In layer 3, the activation is normally computed as a linear combination of the input variables, multiplied by the normalized firing strength of the respective rule. This is changed to the sum of the log firing strength and the linear combination of the input variables.

$$\begin{gathered} \zeta_i = \bar{\omega}_i(\sum_j W_{ij} x_j + b_i) \\ \zeta_i = \log \bar{\omega}_i + (\sum_j W_{ij} x_j + b_i) \end{gathered} \tag{4}$$

Finally, layer 4 computes the class probabilities using the softmax activation function

$$P\left(c_i \middle|\, x\right) = y_i = \frac{\exp\left(\zeta_i\right)}{\sum_j \exp\left(\zeta\right)_i} \tag{5}$$

This architecture mimics the operation of a Fuzzy Inferential System (FIS), which is a rule-based expert system (and thus highly interpretable). Specifically, the layer transfer functions mimic the stages of processing by which an FIS infers an output from its inputs. The hybrid learning rule for ANFIS (a Kalman filter in Layer 4, gradient descent for Layer 1) allows the fuzzy rule-base to be induced from a dataset. There is a

1:1 correspondence between ANFIS and the FIS it mimics, and so we can directly translate a trained ANFIS into fuzzy rules. However, this only applies to the direct inputs of ANFIS - which are the outputs of the final CNN layer. We defer further discussion of this point to our discussion of future work in Section 6.

Note that ANFIS does not provide a shortcut to the softmax layer. Layer 3 of ANFIS implements a linear combination of the network inputs for each individual rule. In the basic ANFIS, the rule is weighted by its firing strength (computed in Layer 2), and then passed to a summation. In our implementation, we have taken the logarithm of the network signals, and so we add the log of the firing strength to the weighted sum in Layer 3, and then pass the sum to the softmax function in Layer 4. As the logarithm function is monotonic increasing, the ranks of the softmax outputs should be the same as for the original ANFIS, even if the logit values have changed.

C. Learning Algorithm

Our fuzzy classifier will use the Adam [27] algorithm for learning the adaptive parameters of the network. The gradients of the linear fully connected layer of DCNFIS are easily calculated. The gradients passed to the convolutional layers are calculated as follows:

$$\begin{aligned}\frac{\partial \zeta_i}{\partial x_j} &= \frac{\partial}{\partial x_j}\log \omega_i + \sum_j W_{ij} \\ &= \frac{\partial}{\partial x_j}\log\left(M_{ij}\right) + \sum_j W_{ij} \\ &= -2\beta_{ij}\left(x_j - \mu_{ij}\right) + \sum_j W_{ij}\end{aligned} \tag{6}$$

Where the gradients of the membership function parameters can be calculated as follows:

$$\begin{aligned}\frac{\partial \zeta_i}{\partial log \omega_i} &= 1 \\ \frac{\partial \zeta_i}{\partial log\ M_{ij}} &= 1 \\ \frac{\partial \zeta_i}{\partial \mu_{ij}} &= 2\beta_{ij}\left(x_j - \mu_{ij}\right) \\ \frac{\partial \zeta_i}{\partial \beta_{ij}} &= -\left(x_i - \mu_{ij}\right)^2\end{aligned} \tag{7}$$

D. Experimental Setup

The existing literature on all four datasets reports results based on a single-split design; for comparability we will follow this design, employing the same test set of 10,000 images as our out-of-sample evaluation. We test LeNet [36] with our modifications, then ResNet and Wide ResNet. Our tests employ the second version of ResNet20 [41] for MNIST, ResNet29 for Fashion MNIST, and ResNet164 for CIFAR-10 and CIFAR-100. We have tested WRN-28-10 for CIFAR-10, and WRN-28-12 for CIFAR-100 [42].

The LeNet-based architectures, are trained for 200 epochs in batches of 128. The parameters for the Adam optimizer have been set as $\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. For the ResNet and WRN-based experiments the models were trained for 200 epochs in batches of 128. The learning rate has been set to 0.001 initially, and then it will be reduced to 1e-6 after epoch 180. Data augmentation techniques are applied on all datasets except MNIST. All the experiments are replicated ten times.

TABLE I
DCNFIS AND REGULAR-CNNS PERFORMANCE COMPARISON

Dataset	LeNet			
	Regular		DCNFIS	
	AVG	STD	AVG	STD
MNIST	99.21	0.0404	**99.23**	0.462
Fashion-MNIST	**90.36**	0.2364	90.10	0.1445
CIFAR-10	73.11	0.4550	**73.39**	0.9267
CIFAR-100	43.25	0.6594	**43.30**	0.5472
	ResNet			
	Regular		DCNFIS	
	AVG	STD	AVG	STD
MNIST	99.57	0.381	**99.59**	0.0288
Fashion-MNIST	**94.64**	0.0816	94.401	0.2308
CIFAR-10	**93.13**[a]	0.0844	93.02	0.0526
CIFAR-100	74.52	0.0978	74.52	0.1589
	WRN			
	Regular		DCNFIS	
	AVG	STD	AVG	STD
CIFAR-10	96.620	0.23870	**96.680**	0.1789
CIFAR-100	**78.50**[a]	0.1077	77.53	0.3523

a-Significant at α = 0.05

V. EVALUATION

As shown in TABLE I, the performance difference between the original CNNs and their DCNFIS versions is very minor; about half the time, the base architecture was slightly better, and half the time the DCNFIS was better. This is a significant change from [28] and [29], in which the enhanced interpretability of the deep fuzzy system came at the price of clearly losing some overall accuracy.

