A Decision Tree-Initialised Neuro-fuzzy Approach for Clinical Decision Support

Chen, Tianhua; Shang, Changjing; Su, Pan; Keravnou-Papailiou, Elpida; Zhao, Yitian; Antoniou, Grigoris; Shen, Qiang

Published in:
Artificial Intelligence in Medicine

DOI:
10.1016/j.artmed.2020.101986

Publication date:
2021

Citation for published version (APA):
Chen, T., Shang, C., Su, P., Keravnou-Papailiou, E., Zhao, Y., Antoniou, G., & Shen, Q. (2021). A Decision Tree-Initialised Neuro-fuzzy Approach for Clinical Decision Support. *Artificial Intelligence in Medicine*, *111*, [101986]. https://doi.org/10.1016/j.artmed.2020.101986

A Decision Tree-initialised Neuro-fuzzy Approach for Clinical Decision Support

Tianhua Chen[a,*], Changjing Shang[b], Pan Su[c,e], Elpida Keravnou-Papailiou[d], Yitian Zhao[c], Grigoris Antoniou[a], Qiang Shen[b]

[a]*Department of Computer Science, School of Computing and Engineering, University of Huddersfield, Huddersfield, UK*
[b]*Department of Computer Science, Faculty of Business and Physical Science, Aberystwyth University, Aberystwyth, UK*
[c]*Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences, China*
[d]*Department of Computer Science, University of Cyprus, Cyprus*
[e]*School of Control and Computer Engineering, North China Electric Power University, Baoding, China*

Abstract

Apart from the need for superior accuracy, healthcare applications of intelligent systems also demand the deployment of interpretable machine learning models which allow clinicians to interrogate and validate extracted medical knowledge. Fuzzy rule-based models are generally considered interpretable that are able to reflect the associations between medical conditions and associated symptoms, through the use of linguistic if-then statements. Systems built on top of fuzzy sets are of particular appealing to medical applications since they enable the tolerance of vague and imprecise concepts that are often embedded in medical entities such as symptom description and test results. They facilitate an approximate reasoning framework which mimics human reasoning and supports the linguistic delivery of medical expertise often expressed in statements such as 'weight low' or 'glucose level high' while describing symptoms. This paper proposes an approach by performing data-driven learning of accurate and interpretable fuzzy rule bases for clinical decision support. The approach starts with the generation of a crisp rule base through a decision tree learning mechanism, capable of capturing simple rule structures. The crisp rule base is then

*Corresponding author
Email address: T.Chen@hud.ac.uk (Tianhua Chen)

transformed into a fuzzy rule base, which forms the input to the framework of adaptive network-based fuzzy inference system (ANFIS), thereby further optimising the parameters of both rule antecedents and consequents. Experimental studies on popular medical data benchmarks demonstrate that the proposed work is able to learn compact rule bases involving simple rule antecedents, with statistically better or comparable performance to those achieved by state-of-the-art fuzzy classifiers.

Keywords: Clinical decision support, medical diagnostic systems, fuzzy rule-based systems.

---

1. Introduction

With rapid advancement in technology, the healthcare industry has been producing and collecting data at a staggering speed. However, raw data is barely of direct interest to healthcare stakeholders unless potentially useful knowledge is extracted. The advancement of machine learning facilitates the generation of data-driven models to: improve the understanding of disease mechanisms, increase the efficiency in healthcare delivery, reduce overall cost to the healthcare systems and facilitate clinical decision support [1]. Whilst rapidly gaining recognition in the value of data analytics for healthcare, impediments to further adoption also remain, which relate to the black box nature of many machine learning algorithms. As healthcare applications especially in critical use cases usually come with high stakes, interpretable models are necessary to allow the end users to: interrogate, understand, debug and perhaps, improve the underlying machine learning systems employed [2, 3].

Clinical decision making, such as predicting a patient's likelihood of readmission to the hospital, can have an immediate effect on the well-being of the public. Healthcare presents unique challenges for the deployment of machine learning models where the demands for interpretability and performance in general are much higher as compared to most other domains [2]. Given that the cost of misclassification is potentially high, models that are able to express the

inner philological associations in a human-readable way are widely sought in an effort to facilitate the interrogation and validation of learned knowledge. This would significantly help clinicians making informed decisions in combination with medical domain knowledge. Despite new models that exhibit high performance as well as interpretability have been proposed recently, the utility of these models in healthcare has not been convincingly demonstrated [2].

To avoid putting patients at risk, it is crucial that models trained on healthcare data be validated prior to deployment, as the pattern reflected from the training data may not necessarily be representative of the true inner workings of a certain medical condition. Rule-based systems are generally considered interpretable in the sense that the associated if-then statements are able to explicitly set out the conclusion under the given condition. In particular, fuzzy rule-based systems are of a natural appeal to the medical sectors. This is because they support the performance of approximate reasoning, through fuzzy logic, to track how a conclusion is reached, gaining insights into a potentially complex problem and therefore, facilitating the explanation of their solutions [4, 5, 6]. Built on top of fuzzy sets that permit gradual assessment of the membership of set elements, fuzzy systems also enable the tolerance of vagueness or imprecision that may result from linguistic descriptions such as 'sever pain' or 'feel uncomfortable' while enquiring medical symptoms or noise that may result from inaccurate testing results. Having recognised the potential of fuzzy techniques to cope with the challenges raising from healthcare, a number of accurate and interpretable fuzzy systems have been proposed in the literature, for various medical applications (e.g., [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]).

Following such promising research, and working towards providing assistance for clinical decision support, this paper proposes a neuro-fuzzy approach for the acquisition of accurate and interpretable fuzzy rule bases. It works by first discretising each of the continuous medical attributes into a certain number of categorical ones. In so doing, the original continuous data are mapped onto a new data set with only nominal values, enabling rapid generation of a set of crisp rules through the exploitation of advanced decision tree learning. The result-

ing crisp rules are able to reveal the basic relationship between attribute-value pairs, whereas the attribute-value pairs which do not appear in the rules can be removed. The generated crisp rules are then transformed to corresponding fuzzy rules with categorical values replaced by Gaussian membership functions. Finally, the set of fuzzy rules are adapted in the framework of adaptive network-based fuzzy inference system (ANFIS) [18] by the use of gradient descent and least square estimation. This leads to the acquisition of an optimal set of accurate fuzzy rules. An illustrative example is provided to explain the working mechanism of the proposed approach, whilst further systematical experiments demonstrate its superior performance over alternative fuzzy classifiers statistically.

The remainder of this paper is organised as follows. Section 2 introduces the related work on the classification of popular medical problems using fuzzy systems. Section 3 describes the proposed methodology. Section 4 presents and discusses the experimental study. Section 5 concludes the paper and outlines ideas for further development.

2. Related Work

One of the most well-known examples in the healthcare sector that favour approaches with interpretability over pure performance is the adoption of rule-based methods over neural networks, for the task of predicting pneumonia patients as high or low risk for in-hospital mortality [3, 19]. Although a neural network model may show better performance over a rule-based model, investigation into the generated rule set shows one pattern that patients with pneumonia as well as a history of asthma have lower risk of dying from pneumonia than those without asthma. What is learned by the neural net may reflect the true pattern of the training examples, but the extracted knowledge is counter-intuitive. This is due to the fact that patients with asthma history who are suffering from pneumonia are usually directly admitted to ICU, thereby reducing their risk of dying from pneumonia compared to those without asthma. However, models

trained on the prognosis data simply overlook the intermediate processes and incorrectly generalized the data into invalid knowledge, which is difficult for black box models like neural nets to recognize and rectify, thus putting patients at greater risk if practically used.

Computationally speaking, knowledge discovery and learning is supposed to be done with data available at hand. This may not be sufficient sometimes to reveal the true patterns of the underlying medical situations. It is necessary to validate the resulting learned model using medical expertise prior to being put into use for clinical decision making. This in turn demonstrates the necessity of adopting interpretable systems in healthcare decision support, which makes it much easier for clinicians to judge and rectify suspicious knowledge based on domain expertise or even commonsense. Fuzzy rule-based systems allow end-users to interrogate how a conclusion is reached via if-then fuzzy statements. The use of fuzzy sets supported with fuzzy logic makes fuzzy systems more robust in dealing with vague concepts that are commonly used in the linguistic description of symptoms as well as noise that may result from inaccurate testing. It is therefore, not surprising that there exist many fuzzy systems for healthcare in general and for learning fuzzy medical knowledge in particular, in the literature. The remaining of this section briefly reviews related work on the use of fuzzy systems for popular medical applications that are also utilised to perform comparative experimental analysis later.

Diabetes is a serious disease where the blood glucose level is too high and has drawn increasing attention for its worldwide prevalence. The Pima Indian diabetes [20] is a popular open access data set facilitating model building for predicting whether or not a patient has type-2 diabetes based on eight diagnostic measurements. A number of fuzzy rule-based systems have been developed using this dataset, for diabetes decision support [7, 8, 9, 10]. For example, an approach using modified Gini index based fuzzy supervised learning in Quest (SLIQ) decision tree algorithm, in conjunction with principal component analysis, is presented in [11], which outperforms a number of earlier models. A fuzzy ontology-based semantic case-based reasoning system [12] is proposed

and implemented for decision support on diabetes diagnosis, constructed on the basis of a standard medical terminology subset for diabetes diagnosis from systematized nomenclature of medicine-clinical terms. A combination of fuzzy k-nearest neighbour method and artificial immune recognition system is proposed in [21], achieving improved performance while gaining a good tradeoff between classification accuracy and system complexity.

The popular Parkinson's disease (PD) data set [22] is composed of various biomedical voice measurements extracted from voice recordings, with a view to discriminating healthy people from those with PD. In the literature, an effective and efficient system using fuzzy k-nearest neighbour [23] has been proposed for diagnosing Parkinson's disease, which achieves excellent performance, outperforming support vector machine (SVM) based approaches and many others. Another study has looked into the application of fuzzy c-means clustering-based feature weighting for the detection of Parkinson's disease [17], demonstrating that the combination of the proposed weighting method and k-nearest neighbour classifier can lead to very promising results on the classification of PD. Also, a hybrid intelligent system is proposed [16], where principal component analysis and expectation maximization are respectively used to address the multi-collinearity problems and data clustering, followed by the prediction of PD progression using a neuro-fuzzy system or an SVM.

