Preventive Healthcare using Machine Learning

Niharika Hegde, Shishir M, Dr. DV Ashoka
Department of Information Science & Engineering
JSS Academy of Technical Education, Bangalore

***Abstract*- The rapid expansion in the use of Machine Learning and Analytics across various domains has been observed in the past few years, owing to the ever growing availability of data sets and a need for extracting actionable information from this enormous data. With the immense advancements in the healthcare community and the subsequent surge in clinical data, it has become pivotal to use elements of Machine Learning and Analytics for improved medical care and disease detection.**

Conventional healthcare depends heavily on the incidence of a disease and its major symptoms for treatment. This paper, however, strives to provide an insight into how Machine Learning principles can transform Medical Domain into being disease detection and prevention oriented. In this paper, we propose the use of algorithms and techniques to streamline analysis, identification and extraction of disease patterns and markers which could have otherwise been hidden or overlooked.

Preventive Healthcare, therefore, can potentially help saves lives globally by alleviating the stress on the healthcare sector and elevating the quality of patient care.

Keywords: Machine Learning, Healthcare, Big Data, Data Analytics, Preventive Medicine

I. INTRODUCTION

The advancement in technology and digitization has led to the generation of tremendous amounts of data, which for the last two years is estimated to be at approximately 2.5 Quintillion bytes of data being generated every day.[1] About 30% of this data was generated in the Healthcare sector, drawn from various sources such as patient health registers, laboratory test reports, procedural archives or even patient surveys. This data has clinical, monetary, operational value and has to be managed cautiously to make sure not to compromise the data quality.

Traditionally, this vast data was stored as a hard copy, an approach which soon became obsolete due to the Big Data scale growth of data. An alternative to storing all of this information was in the form of electronic health records (EHR), on the cloud.

EHR is composed of data such as comprehensive medical records, genetic conditions, results from examinations, diagnosis notes, data from medical imaging and detailed accounts of other vital information, since modern clinics and hospitals are equipped with devices for constant surveillance and data gathering.

With the advent of complex and advanced Machine Learning and Data Analytics models, these records are of great significance in discovering clandestine details which can assist in providing effective medical care through precision medicine, making healthcare affordable and providing more targeted treatment by focusing on certain groups during model generation.

Hence, the multitude of complex information cannot be captured, processed and analyzed efficiently by humans without the assistance of high powered computing from strategies that derive from a plethora of Machine Leaning and Data Analysis tools. These built-in tools enable computers to uncover information such as risk factors, early markers etc. without being explicitly programmed, which could help doctors in making quick and informed decisions on the diagnosis and treatment of patients by eliminating the risk of an uninformed, rushed diagnosis.

II. PREVENTIVE HEALTHCARE

The current healthcare system focuses on diagnosing and treating diseases only after the first occurrence of major, apparent symptoms. This delay between the incubation period of the disease and symptom manifestation can lead to complications in treatment, and in extreme cases, such as HIV and Ebola, may also cause death.

A 2004 study [2] showed that half of all deaths in the year 2000 in USA were caused due to preventable behaviors and exposures. Some of the causes of death from the study include diabetes, chronic diseases, cardiovascular diseases and other infectious diseases. The incidence of these diseases could have been predicted for a specific patient demographic using Analysis techniques along with Machine Learning Algorithms (such as Regression Analysis) [3], on both structured and unstructured data present.

As the proverb goes, “Prevention is better than cure”, preventive healthcare emphasizes on the early detection and prevention of diseases rather than treatment of the disease.

Some of the methods actively employed in preventive healthcare are regular doctor check-ups, even if the person is healthy, screening of diseases, identification of risk factors, timely immunization and health boosters. [4]

Genetic testing and screening can also be performed for early detection and prevention in patients who are genetically susceptible to disorders such as Cancer, Dementia, and Alzheimer's. [4]

After careful analysis of datasets by Accenture that resulted in a study, which predicted that a few terminally-ill patients often utilize more than 60% of the resources and services a hospital may provide. An important inference made from this study was that if these patients were treated and monitored from home, it could reduce the costs of medical care substantially and also help avert readmission to the hospital.

