IET Software

Special Issue: Knowledge Discovery for Software Development (KDSD)

Using health data repositories for developing clinical system software: a multi-objective fuzzy genetic approach

ISSN 1751-8806
Received on 30th August 2019
Revised 28th November 2019
Accepted on 25th February 2020
E-First on 15th April 2020
doi: 10.1049/iet-sen.2019.0261
www.ietdl.org

Bilal S. Raja[1] ✉, *Sohail Asghar*[1]
[1] *CS Department, COMSATS University Islamabad, Pakistan*
✉ *E-mail: bilalsraja@gmail.com*

Abstract: Evolution of technology has brought a revolution in various fields of sciences and amongst them, healthcare is one of the most critical and sensitive areas because of its connection with common masses' quality of life. The notion of integrating the healthcare system with the latest data repositories is to make disease prediction efficient, transparent, and reusable. Due to data heterogeneity, data repositories along with optimum classifiers help stakeholders to predict the disease more accurately without compromising the interpretability. Evolutionary algorithms have shown great efficacy, accuracy, and interpretability in improving disease prediction for several datasets. However, the quest for the best classifier is still in evolution. In this research, a state-of-the-art medical data repository has been developed to give researchers of medical domain great ease of use in utilizing different datasets governed by a multi-objective evolutionary algorithm using fuzzy genetics. The proposed model called 'MEAF' is evaluated on various public repositories. A subset of these repositories includes breast cancer, heart, diabetes, liver, and hepatitis datasets. The results have been analyzed, which show competitive accuracy, sensitivity, and interpretability as compared to relevant research. A customised software application named 'MediHealth' is developed to supplement the proposed model that will facilitate the domain users.

1 Introduction

Recent years have witnessed the widespread adoption of decision support system (DSS) in many fields such as customer relationship management (CRM) [1], education [2], clinical medicine [3], financial fraud detection [4], intrusion detection [5], and genetic data analysis [6]. Even, it can equitably be expected that the adoption of DSS will increase drastically. Therefore, it becomes hard when the manual processing of the clinical DSS (CDSS) is invoked from the perspective of clinical medicine. The CDSS comprises two major categories of evaluation. One is screening and the other is the diagnosis. Screening is the process of persuasion to identify unknown disease when its symptoms are not unveiled or excluding a suspected disease whose symptoms are not clear. On the other hand, diagnosis evaluates the person by the severity of the disease when its symptoms have been observed. The timely detection of disease can avoid its progression and the patients' life can be saved. Therefore, CDSS holds flair to improve the human health quality as there is a dire requirement of automatic extraction of information. CDSS expedites the stakeholders' (physicians, clinicians etc.) decision-making through knowledge inference and assertion justification. To date, the researchers of the globe have evolved various techniques and algorithms that could be used in screening or diagnosis under medical disease prediction. The plethora of research studies on CDSS is gaining brisk popularity amongst the stakeholders as compared to the other fields of science that yields befitting effective clinical decisions. CDSS can be defined as an 'extraction of implicit potentially useful and novel information from medical data to improve accuracy, decrease time and cost, construct decision support systems with the aim of health promotion' [7]. This domain aims to enhance the effectiveness and decrease miscalculation added by humans in medical decision mining [8], extraction of a relationship between variables, seeking fresh knowledge and risk factors identification [9] with utmost accuracy and interpretability. This is relatively different from other fields of science such as statistics because in this case data involves human life [10]. Patel *et al*. [11] have accentuated the issue of confidentiality and the absence of standardised methodology in evolving an intelligent CDSS. Therefore, for decision analysis, medical data is a cumbersome choice. Medical data is generated from heterogeneous mediums and forms that include pictorial and numerical indices. There is a need to choose a suitable DSS with an explicit algorithm based on the type of data. In [12], the author has highlighted clinical and temporal data as the two major categorisations of medical data that could facilitate the researchers in the field of patient management, diagnosis, monitoring, treatment, and screening. Medical data of different types can be elaborated below:

- Sonography and echo data
- Imaging data (X-ray and magnetic resonance imaging)
- The numeric and quantitative data format for trial purposes
- Qualitative and text-based medical data
- Timestamp-based temporal data
- Time series data such as electroencephalography & electrocardiography
- Protein, microarray, and genetic data

Owing to the heterogeneous and diversified nature of data, this field has become quite intricate. There are four main types of CDSS which are categorised on their operational characteristics:

- *Submissive systems* – unassertive nature and clinical practitioner invokes the system as and when needed.
- *Semi-vigorous systems* – assertive in nature and are invoked automatically as a watchdog. They also generate alerts in certain clinical situations.
- *Vigorous systems* – highly assertive in nature and can prescribe the protocols based on certain standard operating procedure, e.g. additional pathological examination etc.
- *Supervisory control systems* – highly assertive in nature and can intelligently control the highly automated parameters of machinery like ventilators in the emergency room.

The focal contributions of this research are enlisted as follows:

- A unique semantic aware data repository is developed that is generated on the latest graph database technology.

- A multi-objective evolutionary algorithm using a fuzzy genetic model is presented that is capable of handling multiple objectives, i.e. accuracy, sensitivity, specificity, and interpretability.
- Tangible comparison to depict and prove improvement/ superiority of our framework with existing techniques of similar nature available in the literature.
- Software application 'MediHealth' is developed that uses MEAF as a prediction engine/model. The application can be helpful to the physicians/hospitals for clinical decision-making especially in remote areas where specialised healthcare facilities are not available.

The overall composition of paper is organised as the upcoming Section 2 describes the related work followed by the proposed model MEAF that is presented in Section 3. Section 4 provides the experimental evaluation and result details. Discussion on the MediHealth application is presented in Section 5 and the conclusion and future work are elaborated in Section 6.

2 Related work

Several authors have identified major challenges of clinical DSSs which should be handled with zeal. One of the major challenges stated to evolve an efficient decision-making framework is the heterogeneous clinical data, its utilisation, and storage into repository [13]. Another keystone that needs serious deliberation is to observe 'privacy'. This concern has been addressed by different authors in the telemedicine domain [14]. Health care system privacy is also emphasised in [15], where cloud-based health care system OpenNCP is presented. For any CDSS, patients are considered as the pivot of the whole concept of data confidentiality in such integrated systems [16], where a novel digital signature scheme approach is presented. While considering to design an integrated health solution that is based on centralised data management for the medical domain, various aspects of this approach should be kept in mind. Some of these considerations may include computational complexities, multiple access points, and information from multiple sources. Several papers have highlighted the strength and effective utilisation of integrated clinical data management in a semantically aware and standardised way [17–19]. In [20], a review on utilising semantic integration methodologies and ontologies proves the huge potential, but still, there is no gold standard benchmark solution available. Chronic disease management in ambulatory and primary care is presented in this research. To evolve an integrated medical data repository one can enjoy following salient benefits:

1. Any modifications to the semantic model are not trivial, let it be a new deployment or an already deployed framework. This ensures flexibility and reusability. Thus, it is easy to make modifications by making small changes to a dataset because of its interfacing through XML.
2. The model is based on the subject–predicate–object principle, hence it does not requires stored information to be perceived at the time of system development.
3. The model provides a single entry point to update and reflect changes to various layers rather than making changes to multiple layers (which is quite costly). This single point of modification does not require the code routines to be deployed again. Thus the re-compilation step is not required.

Since this research provides a working model of disease prediction using fuzzy genetic, therefore, the main objectives of data repository are depicted below:

Domain entity definition

1. Designing of medical data repository
2. Data validation
3. Semantic interoperability
4. Flexibility and reusability
5. Research resource

The above objectives provide a clear understanding of various benefits that can be availed by deploying a context-aware data repository. These objectives cover various standards that are mandatory for a system that involves data modelling from entity definition to ensuring flexibility, interoperability, and reusability. It is also crucial to ensure that data is relevant, i.e. pertaining to the medical domain and it is in the correct format. This way it becomes easier to certify the cross-platform interoperability of data during the exchange. The semantic interoperability of data is based on a strong subject–predicate–object relationship to ensure maximum accuracy of results. One of the main objectives of this selection is to make sure that these tools are available free of cost for researchers for any future work. As highlighted earlier, the emphasis of this research after developing a context-aware data repository is to propose and develop an appropriate disease prediction model that has the capability of handling multi-objective such as interpretability and accuracy simultaneously. A comprehensive representation based on the literature has been represented in the upcoming section that highlights the pros and cons of various disease prediction frameworks and techniques.

2.1 Classification

In [21], a decision tree-based solution is presented that achieved a high specificity and sensitivity yielding an earlier detection/ diagnosis of prostate cancer. Although, the algorithm uses a greedy approach that is easily understandable, and resilient to replication and data with noise. However, the algorithm is not good with inconsistent data and compromise in accuracy is also exercised. In [22], an artificial neural network (ANN)-based solution is presented that achieved high accuracy in determining heart failure risk by combining ANN with a fuzzy analytic hierarchy process. The research involves more than one layer with a precondition of holding one layer bare minimum and is resilient to data replication with the ability to handle complex relationships, but the model showed vulnerabilities to the data anomalies & hides internal details. In [23], a Rule-based classifier with a genetic algorithm (GA) was used to predict breast cancer, diabetes, and heart disease. Rule base (RB) classification achieved 98.5, 81.5, and 89.5% accuracy, respectively. The RBA is based on if–then–else rules that can be easily understandable and have the resilience to inconsistent data but has compromised accuracy. In [24], the accuracy of 88.4% is achieved in diabetes detection when support vector machine (SVM) is used with back propagation neural network on the PIMA dataset of the Indian female. As SVM is an unorthodox approach and involves mathematical calculations, the scheme is ideal when less training data is available and also suitable for multi-dimensional data but has a limitation of hiding internal details. In [25], a heart disease prediction system is presented that uses Naive Bayes (NB) as a classification algorithm. The predictive model showed remarkable results due to its statistical nature and resilience to all sorts of data abnormalities, but NB has a tendency to compromise in accuracy and requires an earlier probability for better results. The research in [26] compared the results of k-nearest neighbour (k-NN) and SVM for chronic renal disease prediction. k-NN showed better performance due to the elastic nature of the algorithm, but k-NN is susceptible to replication and noise.