We next conduct a statistical analysis of these results. As there are only ten replicates of each experiment, but the variances for the original CNN and the DCNFIS versions can be substantially different, we employ the t-test for unequal variances. At a significance level of α = 0.05, there are only two instances where the differences were significant: CIFAR-10 with ResNet, and CIFAR-100 with WRN. We thus claim that *using the DCNFIS method does not appear to reduce the accuracy of a CNN.*

VI. INTERPRETABILITY

As discussed in section 4, DCNFIS uses the rule-based ANFIS architecture as its classifier component. Each rule in ANFIS is of the form of Eq. 8, and forms a conjunction of antecedent clauses, each represented by one fuzzy subset of the corresponding input dimension. The rule itself thus defines

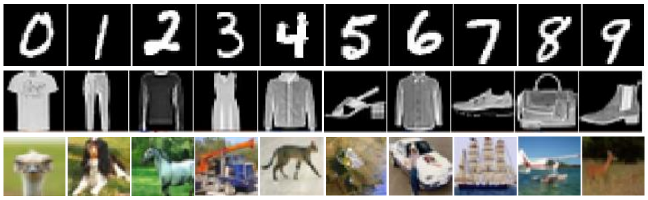

Fig. 2. Class representatives derived from fuzzy rules. Top row: MNIST. Middle row: Fashion-MNIST. Bottom row: CIFAR-10

a fuzzy region of the ANFIS input space. The collection of feature vectors within this fuzzy region can be considered a fuzzy cluster. Following [28] we select the medoid element of a cluster as the representative for that cluster. The following equation shows the if-then rules of our modified ANFIS.

$$\begin{aligned} If\ f_1(p_1)\ \ is\ M_{K,1}\ and\ f_2(p_2)\ \ is\ M_{K,2}\ and\ldots \\ f_D(P_D)\ \ is\ \ M_{K,N}\ \ then \\ \log P(c_K|image) = \sum_i W_{K,i} f_i(P_i) + b_k \end{aligned} \tag{8}$$

In Fig. 2, we present the medoid elements of the ten rules generated for each of our datasets (excluding CIFAR-100). Our approach to XAI is to treat the medoid as a synopsis of the entire cluster, with the saliency map computed for it constituting our explanans for the cluster. The general process is to train DCNFIS, and extract the fuzzy regions for each rule of the classifier. The region is treated as a fuzzy cluster, and the medoid element is identified. Next, we perform a saliency analysis on the medoids, using the Guided Backpropagation algorithm [103]. As the medoid element is arguably the most representative of the whole class, we build our explanations on the saliency maps of the ten medoids. More details on the method, and an empirical evaluation of different saliency analyses, can be found in [28], [29].

The power of DCNFIS in comparison with previous methods described in [28]–[30] is that, by contrast, all the medoids are extracted from the rules of DCNFIS; there is no need for any further post training clustering and classification process.

We next demonstrate our medoid-based explanans on Fashion-MNIST. Our discussion in this section is inspired by our previous analysis of the MNIST Digits dataset in [28]; we focus on examining selected misclassifications in the training dataset, comparing the saliency map of the erroneous examples against the saliency maps for the medoid images in the actual and predicted classes. As a contrast, we also examine selected correct classifications from the training dataset. Following this, we use UMAP visualizations to examine the misclassifications more formally. Finally, we discuss how our explanans were useful in detecting a learning bias being introduced in our model due to a commonly-used preprocessing step.

Fig. 3 presents the class medoids derived from the fuzzy rules for the Fashion-MNIST dataset. The top row is the medoid element of each cluster/class (labeled from left to right T-shirt/Top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle Boot), and the bottom row is the corresponding saliency map.

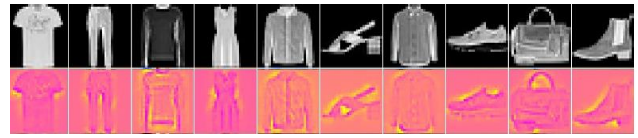

Fig. 3. Class representatives derived from fuzzy rules for Fashion-MNIST dataset. First row: Medoids. Second row: Saliency of Medoids

