A Contemporary Policy to Analyze EMR Records to Forecast Ailments

|  |  |
| --- | --- |
| Navaneeth A V1Department of Master of Computer Applications *Nitte Meenakshi Institute of Technology*Bengaluru, Indiaavnavaneeth25@gmail.com | Vidya Sagar S D2Department of Master of Computer Applications *Nitte Meenakshi Institute of Technology*Bengaluru, Indiavidyasagarsd@gmail.com |
| Shweta Dhareshwar 3Department of Masters of Computer Applications*Nitte Meenakshi Institute of Technology*Bengaluru, Indiashwetashine.dhareshwar@gmail.com | Sreekanth Rallapalli 4Department of Master of Computer Applications*Nitte Meenakshi Institute of Technology*Bengaluru, Indiarsreekanth1@yahoo.com |
| Sashikanth Reddy Avula 5Department of Master of Computer Applications*Nitte Meenakshi Institute of Technology*Bengaluru, Indiaaskr1985@gmail.com | Dileep M R 6Department of Master of Computer Applications*Nitte Meenakshi Institute of Technology*Bengaluru, Indiadileep.kurunimakki@gmail.com |

ABSTRACT

Construction – Diseases those are associated with the way a person or group of people live are known as lifestyle diseases. Healthcare industry collects enormous disease- related data that is unfortunately not mined to discover hidden information that could be used for effective decision making. This study aims to understand, K-means nearest neighbor and use it to predict lifestyle diseases that an individual might be susceptible to. Moreover, it is proposed and simulate an economic machine learning model using EMR data that analyzes an individual’s lifestyle to identify possible threats that form the foundation of diagnostic tests and disease prevention, which may arise due to unhealthy diets and excessive energy intake, physical dormancy, etc. The simulated model will prove to be an intelligent low- cost alternative to detect possible genetic disorders caused by unhealthy lifestyles. The work in the paper majorly concentrates on probable prediction of heart attack occurrence in future based on the previous history.

Keywords— EMR, clustering, K means, PCA, Filteration.

#  INTRODUCTION

 Disease prediction using patient treatment history and health data by applying data mining and machine learning techniques is ongoing struggle for the past decades. Many works have been applied data mining techniques to pathological data or medical profiles for prediction of specific diseases. These approaches tried to predict the reoccurrence of disease. Also, some approaches try to do prediction on control and progression of disease. The recent success of deep learning in disparate areas of machine learning has driven a shift towards machine learning models that can learn rich, hierarchical representations of raw data with little pre processing and produce more accurate results. With the development of big data technology, more attention has been paid to disease prediction from the perspective of big data analysis; various researches have been conducted.

The term diagnosis is used for finding symptoms of disease or analysis of the patients to determine the health conditions. The diagnosis is usually performed through one of these methods, i.e., examining the physical condition of the patient, exploring patient’s history, or from diagnostic tests which are analysed by various healthcare professionals such as dentist, physician, chiropractor, physical therapist, or physician assistant and compounder etc. The patient’s history is frequently saved in the form of a prescription for necessary medications, streamline workflow, and to keep track of the patient’s performance. Initially, the prescription was saved in the form of the paper chart containing the type of diseases, suggested medicines, vaccination dates, treatment plans, and the test results of X-rays specific hospitals. However, in the modern age of the computer, the prescription is saved in a digital format which is known as an electronic medical record (EMR) or electronic health record (EHR).

# Literature Survey

Several research works has been done in the past about estimation of the diseases based on the patient history. The development in the science and technology has provided rich set of facilities to explore new approaches towards healthcare systems. At global and national level tremendous work has been done on healthcare, sanity, livelihood, nutrition and on disease predictions. Typically in prediction systems, if heart attack or cardiac arrest type of cases are considered, then usually previous history of the patient plays a major role, and this data is stored in the electronic form to guess the possible next occurrence of the heart attack like issues with use of various algorithms. Accuracy is the major factor in this kind of research areas.