2.2 Clustering

In [27], a hybrid approach that uses a k-mean and backpropagation neural network to accurately predict heart disease by achieving 97% accuracy on the UCI dataset is presented. This algorithm is good with processing speed, but not feasible for heterogeneous data and is susceptible to data noise. The research in [28] achieved 97% accuracy when the neuro-fuzzy system was applied to chronic kidney disease (CKD) data and hierarchical clustering established a strong association between CKD and diabetes. The algorithm uses a bottom-up approach with graph theory and is less prone to initial values, but is inclined to data noise and involves complexity both in terms of space and time. The research in [29] compared the results of various clustering techniques on a heart disease prediction system. Density based spatial clustering of application with noise Algorithm (DBSCAN) showed compromised results as compare to k-mean. The algorithm is noise resilient and has the ability to handle random density and size. In [30], the author demonstrated that the combination of the fuzzy c-means method and pattern recognition obtained promising results for the classification of Parkinson's disease. The scheme is not feasible for heterogeneous data and prone to data noise.

2.3 Association rule mining

In [31], the author proposes an a priori association rule-based technique to establish a relationship between healthcare parameters with heart disease symptoms for preventive measures. As a priori is a recursive approach which has widespread acceptance due to its straight forward nature, but algorithm lacks in generating complexity in terms of input/output and time. In [32], the author predicts that contagious disease has great tendency to affect the male person in the age range 30–60 which have poor environmental conditions and family history is not important through dynamic itemset counting algorithm. This algorithm is liable to non-heterogeneous data. The researchers in [33] performed an experiment on healthcare databases to mine frequent Itemset and equivalence class clustering and bottom-up lattice traversal (E-CLAT) algorithm outperforms all its competitors as it uses lattice theory and it's a bottom-up approach, but has a limitation of handling large data sets and is complex with respect to space. Research in [34] revealed through direct hashing and pruning algorithm that reduces the number of candidate patterns from the database of the hospital information system. The algorithm is founded on the concept of hashing and has the tendency to handle candidate pattern count, but lacks the handling of anomalies related to the hash table. In the same research, hospital information systems with huge volumes of data makes distance based maximum clique problem (D-CLUB) a better choice for parallel processing and distributed environment due to their dynamic nature.

2.4 Evolutionary algorithms

The conventional algorithms produce reasonable accuracy, but for a specific problem set. Another issue remained unaddressed that is of dealing with more than one objective. In [35], the authors have analysed the prediction of tumour grades using the backpropagation learning methodology and used a multi-layer neural network. The same methodology has been extended in [36] for ovarian cancer but due to its black-box nature, even with elevated accuracies have less impact [37]. In [38], the authors have proposed a neuro-fuzzy inference mechanism that showed outstanding accuracies by employing minimum fuzzy rules. Over time non-dominated sorting GA (NSGA) faced many problems in terms of computational efficiency, lack of elitism, and diversity, which were covered in NSGA-II [39]. Considerable work has already been done on fuzzy-based GAs in a medical DSS. Algorithms such as NSGA, NSGA-II, NSGA-III etc. are proficient only with two to three objectives. It has been learned from the various research evolution that when more than three objectives are exposed to these algorithms the expected outcome is not a success. This is due to the selection pressure loss that results in pulling the population towards the Pareto front (non-dominated solution) [40]. Although these approaches have tried to establish a trade-off in achieving multiple objectives by altering the weights, still a lot is required to address the full spectrum of solution space.

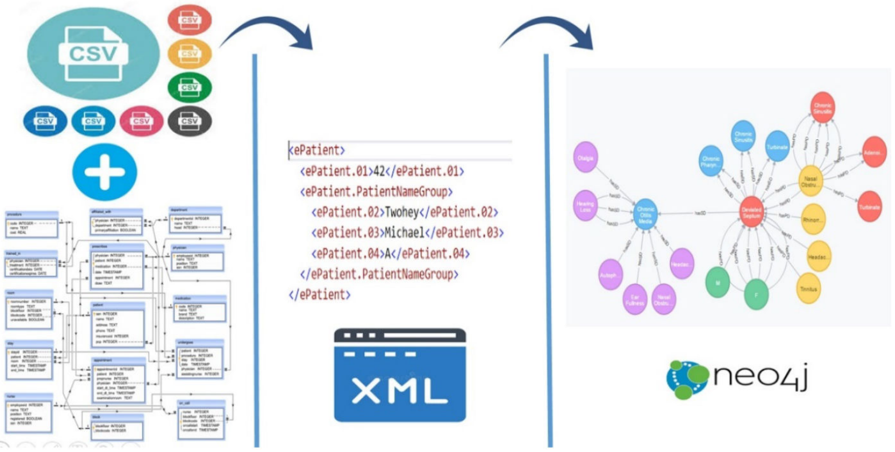
Based on the research done in this field so far, as mentioned in recent literature, the multi-objective algorithms are a good choice with two to three objectives. However, going beyond three objectives algorithms such as NSGA-II fall short with respect to diversity and convergence. This leads to a research direction where the algorithms can handle many objectives. Since the introduction of NSGA-III, a reference point based multi-objective (MO) GA has been presented to preserve the diversity of offspring solution by aiding multiple predefined reference points [41]. In [42], the authors have presented NSGA-III to remove the shortcomings and lacking NSGA-II, such as deficiency of gradual diversity in the contemporary best non-dominated solutions.

In NSGA-III, the preservation of varieties between offspring affiliates is covered by adaptively apprising the number of diversified points of reference. However, still NSGA-III left room for improvement in convergence. In [43], the authors have presented θ-NSGA-III that intended to evolve merging of NSGA-III in MO optimisation, but still needed to integrate diversity improvement for increasing the diversity and convergence of obtained solutions. However, in [44, 45], the research-focus shifted towards improving convergence and the overall performance of the algorithm. In [46], the authors have put forward U-NSGA-III algorithm to solve uni- and MO optimisation problems (MOOP). The model degenerates the equivalent Pareto-optimal solutions for the number of objectives for the problem. The methodology of the proposed research presents an algorithm that can fix the above-mentioned issues of MOOP. The proposed algorithm relates to a fuzzy logic system to gain accuracy. Initially, the results have been compiled with an enhanced and NSGA, where computational efficiency has been improved, keeping a balance between high accuracy and interpretability. The proposed research aims to use the algorithm with a fuzzy inference engine to develop the CDSS with improved accuracy. This study suggests that various areas of research, lead to the development of such GA that is able enough to outperform for multi-objective non optimized (MONO) and MO optimisation problems.

3 MEAF classification in disease prediction – the proposed approach

In order to design a modern repository to support the MEAF framework, it was very challenging to keep in mind several technical aspects. Since this research also provides the implementation of a data repository that is based on most recent graph technologies so it was mandatory to take care of standard aspects including data modelling, the ease of use, modification of various schemes along with message exchanging between various heterogeneous platforms. As shown in Fig. 1, a clinical data repository is presented. This repository is integrated with MEAF and accessed through the application interface MediHealth. Our data model is based on acquisitioning CSV files from a standard repository and then mapping it to XML format for web service interfacing and then to a modern context-aware data repository. To effectively utilise the strength of disease prediction and clinical DSSs a data repository adhering to the above-stated objectives in the related work section must comply. The basic motivation behind the research is to evolve a novel repository that uses flat files, relational database tables and semantically aware content to a single repository that has the capability of handling multiple heterogeneous datasets of the same domain e.g. Liver Disease Dataset with different attribute names and number. The choice of selecting the storage engine of our data repository is either to use a relational standard query language database (SQL DB) relational DB management system (RDBMS) or go with the new concept that is known as not only SQL (NoSQL). RDBMS is the most widely used storage engine, but they tend to suffocate when the requirement is of an evolving data model as in CDSS. Furthermore, RDBMS has no or limited support for context-aware data that has an additional cost of developing a semantic model over the data model to make it context-aware.

The storage and retrieval of NoSQL are quite different from RDBMS. NoSQL comprises graph, column-family, key-value and key-document stores [47]. Some of the key features of using NoSQL consist of cluster-friendliness but lacking in data integrity evaluation. In addition to this, NoSQL attempts to implement data access techniques that are not standardised. Further study revealed that their exist resource description framework (RDF) triple stores. RDF triples work on the notion of the subject–predicate–object model [48]. These triples can be operated in various ways by using import and export features to support multiple formats of data. The above framework is suggested and implemented for the development of a clinical data repository. There are quite some options available to handle context-aware data repository using ontology-based Protégé, Mongo DB, Orient DB, Elastic search but we opted for Neo4J due to its availability and ease of use. After the



```
<ePatient>
  <ePatient.01>42</ePatient.01>
  <ePatient.PatientNameGroup>
    <ePatient.02>Twohey</ePatient.02>
    <ePatient.03>Michael</ePatient.03>
    <ePatient.04>A</ePatient.04>
  </ePatient.PatientNameGroup>
</ePatient>
```

Fig. 1 *Clinical data repository framework*

development of a clinical data repository, the major objective of this research is to evolve a predictive engine for disease classification.

