Intelligent Information System of Heterogeneous Medical Data Analysis

Andriy Yerokhin, Oleksii Turuta,
Andrii Babii
Software Engineering Department
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
andriy.yerokhin@nure.ua

Alina Nechyporenko
Biomedical Engineering Department
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
alina.nechyporenko@nure.ua

Abstract – This paper describes the intelligent information system of heterogeneous medical data analysis for Ear Nose Throat (ENT) domain. Machine learning (ML) methods for medical data processing are analyzed. The scenario of machine learning algorithms for data processing in intelligent information system was developed. Supervised learning algorithms for classification task were implemented, their efficiency was validated. The proposed approach allows to improve the decision-making process and provide pre-surgical and post-surgical assessment of patients.

Keywords — Heterogeneous data, Intelligent information system, Machine learning, Feature extraction, Classification.

I. INTRODUCTION

Medical data are characterized by complexity, inaccuracy, heterogeneity, the presence of hidden dependencies, often their distribution is unknown. Correlations between diseases, symptoms, clinical data and patient's subjective assessments have a high complexity that cannot be fully comprehended by humans anymore. ML methods are very useful for obtaining these correlations. Current paper is devoted to ENT domain. State-of-the-art approaches in diagnostic of rhinological pathologies is set of subjective and objective methods. Among them 4-Phase Rhinomanometry (4PR), endoscopic methods, computer tomography (CT), acoustic rhinometry, subjective tests (VAS), SNOT-22 and other [1-3].

Artificial intelligence (AI) and ML methods have become a powerful tool for medical data analysis. The knowledge base requires effective techniques for extracting medical knowledge. Applications of ML are useful for classification, prediction, recognition patterns and knowledge extraction tasks. The cutting edge techniques of the medical image processing are based on ML methods: Support Vector Machines (SVM), statistical learning, artificial neural networks including deep learning, reinforcement learning algorithms, fuzzy inference etc. These techniques are used in decision support system for brain imaging, larynx imaging, abdominal imaging etc. [4-6]. It is indispensable tool for improvement of decision-making process. In paper [7] we demonstrate the potential use of deep convolutional neural network for rhinomanometry measurements processing. Supervised learning algorithms SVM and Random Forest (RF) are used for classification of the rhinological diseases [8]. The peculiarities of the use of fuzzy regression analysis for the rhinological diagnosis are considered in the paper [9]. Neuro-Fuzzy-system approaches in Data Stream Mining, especially Medical Data Mining, are efficient for the text analysis and are described in [10]. The combination of this approaches with Natural Language Processing (NLP) [11] provides the semantic ontology from the raw text descriptions. The application of AI and ML techniques in the rhinology diagnosis is very promising especially in developing approach of making surgical decisions. The implementation of ML methods to support medical decisions and prognosis of rhinologic diseases is interdisciplinary work that includes the collection and processing of data from the different sources: CT scans, computational fluid dynamics (CFD) simulations [12, 13], biomedical time series and other.

An important meaning have the pre- and post-surgery evaluation. The postoperative evaluation in the functional surgery is very significant for the clinical trials. ENT specialist chooses between surgery and conservative treatment based on the different diagnostic tools: CT-scan, acoustic rhinometry, Rhinomanometry, CFD, VAS. We offer the approach based on supervised learning models to estimate the efficiency of the different methods. The proposed approach will provide an improvement of the prediction of human risks related to surgery.

The main goal of current paper is combine diagnosis methods into a intelligent information system and to extend them by approaches to support medical decision and surgical planning.

II. MATERIALS AND METHODS

Input data stored in the Knowledge base (KB). Machine learning algorithms use this KB as input source for the different types of data. There are medical images data (CT-data), time-series data (CFD, Data from the device), text (a radiologist's description) and the relevant structure field values. The result of data processing will be returned to KB. This result can be represented as anatomical landmark, semantic annotation, CT class, semantic description, classification and clustering model parameters. The data processing consists of the two stages: "CT Post Processing" and "Knowledge management". You can see the scenario of ML algorithms for data processing in intelligent information system in the Fig. 1.

We propose CT automated analysis with computer vision (CV) methods the candidate anatomical landmark detector, feature extraction and clustering for anatomical landmark and CT semi-automated analysis supervised by experts are in the block “CT Post Processing”. The detection of the anatomical landmarks is important task of CT analysis. Best practice of CT analysis requires the anatomical landmark identification which needs the improvement of 3D model layers.

