The Leir Retreat Center
Ridgefield, Connecticut

Symposium on Data Science for Healthcare (DaSH)

Proceedings of the 2017 Symposium

Data Quality:
Practices, Technologies and Implications

October 19 – 20, 2017
The Leir Retreat Center, Ridgefield, CT, USA

NJIT New Jersey's Science & Technology University

The Leir Retreat Center

Acknowledgement

Host: The Leir Retreat Center

Sponsor: The Leir Charitable Foundations

Supporter: New Jersey Institute of Technology

Organization

Conference Chair

- Yi Chen - Henry J. Leir Chair in Healthcare, Associate Professor, Martin Tuchman School of Management, NJIT

Advisory Board

- Susan Davidson - Founding Co-Director, Center for Bioinformatics, Weiss Professor, Computer & Info. Science, UPenn Chair, CRA bd of Directors
- Kathy Grise - Senior Program Director, Big Data Initiatives, Future Directions, IEEE
- Bradford Hesse - Chief, Health Communication & Informatics Research National Cancer Institute, National Institute of Health
- Cheickna Sylla - Professor, Tuchman School of Management, NJIT
- René Bastón - Executive Director, Northeast Big Data Innovation Hub, Columbia University

Program Committee

- David G. Belanger - Howe School of Technology Management, Stevens Institute of Technology
- Yi Chen - Martin Tuchman School of Management, NJIT
- Soon Ae Chun - School of Business, The City University of New York
- Gordon Gao - Robert H. Smith School of Business, University of Maryland
- James Geller - Ying Wu College of Computing Sciences, NJIT
- Andrea L. Hartzler - Department of Biomedical Informatics and Medical Education, University of Washington
- Ketan Mane - Health Informatics, Kaiser Permanente
- Rao Praveen - Department of Computer Science & Electrical Engineering, University of Missouri-Kansas City
- Fei Wang - Weill Cornell Medical School, Cornell University
- Christopher C. Yang - College of Computing & Informatics, Drexel University

Web Chair

- Jinhe Shi - PhD Candidate, Ying Wu College of Computing Sciences, NJIT

Table of Contents

Harnessing the Power of Data in Healthcare: Data as a Strategic Asset

CHRISTOPHER JOYCE, Anthem, Inc

A big-data revolution[1]is underway in healthcare, driven by an exponential growth in data from the digitization of existing data and the generation of new data. With this data expansion, healthcare organizations are harnessing the power of data to improve consumer & provider engagement, deliver better health outcomes, and drive down costs. Many organizations are finding success in using advanced analytics to deliver new value, but it is not all about the analytic models. The growth and complexity of data created by and available to healthcare organizations requires that data is managed as a strategic asset. This has put a greater emphasis on data governance, data quality and integrated data solutions within healthcare organizations to ensure they can continue to meet customer expectations, improve service delivery, and enable value creation opportunities through the use of advanced data analytics. In this presentation we will examine ways in which these rich data assets are being leveraged and what organizations are doing to manage this healthcare "data tsunami".

3rd Symposium on Data Science for Healthcare (DaSH). October 19-20, 2017, Leir Retreat Center, CT, USA.

Challenges in Capturing and Analyzing Data from Clinical Care

FRANK A. SONNENBERG, RUTGERS

In[1] theory, the advent of electronic records should provide a wealth of data for measuring and improving quality of care and generating new knowledge. This presentation will discuss the challenges in making this vision a reality which include the complexity of clinical data, lack of interoperability among health information systems and data entry burdens which can adversely affect productivity and the provider-patient relationship.

3rd Symposium on Data Science for Healthcare (DaSH). October 19-20, 2017, Leir Retreat Center, CT, USA.

Public Health Intelligence Platform from Social Health Records (SHR)

SOON AE CHUN, City University of New York

Social media and personal health monitoring devices (e.g., Fitbit) provide abundant patient-generated health-related data. These open health data, generated via patient engagement and sharing, are referred to as *Social Health Records* (*SHR*) as opposed to the EHR (Electronic Health Records) that are created and entered by clinicians. SHRs are changing the healthcare paradigm from the authoritative provider-centric model to a collaborative and patient-oriented healthcare framework. This chapter proposes an *SHR Integration and Analytics Framework* to leverage Social Health Records for gaining insights into population-level and individual-level healthcare practices and behaviors, as well as emotions.

