Machine Learning and Biomedicine

Abdulmajeed Mohammed M Alanazi

Faten Olayan D Alanazi

Abstract

This paper aims to review some of the research studies which have been carried out in the field of biomedicine and machine learning. For this paper, specific search terms were used in Google Scholar search engine and the results were shortlisted according to the year of publication. From the review of the studies available on the subject, it was clear the health industry is getting restructured due to concurrent development of the fields of electronics, communication and computers, and as a result, biomedicine is also in the midst of a data revolution. Authors have talked about the knowledge explosion in biomedicine. Techniques such as predictive and inferential analysis require data such as medical records, imaging data, sequencing data, genotypes and sensor data etc. Machine learning is being extensively used in the health industry and in the field of biomedicine.

Keywords: Machine Learning, Biomedicine, Healthcare, Review

Introduction

As a result of simultaneous growth in fields of electronics, communication and computers, the health industry is being rapidly restructured, fostering greater and more promising smart/mobile health solutions (Zhang, Zhou & Zeng, 2017),. Similar to a lot of other fields, biomedicine is in the midst of a data revolution (Yu, Ma, Fisher, et al., 2018). Comprehensive molecular and clinical datasets— including complete human genomes, gene expression profiles, high-resolution imaging, metabolomics, electronic medical records, and so on—are no longer isolated to a few study participants; in a few years, we will have such comprehensive information for millions of patients (Torkamani et al., 2017 – as cited in Yu, Ma, Fisher, et al., 2018). Many authors have talked about how big data will transform medicine (Obermeyer & Emanuel, 2016).

According to Kamdar, Fernández, Polleres, Tudorache, & Musen (2019), "the 21st century is the age of data and knowledge explosion in biomedicine. Several key events, such as the completion of the Human Genome Project and the advent of next-generation sequencing technologies, the enactment of the Health Information Technology for Economic and Clinical Health (HITECH) Act, and the Internet of Things phenomenon, have led to a significant increase in the volume, velocity, and variety of biomedical data". In their study, Kamdar, Fernández, Polleres, et al. (2019) say that in order to construct a person's complete profile, for performing predictive and inferential analytics, and for investigation of the mechanisms behind any biological event, biomedical researchers have various sources of data at their disposal, such as medical records, imaging data (e.g., X-ray images, MRI scans), claims, sequencing data (e.g., gene expression, DNA methylation, MicroRNA expression, chromatin accessibility data), genotypes, sensor data (e.g., wearable data, social media streams).

There is also a rapid increase in the number of structured, machine-processable knowledge artifacts, as well as an increase in unstructured knowledge sources in the form of publications in biomedicine. Knowledge bases (e.g., DrugBank, UniProt) and ontologies (e.g., Gene Ontology, National Cancer Institute Thesaurus) are widely-used and popular resources in biomedicine, and

contain knowledge pertaining to molecules (e.g., drugs, proteins) and their characteristics, biological pathways, animal models and phenotypes, organs, symptoms, diseases, and adverse reactions. As of January 2019, there are more than 750 ontologies and terminologies in BioPortal, the world's most comprehensive repository of biomedical ontologies. MEDLINE, the largest repository of scientific articles in biomedicine and the primary component of the PubMed search engine, currently contains more than 25 million citations and thousands more are added each day.

In order to be useful, data needs to be analyzed, interpreted, and acted on. Multiple analysis approaches have been proposed for changing the ever-increasing amount of patient data into effective therapies. Most important amongst these is the field of machine learning which has seen rapid advancement in the recent years (LeCun et al., 2015 - as cited in Yu, Ma, Fisher, et al., 2018). There has been a lot of enthusiasm regarding the use of the many-layered 'deep' artificial neural networks, which have taken inspiration from real neural networks and the manner in which the brain processes patterns. After extensive training using examples, "artificial neural networks learn to predict the correct answer—or output—that should be returned for the many possible input patterns" (Yu, Ma, Fisher, et al., 2018). Deep learning approaches have been utilized to recognize objects in images like dogs, people, and faces and to differentiate good from bad moves in games such as chess (Silver et al., 2016 – as cited in Yu, Ma, Fisher, et al., 2018).

