Deep Science Publishing
https://doi.org/10.70593/978-93-49307-36-0_2

DeepScience
Open Access Books

Chapter 2: Leveraging machine learning algorithms to predict, prevent, and personalize treatment for chronic conditions

2.1. Introduction

Chronic diseases account for over half of all global deaths, and they impose a substantial burden on patients, healthcare systems, and economies. Chronic diseases require long-term and complex treatments and are thus costly and more difficult to manage. As one category of chronic diseases, chronic kidney disease affects about a tenth of the global population. A large percentage of patients with chronic diseases have comorbidities. Multiple chronic conditions not only worsen the quality of life of patients but also present an additional challenge to healthcare systems with regard to coordinated and effective long-term care. This challenge is exacerbated by the rapidly aging global population.

Advances in big data analytics provide an unprecedented opportunity to address the above challenge. High-dimensional electronic health records, which contain comprehensive longitudinal health information, are increasingly being collected. Machine learning algorithms can free people from analyzing large-scale, complex, and high-dimensional data and have proven to be powerful tools in producing predictive models, discovering patterns, and generating insights. To this end, predictive models to accurately estimate future risks of developing a chronic condition would be of great use in selecting high-risk cohorts for effective intervention and care. Predictive models to optimally target a high-risk cohort for specific conditions would be valuable in preventing or delaying disease onset. For diagnosed conditions, there is an increasing emphasis on precision medicine. Optimal personalized treatment can significantly improve the outcomes of patients with chronic diseases.

Advances in big data analytics are revolutionizing the way healthcare challenges, particularly those related to chronic diseases, are addressed. With the growing availability of high-dimensional electronic health records (EHRs) that capture detailed, longitudinal patient data, there is an unprecedented opportunity to leverage machine learning to uncover actionable insights. These algorithms excel at analyzing complex and large-scale datasets, enabling the development of predictive models that can accurately estimate an individual's future risk of developing chronic conditions. Such models are instrumental in identifying high-risk cohorts, allowing for targeted preventive interventions and more efficient allocation of healthcare resources. Furthermore, in the era of precision medicine, predictive analytics facilitates the personalization of treatment plans, optimizing care strategies based on individual patient characteristics and improving outcomes for those already diagnosed. This data-driven approach not only enhances disease prevention but also supports tailored, effective management of chronic illnesses.

Fig 2.1: AI in Chronic Disease Management

2.1.1. Research design

Research design. Study design. The study design is a retrospective cohort analysis using machine learning algorithms to predict the future risk of developing epilepsy. The dataset used in this study comes from the Department of For Science, College of Agriculture. From 1988 to 2018, 30 data were extracted from journals, patents, and other sources. Raw data was processed; features were engineered, and uninformative features were removed. Finally, the selected 7 data attributes of the pattern count, similar situation software, similar software, inventors, development stage, times-time, and future risk. Unbalanced classification was adopted in the processing of future risk, marking 526 developers as 1 and 229 developers as 0. 556 copyright warnings were classified as 1 and 199 non-copyright warnings were classified as 0. A total of 785 fraud sites were received, and 402 fraud prevention scites. Classifier. The classification process is divided into four steps. The first step is data enrichment. The data enriched for future risks were the CPC, IPC, and Uiscite. The data enriched for text-copy were Authors and Titles. The data before and after repair was h-1. The second step is to extract the features. The LSA was used to extract textual content; the SDK was used to extract the content of the token trace; and the function was used to fit the unnormalized form of the time attributes. The third step was to train the classifier.

2.2. Understanding Chronic Conditions

Chronic conditions are one of the most significant public health challenges in the 21st century. Defined as a long-term condition that typically advances slowly over time and affects a person throughout life, symptoms for these conditions may be impossible to reverse or even manage. The illnesses cause the most deaths globally. They contributed to the decrease in life expectancy of a large part of the population worldwide. By managing the most common chronic conditions, it is possible to lengthen and improve the quality of life of many people, with a significant positive impact on individuals and on health systems' general performance. The ability to avoid the possible negative outcomes of chronic conditions is hard and complex. Various causes contribute to the growing burden of these diseases, making them difficult to control. Advances in technology and healthcare have allowed for the capture and storage of huge amounts of personal health data. However, clinicians and health authorities often use only a small part of this knowledge to design and implement appropriate intervention strategies. Machine learning algorithms have emerged as powerful tools to analyze and explore the complex relationships between the states of health and the numerous factors possibly influencing them. Thus, these algorithms may help in managing the treatment of chronic conditions better, representing a valuable resource for both patients and health organizations.

