A Data Analytics Pipeline for Smart Healthcare Applications

Chonho Lee, Seiya Murata, Kobo Ishigaki, and Susumu Date

Abstract The rapidly increasing availability of healthcare data is becoming the driving force for the adoption of data-driven approaches. However, due to a large amount of heterogeneous dataset including images (MRI, X-ray), texts (doctor's note) and sounds, doctors still struggle against temporal and accuracy limitations when processing and analyzing such big data using conventional machines and approaches. Employing advanced machine learning techniques on big healthcare data anlaytics supported by Petascale high performance computing resources is expected to remove those limitations and help find unseen healthcare insights. This paper introduces a data analytics pipeline consisting of data curation (including cleansing, annotation, and integration) and data analytics processes, necessary to develop smart healthcare applications. In order to show its practical use, we present sample applications such as diagnostic imaging, landmark extraction and casenote generation using deep learning models, for orthodontic treatments in dentistry. Eventually, we will build smart healthcare infrastructure and system that fully automate the set of the curation and analytics processes. The developed system will dramatically reduce doctor's workload and is smoothly expanded to other fields.

1 Introduction

Increasing demand and costs for healthcare, exacerbated by ageing populations, are serious concerns worldwide. A relative shortage of doctors or clinical manpower is also a big problem that causes their workload to increase and brings a challenge for them to provide immediate and accurate diagnoses and treatments for patients. Most of the medical practices are completed by medical experts backed by their own experiences, and clinical researches are conducted by researchers via painstaking

C. Lee (✉) • S. Date
Cybermedia Center, Osaka University, 5-1 Mihogaoka, Ibaraki, Osaka, Japan
e-mail: leech@cmc.osaka-u.ac.jp; date@cmc.osaka-u.ac.jp

S. Murata • K. Ishigaki
Graduate School of Information Science and Technology, Osaka University, 5-1 Mihogaoka, Ibaraki, Osaka, Japan
e-mail: murata.seiya@ais.cmc.osaka-u.ac.jp

M.M. Resch et al. (eds.), *Sustained Simulation Performance 2017*,
DOI 10.1007/978-3-319-66896-3_12

designed and costly experiments. Consequently, this has generated a great amount of interests and motivation in providing better healthcare through smarter healthcare systems.

Nowadays, a huge amount of healthcare data, called Electronic Health Records (EHR), has become available in various healthcare organizations, which are the fundamental resource to support medical practices or help derive healthcare insights. The increasing availability of EHR is becoming the driving force for the adoption of data-driven approaches. Efficient big healthcare data analytics supported by advanced machine learning (ML) and high performance computing (HPC) technologies brings the opportunities to automate healthcare related tasks. The benefits may include earlier disease detection, more accurate prognosis, faster clinical research advance and the best fit for patient management.

While the promise of big healthcare data analytics is materializing, there is still a non-negligible gap between its potential and usability in practice due to various factors that are inherent in the data itself such as high-dimensionality, heterogeneity, irregularity, sparsity and privacy. To make the best analytics, all the information must be collected, cleaned, integrated, stored, analyzed and interpreted in a suitable manner. The whole process is a data analytics pipeline where different algorithms or systems focus on different specific targets and are coupled together to deliver an end-to-end solution. It can also be viewed as a software stack where at each phase there are multiple solutions and the actual choice depends on the data type (e.g. image, sound, text, or sensor data) or application requirements (e.g. predictive analysis or cohort analysis).

In this paper, we describe the data analytics pipeline consisting of *data curation phase* with cleansing, annotation and integration, and *data analytics phase* with analytics methods and visualization tools, which are necessary processes to develop healthcare applications. In order to show its practical use, we present three example applications using deep learning methods for orthodontic treatments in dentistry. The applications try to automate the following tasks such as (1) computing Index of Orthodontic Treatment Needs (IOTN)[1] from facial and oral photos; (2) extracting facial morphological landmarks or features from X-rays called Cephalograms; and (3) generating casenote where the first doctor's observation is written based on the diagnostic imaging such as (1) and (2).