In CDSS, the accuracy and interpretability co-exist well critically. The comprehensive literature review evinces that there is still a need for significant improvement in the methodology of accuracy, interpretability, sensitivity, and specificity. The earlier designing of RB from data had more focus on high accuracy that has resulted in the framework with a high number of rules representing a huge number of previously generated occurrences.

In the later stage, explicitly manual removal of less active rules has been done for improving the interpretability, while keeping in view that the accuracy has not been compromised significantly. Various approaches such as neuro-fuzzy have been implemented in the area of the medical DSS [49–51] to address these issues. Another methodology has been presented in [52], which removed previously generated occurrences by taking an intersection of the related fuzzy set or by associating attuned data cluster through orthogonal transformation. Accuracy could be improved by applying fuzzy data clustering techniques that have been presented in [53]. Most of the methodologies had catered to the issues of interpretability and accuracy independently. Hence, they were unable to propose a comprehensive framework that could address both the objectives for the finest solution. To effectively address the above-stated issue, single and MO GAs have been used in [54, 55]. The challenge was still unachievable. For this reason, the dire need of a comprehensive mechanism defining a fitness-function that could optimise two major objectives, in making effective clinical decisions; interpretability and accuracy concurrently. The weighted sums of various components representing the interpretability and accuracy have been used instead of the fitness function definition.

Although these approaches have tried to establish a trade-off in achieving multiple objectives by altering the weights, still a lot is required to address the full spectrum of the search space. Pareto-optimal MO evolutionary optimisation methods were of interest to the domain experts of CDSS. Jin and Sendhoff [56] highlighted the computational complexity of proposed solutions by augmenting the strong fuzzy partitioning in their methodology. The above three problems have degraded the overall performance of CDSS techniques. There is a need to address these issues so that the resulting system would be accurate, more interpretable and efficient. The research questions that constitute the foundation of our proposed approach are given as follows:

Research question 1: Is it possible to evolve a framework that can improve the already existing methodologies of multi-objective optimisation problem by maximising the distance between the nearest neighbour and the non-dominated solution by applying fuzzy rule-based classification (FRBC)?

Research question 2: Is it possible to ingest the improved generalisation of some advanced NSGA that can produce the final selection with well-balanced distribution in the objective space and high spread to maximise the accuracy along with interpretability?

An overview of the proposed integrated CDSS-based framework for disease classification using multi-objective genetic fuzzy optimisation and its main components have been provided. The identified research problems in this specialised domain could be solved with various techniques/methods. However, we have learned from the existing techniques in the literature and hence propose a multi-objective evolutionary fuzzy rule-based GA. It is learned from the implementation of some techniques that the solution of clinical domain decision support with MO can be efficiently addressed by applying the proposed approach. The conceptual framework of our model is shown in Fig. 2.

To apply the proposed framework in action, an integral step is to pre-process the clinical data repository tuples for outliers detection and missing values imputations. This important step is explained in the upcoming section.

3.1 Data preprocessing

3.1.1 Missing data acquisition using the k-NN method: The proposed method uses the k-NN approach for acquisitioning the missing values from the dataset. The method is invoked over a group of tuples that have missing patterns; the *N* nearest recognisable attribute value is selected from the set of training data for missing value identification. When the *N* neighbour becomes recognisable, the missing attribute also becomes recognised. The equation calculates the distance between the variables of different types using heterogeneous Euclidean overlap. For instance, if $d(x_a, x_b)$ is the distance between variables x_a and x_b then the computation required for distance calculation is

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{n} d_j(x_{aj}, x_{bj})^2} \quad (1)$$

Here, the distance is calculated for the jth attribute. The overlapping distance is denoted by d_o and it holds a value 0 for the same quantitative features and 1 otherwise

$$d_o(x_a, x_b) = \begin{cases} 0 & x_{aj} \\ 1 & x_{bj} \end{cases} \quad (2)$$

Noise removal assists in the datasets' reduction. *K* mean clustering technique divides the data into *k* clusters. The centroid of each cluster is calculated and a benchmark value gets selected. If the attributes distance from the centroid is greater than the benchmark value, the attribute is considered as 'noise' and therefore, is eliminated. The following formula is used to calculate the number of clusters *K* from a number of data points *n*:

$$K \cong \sqrt{\frac{n}{2}} \quad (3)$$

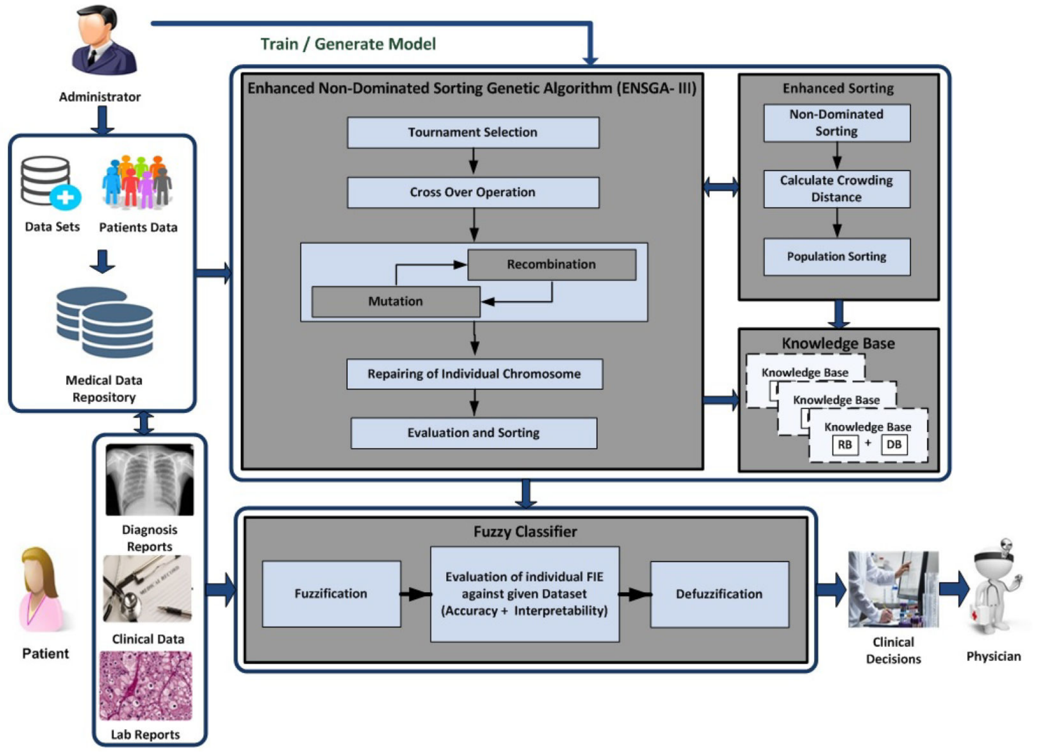


Fig. 2 *Multi-objective evolutionary algorithm using fuzzy genetic — MEAF framework*

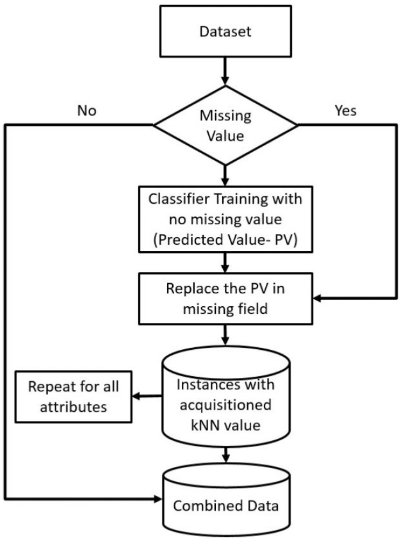

Dataset
No
Missing Value
Yes
Classifier Training with no missing value (Predicted Value- PV)
Replace the PV in missing field
Repeat for all attributes
Instances with acquisitioned kNN value
Combined Data

Fig. 3 *Missing data acquisition*

Ideally, the benchmark value is selected by calculating the mean of all the values within a cluster and is provided by the user most of the time.

3.1.2 Outlier detection: In medical data, the outliers are due to the patient's abnormal condition, equipment malfunctioning or recording error. In the proposed model, we have used the most commonly used outlier elimination method; the ‘Grubbs test’.

The formula to detect the outlier is as under

$$G = \frac{\max_{i = 1\ldots n}^{|xi - x|}}{\sigma} \tag{4}$$

The blueprint of missing value acquisition is given in Fig. 3 that removes the noise by invoking the clustering approach.

