A CONTEMPORARY POLICY TO ANALYZE EMR RECORDS TO FORECAST AILMENTS

Abstract

Construction Lifestyle-related illnesses refer to medical conditions that are linked to the habits and choices of individuals or groups. Despite the vast collection of disease-related information in the healthcare sector, there is an untapped potential to extract concealed insights that could enhance informed decision-making. This research seeks to explore and apply the K-means nearest neighbor technique to predict the likelihood of individuals developing lifestyle-related ailments. Additionally, it aims to propose and simulate an economically viable machine learning model that utilizes Electronic Medical Record (EMR) data. This model would analyze an individual's way of life to detect potential risks, which serve as the basis for devising preventive measures and diagnostic tests. These risks often stem from factors such as unhealthy eating patterns, excessive calorie consumption, sedentary behavior, and so on. By creating this simulated model, an intelligent and cost-effective solution is presented for identifying potential hereditary disorders stemming from unhealthy lifestyles. The central focus of the paper revolves predominantly around forecasting the potential occurrence of heart attacks in the future, drawing insights from past medical histories.

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

Authors

Navaneeth A. V
Assistant Professor
Department of Masters of Computer Applications
Nitte Meenakshi Institute of Technology
Bengaluru, Karnataka, India.
avnavaneeth25@gmail.com

Vidya Sagar S. D
Assistant Professor
Department of Masters of Computer Applications
Nitte Meenakshi Institute of Technology
Bengaluru, India
vidyasagarsd@gmail.com

Dileep M. R
Assistant Professor
Department of Masters of Computer Applications
Nitte Meenakshi Institute of Technology
Bengaluru, India
dileep.kurunimakki@gmail.com

Sashikanth Reddy Avula
Assistant Professor,
Department of Masters of Computer Applications
Nitte Meenakshi Institute of Technology
Bengaluru, India
askr1985@gmail.com

Sreekanth Rallapalli
Assistant Professor,
Department of Masters of Computer Applications
Nitte Meenakshi Institute of Technology
Bengaluru, India
sreekanth1@tyahoo.com

Copyright © 2024 Authors

Shwetha Dhareshwar
Assistant Professor
Department of Masters of Computer
Applications
Nitte Meenakshi Institute of Technology
Bengaluru, India
shwetashine.dhareshwar@gmail.com

Copyright © 2024 Authors

I. INTRODUCTION

For several spans, there has remained a continuous effort to forecast diseases through the utilization of patient diagnosis antiquities and health-related information, employing data mining and machine learning procedures. Abundant endeavors have pragmatic data mining practices to compulsive data or health summaries with the goal of forecasting specific illnesses. These methods have aimed to predict the recurrence of ailments, and some have even focused on anticipating the progression and control of diseases. The recent breakthroughs in deep learning across diverse domains of machine learning have prompted a swing concerning employing machine learning mockups capable of comprehending intricate, layered representations of unprocessed data, thereby yielding more precise outcomes with minimal preprocessing requirements. As the field of big data technology advances, increased emphasis has remained placed on illness forecast after the vantage point of extensive data analysis. This shift has led to a multitude of research endeavors dedicated to exploring and unraveling disease prediction strategies within the realm of big data analysis.

The term "diagnosis" pertains to the identification of disease symptoms or the analysis of a patient's condition to ascertain their state of health. Diagnosis is typically accomplished through one of these methodologies: observation of the patient's physical state, exploration of the patient's remedial antiquity, or through indicative assessments that are evaluated by numerous healthcare practitioners such as physicians, dentists, chiropractors, physical therapists, physician assistants, and pharmacists, among others. Patient histories are often preserved in the form of prescriptions, which serve several purposes including administering necessary medications, streamlining healthcare workflows, and monitoring the patient's progress. Initially, these prescriptions were documented using paper charts that detailed the nature of the ailments, recommended medications, vaccination schedules, treatment strategies, and the outcomes of X-ray tests, particularly in specific healthcare facilities. However, in the contemporary era of computing, prescriptions are now stored in a numeral arrangement referred to as an electronic health record (EHR) or electronic medical record (EMR).