The remainder of this paper is organized as follows. Section 2 introduces some requirements in handling healthcare data and describes the proposed data analytics pipeline consisting of data curation and analytics processes. Section 3 presents example applications using deep learning models, such as diagnostic imaging, landmark extraction and casenote generation for orthodontic treatments in dentistry, followed by conclusion.

[1] IOTN [1] is one of the severity measures for malocclusion and jaw abnormality, which determines whether orthodontic treatment is necessary.

2 Healthcare Data Analytics Pipeline

This section describes the proposed healthcare data analytics pipeline. As illustrated in Fig. 1, it consists of two phases, *data curation phase* (Sect. 2.1) and *data analytics phase* (Sect. 2.2). Processing big data is supported by high performance computing resources.

2.1 Data Curation Phase

Before available data is directly processed for analysis, data needs to go through several steps to refine it according to application requirements. Data curation phase prepares necessary data in a suitable format for further analysis. Firstly, data needs to be acquired and extracted from various data sources. Secondly, obtained raw data is probably heterogeneous, composed of structured, unstructured and sensor data, and also typically noisy due to inaccuracies, missing, biased evaluations, etc.

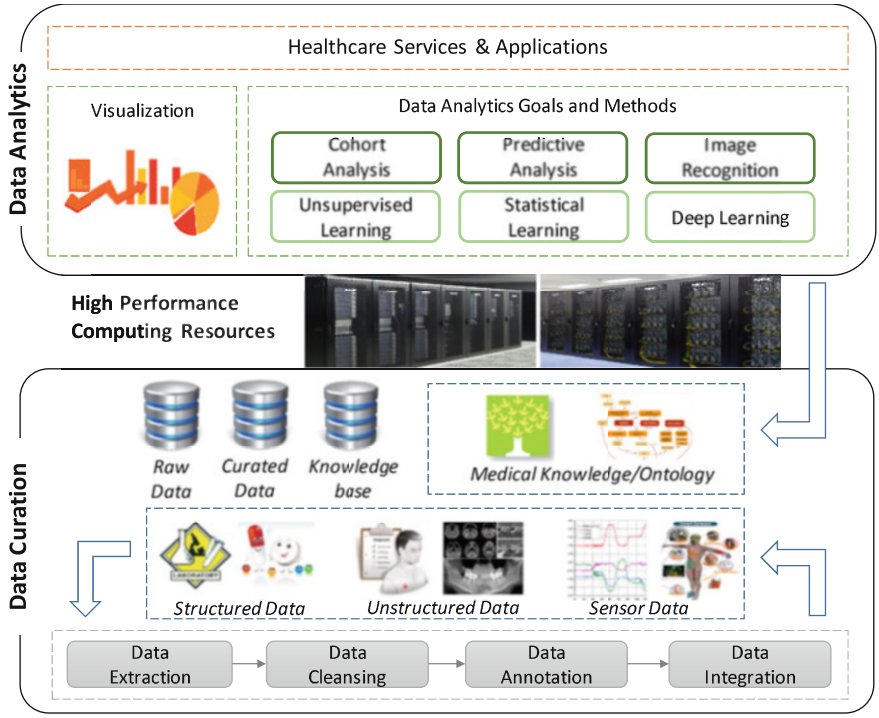

Fig. 1 An illustration of the proposed data analytics pipeline consisting of data curation and analytics, necessary processes to develop healthcare applications

Hence, data cleansing is required to remove data inconsistencies and errors. Thirdly, data annotation with medical experts' assistance contributes to the effectiveness and efficiency of this whole process. Fourthly, data integration combines various sources of data to enrich information for further analysis. Finally, the processed data is modelled and analyzed, and then analytics results are visualized and interpreted.