Ahmed et al [1], has made a survey on EMR/ESD: techniques, complications, and evidence. Arend et al [2], has made a detailed work on EMR 20006-012: A phase II randomized double-blind placebo controlled trial comparing the combination of pimasertib (MEK inhibitor) with SAR245409 (PI3K inhibitor) to pimasertib alone in patients with previously treated unresectable borderline or low grade ovarian cancer. Chakravarthy et al [3], has described Progression from early/intermediate to advanced forms of age-related macular degeneration in a large UK cohort: rates and risk factors. Desai et al [4], has done a detailed survey on Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. Enaizan et al [5], has worked on Electronic medical record systems: Decision support examination framework for individual, security and privacy concerns using multi-perspective analysis. Enaizan et al [6], has done a extensive work on Effects of privacy and security on the acceptance and usage of EMR: the mediating role of trust on the basis of multiple perspectives. Feldman et al [7], has demonstratedTurning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. Imel et al [8], demonstrated Characterizing patients initiating abaloparatide, teriparatide, or denosumab in a real-world setting: a US linked claims and EMR database analysis. Ma et al [9], worked on Distilling Knowledge from Publicly Available Online EMR Data to Emerging Epidemic for Prognosis. Madden et al [10], has been worked on Telehealth uptake into prenatal care and provider attitudes during the COVID-19 pandemic in New York City: a quantitative and qualitative analysis. Miled et al [11], has made extensive work on Predicting dementia with routine care EMR data. Mollart et al [12], has made a detailed literature survey on Introduction of patient electronic medical records (EMR) into undergraduate nursing education. Morkem, et al [13], described Validation of an EMR algorithm to measure the prevalence of ADHD in the Canadian Primary Care Sentinel Surveillance Network. Rayner et al [14], has illustrated the patient journey through the care continuum: Leveraging structured primary care electronic medical record (EMR) data in Ontario, Canada using chronic obstructive pulmonary disease as a case study. Rozenfeld et al [15], has demonstrated a model of disparities: risk factors associated with COVID-19 infection. Sinaga et al [16], demonstrated Unsupervised K-means clustering algorithm. Sun et al [17], has made detailed work on Analysis of PAHs in oily systems using modified QuEChERS with EMR-Lipid clean-up followed by GC-QqQ-MS. Turk et al [18], described Intellectual and developmental disability and COVID-19 case-fatality trends. Yu et al [19], demonstrated the Development of an online health care assessment for preventive medicine: a machine learning approach. Zhang et al [20], has demonstrated Ensuring electronic medical record simulation through better training, modeling, and evaluation.

In this paper the work has been done on the prediction of the heart attack based on EMR data by considering various parameters such as heart rate, systolic blood pressure, diastolic blood pressure, breathing frequency, temperature, hemoglobin, perfusion index, oxygen saturation. The details listed from the above parameters are fed as input to the k-means clustering algorithm and based on the outputs form the algorithms the results are drawn.

**III PROPOSED SYSTEM**

This project, we propose For the disease prediction, we use K-Nearest Neighbor (KNN) and Decision tree (CNN) machine learning algorithm for accurate prediction of disease. For disease prediction required disease symptoms dataset.

In this general disease prediction the living habits of person and checkup information consider for the accurate prediction. The accuracy of general disease prediction by using Decision Tree is 84.5% which is more than KNN algorithm.

After general disease prediction, this system able to gives the risk associated with general disease which is lower risk of general disease or higher.

**ADVANTAGES OF PROPOSED SYSTEM**

In proposed system we are using many machine learning applications is utilized to construct such classifier that can separate the data based on their characteristics. Data set is partitioned into two or more than two classes.

Such classifiers are utilized for medical data investigation and disease prediction. Today machine learning is present everywhere so that without knowing it, one can possibly use it many times a day. CNN uses both the structured and unstructured data of a hospital to do classification.

While other machine learning algorithms only work on structured data and time required for computation is high also they are lazy because they store entire data as a training dataset and uses complex method for calculation.

### **Table 1: Sample Data.**



### **Table 2: Sample Data with Parametric Representation.**





**Figure 1 : Clustered Representation**

All parameters in the table should be mapped with the diagram as shown above in different colours with boundary values and thresholds. All the boundary values and threhodls should be referred from standard medical records.

# IV PROPOSED ALGORITHM

Algorithm: K-means Imposed on EMR data

Step 1 − First, we need to specify the number of clusters, K, need to be generated by this algorithm.

Step 2 − Next, randomly select K data points and assign each data point to a cluster. In simple words, classify the data based on the number of data points.

Step 3 − Now it will compute the cluster centroids.

Step 4 − Next, keep iterating the following until we find optimal centroid which is the assignment of data points to the clusters that are not changing any more

• 4.1 − First, the sum of squared distance between data points and centroids would be computed.

• 4.2 − Now, we have to assign each data point to the cluster that is closer than other cluster (centroid).

• 4.3 − At last compute the centroids for the clusters by taking the average of all data points of that cluster.

K-means follows Expectation-Maximization approach to solve the problem. The Expectation-step is used for assigning the data points to the closest cluster and the Maximization-step is used for computing the centroid of each cluster.

While working with K-means algorithm we need to take care of the following things −

• While working with clustering algorithms including K-Means, it is recommended to standardize the data because such algorithms use distance-based measurement to determine the similarity between data points.

• Due to the iterative nature of K-Means and random initialization of centroids, K-Means may stick in a local optimum and may not converge to global optimum. That is why it is recommended to use different initializations of centroids.