The mammographic mass data set [22] is formed to identify the severity of a mammographic mass lesion based on the patient age and standard attributes from BI-RADS [24]. In general, it is highly recommended in decision-making to adopt a natural language structure in knowledge representation in order to aid in diagnosis for radiologists and physicians. For instance, a new knowledge-based system [13] is developed which integrates clustering, noise removal, and fuzzy rule-based techniques, achieving high prediction accuracy. A method for fuzzy characterization of the main linguistic terminological descriptors in the evaluation of breast nodules and calcifications has been reported [14]. Also, an expert system for the diagnosis of breast cancer [25] has been developed using a neuro-fuzzy mechanism, which prevents unnecessary biopsy and may

be adapted to train relevant medical students given its explanatory power. A fuzzy Gaussian mixture model [15] has been put forward that combines Gaussian mixture model and fuzzy rule-based systems to classify detected regions in mammogram images (into malignant or benign categories). This helps improve the diagnostic accuracy and reliability of radiologists while performing image interpretation for breast cancer diagnosis.

Regarding breast cancer diagnosis, there are indeed many approaches proposed in the literature. Here, two more relevant data sets are used: i) the Breast Cancer Wisconsin (Original) data set which is useful to systems aimed at examining patients who have undergone surgery for breast cancer, and ii) the Breast Cancer Wisconsin (Diagnostic) data set which extracts features from a digitized image regarding the fine needle aspirate of a certain breast mass. Based on the application of Naive Bayes approximation, a fuzzy system has been established [26] for classification of breast cancer patients with optimal interpretability without significantly losing the performance as compared to that of the state-of-the-art methods. A fuzzy quantification subsethood-based algorithm [27] has been extended to develop a novel class assignment procedure for breast cancer diagnosis, showing how fuzzy quantifiers may be utilized in a subsethood based algorithm to strengthen both classification accuracy and interpretability [28]. Furthermore, a technique for fuzzy rule-based non-linear transformation has been introduced to reinforce classification related information from given breast cancer data, thereby improving on the classification performance [29].

Last but not least, clinical data acquired in the areas of appendicitis, blood transfusion, thyroid gland and vertebral column are also popular in medical applications of intelligent systems. These are also to be addressed in this research as with the other problems outlined above. In addition to dedicated approaches for addressing specific medical situations, a number of fuzzy systems [30, 31, 32, 33, 5, 34] have been proposed for a range of different medical problems. For instance, a novel interpretable fuzzy rule-based classifier termed C45-IFRC is presented in [30] in order to seek out a trade-off between accuracy and complexity of the eventually induced fuzzy rule base. It works by

mapping each coarsely learned C4.5 decision tree rule in the knowledge base onto a set of potentially useful fuzzy rules, which is subsequently optimised by genetic algorithm. The generated fuzzy system is interpretable and accurate, which has been tested against various medical cases. A steady-state algorithm for extracting fuzzy classification rules from data (SGERD) is described in [33] that is able to extract a compact set of fuzzy rules by exploiting specific rule and data dependent parameters, resulting in short, accurate and interpretable fuzzy rules that fit clinical decision support. Fuzzy pattern trees (FPT) have also been introduced as a novel type of fuzzy system which has shown superior performances on a wide range of applications including various medical problems. Details of this approach is beyond the scope of this paper, but can be found in [31, 35].

3. Proposed Methodology

The key to accomplish the task of learning a fuzzy rule-based system for clinical decision support is to find a finite set of fuzzy production or if-then rules capable of classifying a given input. Without losing generality, the classification system to be modelled is herein assumed to be multiple-input-single-output, receiving n-dimensional input patterns and producing one output which is determined to be one of the pre-specified M classes. The fuzzy rule set to be induced is required to perform the mapping $\varphi : X^n \rightarrow Y$, where $X^n = X_1 \times X_2 \times \cdots \times X_n$ with $X_1, X_2, ..., X_n$ being the domains of discourse of the input variables, and Y represents the set of possible output classes of a cardinality of M. Following the general supervised learning approach, the behaviour of the classification system is trained through the use of a set of input-output example pairs E, where for each instantiation of the input variables $\bar{x}^p = (x_1^p, x_2^p, ..., x_n^p)^T, x_i^p \in X_i, i = 1, 2, ..., n$, an associated class $y^p \in Y$ is indicated.

Owing to its capability to approximate nonlinear functions to any degree of accuracy in any convex compact problem domain, while being of a fair computational efficiency [36], knowledge or rule base consisting of Takagi-Sugeno-Kang

(TSK) fuzzy if-then rules is adopted in this paper. In general, a TSK fuzzy if-then rule F_j can be represented as follows:

$$\text{If } x_1 \text{ is } A_{j1} \text{ and } ... \text{ and } x_n \text{ is } A_{jn}, \text{Then } z = f(x_1, \ldots, x_n) \tag{1}$$

where $j = 1, 2, .., N$, with N denoting the number of all such fuzzy rules within the system; $x_i, i = 1, \ldots, n$ are the underlying domain variables, jointly defining the n-dimensional pattern space and respectively taking values from X_i; $A_{ji} \in X_i$ denotes a fuzzy set that the variable x_i may take; and z_j is the consequent of the fuzzy rule describing the output of the model, which is a polynomial function of the input variables under TSK rule structure. For classification, an output computed from such a polynomial function is mapped onto one of the class indices specified in the class set Y as previously defined.

The proposed approach works by first discretising each of the continuous domain attributes into a certain number of categorical ones, such that the original continuously valued data is mapped onto a new one involving only symbolic value (with their underlying real meaning unaltered of course). The generated crisp rules are able to reveal the basic relationship between attribute-value pairs and are then transformed to corresponding fuzzy rules with categorical values replaced by Gaussian membership functions. Finally, the set of fuzzy rules is adapted by the use of an ANFIS neuro-fuzzy system satisfying gradient descent and least square estimation. This is in order to acquire an optimal set of accurate fuzzy rules. The following subsections describe the details of the proposed approach.

3.1. Data Discretisation

The first step of the proposed approach is initialised by discretising each of the available continuous variables into a corresponding categorical one. This may be implemented by resorting to physicians domain expertise or using a computational algorithm such as [37]. For easy implementation, this paper adopts the simple technique that partitions the universe of discourse of an individual attribute into a certain number of equal intervals (assuming that the attributes are

uniformly distributed). Each interval length is set to: $intl = \frac{max(X_i)-min(X_i)}{L}$ where X_i is the domain of attribute x_i with $max(X_i)$ and $min(X_i)$ being its maximum and minimal value, respectively; and L is the user-defined number of partitions. Any original variable (in terms of its value) x_i is mapped onto the integer k, if $intl_i^k \leqslant x_i < intl_i^{k+1}, k \in [1, \ldots, L+1]$, where $intl_i^k$ is the k-th interval point for attribute x_i. In so doing, the original continuous attributes are transformed into ordinal integers. For discrete and nominal variables (e.g., gender that only takes male or female as its value), they remain unchanged.

Such a heuristic partitioning is unlikely to be optimal, but the discretisation will be optimised later with an adaptive method. However, this simple implementation comes with two advantages: First, each interval can be readily assigned with a linguistic label that is of interpretable meaning, instead of taking a pure numerical number that hardly makes any sense especially to non-experts in the domain. In particular, the readable annotations make more direct link with practical situations where for example, patients describe their symptoms or clinicians verbally explain the severity of a medical condition. Second, from computational perspective, the small number of categorical values help expedite the construction of a decision tree for the acquisition of an initial crisp rule base as to be introduced next.

Note that when applying the proposed approach in a real clinical setting, the fuzzy quantities and their linguistic fuzzy labels will be defined, and interpreted, in consultation with the medical professionals. In so doing, the specification of variable discretisation will reflect the domain expertise, thereby describing variables in the problem domain directly using terms which have predefined semantic meanings.

3.2. Crisp Rule Generation with Decision Tree Learning

Once the discretisation of the original data set has been carried out, a set of crisp rules can be generated using a decision tree learning mechanism such as the Classification and Regression Tree (CART) algorithm [38]. The basic working of this learning method starts with the full data set at the root node

and iteratively applies the following Gini index to split the node:

$$Gini(S) = \sum_{i=1}^{M} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{M} p_i(1 - p_i) = \sum_{i=1}^{M} p_i - \sum_{i=1}^{M} p_i^2 = 1 - \sum_{i=1}^{M} p_i^2 \tag{2}$$

where S denotes the current data set for which this index is calculated; M is the number of class labels; $p_i, i \in \{1, \ldots, M\}$ is the probability of an object with the label i being randomly chosen; and $\sum_{k \neq i} p_k = 1 - p_i$ represents the probability of a mistake in categorising an object. It can be seen that the Gini index reaches its minimum when all cases fall into a single category, and maximum when all items are equally distributed among all classes. As such, this index can be used to capture the amount of uncertainty in a dataset, measuring how often a randomly chosen object from the dataset may be incorrectly labelled, if it is randomly labelled according to the distribution of all the labels in the data.

Note that the inputs used to construct the decision tree are transformed categorical values during the training stage. Once the training is completed and a new instance is present in a query for a decision, its original crisp input value will first be converted to the category that represents the corresponding interval, which can then be used to match against existing rules. As such, this supports the learning of rules that involve intervals as variable values, instead of crisp cut points. In running the model, when an input is present, the values of individual variables are each a crisp value. These crisp values are checked against the interval values of their corresponding variables within every rule in the rule base to decide whether that rule is to be fired or not. Obviously, if each crisp value falls within the corresponding interval, the rule is activated, else it is not.

At each split, a decision tree node is generated with the attribute for which the resulting Gini index is minimum. The same procedure is then iterated on each of its subsets using the remaining attributes. When there are no more attributes to be selected for further split or every element in the subset belongs to the same class, a complete decision tree is generated, which can be easily

transformed into a set of crisp rules by retrieving paths from each leaf node backwards through its parent recursively to the root node. Without losing generality, denote a generated crisp rule $C_j, j = 1, 2, ..., N$ (with N representing the number of all crisp rules available) as follows:

$$\text{If } x_1 \text{ is } I_{j1} \text{ and } ... \text{ and } x_n \text{ is } I_{jn}, \text{ Then class is } y^{C_j} \tag{3}$$

where $x_1, x_2, ..., x_n$ represent the underlying domain attributes; $I_{ji}, i \in \{1, 2, ..., n\}$, is the crisp interval of the antecedent attribute x_i that may be associated or assigned with a meaningful label for linguistic interpretation; and y^{C_j} is a class label, expressing the rule consequent.