1 INTRODUCTION

There is a large amount of health information available for any patient to address his/her health concerns. The freely available health datasets include open government health data sets, at the national, state or community level, such as OpenHealthdata.gov ranging from Medicare data to epidemiology; Web health resources curated by experts such as WebMD; and the personal health records shared by the patients on open or registered online social media services such as *PatientsLikeMe*. These are so-called open health data, which are readily accessible and downloadable. The patient generated and shared data include the conditions, treatments, side effects, health histories, and personal physical, psychological, emotional and relationship experiences of individual patients. This data resembles the Electronic Health Record (EHR), which is defined as an electronic version of a patient's medical history that is collected and maintained by the provider (e.g. clinicians) over time. The EHR system allows capturing the key administrative and clinical data relevant to that person's care, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports. Since the open online health records shared by patients or family care givers capture similar data about the patients, we call this Social Health Record (SHR) to distinguish from the closed EHR.

The SHR is capturing many instances of personal healthcare experiences, practices and other health-related behaviors, while the EHR is capturing the clinical data necessary to provide care. Even though a doctor prescribes a

medicine X, the patient may consume a substitute medicine Y. The intention of sharing the Social Health Records is support-oriented with information and experience sharing, while the EHR is primarily care-oriented to address the conditions. The SHR expresses emotional and psychological attitudes, opinions and comments in ordinary language riddled with ambiguities, while the EHR may capture mostly the factual statements in expert language to avoid vagueness or ambiguities. Another difference is that the EHR is hard to share, protected by the HIPAA and HITECH regulations and locked into different EHR systems which creates a silo-effect of causing difficulty of interoperation and sharing. On the other hand, the SHR is open system based on online services, so the data is easily shared over lightweight clients, e.g. a Web browser.

The EHR focuses on individual patients, and it is difficult to connect and aggregate EHRs of many patients, unless one has all the access privileges. On the other hand, the SHRs are inherently crowd source data due to their base in social media so they can reveal the aggregated information of the crowd. For instance, the forum entry SHR data in one community group (e.g. cancer patient groups) from many patients may easily reveal the major types of issues and popular treatment options of many patients. The SHRs can provide a unique opportunity to look into healthcare from the patients' perspectives to identify healthcare related issues and improve quality of care. The SHR data from the crowd can facilitate the ability to "connect the dots" among and across many patients and allow gaining public health intelligence and insights, such as detecting disease outbreaks, and understanding population-related health trends. The crowd sourced SHRs can be a great asset for public health intelligence. Some examples of potential healthcare benefits of aggregating and mining SHRs include:

- determine which health topics are of greatest current concern
- identify a high-risk group of patients
- identify health trends both in the general public and at the individual level
- identify how patients view or feel about particular treatments and practices
- track adverse drug events
- identify the perceived quality of healthcare services, e.g., most desirable outcomes
- create education campaigns and interventions
- offer insights into the relationship between an individual's health and their everyday lifestyles
- reveal patients' attitudes towards health

These datasets can help to assess and improve healthcare quality, as well as help to modify health-related policies. However, these data sets are not integrated and varied in formats and quality, an information seeker has to spend time visiting many, possibly irrelevant, Websites, and has to select information from each and integrate it into a coherent mental model. The public health intelligence requires,

as many applications do, integrated data from disparate data sources, to provide value for different communities and users concerned with questions about public health statistics, trends, correlations, and distributions.

2 Social Health Knowledge Graph

In order to query any individual social health related record or to gain public health intelligence, we developed a social health knowledge graph, which serves as integrated knowledge base consisting of health records, extracted from multiple user generated health contents on their social media data sources and data and expertise (knowledge) from other open health data sources. We use a lightweight ontology that contains the health record-related concepts and relationships and which serves as semantic schema for integration. We discuss an approach to integrating these openly available but widely dispersed health data sources, where health data is created and shared by patients voluntarily, and open knowledge and expertise shared by healthcare providers and professionals [4][5]. The goal of developing the integrated data sources is to provide answers to information and knowledge needs of end users, to provide insights on public health through diverse analytics on social behaviors, and behavior models learned from the social data to predict trends. The insights are presented to convey intuitive understanding of the public health trends and alerts for physicians, healthcare staff, health policy workers, and individual patients.