II. 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] conducted an investigation focusing on records of medical health and health records electronically, exploring techniques, complications, and supporting evidence. Arend et al [2] conducted an in-depth study titled "EMR 20006-012," a phase II randomized trial that compared the effectiveness of pimasertib in combination with SAR245409 against pimasertib unaided in patients through beforehand preserved unresectable disputed or inferior ovarian cancer. Chakravarthy et al [3] provided insights into the progression of age-related

Copyright © 2024 Authors

macular deterioration, delving into the transition after initial/middle to progressive procedures inside a substantial UK unit, elucidating taxes then danger issues. In a comprehensive survey, Desai et al [4] compared traditional models and machine learning methods for predicting heart failure outcomes by amalgamating administrative claims with electronic medical records. Enaizan et al [5] investigated electric medicinal best schemes, offering a outline for decision support while addressing separate, safety, and confidentiality anxieties through a multi-perspective investigation. In another study, Enaizan et al [6] extensively explored the impact of privacy and safety on the receipt and practice of EMRs, highlighting the arbitrating part of belief across various viewpoints. Feldman et al [7] demonstrated techniques for transforming substantial datasets into more manageable forms while focusing on k-means, PCA, and projective clustering. Imel et al [8] characterized patients initiating treatment with abaloparatide, teriparatide, or denosumab within a real-world setting through analysis of linked claims and EMR databases. Ma et al [9] distilled insights from publicly available online EMR data to forecast emerging epidemics for prognosis. Madden et al [10] explored the integration of telehealth into prenatal care and examined provider attitudes during the COVID-19 pandemic in New York City, using both quantitative and qualitative analyses. Miled et al [11] dedicated significant effort to predicting dementia utilizing routine care EMR data. Mollart et al [12] conducted a comprehensive literature review centered on the incorporation of patient electronic medical records into undergraduate nursing education. Morkem et al [13] validated an EMR algorithm to gauge the prevalence of ADHD within the Canadian Primary Care Sentinel Surveillance Network. Rayner et al [14] depicted the patient trip finished the upkeep range, leveraging organized main upkeep EMR information to study chronic obstructive pulmonary disease in Ontario, Canada. Rozenfeld et al [15] established a model for discerning disparities, identifying COVID-19 infection risk factors. Sinaga et al [16] demonstrated an unsupervised K-means clustering algorithm. Sun et al [17] meticulously analyzed the analysis of polycyclic aromatic hydrocarbons (PAHs) in oily systems through modified QuEChERS with EMR-lipid cleanup followed by GC-QqQ-MS. Turk et al [18] delved into the trends of case fatality associated with intellectual and developmental disabilities in the context of COVID-19. Innovatively, Yu et al [19] developed an online healthcare assessment for preventive medicine using a machine learning approach. Zhang et al [20] highlighted the importance of ensuring the accuracy of electronic medical record simulations through enhanced training, modeling, and evaluation techniques.

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 proportion, systolic blood pressure, diastolic gore heaviness, breathing frequency, temperature, hemoglobin, perfusion index, oxygen saturation. The details listed from the above parameters are fed as input to the k-means clustering procedure and grounded on the outputs form the algorithms the results are drawn.

III. PROPOSED SYSTEM

In this endeavor, we present a project aimed at disease prediction. To achieve this goal, we employ the K-Nearest Neighbor (KNN) and Decision Tree (CNN) machine learning algorithms, offering a robust and precise disease prediction mechanism. The crux of our approach revolves around utilizing a dataset encompassing disease symptoms. Within this comprehensive disease prediction framework, we factor in an individual's lifestyle habits and medical examination details to enhance the accuracy of our predictions.

