Exploring Heterogeneous Data Processing to Improve Clinical Applications

Rodrigo F. Rönnau[1], Sandro J. Rigo [1], Marta Bez[2], Jorge L.V. Barbosa [1]
[1]Universidade do Vale do Rio dos Sinos (UNISINOS)
[2]Universidade FEEVALE

rodrigo.ronnau@gmail.com, martabez@feevale.br, {rigo, jbarbosa}@unisinos.br

Abstract: *Computer systems have been used widely in health care quality improvement. In general, these systems do not support different data formats, and this consists of a severe limitation. This paper presents a model which makes possible the use of different data formats to provide and integrate information that supports medical specialists' activities. Two prototypes were built using the model, which aims to exemplify its benefits and enable its evaluation. To evaluate the proposed approach, besides the developed applications, twelve health professionals and thirty-five computer professionals answered a questionnaire about the prototypes. The two distinct questionnaires with participants evaluation and the development of the prototypes allowed the identification of the perceived contributions in both software development and clinical support application areas.*

Resumo*: Os sistemas de computadores têm sido amplamente utilizados na melhoria da qualidade dos serviços de saúde. Em geral, esses sistemas não suportam diferentes formatos de dados e isso consiste em uma limitação severa. Este artigo apresenta um modelo que possibilita o uso de diferentes formatos de dados para fornecer e integrar informações que apoiam as atividades de médicos especialistas. Dois protótipos foram construídos usando o modelo, que visa exemplificar seus benefícios e permitir sua avaliação. Para avaliar a abordagem proposta, além das aplicações desenvolvidas, doze profissionais de saúde e 35 profissionais de informática responderam a um questionário sobre os protótipos. Os dois perfis distintos dos participantes do questionário e o desenvolvimento dos protótipos permitiram avaliar as contribuições percebidas nas áreas de desenvolvimento de software e aplicação de suporte clínico.*

1. Introduction

Medical information systems [1–3] usually deal with only one data input category, such as texts [3–5] or images [2,6,7]. However, data integration enables improvements in patient care [8–10]. Prior treatment records and other complementary clinical data provide broader support to the health professionals and can foster more specific diagnostic and treatment recommendation procedures for patients, reducing failures and improving health services quality [11–13].

Some articles address the use of heterogeneous data in clinical support systems [14,15], describing information extraction to determine relationships between them and possible diseases. Also, these systems integrate data with the patient's history to generate a differential diagnosis. Another approaches [16,17] uses features extracted from audio, video, Magnetic Resonance Image (MRI), Diffusion Tensor Imaging (DTI), and text into machine learning algorithms to detect people with depression or to distinguish biomarkers to support the diagnostic process.

Nevertheless, the study of literature about this topic identifies a gap related to the aggregation of different data formats to foster the support of health care to provide more comprehensive and accurate assistance to diagnosis activities [9,11,15,17–20]. The most known applications in the medical field can handle a reduced or a unique data format, while the medical practice demands integration of data originated in different formats. In this context, some authors proposed a framework to lead clinical decision systems to combine some sources of information available. Williams et al. [8] proposed a modular framework for the development of clinical decision support systems, with emphasis on the need to use plug-and-play standards. The acquisition and merging of medical information from the patient's bedside devices, databases about the patients, and manual information inserted by the clinical team, are the main components of the framework. Maglogiannis et al. [11] presented a distributed architecture for health record systems, capable of integrating the processing of different information on patients with melanoma. This improvement was possible with the exchange of data between different clinical information systems, demonstrating the possibility of interoperability among participating institutions.

In their project, Hazlehurst et al. [18] described a web platform for Comparative Effectiveness Research (CER). The platform's processing flow has tasks such as extraction, modeling, aggregation, and data analysis. Another work [19] detailed a medical information retrieval system with support for multimodal searches of clinical cases. Puppala et al. [9] described the development of an integrated computing environment, called METEOR, composed of two components: the enterprise data warehouse and an intelligence and analysis software to the practice of evidence-based medicine. Vizza et al. [20] described a framework that aims to define a general-purpose framework for managing images and their respective annotations. The model allows the visualization of clinical data, research, and integration of different information, such as patient history and laboratory data, into a single information system.