Each Social Health Record (SHR) is modeled as a Linked Data assertion represented as a triple <*subject, predicate, object*>, denoting the atomic knowledge unit which states that the "subject" entity is related to the "object" entity by the "predicate" relationship. The subject or object represents a class in the ontology, and a predicate is a property of a class or between classes which states the relationship in existence between two entities. To instantiate the health record model, we extracted the health-related concepts with their URI's and represented them as triples. In order to integrate disparate data sources, entity resolution is used to recognize the terms from different resources that actually represent the same concept. For instance, consider a term for a condition, "Human immunodeficiency virus," extracted from PatientsLikeMe and a term, "HIV," retrieved from the CDC website [1].

3 Social Health Analytics Platform & Applications

To enable end users like health officials or epidemiologists to draw public health intelligence to better understand the population's health status or to get data-driven insights into the social health behaviors, the social health analytics platform is proposed. Figure 1 shows the major components consisting of data extraction, linking and discovering additional links through inference to construct an integrated connected knowledge graph, and the analytics component where the

machine learning component builds the models to automate the data processing to not only summarize, but also to predict sentiments, and diseases that may be correlated with other diseases.

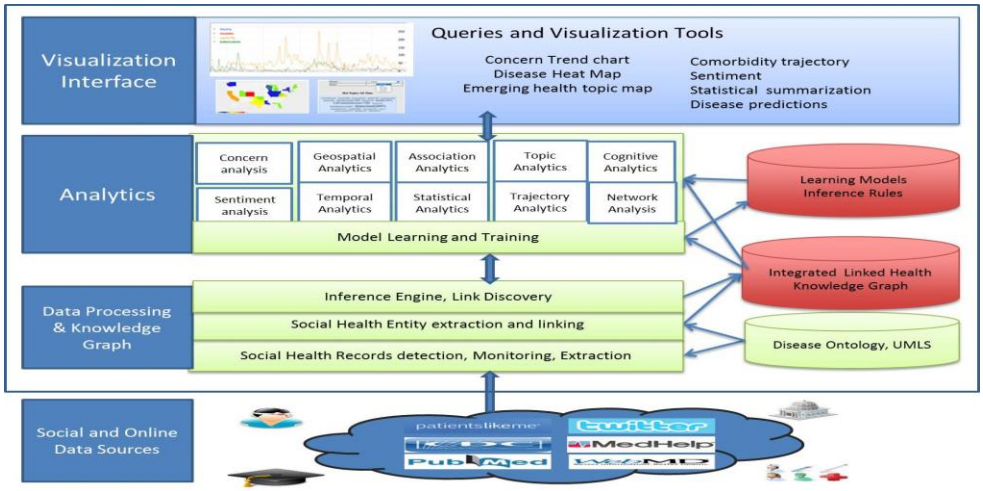


Fig. 1. Social Health Analytics Platform Architecture

Several applications are presented on the social health analytics platform, such as

- Social InfoButtons which provides the public health intelligence, e.g. social health practice, geospatial distribution, and temporal trends of disease.[8]
- Public health concern index which is a measure of public sentiments over disease spreads and effects [3,7]
- Comorbidity trajectory to predict public health trends and prevention [6]
- Detection of real time public health issues (e.g. drug abuse) [2, 9]

4 CONCLUSIONS

Public health intelligence can be gathered from the Social Health Records shared by individuals on online social media, combined with the authoritative data shared by medical experts. We have presented a Social Health Analytics Platform for enabling the use of semantics in the analysis of Social Health Records to gain population level health intelligence.

ACKNOWLEDGMENTS

The research work was partially funded by PSC-CUNY Research Foundation under the award numbers #64266 and #65232. We acknowledge our collaborators at NJIT, J. Geller and X. Ji, for their contributions to the work.

REFERENCES

[1] Behavioral Risk Factor Surveillance System. http://www.cdc.gov/brfss/. Accessed 04/14/2014

[2] Ji X, Chun SA and Geller J. Epidemic outbreak and spread detection system based on twitter data. Proceedings of the First international conference on Health Information Science. Beijing, China. 2012, p. 152-63.

[3] Ji X, Chun SA and Geller J. Monitoring public health concerns using Twitter sentiment classifications. Proceedings of IEEE International Conference on Healthcare Informatics. Philadelphia, PA. 2013, p. 335-44.

[4] Chun, S.A., MacKellar, B.: Social health data integration using semantic Web. Proceedings of the 27th Annual ACM Symposium on Applied Computing, Trento, Italy, (2012)

[5] Ji X, Chun SA, Cappellari P, and Geller J. (2017) Linking and Using Social Media Data for Enhancing Public Health Analytics, Journal of Information Science, Volume 43, Issue 2, April 2017: pp. 221-245.