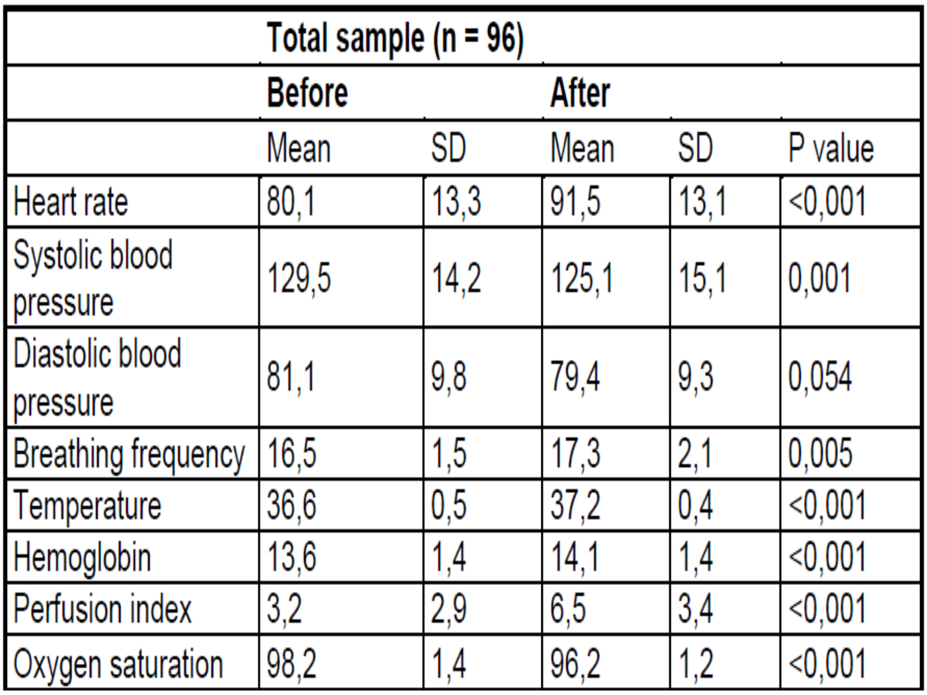
By harnessing the power of Decision Tree algorithm, our model attains a notable accuracy rate of 84.5%, surpassing the performance of the KNN algorithm. Beyond the realm of general disease prediction, our system extends its capabilities to ascertain the associated risk level linked with these predicted diseases. This assessment essentially distinguishes between lower and higher risks of contracting these general diseases, thus contributing to a more informed and nuanced understanding of an individual's health status.

Advantages of Proposed System:

In the proposed system, we employ a range of machine learning applications to create classifiers capable of segregating data based on distinct attributes. The dataset is divided into multiple classes, typically two or more. These classifiers find significant application in analyzing medical data and forecasting diseases. In the contemporary landscape, machine learning has permeated various facets of our lives, often without our conscious recognition, and is frequently employed multiple times daily.

Convolutional Neural Networks (CNN) serve as a prime example, leveraging both structured and unstructured data from hospital records to execute classification tasks. Unlike alternative machine learning algorithms that solely function with structured data, CNN can process various types of data. However, it's worth noting that many other algorithms necessitate substantial computation time, are resource-intensive due to storing entire datasets for training, and adopt intricate methodologies for calculations..