The literature analysis identifies a context in which there is not yet the necessary support for the needs of heterogeneous data integration to support the medical activities related to the diagnosis. A broader set of options in data processing and integration can be related to better resources to support medical activity. The current study aims to propose a model called Heterogeneous Data Processing for Clinical Support Applications (HDPCSA) to support different decision-making applications in the health area, which can aggregate features extracted from different data sources, storing it in a structured and standardized database to facilitate the incorporation of new software features.

2. Methods

The HDPCSA model is composed of four layers: input, feature extraction, feature repository, and applications. Each layer could be extended to support instances of corresponding elements, which implement the use of different input types, representation patterns, and applications. Figure 1 shows the model layers.

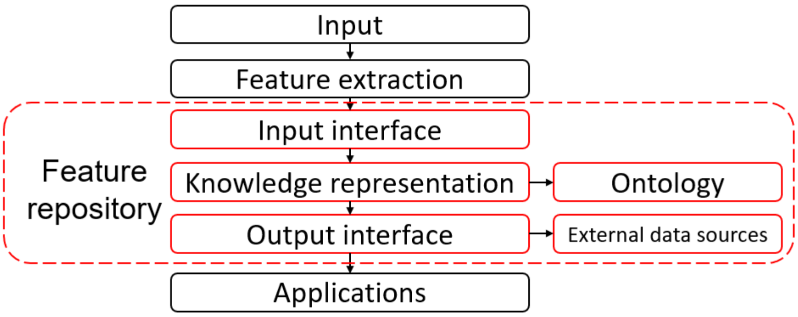


Fig 1 Proposed model general architecture.

The model allows the change on a component without modifying the others, since all blocks are independent. It is also possible to use the same element in different operations, which means that the model allows the reuse of already developed functions. A brief example of systems build with of proposed model could be seen in Figure 2.

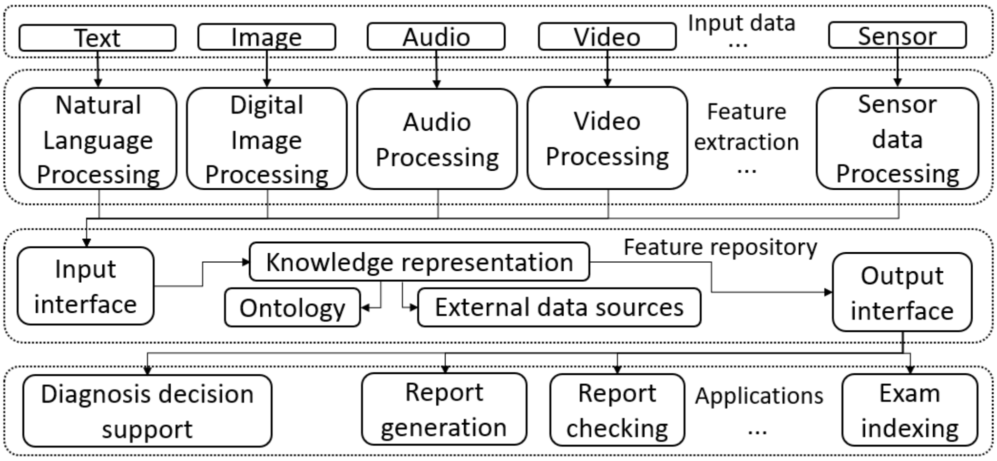


Fig 2 Examples of elements that may exist in each layer of the proposed model.

The Input group elements represent the different data source types that could be used to feature extraction. Each input data type is processed by a specific software routine, aiming to extract the information that could be used by the applications. This process occurs in the Feature Extraction layer, which has specific functions to handle these tasks. Feature Repository groups all components related to the storage, visualization, and access to the data. To enable interoperability between health systems, we used a recognized representation standard. Among the available standards in the literature, we highlight both openEHR and HL7 FHIR [22].

To reduce the complexity of the repository representation pattern, we defined two interfaces: input and output. These boundary interface elements link the public list of information available for use with the corresponding elements according to the representation pattern selected—for instance, the name of a patient. The way of the feature's repository stores this information is not relevant to the other elements of the proposed model, hence both input and output interfaces should adapt the content to the standard in use. Therefore, to see and use stored features, it may be necessary to know the pattern specification employed in the Feature Repository. Despite this, we could reduce the need for previous knowledge about the pattern using ontologies as an alternative way to visualize the structure and contents of the repository. This type of representation tends to be more user-friendly and intuitive for people who are unfamiliar with the features and their corresponding attributes specified in the HL7 FHIR, for instance. An example of an ontology that represents information from a medical report in that format is showed in Figure 3.