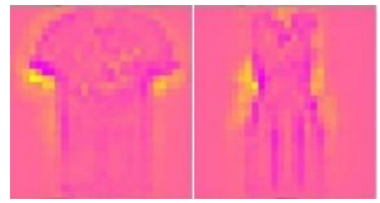

Fig. 4. Difference between T-Shirt and Dress

A. Fashion-MNIST

One of the challenges with Fashion-MNIST is the similarity of some of the classes to each other. The images and labels were taken from the Zalando online store; the labels specifically are the "silhouette code" for that item, which is manually assigned by Zalando's fashion staff (and then cross-checked by a separate team). Thus, while the labels are reliable, the distinctions between some classes are plainly finer than others. Coats and Pullovers, for instance, are only distinguished by the presence of a vertical zipper or line of buttons in the former, while the difference between a Coat and a Sandal is far more dramatic. The distinction between T-Shirt/Top and Dress is another example. As shown in Fig. 3, the shapes formed by yellow pixels (strong negative impact on class assignment) are the main difference between these two classes. In Fig. 4 we have focused on this difference. For class Dress the neural network doesn't care about having a short or long sleeve. It focuses on detection of a long almost vertical yellow line which starts from the axilla and ends at the high hip. For T-Shirts the vertical yellow pixels starting from the axilla peter out around the mid-torso. Meanwhile, a strong region of yellow pixels can be observed directly under the cutoff of each sleeve; long sleeves would thus strongly contraindicate the T-Shirt/Top class.

Our analysis of DCNFIS misclassifications begins with an examination of misclassified images, presented in Fig. 5. We selected 18 misclassified examples from the Fashion-MNIST training data. Note that this particular model has a training accuracy of 98.096 percent, so only 1,146 samples out of 60,000 are misclassified. In this figure the original samples and their saliencies are shown in the first and second columns. In the third column we show the saliency of the medoid of the actual (label) class while in the fifth column we show the saliency of the predicted class. In columns four and six we overlay the saliencies of the sample on the saliencies from the label and prediction class medoids. For example, S1 is a T-Shirt/Top sample with long sleeves which is classified as a

Fig. 5. Images Misclassified by DCNFIS

Pullover, and S2 is a T-Shirt/Top sample which has been found to be more similar to a Dress. This latter is an exemplar of a particularly challenging aspect of the T-Shirt/Top class, which we discuss in Section VI.B. What we see throughout these images is that the saliencies of each sample are noticeably more similar to the medoid of the predicted class (and the high-importance pixels especially so) than to the medoid of the labelled class. As in [28], this evidence tends to support our contention that the medoid saliences effectively capture the classification decisions of DCNFIS.

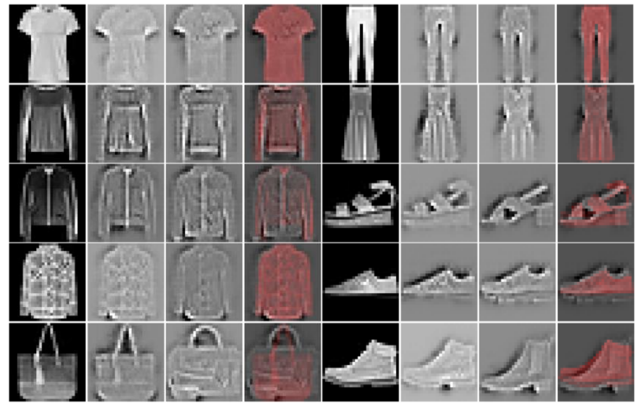

Fig. 6. Samples Correctly Classified by DCNFIS

Fig. 6 shows 10 samples (one for each class) that are correctly classified by DCNFIS. Each row shows samples of two different classes. For each class, the first image shows the original sample, and the second image shows the saliency of the sample. The third image is the saliency of the class medoid, and the two saliences are overlaid in the fourth image, similar to columns of Fig. 5. As shown in the figure, the sample and medoid saliencies are very similar. For example, an ankle boot (bottom right) sample has been classified by detection of the heel and upper surface of the boot.

Hendricks et al. [104] proposed two criteria for explanations in image classification problems: they must be class discriminative, and image relevant. In that work, Hendricks et al. build natural language texts as explanations, so image relevance is a significant challenge. However, the more general point they raise is that explanations must refer to the specific content of the image in question. Understood in this fashion, we argue that our discussion above shows that the medoid-based approach is indeed image relevant. Indeed, each row of Fig. 5 compares the saliencies of a specific image against the actual and predicted class medoid saliencies, and highlights their similarities and differences. Class discriminativeness, meanwhile, is demonstrated in Fig. 3.

B. Automated Feature Selection in DCNFIS

One of the principal advantages of CNNs is that the convolutional component acts as an automated feature selection algorithm. As DCNFIS does not change the convolution components of the base architecture, we expect that it will remain effective in this role. However, as DCNFIS is an end-to-end trainable algorithm, this expectation needs to be checked; the error signals being propagated through the fuzzy classifier component will, after all, be different than those propagated from layers of dense and SoftMax neurons. In this section, we thus evaluate this expectation on the Fashion-MNIST dataset, by visualizing the original dataset, and the outputs of the trained convolutional component, using the Uniform Manifold Approximation and Projection (UMAP) algorithm. We then further investigate selected misclassifications in these visualizations.