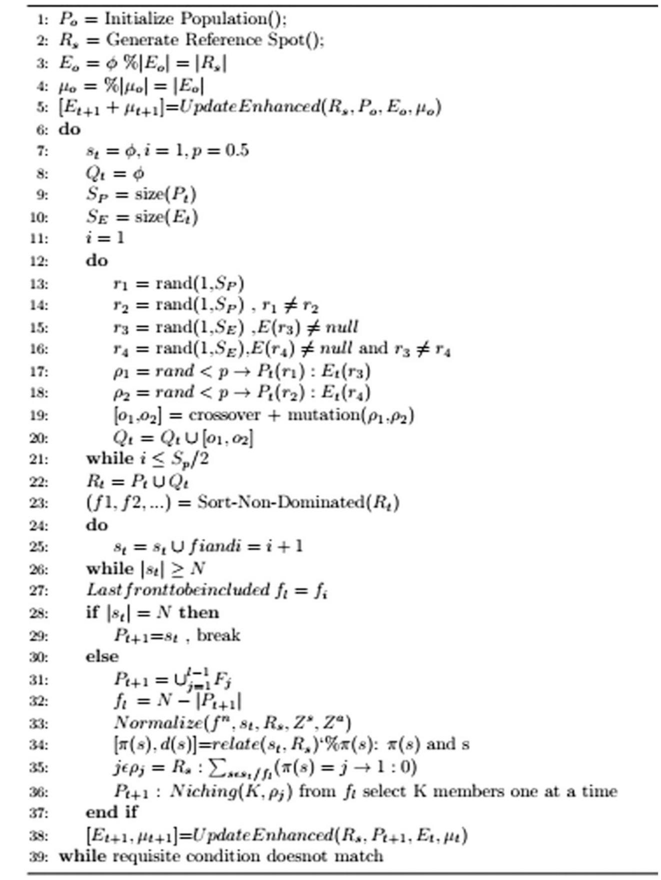
3.2 Algorithm description

The fuzzy GA is applied to the dataset with two fitness measures that are interpretability and accuracy. First, the numbers of fuzzy rules are extracted from the dataset. This extraction is done by passing the data from a novel and user-defined membership function of the fuzzy system.

Then, a multi-objective genetic evolutionary algorithm is used to search for Pareto optimal solutions with respect to the maximisation of both interpretability and accuracy. The proposed framework improves the performance of a non-dominant sorting GA by commissioning a bi-mechanism. First commissioning of an enhanced archive to safeguard enhanced members of the population, which have been removed by the algorithm's selection procedure. Second, is the selection mechanism of a new parent to advance the diversity of the parent population. Moreover, the member function enhancement and customisation are also carried out by developing a comprehensive and specific fuzzy inference engine that is going to focus on two main objectives interpretability and accuracy. The implementation is carried out in MATLAB and a fuzzy inference engine is developed that interacts and interprets the fuzzy rules in the rule base of the fuzzy system. Next, tuning of a membership function of FRBC is done using the GA. Parameters of fuzzy membership functions and rules in the rule base are determined using GA. The recombine and mutation process will produce an ideal offspring population by randomly selecting the parent population along with the enhanced population. The further tournament selection mechanism is applied to filter the fittest for the Pareto-optimal solution as specified in algorithm 1 (see Fig. 4).

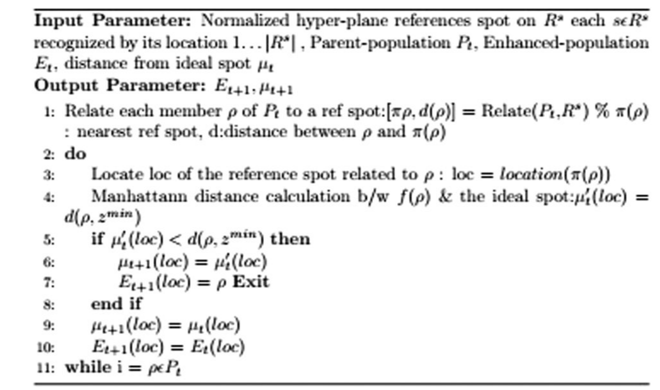
The update enhanced procedure will input reference points on the normalized hyper plane, each recognized by its parent population, enhanced population distance to the ideal point by finding the Manhattan distance and its location as specified in Algorithm 2 (see Fig. 5).

The selection of parent populations is carried out by normalising the probability from the current and enhanced population archive. To improve the diversity, minimising the chance of selecting the parent that is affiliated with the same reference point at most one enhanced member is used.



```
$P_o$ = Initialize Population();
$R_s$ = Generate Reference Spot();
$E_o = \phi$ %$|E_o| = |R_s|$
$\mu_o =$ %$|\mu_o| = |E_o|$
$[E_{t+1} + \mu_{t+1}]$=$UpdateEnhanced(R_s, P_o, E_o, \mu_o)$
do
    $s_t = \phi, i = 1, p = 0.5$
    $Q_t = \phi$
    $S_P$ = size($P_t$)
    $S_E$ = size($E_t$)
    $i = 1$
    do
        $r_1$ = rand(1,$S_P$)
        $r_2$ = rand(1,$S_P$) , $r_1 \neq r_2$
        $r_3$ = rand(1,$S_E$) ,$E(r_3) \neq null$
        $r_4$ = rand(1,$S_E$),$E(r_4) \neq null$ and $r_3 \neq r_4$
        $\rho_1 = rand < p \rightarrow P_t(r_1) : E_t(r_3)$
        $\rho_2 = rand < p \rightarrow P_t(r_2) : E_t(r_4)$
        $[o_1,o_2]$ = crossover + mutation($\rho_1$,$\rho_2$)
        $Q_t = Q_t \cup [o_1, o_2]$
    while $i \leq S_p/2$
    $R_t = P_t \cup Q_t$
    $(f1, f2, ...)$ = Sort-Non-Dominated($R_t$)
    do
        $s_t = s_t \cup fiandi = i + 1$
    while $|s_t| \geq N$
    $Last fronttobeincluded\ f_l = f_i$
    if $|s_t| = N$ then
        $P_{t+1}$=$s_t$ , break
    else
        $P_{t+1} = \cup_{j=1}^{l-1} F_j$
        $f_l = N - |P_{t+1}|$
        $Normalize(f^n, s_t, R_s, Z^s, Z^a)$
        $[\pi(s), d(s)]$=$relate(s_t, R_s)$'%$\pi(s)$: $\pi(s)$ and s
        $j\epsilon\rho_j = R_s : \sum_{s\epsilon s_t/f_l}(\pi(s) = j \rightarrow 1 : 0)$
        $P_{t+1}$ : $Niching(K, \rho_j)$ from $f_l$ select K members one at a time
    end if
    $[E_{t+1}, \mu_{t+1}]$=$UpdateEnhanced(R_s, P_{t+1}, E_t, \mu_t)$
while requisite condition doesnot match
```

Fig. 4 *Algorithm 1. Enhanced NSGA*



```
Input Parameter: Normalized hyper-plane references spot on $R^s$ each $s\epsilon R^s$
recognized by its location 1...$|R^s|$ , Parent-population $P_t$, Enhanced-population
$E_t$, distance from ideal spot $\mu_t$
Output Parameter: $E_{t+1}, \mu_{t+1}$
Relate each member $\rho$ of $P_t$ to a ref spot:$[\pi\rho, d(\rho)]$ = Relate($P_t$,$R^s$) % $\pi(\rho)$
    : nearest ref spot, d:distance between $\rho$ and $\pi(\rho)$
do
    Locate loc of the reference spot related to $\rho$ : loc = $location(\pi(\rho))$
    Manhattann distance calculation b/w $f(\rho)$ & the ideal spot:$\mu_t'(loc) = d(\rho, z^{min})$
    if $\mu_t'(loc) < d(\rho, z^{min})$ then
        $\mu_{t+1}(loc) = \mu_t'(loc)$
        $E_{t+1}(loc) = \rho$ Exit
    end if
    $\mu_{t+1}(loc) = \mu_t(loc)$
    $E_{t+1}(loc) = E_t(loc)$
while i = $\rho\epsilon P_t$
```

Fig. 5 *Algorithm 2. Update enhanced mechanism*

Table 1 Fuzzy rule base Wisconsin breast cancer dataset

No.	Fuzzy rules for classification
1	IF *bare nuclei are small & cell size is small* THEN *benign*
2	IF *bare nuclei are large & cell size is large* THEN *malignant*

4 Results

The experimental setup focuses on the result generation and comparison with one of the latest research of the multi-objective genetic fuzzy optimisation approach presented in [23]. It is opined from the literature that the non-dominated sorting algorithm performs better when the following variation operators are selected optimally.

- Simulated binary crossover
- Differential evaluation operator
- Polynomial mutation (random) to produce the offspring.

The proposed framework has focused on Wisconsin Breast Cancer Dataset, which is multivariate and obeys normal distribution rule by applying the random oversampling technique to solve class imbalance issues. The malignant cases occupy 35% of the dataset and the remaining 65% represent the benign cases. The clusters are found through the random sampling technique. Selecting the right number of features is the most vital concept of data sciences and is the first stage of model designing. Unnecessary/irrelevant features in the model design can hamper the accuracy of the model. If there are more number of features there will be more number of rules to handle. Our proposed model unearths the trade-off between accuracy and interpretability by selecting an appropriate number of features and the results validate that with the right selection of candidate features improvement in accuracy is achieved. More features generate more computational complexity to the model so this pre-processing step is significant for classification algorithms. To expedite our approach, a more detail experiment is carried out on the uni-linear test split of the repository. The learning process of the experiment is conducted through 1K generations and the preliminary population has 10K individuals. The recursive crossover and mutation probabilities are adjusted to 0.7 and 0.5 with a selective pressure 2. To obtain the best Pareto-front approximation on the multi-objective optimisation approach an enhanced non-dominated sorting algorithm is commissioned. The major objectives of the Pareto-front are interpretability and accuracy of the derived solution.