2.2.1. Definition and Classification

The prediction on chronic diseases has gained increasing attention from physicians. Around 40 million people lose their lives due to chronic diseases every year. The key to prolong the life of sufferers is to make an early and accurate diagnosis, and in the meantime, it can provide some beneficial treatment suggestions for practical applications. As a kind of data-driven method, a machine learning algorithm is an effective way to model and analyze the relation of a large amount of medical data and disease label information that can be utilized to predict the risk of chronic disease in an individual patient. Little exists about leveraging machine learning algorithms to predict, prevent, and personalize treatment for these conditions using the above frameworks.

A classification model is built on the chronic diseases data which consist of diabetes, acute, or congestive heart failure. A Binary Classification model tries to deduce from the input values given for training, and a Multi-class Classification model presses on with prediction of categorical value other than Binary Class process. In addition, a Multi-tag classification model in the sense that its accurate forecasting requires gathering of a set of target tags rather than simply tagged value. In BI classification models try to forecast either diabetes or not with regards to the given input data or feature vector. Actually, the machine learning approach aims to predict whether the chance of loan payment has binary output 1 (if the subject payment has been done), otherwise as binary output 0 (if the subject payment has not been done) in a given set of input parameters. The list of parameters which compose input feature vector to model involve Age, Maximum RT, Minimum RT, Average RT, Velocity of Rt-ankle joint, Maximum EDA, Minimum EDA, and Average EDA.

2.2.2. Prevalence and Impact

Of the top five health conditions in the United States, four—heart disease, cancer, stroke, and diabetes—are chronic. Chronic conditions are represented in the remaining conditions within the top ten—benign prostatic hyperplasia, osteoporosis, and hypertension. The probability of living one's life without chronic conditions is less than fifty-fifty. The probability of living one's life with someone who does not have chronic conditions is negligible. The prevalence continues to grow as the population ages. In 2000, 125 million Americans over the age of eighteen lived with at least one chronic condition. By 2050, the number is expected to reach 171 million. Chronic conditions now occupy the center of the nation's research interest. Policy Prescription: 1. Link bio-preparedness research and bio-defense research. 2. Increase epidemiological research.

Chronic conditions interfere with life's daily functions and have a profound impact on quality of life. The cumulative burden of the conditions for a particular individual is the

product of (1) an increase in the number of conditions and (2) an increase in the DET/DLT ratio. Chronic conditions are also distinguished by their potential to lead to catastrophic health expenditures. Encouraging the insurance industry to increase the number of catastrophic conditions on which they sell insurance may have further reduced the demand for health coverage. Government-provided insurance crowds out the private market. As Medicare will soon cover nearly half the population, individuals may have little incentive to purchase supplemental or catastrophic coverage from the private market that necessitates considerable time and effort.