**Figure 2: Imposeing K-means Clustering Algorith on EMR Data.**

The k-means algorithm is a simple iterative method to partition a given dataset into a specified number of clusters, k. This algorithm has been discovered by several researchers across different disciplines. The algorithm operates on a set of d- dimensional vectors, D = {xi | i = 1, . . . , N}, where xi ∈ Rd denotes the ith data point. The algorithm is initialized by picking k points in Rd as the initial k cluster. Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times.

# V RESULTS

As of my last update in September 2021, there have been several applications of k-means clustering on Electronic Medical Record (EMR) data. EMRs contain a vast amount of patient information, and clustering techniques like k-means can help uncover patterns and group patients with similar characteristics. Here are some of the common applications: Patient Segmentation: K-means clustering can be used to segment patients into distinct groups based on their medical histories, demographics, or specific health conditions. This information can be valuable for personalized healthcare, targeted interventions, and improved patient outcomes.Disease Pattern Identification: By applying k-means clustering to EMR data, healthcare providers and researchers can identify patterns associated with specific diseases. This can lead to better understanding, early diagnosis, and effective treatments for various medical conditions.Healthcare Resource Utilization: K-means clustering can help identify patient groups with similar healthcare resource utilization patterns. This information can be used for optimizing resource allocation and improving hospital management.Medication Response Analysis: Clustering patients based on their response to certain medications can provide insights into drug effectiveness and possible side effects. It can help identify subgroups of patients who might benefit more from particular treatments.Risk Stratification: K-means clustering can aid in categorizing patients into low, medium, and high-risk groups for specific health outcomes. This information can be used for preventive care and intervention strategies. Chronic Disease Management: Applying k-means clustering to EMR data can facilitate the development of targeted chronic disease management plans. By identifying patient groups with similar risk profiles, healthcare providers can create personalized care strategies. Patient Readmission Prediction: K-means clustering can assist in identifying factors that lead to higher readmission rates for certain patient groups. This information can be used to design interventions aimed at reducing readmission rates and improving patient care.It's important to note that the results and applications of k-means clustering on EMR data may vary depending on the quality and size of the data, the choice of features, and the specific research questions being addressed. Additionally, advances in machine learning and data science since my last update may have led to further developments and refined approaches in this field.

# VI. CONCLUSION

A lot of work has been done for automatic extraction of useful information from electronic health records, clinical notes and discharge summaries. The physician uses features extracted information as an input for the automatic diagnosis of disease. Knowledge bases initially performed this extraction. In the present era, various rule-based learning, machine learning, and deep learning concepts are used for the extraction and diagnosis of diseases. There are several challenges associated with this automatic extraction including missing values, incomplete information, and data abundance. We have reviewed recent research for the automatic diagnosis of various diseases from electronic medical records. We categorized our work into three classes 1) Rule-Based Methods, 2) Machine Learning Methods, and 3) Deep Learning Methods. These categories are further divided into subcategories based on the proposed algorithm. In this review, we tried to cover almost all the latest and existing research of automatic diagnosis from electronic records. We presented the benefits, limitations, and future directions of various data-driven methods, dataset employed and focused disease. Moreover, we tried to establish a professional structure to familiarize with an up-to-date automatic disease diagnosis technique.