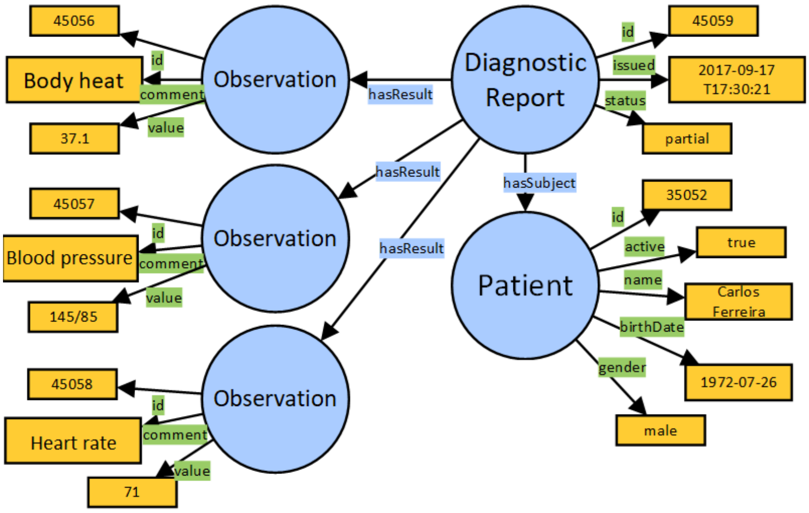


Fig 3 The EHR ontology representation examples.

The Feature Repository layer has the last element the External Databases, an input and output gate for the information stored in the repository. This element's objective is the integration with external sources to allow the use of extracted features by third-party systems. The last group of elements in the model contains the applications that make use of the data stored in the Feature Repository.

2.1 Prototypes

We developed two prototypes with the HDPCSA model. These prototypes contain the fundamental elements to exemplify their use and enable the realization of proofs of concept. Subsequently, we presented the model and the implemented prototypes to professionals who answered a quiz about it. The first prototype allows detecting if all valuable information from medical images is present in the textual report or if the stored data is wrong. Figure 4 shows the proposed model with the already developed first prototype application elements in red.

