UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DIPARTIMENTO DI INFORMATICA

PhD Program in Computer Science and Mathematics
XXXIV cycle

Research Project

PhD Student: Gianluca Zaza

Supervisor: Prof. Giovanna Castellano

Coordinator: Prof. Maria F. Costabile

PhD student signature Gianluca Zaza

Supervisor signature

1. Research title:

Computational Intelligence and Applications in Healthcare.

2. Research area:

Computational Intelligence, Fuzzy Modelling, Medical Informatics, Medical imaging, Personal Healthcare, Pervasive Monitoring.

3. Research motivation and objectives

Medicine is one of the fastest growing fields when compared to other computer-aided technologies. In recent years, healthcare systems have seen exponential grow up due to the increase in technology. In fact, today, thanks to devices such as a smartwatch, a wearable device, a webcam etc. all people can check their own health status independently. All these sensors and devices together have led to new scenarios, thus creating new paradigms such as pervasive monitoring and personalized medicine. These sensors produce huge amounts of data that can be sent to physicians through the network. The physicians should be able to examine and analyse the data of their patients in a timely manner by monitoring them remotely. Traditional data analysis techniques are unsuitable to extract useful information from these data, thus automatic mechanisms capable to deal with large collections of data are necessary. Intelligent data analysis combines the human expertise, in this field the physician, and computational models for accurate and in-depth data analysis.

The physicians use an information system during the diagnosis process. This information system either helps them for recording the data or supports the decision-making process. In both ways, the huge volume and different types of data makes the decision-making process complex and uncertain. Intelligent data analysis (IDA) aims at combining human expertise and computational models for advanced data analysis, in order to narrow the gap between data gathering and their comprehension. In the medical field, more than in others, this interaction is mandatory: on the one hand the experts need automatic tools to transform raw and complex data into easily interpretable information, on the other hand algorithm outputs alone are not sufficient for medical diagnosis, since expert knowledge is needed to understand them.
Traditional machine learning methods are very accurate, but are also non-transparent models, i.e. it is not clear what information in the input data makes them actually arrive at their decisions. Hence, they are not acceptable in medical diagnosis where every decision should be made interpretable for appropriate validation by a human expert. The use of explainable models which can be interpreted and verified by medical experts is an absolute necessity in healthcare. Moreover, in the medical domain both patient information and the reasoning used by the physician to extract conclusions about patients' health are inherently uncertain and vague. Hence the representation of medical knowledge and decision making in the presence of uncertainty and imprecision, are of fundamental importance to derive suitable models for medical decision making.

Among CI paradigms, Fuzzy logic is a powerful tool for dealing with uncertainty and develop interpretable models for medical decision support. Indeed, fuzzy systems use linguistic terms to represent the patients' symptoms, and a fuzzy inference mechanism to derive a suggestion about the health status of the patient. The domain knowledge is embedded into the system in form of fuzzy rules that are easy to read and understand.

In this project, the focus will be on investigating the power of fuzzy methods to deal with uncertainty and develop interpretable models in medical problems. The main goal of the research will be the development of methods able to support the medical diagnosis and to provide new solutions in the field of Personal and Pervasive healthcare.

Different challenges in healthcare will be investigated:
1) the effective and efficient analysis of large collections of clinical data, including physiological signals and images;
2) the automatic creation of diagnostic fuzzy rules from data;
3) the automatic segmentation of medical images to support diagnosis;
4) the development of low-cost and non-invasive solutions for personal healthcare;
5) the development of efficient algorithms to compute the vital parameters and predict health status from remote photoplethysmography signals;
6) the development of algorithms based on deep learning to solve prediction/classifications task in the medical domain.

4. State of the art

Computational Intelligence (CI) is the theory, design, application, and development of biologically and linguistically motivated computational paradigms. The three main pillars of CI are Neural Networks, Fuzzy Systems, and Evolutionary Computation. Computational Intelligence techniques are successfully applied in many domains, healthcare is one of these. In particular, in the recent years we have witnessed an exponential growth of studies that deal with the use of Fuzzy Logic (FL) in the medical domain. Drawing from literature, FL can be defined as the effort of emulating human reasoning model, which uses linguistic variables and concepts, into computer applications. Using FL, computers can process linguistic variables and their degrees of memberships rather than crisp numbers and precise definitions. Especially in medical science, in most of the cases, it is impossible to give exact definitions or descriptions for medical concepts and relationships between these concepts.

