Applications of Machine Learning in Medical Diagnosis

Article · October 2017

CITATIONS
0

READS
167

5 authors, including:

Marcelo Gagliano
University of Auckland
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

John Pham
University of Auckland
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Boyang Tang
University of Auckland
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Hiba Kashif
University of Auckland
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Applications of Machine Learning in Medical Diagnosis

Marcelo Gagliano
Department of Computer Science
University of Auckland
mgag042@aucklanduni.ac.nz

John Van Pham
Department of Electrical and Computer Engineering
University of Auckland
dpha010@aucklanduni.ac.nz

Boyang Tang
Department of Computer Science
University of Auckland
btan766@aucklanduni.ac.nz

Hiba Kashif
Department of Computer Science
University of Auckland
hkas238@aucklanduni.ac.nz

James Ban
Department of Computer Science
University of Auckland
jban997@aucklanduni.ac.nz

I. Introduction

Medical diagnosis is the process of determining the cause of a patient's illness or condition by investigating information acquired from various sources including physical examination, patient interview, laboratory tests, patient's and the patient's family medical record, and existing medical knowledge of the cause of observed signs and symptoms. Getting a correct diagnosis is the most crucial step in treating a patient as it allows physicians to find the best treatment for the patient's condition. However, it is a complicated process and requires lots of human effort and time. Due to that complex nature, it is error-prone. Thus, misdiagnosis is very common. According to the World Health Organization, 5% percent of the outpatient encounters were misdiagnosed in 2015. This is a worrisome, especially when people lives are at stake.

A diagnosis of a disease or a condition relies on information which contains factors that makes getting it correct challenge. These factors include ambiguity, uncertainty, conflicts, and resource and organizational constraints. A lot of symptoms are nonspecific and variable, depending on the person. Many diagnostic tests are expensive, not regularly done and often don't give a black or white answer. Furthermore, physicians are usually prone to cognitive bias and incorrect applications of heuristics during the diagnosis stage. They are more biased towards disease or conditions which they have diagnosed in the past. They often trust the initial diagnostic impression, even though the further information might not support that initial assumption.

The rapid development in the fields of Artificial Intelligence, especially Machine Learning (ML), and Datamining allow technology and healthcare innovators to create intelligent systems to optimize and improve the current processes. Now, ML has been applied in a variety of area within the healthcare industry such as diagnosis, personalized treatment, drug discovery, clinical trial research, radiology and radiotherapy, smart electronic health record, and epidemic outbreak prediction. In medical diagnosis, ML and Datamining algorithms are particularly useful. They can quickly capture unforeseen patterns within complex and large datasets. With unbiased and balanced datasets, machine learning algorithms can mitigate the aforementioned cognitive bias problem and produce higher accuracy. Most of the research in medical diagnosis is being done over image recognition on MRI, Ultrasounds, and Xrays of patients and some pattern recognition ML algorithms are also being implemented on different lab tests like biopsy and blood samples. The most common ML techniques that are being used in Medical systems today are Medical Image Classification and Analysis, Pattern recognition for medical diagnosis, Expert systems for computer-aided diagnosis.

The field of medical diagnosis in medical systems is quite rich with possibilities and its advantages such as cost-cutting, early diagnosis and potentially saving human life; it has several limitations such as privacy. Because of the privacy issues of the patient's sensitive information, extensive data cannot be provided to practice ML algorithms. Another shortcoming is that not a lot of physicians/surgeons are aware of the ML tools available in the market, they will need to undergo proper training to understand the new rising technological applications. When it comes to knowledge sharing it should not be ignored that technical people working on ML applications and algorithms also need to understand complex medical data and relationship between patient's result and final diagnosis taking all the dependencies under consideration.

II. History Overview

The first generation of commercially produced computers came into use in the early 1950s and was immediately adopted by the medical industry as a mean to analyze a significant amount of medical information efficiently. Since then, computing appears in today's medicine in a wide variety of applications, most notable is the medical diagnosis. At the same time, the digital revolution allows a considerable amount of medical data to be stored and disseminated which incentivized the development of ML algorithms to improve the quality of medical diagnosis further.

Initially, ML algorithms were designed and used to model and analyze massive medical datasets. As these ML algorithms become more reliable, they were adopted into tools for assisting medical diagnosis. One of the earliest applications of ML in medical diagnosis take place in the early 1970s with the

development of Internist-1 which is an expert consultant program for diagnosis in general internal medicine [1]. The system uses a probabilistic model which led to the development of Bayesian network and approximate inference. Another solution, Mycin, was also developed in the early 1970s at Stanford University [2]. Mycin was designed to diagnose and propose a treatment for blood-borne diseases. Although these tools had promising results in diagnosing different medical illness and conditions, they were never deployed in any working environment due to a combination of practical and ethical factors.