Methodology

In this paper, we will review some of the research studies which have been carried out on the subject of biomedicine and machine learning. Towards this end, specific search terms were used in Google Scholar search engine. The results of these searches were shortlisted as per the year of publication. For the purpose of this study, only studies published after 2000 were used, in order to examine the phenomenon of biomedicine and machine learning.

Results and Discussion

In the field of medicine, most computer-based algorithms are "expert systems" or rule sets which encode knowledge on a given topic. These are applied to draw conclusions about specific clinical scenarios, for example, detecting drug interactions or judging the appropriateness of obtaining radiologic imaging (Obermeyer & Emanuel, 2016). The authors say that expert systems operate in a manner that any medical student would – by applying general principles about medicine to new patients. But machine learning operates like a resident doctor – by learning rules from data. Algorithms start with patient observations, move on to sifting through numerous variables to look for combinations which can predict results reliably. As per Obermeyer & Emanuel (2016), the process is like traditional regression models, with an outcome, covariates, and a statistical function linking the two. Machine learning handles an extremely large number of predictors, sometimes even more than observations, and then combines them in nonlinear and highly interactive ways. This capacity allows use of new kinds of data (Obermeyer & Emanuel, 2016).

Precision Medicine

Precision Medicine, since its advent, poses a huge opportunity for biomedicine and computational biology. In recent years, numerous computational methods have appeared in biomedicine research, such as medical image analysis, healthcare informatics, and cancer genomics. An increasing number of data mining algorithms have been employed in the prediction tasks of computational biology and biomedicine because medical data needed

prediction and mining works (Zou, Mrozek, Ma & Xu, 2017). In their study, Zou et al. (2017) say that in recent years, advanced data mining techniques have been developed such as affinity propagation. Off late, deep learning seems to be suitable for big data. Parallel mechanism such as Mahout is also developed by the scholar and industry researchers. Zou et al. (2017) say that a growing number of computer scientists are devoted to the advanced large-scale data mining techniques, but application in biomedicine has not fully been addressed.

Machine Learning Empowered Biometric Methods

According to Zhang, Zhou & Zeng (2017), biometrics are attracting intense attention in terms of effective user identification to enable confidential health applications. As assistive services increase exponentially, issues such as connection needs, big data, security and privacy are also increasing. It becomes imperative then to focus on concerns like how to provide confidential biomedicine applications, and how to effectively protect the sensitive data of patients. In their study, Zhang, Zhou & Zeng (2017) say that biometric human identification is attracting great attention as it is focused on the security and privacy issue. The authors say that biometric technology is stronger than the traditional methods such as token (identity card) and knowledge-based (username/password) ones which may be stolen or lost, as biometrics are usually unique to individuals and hence, almost impossible to duplicate. There are several categories of biometric modalities, such as the behavioral and physiological. “The physiological signals can be easily collected by wearable computers which are part of the body sensor network. Then the signals acquired can be processed either by personal digital devices or cloud computing servers, for user identification purpose toward confidential smart digital health” (Zhang, Zhou & Zeng, 2017). Among different physiological signals, the heart electrocardiogram (ECG) and the brain electroencephalogram (EEG) are two main modalities.

According to Zhang, Zhou & Zheng (2017), ECG reproduces electrical behavior of the heart which is controlled by both sympathetic and parasympathetic nerves. To measure the ECG signal, ECG electrodes are used to detect the tiny electrical changes on the human body. These changes are generated by heart muscle’s electrophysiological movements during depolarization and repolarization phases of one heartbeat. Therefore, it is hard to be duplicated and safer than traditional identification methods (Zhang, Zhou & Zheng, 2017). In daily applications, the ECG signal can be easily collected by wearable computers and then sent to cellphone devices or other personal digital platforms. The ECG signal can be collected at any time since the live human body continuously generates heart electrical signals which are propagated to all the body parts (Zhang, Zhou & Zheng, 2017).