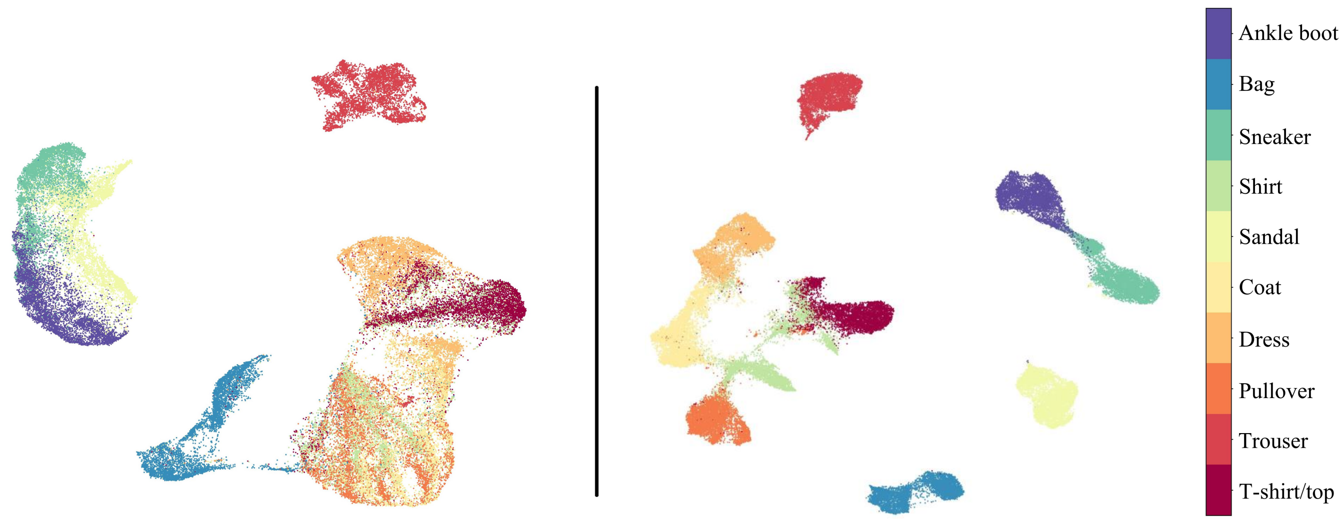

Fig. 7. Left) 2-D UMAP visualization of Fashion-MNIST Dataset. Right) 2-D UMAP visualization of the DCNFIS Classifier Component Inputs (outputs from ResNet 29)

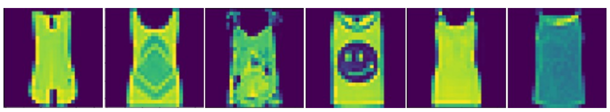

Fig. 8. Visualization of 6 samples of Top images out of 32

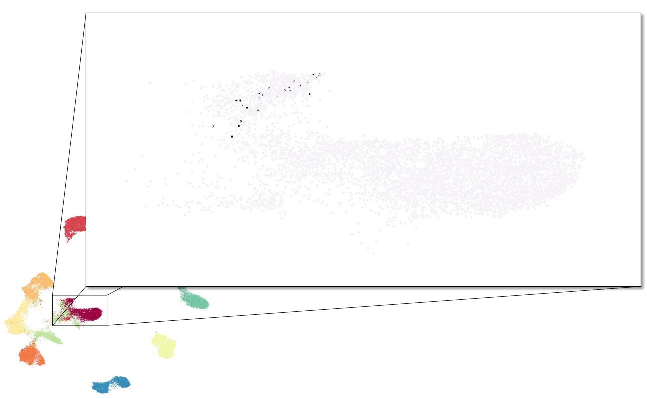

Fig. 9. Visualization of 32 selected Top samples out in the T-shirt/Top class region

Left figure in Fig. 7 shows a 2-dimensional visualization of the Fashion-MNIST training dataset, created using UMAP [105] with its default parameters. The Shirt, Coat, Dress, Pullover, and T-shirt/Top classes are badly comingled at the lower right, while the Ankle Boot, Sneaker and Sandal classes also overlap heavily at the upper left. Only the Trouser and Bag classes are well-separated. Right figure in Fig. 7 is a UMAP visualization (using the same parameters) of the training data output of the convolutional component of a trained DCNFIS network (i.e. the input to the fuzzy classifier). Plainly, there is a substantial improvement in the separation of the different classes (interpreting the compactness of the classes is more difficult, as UMAP is a nonlinear projection that balances preservation of local and global structure).

However, an important characteristic of our medoid-based explanans can be observed in the T-shirt/Top class. We manually selected the sequentially first 32 images in this class that do not resemble T-shirts (i.e. that are sleeveless; Fig. 8 presents 6 such examples). All but one of the 32 are correctly classified, but they do not closely resemble the T-shirt/Top medoid from Fig. 3 (leftmost image). When we highlight these 32 examples in Fig. 9, we find that they all occur in the "upper" lobe of the class distribution, which appears to hold a concentration of examples at some distance from the class medoid. In this context, the misclassification of a Top as a Dress in Fig. 4 is revealing; sleeveless tops seem to lie in a different grouping from the class medoid. It is possible that splitting the T-shirt/Top class into separate clusters might yield a new medoid that better represents the Tops. Put another way, the T-shirt/Top class may well consist of multiple disjuncts. In general, our medoid-based approach is likely to struggle with such classes, especially when the class is arguably a conflation of multiple real-world concepts.