Through the years, as ML and data mining continue to develop, more advanced and powerful algorithms such as k-nearest neighbor, decision trees, and artificial neural network, were applied to solve more complicated and specialized medical diagnosis problems. Some of the interesting works are described in the following subsections.

A. Intelligent Heart Disease Prediction System

Heart diseases include various medical conditions that affect the circulatory system, which consists of heart and blood vessels. Coronary artery disease is the leading form of heart disease in which blood vessels, most commonly artery, becomes blocked preventing oxygen and nutrients from getting to the heart, which ultimately leads to heart attacks. It is the leading global cause of death with 17.3 million deaths each year, a number that is expected to grow to more than 23.6 million by 2030 [3]. Hence, many studies have been undertaken to apply ML and data mining techniques to assist practitioners in diagnosing heart diseases.

Parthiban et al. in [4], developed a prototype Intelligent Heart Disease Prediction System with coactive neuro-fuzzy inference system (CANFIS) and genetic algorithm using historical disease databases to make intelligent clinical decisions. The CANFIS model integrates adaptable fuzzy inputs with a modular neural network to rapidly and accurately approximate complex functions. Genetic algorithms were used for auto-tuning the CANFIS parameters and selection of optimal feature set, thus, reduce the training time and enhance the performance. The performances of the CANFIS model were evaluated regarding training performances and classification accuracies, and the results showed that the proposed CANFIS model has great potential in predicting the heart disease.

B. Histopathology Cancer Image Classification, Segmentation, and Clustering

Unlike other diseases, cancer has been characterized as a group of heterogeneous disorders with various subtypes. Normal cells in human bodies have their specified functionalities, and they divide under a regular principle called the cell cycle, the new cells will take the place of the old ones. When cancer happens, the cells will lose the control of growth and spread, and the cancer cells may travel to any parts of our bodies. Most cancers form a lump called a tumor, and the tumor can be classified as either benign or malignant, only the malignant tumors are cancers.

Cancer tissues in histopathology images exhibit abnormal patterns, provide a mean to differentiate abnormal tissues from the normal ones. High-resolution histopathology images and the accessible digitization of histopathology due to the development of specialized digital microscope escalates the development of new systems for classifying and grading cancer histopathology images. Ultimately, becomes a vital technology for identifying and analyze cancers.

Yan Xu, et al. in [5] proposed a new learning method, multiple clustered instance learning (MCIL), to classify and annotate cancer cells. This process takes a weakly supervised learning approach, namely multiple instance learning (MIL), and combines it with clustering concepts. It simultaneously performs image-level classification (cancer vs. non-cancer image), pixel-level segmentation (cancer vs. non-cancer tissue), and patch-level clustering (cancer sub-classes). The experimental results show MCIL is more efficient and effective than others system with a similar approach in detecting colon cancer.

C. Application of ML in Breast Cancer Diagnosis

Breast cancer has become a major cause of death in women today. Thus, the common awareness of the potential benefits of early detection of breast cancer has increased dramatically. An effective and reliable predictive model for early diagnosis can significantly decrease the harm to women, the prior objective of the prediction is to figure out whether patients are assigned into a benign group (non-cancerous) or malignant group (cancerous) [6]. In the current case, the predictions can be regarded as classification problems.

In the study of [6], various ML / Data Mining techniques have been proposed to strengthen the breasts cancer early detection and prognosis. In general, data mining can be used to extract patterns and illustrate the useful properties of existing data. In the breast cancer research area, the helpful information drawn from different medical datasets are significant for clinical decision and treatment. Data can be examined through various parameters by data mining applications. It is possible to find the association among different events, for instance, whether the breast cancer has an influence on other diseases or has a connection to some clinical factors. It is also possible to find similar trends or objects for a specific purpose, for example, whether a group or cluster of breast cancer patients has similar drug-sensitivity with same treatment plan [6]. On the other hand, the most critical task of data mining is attempting to predict and classify targets. In breast cancer early diagnosis, prediction techniques are used to model the datasets and then assign patients to a cancerous group or non-cancerous group. Therefore, it can be regarded as a basic classification problem, a branch of ML or data mining classification algorithms such as ANNs, Bayesian networks, support vector machines and decision trees are used widely for this purpose. Each method has its merit and weakness, a review given by [6] shows that predictive model constructed by probabilistic neural network had a significantly high classification accuracy for breast cancer early detection. However, only high accuracy is

not enough, using neural networks along with logistic regression for breast cancer diagnosis guarantees the sensitivity and specificity of predictive models [6].