2.1.1 Data Type

Healthcare data, e.g., electric healthcare records (EHR), mainly includes three types of data, namely structured, unstructured and sensor data. *Structured data* includes socio-demographic information and medical features such as diagnoses, lab test results, medications and procedures. Those elements are typically coded in pre-defined forms by a hierarchical medical classification system or IDC-9 and currently IDC-10,[2] and drug databases like First Databank and SNOMED (Systematic Nomenclature of Medicine). *Unstructured data* does not have a specific data model, which includes medical status in a free-text form (e.g., doctors' notes and medical certificates) and non-textual form such as images (e.g., MRI, X-rays) and sounds. *Sensor signals* or data streams are also common in healthcare data with the wide use of sensor devices for monitoring and better response to the situational needs. With the advancement in sensor technology and miniaturization of devices, various types of tiny, energy-efficient and low-cost sensors are expected to be widely used for improving healthcare [2, 3]. Monitoring and analyzing such multi-modal data streams are useful for understanding the physical, psychological and physiological health conditions of patients.

2.1.2 Data Cleansing

Available raw data is typically noisy due to several reasons such as inaccuracies, missing data, erroneous inputs, biased evaluations, etc. Sensor data is also inherently uncertain due to lack of precisions, failure of transmission and instability of battery life, etc. Thus, data cleansing is expected to improve data quality assessed by its accuracy, validity and integrity, which leads to reliable analysis results. It is essentially required to (1) identify and remove inaccurate, redundant, incomplete and irrelevant records from collected data and (2) replace or interpolate incorrect and missing records with reasonably assigned values. This requires us to understand the healthcare background and work with domain experts to achieve better cleansing performance.

[2]International Statistical Classification of Diseases and Related Health Problems.

2.1.3 Data Annotation

Incompleteness is a common issue in terms of data quality. Although the uncertainty of data can be resolved by model inference using various learning techniques, most healthcare data is inherently too complex to be inferred by machines using limited information. In such cases, enriching and annotating data by medical experts are the only choice to help the machine to correctly interpret data. However, the acquisition of supervised, annotated information results in an expensive exploitation of data.

Active learning is one of the approaches to reduce the annotation cost while learning algorithms achieve higher accuracy with few labelled training data. It aims to only annotate the important, informative data instances while inferring others, and thereby the total number of annotated data is significantly reduced. The general solutions may include reducing the uncertainty of training models by uncertainty sampling [4], Query-By-Committee [5], maximizing the information density among the whole query space [6]. Another approach may be to borrow knowledge from related domain(s) such as transfer learning [7]. However, the aforementioned methods have limitations in real healthcare applications due to healthcare data volume, complexity and heterogeneity. The automation of data annotation is still a challenging problem.

2.1.4 Data Integration

Data integration is the process of combining heterogeneous data from multiple sources to provide users with a unified view of these data. Gomez et al. [8] explores the progress made by the data integration community, and Doan [9] introduces some principles as well as theoretical issues in data integration.

Typically, EHR integrates heterogeneous data from different sources including structured data such as diagnoses, lab tests, medications, unstructured free-text data like discharge summary, image data like MRI, etc. Healthcare sensor data is generated by various types of sensor/mobile devices at different sampling rates. The heterogeneity of abundant data types brings another challenge when integrating data streams due to a tradeoff between the data processing speed and the quality of data analytics. The high degree of multi-modality increases the reliability of analytics results, but it requires longer data processing time. The lower degree of multi-modality will improve data processing speed but degrade the interpretability of data analytics results. The efficient data integration helps reduce the size of data to be analyzed without dropping the analysis performance (e.g., accuracy).

2.2 Data Analytics Phase

Data anlaytics phase (the upper box of Fig. 1) applies different analytics methods into the curated data to retrieve medical knowledge. Visualization techniques may also be used to get better understanding of the data. Utilizing high performance

computing resources, we can improve the efficiency of data analysis especially when dealing with a large scale of data. For data privacy, on-demand secure network connection will be established, in which data is located or transferred to compute nodes when only needed. Right after the computation, the connection will be disconnected, and the data will be deleted.