The experiment uses a k-fold cross-validation method. To minimise the associated favouritism with the random splitting of the dataset into training and testing chunks, 10-fold cross-validation is used. The complete data repository is divided into k-subsets, which are mutually exclusive. To further minimise the favouritism equal number of records with similar class distribution is divided. The learning routine is reiterated *k* times in which each one of the *k* chunks is utilised as test data and the left behind chunk as learning data. In every iteration, the solution generated by MEAF with the highest interpretability and accuracy is selected and their average is calculated. Every experiment is reiterated ten times with different spread and the average results from *k* runs are computed. Similarly, for the selection of initial value, each experiment is reiterated for ten times to avoid favouritism in the initialisation process that makes 100 experiments for a given *k*. Each subset has the same class distribution with an equal number of records. For a given number of *k* hundred, cross-validation experiments are conducted to compare it with our benchmark research. The learning to test ratios that are commissioned in this research are (9:1), (4:1), (2:1) and (1:1). The fuzzy rule base from the best Pareto-front solution achieved is depicted in Table 1. The membership function graph is depicted in Fig. 6 for the bare nuclei attribute of the breast cancer dataset that has five distributions.

Similarly, the decision tree and its rule base are extracted from another very important and relevant research [57] of this area in which rules are generated using a fuzzy minimum–maximum neural network combined with a random forest model, classification, regression tree model and the rule base which is shown in Figs. 7 and 8, respectively.

A brief comparison of the quantitative results of the breast cancer dataset is elaborated in Table 2. The results show remarkable improvement in accuracy and interpretability. The graphical elaboration also highlights the same phenomena. Our research has evaluated the context of disease prediction with the latest norms of the research available and the results achieved in Figs. 9–11 indicate that our framework has achieved remarkable accuracy and interpretability while also addressing the sensitivity and specificity issue that is the interim objectives achieved in the process of achieving the main objectives. It is important to mention here that the computational complexity of the GA mainly depends on the fitness function, selection operator, and variation operator. Our approach uses a tournament selection procedure that is further optimised by an enhanced update mechanism. The algorithm yields $\boldsymbol{O(n)}$ computational complexity in the best case scenario when all the above-mentioned parameters are selected optimally which is quite desirable in the complex evolutionary algorithms.

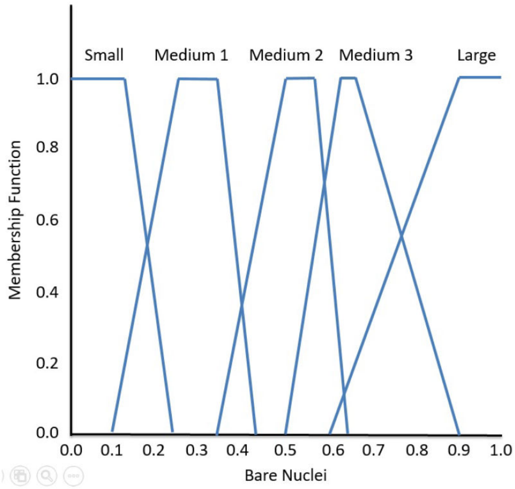


Fig. 6 *Membership function for attribute bare nuclei*

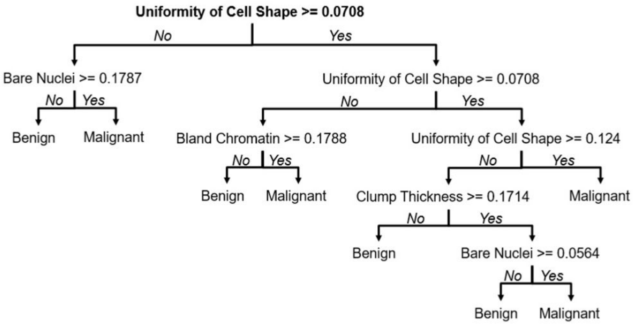
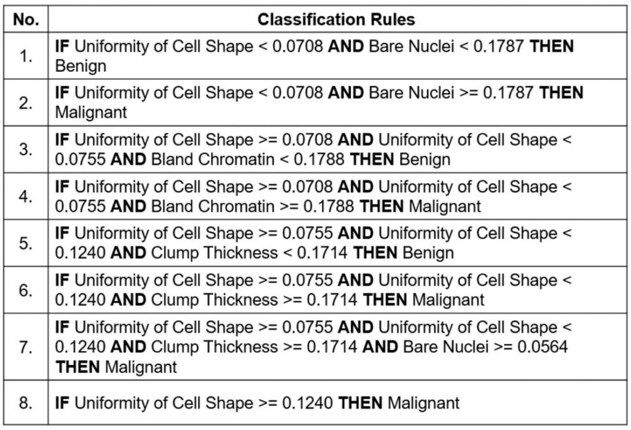

Uniformity of Cell Shape >= 0.0708
No
Yes
Bare Nuclei >= 0.1787
Uniformity of Cell Shape >= 0.0708
No
Yes
No
Yes
Benign
Malignant
Bland Chromatin >= 0.1788
Uniformity of Cell Shape >= 0.124
No
Yes
No
Yes
Benign
Malignant
Clump Thickness >= 0.1714
Malignant
No
Yes
Benign
Bare Nuclei >= 0.0564
No
Yes
Benign
Malignant

Fig. 7 *Decision tree for attribute bare nuclei*



No.	Classification Rules
1.	**IF** Uniformity of Cell Shape < 0.0708 **AND** Bare Nuclei < 0.1787 **THEN** Benign
2.	**IF** Uniformity of Cell Shape < 0.0708 **AND** Bare Nuclei >= 0.1787 **THEN** Malignant
3.	**IF** Uniformity of Cell Shape >= 0.0708 **AND** Uniformity of Cell Shape < 0.0755 **AND** Bland Chromatin < 0.1788 **THEN** Benign
4.	**IF** Uniformity of Cell Shape >= 0.0708 **AND** Uniformity of Cell Shape < 0.0755 **AND** Bland Chromatin >= 0.1788 **THEN** Malignant
5.	**IF** Uniformity of Cell Shape >= 0.0755 **AND** Uniformity of Cell Shape < 0.1240 **AND** Clump Thickness < 0.1714 **THEN** Benign
6.	**IF** Uniformity of Cell Shape >= 0.0755 **AND** Uniformity of Cell Shape < 0.1240 **AND** Clump Thickness >= 0.1714 **THEN** Malignant
7.	**IF** Uniformity of Cell Shape >= 0.0755 **AND** Uniformity of Cell Shape < 0.1240 **AND** Clump Thickness >= 0.1714 **AND** Bare Nuclei >= 0.0564 **THEN** Malignant
8.	**IF** Uniformity of Cell Shape >= 0.1240 **THEN** Malignant

Fig. 8 *Rulebase for Wisconsin breast cancer dataset*

It is evinced from the results that the most transparent and straight forward in terms of accuracy and interpretability is solutions 3 and 4. The solution involves two attributes and two fuzzy rule sets that are arranged in one attribute for one rule giving the highest accuracy that is further illustrated in Table 2. To validate further, MEAF is compared with the schemes of similar nature based on multi-objective, i.e. accuracy, sensitivity, and specificity. A comparison is carried out with FRBC system (FRBCS) presented in [23] and fuzzy min–max-classification and regression tree-random forest presented in [57] and depiction is displayed in Fig. 12.

The results show considerable superiority over the benchmark state of the art literature available.

The main threat to the validity of this research lies in the collection of an appropriate data sample. The field of data sciences relies on the appropriate selection of data. Our proposed scheme mitigates this threat by carefully formalising samples that obey normal distribution rules. This will disqualify all the biases that can be added to converge the model towards a particular belief. Another threat to the validity of our research is to authenticate the results of the proposed model such that the outcome does not converge towards a specific class distribution. To mitigate this we have applied a ten-fold cross-validation scheme that ensures the authenticity of the research through rigorous experimentation. To validate the research further the comparison of the research is carried out with the research of a similar nature. Furthermore, multiple datasets/data samples are invoked to check the diversity of the proposed model. Our research has applied the above-stated rules to mitigate the threats to validity.

5 MediHealth: clinical decision-making application

As stated above that our proposed framework ‘MEAF’ will be supplemented through a software solution with multiple components called ‘MediHealth’. MediHealth is developed with the ideology of supporting the stakeholders in disease diagnosis. This diagnosis has been made much simpler and easier by implementing a user-friendly interface without compromising privacy and data integrity.

The user of the software uses the login screen shown in Fig. 13 with its valid login details to access the system's dashboard.

The dashboard is specially designed by keeping ease of access and simplicity in consideration and is shown in Fig. 14. In order to give stakeholders more clarity, the MediHealth software architecture diagram is shown in Fig. 15. The diagram shows the potential user of the software along with their access rights. One of the major components ‘medical data repository’ is also earmarked as all the patient's data and model training data is stored in this repository.

5.1 Users of mediHealth application

MediHealth application has defined software administrators, physicians, and patients as its main users. These users can use the software according to their assigned privileges and role. The software administrator is the power user with maximum rights. The software administrator has access to data gathering and pre-processing module along with model training and generation module. The physician has the access right to use disease prediction and report generation modules. However, the patient has a viewing role that pertains to information related to his health condition. This ensures data integrity and privacy.

5.2 Modules of MediHealth

MediHealth semantic data repository software is divided into four major modules/components, which are data gathering and preprocessing, prediction model training and generation, disease prediction and reports and returns module. In the next section, a brief explanation is given about each module.