C. Bias in Fashion-MNIST

One of the use cases for XAI is in debugging trained AI systems. Our experiments with DCNFIS unexpectedly offered a demonstration of this use case on the Fashion-MNIST dataset. One of the common preprocessing steps for image data is to subtract the per-pixel mean of the dataset from all images [40] (a zero-mean dataset is somewhat easier to train, as has been well-known for decades [4]). However, in our experiments, we were finding that DCNFIS was noticeably less accurate than the base CNNs. When we look at our saliency maps, a reason for this appears, as in Fig. 10.

What we see in these medoid saliencies (for the Sandal, Sneaker, and Ankle Boot classes) is what appears to be the faint image of some sort of top, indicated by pixels with negative saliency. However, no such structure appears in the actual images. The reason for this lies in an apparent bias of

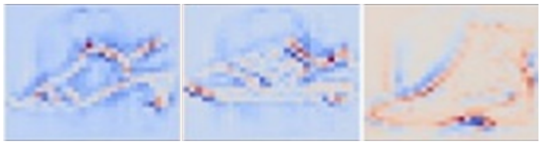

Fig. 10. Medoids of DCNFIS for classes of Sandal, Sneaker, Ankle-Boot from training data with mean subtraction

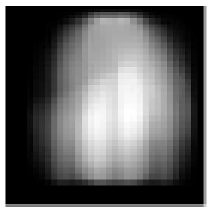

Fig. 11. Mean of training data visualized as an image

the Fashion-MNIST dataset; as has been noted, four of the ten classes are some sort of upper-body garment, and the Dress class is also similar. As a result, when we plot the per-pixel means as an image, we obtain the following (Fig. 11).

TABLE II
CONFUSION MATRIX OF DCNFIS WITH MEAN SUBTRACTION

	T-shirt top	Pullover	Coat	Shirt	Dress	Trouser	Sandal	Sneaker	Ankle boot	Bag
T-shirt top	5671	41	2	227	56	0	1	0	0	2
Pullover	58	5601	176	130	32	1	0	0	0	2
Coat	1	29	5833	84	49	2	0	0	0	2
Shirt	326	94	109	5377	90	1	0	0	0	3
Dress	21	6	68	27	5875	2	0	0	0	1
Trouser	1	1	1	3	15	5978	0	0	0	1
Sandal	0	0	0	0	0	0	5981	16	0	2
Sneaker	0	0	0	0	0	0	15	5925	60	0
Ankle boot	0	0	0	0	0	0	19	105	5874	2
Bag	0	0	1	1	2	0	0	0	0	5996

TABLE III
CONFUSION MATRIX OF DCNFIS WITHOUT MEAN SUBTRACTION

	T-shirt top	Pullover	Coat	Shirt	Dress	Trouser	Sandal	Sneaker	Ankle boot	Bag
T-shirt top	5759	19	3	192	21	0	1	0	0	5
Pullover	35	5809	75	72	9	0	0	0	0	0
Coat	0	33	5888	55	24	0	0	0	0	0
Shirt	156	75	77	5645	44	1	0	0	0	2
Dress	7	5	28	33	5926	0	0	0	0	1
Trouser	0	0	0	1	4	5995	0	0	0	0
Sandal	0	0	0	0	0	0	5984	13	3	0
Sneaker	0	0	0	0	0	0	11	5942	46	1
Ankle boot	0	0	0	0	0	0	3	84	5913	0
Bag	0	0	1	1	1	0	0	0	0	5997

We see what seems like the silhouette of a long-sleeved top garment of some sort, with pant-like structures in the lower 2/3rds of the image. The outer edges of the pants, however, seem to line up with the presumptive axilla on the top garment, thus reinforcing its shape. This observation caused us to run a further set of experiments, without mean removal. In Tables II and TABLE III, we present the confusion matrices produced by DCNFIS on the training dataset with and without mean subtraction, respectively. Plainly, eliminating mean removal improved our accuracy; the total impact was four tenths of a percentage point on the test data, which is greater than the standard deviation of our ten replications (see Table I). In other words, the error introduced by mean subtraction was sufficient to significantly alter the outcomes of our experiments. Thus, our observations in this section serendipitously demonstrate that our medoid-based explanans on DCNFIS can be effective in debugging an AI

VII. CONCLUSION

In this paper we have proposed a novel deep fuzzy network, which replaces the dense layers at the terminal end of a deep CNN with an ANFIS network. The architecture is end-to-end trainable, remains as accurate as the base CNNs is built from, and our medoid-based saliency-map explanations derived from the fuzzy rules seem more effective than extracting saliency maps for each data sample.

In future work, we will explore fuzzy classifiers with multiple clusters mapping to a single class. We will also explore a radically different explanation mechanism: since we already have a fuzzy inference system as the final classifier for this network, it should be possible to back-propagate the fuzzy sets to the input feature space, forming a new set of fuzzy rules directly mapping inputs to the final network outputs as a linguistic explanation

REFERENCES

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[2] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[3] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. rahman Mohamed, G. Dahl, and B. Ramabhadran, "Deep Convolutional Neural Networks for Large-scale Speech Tasks," *Neural Networks*, vol. 64, 2015.