3.3. Conversion of Crisp Rules into Fuzzy Rules

The above data-driven set of crisp rules can then be converted into a set of corresponding fuzzy rules prior to further optimisation. From the viewpoint of rule structures, a rule is made up of an antecedent and a consequent part, be it fuzzy or crisp. Both the fuzzy and crisp rule antecedents are of a conditional statement form, describing the values that the antecedent attributes should take in order to derive the corresponding consequent, which are connected by logical operators. The only difference is that crisp intervals (that are each interpreted with a symbolic or integer term) are utilised as the conditions for the corresponding attributes, whereas attributes in a fuzzy rule are depicted by fuzzy sets (that are normally interpreted with a linguistic label). Therefore, a straightforward approach is to replace each crisp interval with a fuzzy membership function.

In general, the specific membership function used to describe a fuzzy set should be more carefully considered in relation to underlying knowledge for expression. In this paper, in the absence of such prior knowledge, a crisp interval I_i, $i \in \{1, 2, ..., n\}$ as shown in Eqn. 3 is replaced with a Gaussian membership function $\mu_{A_i}(x_i) = e^{-(\frac{x-c_i}{\sigma_i})^2}$ due to its popularity, where c_i and σ_i are the mean value and standard deviation of the Gaussian membership function. Here, the mean value is set to the average of those values belonging to the corresponding

crisp interval I_i, such that

$$c_i = \frac{\sum_{x_i \in I_i} x_i}{|\{x_i \in I_i\}|} \tag{4}$$

Similarly, the standard deviation is computed by

$$\sigma_i = \sqrt{\frac{\sum_{x_i \in I_i} (x_i - c_i)^2}{|\{x_i \in I_i\}|}} \tag{5}$$

Once the process of replacing crisp intervals with the above Gaussian membership functions is complete, the transformation of the entire crisp rule antecedent finishes with the logical conjunctive connector 'AND' in the original crisp rules replaced with a T-norm operator that performs fuzzy conjunction, implemented by the product operation in ANFIS (or typically by minimum in Mamdani models). The consequent of a crisp rule with a decision class is then directly mapped onto that of the corresponding fuzzy rule. Although a TSK fuzzy rule could take higher orders, a zero order polynomial TSK rule is adopted in this paper, given the application problem is to perform classification. That is, the integer that represents the decision class in the crisp rule is taken as the bias term in the fuzzy rule. The resulting mapped fuzzy rule from an original crisp rule can thus, be generally represented as

$$\text{If } x_1 \text{ is } e^{-(\frac{x-c_1}{\sigma_1})^2} \text{ and ... and } x_n \text{ is } e^{-(\frac{x-c_n}{\sigma_n})^2}, \text{Then } z = r \tag{6}$$

where $e^{-(\frac{x-c_i}{\sigma_i})^2}$ is the fuzzy membership function for attribute $x_i, i = 1, \ldots, n$ with c_i and σ_i calculated as above, and r is the integer that represents the decision class of the corresponding crisp rule.

Note that running the conventional method of grid partitioning [39] of each and every input space may suffer from the curse of dimensionality as the number of inputs increases. Therefore, instead of considering all of the possible combinations of the input and class attributes, it is herein proposed to utilise the existing crisp rules, which have been generated by decision tree learning and which are able to efficiently and sufficiently generalise the given data to guide the transformation, without resorting to pure and brute force search. Being fundamentally data-driven, such a rule generation method will omit the empty

parts of the input space, substantially expediting the subsequent optimisation process.

Note also that, whilst in well-experienced domains certain numerical intervals or even single numeric numbers have been widely adopted, there are many scenarios where domains are less understood or where variables are difficult to interpret. For instance, the mood of an individual may be asked when diagnosing an individual's mental well-being that may include options such as very unhappy, unhappy, ok, happy, very happy. In case where the examinee feels difficult to give a precise answer, especially when they feel in-between neighbouring options such as happy and very happy, the use of fuzzy sets supports capturing both concepts, though to different degrees, while reflecting such uncertainty. This helps enhance the tolerance level of capturing and representing linguistic imprecision that often arises from clinical data.

Of course, the interpretability does not just lie in how to label the fuzzy intervals, which conventional discretisation may also achieve, but also in enabling any subsequent pattern matching to be performed partially. This helps to allow for the aggregation of possible conclusions from firing multiple rules as opposed to just one single specific rule as in conventional rule-firing situations, thereby reducing the adverse effect of any bias towards a certain value definition/discretisation.

3.4. Optimisation of Transformed Fuzzy Rules with ANFIS

ANFIS [18] is a popular TSK fuzzy inference system built under the generic framework of artificial neural networks, capturing the benefits of both neural networks and fuzzy logic. Once the acquisition of a set of fuzzy rules has been achieved (through converting a set of crisp rules) they can be utilised to initialise the ANFIS structure to enable further modification or optimisation, by exploiting the inherent adaptive mechanism of the network given any further training data.

To simplify the illustration of the optimisation process, suppose that there are only two crisp rules learnt by CART for a two-input and one-output problem.

The two converted zero-order TSK fuzzy rules are presented as follows:

$$\begin{aligned} &\text{Rule 1: If } x \text{ is } A_1 \text{ and } y \text{ is } B_1, \text{Then } z_1 = r_1 \\ &\text{Rule 2: If } x \text{ is } A_2 \text{ and } y \text{ is } B_2, \text{Then } z_2 = r_2 \end{aligned} \tag{7}$$

The structure of the neuro-fuzzy ANFIS that is equivalent to this flat fuzzy TSK rule base is shown in Fig. 1, where square nodes stand for the network nodes which contain parameters that can be adapted, and circle nodes represent fixed ones without modifiable parameters. For completeness, details of individual layers within the ANFIS are briefly summarised below.

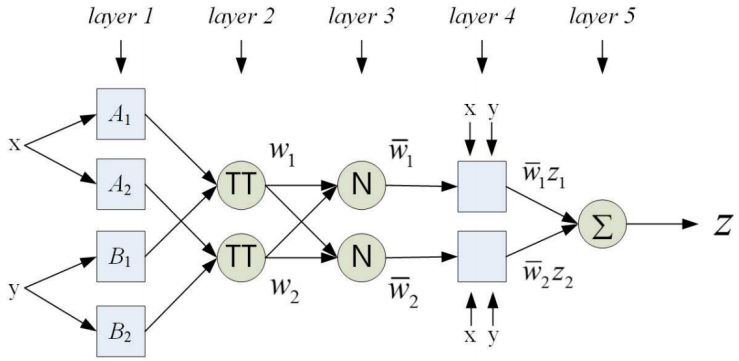

Figure 1: Illustrative ANFIS structure

Layer 1: Every node i in this layer is a square node with the following function:

$$O_i^1 = \mu_{A_i}(x) \tag{8}$$

where x denotes an input variable to this node, and A_i denotes a fuzzy set that may be taken by the variable, which is defined by a Gaussian membership function as previously stated:

$$\mu_{A_i}(x) = e^{-(\frac{x-c_i}{\sigma_i})^2} \tag{9}$$

In particular, c_i and σ_i are the parameters associated with the corresponding variable, representing the mean value and standard deviation of the Gaussian membership function, which are initialised as Eqn. (4) and Eqn. (5) respectively. These parameters are named premise parameters and are to be tuned in

subsequent layers. Note that other continuous and piecewise differentiable functions, such as trapezoid or triangular functions may also be utilised if desired.

Layer 2: Every node in this layer is a circle node which accumulates the incoming values through multiplication and outputs the product. The output w_i in this layer acts as the firing strength of a certain rule, namely, Rule $i, i = 1, 2$ for the present example. That is,

$$w_i = \mu_{A_i}(x) \times \mu_{B_i}(y) \tag{10}$$

Layer 3: Each node in this layer is a circle node, computing the ratio of the ith rule's firing strength to the sum of all rules' firing strengths:

$$\overline{w_i} = \frac{w_i}{\sum_{j=1}^{N} w_i} \tag{11}$$

where again, $i = 1, 2$ and the number of rules $N = 2$, for this particular example. The outputs of this layer are normalised firing strengths from the preceding layer.

Layer 4: Each node i in this layer is a square node with the following function:

$$O_i^4 = \overline{w_i} z_i = \overline{w_i}(r_i) \tag{12}$$

where $\overline{w_i}$ is the output of layer 3, and r_i is the parameter to be adjusted which appears in the rule consequent and is therefore, referred to as the consequent parameter. Note that if higher order TSK rules are used, more degree of freedoms are imposed on the underlying system, resulting in more consequent parameters to be tuned. For example, the consequent $z_i = p_i x + q_i y + r_i$ if first order rule applied, then p_i, q_i are also adjustable parameters. For classification problems, zero-order is sufficient while being computationally simplest.

Layer 5: The single node in this layer, the output layer, is a circle node that computes the overall output in response to all current inputs, defined as the summation of all incoming values, i.e.,

$$O_1^5 = \sum_i \overline{w_i} z_i = z \tag{13}$$

Within an ANFIS, the parameters, including both premise and consequent ones, are trained using a hybrid learning method combining gradient descent and least square estimation. Particularly, each epoch of the hybrid learning procedure is composed of a forward pass and a backward pass. In running the forward pass, the antecedent parameters are fixed and a vector of input values is presented, and then the error between the actual output and the target output is calculated. In the backward pass, the error computed at the last forward pass is propagated backwards, from the output end towards the input end while fixing the consequent parameters, by the gradient method. The details of such an iteration of forward and backward computation is beyond the scope of this paper, but can be found in [18].

3.5. Novel Points of Proposed Approach

A variation of the CART + ANFIS combination exists in the literature [41]. However, the present work differs from that approach significantly, from a number of viewpoints, as detailed below.

First, a sigmoid function was used in [41] that directly fuzzifies the cut-point generated by CART to a fuzzy number. This may complicate the understanding of the overall process of fuzzy value specification, as such a number can only be explained locally because the same variable may end up with having different fuzzy quantity spaces when it appears in the antecedents of different rules. Discretising a variable domain into a fixed set of categories, which are then converted into fuzzy sets, simplifies the process of associating individual labels to the discretised values from a holistic viewpoint.