[6] Xiang Ji, Soon Ae Chun and James Geller (2016) Predicting Comorbid Conditions and Trajectories using Social Health Records, IEEE Transactions on Nanobioscience Vol 15 Issue 4, June 2016:371-379.

[7] Ji, X., Chun, S.A., Wei, Z., Geller, J.: Twitter sentiment classification for measuring public health concerns. Social Network Analysis and Mining 5, 1-25 (2015)

[8] Ji, X., Chun, S.A., Geller, J.: Social InfoButtons: integrating open health data with social data using semantic technology. Proceedings of the Fifth Workshop on Semantic Web Information Management, New York, New York (2013)

[9] Phan, N, Chun, SA, Bhole, M & Geller J. Enabling Real-Time Drug Abuse Detection in Tweet, Proceedings of ICDE, Workshop on Health Data Management and Mining (HDMM) , San Diego, April 30, 2017: 1510-1514.

Adoption of Health Information Exchanges and Physicians' Referral Patterns: Are they Mutually Reinforcing?

SAEEDE EFTEKHARI, SUNY at Buffalo, Buffalo, NY

NIAM YARAGHI, Brookings Institution, Washington DC

RAM GOPAL, University of Connecticut, Storrs, CT

RAM RAMESH, SUNY at Buffalo, Buffalo, NY

This[1] research studies how Health Information Exchanges (HIE) implemented in the U.S. healthcare system impact physicians' referral patterns. Referrals are an important function of healthcare services, and HIE can significantly impact healthcare outcomes due to referrals. We contend that primary care physicians who are HIE members tend to refer their patients to specialists who are also members. Further, referrals between a member and a non-member influence the non-member to adopt HIE. To investigate this reciprocal association, we develop a novel methodology comprising of a Mechanism View and a Trajectory View of this association. While the mechanism view models causal and reverse-causal associations between HIE adoption and referral patterns using panel data, the trajectory view models the transformation process in which referrals and HIE adoption co-evolve between instances of panel observations. We establish that HIE adoption and referral patterns evolve concomitantly. This study has significant implications for healthcare policy-making.

Key Words: Health Information Exchanges (HIE), Physicians' Referral Network, Physicians' HIE Adoption

Health Information Exchange (HIE) platforms have become an important part of the United States healthcare reform since the enactment of Health Information Technology for Economic and Clinical Health Act (Vest et al. 2010). The main purpose of an HIE is to enable healthcare providers to access and share all relevant clinical information pertaining to patients' care. The absence of previous clinical information adversely affects the clinical quality and economic efficiency of healthcare services (Forster et al. 2003; Kohn et al. 2000). Through the HIE platforms, patients' medical history can be shared electronically with providers and

3rd Symposium on Data Science for Healthcare (DaSH). October 19-20, 2017, Leir Retreat Center, CT, USA.

organizations across the continuum of care (Vest et al. 2015). Availability of earlier medical information through HIE leads to more inclusive information and enhancement of quality of care (Branger et al. 1994; Kaelber et al. 2007; Smith et al. 2005).

Since the implementation of HIEs in the U.S healthcare system, a significant number of studies have proposed that these platforms can help physicians make better decisions, save more lives, and reduce costs (Frisse et al. 2012; Overhage et al. 2005; Vest et al. 2015;Walker et al. 2005). Specifically, a few studies have shown that HIEs can reduce unnecessary medical procedures. Bailey et al. (2013) showed that HIEs are associated with decreased diagnostic imaging. Lammers et al. (2014) also found evidence suggesting that using an HIE platform reduces repeat-imaging among patients visiting multiple emergency departments. Recently, Yaraghi (2015) empirically showed that HIE usage is associated with reduction in the expected total number of laboratory tests and radiology examinations ordered at the emergency departments. Although these studies demonstrate the significant impact of HIE platforms on healthcare outcomes, further investigations are needed to determine the effects of HIEs on the healthcare service processes leading to tangible outcomes. This research is motivated by two major shortcomings in the current literature on the impact of HIEs on healthcare outcomes.