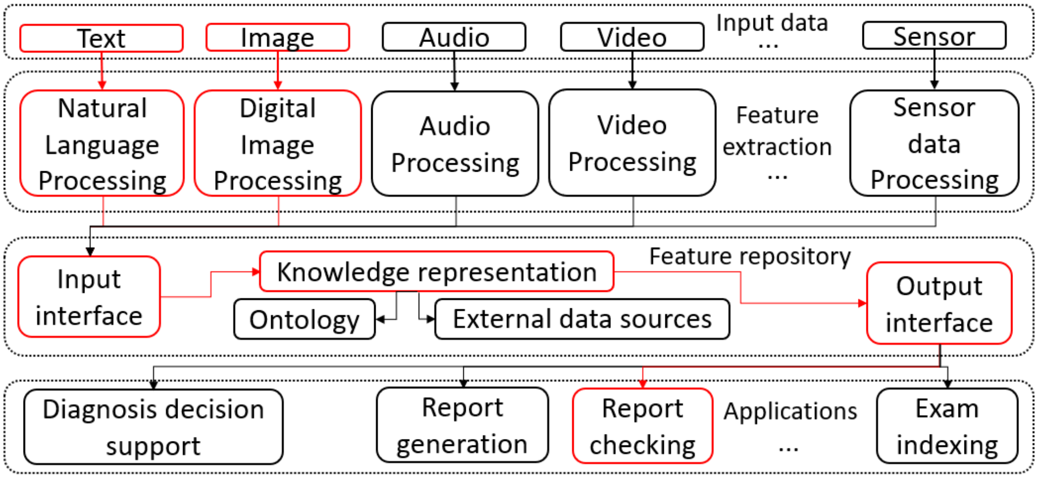


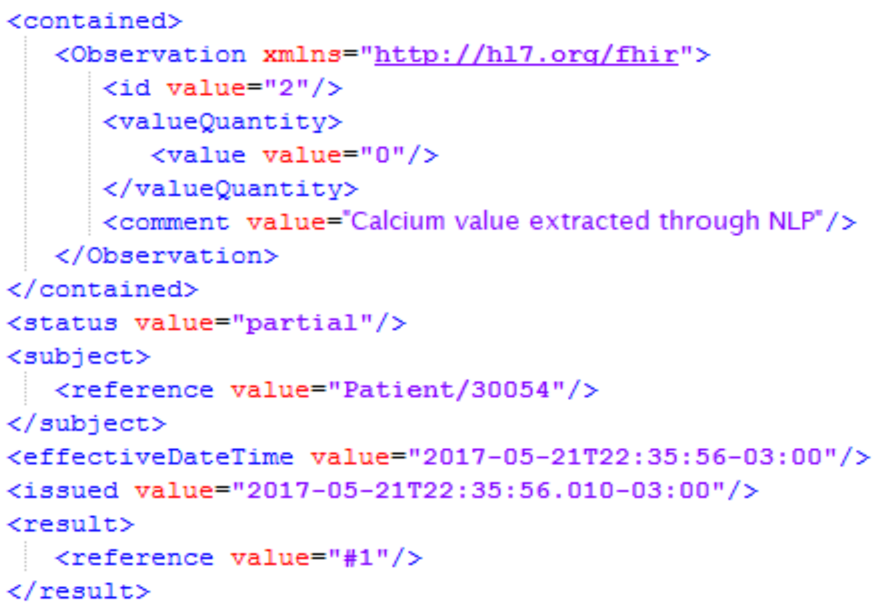
Fig 4 Model elements developed in the first prototype.

The first entry contained the medical reports and intended to identify portions of text on natural language that indicate the presence of calcifications. As Cai et al. [23], the action flow uses the syntax analyzer PALAVRAS [24] besides the UMLS Metathesaurus Browser lexicon. The other entry type used was thorax computed tomography scans. The developed code to process that data can identify calcification points in the cardiac region. We used the image processing tool described by Reis et al. [25], PixelMed Java DICOM Toolkit library[1] and DICOM pattern to allow files access. The Feature Repository group has an element called Representation Pattern, developed according to the HL7 FHIR specification by using the HAPI FHIR library [2]. The extracted features from images and texts were inserted in the database and become available to the reports checking application. We used the exam identifier to query extracted feature repository, allowing the comparison between the result obtained by the processing of images with those from reports.

In the next stage, the system scans each input type processing for any differences in the observations to trigger an alert, which reports the identified situation to the user. An example of an XML fragment with the information contained in the Feature Repository for an exam could see in Figure 5.

[1] Available at http://www.pixelmed.com/dicomtoolkit.html. Accessed 25 July 2018.

[2] Available at http://hapifhir.io/. Accessed 25 July 2018.



```
<contained>
   <Observation xmlns="http://hl7.org/fhir">
      <id value="2"/>
      <valueQuantity>
         <value value="0"/>
      </valueQuantity>
      <comment value="Calcium value extracted through NLP"/>
   </Observation>
</contained>
<status value="partial"/>
<subject>
   <reference value="Patient/30054"/>
</subject>
<effectiveDateTime value="2017-05-21T22:35:56-03:00"/>
<issued value="2017-05-21T22:35:56.010-03:00"/>
<result>
   <reference value="#1"/>
</result>
```

Fig 5 XML fragment with information contained in the Feature Repository.

The second prototype is an application that generates a preliminary report of a specific patient, which contains phrases referring to their current health condition based on information obtained through imaging and sensors. As shown in Figure 6, the second prototype is composed of specific modules (highlighted in orange) and others from the first prototype (highlighted in red). The obtained data from sensors were body temperature, blood pressure, and heart rate.