2.3. Machine Learning Fundamentals

Healthcare data are among the hardest-to-analyze data. In the healthcare sector, research not only needs to follow strict ethics and laws but also has to face cross-disciplinary challenges (Elenitoba-Johnson & Brown, 2020; Acharya et al., 2023; Higgins et al., 2023). Research data can be roughly classified into two types: survey data and clinical data. Survey data typically contain questions asking for opinions or behaviors to self-report. These self-reports may be subjective or noisy. In contrast, clinical medical data come from diagnoses, exams, and treatments. Before the year 2000, most clinical data were collected in paper documents. These healthcare data contain mixed information like blood reports, medical imaging reports, pathology reports, genotype, etc. Today, they are gradually transformed to electronic formats like Discharge Summary or Electronic Health Record, however only about 37% of the cross-institutional medical records were clean data. The time span, i.e. the age of medical record, is also a factor. Health reports from an old period are less reliable since doctors have no technology for detailed exams. Besides, due to competitions, trustability, tenure and daily work, patients usually switch to another healthcare provider, resulting in observable change in healthcare records. There is a ratio of 200 to 1600 days that a patient revisits a healthcare provider with the same (pseudo)personal ID found in the data of more than one provider. While handling medical data, there are strict rules to be followed with the Health Insurance Portability and Accountability Act. For health data in the US, it is not allowed to contain 18 kinds of identifiers to protect the privacy of patients, making the identities of the patients concealed. While the data preprocessing step is simple, the above rules make it tedious to write the data out for storage or input it to the model. Moreover, missing values are common in medical data and almost everything in medicine is imbalanced, such as experiment data, class distribution, prediction accuracy and efficacy. In addition, the prediction might take time, but there is a limit between when the prediction can be made and the time it is used, therefore, the data available to the model is time dependent.

2.3.1. Overview of Machine Learning

Machine learning (ML) is in its renaissance era and holds enormous potential in offering personalized and innovative treatments to a variety of chronic conditions (Li et al., 2023; Tao et al., 2023). Machine learning involves the development of a computational model, also known as a learner or algorithm, to perform a task. The development of this mathematical model that is capable of performing a particular task is in part made possible by the underlying representational power of the model and by providing numerical data to the model that defines the problem. Machine learning effectively makes the computer learn the parameters of its model from the incoming data rather than having the model's parameters explicitly defined by reflecting the desired features of the model.

During training, the model adjusts its weights (parameters) in an iterative manner to minimize its loss function. Loss is a measure of how far the model's predictions are from the true ground-truth label of the training data. In simple terms, the loss can be considered the "penalty" the model incurs when making wrong predictions with increasing severity as the discrepancy between prediction and ground-truth label increases. After the model is trained, it can be used to assign predictions to new data not seen during training. The numerical or categorical output of the model might consist of the predicted class label or it may consist of a continuous variable to predict a quantity. Finally, the output of the model is processed in a post-processing step to derive insights, whether that be the feature weighting or decision taken.

Fig 2.2: Machine learning in healthcare

2.3.2. Types of Machine Learning Algorithms

A machine learning algorithm is a method that evaluates an artificial neural network. The quality of this evaluation, also known as training, depends on the data with which the network is trained. Once the network is trained, it can start making predictions based on new data. It can adjust its own weights and biases in response to the data with which it is trained, discovering the underlying structure and patterns. There are many machine learning algorithms to choose from, such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), and Naïve Bayes (NB). Of the top 10 algorithms, Two-Class Boosted Decision Trees (BT) has the highest average AUC, precision, sensitivity, specificity, F1, and MCC. In addition, ensemble learning algorithms have been known to improve the performances of individual classifiers in terms of accuracy, sensitivity, and specificity, while also decreasing the need for calibration between the model hyperparameters. RandomForest, an ensemble-based BT classifier, improves BT based on two-class decision trees by creating a large number of trees using bootstrapping and random selection of predictive attributes to fit the training data category. The predictions of a test instance from all base classifiers are combined by some strategy, such as averaging, majority voting, or genetic algorithms, to form the ensembled prediction. This strategy usually provides better results than each classifier. The outputs indicate the relevance of attributes relative to the target attribute. Gradient Boosted Trees (GBT) is also an ensemble learning algorithm based on weighted sum (boosting) of BT applied on a number of prior bootstrapped instances with replacement resulting in D decision trees, where weights are the number of participants in each category and bias is the logarithm of the rate of correct positive predictions to all positive predictions. Random discards (feature manipulation) are widely used because it helps create a stable model, allows for parallel learning, and improves the generalization error of the trees. Bagged and PM algorithms are also applicable to the feature manipulation on BT classifier of GBT ensemble.