The information from the medical sector, hence, has the potential to cause major changes in the economy, as suggested by the McKinsey Report [5], which predicts Data Analysis and Machine Learning adding close to $300 billion value to the US Healthcare Sector annually. This kind of a surge in the revenue towards healthcare has innumerable benefits associated with it, such as effective healthcare, reduction in treatment costs, reduction in cost of medication and mortality rate.

With Machine Learning and Data Analytics powering Preventive Healthcare, outbreaks of diseases locally can be prevented from being full-fledged global epidemics. For example, Google's Flu Trends and Global Viral Forecasting Initiative (GVFI) are services for predicting the occurrence of the flu, which implement complex aggregated data processing techniques on information such as the location, previous outbreaks in the geographical area and susceptibility of residents due to gender, ethnicity, genetic predisposition or any other unknown factor. This would help in predicting and preventing the outbreak of a global epidemic and could save countless lives.

Another major contribution of the Preventive Healthcare Approach is towards reducing the stress on the healthcare domain globally, which is made possible by effective and relevant analytical predictions. In the long run, this insinuates globally affordable healthcare and medication. The impact this could have for developing countries on overcoming limited access to medical care would be massive.

This can be attributed to the fact that many healthcare organizations all over the world use Analytics and Machine Learning to create forecasts and models on the how to improve and utilize the resources available to treat the 33 million people in these countries. [6]

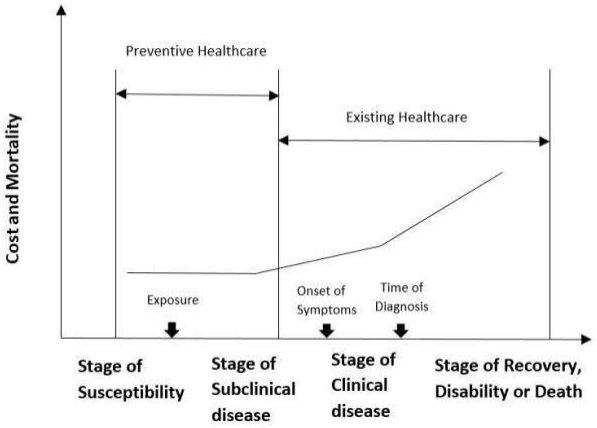

Fig 1: Disease Timeline

III. MACHINE LEARNING

Machine learning is defined as the science of getting computers to learn with data and recognize patterns automatically by experience, without having to be programmed explicitly.

This process of learning is referred to as 'Training', and the resulting outcome is a 'Model'.

The model operates on vast data and discovers insights based on what it had previously learnt.

There are two factors that are crucial for the successful application of Machine Learning:

- Efficient algorithms
- Rich data sets

Machine learning determines a set of rules using the data provided. The more data a Machine learning model is fed, the more complex the rules and the results are more accurate in predictions.

Machine Learning Models:

Machine Learning consists of many models for various types of data sets. Each these models trace their own algorithmic approach to interpret and learn from the data set fed to them.

In this Machine Learning Workflow, Training is one of the stages, wherein a data scientist develops, tests and validates a mathematical model that can formulate a target value or attribute.

Following are some of the models:

A. Classification Model

Classification predictive modelling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

The output variables are called as labels or categories.

Algorithms for classification are commonly used in defect prediction and classification, Neural Networks, Naive Bayes and Decision Trees. [7]

The motive for the classification model is to determine a label or category.

For instance, if we had to determine non-invasively whether a tumor or a lump is benign or malignant, the model would be fed with data sets consisting of attributes

such as patient's medical notes, symptoms, incubation period, lifestyle choices, as well as results from previous such incidents.

The model would then classify it either as being a benign or malignant tumor.