REFERENCES

1. Ahmed, Yahya, and Mohamed Othman. "EMR/ESD: techniques, complications, and evidence." Current Gastroenterology Reports 22, no. 8 (2020): 1-12.
2. Arend, Rebecca C., Allison M. Davis, Przemyslaw Chimiczewski, David M. O'Malley, Diane Provencher, Ignace Vergote, Sharad Ghamande, and Michael J. Birrer. "EMR 20006-012: A phase II randomized double-blind placebo controlled trial comparing the combination of pimasertib (MEK inhibitor) with SAR245409 (PI3K inhibitor) to pimasertib alone in patients with previously treated unresectable borderline or low grade ovarian cancer." Gynecologic oncology 156, no. 2 (2020): 301-307.
3. Chakravarthy, Usha, Clare C. Bailey, Peter H. Scanlon, Martin McKibbin, Rehna S. Khan, Sajjad Mahmood, Louise Downey et al. "Progression from early/intermediate to advanced forms of age-related macular degeneration in a large UK cohort: rates and risk factors." Ophthalmology Retina 4, no. 7 (2020): 662-672.
4. Desai, Rishi J., Shirley V. Wang, Muthiah Vaduganathan, Thomas Evers, and Sebastian Schneeweiss. "Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes." JAMA network open 3, no. 1 (2020): e1918962-e1918962.
5. Enaizan, Odai, Ahmad A. Zaidan, NH M. Alwi, Bulat B. Zaidan, Mohammed Assim Alsalem, O. S. Albahri, and A. S. Albahri. "Electronic medical record systems: Decision support examination framework for individual, security and privacy concerns using multi-perspective analysis." Health and Technology 10, no. 3 (2020): 795-822.
6. Enaizan, Odai, Bilal Eneizan, Mohammad Almaaitah, Ahmad Tawfig Al-Radaideh, and Ashraf Mousa Saleh. "Effects of privacy and security on the acceptance and usage of EMR: the mediating role of trust on the basis of multiple perspectives." Informatics in Medicine Unlocked 21 (2020): 100450.
7. Feldman, Dan, Melanie Schmidt, and Christian Sohler. "Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering." SIAM Journal on Computing 49, no. 3 (2020): 601-657.
8. Imel, E. A., K. Starzyk, R. Gliklich, R. J. Weiss, Y. Wang, and S. A. Williams. "Characterizing patients initiating abaloparatide, teriparatide, or denosumab in a real-world setting: a US linked claims and EMR database analysis." Osteoporosis International 31, no. 12 (2020): 2413-2424.
9. Ma, Liantao, Xinyu Ma, Junyi Gao, Xianfeng Jiao, Zhihao Yu, Chaohe Zhang, Wenjie Ruan, Yasha Wang, Wen Tang, and Jiangtao Wang. "Distilling Knowledge from Publicly Available Online EMR Data to Emerging Epidemic for Prognosis." In Proceedings of the Web Conference 2021, pp. 3558-3568. 2021.
10. Madden, Nigel, Ukachi N. Emeruwa, Alexander M. Friedman, Janice J. Aubey, Aleha Aziz, Caitlin D. Baptiste, Jaclyn M. Coletta et al. "Telehealth uptake into prenatal care and provider attitudes during the COVID-19 pandemic in New York City: a quantitative and qualitative analysis." American journal of perinatology 37, no. 10 (2020): 1005-1014.
11. Miled, Zina Ben, Kyle Haas, Christopher M. Black, Rezaul Karim Khandker, Vasu Chandrasekaran, Richard Lipton, and Malaz A. Boustani. "Predicting dementia with routine care EMR data." Artificial intelligence in medicine 102 (2020): 101771.
12. Mollart, Lyndall, Rachel Newell, Sara K. Geale, Danielle Noble, Carol Norton, and Anthony P. O'brien. "Introduction of patient electronic medical records (EMR) into undergraduate nursing education: an integrated literature review." Nurse education today (2020): 104517.
13. Morkem, Rachael, Kenneth Handelman, John A. Queenan, Richard Birtwhistle, and David Barber. "Validation of an EMR algorithm to measure the prevalence of ADHD in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN)." BMC medical informatics and decision making 20, no. 1 (2020): 1-8.
14. Rayner, Jennifer, Tanya Khan, Carmen Chan, and Chen Wu. "Illustrating the patient journey through the care continuum: Leveraging structured primary care electronic medical record (EMR) data in Ontario, Canada using chronic obstructive pulmonary disease as a case study." International Journal of Medical Informatics 140 (2020): 104159.
15. Rozenfeld, Yelena, Jennifer Beam, Haley Maier, Whitney Haggerson, Karen Boudreau, Jamie Carlson, and Rhonda Medows. "A model of disparities: risk factors associated with COVID-19 infection." International journal for equity in health 19, no. 1 (2020): 1-10.
16. Sinaga, Kristina P., and Miin-Shen Yang. "Unsupervised K-means clustering algorithm." IEEE Access 8 (2020): 80716-80727.
17. Sun, Yaqing, and Shimin Wu. "Analysis of PAHs in oily systems using modified QuEChERS with EMR-Lipid clean-up followed by GC-QqQ-MS." Food Control 109 (2020): 106950.
18. Turk, Margaret A., Scott D. Landes, Margaret K. Formica, and Katherine D. Goss. "Intellectual and developmental disability and COVID-19 case-fatality trends: TriNetX analysis." Disability and Health Journal 13, no. 3 (2020): 100942.
19. Yu, Cheng-Sheng, Yu-Jiun Lin, Chang-Hsien Lin, Shiyng-Yu Lin, Jenny L. Wu, and Shy-Shin Chang. "Development of an online health care assessment for preventive medicine: a machine learning approach." Journal of medical Internet research 22, no. 6 (2020): e18585.
20. Zhang, Ziqi, Chao Yan, Diego A. Mesa, Jimeng Sun, and Bradley A. Malin. "Ensuring electronic medical record simulation through better training, modeling, and evaluation." Journal of the American Medical Informatics Association 27, no. 1 (2020): 99-108.