2.4. Predictive Analytics in Chronic Conditions

Predictive modeling and artificial intelligence (AI) driven by electronic health record (EHR) data have great potential to improve care in a wide range of clinical areas including the diagnosis and risk assessment of acute and chronic conditions, lifestyle management, home monitoring, and treatment optimization. Type-2 Diabetes Mellitus (T2DM) is a common chronic condition characterized by resistance to insulin and its associated metabolic dysregulation. Recent advances in the collection and storage of complex longitudinal data enable new opportunities to develop more sophisticated algorithms and tools. Clinical decision support (CDS) is a promising approach to diagnosing and informing the care of chronic conditions such as T2DM, and machine

learning (ML)-based CDS tools are being explored for providing pharmacotherapy recommendations and predicting the risk of subsequent complications. Additionally, to improve care in the outpatient setting, there is ongoing interest in how complex predictive algorithms may be constructed to better manage patients with T2DM. Such predictors can be used to generate early warnings or inform clinical decision-making by providing personalized information about the likelihood of a potential event, side-effect, or to optimize treatment policies. Despite these interests and ongoing research investments, challenges still remain in developing AI-driven predictive analytics and CDS tools.

2.4.1. Risk Stratification Models

Patient data contain information derived from screening, diagnostic testing, and follow-up evaluations before and after a particular disease. Most patients progress well with treatment, but some encounter adverse events, either due to personal sensitivity to drugs or idiosyncratic drug reactions. The goal is to link the patient data with their post-treatment outcome observed in hospitals in order to isolate predictive indicators of potential health issues or underperformance after treatment. A disease risk stratification model can be trained using patient post-treatment outcome on a disease of interest. Given the training set and the patient data, the model output is the predicted risk of the patient in 30 days upon treatment of the disease. The cohort of patients based on the predicted risk is then ranked and grouped into bins which are used as the risk stratification output. Then methods for (1) training the risk stratification model using a distributed SQL database system and (2) computing the risk stratification performance using the distributed SQL query engine are discussed. Finally, insights gained from exploring the patient hospital claims data are presented. Taking advantage of the structured query language (SQL), an easy-to-understand modeling language in managing and analyzing relational databases, effectiveness can be shown in training a machine learning model using the SQL UP database system. Another easy-to-understand modeling language, SQL procedures, can be extended to sophisticated modeling to accelerate a general practitioner and specialist's algorithm implementations in clinical big data analysis. Being a model validation tool to deliver clinical risk factors in a big data environment and Admissions: Length of Stay and readmission forecasting can be important with potential impact for others who will leverage intensive SQL analytics for big data research.

2.4.2. Early Detection Techniques

Many people establish one or more chronic diseases like Arrhythmia, Asthma, Bronchitis, Chronic kidney diseases (CKD), Cardiovascular diseases, Chronic obstructive pulmonary diseases (COPD), Diabetes. Chronic diseases may last a long time but may likewise go ahead and off periods. The manifestation of a disease defining chronicity on a long time base acts either as intensity of severity, frequency, or duration. Sporadic side effects of treatment over a lifespan explain chronicity in some diseases. Fine mapping on symptoms resorts to chronic conditions because disease mapping defines the essential cause of them. Always existing diseases are chronic and persistent manifestations enough to harm life. But some people lack acute topological conditions by not converting. Most frequently because of a chronic condition; ex- patients diagnosed with copd, Patients diagnosed with arthritis. But a chronic condition can be either detected as a result of complaints of the patient over long periods without a discrete episode or classified as chronic with direct monitoring and examination. Within this work, Chronic diseases like the most Stylish one (asthma), the most covered by the literature (cancer, diabetes), the most life-threatening one (cardiovascular diseases), CKD covering seasonal diseases (flu, gastrointestinal disease), and Malignant dangerous disease types are taken into consideration. The techniques employed in the early prediction of diseases cover most troubling diseases for the state of existence of decision-making systems. There is a lengthy consumer story on the early diagnosis of cancer because of its malignant state and subsequent death necessitating a name of the illness with hope for not being so compatible with life. With the prevalence of technology, repeated studies classified as similar results because of the same datasets can ease the early diagnosis of cancer frequently produced in diet and lifestyle and treated earlier with prognostic symptoms such as diabetes, can draw more interest.

