International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 10, pp: 355-360 www.ijaem.net ISSN: 2395-5252

The application of data science and Machine Learning to healthcare - Research perspectives

Mr. M Venkat Ramarao
4th Year Department of CSE, PVP Siddhartha Institute of Technology, Vijayawada,

----
Date of Submission: 20-11-2020 Date of Acceptance: 10-12-2020

ABSTRACT: In these days of Information Technology it may not be hyperbole if someone sates that our lives will come to stand still if every software machine in the universe stops working. As people are habituated to more utilization of mobiles, Laptops, gadgets and other electronic devices, people are getting more health problems. Health care is in the middle of stagy changes at many levels and emphasizing us to get the data that gives deeper understanding of health problems. The various types of data such as genetic with sensor health information are having major impact on methods of how disease is treated and diagnosed .The objective of this paper is to emphasize on Insight into the application of data science and Machine Learning to healthcare - Research perspectives.

Many of researchers with expertise in machine learning, data engineering, soft computing; Swarm Intelligence, IoT, biomedical & medical informatics, statistics, behavioral and decision sciences, and medicine are trying to apply the machine learning and Data Science for healthcare.We work on developing cutting-edge methodologies to derive insights from diverse sources of health data, to support use cases in personalized care delivery and management, real world evidence, health behavior modeling, cognitive health decision support, and translational informatics. The key concepts are Data collection of the patients, Analyzing the patient similarity, prediction or Estimation modeling, Model of deeper disease modeling,

KEYWORDS:data engineering, Health care, Research perspectives

I. DATA SCIENCE:

Technology **has** revolutionized our lives completely today we cannot think of living without a television, mobile phones and the latest addition' our addiction to the internet.The usage of these electronic gadgets, release a high amount of radiation which causes different health issues. They can be solved using different technologies like ML and Data science.

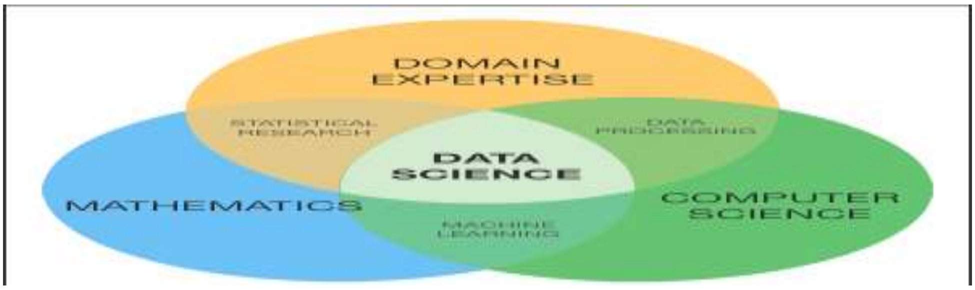

Fig 1.Data Science

Data Science is quite a large and diverse field. As a result, it is really difficult to be a jack of all trades. Traditionally, Data Science would focus on mathematics, computer science and domain expertise. While I will briefly cover some computer science fundamentals, the bulk of this blog will mostly cover the mathematical basics one might either need to brush up on (or even take an entire course).