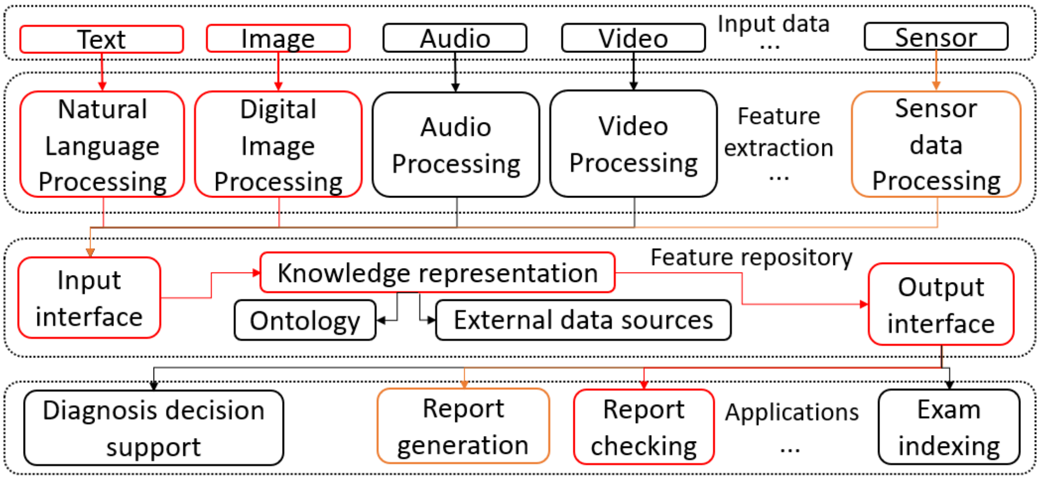


Fig 6 Elements added to the system in the development of the second prototype.

The entire Feature Repository structure of the first prototype was reused, which corroborates intending to provide simple integration and component reuse. After the input processing inserts the features into Feature Repository, the model allows a query though both patient's or exam's identifier to obtain the repository features and then apply the defined rules to determine the report text sentences (Table 1). The prototype rules were defined using current state-of-art literature and American Heart Association[3]. Also, we used the body temperature rules described by Sund-Levander et al. [26]. The proposed work built the prototypes according to the HDPCSA architecture, which allows the processing of three input types (image, text, and sensors) and provides data storage in a standardized repository using the HL7 FHIR specification. One prototype performs medical report checking while the other executes the medical report text generation. The prototypes use elements from all the model layers, thus showing each component's role, relevance, and relationship with the other groups that complete the proposed architecture.

Table 1 Elements added to the system in the development of the second prototype.

Features	Rules	Sentences
Calcification	If = 0	No points of calcification were detected in the cardiac region
	If = 1	Calcification points were detected in the cardiac region
Body temperature	If <= 37.5	Normal temperature (Celsius scale)
	If > 37.5	Fever (Celsius scale)
Heart rate		Heart rate (bpm value)
Blood pressure	S >= 180 or D >= 110	Hypertensive crisis (mmHg value)
	S >= 160 or D >= 100	Stage 2 hypertension (mmHg value)
	S >= 140 or D >= 90	Stage 1 hypertension (mmHg value)
	S >= 120 or D >= 80	Pre-hypertension (mmHg value)
	Others	Normal blood pressure (mmHg value)

3. Evaluation

To evaluate the HDPCSA model, we used questionnaires developed based on Davis [27], which was later extended by Venkatesh and Davis [28]. We applied questionnaires to both health and technology professionals after attending a proposed model and prototypes presentation. Two forms were defined, one per area (computing and health), which each consists of a series of statements about the research topic, model, and the elements that make up the questionnaires. Each evaluation participant chooses the option that more accurately reflects his or her opinion on each of the questions according to the Likert [29] scale. This step aimed to check model acceptability based on these professional reviews and their perspective on usability, utility, and contributions.

The health professionals' evaluation stage took place in a face-to-face meeting with 12 participants. The requirement for participation in this stage is to know clinical support applications. After the presentation, the guests answered a questionnaire containing six multiple-choice (mandatory) questions and three (optional) discursive questions. On the technology professionals, the stage of evaluation occurred through three face-to-face meetings with a total of 35 participants. The requirement for participation in this evaluation stage is to be very familiar with the development of computational applications. After the presentation, the participants answered a questionnaire

[3] Available at http://www.heart.org/HEARTORG/Conditions/HighBloodPressure. Accessed 25 July 2018.

containing eight multiple-choice (mandatory) questions and a discursive question (optional). In the following section, we discuss and present the results obtained through the answers contained in the applied questionnaires.