2.5. Preventive Measures through Machine Learning

Publications on healthcare services have shown that chronic conditions impose a major economic burden on many health systems. The number of patients with a chronic condition is growing every year. Recently, with the advancements of large-scale Electronic Health Records (EHRs) and machine learning strategies, it is possible to leverage this patient data to predict and prevent the onset of chronic diseases. The availability of personal data allows designing treatment plans tailored to the patient's individual risk profile. It examined how to allocate preventive treatment to prediabetic patients with application to diabetes mellitus type II. In a large-scale analysis on EHR data, a cohort of prediabetic patients is studied who have notifications of their condition and are not treated with Metformin. Previous to the onset of diseases, the cases and matched controls are identified and a data-driven decision model is proposed for

allocating the preventive treatment on the individual level. The cost-effects of preventing patients from developing diabetes mellitus type II are discussed and it is shown that this is possible in a large number of patients.

For the assignment, it has been considered how the preventive treatment might be allocated to prediabetic patients who are likely to develop the disease. A group of patients is considered, which have some notifications of their condition and are not treated with Metformin. On the individual level, data-driven decision rules are proposed, suggesting these patients which patients shall be treated with the drug and which should rather not, and it is discussed how the allocation of these drugs affects the probability of patients developing diabetes mellitus type II. On the cohort level, a Bayesian approach is pursued assuming a budget of treatment costs: For how many patients is the prevention possible, i.e. how many patients, who would otherwise develop the disease, can be prevented?.

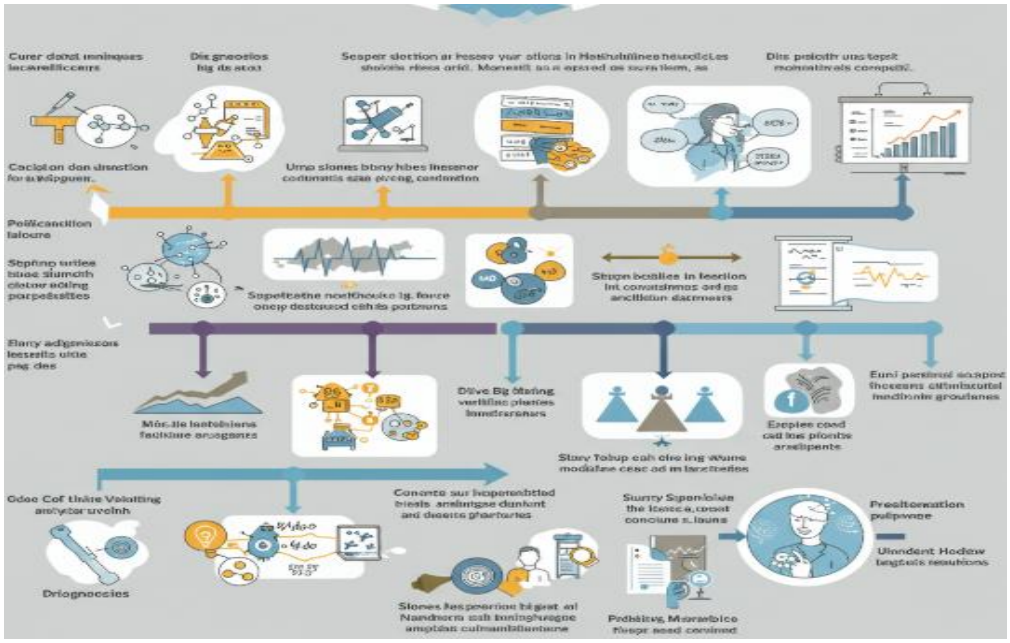

Fig : Evolution of machine learning healthcare analytics research

2.5.1. Identifying At-Risk Populations

The manner and extent to which the prevention of chronic conditions can be studied and implemented change as a response to a number of variables, such as lifestyle, genetics, and environmental factors, among others. Utilizing medically integrated devices, it is possible to collect data on a great number of variables related to one's health, such as

oxygen levels, activity levels, and heart rates, to name a few. By leveraging that data and presenting it to machine learning algorithms, it becomes plausible to predict future chronic conditions of the user currently providing their health-related data. This can be either a patient of a primary health care provider or various other users actively wearing medical or consumer-grade health devices. Using independent health department data, modeling can be done on a particular diagnosis and also on prevention strategies that a care provider could undertake to prevent its population from developing the chronic condition predicted with the usage of respective modeling work.