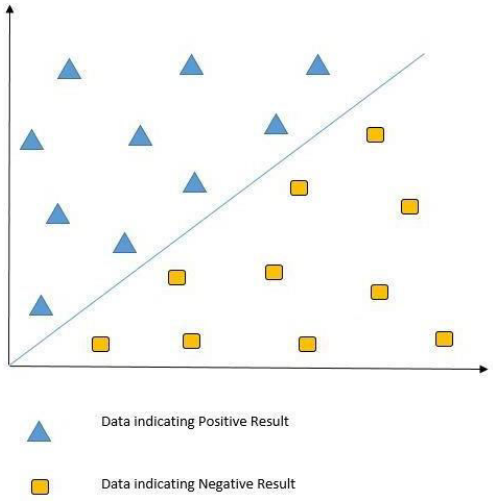

Fig 2: Classification Model

B. Clustering Model

Clustering model focuses on finding patterns in the data sets. It is useful in cases where there is no expected outcome and the objective is to uncover any distinctive, hidden patterns.

For example, if a healthcare provider wishes to see if there is a pattern among the patients over the past year, the model would be provided with patient records as the data set.

The model might detect that people living in a specific locality are more susceptible to contract a particular disease. This insight can lead to further investigation to determine the cause of the occurrence and help avoid future manifestation.

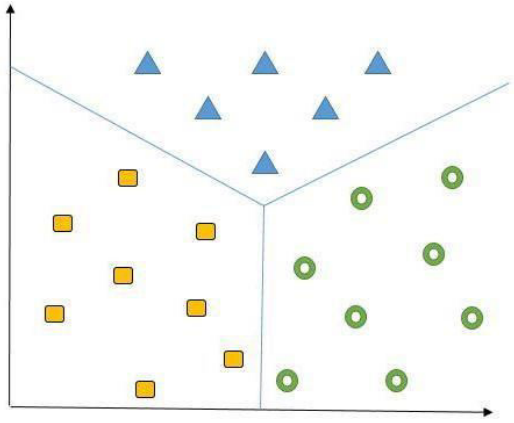

Fig 3: Clustering Model

C. Regression Model

Regression model is used in the cases where we need to determine numeric values. Using the regression model, a data scientist would be able to predict for instance, the projected duration of treatment.

For example, if the objective is to determine how many days it would take for a patient with a chronic condition to return to the hospital for treatment, we provide the model a labelled data set, it comprises of information such as age, gender, condition, records from prior treatment courses etc.

The model is also used for predicting a continuous quantity as output, which is a standard problem in the case of Regression.

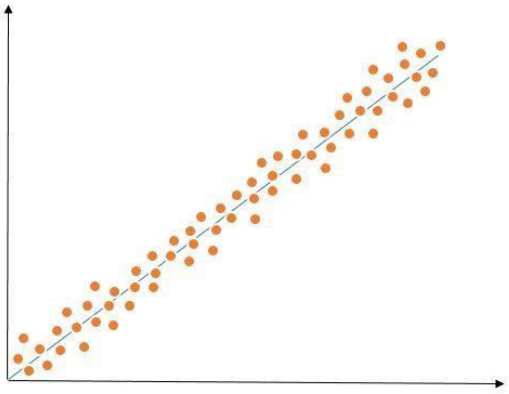

Fig 4: Regression Model

IV. APPLICATIONS OF MACHINE LEARNING IN HEALTHCARE

A. Wearable Devices

With the advancements in technology, everyday devices have become smarter and are capable of channeling large amounts of data. This means that we can now continuously measure various vital signs of the body, such as heart beat, blood glucose, stress levels, sleep patterns etc. This data generated every day approximately equals 2 terabytes.

Using Machine Learning, we can analyze the above information constantly and detect even the slightest changes, and predict possible disorders.

Sudden Cardiac Arrest (SCA), which is a cardiovascular disease, claims 4280 lives for every 1 lakh population in India alone, [8] and causes 3 million deaths around the world. It is caused due to an abnormality in the electrical functioning of the heart and lacks any obvious external symptoms, which adds to the complexity of its detection and treatment.