According to Zhang, Zhou & Zheng (2017), EEG signal is another preferred modality which can be used in applications like seizure detection, sleep quality monitoring and emotion tracking. Zhang et al. (2017) say EEG signal is also unique to individuals and is difficult to be duplicated as the underlying signal generation mechanism is extremely complicated. The EEG signal is collected by EEG electrodes placed on the head fixed by a specific EEG cap (Zhang, Zhou & Zheng, 2017). EEG signal is usually highly weak and of a log signal-to-noise ratio, therefore, many EEG electrodes are still used in practical application scenarios to obtain more redundant information for performance enhancement purpose. Of course, this may lower the wearability. But considering EEG signal is highly difficult to be duplicated, it is still attracting more and more attentions in human identification applications (Zhang, Zhou & Zheng, 2017). A detailed comparative analysis of these two modalities, the ECG and EEG, can successfully advance the practical application of an appropriated signal in universal assisted personal devices, for

confidential biomedicine purposes. In their study, Zhang, Zhou & Zheng prove that ECG is stronger than EEG levering a high signal-to-noise ratio and successfully extracted features. According to the authors, "properly selected biometric empowered by an effective machine learner owns a great potential, to enable confidential biomedicine applications in the era of smart digital health" (Zhang, Zhou & Zheng, 2017).

Visible Machine Learning for Biomedicine

According to Yu, Ma, Fisher, et al. (2018), a chief motivation of artificial intelligence lies in interpreting patient data to successful therapies. Machine learning models face specific challenges in biomedicine including handling of extreme data heterogeneity. +

In their study, Yu, Ma, Fisher, et al (2018) say that due to the simultaneous progress in biomedical data and computer science, it is imperative to know to what degree will machine-learning models be successful at interpreting the enormous amounts of biomedical information being generated every day – specifically, if huge patient datasets, provided as inputs to deep neural networks, will be adequate to produce the next generation of dependable and accurate intelligence infrastructure for comprehending and treating disease? (Yu, Ma, Fisher, et al., 2018). The authors state that these models will not be adequate primarily due to the extremely high complexity of biological systems which will essentially limit the applications of current machine learning in patient data. Hence, in their study, Yu, Ma, Fisher, et al. (2018) highlight a new generation of 'visible' approaches focused to streamline the structure of machine-learning models with an increasingly extensive knowledge of biological mechanism. As per the authors, machine learning will not substitute the need for experimental cell and tissue biology; it will be substantially enabled by such knowledge, given the right visible intelligence infrastructure.

Deep Learning and Biomedicine

Deep Learning has become an increasingly popular Machine Learning approach in the last decade. According to Bacciu, Lisboa, Martín, Stoean & Vellido (2018), it's success primarily due to the fact that its internal representation is in the form of high-level features, which permits the modelling of difficult problems, alongside the smart initialization of other deep structures. As per the authors, "staging the difficult task of efficient feature selection by using multiple layers has been crucial in solving extremely difficult problems of image classification, or Natural Language Processing (NLP) by means of Convolutional Neural Networks (CNNs) and Deep Recurrent Neural Networks (RNN), respectively". The success achieved in such complicated problems has generated a great interest not only in the academic community but also in industry, with many private companies involved in the development of commercial products based on deep learning (Bacciu, Lisboa, Martín, Stoean & Vellido, 2018). According to Hadjiiski, Wang, Drukker, Cha, Summers, & Giger (2018), deep learning is a subfield of machine learning, and machine learning is a field within Artificial Intelligence. Deep learning consists of massive multilayer networks of artificial neurons that can automatically discover useful features, that is, representations of input data needed for tasks such as detection and classification, given large amounts of unlabeled or labeled data (Hadjiiski, Wang, Drukker, Cha, Summers, & Giger, 2018).