5.2.1 Data gathering and pre-processing module: This module sanctions software administrators to ingest the patient's data. The data should comply with the integrity checks and some predefined rules and settings. Pre-processing refers to a process where various data validation and settings are done using software routines that purifies the data from missing values and outliers. This is to make sure that entered data has the right format for further operations that are necessary to make this application work efficiently. All attributes of ‘patient’ are saved in the data repository according to the defined format. In order to uniquely identify each patient in all subsequent references to avoid any conflicting tuples, a unique identification number is used, i.e. patient ID (PID).

For better understanding and clarity, the relevant interface is shown in Fig. 16. It is evident that as soon as the patient's data is entered into the ‘MediHealth’ software, PID is generated automatically through a procedure call. Once this is done, the next step is to pre-process this data for further refinement. The application also ensures privacy and creates/manages various users along with their specified roles. The access rights are also defined

Table 2 Empirical evidence of accuracy and interpretability Wisconsin cancer dataset

Sr. No.	Interpretability measures				Objective function complements			
					FRBCS [23]		MEAF	
	R	*N*(atr)	*N*(fs)	*N*(atr/*R*)	1 − *Q*(int) = *Q*(cplx)	1 − *Q*(acc − lrn)	1 − *Q*(int) = *Q*(cplx)	1 − *Q*(acc − lrn)
i.	2	1	1	1	0	0.0746	0.0358	0.0789
ii.	2	2	2	1	0.0492	0.0396	0.0512	0.0493
iii.	3	2	3	1	0.0568	0.0365	0.0594	0.0411
iv.	4	3	4	1	0.106	0.0301	0.1121	0.03255
v.	5	3	5	1	0.1136	0.0285	0.1149	0.02947
vi.	5	3	5	1.2	0.1219	0.0269	0.1311	0.0286
vii.	7	3	6	1.2	0.1331	0.0253	0.1432	0.0259
viii.	8	3	7	1.3	0.1444	0.0238	0.1563	0.02254
ix.	7	4	7	1.4	0.1883	0.0222	0.1942	0.02486
x.	6	4	8	1.5	0.1988	0.0206	0.2014	0.02156
xi.	8	4	9	1.5	0.2064	0.019	0.2473	0.0204
xii.	6	5	9	1.6	0.255	0.0174	0.2584	0.01963
xiii.	8	5	10	1.6	0.2608	0.0158	0.2891	0.01594
xiv.	8	6	11	1.7	0.3153	0.0142	0.3467	0.01498

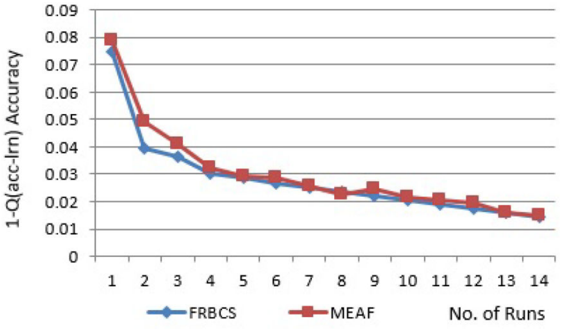


Fig. 9 *Accuracy graph of Wisconsin breast cancer dataset*

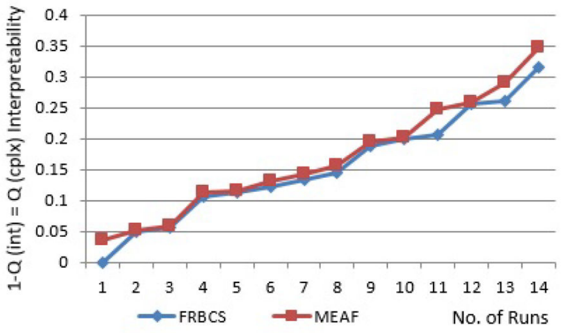


Fig. 10 *Interpretability graph of Wisconsin breast cancer dataset*

according to the role and they are set by software administrator dynamically. After the data entry of the patient, the admin of the MediHealth application follows the process by clicking on the 'Add New Patient' button. The entered data is ingested in the repository in a defined format. The permanent storage of patient data makes the system non-volatile and the data remains available to the prediction model for future training and references.

5.2.2 Prediction model training and generation: MediHealth software is defined in a modular manner. After the successful entry of the patient into the system, the next step is to proceed to prediction model training and generation. The vital function of this module is to perform various operations on the historical data/datasets already stored in the medical data repository after outlier and missing value elimination. Based on these values, our proposed framework 'MEAF' will be invoked to train itself for disease prediction with accuracy. The interface of this module is shown in Fig. 17. Load DB button will load the above-mentioned training data from the repository and the train model button will invoke a software routine that trains the MEAF for disease prediction.

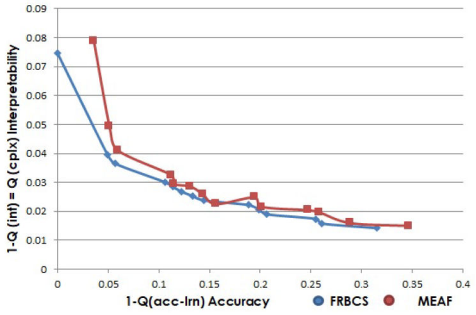


Fig. 11 *Accuracy interpretability trade-off of Wisconsin breast cancer dataset*

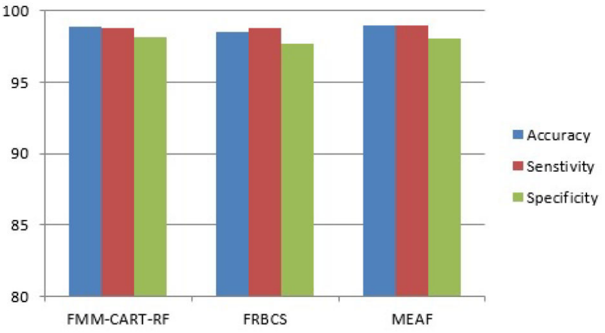
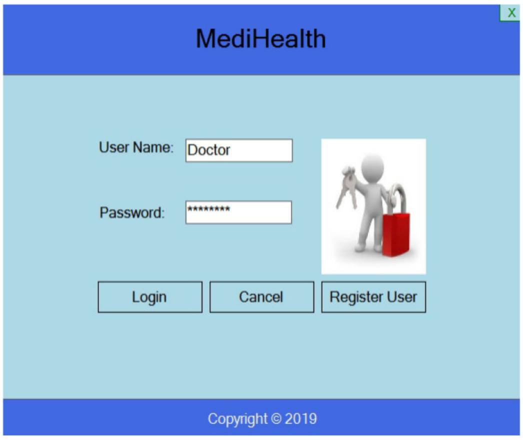


Fig. 12 *Accuracy, sensitivity and specificity compared with state-of-the-art techniques*


MediHealth

User Name: Doctor

Password: ********

Copyright © 2019

Fig. 13 *Login screen of the application*

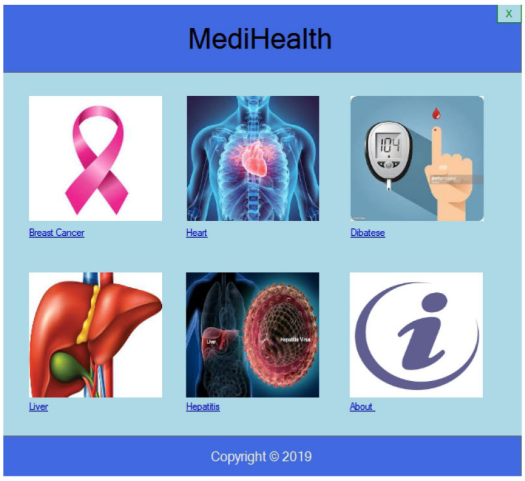


Fig. 14 *Dashboard of application*

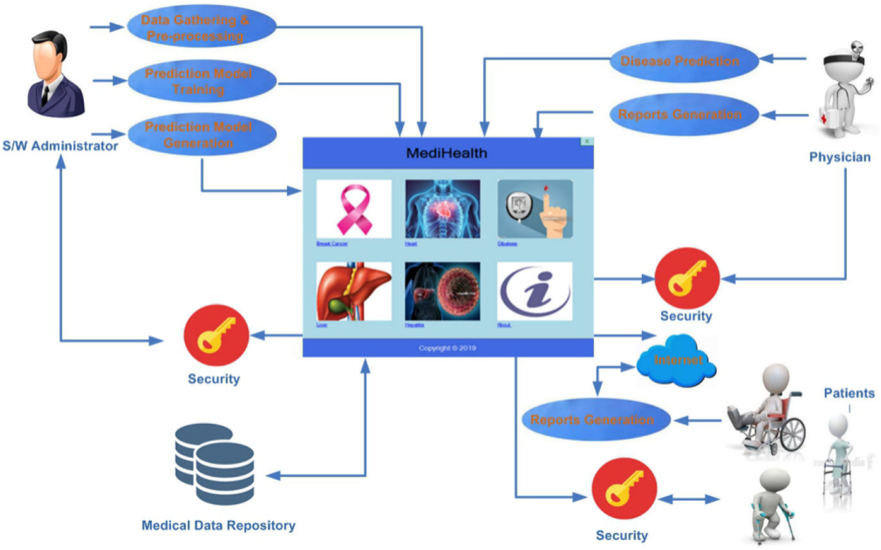
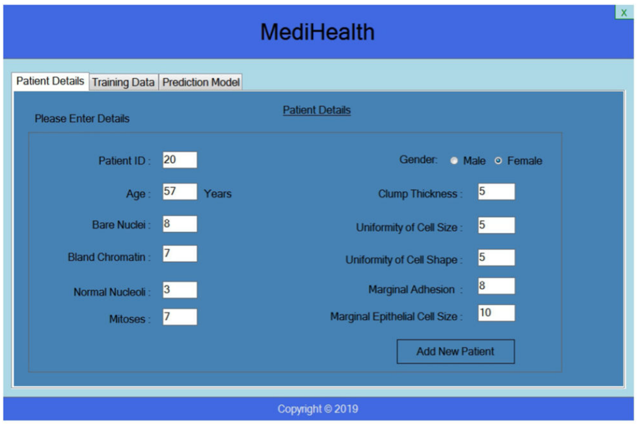