Data Science Life Cycle: The life cycle of data science can be represented as

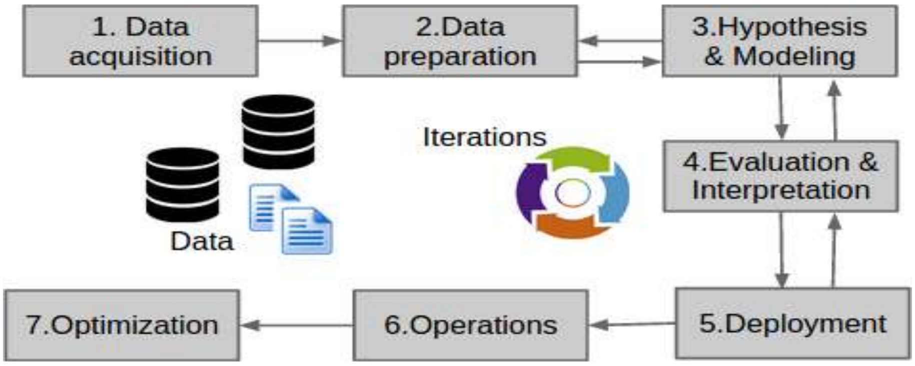

Fig 2. Data Science life cycle

The data science life-cycle phases are
1. Data acquisition 2. Data preparation 3. Hypothesis and modeling
4. Evaluation and Interpretation 5. Deployment 6. Operations7. Optimization
2. Data science on health care:Medicine and healthcare is a revolutionary and promising industry for implementing the data science solutions. Data analytics is moving the medical science to a whole new level, from computerizing medical records to drug discovery and genetic disease exploration. And this is just the beginning. HealthCare and data science are often linked through finances as the industry attempts to reduce its expenses with the help of large amounts of data. Data science and medicine are rapidly developing, and it is important that they advance together.
Tips on implementing data science in health care:

1. Take a holistic view2.Be transparent3.Invite clinical judgment4.Build relationships

II. USE CASES OF DATA SCIENCE:

The following article discusses the use cases of data science with the highest impact and the most significant potential for future development in medicine and healthcare.

Medical image analysis:

The healthcare sector receives great benefits from the data science application in medical imaging. There is a lot of research in this area, and one of the major studies is Big Data Analytics in Healthcare, published in Biomed Research International. According to the study, popular imaging techniques include magnetic resonance imaging (MRI), X-ray, computed tomography, mammography, and so on. Numerous methods are used to tackle the difference in modality, resolution, and dimension of these images. Many more are being developed to improve the image quality, extract data from images more efficiently, and provide the most accurate interpretation. The deep-learning based algorithms increase the diagnostic accuracy by learning from the previous examples and then suggest better treatment solutions. The most popular image-processing techniques focus on enhancement, segmentation, and demising that allows deep analysis of organ anatomy, and detection of diverse disease conditions.

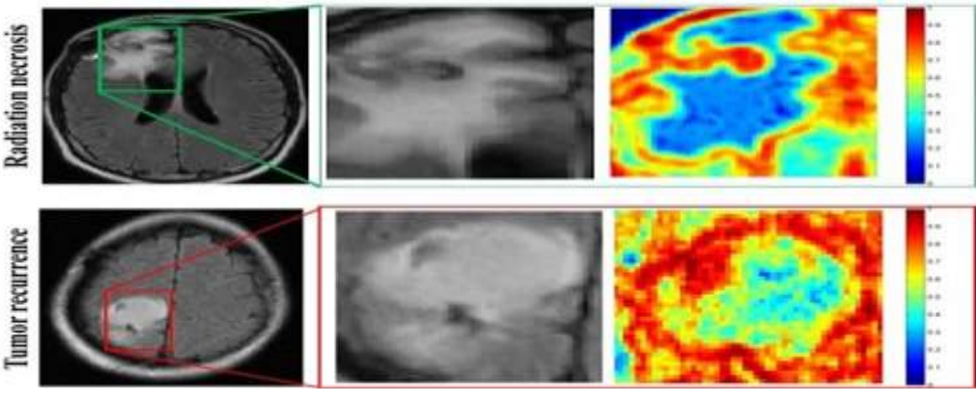

Fig3: Medical image analysis

The most promising applications aim to detect tumors, artery stenosis, organ delineation, etc. Different methods and frameworks contribute to medical imaging in various aspects. Hadoop, a popular analytical framework, employs MapReduce to find the optimal parameters for tasks like lung texture classification. It applies machine learning methods, support vector machines (SVM), content-based medical image indexing, and wavelet analysis for solid texture classification. Other examples include iDASH (integrating data for analysis, anonymization, and sharing) used for biomedical computing, HAMSTER/MPI Graph Labfor processing large images, and more.The data science predictive analytics methods learn from historical data and make accurate predictions about the outcomes. They process the patient data, make sense of clinical notes, find the correlations, associations of symptoms, familiar antecedents, habits, diseases, and then make predictions. The impacts of certain biomedical factors such as genome structure or clinical variables are taken into the account to predict the evolution of certain diseases. Common cases include the prognosis of disease progress or prevention to reduce the risk and the negative outcomes. The main benefit is the improvement of the quality of life for patients and the quality of working conditions for doctors.

