frontiers | Frontiers in Neuroscience

TYPE Brief Research Report
PUBLISHED 25 October 2022
DOI 10.3389/fnins.2022.1031732

OPEN ACCESS

EDITED BY
Weidong Gao,
Beijing University of Posts and Telecommunications (BUPT), China

REVIEWED BY
Ruiquan Ge,
Hangzhou Dianzi University, China
Ruxin Wang,
Shenzhen Institutes of Advanced Technology (CAS), China

*CORRESPONDENCE
Li Xu
shirleyxu@sjtu.edu.cn
Hong Liu
liuh@sumhs.edu.cn

SPECIALTY SECTION
This article was submitted to Neural Technology, a section of the journal Frontiers in Neuroscience

RECEIVED 30 August 2022
ACCEPTED 06 October 2022
PUBLISHED 25 October 2022

CITATION
Xiong H, Chen H, Xu L, Liu H, Fan L, Tang Q and Cho H (2022) A survey of data element perspective: Application of artificial intelligence in health big data.
Front. Neurosci. 16:1031732.
doi: 10.3389/fnins.2022.1031732

A survey of data element perspective: Application of artificial intelligence in health big data

Honglin Xiong[1], Hongmin Chen[1], Li Xu[1]*, Hong Liu[2]*, Lumin Fan[2,3], Qifeng Tang[4,5,6] and Hsunfang Cho[5,6]

[1]Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China, [2]Business School, University of Shanghai for Science and Technology, Shanghai, China, [3]Operation Management Department, East Hospital Affiliated to Tongji University, Shanghai, China, [4]Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, China, [5]National Engineering Laboratory for Big Data Distribution and Exchange Technologies, Shanghai, China, [6]Shanghai Data Exchange Corporation, Shanghai, China

Artificial intelligence (AI) based on the perspective of data elements is widely used in the healthcare informatics domain. Large amounts of clinical data from electronic medical records (EMRs), electronic health records (EHRs), and electroencephalography records (EEGs) have been generated and collected at an unprecedented speed and scale. For instance, the new generation of wearable technologies enables easy-collecting peoples' daily health data such as blood pressure, blood glucose, and physiological data, as well as the application of EHRs documenting large amounts of patient data. The cost of acquiring and processing health big data is expected to reduce dramatically with the help of AI technologies and open-source big data platforms such as Hadoop and Spark. The application of AI technologies in health big data presents new opportunities to discover the relationship among living habits, sports, inheritances, diseases, symptoms, and drugs. Meanwhile, with the development of fast-growing AI technologies, many promising methodologies are proposed in the healthcare field recently. In this paper, we review and discuss the application of machine learning (ML) methods in health big data in two major aspects: (1) Special features of health big data including multimodal, incompletion, time validation, redundancy, and privacy. (2) ML methodologies in the healthcare field including classification, regression, clustering, and association. Furthermore, we review the recent progress and breakthroughs of automatic diagnosis in health big data and summarize the challenges, gaps, and opportunities to improve and advance automatic diagnosis in the health big data field.

KEYWORDS

healthcare big data, machine learning, automatic diagnosis, healthcare informatics, data elements, artificial intelligence

Introduction

The global "digital divide" *status quo* is quickly changing with the progress in artificial intelligence (AI) technologies and their application area expansion. Nowadays, AI has been widely researched and achieves great success in recent years, and the heart of AI technologies is machine learning (ML) algorithms (Xiong et al., 2022). With the development of the digital economy, Internet, Internet of things (IoT), mobile Internet, and cloud technologies, the application of AI based on health big data presents an explosive increase in recent years (Gokmen and Vlasov, 2016; Dolley, 2018; Ngiam and Khor, 2019; Ye et al., 2021; Weerasinghe et al., 2022). Besides, large amounts of personal health records (PHRs), electronic medical records (EMRs), and electronic health records (EHRs) in hospitals, many governments, and health organizations built the public health monitoring system to collect health data (Heart et al., 2017), such as NEDSS (National Electronic Disease Surveillance System) (National Electronic Disease Surveillance System Working Group, 2001), ProMED-mail (Yu and Madoff, 2004), GPHIN (Global Public Health Intelligence Network) (Dion et al., 2015), HealthMap (Freifeld et al., 2008), MediSys (Linge et al., 2010) and BioCaster (Collier et al., 2008). Among the public health regulatory systems, the representative system is NEDSS. It first defined the standard data protocol to ensure the medical or healthcare data with the identical data format collected across the country. Then, it enables large organizations to upload data automatically through electronic data interchange. The system mainly focused on the collection, exchange, and reporting of diseases and is lagging behind in knowledge mining and early disease warning. Meanwhile, the Internet giants like Google, Facebook, and Twitter collected large amounts of Internet social network data through their products and achieved influenza and other infectious diseases for early warning and tracking (Ginsberg et al., 2009; Signorini et al., 2011; Oflac et al., 2015). Google developed flu outbreak forecast software Google Flu, and the corresponding research result was published in Nature which invoked a large influence on the academic community (Oflac et al., 2015). However, recent research showed that the above-mentioned model in the prediction of flu outbreak existed big defects due to the instability of social network data (Lazer et al., 2014). Intel and IBM companies have also tried to use AI technologies for diabetes control (Nachman et al., 2010; Neuvirth et al., 2011) and the research results were published at the top conference of KDD (Knowledge Discovery in Database) (Neuvirth et al., 2011).