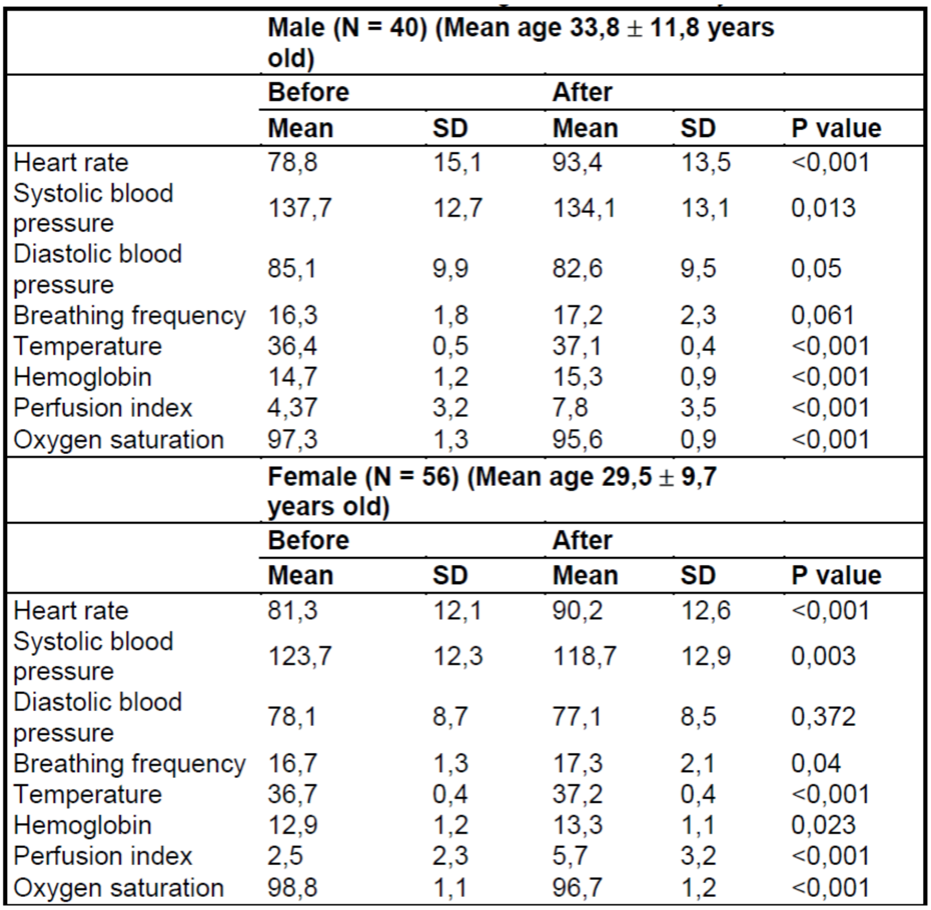
Table 1: Sample Data.



	Total sample (n = 96)				
	Before		After		
	Mean	SD	Mean	SD	P value
Heart rate	80,1	13,3	91,5	13,1	<0,001
Systolic blood pressure	129,5	14,2	125,1	15,1	0,001
Diastolic blood pressure	81,1	9,8	79,4	9,3	0,054
Breathing frequency	16,5	1,5	17,3	2,1	0,005
Temperature	36,6	0,5	37,2	0,4	<0,001
Hemoglobin	13,6	1,4	14,1	1,4	<0,001
Perfusion index	3,2	2,9	6,5	3,4	<0,001
Oxygen saturation	98,2	1,4	96,2	1,2	<0,001

Copyright © 2024 Authors

Table 2: Sample Data with Parametric Representation.



	Male (N = 40) (Mean age 33,8 ± 11,8 years old)				
	Before		After		
	Mean	SD	Mean	SD	P value
Heart rate	78,8	15,1	93,4	13,5	<0,001
Systolic blood pressure	137,7	12,7	134,1	13,1	0,013
Diastolic blood pressure	85,1	9,9	82,6	9,5	0,05
Breathing frequency	16,3	1,8	17,2	2,3	0,061
Temperature	36,4	0,5	37,1	0,4	<0,001
Hemoglobin	14,7	1,2	15,3	0,9	<0,001
Perfusion index	4,37	3,2	7,8	3,5	<0,001
Oxygen saturation	97,3	1,3	95,6	0,9	<0,001
	Female (N = 56) (Mean age 29,5 ± 9,7 years old)				
	Before		After		
	Mean	SD	Mean	SD	P value
Heart rate	81,3	12,1	90,2	12,6	<0,001
Systolic blood pressure	123,7	12,3	118,7	12,9	0,003
Diastolic blood pressure	78,1	8,7	77,1	8,5	0,372
Breathing frequency	16,7	1,3	17,3	2,1	0,04
Temperature	36,7	0,4	37,2	0,4	<0,001
Hemoglobin	12,9	1,2	13,3	1,1	0,023
Perfusion index	2,5	2,3	5,7	3,2	<0,001
Oxygen saturation	98,8	1,1	96,7	1,2	<0,001