For cancer, just a simple diagnosis is not enough. More importantly, it is worth to take care of the prognostic prediction after the malignant lump has been surgically excised [7]. To be more specific, predicting the outcome of recovery by a particular treatment plan is helpful for further clinical research. Furthermore, the prediction of whether a patient will recur at a specific time is also important. Prognosis helps in determining the case for whether a patient is recurring or not as well as the case for the time to recur. The study of [6] summarizes that the breast cancer prognosis is mainly analyzed under ANNs, this method provides an efficient way to classify patients and predict the survival time with high accuracy. Last but not least, there is a beneficial step in the data mining procedure called data pre-processing. In cancer research area, the preparation of data is particularly important, as a suitable dataset after cleaning up and transforming irrelevant or invalid data helps in modeling a more sensitive predictor of cancer diagnosis.

D. An Ensemble Model for Diabetes Diagnosis in Large-scale and Imbalanced Dataset

The growing threat of diabetes has been seen as one of the most challenging health challenges for humans in the future, as this kind of disease spreads worldwide and exceeds expectations in any previous year. Diabetes consists of three types, which are Type 1 diabetes, Type 2 diabetes, and Gestational diabetes, and diabetes is caused by a fault in the insulin production of the body or by the inefficient use of the insulin produced [8]. According to the statistics, 80% of diabetes complications can be avoided or controlled under a low-level risk. Thus, it is valuable to pay more attention to early diagnosis and timely detection.

Some former techniques such as Oral glucose tolerance test (OGTT) and fasting plasma glucose (FPG) have been commonly used for clinical diagnosis, but they are either inconvenient or not sensitive enough. Moreover, manually analyzing medical statistics would undoubtedly put extra burdens on researchers. For above reasons, applications of ML and data mining are used to form an intelligent prediction model for early diagnosis. In the case of diabetes, real medical datasets are more suitable to reflect the actual situation of patients, but the collection process is arduous and time-consuming. On the other hand, a large scale of data must be enormously diverse and complex, and the medical data of diabetes are more likely to be imbalanced. This is because vast majority of diabetes is type 2 diabetes, which means the class may have more instances than others. A novel ensemble method xEnsemble was proposed by [8], which is efficient for diabetes diagnosis in a large-scale and imbalanced dataset.

The main idea behind xEnsemble is to build a capable system that has both low variance and low bias. An ensemble threshold is introduced to observations selection which helps in resampling balanced subsets, and for every base individual classifier, it will learn a different aspect of original majority class. It is commonly known that boosting is mainly used to decrease bias while bagging is typically used to reduce variance. The study of [8] shows that the combination these two strategies facilitate the xEnsemble model to be more precise in diagnosing diabetes. Even though the simple ML methods such as CART, logistic regression, and SVM are widely used, but the xEnsemble model gives a more significant outcome of performance evaluation.

III. TOOLS AND CASES

The ML field has attracted several players in the recent years, due to the business perspectives that it represents. To identify and diagnose diseases is at the forefront of ML research in medicine. According to [10], more than 800 drugs and vaccines to treat cancer were in trial and ML is the only feasible alternative to analyze the amount of data generated by these experimental treatments.

Applying ML can also unlock the following benefits to the patients, doctors, and vendors [9].

- Right living, where the patients can build value by taking an active role in their treatment, including disease prevention;
- Right care, which involves ensuring that the patient gets the most appropriate treatment available;
- Right provider, directing the patient to the best professionals;
- Right value, where the providers and payors look to increase the healthcare value but also preserving its quality continuously;
- Right innovation, which involves the identification of new therapies.

In this paper, we will present some companies and their services towards the opportunities above.

E. IBM

IBM, the American multinational technology company, started in the Information Technology Medical market, after buying purchased Merge Healthcare, a provider of enterprise imaging systems. This purchase was an enabler of an IBM product, Watson, because gave to it access to its extensive database of radiology records. All this information was used to train the AI in evaluating patient data and get better at reading imaging exams. After that, IBM established a partnership with Medtronic to make sense of diabetes and insulin data in real time and bought out healthcare analytics company Truven Health. Also, to maintain this inflow of information that enriches Watson's knowledge base and improves its accuracy, IBM imposed that all vendors are required to share access to all the patient data and imaging studies they have access to.

