pHealth 2022
B. Blobel et al. (Eds.)

doi:10.3233/SHTI220983

Development and Implementation of the Data Science Learning Platform for Research Physician

Lejla BEGIC FAZLIC[a,1], Marvin SCHACHT[a], Marlies MORGEN[a], Anke SCHMEINK[b], Robert LIPP[b], Lukas MARTIN[c], Thomas VOLLMER[d], Stefan WINTER[d] and Guido DARTMANN[a]

[a]*ISS, Trier University of Applied Sciences, Trier, Germany*
[b]*Chair of Information Theory and Data Analytics, RWTH Aachen University, Aachen, Germany*
[c] *Department of Intensive and Intermediate Care, University Hospital Aachen*
[d] *Philips GmbH Innovative Technologies, Aachen, Germany*

Abstract. Data analysis and their application are the unavoidable factors in the activities analyses in health care. Unfortunately, the acquisition of data from large available medical databases is a complex process and requires deep knowledge of computer science and especially knowledge of tools for data management. According to the European General Data Protection Regulation, the problem becomes much more complex. Recognizing these problems and difficulties, we have developed a Data Science Learning Platform (DSLP) that primarily targets practitioners and researchers but also the computer science students. Using our proposed tool chain together with the developed graphical user interface, data scientists and research physicians will be able to use available medical databases, apply and analyze different anonymization methods, analyze data according to the patient's risk and quickly formulate new studies to target a disease in a complex data model. This article presents a clinical research discovery toolbox that implements and demonstrates tools for data anonymization, patient data visualization, NLP-tools for guideline search and data science learning tools.

Keywords. Healthcare, risk prediction, Natural Language Processing, Fuzzy logic

Introduction

Data analysis and their application are becoming an essential aspect in the domain of health care, especially in the field of intensive care. The developed data analytic tools can provide a basis for classification problems in health care, e.g., the classification between normal data and anomalies. These tools will be furthermore useful to identify sub-groups of patients in a given population. Due to the large amounts of data, medical doctors are facing challenges to recognize symptoms and to identify disease in early stage as well as to choose optimal treatment [1].

[1] Lejla Begic Fazlic, ISS, Trier University of Applied Sciences, Trier, Germany, E-mail: l.begic@umwelt-campus.de

The data is often very complex and exists in different medical systems and the doctors need a fast access to the right information at the right time. Even with very experienced physicians, this can be a complex task. It is essential for researching physicians to have available test data that they can easily statistically analyze and combine. Moreover, the results provided by machine learning, e.g., disease prediction or risk, could also help them to build more efficient models by adjusting patients parameters on one side, and to combine it with guideline recommendations on the other side. In the recent years, there are a lot of studies that use the power of Artificial Intelligence for the design of risk prediction models [2], [3][4]. The Risk Based Toolbox was created following systematic literature review and recommendation as well as survey of clinical trial units [5]. A detailed review and the principles of anomaly detection algorithms in the different types of medical problems are described in [6]. Each of these algorithms work well on special types of data and computer scientist and statisticians were in the most cases the targeted user group for this type of research. There are furthermore numerous published examples of powerful computational tools - clinical decision support systems - installed in hospital, that can for example identify critical values automatically and help clinicians in decision [7],[8]. The semantic CDSS systems based on ontology and application of fuzzy ontological reasoning in the medical guidelines are presented in recent studies [9], [10]. The different anonymization methods are earlier presented in [11][12][13] and implemented in anonymization tools [14]. We recognized the lack of integration of the mentioned models into a useful research and discovery tool that will be primarily intended for analyzing, learning, researching and testing the results.

Our goal is that the proposed model and toolbox primarily target learning clinical researchers who are newcomers to the two areas; artificial intelligence and medicine. The computer science students and data scientists will also find this toolbox helpful with their studies to understand and analyze the results provided by AI algorithms, to understand the process of anonymization, machine learning, NLP and fuzzy logic and to be able to use and to easier analyze big data. To ensure that the data, which is continuously collected, can also be used by researching physicians in an appropriate manner to their tasks, a Graphical User Interface (GUI) is developed that can also be operated efficiently and effectively by this user group.

1. Material and Methods

According to the European General Data Protection Regulation (GDPR), the health data are treated “as a "special category" of personal data which is considered to be sensitive by its nature” [15]. For demonstration purpose, we used the MIMIC-III 100 patients database provided by [16][17]. As the MIMIC-III patients’ data are already anonymized, in our application we used synthetic tool [18] to generate patient data and to test and to improve the anonymization methods.