2.2.1 Analytics Methods

Among a variety of anlaytics methods, the actual choice of algorithms or solutions depends on the data type (e.g. image, sound, text, sensor data) and/or application requirements (e.g. cohort analysis, predictive analysis, image recognition). In this section, we shall introduce a few basic methods to solve some healthcare problems as shown in Fig. 1.

- *Cohort Analysis*: Cohort analysis is a technique to find risk factors in a particular group of people, who have certain attributes or conditions such as birth, living area, life style, medical records, etc. The group is compared with another group who are not affected by the conditions. Long term statistical investigation will assess the significant differences between them. For the cohort analysis, clustering or unsupervised learning is the most popular method to divide people into particular groups under the certain conditions. For example, Sewitch [10] identifies multivariate patterns of perceptions using clustering method. Five different patient clusters are finally identified and statistically significant inter-cluster differences are found in psychological distress, social support satisfaction and medication non-adherence.
- *Predictive Analysis*: Disease progression modeling (DPM) is one of the predictive analysis, which employs computational methods to model the progression of a specific disease [11]. Reasonable prediction using DPM can effectively delay patients' deterioration and improve their healthcare outcomes. Typically, statistical learning methods are applied to find a predictive function based on historical data, i.e., the correlation between medical features and condition indicators. For example, Schulze [12] uses a multivariate Cox regression model that computes the probability of developing diabetes within 5 years based on anthropometric, dietary, and lifestyle factors.
- *Image Recognition*: Analyzing medical images such as X-ray, MRI, etc. are beneficial for many medical diagnosis and a wide range of the studies focus on classification or segmentation tasks. The recent breakthrough in image recognition technology using deep convolutional neural network (CNN) model [13] brings further improvement in diagnostic imaging that can diagnose the presence of tuberculosis in chest X-ray images [14], detect diabetic retinopathy from retinal photographs [15], as well as locate breast cancer in pathology images [16]. A model that combines deep learning algorithms and deformable models is developed in [17] for fully automatic segmentation of the left ventricle from cardiac MRI datasets.

Although a variety of data analytics and machine learning (ML) tools are available, there still exists an obstacle for doctors to fully utilize the tools due to the lack of the usability. Besides, it is difficult for them to manage compute resources suitable for executing the analytics methods. Hence, in near future, high performance infrastructure and system that operate fully or semi-fully automated big healthcare data curation and analytics, are eagerly desired in medical environment so that any doctors and/or researchers efficiently conduct their own data analytics.

3 Smart Orthodontic Treatment in Dentistry

The recent breakthrough in image recognition using deep learning techniques brings further improvement in diagnostic imaging. The diagnostic imaging is eagerly desired in the field of orthodontics as well, along with increasing demands for dental healthcare, becoming one of the regular life health factors. For example, the remote diagnostic imaging can evaluate malocclusion and jaw abnormality that are the causes of masticatory dysfunction, apnea syndrome and pyorrhea, etc.

In orthodontic clinic, a patient is generally taken his/her facial and oral images from all directions (as shown in Fig. 2) and given the first observation. Looking at the images, doctors spend time discussing the observation and create the medical