First, while prior studies have focused on the impact of HIEs on the providers' decision to perform medical services on patients, there is no study that examines its impact on providers' decision to refer patients to other providers. Referral decision making is an important process in the framework of healthcare services, and consequently, HIEs could significantly impact the healthcare outcomes due to referrals. Referrals usually occur when a primary care physician determines that specialist care is needed during the course of patient care. In the United States, more than a third of patients are referred to specialists each year, and more than half of outpatient visits are specialist visits (Mehrotra et al. 2011). In a referral process, a primary care physician usually has to decide whom to refer, among a set of possible specialists. A physician's referral decision has important consequences on the clinical outcomes (Barnett et al. 2012). To the best of our knowledge, there has not been any study that focuses on whether a primary care physician's adoption of the HIE system impacts her referral decisions. In this paper, we propose and test a model to examine how the adoption of an HIE system by primary care physicians affects their subsequent choices of the specialists in the referral process.

Second, prior studies primarily focused on the benefits of the HIE system for insurance providers (payers) and patients while neglecting to examine the benefits of HIE for medical providers. In fact, the HIE system is a multisided platform that

connects four major sides: patients, electronic medical data providers (such as pathology and radiology centers), medical providers, and insurance companies (Yaraghi et al. 2014). One of the major challenges in community-wide growth of HIE platforms is that the benefits of HIE with regard to these four sides have neither been fully explored, nor sufficiently demonstrated. As a result, medical providers are not encouraged enough to adopt an HIE or to increase their level of HIE usage. Hence, there has been a growing interest among federal, state-level, and HIE business policy makers in promoting HIE adoption and usage by demonstrating its value to its different sides. This study contributes to this effort by investigating how HIE adoption can influence the referral decisions of primary care physicians and the shifts in the flow of patients that specialists receive as a consequences of such decisions.

This study is grounded in actual HIE adoption behaviors and referral practices of healthcare providers. Most prior studies on referral processes are survey-based. In this respect, this research adds the significant new dimension of an observational study using actual referral data that complements the current literature and enhances our understanding of referral patterns. Furthermore, this research develops a novel dual-perspective analytical methodology to understand the relationships between HIE adoption and referral patterns. We outline these research contributions as follows.

In this research, we use three publicly available datasets: the longitudinal referral dataset and the physicians' attribute dataset from the Centers for Medicare & Medicaid Services (CMS), and HIE adoption dataset provided by HEALTHeLINK, the regional health information organization of Western New York. Terming the physicians who have adopted HIE as HIE members, we essentially hypothesize that the primary care physicians who are HIE members tend to select the specialists who are also HIE members rather than non-members when they refer their patients. To examine this hypothesis, we also take into consideration the possibility of reverse causality: when a non-member physician interacts with a member in the referral process, the non-member is also influenced to adopt HIE as a result. To test this relationship, we develop an analytical methodology comprising of two perspectives: a Mechanism View and a Trajectory View of the relationship. The mechanism view models the causal and reverse-causal associations of the relationship between HIE adoption and referral patterns concomitantly using panel data analysis; the trajectory view models the same associations that could randomly occur between successive instances of panel observations using network data analysis. More specifically, the mechanism view explains the differences between the values of a dependent variable observed at different points in time using predictors in a causal modeling framework; however, the trajectory view explains

the same differences by using predictors in a network process evolution framework. These predictors are the factors that govern the underlying evolution of the network over time and are analogous to those in the causal modeling framework. Together, the two views complement each other by presenting macro- and micro-level perspectives on the underlying relationship.

The mechanism view involves a Difference in Difference (DID) analysis of the panel data to examine the impact of primary care physicians' adoption of HIE on their subsequent referral patterns. The trajectory view involves a Social Network Analysis (SNA) of the network structure underlying the panel data to examine the dynamics of referral patterns among HIE members. In this network, nodes represent primary care or specialist physicians, and the edges represent referrals. The network is stochastic since the edges would change over time. We employ the stochastic network modeling tool known as "Simulation of Investigation for Empirical Network Analyses (SIENA)" to study the changes in referrals among physicians over time. SIENA is a statistical tool designed to analyze longitudinal network data, i.e., two or more sets of observations over time (Snijders et al. 1997).

In this study, the mechanism view reveals the following findings: (a) when a physician adopts HIE, her tendency to refer to other HIE members increases, (b) when a non-member physician refers to HIE members, her tendency to adopt HIE increases, and (c) physicians who have adopted earlier have a higher tendency to refer to HIE members than those who adopted later. The trajectory view reveals: (i) physicians who are HIE members tend to select members in the referral network rather than non-members to refer, and (ii) non-member physicians who interact by way of referring or receiving patients to and from HIE members are more likely to adopt HIE later. The two sets of conclusions are both complementary and mutually reinforcing. The referral networks are central to the economic viability of HIE platforms. The results of this study establish the relationship between HIE membership the evolution of referral networks. Further, these results also establish the economic benefits derived by different sides of the HIE platform. Together, these lead to solving the problems of economic barriers to effective sharing of health information by providers via HIE and developing sustainable HIE business models.