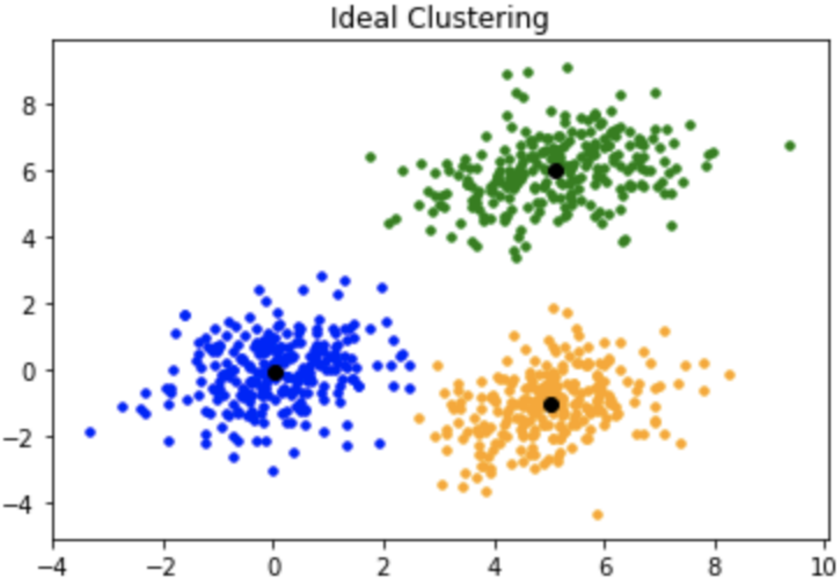


Figure 1: Clustered Representation

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

K means clustering is a prevalent unsupervised, machine learning algorithm used for federation alike information points hooked on bunches. Here's a step-by-step breakdown of how the K-means procedure mechanisms:

1. **Initialising:** Select the amount of groups, K, required to divide your data into.
   Randomly select K data points from your dataset as initial cluster centroids.

2. **Assignment:** Aimed at respectively information argument in your dataset, compute the space to individually of the K centroids.
   Assign the data point to the cluster whose centroid is closest (usually using Euclidean distance).

3. **Update Centroids:** After all data points are assigned to clusters, estimate the despicable of altogether information points in each cluster.
   Set these mean values as the new centroids of their respective clusters.

4. **Repeat Assignment and Update:** Repeat the assignment step (step 2) using the updated centroids.Repeat the centroid update step (step 3) with the newly assigned data points.

5. **Convergence:** Continue iterating between assignment and centroid update till a ending standard is encountered. This principle might be a supreme amount of repetitions or till centroids no lengthier alteration meaningfully.

6. **Final Clusters:** As soon as the procedure touches, the information opinions are alienated hooked on bunches founded on their resemblance to centroids.

   The K-means algorithm goals to minimalize the amount of sharpened detachments among information opinions and their assigned centroids. This process usually leads to points within the same cluster being closer to each other than to opinions in additional bunches.

   However, it's significant to memo that K-means gathering is subtle to the original assignment of centroids. Dissimilar initializations strength lead to different final cluster assignments. To mitigate this, you can achieve manifold innings of the procedure through dissimilar initializations and choice the answer through the lowermost total squared coldness.

   Also, the special of the quantity of bunches (K) is vital and might require area information or techniques like the prod technique to control the optimum importance.