Fuzzy logic and fuzzy set theory have proved to be a remarkable tool for building intelligent decision-making systems based on healthcare practitioner knowledge and observations. In particular, fuzzy rule-based systems have been proposed to support diagnosis and monitoring of different diseases.

In [1] Husain et al. propose a fuzzy logic-based home healthcare system for chronic heart disease patients (in stable conditions) for out-of-hospital follow-up and monitoring. The system is composed of a local part and a remote part. The local part deals with the collection of information from the sensors connected to a patient, and the remote part enables storing and distributing the data to remote service seekers like emergency service providers, doctors, and insurance providers. An Arduino-based data aggregator is used to collect the sensor-data before sending to the data processing unit. It also processes the collected raw data to generate meaningful information that can be understood by specialists and doctors. Then, it displays the processed information and sends it to the remote servers. The data are uploaded to cloud through multi-hop wireless communication from the data aggregator and may be accessed and visualized by caregivers. Additionally, the data may be used to detect anomalies and generate alerts. In order to develop the fuzzy scheme, input variables such as blood pressure, blood oxygen saturation, body temperature, ECG, heart rate, respiration rate are introduced, along with their memberships functions. The output variable refers to cardiovascular risk. More than one thousand fuzzy rules for heart disease detection have been identified.

In [2], Baihaqi et al. compare the performance of C4.5, CART, and RIPPER to generate fuzzy rules and apply the imperialist competitive algorithm (ICA) for the optimization of fuzzy membership functions to be used on the fuzzy expert system. There are four steps in this study, the first is data preprocessing that consists of filling missing values, removing outlier, and normalization. The second step is using C4.5, CART, and RIPPER to generate fuzzy rules and extract them after the elaboration of the classification results. In the third step, the fuzzy rules are transformed into a fuzzy set. Lastly, the imperialist competitive algorithm is used to optimize the fuzzy membership functions. Data sets from the Hungarian Institute of Cardiology, Budapest, and the Cleveland Clinic datasets from the University of California at Irvine were used to evaluate the proposed method. The first dataset includes 294 records while the second dataset includes 303 records, so 597 records were used in this research. Among the 76 attributes in the data sets, 13 input attributes consist of age, blood pressure, serum cholesterol, maximum heart rate, sex, type of chest pain, fasting blood sugar, resting ECG, exercise-induced angina, old peak, slope, fluoroscopy, and thallium scan. The output variable is the angiography status. The results of this study show that the combination between C4.5 algorithm and a fuzzy expert system achieves the highest accuracy of 81.82%.

Another work is [3] by El-Sappagh et al. The goal of this paper is to design and implement two frameworks based on two interesting techniques, such as fuzzy analytical hierarchy process (FAHP) and an adaptive neuro-fuzzy inference system (ANFIS). Diagnostic real data of 119 cases infected by chronic viral hepatitis C are used to train and test both the FAHP and ANFIS. The FAHP framework is composed of 4 phases: (1) Determine diagnosis criteria and sub-criteria, (2) determine weight of criteria and sub-criteria, (3) design fibrosis diagnosis formula, and (4) calculate patient disease diagnosis. Criteria and sub-criteria weights are based on the opinions of two domain experts. The ANFIS model is trained and tested by using different subsets of 119 real HCV cases. Results are later compared with the diagnostic conclusions of a medical expert and other three medical and fuzzy techniques. The resulting system achieved a high accuracy of 93.3%.

In [4], Verma et al. present a novel hybrid method for Coronary artery disease (CAD) diagnosis. This paper includes risk factor identification using correlation-based feature subset (CFS) selection with particle swarm optimization (PSO) search method and K-means clustering algorithm. Supervised learning algorithms such as multi-layer perceptron (MLP), multinomial logistic regression (MLR), fuzzy unordered rule induction algorithm (FURIA) and C4.5 are then used to model CAD cases. Clinical data of 335 suspected CAD patients were collected and represented using 26 features. On this dataset MLR achieves highest prediction accuracy of 88.4%. Afterwards, the authors have tested this approach on benchmarked Cleaveland heart disease dataset that is composed of 14 features and 303 instances. In this case, also, MLR outperforms other techniques.