Machine Learning algorithms for Healthcare Data analytics can be seen in the below figure.

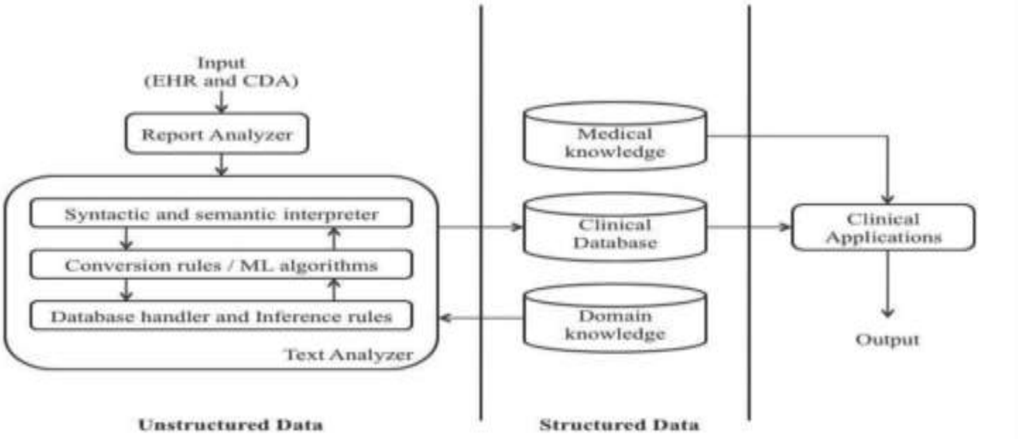

Fig4: Algorithms for Health care

III. DATA SCIENCE OF HEALTHCARE DATA ANALYTICS:

There has been information explosion of big data in the healthcare field. Traditional technologies adopted earlier to analyze genomics, DNA, and cancer with trial and methods through Human Genome Project have taken more than a decade to understand and analyze the composition of DNA and the patterns of the data. Big Data Analytics introduced revolutionary tools and techniques to analyze the chronic diseases for prevention and cure. Genome sequencing has been used to understand the potential root causes of tumor growth causing cancer. The data has grown exponentially from terabytes to Exabyte's. The healthcare data from X-Rays, CT scan and MRI has increased by leaps and bounds concerning the volume of the big data. The advanced technologies of medicine through big data analytics allowed to diagnose the patients records and perform a comparison to a global population to separate the noises from the signal to understand the trends of the tumor growth which was not possible earlier and speed up the diagnosis and treatment. Though there are several theories and techniques that can be applied for the diagnosis of the illnesses, this paper briefly reviews some of the key techniques.

Electronic Health Records: The electronic health record is one of the methods to maintain the entire history of the patient records for analyzing the data for the future as well. The EHR contains significant big data in the form of X-Rays, key observations of the physicians on the fitness and vital signs. The modern big data EHR systems have disparate channels of data sources from the pharmacy, nursing, radiology units, and hospitals through connected network. There are a number of forms to be filled through administration work for registration in a conventional setting of the clinic. However, EHR can automate a large portion of such administrative work including the admission and discharge forms of the patients, billing, and invoicing of the patient check-in and checkout, demographics of the patient. The laboratory systems, pharmacy systems, computerized physicians order entry system, coding systems to organize the healthcare data into particular categories for efficient analysis, and radiology systems are integrated into centralized EHR systems as well. EHR systems data can be both structured and unstructured combining RDBMS and NoSQL data processing techniques.

IV. PHENOTYPING ALGORITHMS THROUGH MACHINE LEARNING FOR DIAGNOSING THE DISEASES:

Phenotyping algorithms can be implemented on EHR data on the disease samples from the hospitals to diagnose the diseases. The unstructured data contains large amount of texts from the physicians' notes, diagnostics, and vital signs records. A Phenotyping algorithm is a special technique that sifts through number of clinical data points through the coding systems with particular billing codes, radiology results, and natural language processing of the large amount of texts from the physicians. Machine learning algorithms with supported vector machine can be applied in identifying the rheumatoid arthritis with the combination of prescription records of the patients to improve the accuracy of predictive models of disease. As an example, usage of hypoglycemic agents from the prescription can suggest the indication of pre-existing condition of diabetes. The Phenotyping algorithm can be applied for cataracts surgery. The EHR data is also combined with biological data banks. A number of machine learning algorithms for various Phenotyping with the aid of coding systems to diagnose the illnesses such as atrial fibrillation, dementia, clopidogrel metabolizers, Type 2 diabetes, sclerosis, and Crohn's disease can be applied to diagnose and detect the diseases.