The enclosed program was created in the object-oriented C\# programming language using the .NET framework, where windows are GUI controls of the .NET Framework. Microsoft Visual Studio 2019 was used as the development environment. For testing data, the open-source synthetic data generator Synthea was used [18]. The software is integrated with Python code where the model for risk prediction [19] is investigated. PostgreSQL as open-source object-relational database and the Matlab Fuzzy rule database are used for the primary data management. The code for our tools can be found at https://gitlab.rlp.net/l.begic899724/dslp.

2. Methodology and Design

We developed a graphical user interface that consists of the four-parts anonymization tool, patient simulator tool, risk prediction tool and guideline tool (Figure 1).

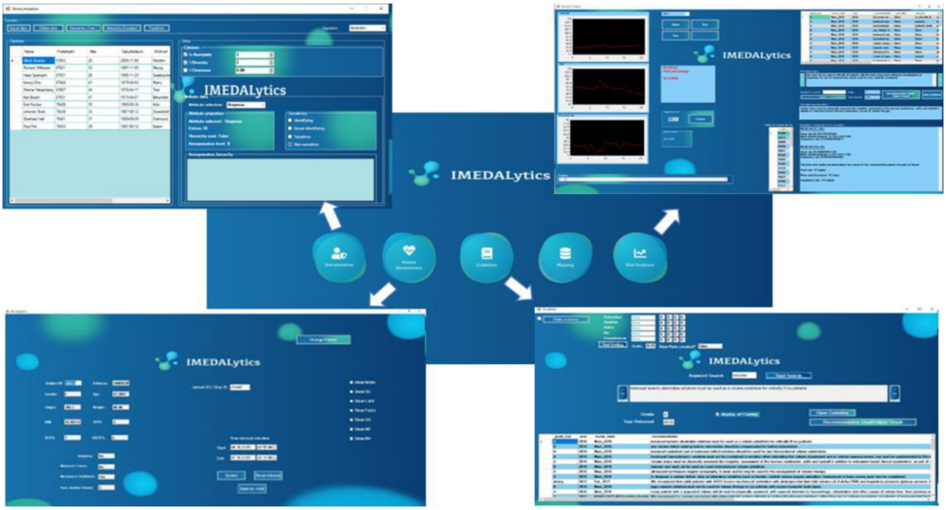
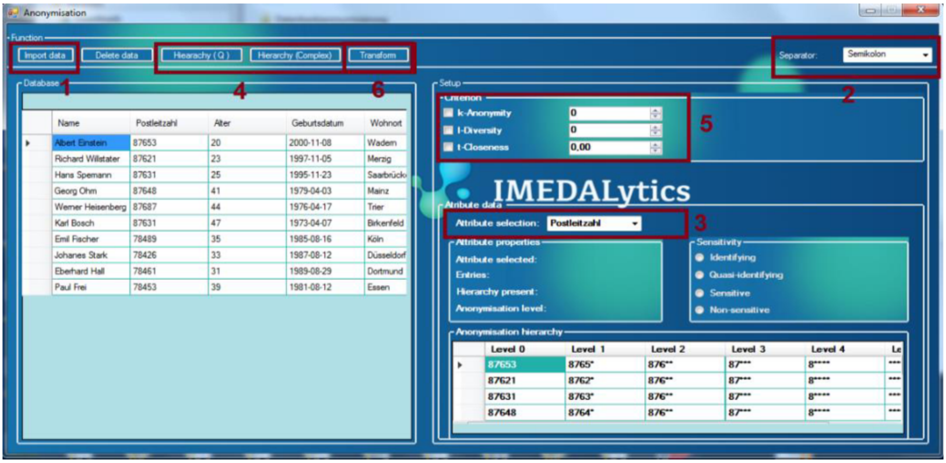


Figure 1. Data Science Learning Toolbox.

The first anonymization tool contains five different parts as it is presented in Figure 2. The first functionalities (No. 1 and No. 2 in Figure 2) represent the data import process, where the files' format can be a CSV-file generated from the various data sources. The properties of the attributes of the data are displayed in No. 3 in Figure 2.