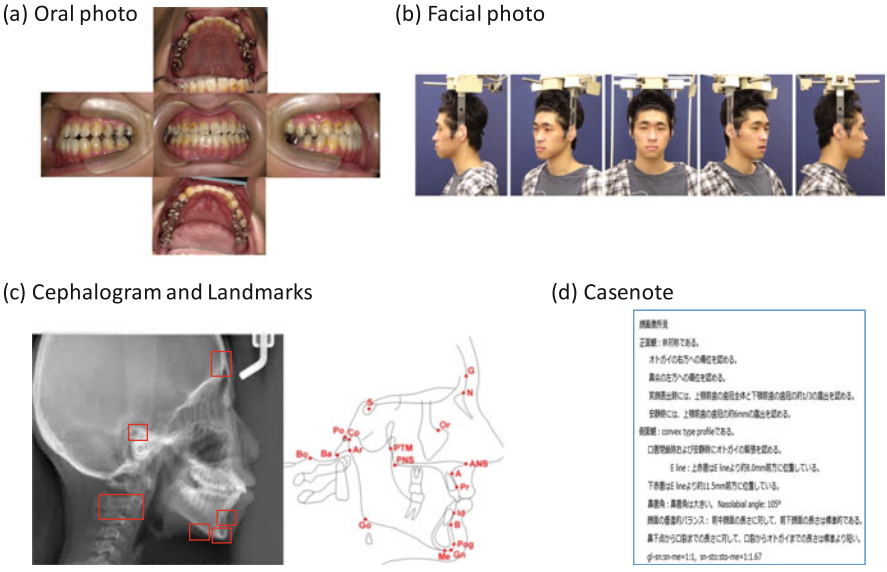

Fig. 2 Sample dataset collected for orthodontic treatments. (**a**) Oral images taken from five directions, (**b**) facial images taken from five directions, (**c**) Cephalogram [18], Morphological landmarks [19] and example patch images (*red boxes*) (**d**) casenote or the first doctor's observation

records including diagnosis, treatment plan and progress checkup, etc. This process is certainly necessary for providing objective diagnosis that is important for both doctors and patients because the diagnosis directly affects to the treatment plan, treatment priority, and insurance coverage; but, it takes a great deal of time. In Osaka University Dental Hospital, over thousands of patient visits are counted including about 100 new patients per year. It is overburdened for doctors to properly manage a sequence of tasks such as diagnosis, treatment, progress checkup and counselling for all patients. Especially, doctors spend a lot of time and effort to diagnose by manually looking at massive number of images; for instance, it takes about 2–3 h for just one patient's case. The automation of diagnostic imaging is highly expected to assist doctors reducing their workload and providing objective diagnosis.

We try to develop a high performance infrastructure that operates big healthcare data analytics systems, especially for orthodontic treatments in dentistry, which automate medical tasks such as diagnostic imaging, landmark extraction and casenote generation. Due to a large amount of heterogeneous dataset including images (facial/oral photo, X-rays) and texts (casenote), doctors struggle against temporal and accuracy limitations when processing and analyzing those data using conventional machines and approaches. We believe that advanced machine learning techniques supported by Petascale high performance computing infrastructure remove those limitations and help find unseen healthcare insights. We evaluate the practical use of DL models in medical front and show its effectiveness.

In this paper, we consider three example applications dealing with medical images and casenotes (text), as illustrated in Fig. 3. Section 3.1 (App1) explains how to compute the score of orthodontic treatment needs from facial and oral images. Section 3.2 (App2) shows how to retrieve facial morphological landmarks from X-rays called Cephalograms. Section 3.3 (App3) describes how to generate casenotes where the first doctor's observation is written.

3.1 Assessment of Treatment Need (App 1)

This application tries to automate the assessment of Index of Orthodontic Treatment Needs (IOTN) [1], one of the severity measures for malocclusion and jaw abnormality, which determines whether orthodontic treatment is necessary. Providing orthodontic treatments at appropriate timing is very important for patients to prevent a masticatory dysfunction. Generally, a primary care doctor or general dentist assesses the IOTN of his/her patient, and if the severity is high, he/she refers the patient to the other specialist for further treatments. However, there is a problem that many patients tends to miss the appropriate treatment timing due to an incorrect assessment by an inexperienced doctor. The automation of the IOTN assessment helps provide an objective assessment and train such inexperienced doctors.