Second, the integer in a zero-order TSK fuzzy rule consequent may be interpreted as a certain decision (e.g., a class for a classification task) and the corresponding decimal figures as the rule confidence (e.g., a certainty factor for the class in performing classification). This facilitates the explanation of each resulting individual fuzzy rule. However, a first-order TSK rule as adopted in [41] does not offer such interpretability regarding the rule consequent, thus potentially damaging the readability of the overall fuzzy system. Also, as to be

justified in Section 4.4, which was not considered at all in [41], the use of additional parameters in a first-order TSK rule significantly expands the hypothesis space, adding further run-time costs, and may cause serious overfitting of the learnt model.

Last but not least, the method of [41] is devised specifically for fault diagnosis of induction motors, aiming at providing accurate decisions without due consideration of interpretability. Yet, the proposed approach aims to learn an interpretable fuzzy system with application to a range of popular medical diagnostic problems. The underlying motivation for the present work rests in the development of an interpretable fuzzy system for clinic decision support. Indeed, the employment of the specific discretization mechanism and zero order TSK structure reflects such design intentions.

4. Experimentation

This section presents an experimental analysis of the proposed approach, supported with comparative studies with respect to popular techniques selected from the existing literature.

4.1. Experimental Setup

To demonstrate the efficacy of the present work for clinical decision support, experiments are performed on nine medical benchmark data sets taken from UCI machine learning repository [22]. A summary of the characteristics of these data sets is given in Table 1. As the range of different attributes vary significantly, a preprocessing step is to normalise each attribute so that their normalised values fall within the range of $[0, 1]$. This facilitates better comparisons. Data normalisation is a common approach in machine learning, though this may affect the model's interpretability. Fortunately, this issue can be addressed in a straightforward manner, by mapping any derived (normalised) fuzzy sets back onto their original domains once the training is completed.

In the absence of testing data for the performance evaluation of the proposed approach, stratified tenfold cross-validation (10-CV) is employed for result validation. In 10-CV, a given data set is partitioned into ten subsets. Of the ten, nine subsets are used to perform training, where the proposed approach is used to generate a fuzzy rule base, and the remaining single subset is retained as the testing data for assessing the learned classifier's performance. This cross-validation process is then randomly repeated ten times in order to lessen the impact of any random factors; results of these 10×10 cross-validations are then averaged to produce each final experimental outcome reported below (except for the illustrative showcase as given in Section 4.2).

Table 1: Summary of Data Sets Used

Data Set	**Abbreviation**	**#Attribute**	**#Instance**	**#Class**
Appendicitis	APN	7	106	2
Blood Transfusion	BLD	4	748	2
Mammographic Mass	MM	5	961	2
Parkinson's Diseases	PD	22	195	2
Pima Indians Diabetes	PID	8	768	2
Thyroid Gland	TG	5	215	3
Vertebral Column	VC	6	310	2
Wisconsin (Diagnostic) Breast Cancer	WDBC	30	569	2
Wisconsin (Original) Breast Cancer	WOBC	9	699	2

In an effort to examine the effect of domain discretisation, or the number of discretised intervals upon the resulting fuzzy rules, seven different bin numbers are tested, where each of the pattern spaces is divided into K ($K = 3, 4, 5, 6, 7, 8$) equal intervals following the ideas as discussed in Section 3.1. This allows the performance of the proposed method to be investigated for fine partitions (such as when $K = 8$) as well as for coarse partitions (when $K = 3$). Note that given a K, in theory, the total number of rule antecedent combination would be K^n, where n stands for the number of input attributes. However, a fuzzy rule is produced only when there is a corresponding crisp rule generated by a certain data-driven crisp rule-based learning mechanism, with each crisp rule created to cover at least one given training data.

4.2. Illustrative Example

To demonstrate the proposed approach at work for effectively aiding in clinical diagnosis, an illustrative example is performed on the popular Pima Indian diabetes data set first.

The decision tree algorithm utilised in this paper is the popular Classification and Regression Trees (CART), which is characterised by its construction of a binary tree with each internal node having exactly two outgoing branches. In particular, each original continuous attribute is herein discretised into a categorical one with 3 equally spaced bins. As an example, the resulting decision tree structure, which is taken from a single fold out of the complete 10-CV, is shown in Fig. 2, where '0' and '1' in the leaf nodes stand for negative and positive test respectively; and 'low', 'medium' and 'high' are the labels used to denote the corresponding discretised crisp intervals. Note that certain attributes may take more than one interval as its value (e.g., Glucose can take either low or medium in its left branch), which is attributed to the mechanism that CART grows the trees. However, this can still be transformed directly into a rule base with each attribute only taking a single value as follows:

- Rule 1: If Glucose is *low*, Then test *negative*;
- Rule 2: If Glucose is *medium*, Then test *negative*;
- Rule 3: If Glucose is *high* and BMI is *low*, Then test *negative*;
- Rule 4: If Glucose is *high* and BMI is *medium*, Then test *positive*;
- Rule 5: If Glucose is *high* and BMI is *high*, Then test *positive*.

In this example, as a side effect, the CART algorithm may also be interpreted as a feature selection technique, resulting in the use of two attributes only out of the original eight. With only five crisp rules generated, this significantly reduces the problem space, which could have been as many as $3^8 \times 2$ if the conventional grid partitioning were used for rule generation [39]. The above crisp rule base is then transformed into a fuzzy one with crisp intervals replaced by Gaussian

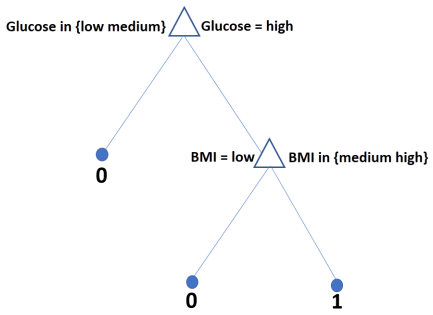

Figure 2: Generated decision tree

membership functions as specified in Section 3.3. The fuzzy rule base and the fuzzy inference process are illustrated in Fig. 3, and the learned decision tree shown in Fig. 2. Each row in Fig. 3 represents a fuzzy rule converted from the corresponding crisp rule generated by CART. The rectangles in each column sets out the learnt fuzzy membership functions that specify the values taken by each linguistic variable concerned. Considering only two features utilised in this example, only two columns of rectangles are used in the figure to form rule antecedents. Note that for the attribute *BMI*, no value is present in column two on the first two rows, indicating that no involvement of the second variable in the corresponding rule antecedents. The transformed rule base serves as the input to the neuro-fuzzy ANFIS structure for optimisation, as shown in Fig. 4. ANFIS then fine-tunes both the antecedent and consequent parameters based on the existing rule base.

To demonstrate how such fuzzy rule base may be utilised to aid clinical decision making, consider an incoming patient with the following testing values: $(\#Pregnance = 2, Glucose = 120, BloodPressure = 61, SkinThickness = 50.1, Insulin = 423, BMI = 36.2, DiabetesPedigree = 1.25, Age = 51)$. The trained fuzzy system performs approximate reasoning while helping clinicians derive a diagnosis. The first step is to take the crisp input and determine the degree to which they belong to each of the appropriate fuzzy sets by matching it against the membership functions of the respective variables. As shown in

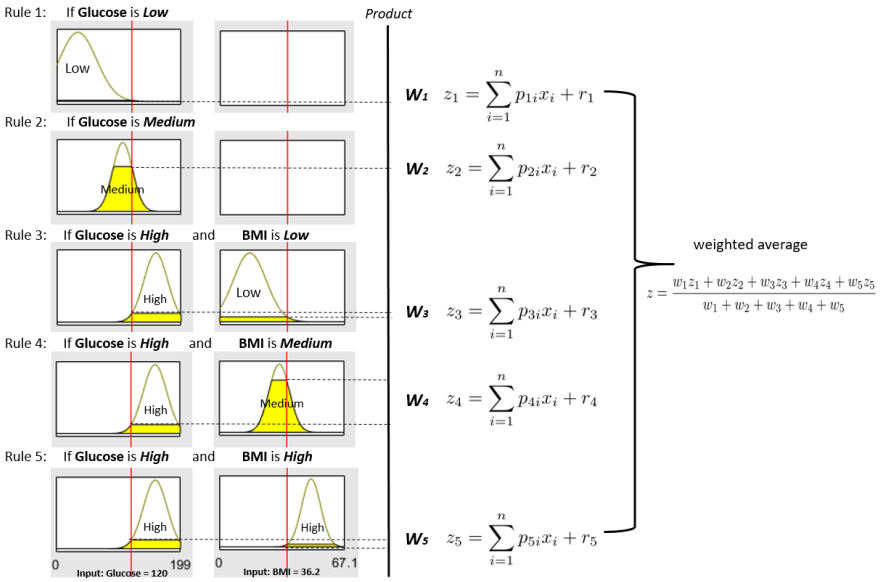

Figure 3: Fuzzy rule base and inference process

Fig. 3, consider the variable $Glucose = 120$, it intersects with two fuzzy sets (*Medium* and *High*, but not *Low*) and the yellow-coloured patches under these Gaussian membership function curves reflect the two respective matching degrees. Similarly, for $BMI = 36.2$, it intersects with all three fuzzy sets. Note again that all membership functions can be annotated with linguistic labels such as *Medium* and *High* as per this example, aiding clinicians (who are not necessarily experts in fuzzy systems) in the interpretation of extracted knowledge and the patients in the understanding of their medical situations concerned.

Having obtained the matching degree to which each antecedent variable is satisfied with regard to any given rule, the fuzzy operator such as the product operation (as it is used in this paper) is applied to compute the overall firing degree of all antecedent variables within the rule. This overall matching degree is commonly referred to as the rule's activation strength in response to the given testing input, which is subsequently normalised to compute the ratio (or relative contribution) of this rule's firing strength over the sum of all rules' firing strengths. The output of the entire fuzzy inference process is the average of the

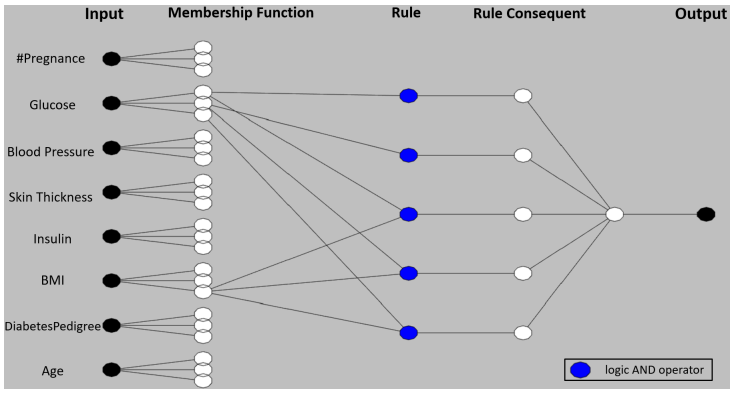

Figure 4: ANFIS structure

outcomes produced by the individual matched rules respectively weighted by their corresponding normalised firing degrees. As the consequent of each rule is calculated to be the zero-order polynomial of the input values, the final output for the present example is 0.128. This may be rounded to integer 0 to signify a negative diagnostic outcome, with a confidence level of (1 - 0.128/1 = 0.872).