Separator: Semikolon

	Name	Postleitzahl	Alter	Geburtsdatum	Wohnort
▸	Albert Einstein	87653	20	2000-11-08	Wadern
	Richard Willstater	87621	23	1997-11-05	Merzig
	Hans Spemann	87631	25	1995-11-23	Saarbrück
	Georg Ohm	87648	41	1979-04-03	Mainz
	Werner Heisenberg	87687	44	1976-04-17	Trier
	Karl Bosch	87631	47	1973-04-07	Birkenfeld
	Emil Fischer	78489	35	1985-08-16	Köln
	Johanes Stark	78426	33	1987-08-12	Düsseldorf
	Eberhard Hall	78461	31	1989-08-29	Dortmund
	Paul Frei	78453	39	1981-08-12	Essen

Criterion
☐ k-Anonymity 0
☐ l-Diversity 0
☐ t-Closeness 0,00

Atribute data
Attribute selection: Postleitzahl

Attribute selected:
Entries:
Hierarchy present:
Anonymisation level:

● Identifying
● Quasi-identifying
● Sensitive
● Non-sensitive

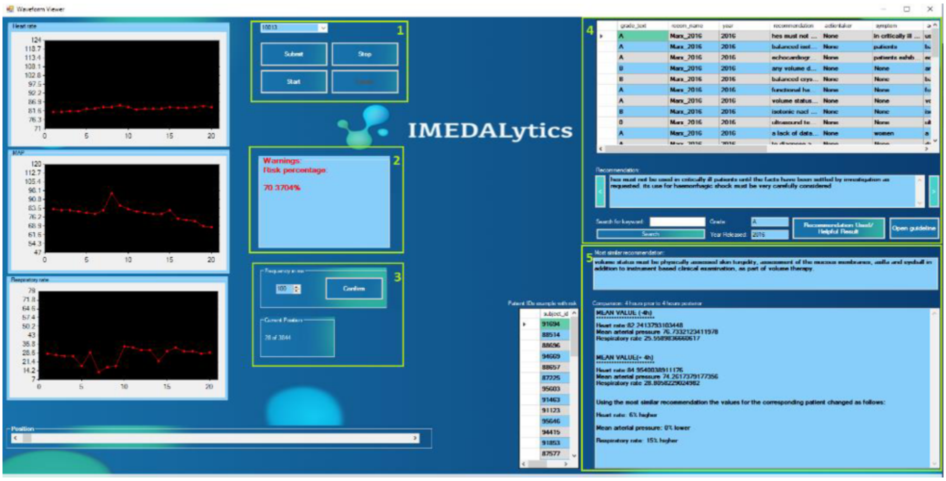
	Level 0	Level 1	Level 2	Level 3	Level 4	Le
▸	87653	8765*	876**	87***	8****	***
	87621	8762*	876**	87***	8****	***
	87631	8763*	876**	87***	8****	***
	87648	8764*	876**	87***	8****	***

Figure 2. Anonymization toolbox

To anonymize the data, a hierarchy is required for each quasi-identifying attribute (No. 3 and No. 4 in Figure 2). At this stage, the application offers the different anonymizations' algorithms: k-Anonymity [11], Distinct l-Diversity [12] and t-Neighborhood [13] (No.5 in Figure 2). The button marked as No.6 in the same Figure is used to transfer the input data and the specified information to the next level, in which the individual created nodes are checked via the OLA algorithm [20]. The patient simulator toolbox simplifies the procedure of formulating complex SQL queries to get patients' results for different parameters like lab events, vital signs or the parameters that integrated more data to construct an event e.g., mechanical ventilation or medication treatment. The users in a simple way pick up and match the patient data according to the desired criterion and have the statistical analysis and visualization for every available parameter over the different time interval. The guideline tool represents a recommendation engine where over 1020 different recommendations from 45 guidelines [21] are stored in a database and can be used by different search criteria.

Additionally, using our recently published NLP-FUZZY algorithm [19] we created a procedure for automatic recommendation extraction from guidelines and insertion of recommendations from guidelines to the database. The proposed NLP-FUZZY algorithm combines capabilities of Natural Language Processing (NLP) and Fuzzy Logic approaches. In the first step, the NLP-FUZZY performs a semantic extraction of medical guidelines using a bi-directional Long Short-Term Memory (LSTM). Subsequently, using the extracted semantic, it creates fuzzy rules, which are able to recognize new cases in a learning domain while predicting and extracting the grade of recommendation.