An approach that could help in early detection and prevention of SCA is by identifying high-

risk patients who are most susceptible to SCA, such as patients with history of drinking, smoking, drug abuse, diabetes, high blood pressure etc. [9]. These patients are then constantly monitored by the means of Wearable Devices. The data from these devices are scrutinized by Machine Learning and Analytics Tools to monitor heart activity and detect even the slightest of arrhythmia, as prolonged ventricular arrhythmia is often said to prelude SCA. [9]

Such predictions on the basis of Heart Rate Variability (HVA) could help save the lives of patients by curbing the delay of receiving medical care.

B. Precision Medicine

With the help of Electronic Health Records (EHR) which store information such as clinical treatment history, genetic conditions, allergies and test results, Physicians would now be able to use Machine Learning to process this information for a more tailored and accurate medical treatment.

For example, IBM Watson traces its first major commercial application in rifling through 200 million pages of patient records, medical literature and reference materials to enable the feature of self-reporting of symptoms or major risk-factors to the Doctor. This Machine Learning and Analytics application could help in accurate decision making regarding the diagnosis and treatment by predicting the occurrence of diseases, and also to find the trends and treatments that are more successful, which could save time and money for the patient and the Doctor by significantly improving the scenario of a more targeted treatment approach.

C. Varied Dataset Analysis for Diagnosis and Treatment

Machine Learning is a technology that has made it possible for the secure querying of data using a wide array of Analytical Tools that are capable of parsing and scanning an enormous amount of information, in different forms and from various scattered sources.

For breast cancer diagnosis, oncologists and radiologists have been using the technique of labelling focus points on high-resolution mammograms and then feeding it to systems. This process has various drawbacks, the major one being that it's a protracted, exhaustive and quite an outdated approach.

With various tools and equipment available for the diagnosis of cancer tissue, a new approach, called the "semi supervised learning" [10] has gained momentum. In this technique, a small amount of carefully labelled and diagnosed mammograms, along with unlabeled data together are used to enhance data processing capabilities and thereby, the performance in terms of self-learning and prediction.

Due to the capabilities of Machine Learning and Analytics extended to the massive quantity of medical research data available, it has also led to unexpected benefits such as the discovery that Desipramine, which was originally used only an anti-depressant, also has the ability to help treat certain types of lung cancer.

D. Preventive Healthcare

Using Machine learning, medical personnel can now process the vast information in health records and identify and avoid the occurrence of a disease.

This application is particularly helpful in cases where the patient is suffering from multiple conditions, has multiple symptoms and a complex or incomplete medical history.

With the help of machine learning, we will now be able to predict if patient is at the risk of diabetes and take precautionary measures such as testing for underlying conditions and begin weight management.

E. Makes Healthcare Affordable

Preventative medicine has led to an overall reduction in healthcare costs.

For example, Clover Health, a health insurance provider which makes use of data driven analytics, reports 50 percent lesser hospital admissions and 34 percent fewer hospital readmissions.

It implements algorithms that are capable of identifying at-risk patients, and ensures that necessary steps are taken.

Further, it aids physicians in making decisions, thereby resulting in significant savings.

Processing of data sets to optimize the knee replacement process has helped in saving over $1.2 million in a year.

F. Gene Sequencing

The Human Genome Project, which took nearly 13 years to complete, was a laudable endeavor towards identification as well as mapping of all the genes in the Human Genome and to determine the sequence of the 3-billion nucleotide base-pairs which make up the human DNA.

Now, with the use of Machine Learning and Analytics, there is a possibility for a major breakthrough in bio-medicine in regards to genetic disease research, when used for Gene Sequencing, coupled with the data from the abundance of patient records and clinical data.

The use of algorithms such as Weighted Voting, K- Nearest Neighbor and Support Vector Machines could also help expedite the process of Gene Sequencing of other species, which could help shed more light on their physical, functional and structural composition and lead to more discoveries pertaining to various realms.

V. CHALLENGES OF MACHINE LEARNING IN HEALTHCARE

With the Healthcare information being generated crossing over in size to hundreds of Petabytes [10], the complexity of the task of Analysis and Modelling is also on the rise.