Concerning the specific questions presented to health professionals, most of the group who answered the questionnaire agree, partially or totally, that there is a shortage of health applications that use multiple data sources (75%), that the use of heterogeneous data can bring improvements to existing systems (91.7%), and that is interesting to develop new applications that integrate different information (100%). These percentages demonstrate that there is a space for the evolution of clinical support systems that use heterogeneous data. The evaluators of this area were also asked to highlight the types of tests that they consider relevant for clinical support applications. The most cited examples of data were laboratory tests, imaging (magnetic resonance, computed tomography, radiography), physical examination, vital signs, and prescriptions. All types of information indicated can be used in systems built based on the proposed model.

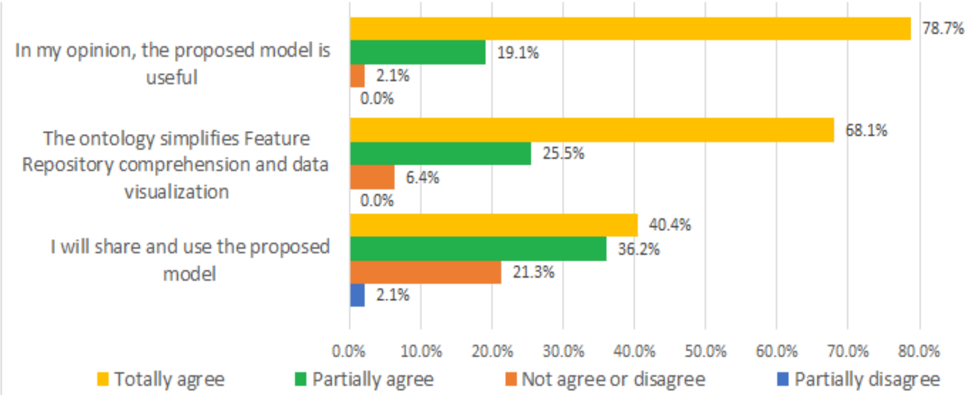


Figure 7: Equivalent responses grouping results.

Concerning the specific questions presented to technology professionals, a large set who participated in the evaluation agree that the model could simplify the development and expansion of clinical support systems (97.1%). That proposed model has a broader architecture than the others currently in use (80%). Also, the answers appoint that the model is clearly defined (85.7%) besides has modular and expansive architecture, which enables both the addition and replacement of elements independently (94.3%). Finally, almost all participants in this area understand that input and output interfaces allow the adaptation of already developed researches in a more straightforward way (97.1%). The last item in this specific questionnaire is a space for participants to insert suggestions for the presented model. The highlight in this area was the positive comments related to the use of the ontology as an alternative way to visualize the structure and content of the characteristics repository.

The last group of questions was the same in both questionnaires, and these questions were related directly to the proposed model and could be considered equivalent, which allows the comparison of results from both groups (health and information technology professionals) in Figure 7. The answers appointed that the proposed model is useful (97.9%) and that the participants would use and share it (76.6%). About 93.6% of answers appointed that the existence of an alternative way to visualize the data through ontologies, make the access and understanding simpler to the contents stored in Features Repository than the XML or JSON visualizations of HL7 FHIR pattern used in prototypes.

4. Discussion

The current work made two prototypes of clinical support applications, making use of information extracted from the processing of different data types. The first one performs medical report checking using as input the computed tomography images and their corresponding text. The second prototype shows the generation of a preliminary textual report of the patients using as input medical images and sensors data. In both cases, each element was independently implemented, either integrally by the authors or based on applications already validated and well-known in the literature. Therefore, it shows the modular and extensible features of the proposed architecture. Hence, the elements can be replaced by new software features without requiring changes in other components. The use of different data types as input and the use of a recognized representation pattern, such as that used in prototypes (HL7 FHIR), fill the research gap described. It also indicated the option to interact with external databases, allowing the use of information in other repositories, or sharing the data stored in the local repository with third-party applications.

The HDPCSA model evaluation was through a questionnaire applied to 47 specialists from health and computing domains. The health specialists' answers showed that there is a shortage of computational applications that make use of multiple input sources in an integrated way besides that the use of different data sources could contribute to the achievement of advances and improvements in the patient’s quality of care. Also, the technology professionals highlighted model benefits such as modularity and the possibility of adding new features or extending the existing ones. Both specialists groups highlight the use of ontology as an alternative way to visualize the data as a positive, making the access and understanding more simple to the contents stored in Features Repository.