A recent case where IBM Watson was used for medical diagnosis was in Memorial Sloan Kettering (MSK)'s Oncology department. The clinicians trained Watson to interpret cancer patients' clinical information and identify treatment options, based on the most recent medical literature presented in its knowledge base. To the clinicians, keeping up with the medical research can take as many as 160 hours a week, so only 20 percent of the knowledge that clinicians use today is based on

medical research [11]. The application of Watson allowed this number to increase drastically.

F. HearFlow

Root HeartFlow, Inc. is a medical technology company specialized in solutions to diagnose and treat cardiovascular disease. Its primary product is called HeartFlow® FFRct Analysis, a non-invasive solution that evaluates whether an individual has significant coronary artery disease, creating a personalized 3D model of the patient's arteries based on both anatomy and physiology. To achieve that, it uses CT scans, as training data, with deep learning algorithms. The output is a detailed mapping of the patient's heart, able to expose the health of the blood flow.

Until recently, the best test for it was an angiogram, which is an invasive and costly procedure. It improves both clinical outcomes and the patient experience while reducing the cost of care, addressing most of the perspectives presented in the introduction to this section.

According to [12], the usage of FFRct increased the specificity, the positive and negative predicted value and the accuracy of the diagnosis, as shown in the table below:

Year	2011		2012		2013	
Metrics	CT	FFR_{ct}	CT	FFR_{ct}	CT	FFR_{ct}
Sensitivity	94%	93%	84%	90%	94%	86%
Specificity	25%	82%	42%	54%	34%	79%
PPV	58%	85%	61%	67%	40%	65%
NPV	80%	91%	72%	84%	92%	93%
Accuracy	61%	81%	64%	73%	53%	81%

G. Enlitic

Enlitic is an American company that applies deep learning techniques and image analysis to process the vast stores of medical data collected from a given client, structured or not. Working with the Radiological Society of North America (RSNA), Enlitic used Deep Learning to create a classifier able to recognize suspicious masses on radiological scans that are likely cancerous. To do that, it used the database of medical imaging from RSA, to train the model. The inputs were previous scans, health information of the patients, who were anonymous to preserve their identities, the diagnose made, the recommended treatment and the success of it. The resulting model was able to increase the accuracy of the diagnostics of tumors.

H. Google

Research at Google is an engineering organization that deploys small teams to explore new ideas and conducts research. One of these groups, 'Google Brain Team' focuses on artificial intelligence and ML. In particular, this team focused on utilizing the deep learning in computer vision, where the identification of dog breeds are cars were expanded into the identification of disease in medical images [13].

In particular, the team focused on the diagnosis of diabetic retinopathy. The usual diagnosis involves a highly trained doctor examining a retinal scan of the eye. A treatment is available when detected early, but if left undetected, this condition can progress to irreversible blindness. Hence an early diagnosis is extremely valuable [13].

In the research, the team trained a type of neural network optimized for image classification called a deep convolutional neural network with data set of 128 175 retinal images. The resultant data set was then validated using 2 separate data sets graded by at least 7 US board-certified ophthalmologist. The team has found that the algorithm had high sensitivity and specificity for detecting diabetic retinopathy. However, they concluded that more research is required to determine the feasibility of using the algorithm in clinical settings [14]. Hence, Research at Google has partnered with healthcare providers such as Stanford Medicine, Chicago Medicine and University of California to combine ML with clinical expertise to do further research on prediction through medical records and improve patient outcomes in actual treatments [13].

I. Microsoft

There are several teams at Microsoft that are working on cancer. There is a team that utilizes ML and natural language processing to create an individualized treatment for patients. There is another team that focuses on using ML with computer vision to monitor tumor progression. There is also a team that works on algorithms to help determine the pathophysiology of cancer and treatment options available to treat cancer.

In particular, a project named InnerEye is an assistive AI for cancer detection. InnerEye utilizes image analysis and ML to help identify and diagnose cancer in patients [16]. Given a CT (computed tomography) or MR (magnetic resonance) scan, a clinician has to identify different organs. This is done by an algorithm called Decision Forests, which was initially used in skeletal tracking of Microsoft Kinect, an entertainment device [17]. InnerEye can identify organs, and anomalies in CT scans in around 30 seconds with just a little human assistance [16]. This would have taken hours by even the most skilled oncologists. This allows more time to be spent on the individual treatment plan for each patient.

InnerEye works by supervised learning, where the doctors can fine-tune the results and feed the data back to refine the algorithm. InnerEye is not a replacement for doctors but a tool for doctors to do their jobs more efficiently and effectively [16].

IV. CONCLUSION

ML medical diagnosis is a fertile field with a tendency to transform Medical Science for better. The work in this field is growing in fast pace with potential in every different disease of diagnosis. The application of ML algorithms increases the quality of the diagnostic.