Demo Proposal: Lightweight Deep Learning on Smartphone for Early Detection of Skin Cancer

PRANJAL SAHU, State University of New York at StonyBrook
HONG QIN, State University of New York at StonyBrook
DANTONG YU, New Jersey Institute of Technology

Early diagnosis of Melanoma skin cancer is essential to cure this most malignant form of cancer and improve the five-year surviving rate. Medical practitioners often rely on visually inspecting skin lesions and their medical images. In this paper we propose a smartphone based skin cancer detection system that utilizes deep learning and low-cost camera to take the snapshots of suspected skin lesions and distinguish between malignant and benign melanoma skin images. Furthermore, we demonstrate that the same machine learning pipeline can be integrated into many low cost computing devices, for example, Raspberry Pi, to train the deep networks and perform inference with the trained network models. To mitigate the computation complexity, we employ transfer learning to leverage a well-trained inception v3 deep network model and retrain it with our target image dataset on hand-held computing platform. Experiments show that the network model based on transfer learning attains the accuracy of 76%.

1 INTRODUCTION

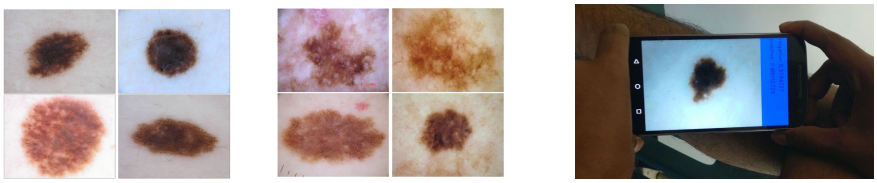
The number of skin cancer cases reported each year is more than that of the lung, breast, colon, prostate cancer cases combined. Melanoma skin cancer is the most fatal form of skin cancer since it can grow to vital organ and become incurable in its late stages. According to the estimation of the skin cancer organization cancer [2], melanoma skin cancer kills 10,130 people annually in the US only. Melanoma in early stages is curable and the visual images of skin provide a sound medium and viable approach to identify it. Hence for this reason, dermatologist often recommend regular skin exams. Inspired by the recent work on machine-learning assisted medical imaging analysis [1], we propose a personal skin cancer diagnostic method that is based on deep learning and smartphone technology. In this paper we demonstrate to use an Android smartphone to identify melanoma skin cancer with a high accuracy. With recent success of Android applications, for example, Prisma [7], that have introduced to millions of people the power of artificial intelligence, we believe our proposed application will democratize the access to preventive health care. Furthermore, we show the feasibility of an cost-effective Raspberry PI cluster of training a machine learning model that can be used

 3rd Symposium on Data Science for Healthcare (DaSH). October 19-20, 2017, Leir Retreat Center, CT, USA.

later for skin cancer detection.

Our contributions include:

- Applying a transfer learning approach to retrain the Inception V3 deep net on the ISIC skin dataset [3] and achieved 76% accuracy without incurring a computation-intensive task of training complex network.
- Designing an Android application for deep learning inference and prediction on malignant and benign skin images.
- Demonstrating the effectiveness and viability of a Raspberry PI cluster to train/retrain deep neural nets.



(a) Malignant skin samples (b) Benign skin samples (c) Android application classifying skin

Fig. 1. (a) and (b) shows some skin samples of Melanoma cancer and (c) shows a sample Android application classifying a skin portion

2 DEEP LEARNING BACKGROUND

Recent advancements in computer vision domain due to the success of deep learning is phenomenal. Many tasks, including image detection, localization, and recognition, have witnessed considerable improvements in accuracy. However all of these fields require feeding a huge amount of training data into the deep learning architectures, which requires the network models to be trained on high performance computers with multiple GPUs and large size memory. Community often adopts crowd sourcing to expand training data for these complex network architectures, a practice that is not viable for the domain of medical imaging because it requires the involvement and supervision by domain experts. Consequently, the lack of professionally curated data greatly affects the generalizability of the large network architecture that is trained for the field of medical imaging from scratch. To mitigate this problem, a fine-tuning technique [5] is applied to leverage a large general deep network architecture that is pre-trained with a large dataset with easy access and retrain it with (limited) domain data for a specific purpose. In this paper, we apply this type of transfer learning in examining medical images and classifying them as “benign” or “malignant”. In particular, we remove the last layer from an existing model and replace it with a new layer that needs to be trained with the domain data. The fine tuning requires a much less amount of training data and incurs a significantly lower computation time than does a full-scale training, and hence makes the

task for learning from medical images much easier and more accessible.