The last toolbox (Figure 3) brings together the results of merged waveform patient's data from the patient simulator toolbox, the risk prediction algorithm based on Dynamic Time Warping- Dynamic Barycenter Averaging (DTW-DBA) [4] approach and the medical guidelines engine. In the learning phase of the DTW-DBA algorithm, a statistic approach in combination with DTW is used to merge all patients in "positive on disease" and "negative on disease" classes and create a Dynamic Time Warping Barycenter Averaging analysis for each feature [19]. In the second classification phase, validation data sets are used to validate the precision of classification (No.1 - No. 2 in Figure 3). Additional adjustments such as frequency and positioning in the graph gives to user an opportunity to analyze the graph and risk results in selected segments as well as in chosen time slots (No. 3 in Figure 3). In the risk simulation, the stable vital parameters are presented with the blue line, and in the moment where risk is recognized, the graphs' lines change their color in red. When the risk is recognized, the risk percentage is shown and in the same moment the recommendations connected with the disease are offered to the clinician (No. 4 in Figure 3). Here the complete statistical analysis over the patients with the most similar symptoms are done and according to it, the most appropriate recommendation is singled out (No. 5 in Figure 3). Expected results for vital sign parameters after the suggested recommendation (based on previous statistical analysis over the most similar patients) are also presented to the researcher. The information about the recommendation strategy and the statistical analysis over the patients with the similar patient condition gives the young physician the opportunity to analyze different risk patterns and possibly find out some hidden properties in patients states. The developed tools in this paper are prototypes and research demonstrators for training purposes and showing the work process of a future research tool chain. However, further research is needed especially in the field of the clinical applications and use-cases.



10013

Warnings:
Risk percentage:

70.3704%

Frequency in ms: 100

Current Position: 78 of 3044

Patient IDs example with risk

subject_id
91694
88514
88696
94669
88657
87225
95603
91463
91123
95646
94415
91853
87577

grade_text	room_name	year	recommendation	action taker	symptom	a
A	Marx_2016	2016	hes must not ...	None	in critically ill ...	us
A	Marx_2016	2016	balanced isot...	None	patients	ba
A	Marx_2016	2016	echocardiogr...	None	patients exhib...	ec
B	Marx_2016	2016	any volume d...	None	None	ar
B	Marx_2016	2016	balanced crys...	None	None	ba
A	Marx_2016	2016	functional he...	None	None	fu
A	Marx_2016	2016	volume status...	None	None	vo
B	Marx_2016	2016	isotonic nacl ...	None	None	is
B	Marx_2016	2016	ultrasound te...	None	None	ul
A	Marx_2016	2016	a lack of data...	None	women	a

Recommendation:
hes must not be used in critically ill patients until the facts have been settled by investigation as requested. its use for haemorrhagic shock must be very carefully considered

Search for keyword:
Grade: A
Year Released: 2016

Most similar recommendation:
volume status must be physically assessed skin turgidity, assessment of the mucous membranes, axilla and eyeball in addition to instrument based clinical examination, as part of volume therapy.

Comparison: 4 hours prior to 4 hours posterior

```
MEAN VALUE (-4h)
*********************
Heart rate 82.2413793103448
Mean arterial pressure 76.7332123411978
Respiratory rate 25.5589836660617

MEAN VALUE(+ 4h)
*********************
Heart rate 84.9540038911176
Mean arterial pressure 74.2617379177356
Respiratory rate 28.8058229024982

Using the most similar recommendation the values for the corresponding patient changed as follows:

Heart rate: 6% higher

Mean arterial pressure: 0% lower

Respiratory rate: 15% higher
```

Figure 3. Risk prediction toolbox.

Here especially knowledge from medical experts is required to identify/classify, e.g., anomalies in the time series and to connect them with a real medical event. This solution represents a useful tool to connect experts from the computer science with health experts and to evaluate upcoming novel data analysis approaches and algorithms.

3. Conclusion and Future Work

In our work, we presented the development and implementation of the Data Science Learning Platform (DSLP) for research physicians. The first purpose of the DSLP is to offer research physicians the possibility to simplify the data acquisition from Healthcare Big Data without knowing complex SQL queries and other computer science techniques. The combination of Natural Language Processing and Fuzzy logic is demonstrated in real guideline examples and the automatic extraction and insertion of the guidelines are provided for research. The last tool in our DSLP is an integrated risk prediction algorithm based on DTW-DBA approach. Here, the physicians have an opportunity to analyze the patient's risk, analyze suggested recommendations as well as estimated vital sign parameters after recommendation application. The next step in DSLP will be to implement synthetic data platform using ML techniques. Here the research physician will have opportunity to generate synthetic data to match sample data, where the important properties of sample data are reflected in synthetic data.