Fig. 15 *Architecture of MediHealth application*



MediHealth

Patient Details | Training Data | Prediction Model

Please Enter Details

Patient Details

Patient ID : 20

Gender: ○ Male ● Female

Age : 57 Years

Clump Thickness : 5

Bare Nuclei : 8

Uniformity of Cell Size : 5

Bland Chromatin : 7

Uniformity of Cell Shape : 5

Normal Nucleoli : 3

Marginal Adhesion : 8

Mitoses : 7

Marginal Epithelial Cell Size : 10

Add New Patient

Copyright © 2019

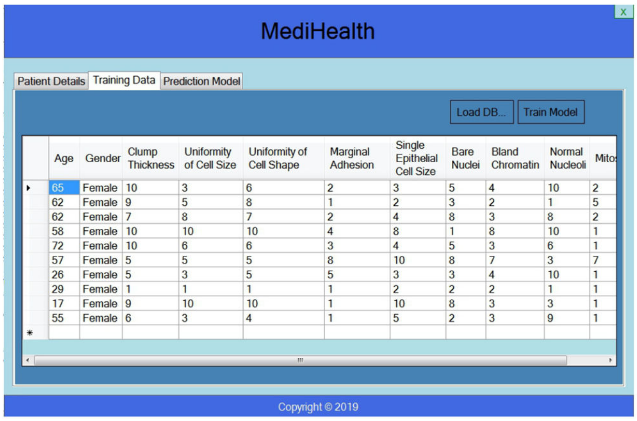
Fig. 16 *Enter patient interface*

5.2.3 Disease prediction module: Since the MEAF framework has been designed to help medical professionals and researchers to predict disease by improving accuracy and interpretability simultaneously using evolutionary algorithms. Its invocation is significant in a way that it helps the practitioners/physicians of the remote area to gain maximum benefit from the model. These remote areas have no or limited access to specialist doctors. Module name 'disease prediction' provides the refined output of data that will be used by the doctor to predict disease by augmenting their expert knowledge of the domain with prediction results of the model. Fig. 18 shows this working as part of the MediHealth application. This easy to use interface will require the doctor to load particular data of a patient and after clicking on 'Diagnose Patient', the doctor can see whether the patient has symptoms of a disease or any further tests are to be conducted before improved diagnosis. Fig. 19 shows the confidence/belief level of MEAF on the prediction results. Medihealth application also generates various alerts based on filtered results to support the decision-making process. Thus, providing all necessary data refinements necessary for decision-making to predict a particular disease.

5.2.4 Reports and return: One of the key area/module of the MediHealth application is the 'Reports and Returns'. This module is responsible for generating a specific and targeted information based on various operations that are performed on available data. This report is based on the individual patient and its historical data values about the prediction of a particular disease. Reports and return module allows the patients to generate their own report by successfully log into the MediHealth application through the login interface. For simplicity, the module declares the patient as healthy or sick.

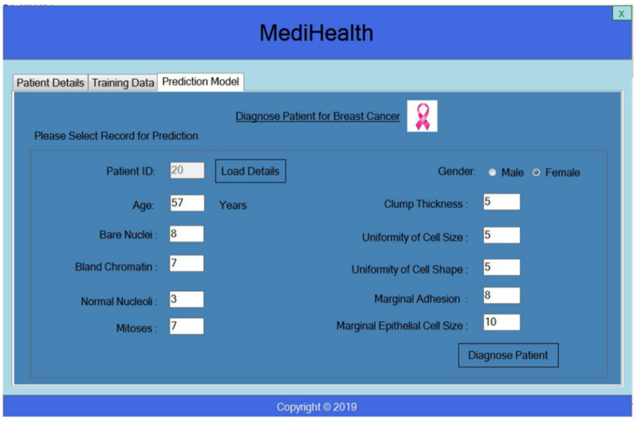
6 Conclusion and future work

This research proposes a framework that has the ability to apply evolutionary algorithms with a multi-objective evolutionary scheme using fuzzy genetic. A framework named 'MEAF' is tested with a breast cancer dataset. Our comparative analysis with other classifiers/frameworks showed that MEAF has achieved the



	Age	Gender	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mito
▸	65	Female	10	3	6	2	3	5	4	10	2
	62	Female	9	5	8	1	2	3	2	1	5
	62	Female	7	8	7	2	4	8	3	8	2
	58	Female	10	10	10	4	8	1	8	10	1
	72	Female	10	6	6	3	4	5	3	6	1
	57	Female	5	5	5	8	10	8	7	3	7
	26	Female	5	3	5	5	3	3	4	10	1
	29	Female	1	1	1	1	2	2	2	1	1
	17	Female	9	10	10	1	10	8	3	3	1
	55	Female	6	3	4	1	5	2	3	9	1
*											

Fig. 17 *Model training and generation interface*



Please Select Record for Prediction

Patient ID: 20

Gender: ○ Male ● Female

Age: 57 Years

Clump Thickness : 5

Bare Nuclei : 8

Uniformity of Cell Size : 5

Bland Chromatin : 7

Uniformity of Cell Shape : 5

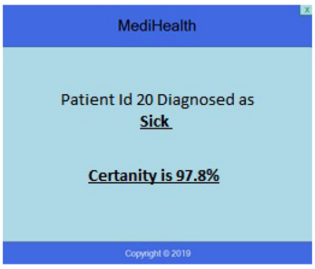
Normal Nucleoli : 3

Marginal Adhesion : 8

Mitoses : 7

Marginal Epithelial Cell Size : 10

Fig. 18 *Patient diagnoses interface available to the doctor*



Patient Id 20 Diagnosed as
Sick

Certanity is 97.8%

Fig. 19 *Result interface displaying a model belief*

highest accuracy and interpretability when compared with distinguished existing techniques used to address some problem domain. The working of MEAF has been supplemented further by designing and implementing an application named 'MediHealth'. This application has been implemented by incorporating the latest graph database technology so that decision-making is easy and accurate. This application is developed by keeping in mind ease of use for medical doctors to predict the disease by looking at refined data of various patients including their historical perspective. This application also allows patients to view their report in an easy and understandable way. For the future, researchers and developers can work on providing a publicly available web server via the internet to access the application.

7 References

[1] Ngai, E.W., Xiu, L., Chau, D.C.: 'Application of data mining techniques in customer relationship management: A literature review and classification', *Expert Syst. Appl.*, 2009, **36**, (2), pp. 2592–2602

[2] Romero, C., Ventura, S.: 'Educational data mining: a review of the state of the art', *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, 2010, **40**, (6), pp. 601–618

[3] Bellazzi, R., Zupan, B.: 'Predictive data mining in clinical medicine: current issues and guidelines', *Int. J. Med. Inform.*, 2008, **77**, (2), pp. 81–97

[4] Ngai, E.W., Hu, Y., Wong, Y.H., *et al.*: 'The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature', *Decis. Support Syst.*, 2011, **50**, (3), pp. 559–569

[5] Pietraszek, T., Tanner, A.: 'Data mining and machine learning towards reducing false positives in intrusion detection', *Inf. Sec. Tech. Rep.*, 2005, **10**, (3), pp. 169–183

[6] Jiang, D., Tang, C., Zhang, A.: 'Cluster analysis for gene expression data: a survey', *IEEE Trans. Knowl. Data Eng.*, 2004, **11**, pp. 1370–1386

[7] Esfandiari, N., Babavalian, M.R., Moghadam, A.M.E., *et al.*: 'Knowledge discovery in medicine: current issue and future trend', *Expert Syst. Appl.*, 2014, **41**, (9), pp. 4434–4463

[8] Yeh, D.Y., Cheng, C.H., Chen, Y.W.: 'A predictive model for cerebrovascular disease using data mining', *Expert Syst. Appl.*, 2011, **38**, (7), pp. 8970–8977

[9] Mansingh, G., Osei-Bryson, K.M., Reichgelt, H.: 'Using ontologies to facilitate post-processing of association rules by domain experts', *Inf. Sci.*, 2011, **181**, (3), pp. 419–434

[10] Cios, K.J., Moore, G.W.: 'Uniqueness of medical data mining', *Artif. Intell. Med.*, 2002, **26**, (1–2), pp. 1–24

[11] Patel, V.L., Shortliffe, E.H., Stefanelli, M., *et al.*: 'The coming of age of artificial intelligence in medicine', *Artif. Intell. Med.*, 2009, **46**, (1), pp. 5–17

[12] Lavrač, N.: 'Selected techniques for data mining in medicine', *Artif. Intell. Med.*, 1999, **16**, (1), pp. 3–23