[4] S. Haykin, *Neural Networks and Learning Machines, 3rd Ed.* Upper Saddle River: Pearson Education, Inc., 2009.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.

[6] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," in *21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15), Proceedings*, 2015.

[7] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MÃžller, "How to explain individual classification decisions," *Machine Learning Research 11(Jun).*, 2010.

[8] L. R. Ye and P. E. Johnson, "The impact of explanation facilities on user acceptance of expert systems advice," *Mis Quarterly*, pp. 157–172, 1995.

[9] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework." *Int Conf on Knowledge Discovery and Data Mining*, 1996.

[10] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[11] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision making and a "right to explanation"," *AI Magazine*, 2017.

[12] M. Yeganejou and S. Dick, "Explainable Artificial Intelligence and Computational Intelligence: Past and Present," *HANDBOOK ON COMPUTER LEARNING AND INTELLIGENCE: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation*, pp. 3–55, 2022.

[13] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger, "Explainable AI: the new 42?" in *International cross-domain conference for machine learning and knowledge extraction*. Springer, 2018, pp. 295–303.

[14] H. Ishibuchi and Y. Nojima, "Discussions on Interpretability of Fuzzy Systems using Simple Examples." in *IFSA/EUSFLAT Conf.*, 2009, pp. 1649–1654.

[15] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, 2018.

[16] A. B. Tickle, R. Andrews, M. Golea, and J. Diederich, "The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks," *IEEE Transactions on Neural Networks*, 1998.

[17] J. R. Zilke, E. Loza Mencía, and F. Janssen, "DeepRED – Rule Extraction from Deep Neural Networks," *Discovery Science 19th International Conference*, 2016.

[18] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[19] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, 2015.

[20] S. K. Pal and S. Mitra, "Multilayer Perceptron, Fuzzy Sets, and Classification," *IEEE Transactions on Neural Networks*, 1992.

[21] J. S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System," *IEEE Transactions on Systems, Man and Cybernetics*, 1993.

[22] M. J. Gacto, R. Alcalá, and F. Herrera, "Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures," *Information Sciences*, 2011.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[24] H. Ishibuchi and Y. Nojima, "Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning," *International Journal of Approximate Reasoning*, vol. 44, no. 1, pp. 4–31, 2007.

[25] F. Chollet, *Deep learning with Python*. Simon and Schuster, 2021.

[26] M. Yeganejou, R. Kluzinski, S. Dick, and J. Miller, "An End-to-End Trainable Deep Convolutional Neuro-Fuzzy Classifier," in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2022, pp. 1–7.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] M. Yeganejou, S. Dick, and J. Miller, "Interpretable Deep Convolutional Fuzzy Classifier," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 7, pp. 1407–1419, 2020.

[29] M. Yeganejou and S. Dick, "Improved deep fuzzy clustering for accurate and interpretable classifiers," in *2019 IEEE international conference on fuzzy systems (FUZZ-IEEE)*. IEEE, 2019, pp. 1–7.

[30] ——, "Classification via deep fuzzy c-means clustering," in *IEEE International Conference on Fuzzy Systems*, 2018.

[31] M. Yeganejou, "Interpretable Deep Covolutional Fuzzy Networks," Ph.D. dissertation, University of Alberta, 2019. [Online]. Available: https://era.library.ualberta.ca/items/3e81868d-b0d8-4a25-a154-8d1f51ee293e

[32] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Fluids Engineering, Transactions of the ASME*, 1960.

[33] S. Ovchinnikov, "Similarity relations, fuzzy partitions, and fuzzy orderings," *Fuzzy Sets and Systems*, vol. 40, no. 1, pp. 107–126, 1991.

[34] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning-I," *Information Sciences*, 1975.

[35] L. Zadeh, "Fuzzy sets," *Information and Control*, 1965.

[36] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[38] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[39] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]," *IEEE Computational Intelligence Magazine*, 2010.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] ——, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[42] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[43] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, 1997.

[44] A. Sarabakha and E. Kayacan, "Online Deep Fuzzy Learning for Control of Nonlinear Systems Using Expert Knowledge," *IEEE Transactions on Fuzzy Systems*, 2020.

[45] C. L. Chen, C. Y. Zhang, L. Chen, and M. Gan, "Fuzzy Restricted Boltzmann Machine for the Enhancement of Deep Learning," *IEEE Transactions on Fuzzy Systems*, 2015.

[46] R. R. Yager, "Pythagorean membership grades in multicriteria decision making," *IEEE Transactions on Fuzzy Systems*, 2014.

[47] Y.-J. Zheng, W.-G. Sheng, X.-M. Sun, and S.-Y. Chen, "Airline Passenger Profiling Based on Fuzzy Deep Machine Learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

[48] A. Shukla, T. Seth, and P. Muhuri, "Interval type-2 fuzzy sets for enhanced learning in deep belief networks," in *IEEE International Conference on Fuzzy Systems*, 2017.