According to Bacciu, Lisboa, Martín, Stoean & Vellido (2018), one of the most popular applications of deep learning to biomedicine is linked to the processing of sequential data. Sequential data are the simplest form of structured data (arising naturally as the result of many biological, physical and chemical processes) and are a natural representation, for nucleotide

compounds (e.g. DNA, RNA) and for physiological signals (e.g. ECG, EEG, MEG data) (Bacciu, Lisboa, Martín, Stoean & Vellido, 2018).

According to Bacciu, Lisboa, Martín, Stoean & Vellido (2018), Convolutional Neural Network (CNN) models have found extensive application to genomics, especially for the prediction of protein binding sites. In the DeepBind and DeepSEA methods, it has also been projected as a technique to envisage the outcome of wild-type mutation on binding site prediction, contributing to the interpretability of the learned model. Despite their limitation in treating only fixed length subsequences, CNNs have become the reference deep learning model for genomic studies (Bacciu, Lisboa, Martín, Stoean & Vellido, 2018).

According to Ravi, Wong, Deligianni, Berthelot, Andreu-Perez, Lo, & Yang (2017), CNNs have had the highest influence within the field of health informatics. CNN's architecture can be defined as an interwoven batch of "feed-forward layers implementing convolutional filters followed by reduction, rectification or pooling layers", where every layer in the network creates a high-level abstract feature (Ravi, Wong, Deligianni, Berthelot, Andreu-Perez, Lo, & Yang, 2017). This biologically-inspired architecture looks like the process in which the visual cortex assimilates visual information in the form of receptive fields. Other credible architectures for deep learning include those based on the compositions of restricted Boltzmann machines (RBMs) such as deep belief networks (DBNs), stacked Autoencoders functioning as deep Autoencoders, extending artificial NNs with many layers as deep neural networks (DNNs), or with directed cycles as recurrent neural networks (RNNs) (Ravi, Wong, Deligianni, Berthelot, Andreu-Perez, Lo, & Yang, 2017). Latest developments in Graphics Processing Units (GPUs) have also had an important influence on the practical acceptance and acceleration of deep learning. In fact, many of the theoretical ideas behind deep learning were proposed during the pre-GPU era, although they have started to gain importance in the last few year (Ravi, Wong, Deligianni, Berthelot, Andreu-Perez, Lo, & Yang, 2017)s. Deep learning architectures such as CNNs can be highly parallelized by transferring most common algebraic operations with dense matrices such as matrix products and convolutions to the GPU (Ravi, Wong, Deligianni, Berthelot, Andreu-Perez, Lo, & Yang, 2017).

Deep Recurrent Neural Networks (RNNs), such as the Long Short Term Memory (LSTM), have found less application in genomics, even though they are good at modeling variable length sequences (Bacciu, Lisboa, Martín, Stoean & Vellido, 2018). This is predominantly due to a prevalent belief in the biomedical community that RNNs are difficult to train. An attempt to bring the LSTM model into the biomedical community is put forward in a study by Che, Purushotham, Khemani, and Liu (2015), where an approach to learn interpretable features from an LSTM trained on real-world clinical time-series is proposed (- as cited in Bacciu, Lisboa, Martín, Stoean & Vellido, 2018). Along the same line, Lipton, Kale, Elkan, and Wetzell (2016) proposed the use of a LSTM to classify 128 diagnoses from multivariate clinical time series collected in an intensive care unit (ICU) (- as cited in Bacciu, Lisboa, Martín, Stoean & Vellido, 2018). Other authors discuss more controlled use of RNN models for scoring stress levels from heart-rate information (Bacciu, Colombo, Morelli, and Plans, 2018 - as cited in Bacciu, Lisboa, Martín, Stoean & Vellido, 2018).