In [5] the aim is to discriminate post-stroke dementia using the electroencephalogram (EEG) signal of 5 patients with vascular dementia (VaD), 15 patients with stroke-related mild cognitive impairment (MCI), and 15 control normal subjects during working memory (WM) task. The authors used independent component analysis (ICA) and wavelet transform (WT) as a hybrid preprocessing approach for EEG artifact removal. They extracted three different features from the cleaned EEG signals: spectral entropy (SpecEn), permutation entropy (PerEn) and Tsallis entropy (TsEn). Afterward, the authors applied two classification schemes, such as support vector machine (SVM) and $\boldsymbol{k}$-nearest neighbors ($\boldsymbol{k}$NN) with fuzzy neighborhood preserving analysis with QR decomposition (FNPAQR) as a dimensionality reduction technique. The dimensionality reduction technique increased the SVM classification accuracy from 82.22% to 90.37%, and from 82.6% to 86.67% for kNN. This study suggests that dimensionality reduction technique and SVM classifier might be useful in discriminating the post-stroke dementia patients using EEG signal analysis.

Medical imaging is also a strong supporting element in medical decision-making. Two-dimensional or three-dimensional medical images are generated by magnetic resonance imaging, computed tomography, digital mammography, positron emission tomography tests. These images can only be usable in medical decision support after a proper processing devoted to extract regions of interest. FL techniques are also well used in this area. Indeed, medical images are full of imprecise conditions and vagueness. Usually medical images present noise and inaccuracies in the definition of edges, so traditional image processing techniques may not be adequate. Fuzzy methods appear to be suitable for medical imaging.

Hien et al. [6] present a new approach to Magnetic resonance image (MRI) edge detection issue. The method proposed is composed of three stages. Firstly, the Semi Translation Invariant Contourlet Transform (STICT) is used to improve the quality of the original MRI. Secondly, the result of the first stage is subjected to image segmentation by using Fuzzy C Means (FCM) clustering method. Finally, the Canny edge detection method is applied to detect the fine edges. The method proposed performs well because the Canny method is applied for ideal input images which are improved quality and segmented into homogeneous regions thanks to the STICTFCM.

In [7] the aim is to develop an automated system for mass segmentation in mammograms using the Fuzzy C-means (FCM) algorithm. The proposed method aims at avoiding the problematic estimation of the cluster number in FCM by selecting as input data the set of pixels which are able to provide the information required to perform the mass segmentation by fixing two clusters only. The Gray Level Occurrence Matrix (GLCM) is used to extract the texture features for getting the optimal threshold, which separates between selected set and the other sets of the pixels that influence on the mass boundary accuracy. The 18 mammogram images used in this study are taken from MiniMIAS database. To evaluate the performance of detection all masses are manually marked by the radiologist based on the visual criteria. The results showed a sensitivity of 86.2%, a specificity of 96.4% and an accuracy of 94.6%.

El-Melegy et al. [8] propose a method for kidney segmentation from Dynamic Contrast Enhanced Magnetic Resonance Images (DCE-MRI). In this method, the fuzzy c-means (FCM) algorithm is combined with a geometric deformable model (level set) method to accurately extract the kidney from its background. The FCM algorithm is applied to the input image and the obtained result is used as the initial contour for the level set method. The evolution of the level set boundary is controlled using the kidney shape prior model and the memberships of the pixels computed using the FCM algorithm. This method has been tested on 40 subjects. For each patient, approximately 80 repeated temporal frames were obtained at 3sec intervals. Thus, the dataset for each patient consists of approximately 80 images. The accuracy of the segmentation results obtained is evaluated by computing the Dice Similarity Coefficient (DSC) between the segmented results and the manually segmented images obtained by a medical expert. The average DSC over the 40 subjects is found to be 0.943 ± 0.023 (mean $\pm$ standard deviation).

5. Problem approach

The following methods will be investigated and applied in the project in order to find a solution to the challenges in medicine mentioned above:

- Fuzzy rule-based systems (FRBSs) allow us to deal with the modeling of systems by building a linguistic model which could be interpreted easily by human beings. Fuzzy Modeling may be divided into:
  1. Linguistic Fuzzy Modeling (LFM), aiming to obtain interpretable fuzzy models. It is mainly developed by means of linguistic (or classic Mamdani) FRBSs. Linguistic FRBSs are based on linguistic rules, which antecedent and consequent make use of linguistic variables comprised of linguistic terms and the associated fuzzy sets defining their meanings;
  2. Precise Fuzzy Modelling (PFM), aiming to obtain fuzzy models with good accuracy. It is mainly developed by means of Takagi–Sugeno FRBSs or by means of approximate FRBSs, which differ from the linguistic ones in the use of fuzzy variables, i.e., fuzzy sets without an associated meaning.

- Hybrid Fuzzy approaches, e.g. neuro-fuzzy approaches to learn diagnostic fuzzy rules from data. In practice, data driven neuro-fuzzy model construction algorithms utilize finite data sets to generate simple models with good explanatory predictive power. Due to the inherent transparency properties of a neuro-fuzzy network, a model construction approach should lead also to a rule extraction process that increases model transparency, as simpler models inherently involve fewer rules which are in turn easier to interpret.