[49] S. Feng, C. L. Philip Chen, and C. Y. Zhang, "A Fuzzy Deep Model Based on Fuzzy Restricted Boltzmann Machines for High-Dimensional Data Classification," *IEEE Transactions on Fuzzy Systems*, 2020.

[50] Y. J. Zheng, S. Y. Chen, Y. Xue, and J. Y. Xue, "A Pythagorean-Type Fuzzy Deep Denoising Autoencoder for Industrial Accident Early Warning," *IEEE Transactions on Fuzzy Systems*, 2017.

[51] S. Zhang, Z. Sun, M. Wang, J. Long, Y. Bai, and C. Li, "Deep Fuzzy Echo State Networks for Machinery Fault Diagnosis," *IEEE Transactions on Fuzzy Systems*, 2020.

[52] L. X. Wang, "Fast Training Algorithms for Deep Convolutional Fuzzy Systems with Application to Stock Index Prediction," *IEEE Transactions on Fuzzy Systems*, 2020.

[53] M. Pratama, W. Pedrycz, and G. I. Webb, "An Incremental Construction of Deep Neuro Fuzzy System for Continual Learning of Nonstationary Data Streams," *IEEE Transactions on Fuzzy Systems*, 2020.

[54] S. Liu, G. Lin, Q. L. Han, S. Wen, J. Zhang, and Y. Xiang, "DeepBalance: Deep-Learning and Fuzzy Oversampling for Vulnerability Detection," *IEEE Transactions on Fuzzy Systems*, 2020.

[55] Q. Feng, L. Chen, C. L. Philip Chen, and L. Guo, "Deep Fuzzy Clustering-A Representation Learning Approach," *IEEE Transactions on Fuzzy Systems*, 2020.

[56] P. Hurtik, V. Molek, and J. Hula, "Data Preprocessing Technique for Neural Networks Based on Image Represented by a Fuzzy Function," *IEEE Transactions on Fuzzy Systems*, 2020.

[57] L. Chen, W. Su, M. Wu, W. Pedrycz, and K. Hirota, "A Fuzzy Deep Neural Network with Sparse Autoencoder for Emotional Intention Understanding in Human-Robot Interaction," *IEEE Transactions on Fuzzy Systems*, 2020.

[58] C. Guan, S. Wang, and A. W.-C. Liew, "Lip image segmentation based on a fuzzy convolutional neural network," *IEEE Transactions on Fuzzy Systems*, 2019.

[59] Y. Wang, Q. Hu, P. Zhu, L. Li, B. Lu, J. M. Garibaldi, and X. Li, "Deep Fuzzy Tree for Large-Scale Hierarchical Visual Classification," *IEEE Transactions on Fuzzy Systems*, 2020.

[60] A. I. Aviles, S. M. Alsaleh, E. Montseny, P. Sobrevilla, and A. Casals, "A deep-neuro-fuzzy approach for estimating the interaction forces in robotic surgery," in *2016 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2016*, 2016.

[61] S. Rajurkar and N. K. Verma, "Developing deep fuzzy network with Takagi Sugeno fuzzy inference system," in *IEEE International Conference on Fuzzy Systems*, 2017.