Note that given the normalised firing strength of each rule, the overall diagnostic outcome may be interpreted as the total accumulated contribution made by individual rules that match at least partially with the patient's symptoms. As such, such results can help clinicians decompose their overall decision into component sub-decisions for further analysis. Another strength of such fuzzy decision support systems is their ability to perform approximate reasoning, simultaneously firing multiple rules that imprecisely match given symptoms. This may help mitigate the sensitivity of a crisp rule-based system to noisy outliers that often arise from medical data.

4.3. Performance vs. Number of Discretised Intervals

Table 2 presents the performance variations of the proposed approach in relation to the number of partitioned intervals of the feature space. As can be seen (and can be expected), for each individual data set, the performance may be affected significantly by the bin number used. For example, in the case of

Table 2: Performance based on 10*10-CV in response to variation of K

Data Set	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8
APN	82.78 ± 1.70	**85.67 ± 2.06**	79.19 ± 2.86	81.48 ± 2.22	75.55 ± 1.81	73.26 ± 3.45
BLD	76.82 ± 0.60	**79.09 ± 0.46**	76.83 ± 0.27	76.79 ± 0.63	76.38 ± 0.45	77.07 ± 0.96
MM	79.88 ± 0.48	**80.16 ± 0.38**	79.30 ± 0.67	78.91 ± 0.69	73.03 ± 2.44	78.62 ± 1.36
PD	**88.64 ± 1.67**	86.76 ± 1.27	87.29 ± 1.57	83.27 ± 2.43	82.55 ± 1.18	78.25 ± 1.97
PID	74.80 ± 0.95	**75.36 ± 0.89**	73.18 ± 0.71	72.44 ± 0.76	73.61 ± 1.04	72.34 ± 0.64
TG	91.49 ± 1.00	89.82 ± 1.03	**91.68 ± 0.75**	90.99 ± 0.80	91.09 ± 0.86	86.50 ± 1.43
VC	77.32 ± 1.43	**79.45 ± 1.34**	75.87 ± 1.70	74.81 ± 2.16	78.16 ± 1.65	77.16 ± 1.69
WDBC	93.70 ± 0.69	**93.74 ± 0.31**	93.67 ± 0.62	92.78 ± 0.87	90.93 ± 0.73	91.05 ± 0.99
WOBC	95.79 ± 0.36	**95.91 ± 0.36**	93.99 ± 0.46	93.65 ± 0.90	92.80 ± 0.49	92.58 ± 0.92
Average	84.580 ± 0.989	**85.107 ± 0.900**	83.444 ± 1.066	82.791 ± 1.273	81.567 ± 1.182	80.759 ± 1.490

APN, its performance falls in the range between %73.26 and %85.67. Generally speaking, the coarse partitions such as $K = 3$ or 4, outperform those with finer partitions such as $K = 7$ or 8. With the bin number accelerates from $K = 4$, the averaged performance gradually drops. This is likely attributed to overfitting that results from the use of a rule base that is of a much greater size. Fig. 5 shows that the averaged rule base size (in terms of the number of rules contained) increases along with the bin number K, across all data sets.

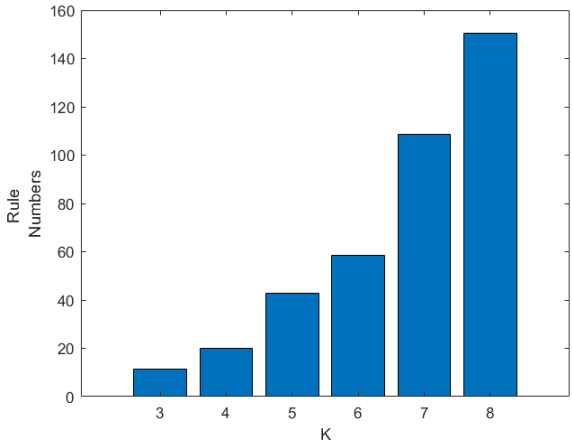

Figure 5: Rule number vs. bin number

The problem of overfitting can be reflected by plotting the average training and testing performances in variation to the bin number K. As shown in Fig. 6,

with the increment of the number of discretised intervals from $K = 4$ upwards, the training accuracy builds up, whereas the corresponding testing accuracy drops. Overall, when $K = 4$, the proposed method performs the best for 7 out of 9 data sets according to Table 2. It also achieves the best average performance as well as the smallest average standard deviation when $K = 4$. Although a finer partition may be necessary for more complicated medical problems, this result indicates that a relatively coarse partition with $K = 4$ works better for the underlying medical data sets. From interpretability viewpoint, use of a smaller bin number offers an easier interpretation while making the terms employed more distinguishable. Besides, an overly large bin number is not encouraging in practical clinical settings owing to psychological theory, nor in the efficiency of computational implementation.

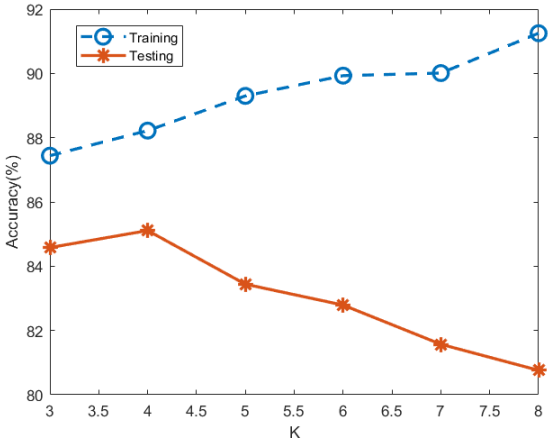

Figure 6: Rule numbers in variation to bin number

4.4. Linear vs. Zero-order Rule Consequent

In general, TSK fuzzy rules used in many ANFIS usually adopt the first order polynomial rule consequent [18]. Whilst it has a natural appeal to use zero-order for the classification tasks that are of direct interest regarding the present application problem, it is interesting to examine the effect of the rules

that do employ polynomial consequents. Without losing generality, a TSK fuzzy rule with first order polynomial consequent can be represented by

$$\begin{aligned} &\text{If } x_1 \text{ is } e^{-(\frac{x-c_1}{\sigma_1})^2} \text{ and } ... \text{ and } x_n \text{ is } e^{-(\frac{x-c_n}{\sigma_n})^2}, \\ &\text{Then } z = p_1 x_1 + \cdots + p_n x_n + r \end{aligned} \tag{14}$$

where the rule antecedent takes exactly the same structure as that in Eqn. 6, but the rule consequent is represented as a linear combination of the input values. This significantly expands the hypothesis space whereby each input is associated with an additional parameter to tune in the search space. Owing to limited space, the settings in which the number of feature partitions is 4 (which has been empirically shown earlier to have led to the best result) are used to run this particular study.

Table 3: Zzero order vs. first order fuzzy rules ($K = 4$)

	Zero Order Rule Consequent			**Linear Rule Consequent**		
Data Sets	**Trn Acc**	**Tst Acc**	**Time**	**Trn Acc**	**Tst Acc**	**Time**
APN	91.00	85.67	9.43	97.11	71.69	16.54
BLD	80.57	79.09	22.20	82.00	77.55	79.80
MM	81.98	80.16	58.07	85.39	78.63	438.83
PD	93.19	86.76	32.80	99.98	76.29	2138.95
PID	77.70	75.36	38.14	83.84	71.04	249.93
TG	93.05	89.82	9.91	97.59	87.40	25.19
VC	82.52	79.45	13.49	90.66	76.58	55.97
WDBC	96.49	93.74	113.32	99.98	76.50	18082.37
WOBC	97.50	95.91	38.27	99.27	90.03	435.73
Average	88.222	85.107	37.291	92.869	78.412	2391.479

Table 3 shows the results where the zero order and linear rule consequent are used respectively. Note the results displayed come from the same random runs of 10-CV, which implies both methods are initialised using the exact fuzzy rule bases converted from CART. It is not surprising that training accuracies of those with linear rule consequent are better across all clinical data sets than those with zero order rule consequent, owing to a larger hypothesis space being covered by exploiting more parameters. This of course comes with greater computational burden, resulting in over 60 times more efforts of that required by running the zero order counterpart. Such significant differences in time consumption may be

exaggerated even more for higher dimensional data sets such as WDBC (which has 30 attributes), consuming over 100 times more resources if linear consequent is applied.

Unfortunately, the more expressive representation of the hypothesis space and significantly more consumption of computation and run time efforts do not offer better performance for unseen testing data. The testing accuracies with linear consequent actually drop across all data sets, resulting in much worse overall performance (%78.412) in comparison to that (%85.107) achievable when the zero-order consequent representation is used. With the exact same initial fuzzy rule bases to tune, the plunging performance with linear consequent is likely attributed to overfitting that results from excessive degrees of freedom. Apart from the viewpoint of classifier generalisation capability, using ANFIS involving just zero order rule consequent allows class labels to be represented as integers, simplifying the coding while attaining model interpretability.