According to Bacciu, Lisboa, Martín, Stoean & Vellido (2018), the popularity of deep learning has increased exponentially, mainly due to its ability to develop images independently from human intervention, including intrinsic strength to variations in position, rotation, scale, perspective and occlusion. As a result, these characteristics appeared as predominantly

appreciated in the medical sector. Modernization and the constant use of imaging devices has resulted in an ever-increasing amount of image data available for analysis. "Time-saving decision support in this area was achieved by machine learning techniques before the arrival of deep learning, but with a different supplementary human cost, i.e. that of highlighting the regions of interest in the images, of handcrafting the relevant features for the diagnosis and of labelling each image" (Bacciu, Lisboa, Martín, Stoean & Vellido, 2018). This has proved to be a turning point for deep learning, as it has emerged as the true provider of image processing and image interpretation. Deep learning automatically learns the optimal attributes from available images and benefits from large amounts of available data. This has led to deep learning entering the realm of medical imaging successfully, with its applications ranging from landmark detection and tissue segmentation to diagnosis and prognosis (Bacciu, Lisboa, Martín, Stoean & Vellido, 2018).

According to Wang (2016), "deep learning is not only a new wave of research, development and application in the field of medical imaging (and other imaging fields such as homeland security screening) but also a paradigm shift. This could be the magic wand to achieving optimal results cost-effectively, especially from huge and compromised data, as well as for problems that are nonlinear, nonconvex, and overly complex".

According to Esteva, Robicquet, Ramsundar, et al. (2019), deep-learning techniques for healthcare include use of deep learning in computer vision, natural language processing, reinforcement learning, and generalized methods. Such computational techniques can influence a few crucial areas of medicine and explore how to build end-to-end systems.

Conclusion

This paper reviewed some research studies which have been carried out in the field of biomedicine and machine learning. From the review of the studies available on the subject, it was clear the health industry is getting reorganized due to parallel development of the fields of electronics, communication and computers, and as a result, biomedicine is also in the midst of a data revolution. Authors have talked about the knowledge explosion in biomedicine. Techniques such as predictive and inferential analysis require data such as medical records, imaging data, sequencing data, genotypes and sensor data etc. Machine learning is being extensively used in the health industry and in the field of biomedicine.

References

Bacciu, D., Lisboa, P.J.G., Martín, J.D., Stoean, R., & Vellido, A. (2018, April 25-27). Bioinformatics and Medicine in the Era of Deep Learning. In *Proceedings of the 2018 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2018)*, Bruges, Belgium.

Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*, 24-29.

Kamdar, M.R., Fernández, J.D., Polleres, A., Tudorache, T., & Musen, M.A. (2019). Enabling Web-scale data integration in biomedicine through Linked Open Data. *npj Digital Medicine, 2*(90).

Obermeyer, Z., & Emanuel, E.J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England Journal of Medicine, 375*(13), 1216-1219.

Hadjiiski, L.M., Wang, X., Drukker, K., Cha, K.H., Summers, R.M., & Giger, M.L. (2019). Deep learning in medical imaging and radiation therapy. *Medical Physics Journal, 46*(1).

Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G-Z. (2017). Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics, 21*(1), 4-21.

Wang, G. (2016). A Perspective on Deep Imaging. *IEEE Access, 4*, 8914-8924.

Yu, M.K., Ma, J., Fisher, J., Kreisberg, J.F, Raphael, B.J., & Ideker, T. (2018). Visible Machine Learning for Biomedicine. *Cell, 173*, 1562-1565.

Zhang, Q., Zhou, D., & Zeng, X. (2017). Machine Learning-Empowered Biometric Methods for Biomedicine Applications. *AIMS Medical Science, 4*(3), 274-290.

Zou, Q., Mrozek, D., Ma, Q., & Xu, Y. (2017). Scalable Data Mining Algorithms in Computational Biology and Biomedicine. Hindawi BioMed Research International, 2017.