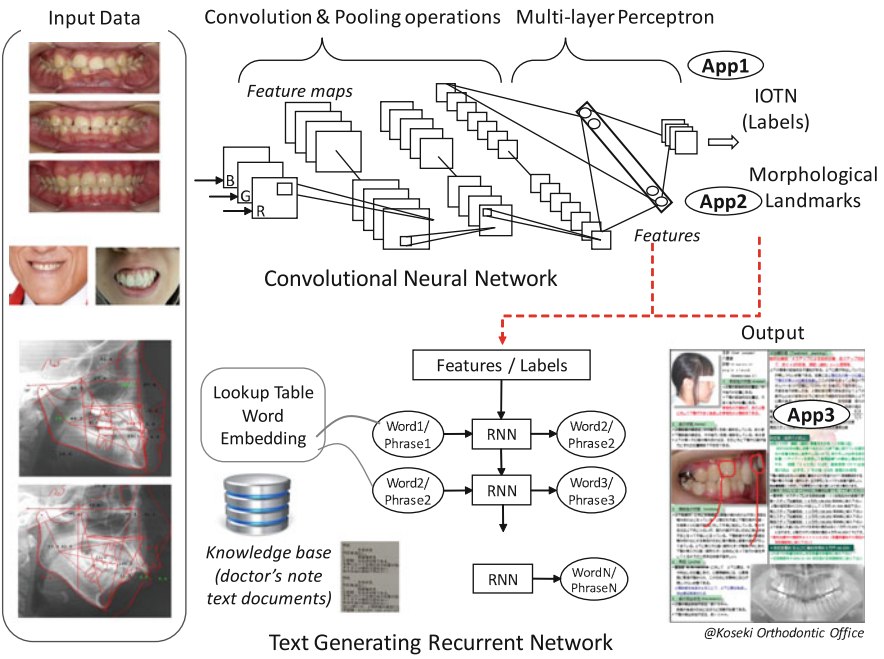

Fig. 3 An illustration of deep learning models that perform diagnostic imaging (App1), landmark extraction (App2) and casenote generation (App3) for orthodontic treatments

We collect oral and facial images of over a thousand patients, taken from five different directions, as shown in Fig. 2a, b. Unlike typical image classification problems where each image is paired with one class or label, one class, i.e., a severity value, is paired with a set of images of a patient. We design a parallel convolutional neural network (CNN) model that independently runs multiple CNNs, each of which deals with images taken from each direction, and then concatenates feature vectors (i.e., outputs of the multiple CNNs). The concatenated feature vector is input to a multi-layer perceptron whose output is one of IOTN levels.

3.2 Morphological Landmarking (App 2)

Cephalometric analysis is also a significant diagnosis necessary for further orthodontic treatments. It helps in classification of skeletal and dental abnormalities, planning treatment of an individual, and predicting growth related changes. This application tries to automate facial morphological landmark detection in Cephalometric X-ray images (Fig. 2c).

We consider a landmark as an image patch, i.e., a sub-image of the whole cephalometric image, which includes the landmark. Collecting a bunch of patches for several landmarks from different patients, we train a CNN-based model to recognize whether given sub-images (i.e., regions) include the landmarks. The model outputs an N dimensional vector at the last layer, and each vector element represents the probability that given patch includes a corresponding landmark. Compared to image patches, the whole cephalometric image resolution (or the number of pixels) is normally high. In order to speed up the recognition speed, we distribute the sub-images over multiple nodes and run the model independently. Candidate regions of landmarks will be selected from each of the nodes. Then, based on the probability associated to the candidates, we determine the most likely region as the target landmark.

3.3 Casenote Generation (App 3)

In general, generating casenote is a time consuming work for doctors. For instance, doctors regularly gather together to discuss the results of diagnostic imaging such as App1 and App2, and prepares casenotes including the diagnosis, treatment plan and priority, etc. It often takes about a few hours to generate the casenote for just one patient. This application tries to automate the process of the casenote generation.

We collect over a thousand of casenotes (Fig. 2d) in addition to the oral and facial images of the corresponding patients. Inspired by a related work [20] that describes the content of an image, we design a hybrid model using CNN and Recurrent neural network (RNN) that inputs both images and casenotes. The hybrid model will lean how the casenote was written according to the diagnostic imaging; in other words, the model will find the association rules between features in images and words in casenotes.