The genetic variants can be studied for diagnosing the illnesses through univariate and multivariate analysis of genome-wide association studies through machine learning algorithms based on the disease-phenotype attributes. The predictive model has to be optimized to avoid overfitting and under fitting by choosing the best-fit statistical model with accuracy of prediction model. The methods applied for reinforcement learning are supervised machine learning by building a phenotype of the genotype through a labeled training set data to identify the genetic interactions through the analysis of high-dimensional genetic datasets. Risk models are built for the genetic risk prediction. Bayesian statistics with machine learning techniques can be applied to calculate the posterior distribution. Other methods such as linear regression, logistic regression, and Elastic Net with the variants of support vector machine can be applied for modeling the continuous attributes of the phenotypes.

V. DECISION TREES IN HEALTHCARE FIELD:

Decision trees are heavily leveraged in the diagnosis of illnesses in healthcare field. In certain cases, the diagnosis requires constant monitoring of autonomic neuropathy. In the healthcare field, sensors constantly collect the big data from the subject to identify the patterns in the chunks of data sets and for further processing of this data through machine learning algorithms. Identification of cardiovascular autonomic neuropathy through sensors data is the key to understand the vital signs of diabetes. The analysis on this data can be performed through decision trees and ensemble methods. This analysis aids to provide advanced diet and treatment plans for the subject. The research study was conducted by gathering the data from the mobile devices and further the following decision tree and ensemble methods were applied.

· ADTree This technique creates a way two-classification of the problems for generating an alternative decision tree to boost the machine learning.

· J48 Both pruned and unpruned trees are leveraged with this c.45 decision tree classifier.

· NBTree Naïve Bayes Algorithm is applied to generate the decision tree in this instance.

· SimpleCart In this classifier model, the complexity of the pruning is reduced by generating the decision tree.A large number of ensemble methods are applied such as bootstrap aggregation through resampling of the labeled data with randomization through bagging technique. Boosting the algorithms can aid the sequence of the classified on the trained datasets to accelerate the outputs to the next classifier in succession. Wagging, multiboosting, and adaboost are few other methods that are applied in this research method.

Bayesian networks: Big data analytics can aid in identifying the global outbreaks such as flu based on the anonymized electronic health records of the individuals. The Department of Defense's Science and Technology from Victoria, Australia has invented an analytic tool EpiDefend and EpiAttack to identify and target the outbreaks occurring globally through Bayesian network machine learning algorithms. The big data is collected through large-scale environmental data for flu, and influenza, hazardous biological agents, and various other outbreaks. The results are drawn through the probabilistic approach of Bayesian networks. The Bayesian network takes the time series of the electronic health records into consideration to track the patterns and trends of the epidemics. The researchers from Defense's Science and Technology from Victoria leveraged Markovian dynamic Bayesian network approach, text mining to sift through the keywords from the telephone calls to determine epidemics such as anthrax. Particle filtering, Dynamic Bayesian network, and subject-level Bayesian Network algorithms are applied to a large population to determine the outbreaks of epidemics. The text mining is applied on a time series WSARE data sets from the emergency departments that collected the data for the preliminary investigation of the outbreak

Expected Results:1.consolidating differing understanding credits to create comparability investigation by applying propelled machine learning techniques to distinguish exactness partners, joined with displaying approaches for customized prescient models equipped for recognizing tolerant dimension rankings of hazard factors,

2. Ways to deal with location challenges in creating powerful and proficient prescient models from observational medicinal services information in various use cases. Precedents incorporate grid based techniques to address shortage, include, highlight choice, versatile prescient demonstrating stage, customized prescient displaying utilizing exactness associates, and perform multiple tasks learning for complete hazard evaluation, change into wellbeing informatics ,Intellectual choice help,

3. Understanding illness beginning, attributes of ailment stages, rate of movement from asymptomatic to symptomatic sickness, from prior to progressively extreme stages, and factors that impact malady movement pathways.