- Deep learning refers to Neural Networks using many hidden neurons and layers—typically more than two—that provide effective high-level abstraction of raw data or images. This high level of abstraction leads to an automatic feature set, which otherwise would required hand-crafted or bespoke features. In domains such as health informatics, the automatic generation of features without human intervention has many advantages. For instance, in medical imaging, it can generate features that are more sophisticated and difficult to elaborate in descriptive means.

- Fuzzy clustering is widely used for medical image segmentation where data elements can fit more than one cluster and membership level is linked with each element. The belongingness of each image pixel to a region is never crisply defined and hence the introduction of fuzziness allows to take into account this uncertainty. Data points are combined to form individual clusters in the feature domain. This is the fundamental principle for Fuzzy C-Means (FCM).

Incorrect FCM clustering results are obtained in case the image is corrupted with noise because of its anomalous feature data.

- Fuzzy cognitive maps (FCM) in medical decision-making, where the highest advantage is to model the interaction between concepts. Nodes for the FCM stand for the concepts that are used to describe the behavior of the system and are connected by signed and weighted arcs, thus representing the causal relationships that exist between the concepts. FCM is an approach for knowledge representation and inference, which is essential to any intelligent system. It may help describe the schematic structure and represent the causal relationships among the elements of a given decision environment, and the inference can be computed by a numeric matrix operation.

Moreover, to develop methods that solve diagnostic (classification) tasks on data from real contexts, we plan to undertake the following steps:
1) Gather data from real contexts about healthy people as well as patients affected by specific diseases;
2) Preprocess and integrate such data;
3) Organize data in a suitable way;
4) Implement algorithms which perform classification/diagnosis tasks.

6. Expected results

The computational models developed during the research activity will be employed in several healthcare applications, such as:

- Diagnosis of cardiovascular disease, to create a Decision Support System (DSS) for the physician. Thus, the physician is assisted while he makes a medical examination. DSSs will allow acquiring information, such as age, sex, heart rate, blood oxygen saturation etc. This information is combined to create a DSS able to predict cardiovascular diseases. Prevention helps increase people's life quality, thus allowing healthcare systems to save financing resources.

- Medical image segmentation allows to process images to segment the desired organs or structures from the image series. Usually, to perform the manual segmentation medical technicians need to sketch the contours slice by slice using pointing devices such as a mouse or a trackball. This procedure is very time consuming and the results may suffer from intra - or interobserver variability. However, it is a necessary process of automatic or semi-automatic detection of boundaries within images. Furthermore, many different modalities (X-ray, CT, MRI, microscopy, PET, SPECT, Endoscopy, OCT, and many more) are used to create medical images. The result of the segmentation can then be used to obtain further diagnostic insights. Possible applications are automatic measurement of organs, cell counting, or simulations based on the extracted boundary information.

- Biometrics is used in healthcare primarily in the field of security. For example, it is used in digital medical records, which are one of the most valuable personal documents. These documents have to be accurate and complete, and doctors need to access to them quickly. A fail in security and proper accounting of data can lead, on the one hand, to a timely and accurate diagnosis; on the other hand, it can enable health fraudulence. A lot of biometric access control solutions are available on the market, such as:

- Access control to computer systems, through USB fingerprint readers, voice and face recognition software using standard camera and microphone hardware, etc.;
- Door security: doors with biometric locks using iris recognition, fingerprint readers, etc.;
- Portable media such as USB sticks and mobile hard drives with integrated biometric access control and mostly encrypting your data using a built-in algorithm;
- Safes with biometric locks.

- Pervasive-Healthcare systems to increase the monitoring of patients at home. The integration of mobile technology and broadband communications, as well as the proliferation of innovative medical devices, have resulted in the development of pervasive healthcare. Thanks to this development, patients can benefit from healthcare services at any time, without any restriction of place, time and quality. Not only does pervasive healthcare development depend on the evolution of technology, but it also depends on the increasing acceptance of patients to embrace this technology. These systems can gather multiple clinical parameters and are able to operate autonomously without disturbing the patients' lives. These solutions have focused mainly on risk disease management. The benefits that these systems bring are: increase in the speed in diagnosing diseases, continuous patients' data monitoring through the sensors, reduction of the number of visits, and saving of financial resources etc.

