Big Data and Information Analytics

BDIA, 5: 58–75.
DOI: 10.3934/bdia.2020005
Received: 30 June 2020
Accepted: 15 October 2020
Published: 29 October 2020

http://www.aimspress.com/journal/BDIA

Review

Machine learning approach on healthcare big data: a review

M Supriya[1,*] and AJ Deepa[2]

[1] Research scholar, Anna University, Chennai, Tamilnadu, India
[2] Department of Computer Science and Engineering, Ponjesly Engineering College, Kanyakumari, Tamilnadu, India

* **Correspondence:** Email: smily.supriya@gmail.com.

Abstract: In the past few years, big data has flattering more dominant in healthcare, due to three major reasons, such as the huge amount of data available, expanding healthcare costs, and a target on personalized care. Big data processing in healthcare refers to generating, collecting, analyzing, and holding clinical data that is too vast or complex to be inferred by classical means of data processing methods. Big data sources for healthcare include, the Internet of Things (IoT), Electronic Medical Record/Electronic Health Record (EMR/EHR) contains patient's medical history, diagnoses, medications, treatment plans, allergies, laboratory and test results, genomic sequencing, Medical Imaging, Insurance Providers and other clinical data. This paper discusses different machine learning algorithms that were applied to various healthcare data. Also, the challenges of processing, handling big data, and their applications. The scope of the paper is to elaborate on the application of machine learning algorithms and the need for handling and utilizing big data from a different perspective.

Keywords: big data; Electronic Health Record (EHR); machine learning (ML)

1. Introduction

1.1. Big data in healthcare

Big data originally illustrate the volume, velocity, and variety of data that comes from various data production time by health care providers that contain information relevant to a patient's care, including to demographics, diagnoses, medical procedures, medications, vital signs, immunizations, laboratory results, and radiology images. With the development of medical data-collecting, electronic health sources such as sensor devices, streaming machines, and high throughput instruments are accumulating more. This healthcare big data is used for various applications like diagnosis, drug discovery, precision medicine, prediction of disease, etc., Chandu Thota et al. [1]. Big data has been playing a crucial role in a variety of environments such as healthcare, scientific research, industry, social networking, and public administration [1]. Big data can be classified by 5Vs as follows:

Volume: The big volume undoubtedly represents big data. To process the vast amount of data like text, audio, video, and large size images the traditional data processing platforms and techniques have to be strengthened. In a healthcare database, information such as personal information, radiology images, personal medical records, genomics, and biometric sensor readings, etc. are being included gradually. All this information cumulatively increases the size and complexity of the database to a great extent.

Velocity: The rate at which the data is generated truly represents big data. The data burst of social media has changed and causes variety in data. Most of the health data are in form of paper files, X-ray films, and scripts and the growth rate of such data is now immensely increasing.

Variety: The variety of data undoubtedly represents big data. For example, the different types of data formats include database, excel, and CSV, which can be stored in a plain text file. In existence, health data also are structured, unstructured, and semi-structured. An example of structured information is clinical data, instead, data such as doctor notes, paper prescriptions, office medical records, images, and radiograph films are unstructured or semi-structured.

Veracity: This veracity of data truly represents bug data. It represents data understandability, not data quality. In healthcare data, the veracity feature gives information certificate about correct diagnosis/treatment/prescription/procedure/outcomes, etc.

Value: The value of data truly represents big data. The benefits and costs of analyzing and collecting big data are more important things when doing big data analytics. In healthcare, the creation of value for patients should determine the rewards for all other actors in the system. Obtaining high value for patients must become the underlying goal of healthcare delivery.

1.2. Applications of healthcare big data

Big data applications present new opportunities to diverse new knowledge and create innovative methods to improve the quality of health care. Some of the important applications are public health, clinical operations, research and development, patient profile analysis, evidence-based medicine, remote monitoring, etc.. The frameworks and storage system for healthcare big data are illustrated below.

Internet of Things (IoT)—enabled devices has made remote monitoring in the healthcare sector possible, unleashing the potential to keep patients safe and healthy, and empowering physicians to deliver superlative care. It has also increased patient engagement and satisfaction as interactions with doctors have become easier and more efficient. Furthermore, remote monitoring of patient's health helps in reducing the length of hospital stay and prevents re-admissions. IoT also has a major impact on reducing healthcare costs significantly and improving treatment outcomes.

Digital epidemiology has defined epidemiology that uses digital methods from data collection to analyze data. It enhances traditional epidemiological studies, such as case records, case reports, ecological studies, cross-sectional studies, case-control studies, cohort studies, randomized controlled trials, and systematic reviews and meta-analysis. It also makes use of data sources that are originally collected and/or generated for health and non-health-related purposes.

FRED (a framework for reconstructing epidemiological dynamics) is an open-source framework for epidemic modeling, rather than a model of a particular infectious disease. Geographic regions are used in FRED to represent every individual as an agent. Each agent has a set of sociodemographic characteristics and daily behaviors that include age, sex, employment status, occupation, and household location, and membership in a set of social contact networks. This synthetic population data is used to model the disease outbreak in FRED.