The modular and expandable HDPCSA model architecture allows the use of features extracted from the processing of different input types in clinical support applications. Besides, the model provides a framework to assist in the medical application development and integration, making the use of different data types simpler and allowing the addition of new software features to already developed research. This kind of architecture could encourage the use of various works in a complementary way, making them more complete and broader. Therefore, it could increase the integration between existing systems and the development of new applications, thus fostering the computational support in the clinical area.

We highlight the benefits of the model integration with distinct applications and the ability to add both new features and resources, enabling the reuse of already developed methods. In related work, these opportunities were not observed, besides each research deal in a distinct way with the obtained information and the results, not having discussions on how they can share them. The HDPCSA model differential is the layer called Feature Repository, which is composed of the input and output interfaces, representation pattern, external database communication, and ontology. These interfaces promote compatibility between feature extraction, data repository, and applications because they remove some complexity related to information storage patterns. The ontology element is another differential because it allows both the access and view of repository contents in an alternative way.

The main scientific contribution obtained in this current study is the definition of the HDPCSA model elements and relationships, which aims to establish a reference structure that assists experts in building clinical decision support systems using the information of different types and data sources (heterogeneous data). The model also aims to make it easier to extend existing applications with the use of already developed studies, making them more complete.

5. Conclusion

The current study proposed the HDPCSA model, modular, and expandable architecture that makes possible the use of different data formats as input. The extracted information is stored in a structured way to allow the development of clinical decision support applications. As showed by developed prototypes, the use of many data sources made it possible to build a system that makes use of extracted data in a complementary way, improving aspects such as efficiency, accuracy, and personalization of clinical activities.

In this study, we tried to mitigate the research bias and sought to obtain an outline of the current literature related to clinical decision support applications and models that integrates information into clinical applications, such as diagnosis support, patients management, research, and administrative systems. We highlight as main contributions of current work (1) the HDPCSA model architecture, elements, and relationships ; (2) a layer called Feature Repository, which is composed by the input and output interfaces, representation pattern, external database communication, and ontology; (3) the interfaces which are the most prominent elements since they are responsible for the compatibility between the feature extraction, data repository, and applications, while they remove the complexity related to the specification of the information storage pattern; (4) the ontology element that allows access and views the content of repositories in an alternative way.

As future work, we highlight the development of a software component to dynamically create the data manipulation methods available at the input and output interfaces. We developed in the prototypes only the methods necessary to evaluate the proposed model. Therefore, these features could make the addition of an element more accessible to manipulate data types that do not exist in the repository.

References

Fichman RG, Kohli R, Krishnan R. Editorial Overview—The Role of Information Systems in Healthcare: Current Research and Future Trends. Information Systems Research 2011;22(3):419–428.

El-Dahshan ESA, Mohsen HM, Revett K, Salem ABM. Computer-aideddiagnosis of human brain tumor through MRI: A survey and a new algorithm. Expert Systems with Applications 2014;41(11):5526–5545.

Bozkurt S, Gimenez F, Burnside ES, Gulkesen KH, Rubin DL. Using automatically extracted information from mammography reports for decision support. Journal of Biomedical Informatics 2016;62:224–231.

Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology 2016;.

Topaz M, Lai K, Dowding D, Lei VJ, Zisberg A, Bowles KH, Zhou L. Automated identification of wound information in clinical notes of patients with heart diseases. International Journal of Nursing Studies 2016;64:25–31.

Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, Huang CS, Shen D,Chen CM. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Scientific Reports 2016;6.

Firmino M, Angelo G, Morais H, Dantas MR, Valentim R. Computeraided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. BioMedical Engineering Online 2016;15(1).

Williams M, Wu F, Kazanzides P, Brady K, Fackler J. A modular framework for clinical decision support systems. ACM SIGBED Review. 2009;6(2):1–11.

Puppala M, He T, Chen S, Ogunti R, Yu X, Li F, Jackson R, WongSTC. METEOR: An Enterprise Health Informatics Environment to Support Evidence-Based Medicine. IEEE Transactions on Biomedical Engineering 2015;62(12):2776–2786.