2.5.2. Behavioral Interventions

Type 2 diabetes is a chronic disease that affects approximately 30 million adults in the United States and can lead to microvascular complications, neuropathy, and death due to multiple causes if not properly managed. Behavioral interventions can prevent or delay the development of type 2 diabetes. However, human-driven interventions are resource-intensive and costly. In recent years, technology-based interventions such as smartphone apps and wearable devices have started to show promise in scaling such interventions. Based on the characteristics of the program, machine learning algorithms can predict the optimal day of the week and time of day to send automatic messages as well as the method of message delivery that maximizes efficacy .

2.6. Conclusion

Predictive modeling and artificial intelligence (AI) have great potential to improve care in a wide range of clinical areas including diagnosis, risk assessment, pharmacotherapy, lifestyle management, home monitoring, and early interventions. Such AI-driven approaches to improving care could have significant impact if applied effectively in the care of common chronic diseases with high morbidity and mortality such as diabetes mellitus. Clinical decision support (CDS) is a technique for helping healthcare providers follow clinical guidelines and providing patient- and population-specialized care. CDS has been long recognized as a promising approach to informing the care of chronic conditions. In the field of medical informatics, a variety of machine learning (ML)-based CDS tools have been explored for providing pharmacotherapy recommendations and predicting the risk of complications, adverse drug events, or hospitalization at an early stage in the context of chronic disease. Despite the great promise, AI-based CDS for improving the care of chronic diseases is still in early stages and only a limited number of studies have been shown to be effective in practical clinical settings. Current research faces several challenges. One of these is in the establishment of the evidence base for AI-based CDS in a retrospective review, and in effectively evaluating gradually-deployed services.

2.6.1. Emerging Technologies

Biotechnology which combines biological and engineering principles is radically transforming the way researchers understand disease in micro and nano scale. Individual biomolecules' groups called biomarkers can be organized with respect to variation in stress, health, and disease. Furthermore, tissues composed of collections of cell types can be analyzed for spatial expression patterns. Understanding in one type of readout can be transferred to another. For a vast number of reasons, the analytical and conceptual toolkit in biotechnology has the power to evolve at a much faster pace than all its counterparts in software engineering, nanotechnology, and mechanics, and signals biochemical and physical transduce n together with rigorous control and statistical approaches are profoundly lacking in machine learning. Engineering reality from biosignatures demands much more than pattern recognition.

In addition to new data and novel algorithms, new hardware in modern biology is disproportionately important for discovery and treatment of chronic diseases. Technologies from next-generation sequencing, gene editing, protein sequencing to single-molecule counting enable a wealth of new biological signals. The way these signals are transduced also brings to the fore a flood of imaging, optoelectronic, and optomechanical outputs that are unalike today's mass spectroscopy and gel electrophoresis derived readouts. The complexity of these acquisitions from a stochastic, geometric, information theoretic, and thermodynamic standpoint has the potential of exceeding the impact of tissue technologies on machine learning in the last few decades. Further, with the integration of imaging and single-cell technologies combined with a systems biology approach, biotechnologies can illustrate cellular causal influence in a temporal manner.

References

Acharya, U. R., et al. (2023). Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. arXiv.

Elenitoba-Johnson, K. S. J., & Brown, N. A. (2020). Enabling precision oncology through precision diagnostics. *Annual Review of Pathology: Mechanisms of Disease*.

Higgins, N., et al. (2023). Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. arXiv.

Li, L., et al. (2023). Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. arXiv.

Tao, X., et al. (2023). Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. arXiv.