It is widely accepted that health big data have the potential to help physicians to improve diagnosis and aid drug usage. However, there exist many challenges in processing health big data even though researchers have achieved a lot of good results and applications. Except for five major features (5Vs) Volume, Velocity, Variety, Veracity, and Value, health big data have five additional special features (shown in **Figure 1**) as follows:

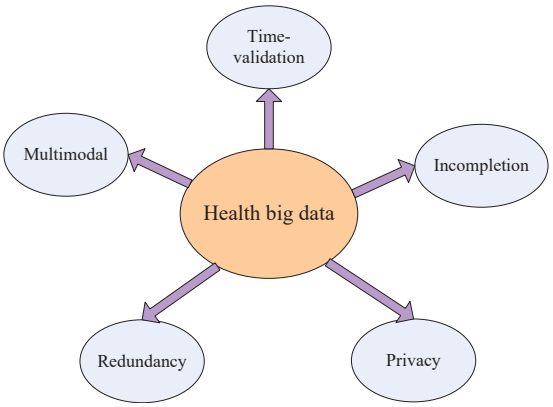

FIGURE 1
Five special features of health big data.

(1) Multimodal: healthcare data consist of text data, image data, and numerical data.

(2) Incompletion: There is a gap between medical data collection and treatment, which cause disease information reflection not enough. At the same time, recording data manually would have deviation, incompletion, and expression uncertainty due to subjective cognition.

(3) Time validation: There is progress between the patient's treatment and the disease outbreak. For example, electrocardiogram (ECG) and electroencephalogram (EEG) are time signals which have strong time-validated properties.

4) Redundancy: There are many same records stored in the healthcare data system. Take EHRs for example, physicians who serve in community hospitals often input multiple records due to unfamiliar computer operations, especially in China.

5) Privacy: It is inevitably related to the patient's private information when researchers process healthcare data. Disclosure of patients' privacy information will hurt patients' lives.

The rest of this paper is organized as follows. Section II introduces big data technologies like Hadoop, Spark, and Storm. The artificial intelligence technologies in health big data are described in section III. Section IV summarizes this paper.

Big data technologies

In the big data era, the traditional data management framework which is based on a relational database management system (RDBMS) is challenged by the increasing data deluge. The old framework is unable to deal with the growing amount of unstructured data. Therefore, new technologies were developed to serve the need for big data management, in the aspects of file systems and programming models. These technologies aim at providing scalability as well as fault tolerance, to handle the huge volume and heterogeneity of big data (Wang, 2017). Big data

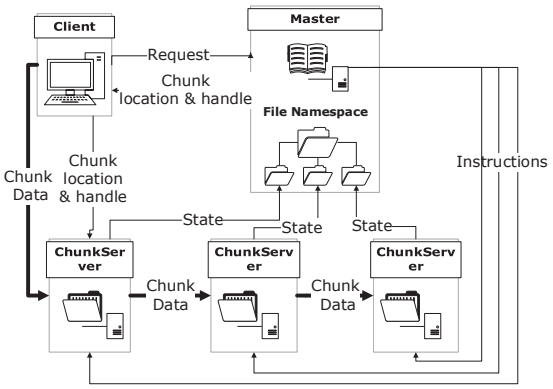

FIGURE 2
Google file system architecture.

have a wide range of applications, including Smart Grid cases, E-health, Internet of Things, Public utilities, Transportation and logistics, and other areas. The following passages introduce newly developed big data technologies in two aspects in detail.

Distributed file system

Although Moore's law promised that the storage capacity of computer chips doubles roughly every 18 months or so, current magnetic storage technology relies on a million atoms per bit, and the quantity of data grows much faster (Bradley, 2017). It is in great demand to develop an efficient and persistent distributed file system. In 2000, Brewer proposed the CAP theorem which states that it is impossible to meet the requirements of consistency, availability, and partition tolerance in an asynchronous distributed read/write system (Wang and Manzie, 2022). As frequent requests are common in big data scenarios, distributed file systems are commonly designed as AP systems, in which only eventual consistency rather than strong consistency is ensured.

As a pioneer in the attempts of providing users with high-performance services with a distributed file system, Google File System (GFS) achieved great success and its concepts were inherited by a lot of its successors. It features work division between control and storage servers, and replication of the same data, to provide performance and reliability in this way (Ghemawat et al., 2003). Its basic architecture and the data flow in it during a writing procedure was shown in **Figure 2**.

The highlight of this architecture includes that the control flow is separated from the data flow which leads to higher performance and that the replicas of data offer both reliability and efficiency under good management. Meanwhile, this system also has limitations in supporting small files, for its specific design purpose to support Google's own service.

Some successors of GFS are generally different implementations of the same idea, for example, HDFS (Kumari and Bucker, 2022) and KosmosFS. Others made some improvements to meet their own demands. Facebook developed Haystack which reduces disk operations for metadata lookups and increases overall throughput to support their Photos application (Beaver et al., 2010). Taobao developed TFS (Fu et al., 2014) which provides significantly higher performance in dealing with small files to support their online shopping service.

In conclusion, after many years of development, distributed file systems are relatively mature, and it is a prevailing trend to develop a customized DFS for a certain field.

MapReduce framework

As scalability and performance are two of the key requirements of a big data system, parallel computing must be implemented to offer these features. However, traditional parallel programming models fail in migrating to big data systems which consist of a massive number of servers over a wide area. In recent years, a lot of programming models were proposed to provide solutions to this specific need.

As the forerunner in distributing heavy computations across thousands of machines, MapReduce abstracted two basic operations from a broad variety of real-world tasks (Kalia and Gupta, 2021). The map function takes an input pair and produces a set of intermediate key/value pairs which will then be grouped and passed to the Reduce function. Reduce function is responsible for merging a set of values for one key to form a possibly smaller set of values. Once programmers give the proper definition to the two operations, the underlying runtime system will automatically parallelize and distribute the computation and handle other details including machine failures and inter-machine communication. The major part of its work procedure is illustrated in **Figure 3**.

Many programming models have been proposed afterward. Some provided considerable improvement to the MapReduce model. Microsoft developed Dryad (Isard et al., 2007) in which a job is abstracted as a directed acyclic graph. Each vertex is a program, and data channels are represented by edges. Higher generality is reached as data channels can be customized to support functions more than Map and Reduce. Spark (Solovyev et al., 2010) introduced an abstraction called resilient distributed datasets (RDDs) and parallel operations on them. An RDD represents a read-only collection of objects across a set of machines. By combining parallel operations based on data, Spark avoids redundant I/O operations and multiplied the performance. Other models focus on specific categories of distributed computing. Pregel (Malewicz et al., 2010) aims at large-scale graph processing, in which poor locality of memory access and very little work per vertex often lead to poor efficiency. Storm, as a stream processing model, offers outstanding performance in event processing and incremental computation.

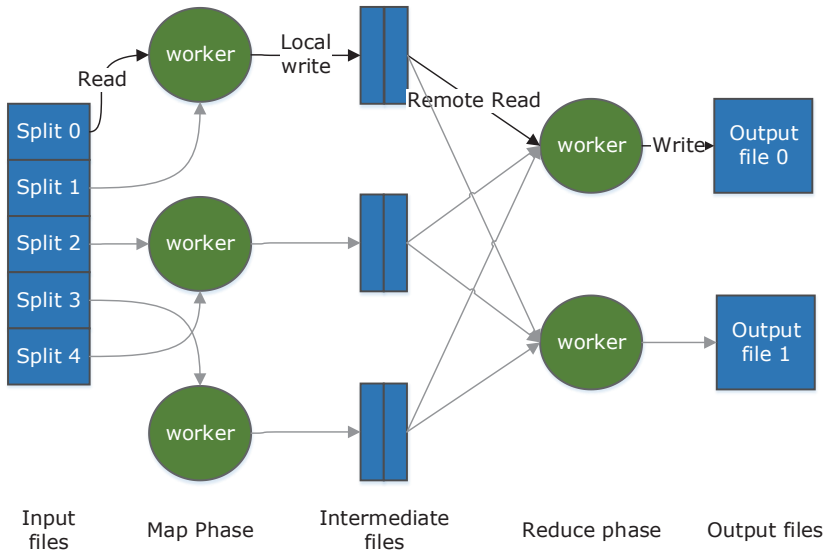

FIGURE 3
MapReduce work procedure.

Artificial intellectual technologies

In the big data era, various public hospitals and private healthcare providers are producing large amounts of data that are difficult to process. Therefore, powerful automatic artificial intellectual algorithms are needed for the analysis and processing of useful information from healthcare data. This information is very precious for healthcare specialists and physicians to apprehend the cause of diseases and for providing better and cost-effective treatment to patients. To improve prediction accuracy, there are various artificial intelligence technologies such as classification, regression, clustering, and association used in healthcare data analysis to increase the healthcare provider's capability for making the decision in regard to patients health. There are large amounts of research resources available regarding artificial intellectual application in health big data which are presented in subsequent sections with their advantages and disadvantages.

Classification

One of the data analysis tasks is classification, which divides data into target labels. Each data point is predicted into the target label by a pattern classifier. For instance, hypertension patients can be classified into three stages of stage 1 hypertension, stage 2 hypertension, and stage 3 hypertension (Wermelt and Schunkert, 2017) on the basis of a supervised classification model. Dataset is often partitioned into a training set, validation dataset, and testing dataset. The training dataset is utilized for training the classifier. The validation dataset is used to tune the classifier parameters to achieve optimal performance. Testing dataset verifies the classification accuracy. **Figure 4** shows the entire flowchart of classification.

In the ML domain, SVM as a supervised classifier is widely used for classification (Tsang et al., 2005). It is widely applied in healthcare data recently. Fei proposed the PSO-SVM model which has a strong global search capability (Fei, 2010), and the PSO-SVM model is applied to the diagnosis of arrhythmia cordis, in which PSO is used to determine the free parameters of the support vector machine (Cuong-Le et al., 2022). The testing results showed that the average classification accuracy is 95.65%. Huang et al. (2008) developed a hybrid SVM-based strategy with feature selection to render a diagnosis between breast cancer and fibroadenoma and to find the important risk factor for breast cancer (Azar and El-Said, 2014). The experimental results showed that the features {HSV-1, HHV-8} or {HSV-1, HHV-8, CMV} could achieve identical high accuracy, at 86% of the average overall hit rate. Zheng et al. (2014) used a hybrid of K-means and support vector machine (K-SVM) algorithms to extract useful information and diagnose the tumor. According to 10-fold cross-validation, the developed methodology which was tested on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the University of California—Irvine ML repository, increased the accuracy to 97.38%. Avci utilized the genetic-support vector machine (GSVM) approach to classify the Doppler signals of heart valve diseases (Gonzalez-Abril et al., 2014). With the combination of feature extraction and classification from measured Doppler signal waveforms, the performance of the GSVM system showed that this GSVM system is effective to detect Doppler heart

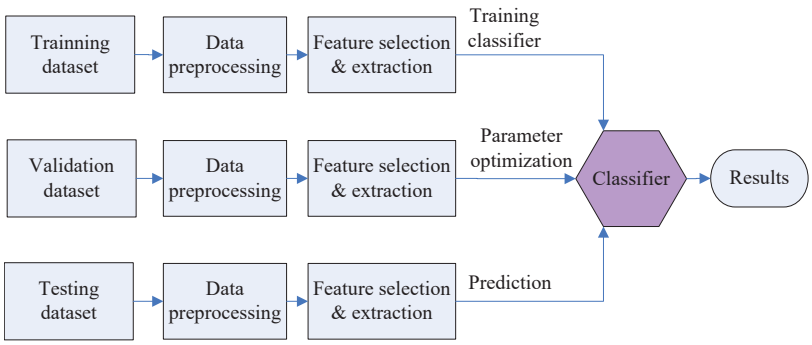

FIGURE 4
Classification flowchart.

sounds. The average rate of correct classification rate was about 95%.

A decision tree (DT) is a common ML method for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition (Kotsiantis, 2013; Loh, 2014). Due to its results with features of human-readable and interpretable, DT is widely used by many researchers in the healthcare field. Khan et al. (2008) proposed to investigate a hybrid scheme based on fuzzy decision trees, as an efficient alternative to predict breast cancer survivability for personalized healthcare. The experimental results showed that, for cancer prognosis, hybrid fuzz decision tree classification can achieve an average accuracy of 85%. Levashenko et al. (2016) proposed fuzzy decision trees in the medical decision-making support system. The classification accuracy of breast cancer was over 96%. Hassan et al. (2011) developed a decision tree with a CART classification algorithm to forecast response to therapy with 200 chronic hepatitis C patients. The overall classification error was 20%, and 80% was the best accuracy. Moon et al. (2012) developed decision tree models for characterizing smoking patterns in older adults. Their results suggest that social workers need to provide more customized and individualized interventions to older adults. Chang and Chen (2009) applied a decision tree and neural network to increase the quality of dermatologic diagnosis. Using sensitivity analysis combined with the decision tree model, on the contrary, has the least accuracy, which is 80.33%.

A neural network (NN) is based on a biological nervous system having multiple interrelated processing elements known as neurons, functioning in unity to solve a classification problem. Rules extracted from the trained model help to improve the interoperability of the learned network (Schmidhuber, 2015). Er et al. (2010) developed an artificial neural network (ANN) to diagnose chest diseases. Sokolov (2018) presents recent research on approaches in autonomous systems for combining multiple modalities for emotion estimation based on neural networks. Sharma and Parmar (2020) utilized a neural network approach to analyze a heart disease dataset the experimental results proved better accuracy (90.76%) than other optimizations. It is applied to heart disease datasets and finds out a good prediction (Sharma and). In the past several years, intricate neural networks have inspired the further development of intelligent systems. Many disciplines, including the complex field of medicine, neuroimaging modalities, and diagnosis of the disease, have taken advantage of the useful applications of artificial neural networks (Yang et al., 2018; Deperlioglu et al., 2020).

Bayesian decision theory is a basic method under the statistic framework, and it is extended easily to do classification tasks (Chickering et al., 2004). Liu and Lu (2009) proposed to use Bayesian belief network (BBN) as decision support for the higher-level risk estimate which can represent the probabilistic relationships between all kinds of health effects and air pollutants. Dawson et al. (2015) used the Bayesian network to produce the baseline distribution by taking the joint distribution of the data and conditioning it on attributes that are responsible for anomaly pattern detection for disease outbreaks. Curiac et al. (2009) analyzed the psychiatric patient data using BBN in making a significant decision regarding patient health suffering from psychiatric disease and performed an experiment on real data obtained from Lugoj Municipal Hospital.

Long et al. (2015) proposed a heart disease diagnosis system using rough set-based attribute reduction and interval type-2 fuzzy logic system (IT2FLS). The experimental results showed that it could efficiently find minimal attribute reduction from the high-dimensional dataset that enhances the performance of the classification system. The use of an interval type-2 fuzzy logic system for the classification of heart disease datasets to handle the uncertainties and noisiness of these datasets was successful. Nahar et al. (2013) presented the potential of an expert judgment-based (i.e., medical knowledge-driven) feature selection process (termed as MFS). The medical knowledge-based feature selection method has shown promise

for use in heart disease diagnostics. The main classification methods used in healthcare big data are shown in **Table 1**.

The application of classification analysis methods in medicine is getting more and more advanced, not only is it used extensively in disease diagnosis, but also there will be more breakthroughs in disease treatment options in future, and all of these expectations become more apparent in the near future.

Regression

Regression analysis is a statistical method to determine the quantitative relationship between two or more variables. Based on observational data, regression analysis could establish appropriate dependencies between variables and analyze the inherent rules of data (Merrick et al., 2022). It is widely used for forecasting in the healthcare field. Gutiérrez et al. (2010) proposed a hybrid multi-logistic methodology, named logistic regression using initial and radial basis function (RBF) covariates. Agarwal (2011). developed weighted support vector regression (SVR) approach for remote healthcare monitoring. Vinsnes et al. (2001) developed a regression analysis approach for healthcare personnel's attitudes to predict nursing assistants' attitudes. Ko and Osei-Bryson (2004) explored the productivity impact of information technology (IT) in the healthcare industry using a regression spline (RS)-based approach. Luo et al. (2012) presented scalable orthogonal regression (SOR) for non-redundant feature selection and its healthcare applications.

Regression analysis can accurately measure the degree of correlation between factors and the degree of the regression fit to improve the effectiveness of prediction, which is of great significance in medical diagnosis. More recently, regression analysis is one of the most frequently used analytical techniques in disease diagnosis and etiology analysis (Hannan et al., 2010; Liu et al., 2018; Jfri et al., 2021).

Clustering

Clustering is an unsupervised learning method that is different from classification. Clustering is a process of classifying data into different classes or clusters, so objects in the same cluster have a large similarity, and objects between different clusters have a large degree of dissimilarity (Caron et al., 2018). Clustering is also used in the healthcare field. Sinaga and Yang (2020) proposed a novel unsupervised k-means (U-k-means) clustering algorithm which automatically finds an optimal number of clusters without giving any initialization and parameter selection. Stein et al. (2007) used data clustering techniques to develop health state descriptions based on data from 66 women who completed the EORTC QLQ-C30 over a 6-month period while receiving chemotherapy for ovarian cancer. Belciug et al. (2010) detected breast cancer recurrence with the help of a clustering-based approach. Zhao et al. (2020) proposed to propose a new deep learning and clustering UDFCMN (Unsupervised Deep Fuzzy C-Means clustering Network) model, to cluster lung cancer patients from lung CT images; these results also indicate that this method has practical applications in lung cancer pathogenesis studies and provide useful guidelines for personalized cancer therapy. Balasubramanian and Umarani (2012) analyzed the impact of fluoride on human health (dental) with the help of a clustering-based method and found meaningful hidden patterns which gave meaningful decision-making to this socio-economic real-world health hazard. In addition, some researchers have also used clustering methods to early detect Alzheimer's disease (Escudero et al., 2011; Holilah et al., 2021). The main clustering methods used in healthcare big data are shown in **Table 2**.

Cluster analysis is essentially finding a statistic that objectively reflects the affinity of an element and then classifying the elements into categories based on this statistic. Cluster analysis decomposes the symptoms of chronic diseases and is used to assess the quality of life in chronic diseases, such as lung cancer; cluster analysis is very effective in assessing these diseases.

Association

Association is one of the most vital approaches to data mining that is used to find out the frequent patterns, and interesting relationships among a set of data items in the data repository. Frequent patterns are patterns that appear frequently in a dataset. The initial motivations of the association rules were raised for the issue of the market basket analysis. The association process analyzes the customer's shopping habits by discovering the association between the different items placed in the "shopping basket" by the customer. The discovery of this association can help retailers understand which goods are frequently purchased by customers at the same time, so as to help them develop better marketing strategies (Tomar and Agarwal, 2013). Association also has a great impact in the healthcare field to detect the relationships among diseases, health status, and symptoms. Nahar et al. (2013) presented a rule extraction experiment on heart disease data using different rule-mining algorithms (*Apriori*, Predictive *Apriori*, and Tertius). Further rule-mining-based analysis was undertaken by categorizing data based on gender, and significant risk factors for heart disease were found for both men and women. Ji et al. (2010) developed a new interestingness measure, exclusive causal-leverage, based on an experience-based fuzzy recognition-primed decision (RPD) model. On the basis of this new measure, a new association rule algorithm is proposed to discover infrequent causal relationships in electronic health databases (Horton et al., 2019). In addition, Soni and Vyas (2010) used the associative method to construct a classifier for predictive analysis in healthcare data mining.

TABLE 1 Main application of classification methods in healthcare big data.

Method	Scenes	Features
SVM	Diagnosis of arrhythmia cordis; diagnosis between breast cancer and fibroadenoma; diagnosis of the tumor; detect Doppler heart sounds and so on	Non-linear mapping; low generalization error rate, fast classification; suitable for small samples, excellent generalization ability, etc.,
DT	Predict breast cancer survivability; medical decision-making support system; characterizing smoking patterns and so on	Simple to understand, easy to explain, visualization, and wide applicability; prone to overfitting, in addition, small changes in the data can affect the results and are unstable
NN	Including the complex field of medicine, neuroimaging modalities, and diagnosis of the disease; image analysis and interpretation	With self-learning function; no *a priori* assumptions about the problem model are required. suitable for some problems with very complex environmental information, unclear knowledge background, and unclear inference rules.
BN	Anomaly pattern detection for disease outbreaks; regarding patient health suffering from psychiatric disease	Distribution of input data in each layer of the network is relatively stable, which accelerates the model learning speed; makes the model less sensitive to the parameters in the network, simplifies the tuning process, and makes the network learning more stable

TABLE 2 Main application of clustering methods in healthcare big data.

Method	Scenes	Features
k-means	Health state descriptions; Alzheimer's disease; health hazards and so on	Fast convergence; better clustering effect; stronger interpretability of the model and so on
Fuzzy C-means	Lung cancer patients from lung CT images	Clustering objectively and accurately

Using the association analysis method to discover the relationship between the attributes in the medical dataset, especially some general factors such as age and smoking habits, and the measured body organ function indices related to the possibility of disease, the doctor can accurately determine the possibility of disease through the patient's characteristics, which is very meaningful for medical diagnosis, and the future application of the association analysis method to predict diseases and develop treatment plans based on vital signs.

In summary, AI as a role in healthcare big data, its effects on the development of the medical industry, applications of AI in medicine, challenges, and promises of both AI and big data with respect to healthcare, and prevailing techniques (methods such as deep neural network, convolutional neural network, and recurrent neural network) and tools for performance optimization of healthcare big data can be used by the medical industry.

Conclusion

This paper investigated the application of AI technologies in health big data based on a data elements perspective. Traditional data management framework, which is based on a relational database management system (RDBMS), is hard to deal with a growing amount of healthcare data. Big data processing frameworks like Hadoop (Spark et al.) are employed in the data preprocess stage to accommodate big data and accelerate computing efficiency. We found that there is no single ML method that gives consistently good results for all kinds of health big data. The performance of ML methods depends on the type of dataset that researchers have taken for doing the experiment. To get the higher performance of ML method, most of the research combined many artificial methods to complement the deficiency of each one called hybrid method or integrated method or assemble method.

In addition, it is well known in the AI field that feature selection and extraction are very important factors that affect the performance of artificial methods. Features are extracted and selected on the basis of healthcare/medical domain knowledge and optimal techniques in normal conditions. For healthcare providers and medical providers, AI technologies are widely utilized to make effective decisions in regard to how to enhance patients' health, how to provide healthcare service at low cost, and how to remind physicians to avoid misusing drugs and misdiagnosing.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

HX analyzed the methods and penned the manuscript. HC, LX, and HL gave important suggestions. QT, LF,

and HFC participated in discussions and provided some literature resources. HC revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Major Research Project of Philosophy and Social Science of the Ministry of Education of PRC (Grant No. 20JZD010).

Acknowledgments

The authors would like to thank Chen and Xu who give their precious suggestions on improving manuscript quality.

Conflict of interest

Authors QT and HFC were employed by Shanghai Data Exchange Corporation.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Agarwal, S. (2011). "Weighted support vector regression approach for remote healthcare monitoring," in *Proceedings of the 2011 international conference on recent trends in information technology (ICRTIT)*, (Chennai: IEEE), 969–974. doi: 10.1109/ICRTIT.2011.5972437

Azar, A. T., and El-Said, S. A. (2014). Performance analysis of support vector machines classifiers in breast cancer mammography recognition. *Neural Comput. Appl.* 24, 1163–1177. doi: 10.1007/s00521-012-1324-4

Balasubramanian, T., and Umarani, R. (2012). "An analysis on the impact of fluoride in human health (dental) using clustering data mining technique," in *Proceedings of the international conference on pattern recognition, informatics and medical engineering*, (Salem: IEEE), 370–375. doi: 10.1109/ICPRIME.2012.6208374

Beaver, D., Kumar, S., Li, H. C., Sobel, J., and Vajgel, P. (2010). "Finding a needle in haystack: Facebook's photo storage," in *Proceedings of the 9th USENIX symposium on operating systems design and implementation (OSDI 10)*, (Vancouver: IEEE), 1–8.

Belciug, S., Salem, A. B., Gorunescu, F., and Gorunescu, M. (2010). "Clustering-based approach for detecting breast cancer recurrence," in *Proceedings of the 2010 10th international conference on intelligent systems design and applications*, (Cairo: IEEE), 533–538. doi: 10.1109/ISDA.2010.5687211

Bradley, D. (2017). Single-atom memory maintains Moore's Law. *Materialstoday* 20:225. doi: 10.1016/j.mattod.2017.04.021

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, (Tel Aviv: IEEE), 132–149.

Chang, C. L., and Chen, C. H. (2009). Applying decision tree and neural network to increase quality of dermatologic diagnosis. *Expert Syst. Appl.* 36, 4035–4041. doi: 10.1016/j.eswa.2008.03.007

Chickering, M., Heckerman, D., and Meek, C. (2004). Large-sample learning of Bayesian networks is NP-hard. *J. Mach. Learn. Res.* 5, 1287–1330.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., et al. (2008). BioCaster: Detecting public health rumors with a Web-based text mining system. *Bioinformatics* 24, 2940–2941. doi: 10.1093/bioinformatics/btn534

Cuong-Le, T., Nghia-Nguyen, T., Khatir, S., Trong-Nguyen, P., Mirjalili, S., and Nguyen, K. D. (2022). An efficient approach for damage identification based on improved machine learning using PSO-SVM. *Eng. Comput.* 38, 3069–3084. doi: 10.1007/s00366-021-01299-6

Curiac, D. I., Vasile, G., Banias, O., Volosencu, C., and Albu, A. (2009). "Bayesian network model for diagnosis of psychiatric diseases," in *Proceedings of the ITI 2009 31st international conference on information technology interfaces*, (Cavtat: IEEE), 61–66. doi: 10.1109/ITI.2009.5196055

Dawson, P., Gailis, R., and Meehan, A. (2015). Detecting disease outbreaks using a combined Bayesian network and particle filter approach. *J. Theor. Biol.* 370, 171–183. doi: 10.1016/j.jtbi.2015.01.023

Deperlioglu, O., Kose, U., Gupta, D., Khanna, A., and Sangaiah, A. K. (2020). Diagnosis of heart diseases by a secure internet of health things system based on autoencoder deep neural network. *Comput. Commun.* 162, 31–50. doi: 10.1016/j.comcom.2020.08.011

Dion, M., AbdelMalik, P., and Mawudeku, A. (2015). Big data: Big data and the global public health intelligence network (GPHIN). *Can. Commun. Dis. Rep.* 41, 209–214. doi: 10.14745/ccdr.v41i09a02

Dolley, S. (2018). Big data's role in precision public health. *Front. Public Health* 68:6. doi: 10.3389/fpubh.2018.0006

Er, O., Yumusak, N., and Temurtas, F. (2010). Chest diseases diagnosis using artificial neural networks. *Expert Syst. Appl.* 37, 7648–7655. doi: 10.1016/j.eswa.2010.04.078

Escudero, J., Zajicek, J. P., and Ifeachor, E. (2011). Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and K-means. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2011, 6470–6473. doi: 10.1109/IEMBS.2011.6091597

Fei, S. W. (2010). Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine. *Expert Syst. Appl.* 37, 6748–6752. doi: 10.1016/j.eswa.2010.02.126

Freifeld, C. C., Mandl, K. D., Reis, B. Y., and Brownstein, J. S. (2008). HealthMap: Global infectious disease monitoring through automated classification and visualization of Internet media reports. *J. Am. Med. Inf. Assoc.* 15, 150–157. doi: 10.1197/jamia.M2544

Fu, S., He, L., Huang, C., Liao, X., and Li, K. (2014). Performance optimization for managing massive numbers of small files in distributed file systems. *IEEE Trans. Parallel Distrib. Syst.* 26, 3433–3448. doi: 10.1109/TPDS.2014.2377720

Ghemawat, S., Gobioff, H., and Leung, S. T. (2003). "The Google file system," in *Proceedings of the nineteenth ACM symposium on operating systems principles*, New York, NY, 29–43. doi: 10.1145/945445.945450

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature* 457, 1012–1014.

Gokmen, T., and Vlasov, Y. (2016). Acceleration of deep neural network training with resistive cross-point devices: Design considerations. *Front. Neurosci.* 10:333. doi: 10.3389/fnins.2016.00333

Gonzalez-Abril, L., Nunez, H., Angulo, C., and Velasco, F. (2014). GSVM: An SVM for handling imbalanced accuracy between classes inbi-classification problems. *Appl. Soft Comput.* 17, 23–31. doi: 10.1016/j.asoc.2013.12.013

Gutiérrez, P. A., Hervás-Martínez, C., and Martínez-Estudillo, F. J. (2010). Logistic regression by means of evolutionary radial basis function neural networks. *IEEE Trans. Neural Netw.* 22, 246–263. doi: 10.1109/TNN.2010.2093537

Hannan, S. A., Manza, R. R., and Ramteke, R. J. (2010). Generalized regression neural network and radial basis function for heart disease diagnosis. *Int. J. Comput. Appl.* 7, 7–13. doi: 10.5120/1325-1799

Hassan, M., Abdalla, M. I., Ahmed, S. R., Akil, W., Esmat, G., Khamis, S., et al. (2011). The decision tree mode for prediction the response to the treatment in patients with chronic hepatitis C. *N. Y. Sci. J.* 4, 69–79.

Heart, T., Ben-Assuli, O., and Shabtai, I. (2017). A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy. *Health Policy Technol.* 6, 20–25. doi: 10.1016/j.hlpt.2016.08.002

Holilah, D., Bustamam, A., and Sarwinda, D. (2021). Detection of Alzheimer's disease with segmentation approach using K-Means Clustering and Watershed Method of MRI image. *J. Phys. Conf. Ser.* 1725:012009. doi: 10.1088/1742-6596/1725/1/012009

Horton, D. B., Bhullar, H., Carty, L., Cunningham, F., Ogdie, A., Sultana, J., et al. (2019). "Electronic health record databases," in *Pharmacoepidemiology*, eds B. L. Strom, S. E. Kimmel, and S. Hennessy (Hoboken, NJ: Wiley), 241–289. doi: 10.1002/9781119413431.ch13

Huang, C. L., Liao, H. C., and Chen, M. C. (2008). Prediction model building and feature selection with support vector machines in breast cancer diagnosis. *Expert Syst. Appl.* 34, 578–587. doi: 10.1016/j.eswa.2006.09.041

Isard, M., Budiu, M., Yu, Y., Birrell, A., and Fetterly, D. (2007). "Dryad: Distributed data-parallel programs from sequential building blocks," in *Proceedings of the 2nd ACM SIGOPS/eurosys European conference on computer systems*, Vol. 2007, (New York, NY), 59–72. doi: 10.1145/1272996.1273005

Jfri, A., Nassim, D., O'Brien, E., Gulliver, W., Nikolakis, G., and Zouboulis, C. C. (2021). Prevalence of hidradenitis suppurativa: A systematic review and meta-regression analysis. *JAMA Dermatol.* 157, 924–931. doi: 10.1001/jamadermatol.2021.1677

Ji, Y., Ying, H., Dews, P., Farber, M. S., Mansour, A., Tran, J., et al. (2010). "A fuzzy recognition-primed decision model-based causal association mining algorithm for detecting adverse drug reactions in postmarketing surveillance," in *Proceedings of the international conference on fuzzy systems*, (Barcelona: IEEE), 1–8. doi: 10.1109/FUZZY.2010.5584288

Kalia, K., and Gupta, N. (2021). Analysis of hadoop MapReduce scheduling in heterogeneous environment. *Ain Shams Eng. J.* 12, 1101–1110. doi: 10.1016/j.asej.2020.06.009

Khan, M. U., Choi, J. P., Shin, H., and Kim, M. (2008). Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2008, 5148–5151. doi: 10.1109/IEMBS.2008.4650373

Ko, M., and Osei-Bryson, K. M. (2004). The productivity impact of information technology in the healthcare industry: An empirical study using a regression spline-based approach. *Inf. Softw. Technol.* 46, 65–73. doi: 10.1016/S0950-5849(03)00110-1

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artif. Intell. Rev.* 39, 261–283. doi: 10.1007/s10462-011-9272-4

Kumari, P. S., and Bucker, N. A. B. A. (2022). Data integrity verification using HDFS framework in data flow material environment using cloud computing. *Mater. Today Proc.* 60, 1329–1333. doi: 10.1016/j.matpr.2021.09.435

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science* 343, 1203–1205. doi: 10.1126/science.1248506

Levashenko, V., Zaitseva, E., Kvassay, M., and Deserno, T. M. (2016). "Reliability estimation of healthcare systems using fuzzy decision trees," in *Proceedings of the 2016 federated conference on computer science and information systems (FedCSIS)*, (Gdansk: IEEE), 331–340. doi: 10.15439/2016F150

Linge, J. P., Steinberger, R., Fuart, F., Bucci, S., Belyaeva, J., Gemo, M., et al. (2010). "MedISys: Medical information system," in *Advanced ICTs for disaster management and threat detection: Collaborative and distributed frameworks*, eds E. Asimakopoulou and N. Bessis (Hershey, PA: IGI Global), 131–142.

Liu, K. F., and Lu, C. F. (2009). "BBN-based decision support for health risk analysis," in *Proceedings of the 2009 fifth international joint conference on INC, IMS and IDC*, Washington, DC, 696–702. doi: 10.1109/NCM.2009.187

Liu, M., Zhang, J., Adeli, E., and Shen, D. (2018). Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. *IEEE Trans. Biomed. Eng.* 66, 1195–1206. doi: 10.1109/TBME.2018.2869989

Loh, W. Y. (2014). Fifty years of classification and regression trees. *Int. Stat. Rev.* 82, 329–348. doi: 10.1111/insr.12016

Long, N. C., Meesad, P., and Unger, H. (2015). A highly accurate firefly based algorithm for heart disease prediction. *Expert Syst. Appl.* 42, 8221–8231. doi: 10.1016/j.eswa.2015.06.024

Luo, D., Wang, F., Sun, J., Markatou, M., Hu, J., and Ebadollahi, S. (2012). "Sor: Scalable orthogonal regression for non-redundant feature selection and its healthcare applications," in *Proceedings of the 2012 SIAM international conference on data mining*, Alexandria, 576–587. doi: 10.1137/1.9781611972825.50

Malewicz, G., Austern, M. H., Bik, A. J., Dehnert, J. C., Horn, I., Leiser, N., et al. (2010). "Pregel: A system for large-scale graph processing," in *Proceedings of the 2010 ACM SIGMOD international conference on management of data*, (New York, NY: ACM), 135–146. doi: 10.1145/1807167.1807184

Merrick, L. F., Lozada, D. N., Chen, X., and Carter, A. H. (2022). Classification and regression models for genomic selection of skewed phenotypes: A case for disease resistance in winter wheat (*Triticum aestivum* L). *Front. Genet.* 13:835781. doi: 10.3389/fgene.2022.835781

Moon, S. S., Kang, S. Y., Jitpitaklert, W., and Kim, S. B. (2012). Decision tree models for characterizing smoking patterns of older adults. *Expert Syst. Appl.* 39, 445–451. doi: 10.1016/j.eswa.2011.07.035

Nachman, L., Baxi, A., Bhattacharya, S., Darera, V., Deshpande, P., Kodalapura, N., et al. (2010). "Jog falls: A pervasive healthcare platform for diabetes management," in *International conference on pervasive computing*, eds P. Floréen, A. Krüger, and M. Spasojevic (Berlin: Springer), 94–111.

Nahar, J., Imam, T., Tickle, K. S., and Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Syst. Appl.* 40, 1086–1093. doi: 10.1016/j.eswa.2012.08.028

National Electronic Disease Surveillance System Working Group (2001). National Electronic Disease Surveillance System (NEDSS): A standards-based approach to connect public health and clinical medicine. *J. Public Health Manag. Pract.* 43–50.

Neuvirth, H., Ozery-Flato, M., Hu, J., Laserson, J., Kohn, M. S., Ebadollahi, S., et al. (2011). "Toward personalized care management of patients at risk: The diabetes case study," in *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, Vol. 8, San Diego, CA, 395–403.

Ngiam, K. Y., and Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* 20, e262–e273. doi: 10.1016/S1470-2045(19)30149-4

Oflac, B. S., Dobrucalı, B., Yavas, T., and Escobar, M. G. (2015). Services marketing mix efforts of a global services brand: The case of DHL Logistics. *Procedia Econ. Finance* 23, 1079–1083. doi: 10.1016/S2212-5671(15)00457-8

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Sharma, S., and Parmar, M. (2020). Heart diseases prediction using deep learning neural network model. *Int. J. Innov. Technol. Explor. Eng.* 9, 124–137. doi: 10.35940/ijitee.C9009.019320

Signorini, A., Segre, A. M., and Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS One* 6:e19467. doi: 10.1371/journal.pone.0019467

Sinaga, K. P., and Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access* 8, 80716–80727. doi: 10.1109/ACCESS.2020.2988796

Sokolov, S. (2018). Neural network based multimodal emotion estimation. *ICAS* 2018, 4–7.

Solovyev, A., Mikheev, M., Zhou, L., Dutta-Moscato, J., Ziraldo, C., An, G., et al. (2010). "SPARK: A framework for multi-scale agent-based biomedical modeling," in *Proceedings of the 2010 spring simulation multiconference*, Orlando, FL, 1–7. doi: 10.1145/1878537.1878541

Soni, S., and Vyas, O. P. (2010). Using associative classifiers for predictive analysis in health care data mining. *Int. J. Comput. Appl.* 4, 33–37.

Stein, K., Sugar, C., Velikova, G., and Stark, D. (2007). Putting the 'Q' in quality adjusted life years (QALYs) for advanced ovarian cancer–An approach using data clustering methods and the internet. *Eur. J. Cancer* 43, 104–113. doi: 10.1016/j.ejca.2006.09.007

Tomar, D., and Agarwal, S. (2013). A survey on data mining approaches for healthcare. *Int. J. Biosci. Biotechnol.* 5, 241–266.

Tsang, I. W., Kwok, J. T., Cheung, P. M., and Cristianini, N. (2005). Core vector machines: Fast SVM training on very large data sets. *J. Mach. Learn. Res.* 6, 363–392.

Vinsnes, A. G., Harkless, G. E., Haltbakk, J., Bohm, J., and Hunskaar, S. (2001). Healthcare personnel's attitudes towards patients with urinary incontinence INFORMATION POINT: Regression analysis. *J. Clin. Nurs.* 10, 455–462. doi: 10.1046/j.1365-2702.2001.00513.x

Wang, L. (2017). Heterogeneous data and big data analytics. *Automat. Control Inf. Sci.* 3, 8–15. doi: 10.12691/acis-3-1-3

Wang, Y., and Manzie, C. (2022). Robust distributed model predictive control of linear systems: Analysis and synthesis. *Automatica* 137:110141. doi: 10.1016/j.automatica.2021.110141

Weerasinghe, K., Scahill, S. L., Pauleen, D. J., and Taskin, N. (2022). Big data analytics for clinical decision-making: Understanding health sector perceptions of policy and practice. *Technol. Forecast. Soc. Change* 174:121222. doi: 10.1016/j.techfore.2021.121222

Wermelt, J. A., and Schunkert, H. (2017). Management of arterial hypertension. *Herz* 42, 515–526. doi: 10.1007/s00059-017-4574-1

Xiong, H., Fan, C., Chen, H., Yang, Y., Antwi, C. O., and Fan, X. (2022). A novel approach to air passenger index prediction: Based on mutual information principle and support vector regression blended model. *SAGE Open* 12:21582440211071102. doi: 10.1177/21582440211071102

Yang, Z., Huang, Y., Jiang, Y., Sun, Y., Zhang, Y. J., and Luo, P. (2018). Clinical assistant diagnosis for electronic medical record based on convolutional neural network. *Sci. Rep.* 8:6329. doi: 10.1038/s41598-018-24389-w

Ye, Y., Shi, J., Zhu, D., Su, L., Huang, J., and Huang, Y. (2021). Management of medical and health big data based on integrated learning-based health care system: A review and comparative analysis. *Comput. Methods Prog. Biomed.* 209:106293. doi: 10.1016/j.cmpb.2021.106293

Yu, V. L., and Madoff, L. C. (2004). ProMED-mail: An early warning system for emerging diseases. *Clin. Infect. Dis.* 39, 227–232. doi: 10.1086/422003

Zhao, Z., Zhao, J., Song, K., Hussain, A., Du, Q., Dong, Y., et al. (2020). Joint DBN and Fuzzy C-Means unsupervised deep clustering for lung cancer patient stratification. *Eng. Appl. Artif. Intell.* 91:103571. doi: 10.1016/j.engappai.2020.103571

Zheng, B., Yoon, S. W., and Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Syst. Appl.* 41, 1476–1482. doi: 10.1016/j.eswa.2013.08.044