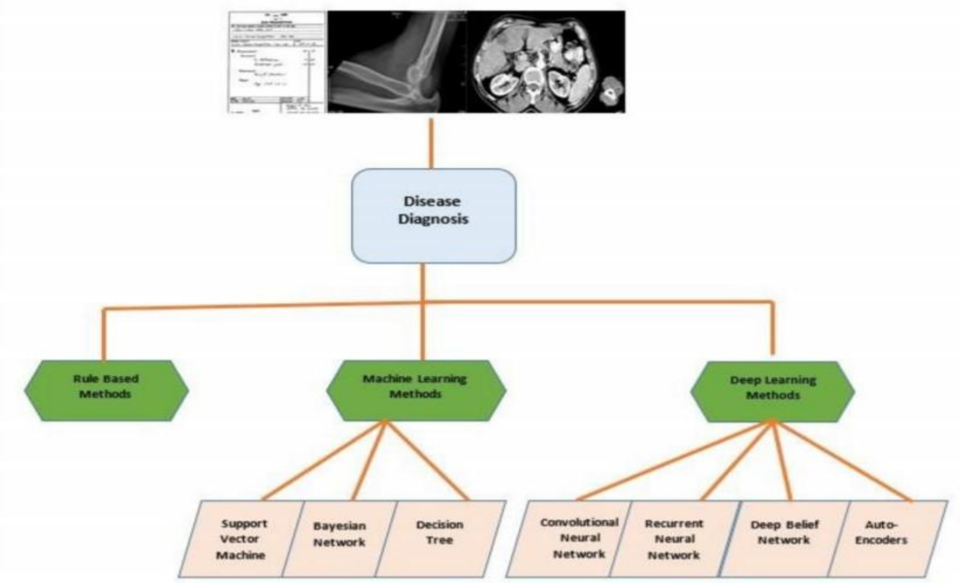


Figure 2: Imposeing K-means Clustering Algorith on EMR Data**.**

The k means procedure is a straightforward iterative technique designed to divider a assumed dataset hooked on a predefined amount of bunches, denoted as 'k.' This procedure has been independently developed by researchers across various fields. It operates on a collection of d-dimensional vectors, represented as D = {xi | i = 1, . . . , N}, where each xi ∈ Rd signifies the individual information opinion. The procedure's initialization phase involves selecting k points in Rd to serve as the initial cluster centroids. Various methods can be employed to choose these starting points. Approaches include random sampling from the dataset, designating them as the outcome of gathering a trivial information subsection, or perturbing the overall 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 Division: K-means gathering container be rummage-sale to section 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

Copyright © 2024 Authors

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 info can be rummage-sale to design interferences meant at plummeting readmission rates and improving patient care. It's significant 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

Extensive efforts have been dedicated to the automated extraction of valuable insights from automated well-being archives, scientific minutes, and release precise. Physicians utilize the extracted features as inputs for automated disease diagnosis. Initially, knowledge bases spearheaded this extraction endeavor. However, in the contemporary landscape, a diverse array of methodologies encompassing rule-based learning, machine learning, and deep learning concepts have taken the reins in the extraction and disease diagnosis domain. This pursuit of automatic extraction is confronted with a range of challenges, encompassing missing data values, incomplete information, and the sheer abundance of data. In our comprehensive review, we have delved into recent research endeavors focusing on the automatic diagnosis of various diseases using electronic medical records. We have meticulously classified our study into three distinct categories: 1) Rule-Based Approaches, 2) Machine Learning Methods, and 3) Deep Learning Methods. These categories have further been segmented based on the specific algorithms proposed. Within this review, our intent was to encompass both contemporary and existing research, thereby offering a comprehensive exploration of automatic disease diagnosis through electronic records. We have presented not only the advantages and drawbacks associated with different data-driven techniques but also laid out potential pathways for future development. Our review encapsulates the datasets employed and the specific diseases under scrutiny. In essence, we have endeavored to establish a structured framework that acquaints readers with the cutting-edge landscape of automatic disease diagnosis methods.

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

Copyright © 2024 Authors

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.

Copyright © 2024 Authors