4. Aknowledgment

The results of this project were generated by funding from the German Federal Ministry of Education and Research (BMBF) under the IMEDALytics grants 13GW0280A, 13GW0280C, 13GW0280D, and 13GW0280E. These results are currently used in the BMEL project KI-Pilot (funding code 2820KI001) and the demonstrator is exhibited in the KI-Reallabor at Trier University of Applied Sciences. We thank the BMEL for the continuation of the research especially in the area of anomaly detection and smart data platforms. We gratefully acknowledge Annika Kaltenhauser (Ergosign, GmbH) for the design elements of the GUI and student Mr. Marcel Simon for providing necessary resources to accomplish this research work.

References

[1] Shakeel D, Shakeel AM. Personalized drug concentration predictions with machine learning: an exploratory study. International journal of basic and clinical pharmacology. 2020: 980.

[2] Dahiwade D, Patle G, Meshram E. Designing Disease Prediction Model Using Machine Learning Approach. 3rd International Conference on Computing Methodologies and Communication (ICCMC); 2019; pp. 1211-1215. doi:10.1109/ICCMC.2019.8819782.

[3] Deepthi Y, Pavan Kalyan K, Mukul V, Radhika K, Kishore Babu D, Krishna Rao N V. Disease Prediction Based on Symptoms Using Machine Learning, Energy Systems. Drives and Automations; 2020, pp. 561-569. doi:10.1007/978-981-15-5089-8-55.

[4] Begic Fazlic L, Hallawa A, Dziubany M, Schmeink A, Lipp R, Peine A, Martin L, Vollmer T, Winter S, Dartmann G. A machine learning approach for the classification of disease risks in time series. 9th Mediterranean Conference on Embedded Computing (MECO); 2020; Budva, Montenegro.

[5] https://ecrin.org/tools/risk-based-monitoring-toolbox

[6] Li D, Chen D, Jin B, Shi L, Goh J,Ng S.K. MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks. ICANN 2019: Artificial Neural Networks and Machine Learning – ICANN; 2019, pp. 703-716. doi:https://doi.org/10.48550/arXiv.1901.04997.

[7] Sutton RT, Pincock D, Baumgart DC et al. An overview of clinical decision support systems: benefits, risks, and strategies for success, npj Digit. Med; Vol3, No 17, 2020. doi:10.1038/s41746-020-0221-y.

[8] Belard A, Buchman T, Forsberg J ~et al. Precision diagnosis: a view of the clinical decision support systems (CDSS) landscape through the lens of critical care, J Clin Monit Comput 31; 2017; pp. 261–271. doi:10.1007/s10877-016-9849-1.

[9] Piotr S. Application of Fuzzy Ontological Reasoning in an Implementation of Medical Guidelines; 6th International Conference on Human System Interactions; 2013. doi:10.1109/HSI.2013.6577845.

[10] El-Sappagh S, Ali F, Hendawi A, Jang JH, Kwak KS. A mobile health monitoring-and-treatment system based on integration of the SSN sensor ontology and the HL7 FHIR standard. BMC Med Inform Decis Mak. 2019 May 10;19(1):97. doi: 10.1186/s12911-019-0806-z.

[11] Sweeney L. k-Anonimity:A model for protecting privacy, International Journal of Uncertainty; Fuzziness and Knowledge-Based Systems. 2002. Vol10; No.5, pp.557-570.

[12] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M, L-diversity: privacy beyond k-anonymity, 22nd International Conference on Data Engineering (ICDE'06).2006; pp. 24-24.doi: 10.1109/ICDE.2006.1.

[13] Li N, Li T, Venkatasubramanian S, t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, IEEE 23rd International Conference on Data Engineering, 2007. pp. 106-115, 2007.doi:10.1109/ICDE.2007.367856.

[14] https://arx.deidentifier.org/

[15] https://globaldatahub.taylorwessing.com/article/health-data-and-data-privacy-challenges-for-data-processors-under-the-gdpr

[16] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000 Jun 13;101(23):E215-20. doi: 10.1161/01.cir.101.23.e215. PMID: 10851218.

[17] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035. doi: 10.1038/sdata.2016.35. PMID: 27219127; PMCID: PMC4878278.

[18] https://github.com/synthetichealth/synthea

[19] Begic Fazlic L, Hallawa A, Schmeink A, Peine A, Martin L, Dartmann G. A Novel NLP-FUZZY System Prototype for Information Extraction from Medical Guidelines, 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, 2019. pp. 1025-1030

[20] Zhang J, Gong X, Zhipeng H, Feng S, An Improved Algorithm for K-anonymity, In book: Contemporary Research on E-business Technology and Strategy; 2012. pp.352-360.

[21] https://www.guidelinecentral.com/