We fine tuned the Inception V3 deep net that was intensively trained for natural image classification on ImageNet dataset [4], re-trained it for the task of skin cancer detection, and deploy the trained model in a hand held smartphone for the purpose of real-time pre-screen and diagnosis.

3 SMARTPHONE FOR HEALTHCARE

To democratize the benefits of health care, we develop an Android application that uses the retrained in-mobile deep neural networks to detect malignant melanoma skin. The primary benefits of in-mobile deployment and inference are that its diagnosis is cost-effective, does not depend on network connectivity, and can be performed virtually everywhere, particularly where in-mobile detection is the only option.
Two technology advancements make this use-case even more plausible: the first one is that Google has been putting effort to make in-mobile machine learning more friendly and develop the Tensorflow Lite for mobile platforms [6], and the second is that edge computing platform, for example Raspberry PI, provides a fully featured Linux OS system that costs much less than state-of-the-art computers.

4 RASPBERRY PI CLUSTER FOR TRAINING

As discussed earlier, considering the current limitations of training deep network we propose to use a portable cluster of raspberry pi nodes, each of which has four cores, one GB RAM and costs around $35, and demonstrate that training a deep net on such a cluster is feasible provided done in a distributed manner. In the remote and rural areas where the availability of smartphone is a privilege, a Raspberry pi cluster can be used in a district clinic to speed up the diagnosis process and provide a doctor with a second opinion on the case.

5 EXPERIMENT AND RESULTS

We conducted our experiment on the dataset provided by the ISIC for the 2017 challenge on melanoma detection and classification. The dataset comprises of 2750 images in total: 521 images being malignant melanoma and 2229 images being benign melanoma. The dataset was divided into three parts with 80% for training, and 10% each for validation and testing. Training was done on a supermicro server with 32 cores. Last layer of Inception V3 net was replaced for a two class classification training. Accuracy of 76% was obtained. ROC curve obtained of the experiment is shown in Fig 4. The trained model was later used to develop an Android application.
The results of our second experiment to compare time required to achieve a certain accuracy between a single node and two node Raspberry PI cluster on MNIST dataset [8] is shown

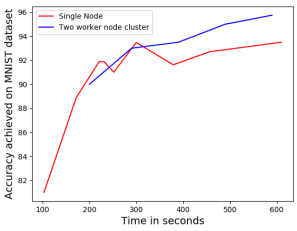


Fig. 2. Accuracy vs time taken to achieve on MNIST dataset

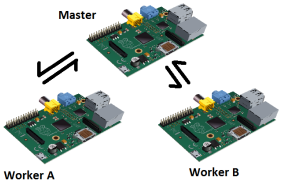


Fig. 3. Raspberry PI cluster

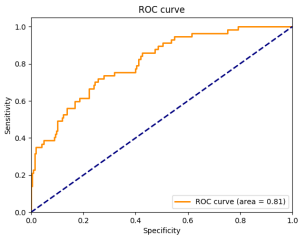


Fig. 4. ROC curve of malignant skin classification on ISIC dataset

in Fig. 2. We omit the details of network architecture used in this experiment due to space limitation.

6 FUTURE WORK

In this paper we have demonstrated skin melanoma detection using smartphone with still images. However, having multiple images of the skin with varying resolution and multiple views can improve classification results. It remains to be seen how much improvement can be obtained by obtaining a video of the skin.

7 CONCLUSION

We have shown the immense potential of smartphone based applications to detect skin cancers. The results demonstrate that a cluster of multiple Raspberry PI nodes can be a good alternative for training deep neural nets, bears a great value proposition in the household smart medical devise and demonstrates a great potential in extending Internet of things to remote areas and developing countries.