Agencies like HIPAA (Health Insurance Probability and Accountability Act) are doing some good work by providing healthcare databases for ML testing and implementation purpose and making sure that the patient's privacy stays intact.

The research work in different diseases, applying appropriate ML approach can lead to a revolutionary Medical Science.

REFERENCES

[1] R. Miller, H. Pople and J. Myers, "Internist-I, an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine", New England Journal of Medicine, vol. 307, no. 8, pp. 468-476, 1982.

[2] W. van Melle, "MYCIN: a knowledge-based consultation program for infectious disease diagnosis", *International Journal of Man-Machine Studies*, vol. 10, no. 3, pp. 313-322, 1978

[3] D. Mozaffarian, E. Benjamin, A. Go, D. Arnett, M. Blaha, M. Cushman, S. de Ferranti, J. Després, H. Fullerton, V. Howard, M. Huffman, S. Judd, B. Kissela, D. Lackland, J. Lichtman, L. Lisabeth, S. Liu, R. Mackey, D. Matchar, D. McGuire, E. Mohler, C. Moy, P. Muntner, M. Mussolino, K. Nasir, R. Neumar, G. Nichol, L. Palaniappan, D. Pandey, M. Reeves, C. Rodriguez, P. Sorlie, J. Stein, A. Towfighi, T. Turan, S. Virani, J. Willey, D. Woo, R. Yeh and M. Turner, "Heart Disease and Stroke Statistics—2015 Update", 2017.

[4] P. Sharma and K. Saxena, "Application of fuzzy logic and genetic algorithm in heart disease risk level prediction", *International Journal of System Assurance Engineering and Management*, 2017.

[5] Y. Xu, J. Zhu, E. Chang and Z Tu, "Multiple Clustered Instance Learning for Histopathology Cancer Image Classification, Segmentation and Clustering, UCLA, 21st June 2012.

[6] Gupta, Shelly, Dharminder Kumar, and Anand Sharma. "Data mining classification techniques applied for breast cancer diagnosis and prognosis." *Indian Journal of Computer Science and Engineering (IJCSE)* 2.2 (2011): 188-195

[7] J. A. Cruz and D. S. Wishart, "Applications of ML in Cancer Prediction and Prognosis," Cancer Informatics, vol. 2, p. 117693510600200, 2006.

[8] X. Wei, F. Jiang, F. Wei, J. Zhang, W. Liao, and S. Cheng, "An Ensemble Model for Diabetes Diagnosis in Large-scale and Imbalanced Dataset," Proceedings of the Computing Frontiers Conference on ZZZ - CF17, 2017.

[9] Groves, P., Kayyali, B., Knott, D., & Kuiken, S. V. (n.d.). *The 'big data' revolution in healthcare - Accelerating value and innovation* (Rep.). A publication sponsored by McKinsey&Company.

[10] P. (2015). *Pharmaceutical Research and Manufacturers of America - 2015 Annual Report* (Rep.). Pharma Foundation.

[11] I. (n.d.). *Memorial Sloan-Kettering Cancer Center: IBM Watson helps fight cancer with evidence-based diagnosis and treatment suggestions* (Rep.). IBM Corporation.

[12] Gonzalez, J. A., MD. (n.d.). *Will CT-FFR Radically Change Our Approach to the Patient with Ischemic Heart Disease?* (Rep.). Scripps Clinic.

[13] "Research at Google", Research.google.com, 2017. [Online]. Available: https://research.google.com/teams/brain/healthcare/. [Accessed: 20- Oct- 2017].

[14] V. Gulshan, L. Peng, M. Coram, M. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. Nelson, J. Mega and D. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs", *JAMA*, vol. 316, no. 22, p. 2402, 2016.

[15] *News.microsoft.com*, 2017. [Online]. Available: https://news.microsoft.com/stories/computingcancer/. [Accessed: 20- Oct- 2017].

[16] G. Lynch, "From Kinect to InnerEye – How Microsoft is supercharging gaming tech with AI smarts to help diagnose cancer", *TechRadar*, 2017. [Online]. Available: http://www.techradar.com/news/from-kinect-to-innereye-how-microsoft-is-supercharging-gaming-tech-with-ai-smarts-to-help-diagnose-cancer. [Accessed: 20- Oct- 2017].

[17] "Microsoft Research Connections: Science at Microsoft - Searching the Human Body", *Microsoft.com*, 2017. [Online]. Available: https://www.microsoft.com/en-us/researchconnections/science/stories/inner-eye.aspx. [Accessed: 20- Oct- 2017].