Some of these challenges are illustrated below:

- Security and privacy are major concerns in the field of healthcare, due to the sensitive nature of information stored in the health records. It is crucial that this information is safeguarded against misuse and must be made available only to authorized personnel.
- Medical data is spread across various organizations such as hospitals, diagnostic centers and labs, and administrative departments. Integrating data from these multiple sources would require the development of new infrastructure where all of these data sources can collaborate.
- In predictive medicine using Machine Learning and Data Analytics, the possibility of datasets with missing, incomplete, noise and non-transparent data is ever-present. The algorithms we implement to extract data or models have to be appropriately equipped with the right tools and means for handling these types of data, without compromising on the overall quality or reliability of the results as any compromise could be catastrophic.
- In case of predictive models, it is crucial to bear in mind the costs of false-positives and false-negatives.
  A false-negative may be detrimental to the patient's heath.
  False-positives may result in unwanted and expensive treatments, which could be inimical to the patient as well as the overall economy.

VI. CONCLUSION

In this paper, we have expounded on the role of Machine Learning in transforming the medical domain from being merely treatment-oriented to disease prediction and prevention oriented.

The advancements in healthcare sector and related industries have contributed to the surge in data streaming in from various devices, wearables, countless patient records, research data and statistics. The tremendous potential that this raw data holds must be extracted by adopting the tools and modelling techniques of Machine Learning and Analytics. The information thus obtained can be streamlined to provide medical practitioners with critical facts about the status of their patients, thus aiding them in making informed decisions regarding treatment.

Machine Learning and Analytics could also influence the state of healthcare globally by providing pharmaceutical companies and healthcare organizations with much needed insights into disease prediction and their targeted treatment, and also help divert resources to these developing countries, which can prove to be critical for the medical care in developing countries.

Hence, Machine Learning has proved to be indispensable in the healthcare sector due to its contribution towards predicting the occurrence of diseases, premeditating the treatment outcomes, making healthcare more effective and affordable.

REFERENCES

[1] https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/

[2] Mokdad A. H., Marks J. S., Stroup D. F., Gerberding J. L. (2004). "Actual Causes of Death in the United States, 2000". Journal of the American Medical Association. 291 (10): 1238–1245

[3] M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, "Disease prediction by machine learning over big data from healthcare communities", IEEE Access, vol. 5, pp. 8869-8879, 2017

[4] Vorvick, L. (2013). Preventive health care. In D. Zieve, D. R. Eltz, S. Slon, & N. Wang (Eds.), The A.D.A.M. Medical Encyclopedia

[5] Manyika J., Chui M., Brown B., et al. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute; 2011

[6] Bram JT, Warwick-Clark B, Obeysekare E, Mehta K. Utilization and monetization of healthcare data in developing countries. Big Data. June 2015;3:59–66.

[7] Ajay Prakash B.V., Ashoka D.V., Manjunath Aradya V.N. (2017) Exploration of Machine Learning Techniques for Defect Classification. In: Vishwakarma H., Akashe S. (eds) Computing and Network Sustainability. Lecture Notes in Networks and Systems, vol 12. Springer, Singapore

[8] Rao BH. Global burden of sudden cardiac death and insights from India. Indian Heart J 2014;66:S18–23.

[9] Loganathan, Murukesan & Htut, Ye & M, Murugappan. (2014). Machine Learning Approach for Sudden Cardiac Arrest (SCA) Prediction Based on Optimal Heart Rate Variability (HRV) Features. Journal of Medical Imaging and Health (0.642).

[10] M. Li, Z.-H. Zhou, "Improve Computer-Aided Diagnosis with Machine Learning Techniques Using Undiagnosed Samples", IEEE Trans. Systems Man and CyberneticsPart A, vol. 38, 2008.

[11] H. Chang Book review: Data-driven healthcare & analytics in a big data world
Healthcare informatics research, 21 (1) (2015), pp. 61-62