4.5. Comparison against Alternative Fuzzy Classifiers

Table 4: Comparison on classification accuracy (%)

Data Set	CART-NFC	C45-IFRC	FPT	SGERD
APN	85.67 ± 2.06	84.34 ± 2.63	**86.66 ± 0.89**	85.04 ± 1.01
BLD	**79.09 ± 0.46**	77.53 ± 0.48	77.42 ± 0.13	76.22 ± 0.18
MM	**80.16 ± 0.38**	79.13 ± 0.79	76.23 ± 0.44	77.39 ± 0.20
PD	**86.76 ± 1.27**	84.33 ± 1.11	85.03 ± 0.71	82.28 ± 1.53
PID	**75.36 ± 0.89**	75.05 ± 0.89	74.13 ± 0.36	70.17 ± 0.69
TG	89.82 ± 1.03	**91.88 ± 1.20**	88.75 ± 0.53	87.23 ± 0.58
VC	79.45 ± 1.34	**80.16 ± 2.16**	74.65 ± 1.48	70.00 ± 1.22
WDBC	93.74 ± 0.31	**94.30 ± 0.53**	93.25 ± 0.54	91.86 ± 0.67
WOBC	**95.91 ± 0.36**	95.15 ± 0.68	95.35 ± 0.23	93.49 ± 0.36
Averaged	**85.107 ± 0.900**	84.652 ± 1.163	83.497 ± 0.591	81.520 ± 0.716

To compare how the proposed CART initialised neuro-fuzzy classifier (CART-NFC) performs against state-of-the-art methods, the following three popular fuzzy classifiers that have been recently proposed and reviewed in Section 2 are also run: fuzzy pattern tree (FPT) [31, 35], C45-IFRC [30], and SGERD [33]. Table 4 presents the accuracies of these algorithms for the medical data sets. It

can be seen that CART-NFC achieves five out of nine best accuracies as well as the highest average accuracy.

In order to verify whether there is indeed any statistically significant difference among the algorithms (namely, CART-NFC, FPT, C45-IFRC and SGERD), non-parametric statistical tests are carried out using the Friedman Aligned Ranks test [40]. Fig. 7 shows the rankings produced by this test, where the bars are proportional to the average ranking obtained for each named algorithm. The lowest bar (implying the best algorithm statistically) achieved by the proposed algorithm agrees with the smallest error that is incurred by running it (see Table 4). To examine whether significant differences exist among the average errors, parameters associated with the outcomes of the Friedman Aligned Ranks test are given in Table 5, where the p value indicates the probability to reject the null hypothesis that there is no significant difference among the three average performances. At the significance level of $\alpha = 0.05$, the null hypothesis is rejected, indicating that there exists significant statistical differences amongst the results attainable by the members of this group of four fuzzy classifiers.

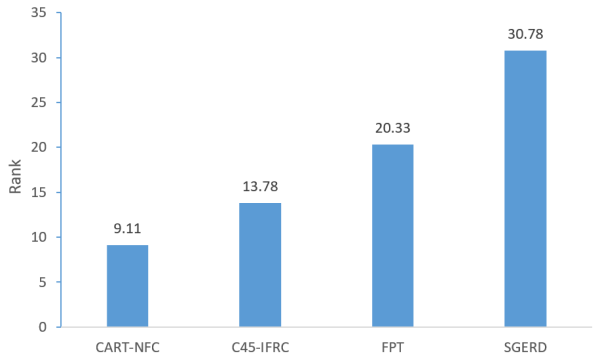

Figure 7: Rankings of CART-NFC, FPT, C45-IFRC and SGERD

Table 5: Friedman Aligned Ranks test

Comparison	**Hypothesis** ($\alpha = 0.05$)	p **value**	**statistic**
CART-NFC, C45-IFRC, FPT, SGERD	Reject	0.00078	16.785

The Friedman Aligned Ranks test is capable of detecting any significant

differences within a certain group. However, it is unable to establish explicit comparisons when using a certain control method over a set of possible alternatives. As CART-NFC achieves the smallest error and is of the lowest ranking bar among the four compared algorithms, it is of a natural appeal to utilise it as the control method in comparison to FPT, C45-IFRC and SGERD. The standard Holm's procedure [40] is applied to run this test, computing the adjusted p values. The results of this investigation are presented in Table 6. Since the p values are smaller than the level of significance specified by $\alpha = 0.05$, the null hypothesis that there exists no significant performance difference between CART-NFC and FPT or between CART-NFC and SGERD is rejected. Thus, it can be concluded statistically that CART-NFC works significantly better than both FPT and SGERD. Despite CART-NFC achieves better average performance than C45-IFRC, no statistical difference between them can be detected under this setting. In a nutshell, these results demonstrate that the present work is at least competitive to the state-of-the-art fuzzy classifiers for clinical decision support.

Table 6: Result of running Holm's procedure

Comparison	**Hypothesis** ($\alpha = 0.05$)	**Adjusted** p **value**	**Statistic**
CART-NFC v.s. SGERD	Reject	0.00004	4.36251
CART-NFC v.s. FPT	Reject	0.04770	2.25956
CART-NFC v.s. C45-DFRC	Accepted	0.34741	0.93962

4.6. Model Complexity

Fuzzy systems adopted in conventional applications (e.g., [41]) typically aim to maximise certain performance metrics, but they may take it for granted that models are transparent given the semantics of the underlying fuzzy sets utilised, thereby overlooking the overall model interpretability. Apart from performance, interpretability of application systems and their reasoning processes should also be taken into consideration while designing a fuzzy model for clinical decision support. This is important to facilitate the interpretation of any resultant fuzzy rules to medical professionals and that of diagnostic outcomes to the patients.

Unlike performance criteria such as accuracy that can be used to objectively measure how good a fuzzy model is, the assessment of system interpretability is a subjective property, largely depending on the person who makes the assessment. Unfortunately, the interpretability may be affected by a range of practical issues [42]. Nonetheless, whilst there generally lacks a commonly accepted mechanism to make an informed objective judgement, the complexities of a resultant fuzzy system are of great significance to be worth careful consideration. The aim is to obtain a fuzzy model equipped with a small number of rules and a small number of antecedents per rule.

Table 7: Comparison on model complexity

	CART-NFC		**C45-IFRC**		**FPT**		**SGERD**	
Data Sets	**Rul**	**Cond**	**Rul**	**Cond**	**Rul**	**Cond**	**Rul**	**Cond**
APN	7.07	1.63	2.76	1.97	2.00	1.50	2.48	2.00
BLD	13.04	2.37	7.51	2.67	2.00	1.50	2.69	2.00
MM	24.82	2.44	10.90	3.15	2.00	1.50	2.02	2.00
PD	21.85	2.36	12.48	4.93	2.00	1.50	2.47	2.00
PID	14.02	2.04	14.66	3.19	2.00	1.50	3.71	2.00
TG	10.47	1.83	8.77	2.60	3.00	3.00	3.26	2.00
VC	12.30	2.27	8.12	3.36	2.00	1.50	2.98	2.00
WDBC	16.81	2.43	18.42	3.63	2.00	1.50	2.18	1.53
WOBC	28.67	2.72	13.40	4.98	2.00	1.50	3.57	2.00
average	16.561	2.231	10.780	3.387	2.111	1.667	2.818	1.948

Table 7 presents an empirical analysis of the complexity of learned fuzzy rule bases, in terms of the average number of antecedent conditions (*Cond*) per fuzzy rule, and average number of rules (*Rul*) per rule base. For *Cond*, C45-IFRC comes last with rule length systematically longer than the proposed approach across all nine data sets. FPT and SGERD return the most compact rules, with both having learned fuzzy rules involving fewer than 2 antecedent conditions. Following these two, the rule base returned by the proposed work also enjoys high structural interpretability, being able to learn rules with just slightly over 2 antecedents on average in length.

For *Rul*, FPT returns rule bases with the smallest size, due to their imposed heuristic nature of setting the number of rules to the number of the classes.

SGERD also generates highly compact rule bases with not only very small rule sizes but also short rules. Yet, both classification performances of FPT and SGERD are statistically worse compared with that of the proposed approach as shown previously. Compared with C45-IFRC, CART-NFC has a few more rules on average, but it uses fewer rule antecedents than those of C45-IFRC. If the total number of antecedents per rule base is summarised by calculating ($\#Rule * \#Cond$), both CART-NFC (36.95 antecedents) and C45-IFRC (36.51 antecedents) are very close without significant differences. Overall, it can be concluded that CART-NFC is able to learn rule bases of a small cardinality (each time returning fewer than 17 rules with just about 2 rule antecedents on average across all medical benchmarks).

Although C45-IFRC [30] and the proposed CART-NFC both adopt the general idea of initialising a preliminary fuzzy system through a crisp rule base, the underlying methodologies of how crisp rules are utilized to initialise the fuzzy system and of how the subsequent optimisation is performed are completely different. For C45-IFRC, each of the generated crisp rules is firstly converted into a set of fuzzy rules involving predefined fuzzy sets through a heuristic mapping procedure; whereas crisp rules in this paper are directly converted into fuzzy rules by replacing crisp intervals with fuzzy sets and logical operators with their respective fuzzy counterparts. A local rule selection procedure is then performed in C45-IFRC to obtain a compact subset of initially mapped fuzzy rules that jointly generalise the capability of the underlying crisp rule. This involves a computational cost of $O(N \times 2^{N^{intl}} \times T)$ at this stage, where N denotes the number of given crisp rules, N^{intl} is the maximum number of the existing crisp intervals for any crisp rule, T is the maximum number of similar fuzzy sets that are allowed per crisp interval. In [30], the population-based genetic algorithm (GA) is used to perform such optimization. For the proposed work, there is no such an intermediate step.

For C45-IFRC, the above procedure is followed by an additional module, where a fine grain tuning of all selected subsets of fuzzy rules is finally carried out with another GA, at a computational cost of $O(d^n * N_r)$, where d is the

maximum number of predefined fuzzy sets per attribute, n is the number of antecedent attributes in the problem domain and N_r is the number of fuzzy rules previously selected through the above step. For CART-NFC, the corresponding computational cost is $O(N_r * 2n) < O(d^n * N_r) + O(N \times 2^{N^{intl}} \times T)$. However, as the proposed work is for ANFIS models, with parameter optimization implemented by least squares and gradient descent, it is expected to converge faster than using population-based GA. In a nutshell, although C45-IFRC and CART-NFC exhibit performances regarding accuracy without statistical difference and the complexities of resultant fuzzy systems are of the same scale, the proposed method comes with lower computational cost and expected faster convergence rate.

Obtaining both a high degree of accuracy and a high degree of interpretability is a challenging contradictory aim and, in practice, one of these two conflicting demands usually prevails over the other [50]. As such, a balanced trade-off between interpretability and accuracy must be considered when designing a fuzzy model for a specific application. Instead of simply resorting to weighting the importance of accuracy/interpretability quantitatively, which may be subject to various experimental settings and hence, can be very difficult to estimate precisely, qualitative decisions for choosing a model most appropriate among those non-dominating solutions (with respect to domain specific requirements) is the way forward [49]. Here, a dominating solution is one that would achieve both higher accuracy and lower complexity than its alternatives given idealized conditions (which may not be easy to confirm or obtain in practice, without comprehensive experimental investigations). To reflect this general observation, further comparative studies are made against the other two alternative methods, namely, FPT and SGERD.