The next step is the anatomy objects detection. It can be divided into the two stages: objects candidate detection and classification with boosting methods. The set of marked objects of 3D model is consolidated into new knowledge stored in KB. The block consists of the four key subblocks: core ML methods, ML Clustering, ML Classification and ML Prediction.

Common ML algorithms process different data type: images, time series, text.

Different methods have wide range of efficiency for the each specific problem. We propose to investigate ML methods and compare the quality for each case. We propose to do the following:

usage of k-means, Fuzzy clustering, DBSCAN, OPTICS algorithms, Kohonen self-organizing map for clustering task;

- usage of regression-based methods for detection global context landmarks and classification-based methods for exhaustive scanning local landmarks for detection anatomical landmarks task;
- usage of Artificial Neural Networks, SVM, RF, Multivariate Discriminant Analysis, Regression methods,

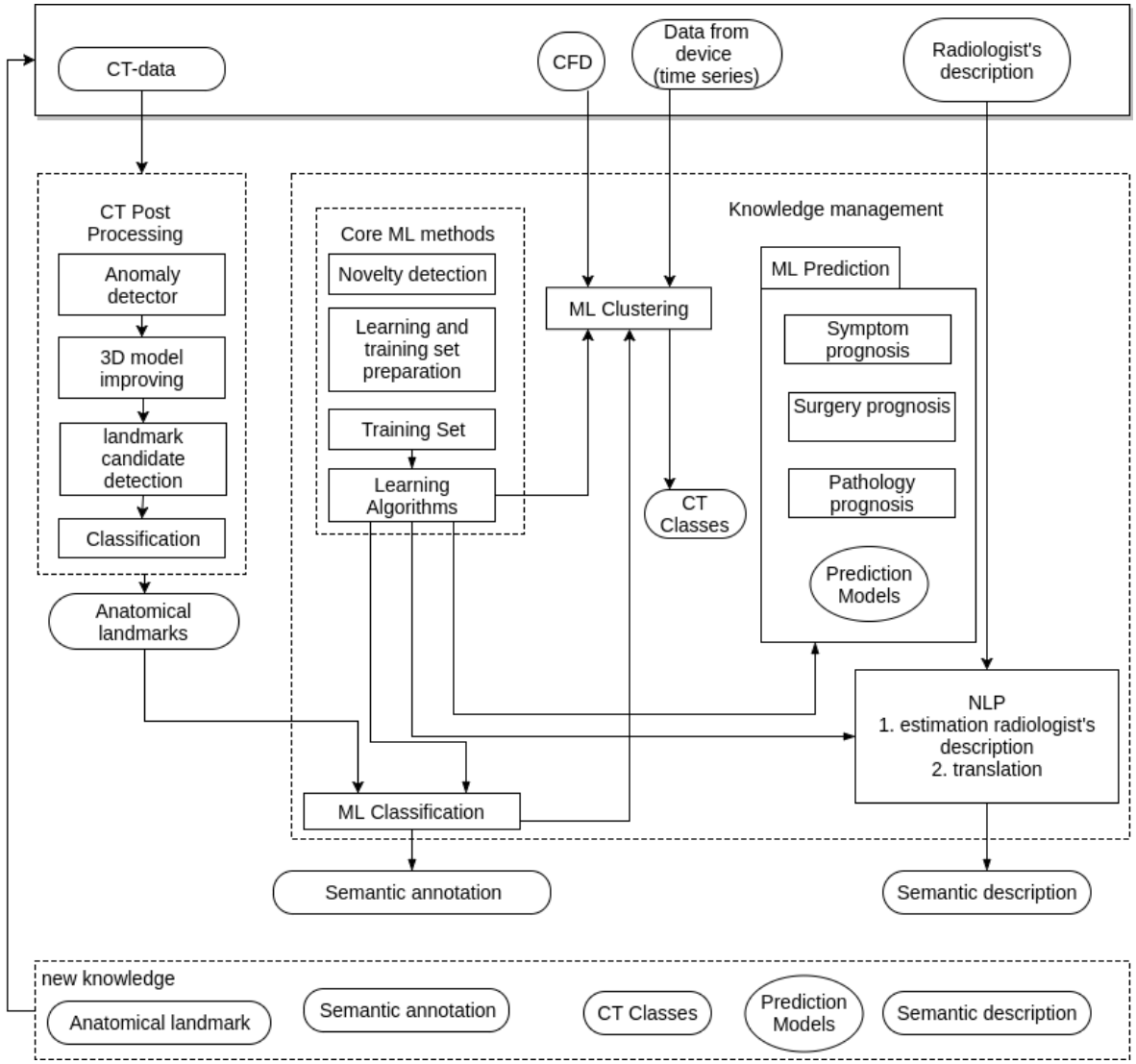

Fig. 1. Scenario of Machine Learning algorithms for data processing in intelligent information system

ML methods can be divided into: supervised and unsupervised, semi-supervised and reinforcement learning.

Regression Trees for classification task;

- usage of classification and regression methods for prediction.

In addition, we offer to use cascade of methods for object detection with strong pattern on CT: CV methods (unsharp masking and SURF/BRISK) for list of feature descriptors and ML (gradient boosting) for object classification.

The core block includes general ML method which is used by another three ML subblocks. The core block is editable. It allows developers to add new methods and improve quality of system. The novelty detection element is designed to identify mistakes or new correct infrequent case in the KB. For example, it can be a human error in supervised learning or undiscovered symptoms of pathology. The novelty detector focuses on this case and requires expert decision. Quality of ML methods work depends on learning and testing of the sets properties. It should be prepared to meet requirements such as accuracy, balance, representativeness.

Each case of classes can be described by the set of the specific features. The extract of correct set of this specific features required of ML methods applying. Success of classification or clustering depends on the feature set. The feature extraction includes the actions with various data transformation in the field of CV, time series data processing or NLP to generate a set of properties for each case.

From various features we should select only significant features. This task can be solved using the expert knowledge collaboration or using automation ML approaches. The "ML Classification" block uses processed CT data, such as anatomical landmark and learned model. The result is the semantic annotation. It consists of new knowledge about CT data, region of interest, pathology.