[62] T. Zhou, F. L. Chung, and S. Wang, "Deep TSK Fuzzy Classifier With Stacked Generalization and Triplely Concise Interpretability Guarantee for Large Data," *IEEE Transactions on Fuzzy Systems*, 2017.
[63] S. Gu, F.-L. Chung, and S. Wang, "A novel deep fuzzy classifier by stacking adversarial interpretable TSK fuzzy sub-classifiers with smooth gradient information," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 7, pp. 1369–1382, 2019.
[64] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka, "Fuzzy qualitative deep compression network," *Neurocomputing*, 2017.
[65] R. Jafari, S. Razvarz, and A. Gegov, "Neural Network Approach to Solving Fuzzy Nonlinear Equations Using Z-Numbers," *IEEE Transactions on Fuzzy Systems*, 2020.
[66] M. A. Islam, D. T. Anderson, A. J. Pinar, T. C. Havens, G. Scott, and J. M. Keller, "Enabling Explainable Fusion in Deep Learning with Fuzzy Integral Neural Networks," *IEEE Transactions on Fuzzy Systems*, 2020.
[67] V. John, S. Mita, Z. Liu, and B. Qi, "Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks," in *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, 2015.
[68] Y. De la Rosa, Erick, Wen, "Data-Driven Fuzzy Modeling Using Deep Learning," *arXiv preprint arXiv:1702.07076 (2017).*
[69] Z. Zhang, M. Huang, S. Liu, B. Xiao, and T. S. Durrani, "Fuzzy Multilayer Clustering and Fuzzy Label Regularization for Unsupervised Person Reidentification," *IEEE Transactions on Fuzzy Systems*, 2020.
[70] K. Lazanyi, "Readiness for Artificial Intelligence," in *International Symposium on Intelligent Systems and Informatics (SISY)*, 2018, pp. 235–238.
[71] M. Ashoori and J. D. Weisz, "In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes," Tech. Rep., 2019.
[72] D. H. McKnight and N. L. Chervany, "What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology," *International journal of electronic commerce*, vol. 6, no. 2, pp. 35–59, 2001.
[73] P. Beatty, I. Reay, S. Dick, and J. Miller, "Consumer trust in e-commerce web sites: A meta-study," *ACM Computing Surveys*, vol. 43, no. 3, pp. 14:1–14:46, 2011.
[74] J. B. Rotter, "A new scale for the measurement of interpersonal trust." *Journal of personality*, 1967.
[75] T. R. Tyler, "Why People Obey the Law," 1990.
[76] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross-discipline view of trust," *Academy of management review*, vol. 23, no. 3, pp. 393–404, 1998.
[77] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model Of Organizational Trust," *Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995.
[78] J. Lyons, N. Ho, J. Friedman, G. Alarcon, and S. Guznov, "Trust of Learning Systems: Considerations for Code, Algorithms, and Affordances for Learning," in *Human and Machine Learning*, J. Zhou and F. Chen, Eds. Cham: Springer International Publishing AG, 2018, pp. 265–278.
[79] J. D. Lee and B. D. Seppelt, "Human Factors in Automation Design," in *Springer Handbook of Automation*, S. Y. Nof, Ed. Berlin: Springer-Verlag, pp. 417–436.
[80] J. B. Lyons and C. K. Stokes, "Human–Human Reliance in the Context of Automation," *Human Factors*, vol. 54, no. 1, pp. 112–121, 2011.
[81] L. Onnasch, C. D. Wickens, H. Li, and D. Manzey, "Human performance consequences of stages and levels of automation: An integrated meta-analysis," *Human Factors*, vol. 56, no. 3, pp. 476–488, 2014.
[82] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society2*, vol. 46, no. 1, pp. 50–80, 4.
[83] B. W. Israelsen and N. R. Ahmed, ""Dave...I can assure you ...that it's going to be all right ..." A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships," pp. 1–37, 2019.
[84] J. Drozdal, J. Weisz, D. Wang, G. Dass, B. Yao, C. Zhao, M. Muller, L. Ju, and H. Su, "Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems," 2020. [Online]. Available: http://arxiv.org/abs/2001.06509%0Ahttp://dx.doi.org/10.1145/3377325.3377501
[85] J. Overton, "Scientific Explanation and Computation." in *ExaCt*, 2011, pp. 41–50.
[86] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.
[87] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel, "The relationship between trust in AI and trustworthy machine learning technologies," in *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 272–283.
[88] G. Dietz and D. N. D. Hartog, "Measuring trust inside organisations," *Personnel Review*, vol. 35, no. 5, pp. 557–588, 2006.
[89] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.
[90] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," 2018.
[91] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," *Information systems research*, vol. 13, no. 3, pp. 334–359, 2002.
[92] D. Gefen, I. Benbasat, and P. Pavlou, "A research agenda for trust in online environments," *Journal of Management Information Systems*, vol. 24, no. 4, pp. 275–286, 2008.
[93] P. B. Lowry, A. Vance, G. Moody, B. Beckman, and A. Read, "Explaining and predicting the impact of branding alliances and web site quality on initial consumer trust of e-commerce web sites," *Journal of Management Information Systems*, vol. 24, no. 4, pp. 199–224, 2008.
[94] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems," 2018.
[95] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The Benefits of Interactions with Physically Present Robots over Video-Displayed Agents," *International Journal of Social Robotics*, vol. 3, pp. 41–52, 2011.
[96] G. J. Nalepa, M. van Otterlo, S. Bobek, and M. Atzmueller, "From context mediation to declarative values and explainability," in *Proceedings of the IJCAI/ECAI Workshop on Explainable Artificial Intelligence (XAI 2018). IJCAI, Stockholm*, 2018.
[97] V. Beaudouin, I. Bloch, D. Bounie, S. Clémençon, F. D'Alché-Buc, J. Eagan, W. Maxwell, P. Mozharovskyi, and J. Parekh, "Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach," *SSRN Electronic Journal*, 2020.
[98] S. R. Islam, W. Eberle, and S. K. Ghafoor, "Towards quantification of explainability in explainable artificial intelligence methods," in *The thirty-third international flairs conference*, 2020.
[99] "Dataset, MNIST." [Online]. Available: http://yann.lecun.com/exdb/mnist/
[100] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.
[101] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," *... Science Department, University of Toronto, Tech. ...*, 2009.
[102] C.-T. Sun and J.-S. Jang, "A neuro-fuzzy classifier and its applications," in *[Proceedings 1993] Second IEEE International Conference on Fuzzy Systems*. IEEE, 1993, pp. 94–98.
[103] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.
[104] L. A. Hendricks, A. Rohrbach, B. Schiele, T. Darrell, and Z. Akata, "Generating visual explanations with natural language," *Applied AI Letters*, vol. 2, no. 4, 2021.
[105] L. McInnes and J. Healy, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.