4 Conclusion

In this paper, we summarize some requirements in handling healthcare data and the data analytics. We propose a data analytics pipeline that consists of data curation with cleansing, annotation and integration, and data analytics processes using several analytics methods and visualization tools. In order to verify the practical use of such data curation and analytics methods in medical front and show its effectiveness, we present example healthcare applications such as diagnostic imaging, landmark extraction and casenote generation using deep learning models, for orthodontic treatments in dentistry.

In future work, we will conduct rigorous experiments to evaluate the results of the curation and analytics and whether they satisfy the application requirements. Eventually, we will build smart healthcare infrastructure and system that fully

or semi-fully automate the set of the curation and analytics processes, where any doctors and/or researchers efficiently conduct their own data analytics, which dramatically reduces their workload. This will be smoothly expanded to other fields such as otolaryngology (ear and nose) and ophthalmology (eye).

Acknowledgements The authors would like to thank Prof. Kazunori Nozaki in Osaka University Dental Hospital, for managing and providing medical dataset for experiments. We also thank Prof. Chihiro Tanikawa in Department of Orthodontics & Dentofacial Orthopedics, Osaka University Dental Hospital, for lending her expertise on the orthodontic treatments in dentistry.

References

1. Brook, P.H., Shaw, W.C.: The development of an index of orthodontic treatment priority. Eur. J. Orthod. **11**(3), 309–320 (1989)
2. Caytiles, R.D., Park, S.: A study of the design of wireless medical sensor network based u-healthcare system. Int. J. Bio-Sci. Bio-Technol. **6**(3), 91–96 (2014)
3. Filipe, L., Fdez-Riverola, F., Costa, N., et al.: Wireless body area networks for healthcare applications. Protocol stack review. Int. J. Distrib. Sens. Netw. (2015). http://dx.doi.org/10.1155/2015/213705
4. Sharma, M., Bilgic, M.: Evidence-based uncertainty sampling for active learning. Data Min. Knowl. Discov. **31**(1), 164–202 (2017)
5. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory (1992)
6. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP08 (2008)
7. Ma, Z., Yang, Y., Nie, F., Sebe, N., Yan, S., Hauptmann, A.: Harnessing lab knowledge for real-world action recognition. Int. J. Comput. Vis. **109**(1–2), 60–73 (2014)
8. Gomez-Cabrero, D., Abugessaisa, I., Maier, D., et al.: Data integration in the era of omics: current and future challenges. BMC Syst. Biol. **8**(2) (2014). http://dx.doi.org/10.1186/1752-0509-8-S2-I1
9. Doan, A., Halevy, A., Ives, Z.: Principles of Data Integration. Elsevier, Amsterdam (2012)
10. Sewitch, M.J., Leffondré, K., Dobkin, P.L.: Clustering patients according to health perceptions: relationships to psychosocial characteristics and medication nonadherence. J. Psychosom. Res. **56**(3), 323–332 (2004)
11. Mould, D.: Models for disease progression: new approaches and uses. Clin. Pharmacol. Ther. **92**(1), 125–131 (2012)
12. Schulze, M.B., Hoffmann, K., Boeing, H., et al.: An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. Diabetes Care **30**(3), e89 (2007)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E. : Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 1 (2012)
14. Lakhani, P., Sundaram, B.: Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology (2017). http://dx.doi.org/10.1148/radiol.2017162326
15. Gulshan, V., Peng, L., Coram, M., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J. Am. Med. Assoc. **316**(22), 2402–2410 (2016)
16. Janowczyk, A., Madabhushi, A.: Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J. Pathol. Inf. **7**(292), 29 (2016)

17. Avendi, M.R., Kheradvar, A., Jafarkhani, H.: A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. Med. Image Anal. **30**, 108–119 (2016)
18. Jimbocho Orthodontic clinic: www.jimbocho-ortho.com. Accessed June 2017
19. Grau, V., Alcaniz, M., Juan, M., Knoll, C.: Automatic localization of cephalometric landmarks. J. Biomed. Inform. **34**(3), 146–156 (2001)
20. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR15 (2015)