For FPT, the proposed method achieves a statistically significant accuracy gain of 1.61% on average. This improvement is of a remarkable impact in practice when such a system is adopted to serve a wide range of patients on a larger scale, implying many more correct diagnostic outcomes as compared to the use of FPT. From the perspective of model complexity, FPT produces models of

an average complexity of 3.52 antecedents as summarized by $(\#Rule * \#Cond)$, which is substantially simpler than what is attained by the proposed method (36.95 antecedents). However, the underlying approach taken here to compare the model complexity is an overly simplified one that counts the generated rule numbers and antecedents only, at a superficial level.

Nonetheless, it should be noted that FPT learns logical connectors that connect adjacent antecedent conditions, which may be rather difficult for domain users to comprehend. For example, a typical fuzzy rule produced by FPT may appear like:

If x_1 is A_1 Einstein-AND x_2 is B_1 Algebraic-OR x_3 is C_1, Then y.

where x_1, x_2, x_3 are the domain variables; A_1, B_1, C_1 are the corresponding fuzzy sets taken by these variables; y is the rule consequent; the connector Einstein-And is supposed to be a logical operator that takes two inputs (a, b) and generates $\frac{ab}{(2-(a+b-ab))}$ as a compounded rule antecedent; and similarly, Algebraic-OR$(a, b) = a + b - ab$. There are even more complicated non-linear logical operators than these that may be used to join antecedents in FPT. Such complicated mathematical interpretation may make the resulting learned rules impractical to interpret without sufficient theoretical backgrounds. Fortunately, the proposed method only uses a basic logical AND operator, significantly facilitating the interpretation of its learned rules. As such, it is fair to conclude that from the interpretability viewpoint of learned rules, the proposed method beats FPT.

Comparing the proposed work with SGERD shows an even more statistically significant result with an average accuracy gain of $\sim$3.6%. Again, such a significant improvement demonstrates practical significance, especially when the system is used to serve large numbers of patients. From the rule base complexity perspective, whilst both approaches learn short rules, SGERD that employs only basic AND logical operator as well, enjoys a low structural complexity by learning rules of 3.52 antecedents on average. Therefore, these two methods have different strengths and may be applied in different scenarios depending on the real needs. When accurate diagnoses become the chief requirement for the

diagnostic system, which is the generally expected in medical applications, the proposed work clearly serves better.

In short, the proposed method outweighs FPT with statistically significant performance on accuracy while generally having higher interpretability at rule level, though its overall model complexity is higher than that of FPT. It also outperforms SGERD, with an even larger margin in terms of model accuracy, but it has a higher model complexity. When compared to C45-IFRC, both approaches learn a rule base of fairly similar model complexity and accuracy, but the proposed method has lower computational overheads. Overall, the proposed method is able to learn an accurate and interpretable fuzzy system model of reasonable complexity (16.56 rules per system with 2.23 antecedents per rule on average), with low computational overheads, while at least having a performance on a par with the state-of-the-art fuzzy model alternatives.

4.7. Effect of Discretisation with k-means

On the assumption that all variables are assumed uniformly distributed, without any optimization (which is purposefully designed so as to enable systematic investigations over a wide range of experimental settings without bias), the previous experimental analysis has already shown that the proposed approach achieves competent results in comparison with alternative popular fuzzy approaches. It is interesting to empirically investigate what if an (at least partially) optimised quantity space is utilized. Without overly complicating the experimental investigation, the illustrative example on the diabetes diagnosis in Section 4.2 is reused here to explore the potential effect of optimised quantity space.

Instead of discretising the domain of each continuous attribute into 3 equally spaced bins as the previous example did, the k-means clustering algorithm [43] is first performed with respect to each attribute, resulting in 3 exclusive clusters. The minimal and maximum values from a certain cluster will then be used to generate the crisp interval, serving as the original input to the decision tree. Fig. 8 summaries the membership functions for the variable *Glucose* un-

der different settings. In particular, Fig. 8(a) depicts the original membership functions where the uniform distribution is adopted, which is then tuned by ANFIS, resulting in Fig. 8(b) with the membership function highly squashed. Fig. 8(c) gives the optimised quantity space and Fig. 8(d) is a minor variation of Fig. 8(c) optimised by ANFIS.

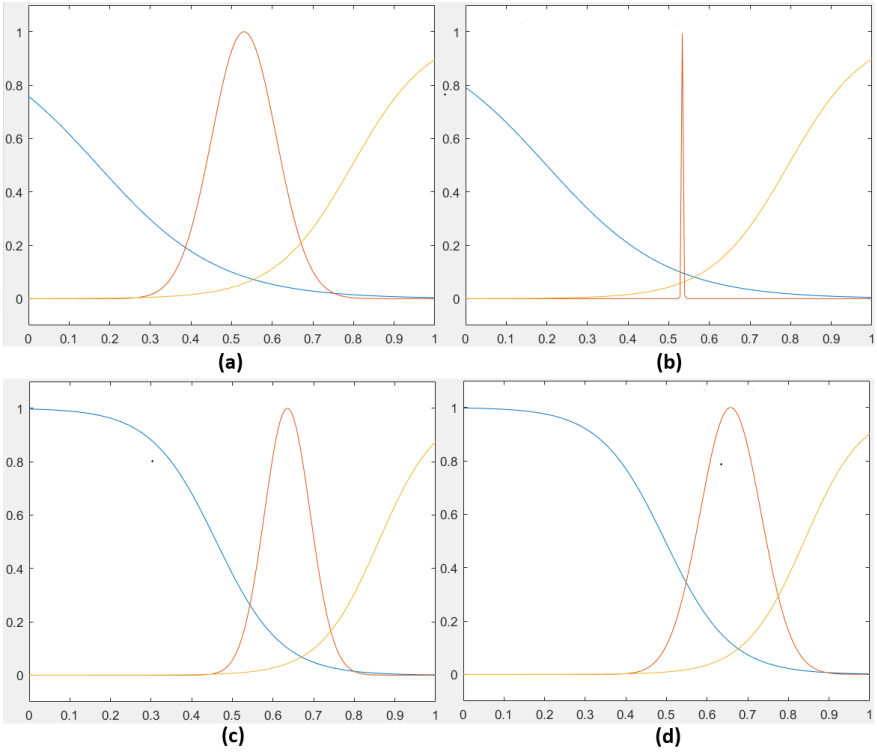

Figure 8: Membership functions of variable Glucose plotted in sub-figures

Unlike the previous set of five crisp rules involving the variables *Glucose* and *BMI*, the resultant crisp rule base with a quantity space specified through k-means optimisation leads to a rules base that is even more compact, with only 3 rules involving just *Glucose* as follows:

- Rule 1: If Glucose is *low*, Then test *negative*;
- Rule 2: If Glucose is *medium*, Then test *negative*;

- Rule 3: If Glucose is *high*, Then test *positive*.

Interestingly, the variable *BMI*, which has been utilised in the construction of the decision three when uniform distribution is adopted, is now completely discarded. This is not surprised, as how the input variables are partitioned will directly affect the selection of attribute-value pair whilst the tree expands, thereby possibly leading to an early stop when using an optimised quantity space.

When an optimised quantity space (as per Fig. 8(c)) is employed, the fuzzy system implemented using these three rules performs exactly well as its counterpart with five crisp rules. Recall that the proposed method achieved better or competent results in comparison with alternative fuzzy methods when considering the initial and the worst case scenario in which uniform distribution of the domain values was assumed. Together, these results demonstrate that the proposed approach can offer an even more compact outcome when the quantity space of the antecedent variables is optimised with a certain optimisation algorithm, without adversely affecting the performance level in terms of model accuracy. This is of great practical significance since more compact rules are easier to understand and the inference processes involving the use of such rules are in turn, easier to explain.

5. Conclusion

This paper has proposed an effective approach for learning a fuzzy rule base, with target applications to clinical decision support. The proposed approach starts with the generation of a crisp rule base from given data using a decision tree learning mechanism, producing basic rule structures that reflect the characteristics between domain inputs and output while having low computational overheads. The crisp rule base is then transformed into a fuzzy rule base with crisp intervals replaced by Gaussian membership functions. This forms the input for subsequent neuro-fuzzy adaption implemented by an ANFIS, optimising the fuzzy rules. The proposed work is able to track back how a diagnosis may be

reached by decomposing the matching degrees of an input observation against each of the available rules in the system. This explicitly shows the contributions made by individual rules towards the overall decision, thereby explaining how the conclusion is reached from the given input. In addition to such transparent inference process, interpretability is also supported by the generation of a rule-based system of reasonable size with short rules. Furthermore, statistical comparative results have shown that the proposed work achieves better or at least, comparable performance to those derived from state-of-the-art fuzzy classifiers.

Whilst promising, interesting work remains for further development. This includes examining the effect of any subsequent fine-tuning of the ANFIS employed, and investigating the use of more powerful data discretisation techniques in conjunction with the exploitation of the bin number for each variable domain. The latter may be implemented by adapting soft clustering techniques (e.g., [44], [45]) that enable the generation of overlapping intervals, which supports the potential integration with the proposed fuzzy classifier given the relevance of the mathematical theories underpinning these methods. Interesting further work also includes working with missing values that often arise in clinical practice by exploiting advanced knowledge interpolation techniques (e.g., [46], [47]).

It is acknowledged that the present interpretability may not be achieved with respect to the clinical practice standard, unless the system is constructed in close consultation with medical professionals. For real applications, domain experts would be required to advise on any specifications and/or constraints that correctly reflect domain expertise while devising the underlying system. Thus, future work also intends to redesign the ANFIS structure by adding an extra linguistic hedge layer before the membership function layer, to describe the level of fulfilment associated with the corresponding fuzzy sets. The implementation of this latter aspect will help enrich the hypothesis space for parameter adjustment, facilitating ANFIS model optimisation effectively without adversely affecting the interpretability of the membership functions used to describe the values of domain variables [48]. Lastly, the proposed approach is herein devised

to develop an interpretable fuzzy system for clinic decision support, the underlying principles appear to be generic. It is therefore very interesting to apply it for addressing different problems such as plant monitoring [6] and network security [51].