4. Medication Similarity Analytics joined with cutting edge machine learning strategies, for example, joint network factorization can help pharmaceutical scientists rapidly distinguish drugs that have comparative qualities to target drugs, supporting three unmistakable, yet similarly essential use-cases: Drug Safety, Drug Repositioning and Personalized Medicine, relevant Modeling

5. Imaginative visual examination stage and UIs that quicken the way toward investigating and mining information to infer new experiences that can be converted into increasingly compelling therapeutics and procedures

6. Consolidating ongoing information from wearable gadgets, self-revealed action and clinical information, enables us to show conduct for both forecast and customized health and wellness techniques tweaked to a person's extraordinary needs**.**

VI. CONCLUSION:

This paper explains how the machine learning and Data Science applied to the health informatics. The data science solutions reshape the medicine industry, uncover new insights, and turn brave ideas into reality. The possibilities for integrating data science and healthcare are expanding as the amount of data is growing faster each day, and the technologies are constantly improving. We covered only a small part of the possible use cases, and the list can be complemented continuously. Many general use cases, like fraud detection and Robotization, apply to healthcare, while some specific cases are inherent only to this industry.

REFERENCES:

[1]. Andrea Manieri et al., "Data Science Professional uncovered: How the EDISON Project will contribute to a widely accepted profile for Data Scientists", Proc. The 7th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2015), 30 November – 3 December 2015.

[2]. Y. Demchenko, E. Gruengard, S. Klous, "Instructional Model for Building effective Big Data Curricula for Online and Campus Education", Proc. 6th IEEE Intern Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15–18 Dec 2014.

[3]. Tomasz WiktorWlodarczyk, Thomas J. Hacker, "Problem-Based Learning Approach to a Course in Data Intensive Systems", Cloud Computing Technology and Science (CloudCom) 2014 IEEE 6th International Conference on., 2014.

[4]. Demchenko Yuri, David Bernstein, Adam Belloum, Ana Oprescu, Tomasz W. Wlodarczyk, Cees de Laat, "New Instructional Models for Building Effective Curricula on Cloud Computing Technologies and Engineering", Proc. Workshop "Requirements Engineering for Cloud Computing (RECC)" in conjunction with The 5th IEEE International Conference

and Workshops on Cloud Computing Technology and Science (CloudCom2013), 2–5 December 2013.

[5]. J. Saltz, "The Need for New Processes Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness", Big Data Conference, 2015.

[6]. J. S. Saltz, I. Shamshurin, "Big data team process methodologies: A literature review and the identification of key factors for a project's success", Big Data (Big Data) 2016 IEEE International Conference on, pp. 2872-2879, 2016.

[7]. N. W. Grady, "KDD meets Big Data", Big Data (Big Data) IEEE International Conference on, 2016.

[8]. J. Saltz, I. Shamshurin, C. Connors, "Predicting data science sociotechnical execution challenges by categorizing data science projects", Journal of the Association for Information Science and Technology, 2017

[9]. J. Payne, N. Grady, H. Parker;, "Analytics Ops for Data Science", Big Data (Big Data) 2016 IEEE International Conference on, 2017.

[10]. “The Data Science Handbook: Advice and Insights from 25 Amazing Data Scientists” by Carl Shan, William Chen, Henry Wang, and Max Song

[11]. “Doing Data Science: Straight Talk from the Frontline” by Cathy O'Neil and Rachel Schutt

[12]. “Numsense! Data Science for the Layman: No Math Added” by Annalyn Ng and Kenneth Soo

[13]. “The Art of Data Science” by Roger D. Peng and Elizabeth Matsui

[14]. “Data Science For Dummies” by Lillian Pierson

[15]. “Big Data For Dummies” by Judith Hurwitz, Alan Nugent, Fern Halper, and Marcia Kaufman

[16]. “Data Jujitsu: The Art of Turning Data into Product” by DJ Patil

[17]. “Big Data: A Revolution That Will Transform How We Live, Work, and Think” by Viktor Mayer-Schonberg and Kenneth Cukier