REFERENCES

[1] Roberto A. Novoa et al Andre Esteva, Brett Kuprel. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542 (Feb. 2017), 115–118. https://doi.org/10.1038/nature21056

[2] Skin Cancer Foundation. 2017. What is Melanoma? (2017). http://www.skincancer.org/skin-cancer-information/melanoma

[3] ISIC. 2017. Skin Lesion Analysis Towards Melanoma DetectionPart 3: Lesion Classification. (2017). https://challenge.kitware.com/#phase/584b0afccad3a51cc66c8e38

[4] Hao Su et al. Olga Russakovsky, Jia Deng. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* 115 (2015), 211–252. https://doi.org/10.1038/nature21056

[5] Lisa Torrey and Jude Shavlik. 2009. *Transfer Learning*. IGI Global. ftp://ftp.cs.wisc.edu/machine-learning/shavlik-group/torrey.handbook09.pdf (handbook).

[6] VentureBeat. 2017. Tensorflow Lite for mobile machine learning. Text. (May 2017). https://venturebeat.com/2017/05/17/android-launches-tensorflow-lite-for-mobile-machine-learning/

[7] Wikipedia. 2016. Prisma application. Text. (2016). https://en.wikipedia.org/wiki/Prisma_(app)

[8] Leon Bottou et al. Yann Lecun. 1998. Gradient based learning applied to document recognition. (1998). http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf

Data Driven Self-Learning for Knowledge Discovery in Health

AIDONG ZHANG, State University of New York at Buffalo

VISHRAWAS GOPALAKRISHNAN, State University of New York at Buffalo

Literature Based Discovery (LBD) [1]refers to the research process of inferring new and interesting knowledge by logically connecting independent fragments of information units through explicit or implicit means. This area of research, which incorporates techniques from Natural Language Processing (NLP), Information Retrieval (IR) and Artificial Intelligence (AI), has significant potential to reduce discovery time in medical and health research fields. Formally introduced in 1986 by Dr. Swanson, LBD has grown to be a significant and a core task for text mining practitioners in the medical domain. Being an inter-disciplinary activity, this has led researchers across related domains to contribute in advancing this field of study. With evidence and semantics playing an important role, LBD is quite different from traditional link prediction tasks in the sense that it does not follow the notion of "end justifies the means". Consequently, each and every step towards a decision needs to have a sound interpretation in the real world. The focus on evidence and intermediary steps have led to multiple works focusing on effective ways to rank the medical concepts, with many of them requiring intense human intervention.

Due to the very nature this discovery process, the task itself presents challenges that require human skills and expertise to circumvent the issues related to scalability. For instance, open-ended discovery questions, like finding all the possible treatments for a particular disease, require careful consideration regarding computing resources as against to closed-ended discoveries like the determination of the possible existence of a relationship between a particular disease and a particular biological substance.

Hence, we need an end to end framework that effectively and efficiently generates various hypotheses, based on the discovery question at hand and then ranks them in a "domain and contextually sensitive fashion". Furthermore, the approach should involve minimal to no user intervention to enable discovery, which is novel and not biased towards user expectation. Towards this end, we need an approach that is able to, whenever possible, learn the association formation process and use this model with contextually sensitive ranking methodologies to generate and rank various

3rd Symposium on Data Science for Healthcare (DaSH). October 19-20, 2017, Leir Retreat Center, CT, USA.

hypotheses. In this talk, we will present our ongoing endeavors in this field. We will discuss how a self-learning based framework for knowledge discovery can be designed to mine hidden associations between non-interacting scientific concepts by rationally connecting independent nuggets of published literature. The self-learning process can model the evolutionary behavior of concepts to uncover latent associations between text concepts, which allows us to learn the evolutionary trajectories of text terms and detect informative terms in a completely unsupervised manner. Hence, meaningful hypotheses can be efficiently generated without prior knowledge. With the capability to discern reliable information from various sources, this self-learning framework provides a platform for combining heterogeneous sources and intelligently learning new knowledge with no user intervention.

Cognitive Computing for Healthcare

ERIC BROWN, IBM

The[1] growing amount of data and information in healthcare creates enormous opportunities for informing decisions and improving outcomes in a wide range of settings. At the same time, the volume, veracity, and variability of the data create enormous challenges for automated analysis and reasoning. Watson Health is applying cognitive computing and advanced big data analytics to address these challenges and build innovative solutions that solve a wide range of pain points for healthcare practitioners. This talk will introduce the challenges and describe some of the solutions we're developing at Watson Health.

3rd Symposium on Data Science for Healthcare (DaSH). October 19-20, 2017, Leir Retreat Center, CT, USA.