- Personal-Healthcare systems to create novel models for personalized diagnostics. These models lead to increase the accuracy and effectiveness of diagnosis. Consequently, the cases of wrong diagnoses have decrease. Moreover, with the increasing strength of calculators able to process large amounts of data, the future of medicine will lead to create a personal treatment for each patient.

7. Phases of the project

1st year: study of the literature, of the state of art and other basic research material

Activity 1.1 study of the state of the art regarding Computational Intelligence in healthcare with particular focus to Fuzzy models;
Activity 1.2 in-depth analysis concerning fuzzy logic and fuzzy rule-based diagnostic systems;
Activity 1.3 attendance of courses and seminars included in the doctorate study plan;
Activity 1.4 participation to international schools and conferences regarding topics relevant for the research theme and the goals planned.

2nd year: development of methods

Activity 2.1 study of the works produced by other researchers with the same goals;
Activity 2.2 design and development of fuzzy methods able to solve recognitions and diagnostic tasks in healthcare;
Activity 2.3 evaluations of the developed methods, comparison with existing approaches and publication of the results.

3^{rd} **year**: application to real-world domains and writing of the doctorate thesis

Activity 3.1 comparison with other research groups works, both national and international, with possible stages in foreign universities or research centers;
Activity 3.2 analysis of the results obtained by applying the developed methods to the selected domains;
Activity 3.3 writing of the doctorate thesis.

8. Result evaluation

To evaluate the results obtained by fuzzy methods on classification (diagnosis) problems, it is possible to use different metrics from literature, such as:

- Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

  For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- Positive Predictive Value (PPV) or Precision is defined as:

$$PPV = \frac{TP}{TP + FP}$$

- Negative Predictive Value (NPV) is defined as:

$$NPV = \frac{TN}{TN + TF}$$

- True Positive Rate (TPR) or Recall is defined as:

$$TPR = \frac{TP}{TP + FN}$$

- True Negative Rate (TNR) is defined as:

$$TNR = \frac{TN}{FP + TN}$$

- F - measure is defined as:

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

in the previously formulations, TP stands for True Positives, FP stands for False Positives, TN stands for True Negatives and FN stands for False Negatives.

9. Possible reference persons external to the department

I hope to identify some possible reference persons external to the department, working in European universities or European research centers, during summer schools or during time spent abroad. Nowadays, possible reference persons external to the department are:

- Martino Pepe, Cardiologia Universitaria, Policlinico di Bari;

- Francesco Masulli, DIBRIS, Università di Genova.

10. References

[1] Aamir Hussain, Rao Wenbi, Zheng Xiaosong, Wang Hongyang, and Aristides Lopes da Silva, "Personal Home Healthcare System for the Cardiac Patient of Smart City Using Fuzzy Logic", Journal of Advances in Information Technology Vol. 7, No. 1, February 2016.

[2] Wiga Maulana Baihaqi, Noor Akhmad Setiawan, Igi Ardiyanto "Rule Extraction for Fuzzy Expert System to Diagnose Coronary Artery Disease", 2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia.

[3] Shaker El-Sappagh, Farman Ali, Amjad Ali, Abdeltawab Hendawi, Farid A. Badira, and Doug Young Suh, "Clinical Decision Support System for Liver Fibrosis Prediction in Hepatitis Patients: A Case Comparison of Two Soft Computing Techniques", Journals & Magazines IEEE Access Volume: 6 ,October 12, 2018.

[4] Luxmi Verma & Sangeet Srivastava & P. C. Negi, "A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data", J Med Syst (2016) 40: 178 DOI 10.1007/s10916-016-0536-z.

[5] Noor Kamal Al-Qazzaz, Sawal Ali1, Siti Anom Ahmad, Javier Escudero, "Classification enhancement for post-stroke dementia using fuzzy neighborhood preserving analysis with QR-decomposition", 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

[6] Nguyen Mong Hien, Nguyen Thanh Binh, and Ngo Quoe Viet, "Edge Detection based on Fuzzy C Means in Medical Image Processing System", 2017 International Conference on System Science and Engineering (ICSSE).

[7] Moustapha Mohamed Saleck, Abdelmajid ElMoutaouakkil, and Mohammed Mouçouf, "Tumor Detection in Mammography Images Using Fuzzy C-means and GLCM Texture Features", 2017 14th International Conference on Computer Graphics, Imaging and Visualization.

[8] Moumen El-Melegy, Rasha Abd El-karim,Ayman El-Baz, Mohamed Abou El-Ghar , "Fuzzy Membership-Driven Level Set for Automatic Kidney Segmentation from DCE-MRI ", 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE).