IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EVALUATION AND MAPPING OF DATA SCIENCE METHODS TO SOLVE HEALTH CONDITIONS OF WOMEN

Sugavasi Nomitha, Student, Computer Science and Technology, Narayanamma Institute of Technology and Science(For Women), Attapur, Hyderabad

Raaga Pravija Gaddam, Student, Computer Science and Technology, Narayanamma Institute of Technology and Science (For Women), Bachupally, Hyderabad

ABSTRACT

Data science is an emerging subject that integrates traditional scientific research with data mining, machine learning, and massive datasets. The healthcare sector produces copious amounts of valuable data on patient demographics, treatment plans, diagnostic outcomes, financial coverage, and more. The healthcare business generates a great deal of data, much of it unstructured and in need of processing, management, analysis, and integration. Results can only be trusted if the data has been managed and analysed effectively. This article explores into healthcare-specific applications of data cleansing, data mining, data preparation, and data analysis. This article offers a glance into the present and future prospects of big data analytics in healthcare by outlining the advantages, describing the frameworks and technique used, analysing the current obstacles encountered, and proposing potential solutions. The primary goal of this work is to evaluate and map data science approaches used to improve women's health, as well as to investigate the nature and efficacy of the issues that have been addressed. Data science techniques, such as text analytics, science mapping, and descriptive assessment, are frequently employed in investigations into women's health. The application of data science techniques to women's health issues is an exciting new field of study since it has the potential to be both effective and cost-efficient. Women bear a disproportionate share of the health consequences of the current COVID-19 pandemic. More needs to be done by policymakers to ensure women have equal access to health care.

Key words: Data science, women, Healthcare, machine learning, algorithms, big data.

1. INTRODUCTION

Novel data-related applications have been developed as a result of today's medical and technological advances [1]. There is a tremendous opportunity for analysis and study using cutting-edge technologies on clinical data generated by the health care industry, such as patient electronic health records (EHR), prescriptions, clinical reports, information about the purchase of medicines, data related to medical insurance, data on investigations, and laboratory reports [2]. The huge data sets can be efficiently analysed using machine-learning techniques. Improved decision-making that benefits patients is the end consequence of attentive data analysis and pattern recognition.

Improved health care, greater life expectancy, earlier disease diagnosis, and more cost-effective treatment of disease are all possible thanks to its ability to shed light on underlying patterns [3]. Health information exchange (HIE) can be set up to consolidate a patient's medical history from multiple sources into a single, easily accessible database that can be accessed by their numerous doctors. Therefore, healthcare organisations should strive to collect all the resources they need to capitalise on big data, which can increase income and profits, stimulate the development of more efficient healthcare networks, and give a variety of additional benefits [4, 5]. In the next decade, data mining methods will pave the way for a new kind of healthcare system that is knowledge-rich and evidence-based.

Smartphone apps that can monitor personal health indicators using sensors and analyzers [6, 7] have increased the importance of big data and its usefulness in healthcare and the medical sciences. Data mining's end game is better customer service, and it does this by analysing massive amounts of user information.

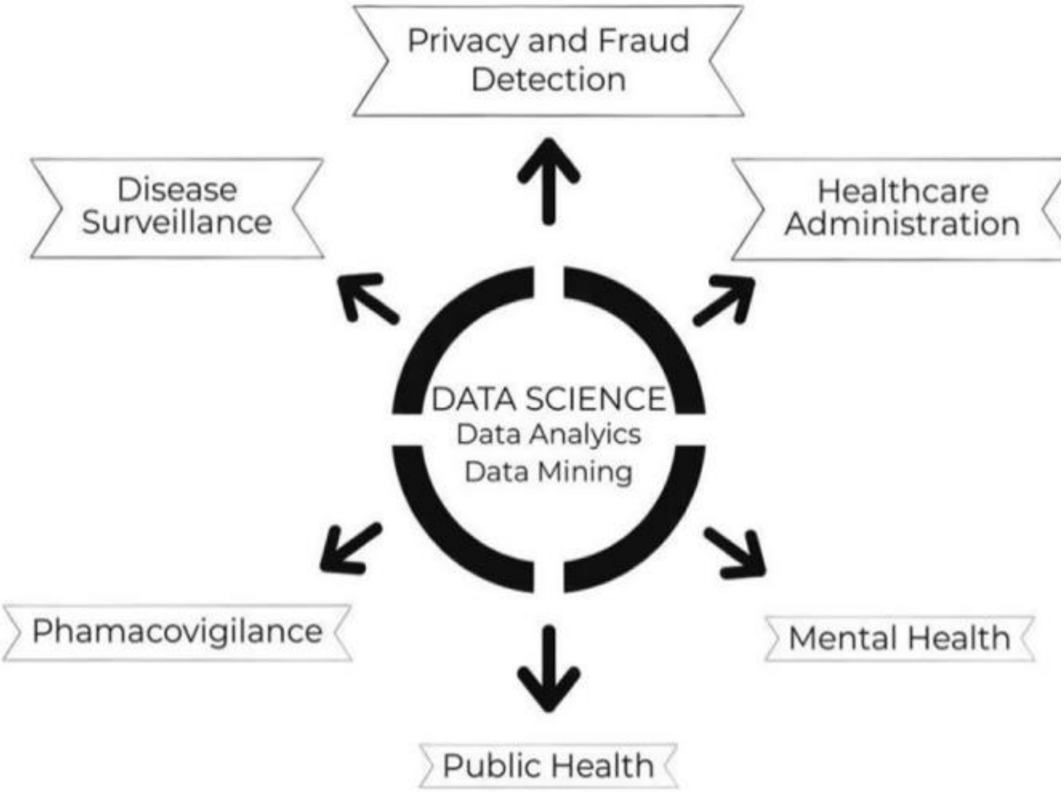


Fig. 1Data science has several uses in the medical field.

Six of the various healthcare analytics applications are shown in Fig. 1 below: Pharmacovigilance, Public Health, Mental Health, Data Privacy, Healthcare Management, and Administration, and Detection of Fraud. Among the many applications of data extraction in research are cloud computing and data deposition, improved quality control, lower costs, more resource utilisation, and patient management.

The use of data science (DS) tools and methodology, as well as big data analytics approaches, has increased in the healthcare and biological sciences as a result of recent advancements in computer processing capacity and the availability of enormous and complicated data.[8] Complementing visual analytics and information visualisation, recent developments in DS and analytics techniques like machine learning, deep learning, and artificial intelligence improve the discovery, sensemaking, understanding, and supply of solutions to complex medical and health problems.[9] Clinical decision-making support systems and health outcomes have recently been shown to benefit from data science tools and approaches. With this goal in mind, we set out to assess the efficacy of cutting-edge algorithms and data science methods in the field of women's health research. We also keep track of the challenges that have been solved, the approaches that have been taken to those difficulties, the outcomes of those approaches, and the areas in which further development is required.

2. LITERATURE REVIEW

It can be thought of as a family saga, with the doctor playing a pivotal role [10] in documenting the onset and resolution of illness in a patient's body through the medium of clinical case records. Digitising medical tests, health records, and investigations is commonplace in the current day due to the availability of electronic technologies and the capabilities of these systems. The term "Electronic Health Records" was first used in 2003 by the Institute of Medicine of the National Academies of Sciences, Engineering, and Medicine to describe the compilation and arrangement of patient medical records. This term, which originally referred to a computerised repository for patient information, was coined by the Institute of Medicine. Vital data and other information about a patient's physical and mental health are included in electronic health records (EHRs), which are digital versions of the patient's medical files. These records are kept in an electronic system that can also be used to communicate between the patient and the medical professionals providing care.

Segmentation, enhancement, and noise reduction are only few of the common uses for the wavelets technology in the field of image processing. Screening, diagnosis, and prognosis are just a few areas of medicine that stand to benefit from the incorporation of AI into the image-processing step. In addition, combining medical images with other forms of data and genomic data would boost diagnostic precision and allow for earlier illness identification [11]. The meteoric rise in the number of healthcare facilities and individuals requiring medical attention has resulted in an improvement in the diagnostic and decision-making capabilities of computerised healthcare settings.

Large amounts of data are produced as a result of telemetry as well as other monitoring devices for physiological indicators. Since the data that are produced are typically only kept for a limited amount of time, in-depth research into the data that is produced is typically overlooked. However, recent developments in data science have been applied to the field of healthcare in an effort to improve data management and the quality of treatment that is provided to patients [12].

Numerous online consultation websites, on the other hand, have trouble luring customers who are willing to pay for their services and keep them operational. Furthermore, health care professionals on these websites face the additional issue of distinguishing themselves from a huge number of other doctors who are able to give services that are comparable [13]. In order to (1) distinguish features associated with actual patient payments from free consultations, (2) investigate the importance and significance of these features, and (3) comprehend the linear or nonlinear relationships between these features and actual patient payments, the authors of [13] used ML methods to sift through mountains of service data.

Data science can aid in the processing, management, analysis, and assimilation of healthcare systems' massive amounts of disparate, organised, and unstructured data. Accurate findings can only be obtained by careful management and analysis of this information. In this article, we will examine the applications of data cleansing, data mining, data preparation, and data analysis in healthcare software programmes. This article explores the current state and prospective development of medical big data analytics. Additionally, it summarises the problems that are currently being addressed and provides an analysis of possible remedies while emphasising the advantages. Data science and big data analytics may provide useful information that might aid policymakers in making informed choices for the healthcare system. It helps paint a fuller picture of doctors, patients, and customers. An increase in the quality of medical care is possible because to new possibilities made possible by data-informed decision making [14].

The purpose of this study is to compile and organise all the existing literature on the topic of improving data analytics technology for the purpose of illness prevention and treatment [15]. The review will begin with a discussion of the difficulties of disease prevention before moving on to more traditional methods. This article provides a high-level overview of recent developments in data analytics algorithms used for disease classification, clustering (unusually high incidence of a particular disease), anomalies detection (detection of disease), and disease association, along with a discussion of the benefits and drawbacks of each approach and some suggestions for making the best preference. Then, we'll talk about some of the newer, more successful approaches to illness prevention. The paper comes to a close with some recommendations as well as open research challenges.

3. DATA SCIENCE TECHNIQUES AND ALGORITHMS

3.1 Algorithms and methods

Data analytics in healthcare encompasses many subfields, including but not limited to: Machine learning, analytics, big data, pattern recognition, and data mining. The medical community is continually interested in the development of more intelligent methods for controlling and preventing disease. The primary objective of illness prevention through data analytics is to use actual patient data to assist lessen the likelihood of sickness. In light of recent developments in machine learning and data analytics for processing complex data, prospects exist for efficient and cost-effective preventative measures capable of managing truly massive data sets. In order to discover useful information in massive data sets, data miners might use a variety of tools and methodologies. This section is broken down into four subheadings according to data mining's use in healthcare, specifically its ability to categorise diseases according to their symptoms. The purpose of this section is not to provide extensive technical details, but rather to highlight the fundamental ideas behind and differences between various algorithms.

- Data, such as medical records, must be classified into distinct categories before analysis can begin.
- In order to make decisions, such as discovering patterns in EHR or predicting readmission, clustering is utilised to organise and separate data.
- Through the use of associative analysis, previously unknown connections between medical data (such as illness prevalence) might be uncovered.
- Detecting anomalies that don't fit any known mould is called anomaly detection.

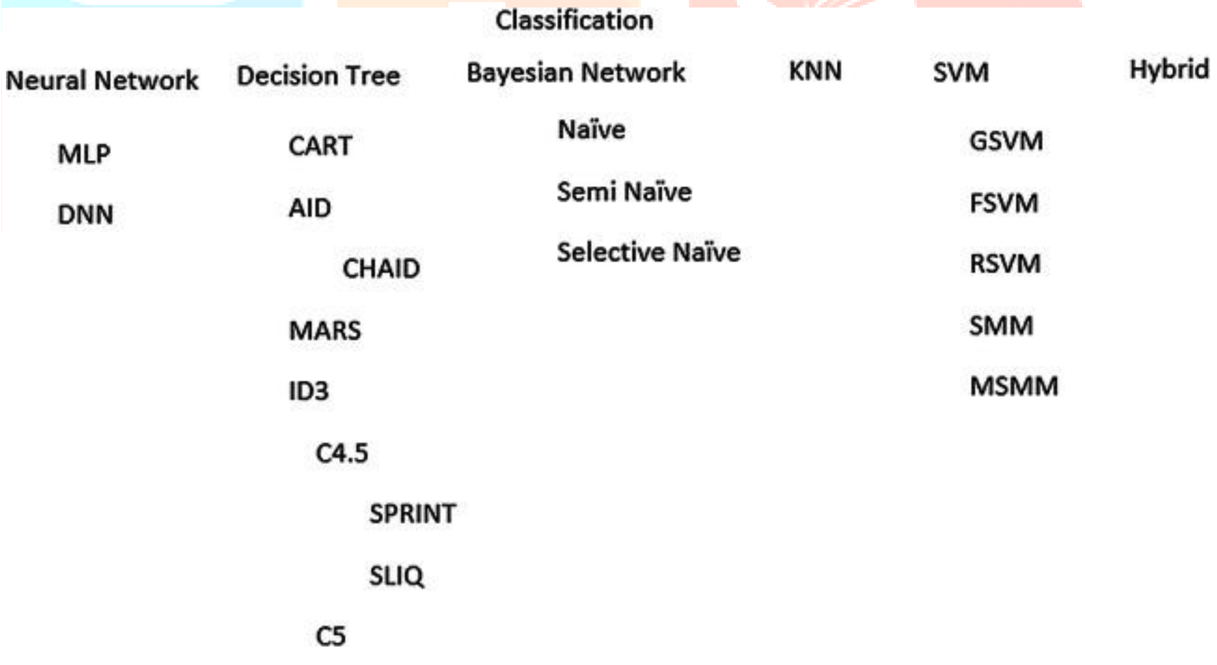


Fig. 2 Classification methodologies

3.2 Anomaly detection

There has been extensive study in the field of anomaly detection because of the practical and potentially lifesaving information it provides in a wide range of settings, from the early detection of an epidemic to the use of electrocardiogram (ECG) signals and other body sensors in patient monitoring. The lack of labelled data and domain-specific normality makes it difficult to design a generic framework for anomaly identification; hence, most anomaly detection approaches are domain-specific. This is a major problem that affects the efficiency of any data mining system. Due to its detrimental effects on data quality and machine learning performance, its detection is crucial. Clearly, this is a matter of the utmost urgency, and it requires careful consideration of every last aspect. Anomaly detection refers to the act of seeking out and detecting data points that do not follow a normal distribution. Point anomalies occur when a single data point deviates from the norm, contextual anomalies occur when a single data point deviates from the norm in a specific context, and collective anomalies occur when a group of data points deviates from the norm.The simplest unsupervised method for discovering outliers in data is clustering (see Fig. 3).

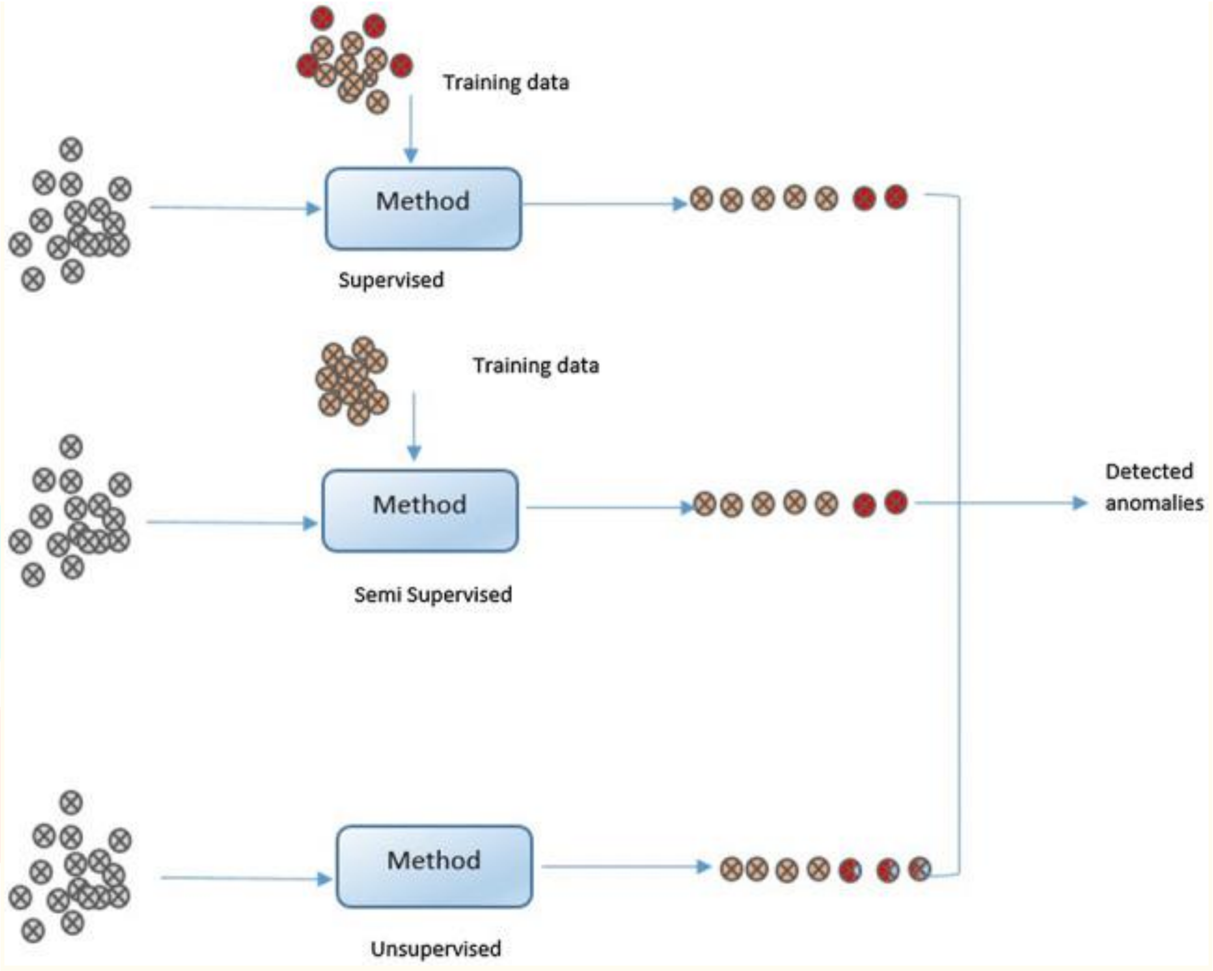


Fig. 3Methods for detecting anomalies in large datasets

The dataset and parameters used to implement a detection approach can have a significant impact on its performance. Several methods have been proposed in the literature for identifying outliers. These include support vector machines (SVMs), replicated neural networks (RNNs), correlation-based detection, density-based techniques (k-nearest neighbour, local outlier factor), deviations from association rules, and ensemble methods (using feature bagging or score normalisation). A system for detecting anomalies can give a rating or a label as its final result. Unsupervised and semi-supervised methods produce scores, which are ordered lists of anomalies assigned to each instance based on the degree to which the instance is seen as anomalous, rather than a binary value (anomalous or not) (Fig. 4).

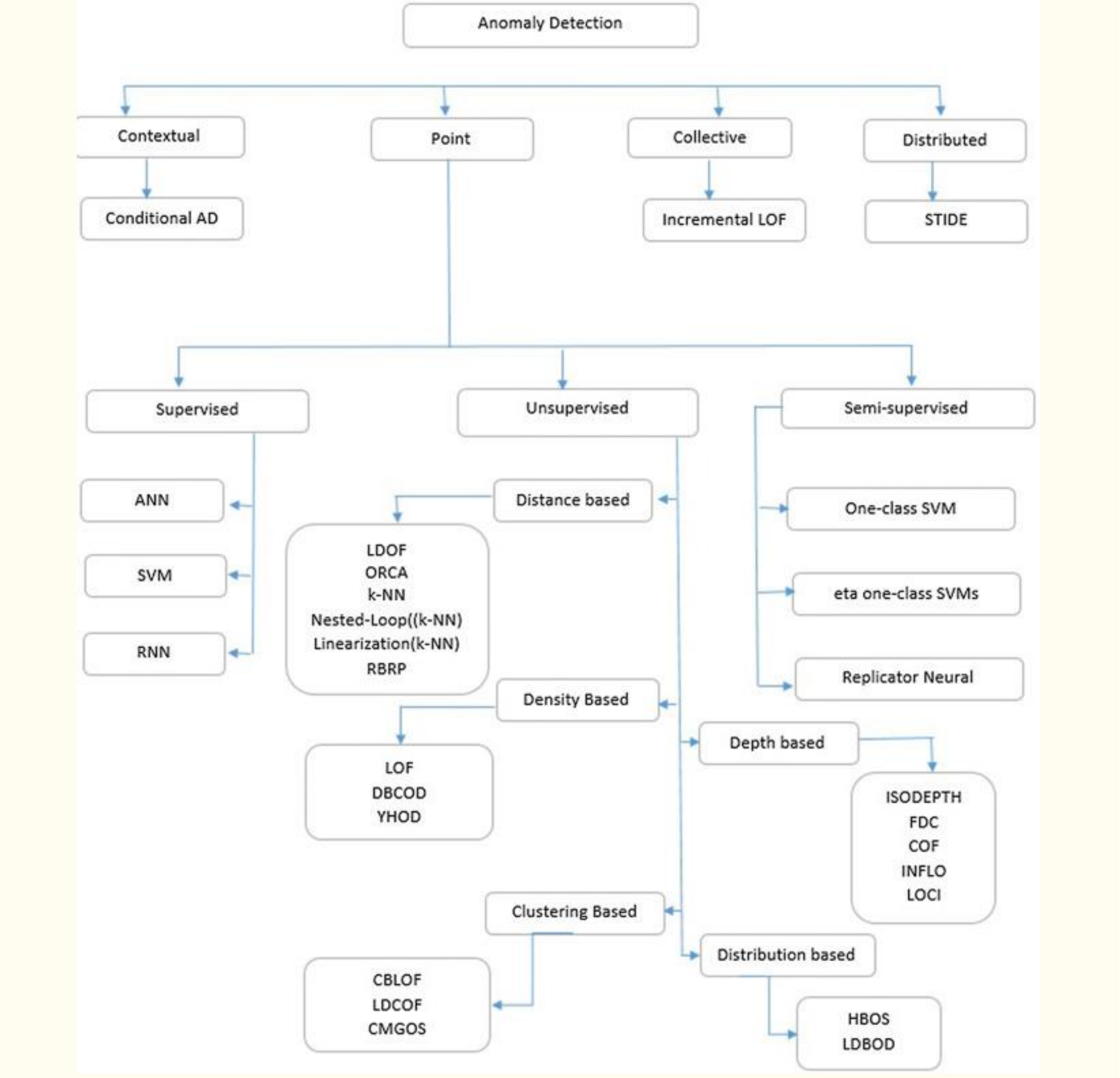


Fig. 4Anomaly detection methods

Anomaly detection using unsupervised k-NN (not k-nearest neighbour classification) is a simple method. It can only detect widespread irregularities, not smaller ones. The k-nearest-neighbors distance is the basis for the anomaly score, which can be calculated either singly (kth-NN) or on average (k-NN). Unique Variable The local outlier factor is a well-known technique for identifying out-of-the-ordinary data points by comparing each data point's local deviation to that of its neighbours. There are three stages to LOF's operation: first, all data points are processed using k-NN; second, the local density is estimated using the local readability density; and third, the LOF score is calculated by comparing the LRD to its neighbour LRD. In reality, the ratio of local density constitutes LOF. Instead of using a k-NN algorithm, the Connectivity-Based Outlier Factor (COF) calculates its value using a spherical density. In contrast to LOF and COF, Cluster-Based Local Outlier Factor (CBLOF) can estimate cluster density. After clustering, anomaly scores are determined by determining how far apart each data point is from the cluster's centre. LDCOF (local density cluster-based outlier factor) is an extension of CBLOF that uses cluster density to categorise clusters as small or large. To address the limitations of CBLOF and unweighted-CBLOF, we propose LDCOF (local density cluster-based outlier factor), an extension of LDLOF that estimates cluster densities based on a spherical distribution of the cluster members. The LDCOF value for this instance is local because it is calculated in relation to the centre of the cluster. Scores are produced by both the LOF and the COF, although neither of these measures is used to establish the threshold. This problem goes away after Outlier Probability is applied. Previous cluster-based anomaly identification methods have been upgraded by the Clustering-based Multivariate Gaussian Outlier Score (CMGOS). Before calculating the covariance matrix, K-means clustering is performed. Estimation of a multivariate Gaussian model is utilised to determine local density, and the Mahalanobis distance is employed to determine distance. The Histogram-based Outlier Score (HBOS) is an unsupervised technique for finding outliers in a histogram. Due to the characteristics' inherent independence, this method is much quicker than multivariate methods. The histogram is calculated for each feature, and the inverse bin height is multiplied by each instance. Unlike other algorithms, which can take hours to process a dataset, HBOS can do it in under a minute thanks to its feature independence.

3.3 Dimensionality Reduction

It is possible to reduce the number of dimensions or variables in a dataset using dimension reduction techniques without losing any of the information contained within the dataset. In order to reduce the complexity of a high number of variables, principal component analysis (PCA) and factor analysis are frequently used. At its core, PCA is all about identifying the principal component (a direction with the largest variance in the dataset) that most accurately represents the data. Principal components are the independent and orthogonal axes with the highest variance that result from rotating the axes of each variable to their highest Eigen vector.

3.4 Support Vector Machines

To learn to categorise new data according to the classes of an existing dataset, the Support Vector Machine (SVM) is a supervised approach.Data is split in half based on the learned hyperplane. It exhibits similar high-level behaviour as C4.5, but without the usage of decision trees. In SVM, data is projected into higher dimensions, and the optimal hyperplane for data classification is determined. If the balls are not too jumbled up, it can be compared to the effect of using a stick to sort red and blue marbles on a table without touching them. When a new ball is added to the table, knowing which side of the stick it was placed on allows one to make an educated guess as to what colour that ball will be. In this setting, the balls can be interpreted as objects, the red and blue colours as indicating two categories, and the stick as symbolising the simplest hyperplane as a line. The hyperplane's function is calculated by support vector machines.

4. RESULTS AND STUDY

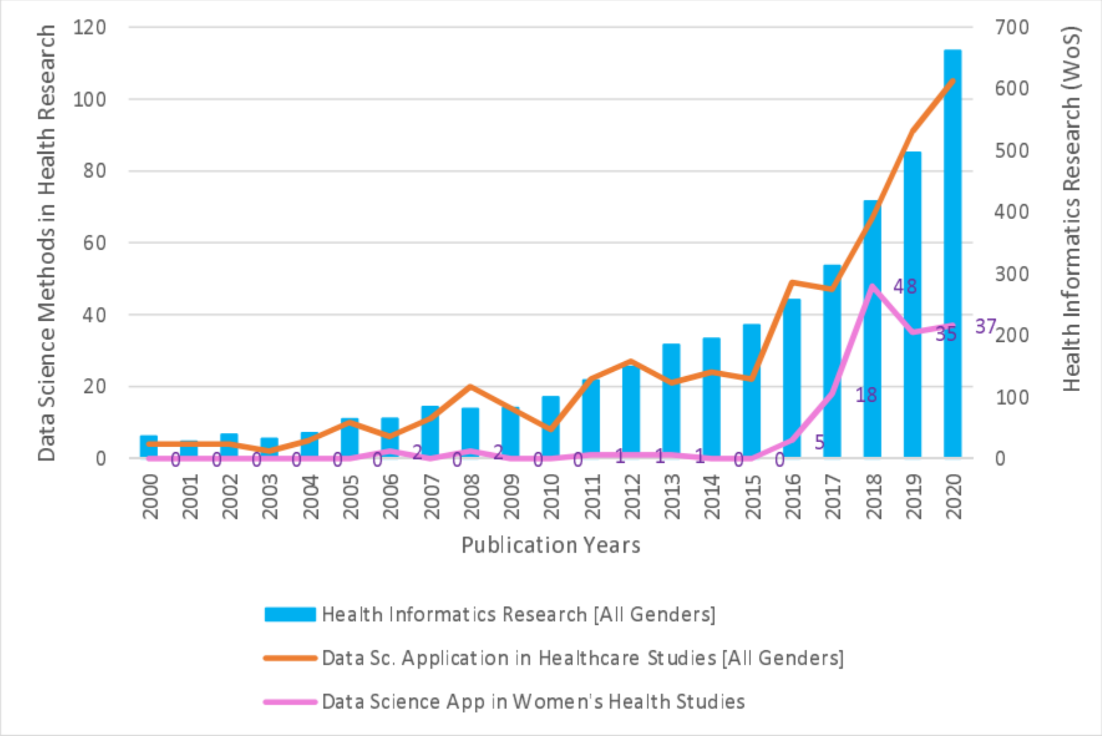
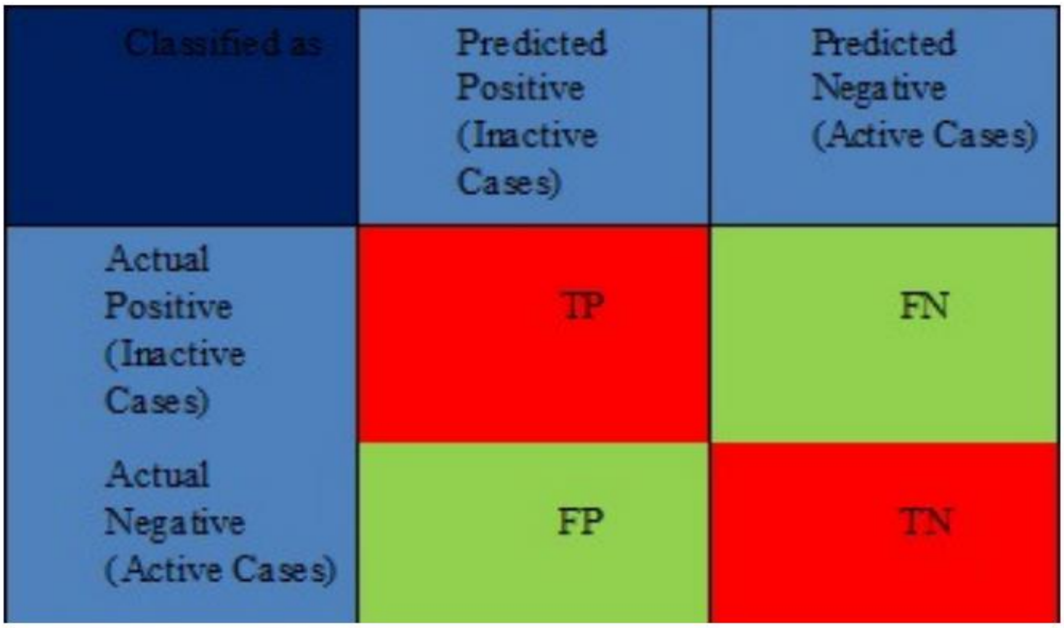


Figure 5: Annual scientific literature production on data science application in woman’s health.

During this time span, 26.6% of DS studies focused on women's health, whereas 73.4% focused on either general health concerns or difficulties involving men (Figure 5).



Classified as	Predicted Positive (Inactive Cases)	Predicted Negative (Active Cases)
Actual Positive (Inactive Cases)	TP	FN
Actual Negative (Active Cases)	FP	TN

Fig. 6. Confusion Matrix for Performance Evaluation.

The accuracy, sensitivity, and specificity of the classifier are measured in relation to the confusion matrix in Fig. 6.

Accuracy: The reliability of a test depends on how well it can identify active cases from inactive ones. It measures how often an individual is able to make an accurate forecast. It is determined using the following equation 1 which is based on the Confusion Matrix.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \textbf{(1)}$$

Where, Number of cases that were accurately classified as "inactive" (true positive, or TP) Incorrectly classifying patients as inactive is what we call a false positive (FN). Number of cases that were successfully diagnosed as inactive (true negative, or TN) The number of times a patient was wrongly labelled as having an active case is the false positive (FP) rate.

Sensitivity: means that a test can reliably identify people who don't engage in any physical activity. The sensitivity equation is given as 2.

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

Specificity: is the test's accuracy in identifying people who engage in regular physical activity. Equation 3 is used to determine a substance's specificity.

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

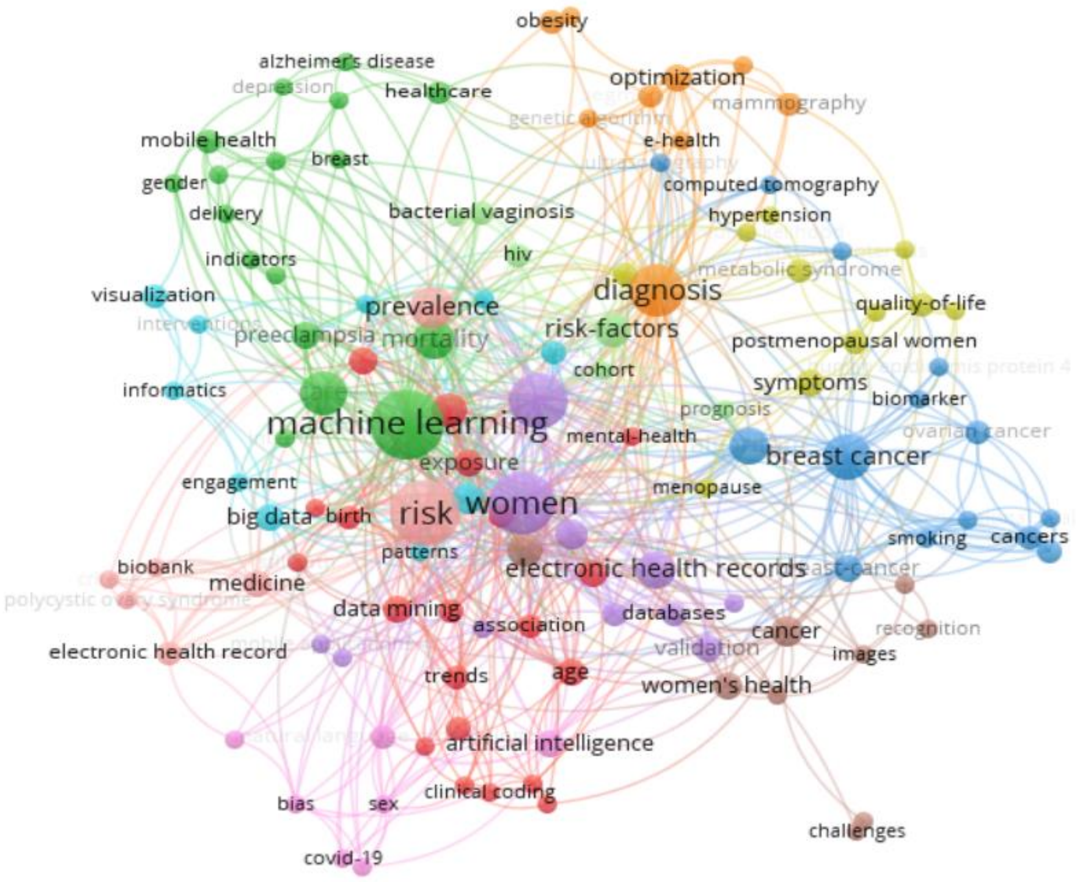


Figure 7. Visualization of the connections and inter-relationships among data science techniques and women's health conditions.

Network analysis performed with VOSVIWER, an open-source science mapping application, shows how various DS methods are interconnected with one another and with the diseases they were designed to study. Breast cancer and pregnancy-related issues are two areas where "machine learning" has seen widespread use. A similar correlation between COVID-19 and cardiovascular disorders is suggested by the persistent SARS-CoV-2 state (Figure 7).

CONCLUSION

Research in this area evaluated DS approaches and algorithms with the aim of improving women's health. Purposes include drawing up diagrams of approaches and problems, doing in-depth analyses of their efficacy, pinpointing what helps reduce health inequalities for women, and suggesting further strategies for doing so. This work accomplishes these aims, sheds light on the history and development of the field, and looks ahead to its promising future by employing straightforward linear trends and forecasting methods. The efficacy of data science approaches and the enhanced outcomes that can aid in solving women's health concerns inspire optimism for narrowing the health gap between sexes, particularly in regards to the illness burden. To improve health care access, however, we need more than just better computers and information systems. One area where policymakers can make a huge difference is in improving health care access. There is a need for additional research into why women are underrepresented in data science-based studies of women's health. The necessity for education on the virtues of the suggested strategies is exemplified by the fact that some women activists are unwittingly pushing against the adoption of data science tools that appear promising for solving women's health problems.

REFERENCES

1. Sengupta PP (2013) Intelligent platforms for disease assessment: novel approaches in functional echocardiography. JACC: Cardiovascular Imaging 6(11):1206–1211. https://doi.org/10. 1016/j.jcmg.2013.09.003

2. Muni Kumar N, Manjula R (2014) Role of big data analytics in rural health care-a step towards svasthbharath. Int J Comp Sci Inform Technol 5(6):7172–7178

3. Ren Y, Werner R, Pazzi N, Boukerche A (2010) Monitoring patients via a secure and mobile healthcare system. IEEE WirelCommun 17(1):59–65

4. IBM Corporation (2013) Data-driven healthcare organizations use big data analytics for big gains. https://silo.tips/ download/ibm-software-white-paper-data-driven-healthcareorganizations-use-big-data-analy

5. Burghard C (2012) Big data and analytics key to accountable care success. IDC health insights :1–9

6. Bollen J, Mao H, Zeng X (2010) Twitter mood predicts the stock market. J Comp Sci 2(1):1–8. https://doi.org/10.1016/j. jocs.2010.12.007

7. Kuehn BM (2013) NIH recruits centers to lead efort to leverage “big data.” JAMA 310(8):787–787

8. Akpan IJ, Akpan AA. Multiple criteria analysis of the popularity and growth of research and practice of visual analytics, and a forecast of the future trajectory. International Transactions in Operational Research. 2021; 1- 24. DOI:10.1111/itor.12952.

9. Abad A, Gerassis S, Saavedra Á, Giráldez E, García JF, Taboada J. A Bayesian assessment of occupational health surveillance in workers exposed to silica in the energy and construction industry. Environmental Science and Pollution Research. 2019 Oct;26(29):29560-9.

10. Atasoy H, Greenwood BN, McCullough JS (2019) The digitization of patient care: a review of the efects of electronic health records on health care quality and utilization. Annu Rev Public Health 40:487–500

11. Bihan K, Lebrun-Vignes B, Funck-Brentano C, Salem JE (2020) Uses of pharmacovigilance databases: an overview. Therapies 75(6):591–598

12. Seshadri DR, Li RT, Voos JE, Rowbottom JR, Alfes CM, Zorman CA, Drummond CK (2019) Wearable sensors for monitoring the physiological and biochemical profle of the athlete. NPJ digital medicine 2(1):1–16

13. Jiang J, Cameron AF, Yang M (2020) Analysis of massive online medical consultation service data to understand physicians’ economic return: observational data mining study. JMIR medical informatics 8(2):e16765. https://doi.org/10.2196/16765.

14. Sri Venkat Gunturi Subrahmanya, “The role of data science in healthcare advancements: applications, benefits, and future prospects”, Irish Journal of Medical Science (1971 -) (2022) 191:1473–1483.

15. Muhammad Imran Razzak, “Big data analytics for preventive medicine ”, Neural Comput Appl. 2020; 32(9): 4417–4451. Published online 2019 Mar 16. doi: 10.1007/s00521-019-04095-y

IJCRT