[13] Drozdowicz, M., Ganzha, M., Paprzycki, M.: 'Semantically enriched data access policies in eHealth', *J. Med. Syst.*, 2016, **40**, (11), p. 238

[14] Layouni, M., Verslype, K., Sandkkaya, M.T., *et al.*: 'Privacy-preserving telemonitoring for ehealth'. IFIP Annual Conf. on Data and Applications Security and Privacy, Berlin, Heidelberg, 2009, pp. 95–110

[15] Larrucea, X., Santamaria, I., Colomo-Palacios, R.: 'Assessing source code vulnerabilities in a cloud-based system for health systems: OpenNCP', *IET Softw.*, 2019, **13**, (3), pp. 195–202

[16] Werlang, F.C., Custódio, R.F., Vigil, M.A.: 'A user-centric digital signature scheme'. European Public Key Infrastructure Workshop, Berlin, Heidelberg, 2013, pp. 152–169

[17] Kelman, C.W., Bass, A.J., Holman, C.: 'Research use of linked health data a best practice protocol', *Aust. N.Z. J. Public Health*, 2002, **26**, (3), pp. 251–255

[18] Aranguren, M.E., Fernández-Breis, J.T., Dumontier, M.: 'Special issue on linked data for health care and the life sciences', *Semant. Web*, 2014, **5**, (2), pp. 99–100

[19] Tilahun, B., Kauppinen, T.: 'Potential of linked open data in health information representation on the semantic web'. SWAT4LS, Paris, France, 2012

[20] Liyanage, H., Liaw, S.T., Kuziemsky, C., *et al.*: 'The evidence-base for using ontologies and semantic integration methodologies to support integrated chronic disease management in primary and ambulatory care: realist review', *Yearb. Med. Inform.*, 2013, **22**, (1), pp. 147–154

[21] Qu, Y., Adam, B.L., Yasui, Y., *et al.*: 'Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients', *Clin. Chem.*, 2002, **48**, (10), pp. 1835–1843

[22] Samuel, O.W., Asogbon, G.M., Sangaiah, A.K., *et al.*: 'An integrated decision support system based on ANN and fuzzy_AHP for heart failure risk prediction', *Expert Syst. Appl.*, 2017, **68**, pp. 163–172

[23] Gorzałczany, M.B., Rudziński, F.: 'Interpretable and accurate medical data classification–a multi-objective genetic-fuzzy optimization approach', *Expert Syst. Appl.*, 2017, **71**, pp. 26–39

[24] Zolfaghari, R.: 'Diagnosis of diabetes in female population of Pima Indian heritage with ensemble of BP neural network and SVM', *Int. J. Comput. Eng. Manage.*, 2012, **15**, pp. 2230–7893

[25] Pattekari, S.A., Parveen, A.: 'Prediction system for heart disease using Naive Bayes', *Int. J. Adv. Comput. Math. Sci.*, 2012, **3**, (3), pp. 290–294

[26] Sinha, P., Sinha, P.: 'Comparative study of chronic kidney disease prediction using KNN and SVM', *Int. J. Eng. Res. Technol.*, 2015, **4**, (12), pp. 608–612

[27] Malav, A., Kadam, K., Kamat, P.: 'Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy', *Int. J. Eng. Technol.*, 2017, **9**, (4), pp. 3081–3082

[28] Chimwayi, K.B., Haris, N., Caytiles, R.D., *et al.*: 'Risk level prediction of chronic kidney disease using neuro-fuzzy and hierarchical clustering algorithm(s), 2017

[29] Bhargava, N., Jain, A., Kumar, A., *et al.*: 'Clustered comparative analysis of security sensor discrimination data'. Proc. First Int. Conf. on Information Technology and Knowledge Management, 2017, pp. 29–33

[30] Rustempasic, I., Can, M.: 'Diagnosis of Parkinson's disease using fuzzy C-means clustering and pattern recognition', *Southeast Eur. J. Soft Comput.*, 2013, **2**, (1), pp. 42–49

[31] Rao, P.S., Devi, T.U.: 'Applicability of apriori based association rules on medical data', *Int. J. Appl. Eng. Res.*, 2017, **12**, (20), pp. 9451–9458

[32] Smitha, T., Sundaram, V.: 'Association models for prediction with apriori concept', *Int. J. Adv. Eng. Technol.*, 2012, **5**, (1), p. 354

[33] Ramaraj, E., Venkatesan, N.: 'An efficient pattern mining analysis in health care database', 2009

[34] Li, J.S., Zhang, Y.F., Tian, Y.: 'Medical big data analysis in hospital information system', *Big Data Real-World Appl.*, 2016, p. 65

[35] Ball, G., Mian, S., Holding, F., *et al.*: 'An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers', *Bioinformatics*, 2002, **18**, (3), pp. 395–404

[36] Tsai, M.H., Lai, C.H., Yu, S.S.: 'A statistical and learning based oncogene detection and classification scheme using human cDNA expressions for ovarian carcinoma', *Expert Syst. Appl.*, 2011, **38**, (8), pp. 10066–10074

[37] Bernd, T., Kleutges, M., Kroll, A.: 'Nonlinear black box modelling–fuzzy networks versus neural networks', *Neural Comput. Appl.*, 1999, **8**, (2), pp. 151–162

[38] Tan, T.Z., Quek, C., Ng, G.S.: 'Ovarian cancer diagnosis by hippocampus and neocortex-inspired learning memory structures', *Neural Netw.*, 2005, **18**, (5–6), pp. 818–825

[39] Deb, K., Pratap, A., Agarwal, S.*, et al.*: 'A fast and elitist multiobjective genetic algorithm: NSGA-II', *IEEE Trans. Evol. Comput.*, 2002, **6**, (2), pp. 182–197

[40] Bechikh, S., Datta, R., Gupta, A.: '*Recent advances in evolutionary multi-objective optimization*', vol. 20' (Springer International Publishing, AG, Switzerland, 2016)

[41] Deb, K.: 'Multi-objective optimisation using evolutionary algorithms: an introduction', in '*Multi-objective evolutionary optimisation for product design and manufacturing*' (Springer, London, 2011), pp. 3–34

[42] Deb, K., Jain, H.: 'An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints', *IEEE Trans. Evol. Comput.*, 2013, **18**, (4), pp. 577–601

[43] Yuan, Y., Xu, H., Wang, B.: 'An improved NSGA-III procedure for evolutionary many-objective optimization'. Proc. 2014 Annual Conf. on Genetic and Evolutionary Computation, Vancouver, Canada, 2014, pp. 661–668

[44] Fraccaro, P., Plastiras, P., Dentone, C.*, et al.*: 'Behind the screens: clinical decision support methodologies–a review', *Health Policy Technol.*, 2015, **4**, (1), pp. 29–38

[45] Ibrahim, A., Rahnamayan, S., Martin, M.V.*, et al.*: 'EliteNSGA-III: an improved evolutionary many-objective optimization algorithm'. 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, Canada, 2016, p. pp. 973–982

[46] Seada, H., Deb, K.: 'U-NSGA-III: A unified evolutionary algorithm for single, multiple, and many-objective optimization'. COIN Report, 2014, 2014022

[47] Distilled, N.: '*A brief guide to the emerging world of polyglot persistence by Pramod J. Sadalage and Martin owler*' (Addison-Wesley, Boston, USA, 2013)

[48] Gandon, F., Schreiber, A.T.: 'Rdf 1.1 xml syntax', 2014

[49] Gorzalczany, M.B.: 'Computational intelligence systems and applications: neuro-fuzzy and fuzzy neural synergisms', *Physica*, 2012, **86**

[50] Gorzalczany, M.B., Gradzki, P.: 'A neuro-fuzzy-genetic classifier for technical applications'. Proc. IEEE Int. Conf. on Industrial Technology 2000 (IEEE Cat. No. 00TH8482), Goa, India, 2000, vol. 1, pp. 503–508

[51] Rutkowski, L.: '*Flexible neuro-fuzzy systems: structures, learning and performance evaluation*', vol. 771 (Springer Science & Business Media, 2006)

[52] Kaymak, U., Babuska, R.: 'Compatible cluster merging for fuzzy modelling'. Proc. 1995 IEEE Int. Conf. on Fuzzy Systems, Yokohama, Japan, 1995, vol. 2, pp. 897–904

[53] Setnes, M., Roubos, H.: 'GA-fuzzy modeling and classification: complexity and performance', *IEEE Trans. Fuzzy Syst.*, 2000, **8**, (5), pp. 509–522

[54] Cordón, O., Herrera, F., Gomide, F.*, et al.*: 'Ten years of genetic fuzzy systems: current framework and new trends'. Proc. Joint 9th IFSA World Congress and 20th NAFIPS Int. Conf. (Cat. No. 01TH8569), Vancouver, Canada, 2001, vol. 3, pp. 1241–1246

[55] Cordón, O., Herrera, F., Hoffmann, F.*, et al.*: '*Genetic fuzzy systems: evolutionary tuning and learning of fuzzy knowledge bases*' vol. **19** (World Scientific, Singapore, 2001)

[56] Jin, Y., Sendhoff, B.: 'Pareto-based multiobjective machine learning: an overview and case studies', *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, 2008, **38**, (3), pp. 397–415

[57] Seera, M., Lim, C.P.: 'A hybrid intelligent system for medical data classification', *Expert Syst. Appl.*, 2014, **41**, (5), pp. 2239–2249