Sylvestre E, Bouzillé G, Chazard E, His-Mahier C. .Combining information from a clinical data warehouse and a pharmaceutical database to generate a framework to detect comorbidities in electronic health records. BMC Medical Informatics and Decision Making 2018;.

Maglogiannis I, Goudas T, Billiris A, Karanikas H, Valavanis I, PapadodimaO, Kontogianni G, Chatziioannou A. Redesigning EHRs and Clinical Decision Support Systems for the Precision Medicine Era. In: Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS) - EANN '15. ISBN 978-1-4503-3580-5; 2015:.

Weese J, Lorenz C. Four challenges in medical image analysis from anindustrial perspective. Medical Image Analysis 2016;33:44–49.

Kansagra AP, Yu JPJ, Chatterjee AR, Lenchik L, Chow DS, Prater AB,Yeh J, Doshi AM, Hawkins CM, Heilbrun ME, Smith SE, Oselkin M, Gupta P, Ali S. Big Data and the Future of Radiology Informatics. Academic Radiology 2016;23(1):30–42.

Morrison JJ, Hostetter JM, Aggarwal A, Filice RW. Constructing a Computer-Aided Differential Diagnosis Engine from Open-Source APIs. Journal of Digital Imaging 2016;29(6):654–657.

Zhao J, Papapetrou P, Asker L, Boström H. Learning from heterogeneoustemporal data in electronic health records. Journal of Biomedical Informatics 2017;65:105–119.

Pampouchidou A, Marias K, Yang F, Tsiknakis M, Pediaditis M, Manousos G, Meriaudeau F, Simos P. Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text. In: AVEC '16. New York,: ACM Press. 2016:27–34.

Wen H, Liu Y, Rekik I, Wang S, Chen Z, Zhang J, Zhang Y, Peng Y, He H. Multi-modal multiple kernel learning for accurate identification of Tourette syndrome children. Pattern Recognition 2017;63:601–611.

Hazlehurst BL, Kurtz SE, Masica A, Stevens VJ, McBurnie MA, PuroJE, Vijayadeva V, Au DH, Brannon ED, Sittig DF. CER Hub: An informatics platform for conducting comparative effectiveness research using multi-institutional, heterogeneous, electronic clinical data. International Journal of Medical Informatics 2015;84(10):763–773.

Mourão A, Martins F, Magalhães J. Multimodal medical information retrieval with unsupervised rank fusion. Computerized Medical Imaging and Graphics 2015;39:35–45.

Vizza P, Guzzi PH, Veltri P, Cascini GL, Curia R, Sisca L. GIDAC: A prototype for bioimages annotation and clinical data integration. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine. ISBN 978-1-5090-1611-2; 2016:1028–1031.

Riegler M, Lux M, Griwodz C, Spampinato C, de Lange T, Eskeland SL,Pogorelov K, Tavanapong W, Schmidt PT, Gurrin C, Johansen D, Johansen H, Halvorsen P. Multimedia and Medicine: Teammates for Better Disease Detection and Survival. Proceedings of the 2016 ACM on Multimedia Conference 2016;:968–977.

Allwell-Brown E. A Comparative Analysis of HL7 FHIR and openEHRfor Electronic Aggregation, Exchange and Reuse of Patient Data in Acute Care. Ph.D. thesis; Stockholm University; 2016.

Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, RybickiFJ, Mitsouras D. Natural Language Processing Technologies in Radiology Research and Clinical Applications. RadioGraphics 2016;36(1):176–191.

Aarhus: Aarhus Reis A, Rönnau RF, Pohren N, Spaniol AR, Becker J, Perez Â, Bez MR,de Oliveira RF. Ferramenta para apoio ao diagnóstico baseada em processamento de imagens de tomografia computadorizada do tórax. IV ERCASEscola Regional de Computação Aplicada à Saúde 2016;:20–21.

Sund-Levander M, Forsberg C, Wahren LK. Normal oral, rectal, tympanicand axillary body temperature in adult men and women: A systematic literature review. Scandinavian Journal of Caring Sciences 2002;16(2):122– 128.

Davis FD. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly 1989;.

Venkatesh V, Davis FD. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Management Science. 2000.

Likert R. A technique for the measurement of attitudes. Archives of Psychology 1932.