Acknowledgments

This work has been supported by the National Science Foundation Program of China (no. 61906181) and the China Postdoctoral Science Foundation (no. 2019M652156). The authors are very grateful to the anonymous reviewers whose review comments have helped improve this work significantly.

References

[1] C. S. Kruse, R. Goswamy, Y. Raval, and S. Marawi, "Challenges and opportunities of big data in health care: a systematic review," *JMIR Medical Informatics*, vol. 4, no. 4, 2016.

[2] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2018, pp. 559–560.

[3] 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 *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1721–1730.

[4] P. Su, C. Shang, T. Chen, and Q. Shen, "Exploiting data reliability and fuzzy clustering for journal ranking," *Fuzzy Systems, IEEE Transactions on*, vol. 25, no. 5, pp. 1306–1319, 2017.

[5] T. Chen, Q. Shen, P. Su, and C. Shang, "Fuzzy rule weight modification with particle swarm optimisation," *Soft Computing*, vol. 20, no. 8, pp. 2923–2937, 2016.

[6] P. Su, Q. Shen, T. Chen, and C. Shang, "Ordered weighted aggregation of fuzzy similarity relations and its application to detecting water treatment plant malfunction," *Engineering Applications of Artificial Intelligence*, vol. 66, pp. 17–29, 2017.

[7] R. Meza-Palacios, A. A. Aguilar-Lasserre, E. L. Ureña-Bogarín, C. F. Vázquez-Rodríguez, R. Posada-Gómez, and A. Trujillo-Mata, "Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus," *Expert Systems with Applications*, vol. 72, pp. 335–343, 2017.

[8] B. Cosenza, "Off-line control of the postprandial glycemia in type 1 diabetes patients by a fuzzy logic decision support," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 693–10 699, 2012.

[9] M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14 650–14 659, 2011.

[10] S. Lekkas and L. Mikhailov, "Evolving fuzzy medical diagnosis of pima indians diabetes and of dermatological diseases," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 117–126, 2010.

[11] V. V. Kamadi, A. R. Allam, S. M. Thummala *et al.*, "A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (pca) and modified fuzzy sliq decision tree approach," *Applied Soft Computing*, vol. 49, pp. 137–145, 2016.

[12] S. El-Sappagh, M. Elmogy, and A. Riad, "A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis," *Artificial Intelligence in Medicine*, vol. 65, no. 3, pp. 179–208, 2015.

[13] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," *Telematics and Informatics*, vol. 34, no. 4, pp. 133–144, 2017.

[14] G. H. B. Miranda and J. C. Felipe, "Computer-aided diagnosis system based on fuzzy logic for breast cancer categorization," *Computers in Biology and Medicine*, vol. 64, pp. 334–346, 2015.

[15] S. Aminikhanghahi, S. Shin, W. Wang, S. I. Jeon, and S. H. Son, "A new fuzzy gaussian mixture model (fgmm) based algorithm for mammography tumor image classification," *Multimedia Tools and Applications*, vol. 76, no. 7, pp. 10 191–10 205, 2017.

[16] M. Nilashi, O. Ibrahim, and A. Ahani, "Accuracy improvement for predicting parkinson's disease progression," *Scientific Reports*, vol. 6, p. 34181, 2016.

[17] K. Polat, "Classification of parkinson's disease using feature weighting method on the basis of fuzzy c-means clustering," *International Journal of Systems Science*, vol. 43, no. 4, pp. 597–609, 2012.

[18] J.-S. Jang, "Anfis: adaptive-network-based fuzzy inference system," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 23, no. 3, pp. 665–685, 1993.

[19] R. Ambrosino, B. G. Buchanan, G. F. Cooper, and M. J. Fine, "The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies." in *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 1995, p. 304.

[20] W. C. Knowler, P. H. Bennett, R. F. Hamman, and M. Miller, "Diabetes incidence and prevalence in pima indians: a 19-fold greater incidence than in rochester, minnesota," *American Journal of Epidemiology*, vol. 108, no. 6, pp. 497–505, 1978.

[21] M. A. Chikh, M. Saidi, and N. Settouti, "Diagnosis of diabetes diseases using an artificial immune recognition system2 (airs2) with fuzzy k-nearest neighbor," *Journal of Medical Systems*, vol. 36, no. 5, pp. 2721–2729, 2012.

[22] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[23] H.-L. Chen, C.-C. Huang, X.-G. Yu, X. Xu, X. Sun, G. Wang, and S.-J. Wang, "An efficient diagnosis system for detection of parksinson's disease using fuzzy k-nearest neighbor approach," *Expert Systems with Applications*, vol. 40, no. 1, pp. 263–271, 2013.

[24] C. Balleyguier, S. Ayadi, K. Van Nguyen, D. Vanel, C. Dromain, and R. Sigal, "BIRADS^{TM} classification in mammography," *European Journal of Radiology*, vol. 61, no. 2, pp. 192–194, 2007.

[25] A. Keleş, A. Keleş, and U. Yavuz, "Expert system based on neuro-fuzzy rules for diagnosis breast cancer," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5719–5726, 2011.

[26] M. Pota, M. Esposito, and G. De Pietro, "Designing rule-based fuzzy systems for classification in medicine," *Knowledge-Based Systems*, vol. 124, pp. 105–132, 2017.

[27] D. Soria, J. M. Garibaldi, A. R. Green, D. G. Powe, C. C. Nolan, C. Lemetre, G. R. Ball, and I. O. Ellis, "A quantifier-based fuzzy classification system for breast cancer patients," *Artificial Intelligence in Medicine*, vol. 58, no. 3, pp. 175–184, 2013.

[28] K. A. Rasmani, J. M. Garibaldi, Q. Shen, and I. O. Ellis, "Linguistic rulesets extracted from a quantifier-based fuzzy classification system," in *2009 IEEE International Conference on Fuzzy Systems*. IEEE, 2009, pp. 1204–1209.

[29] D.-C. Li, C.-W. Liu, and S. C. Hu, "A fuzzy-based data transformation for feature extraction to increase classification performance with small medical

data sets," *Artificial Intelligence in Medicine*, vol. 52, no. 1, pp. 45–52, 2011.

[30] T. Chen, C. Shang, P. Su, and Q. Shen, "Induction of accurate and interpretable fuzzy rules from preliminary crisp representation," *Knowledge-Based Systems*, vol. 146, pp. 152–166, 2018.

[31] R. Senge and E. Hüllermeier, "Fast fuzzy pattern tree learning for classification," *Fuzzy Systems, IEEE Transactions on*, vol. 23, no. 6, pp. 2024–2033, 2015.

[32] H. Ishibuchi, S. Mihara, and Y. Nojima, "Parallel distributed hybrid fuzzy gbml models with rule set migration and training data rotation," *Fuzzy Systems, IEEE Transactions on*, vol. 21, no. 2, pp. 355–368, 2013.

[33] E. G. Mansoori, M. J. Zolghadri, and S. D. Katebi, "Sgerd: A steady-state genetic algorithm for extracting fuzzy classification rules from data," *Fuzzy Systems, IEEE Transactions on*, vol. 16, no. 4, pp. 1061–1071, 2008.

[34] D. García, A. González, and R. Pérez, "Overview of the slave learning algorithm: A review of its evolution and prospects," *International Journal of Computational Intelligence Systems*, vol. 7, no. 6, pp. 1194–1221, 2014.

[35] R. Senge and E. Hullermeier, "Top-down induction of fuzzy pattern trees," *Fuzzy Systems, IEEE Transactions on*, vol. 19, no. 2, pp. 241–252, 2011.

[36] G. Feng, "A survey on analysis and design of model-based fuzzy control systems," *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 5, pp. 676–697, 2006.

[37] D. A. Zighed, S. Rabaséda, and R. Rakotomalala, "Fusinter: a method for discretization of continuous attributes," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 03, pp. 307–326, 1998.

[38] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[39] L.-X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 6, pp. 1414–1427, 1992.

[40] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, no. Jan, pp. 1–30, 2006.

[41] B.-S. Yang, M.-S. Oh, A. C. C. Tan *et al.*, "Fault diagnosis of induction motor based on decision trees and adaptive neuro-fuzzy inference," *Expert Systems with Applications*, vol. 36, no. 2, pp. 1840–1849, 2009.

[42] K. Cpałka, "Design of interpretable fuzzy systems," *Studies in Computational Intelligence*, 684(1), 2017.

[43] D. Steinley, "K-means clustering: a half-century synthesis," *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 1, pp. 1–34, 2006.

[44] T. Boongoen, C. Shang, N. Iam-On, and Q. Shen, "Extending data reliability measure to a filter approach for soft subspace clustering," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 6, pp. 1705–1714, 2011.

[45] K. Thangavel, Q. Shen and A. Pethalakshmi, "Application of clustering for feature selection based on rough set theory approach," *Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 19–27, 2006.

[46] T. Chen, C. Shang, J. Yang, F. Li, and Q. Shen, "A new approach for transformation-based fuzzy rule interpolation," *Fuzzy Systems, IEEE Transactions on*, 2019. [Online]. Available: https://doi.org/10.1109/TFUZZ.2019.2949767

[47] F. Li, Y. Li, C. Shang, and Q. Shen, "Fuzzy knowledge-based prediction through weighted rule interpolation," *Cybernetics, IEEE Transactions on*, 2019. [Online]. Available: https://doi.org/10.1109/TCYB.2018.2887340, 2019

[48] J. G. Marín-Blázquez and Q. Shen, "From approximative to descriptive fuzzy classifiers," *Fuzzy Systems, IEEE Transactions on*, vol. 10, no. 4, pp. 484–497, 2002.

[49] R. Leitch, Q. Shen, G. Coghill, and M. Chantler, "Choosing the right model," *IEE Proceedings-Control Theory and Applications*, vol. 146, no. 5, pp. 435–449, 1999.

[50] O. Cordón, "A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems," *International Journal of Approximate Reasoning*, vol. 52, no. 6, pp. 894–913, 2011.

[51] T. Chen, P. Su, C. Shang, and Q. Shen, "Weighted fuzzy rules optimised by particle swarm for network intrusion detection," in *Fuzzy Systems, 2018 IEEE International Conference on*, IEEE, 2018, pp. 1–7.