The set of different CT classes in semantic annotation is used for clustering CT data and different types of relevant patient data. The semantic annotation extends the searching possibilities in KB. For example, the advanced search by pathology, CT classes, anatomical landmark.

The "ML Clustering" block is designed for data mining of CT data with the semantic annotation and other type of relevant patient information, raw and processed times series data in particular. CT data analysis provides a set of anatomical landmark and features which is unique for patient and based on particular qualities of the patient. Times series data obtained in device and CFD is dependent on this particular qualities of the patient. ML clustering algorithms should solve two task: train the model for clustering and find the influence of important features on clustering class. Finally, we offer to investigate initial parameters of CFD. We should investigate different cases of CFD for one patient. Successful CFD is adequate to in-silico trials. ML clustering method allows to tuning CFD initial parameters by CT data.

The main idea of the "ML Prediction" block is the usage of machine learning methods with all of the available data from the knowledge base to create a forecast for the individual patient. We focus on such tasks:

- prediction of pathology stages;
- prediction of surgery operation;
- prediction of symptom changes.

All kinds of prediction should be based on massive data which describes the process progressing in time. It uses various data sources such as images, text descriptions, values of measurements, time-series data.

It requires implementation of some "Core ML methods" block for multidimensional regression and classification algorithms. The stage of pathology prediction can be useful at surgery planning and for understanding the possible impacts of surgery. Pre-surgery and post-surgery data sets can be used to generate prediction of surgery treatment.

There are text description fields for each patient or CT record in KB. A lot of useful and important information will be stored in unstructured text. For automatic processing we offer to find significant features in textual data. Considering the international collaboration of the project we should provide support on filling KB in different languages. NLP methods allow to generate internal semantic description for different case. For example, the advanced search of radiologic description or by symptom.

III. EXPERIMENTS AND RESULTS

For the classification task the RF, SVM and Deep Convolutional Neural Network (DCNN) were implemented. These methods are stored in the ML core block, subblock "Learning algorithms" and can be reused for classification task of any medical data from Knowledge Base of intelligent information system.

The initial data set was obtained from software/hardware system for rhinomanometric measurements "Optimus". Details of the proposed methods and additional information about dataset properties you can find in [7]. The performance of classification was measured in terms of the percentage of correctly classified images with rhinitis. The highest classification accuracy was obtained using DCNN.
The results of classification you can find in the table 1.

TABLE I. RESULTS OF CLASSIFICATION

Methods	**Accuracy (%)**	
	Learning set	***Test set***
RF	89,4	85,6
SVM	88,2	83,4
DCNN	90,1	88,7

IV. CONCLUSION

In this paper the new intelligent information system for analysis of heterogeneous medical data was proposed. It is a developing of concept of intelligence-based approaches for medical data analysis in ENT domain. The key moment is machine learning methods and algorithms usage for data processing. These methods were adopted for heterogeneous data processing. It's indispensable tool for intelligent data analysis to improve the accuracy and quality of the medical care in the ENT domain. It will allow ENT specialists to optimize the personalized treatment of patient, provide predictive and preventive approaches, avoid unnecessary surgical interventions.

REFERENCES

[1] D. Demirbas, C. Cingi, H. Cakli, E. Kaya, "Use of rhinomanometry in common rhinologic disorders", Expert Rev. Med. Devices, no. 8(6), pp. 769–777, 2011.

[2] C. Chaves, C. Ribeiro de Andrade, C. Ibiapina, "Objective measures for functional diagnostic of the upper airways: practical aspects", Rhinology, Vol. 52, no 2, pp. 99-103, 2014.

[3] K. Vogt, A. A. Jalowayski, W. Althaus, C. Cao, D. Han, W. Hasse, H. Hoffrichter, R. Mosges, J. Pallanch, K. Shah-Hosseini, K. Peksis, K. D. Wernecke, L. Zhang and P. Zaporoshenko, "4-Phase- Rhinomanometry (4PR) – basics and practice 2010", Rhinology Suppl. 21, pp. 1-50, 2010.

[4] K. Suzuki, Machine learning in computer-aided diagnosis: medical imaging intelligence and analysis, University of Chicago, 524 p., 2012.

[5] E. E. Bron, M. Smits, W. J. Niessen, and S. Klein, Feature Selection Based on the SVM Weight Vector for Classification of Dementia, IEEE Journal of biomedical and health informatics, Vol. 19, No. 5, pp. 1617-1626, 2015.

[6] C. Barbalata, L. S. Mattos, Laryngeal Tumor Detection and Classification in Endoscopic Video, IEEE Journal of biomedical and health informatics, Vol. 20, No. 1, pp. 322-332, 2016.

[7] A. Yerokhin, A. Nechyporenko, A. Babii, A. Turuta, A New Intelligence-Based Approach for Rhinomanometric Data Processing, Proc. of IEEE 36th International Conference on "Electronics and nanotechnology", pp. 198-201, 2016

[8] A. Yerokhin, A. Nechyporenko, A. Babii, A. Turuta, I. Mahdalina, Usage of Phase Space Diagram to Finding Significant Features of Rhinomanometric Signals, Proc. of the International Conference on Computer Sciences and Information Technologies, Lviv, Ukraine, pp. 70-73, 2016.

[9] A. L. Yerokhin, A. S. Babii, A. S. Nechyporenko, O. P. Turuta, A Lars-Based Method of the Construction of a Fuzzy Regression Model for the Selection of Significant Features, Cybernetics and Systems Analysis, Vol. 52, Issue 4, pp. 641–646, 2016.

[10] O. Turuta, I. Perova, A.Deineko, Evolving Flexible Neuro-Fuzzy System for Medical Diagnostic Tasks, International Journal of Computer Science and Mobile Computing - JCSMC, Vol. 4, Issue. 8, pp. 475-480, 2015.

[11] N. Indurkhya, F. J. Damerau, Handbook of Natural Language Processing, Second Edition, - Chapman & Hall/CRC Machine Learning & Pattern Recognition, 2nd Edition, 704 p., 2010.

[12] I. Horschler, M. Meinke, W. Schroder, Numerical simulation of the flow field in a model of the nasal cavity, Computers & Fluids 32 (1), pp. 39-45, 2003.

[13] K. Inthavong, J. Wen, J. Tu, Z. Tian, From CT scans to CFD modelling - fluid and heat transfer in a realistic human nasal cavity, Engineering applications of computational fluid mechanics, Vol. 3, No. 3, pp. 321-335, 2009.