NoSQL database is used to store a huge volume of data in a distributed manner. A NoSQL database does not follow any relational schema. NoSQL databases can be classified into four types such as key-value stores, column family database stores, document stores, and graph stores. The key-value stores the data based on key-value pairs and used for small applications. Secondly, the column family database stores huge data into rows as a collection of columns. Thirdly, the document stores huge data related to a document format and is used for semi-structured data. Finally, the graph stores database contains edges between noded with map and query relationships.

Big data may come from various sources such as healthcare, CCTV surveillance, Social networking, machine-generated data, and sensor data. The type of data may be structured and unstructured. For exploiting big data there is a need for big data architecture. Data in the order of 100s of GB does not require any kind of architecture but when it goes beyond this there comes big data architecture. Investments in Big Data Project and have multiple sources of big data need big data architecture. Big data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems [2]. Big data architecture is designed to handle the following types of work:

- Batch processing of big data sources.
- Real-time processing of big data.
- Predictive analytics and machine learning.

The four import layers of big data architecture include, (i) Data Source Layer, (ii) Ingestion Layer, (iii) Manage Layer, (iv) Analyze and Visualize Layer. Some of the state-of-the-art are architectures are Lambda architecture, the service-on-line-index-data (SOLID) architecture, semantic-based architecture for heterogeneous multimedia retrieval, large-scale security monitoring architecture, modular software architecture, etc.. Every architecture has its advantages and disadvantages.

1.3. Big data analytics in healthcare

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured, and unstructured data, from different sources, and in different sizes from terabytes to zettabytes. Big data analytics in healthcare covers integration and analysis of a large amount of complex heterogeneous data such as various—omics data (genomics, epigenomics, transcriptomics, proteomics, metabolomics, interactomics, pharmacogenomics, diseasomics), biomedical data, and HER (Electronic Health Record). Among these EHR places a major role and is also adopted in different countries nowadays. The main objective of HER is to gain actionable big data insights from health workflow.

EHR is an electronic (digital) collection of medical information about a person that is stored on a computer. An EHR includes information about a patient's health history, such as diagnoses, medicines, tests, allergies, immunizations, and treatment plans. EHR can be seen by all healthcare providers who are taking care of a patient and can be used by them to help make recommendations about the patient's care. EHR is also called as EMR (Electronic Medical Record). Every second, dozens of terabytes of data are generated and accumulated from various sources, e.g., internet browsing, social networks, mobile transactions, online shopping, and many others. Indeed, the big data paradigm has taken an expended shape, and the abundance of such structured and unstructured data has made it possible to be open to new perspectives. These new sources of data increase the chances of understanding one's behavior and motivations, identifying instant signals and triggers for someone's interest in a specific offer or product. Getting meaningful insights from abundant and varied amounts of data helps to understand and extract hidden information, which can be used and exploited for the proper improvement of the users' experiences [3]. There is free, open-source, or paid EHR systems, which have a major impact on the development process of any medical organization. The era of big data is driving researchers to think and pick for a wide vision facing the future. A clinical decision support system (CDSS) is an application that analyzes data to help healthcare providers make decisions, and improve patient care. A CDSS focuses on using knowledge management to get clinical advice based on multiple factors of patient-related data. Clinical decision support systems enable integrated workflows, provide assistance at the time of care, and offer care plan recommendations. Physicians use a CDSS to diagnose and improve care by eliminating unnecessary testing, enhancing patient safety, and avoiding potentially dangerous and costly complications. The applications of big data in healthcare include, cost reduction in medical treatments, eliminate the risk factors associated with diseases, prediction of diseases, improves preventive care, analyzing drug efficiency. Some challenging tasks for the healthcare industry are: (i) how to decide the most effective treatment for a particular disease? (ii) How certain policies impact the outlay and behavior? (iii) How does the healthcare cost likely to rise for different aspects of the future? (iv) How the claimed fraudulently can be identified? (v) Does the healthcare outlay vary geographically? [4]. These challenges can be overcome by utilizing big data analytical tools and techniques. There are four major pillars of quality healthcare. Such as real-time patient monitoring, patient-centric care, improving the treatment methods, and predictive analytics of diseases. All these four pillars of quality healthcare can be potently managed by using descriptive, predictive, and prescriptive big data analytical techniques.

1.4. Machine learning for healthcare big data

Machine Learning is a sub-area of artificial intelligence, whereby the term refers to the ability of IT systems to independently find solutions to problems by recognizing patterns in databases. Machine Learning enables IT systems to recognize patterns based on existing algorithms and data sets and to develop adequate solution concepts. Therefore, in machine learning, artificial knowledge is generated based on experience. In machine learning, statistical and mathematical methods are used to learn from data sets. There are two main systems namely symbolic approaches and sub-symbolic approaches. While symbolic systems are, for example, propositional systems in which the knowledge content, i.e., the induced rules and the examples are explicitly represented, sub-symbolic systems are artificial neuronal networks. These work on the principle of the human brain, whereby the knowledge contents are implicitly represented. The critical issues of machine learning for big data are large scale of data, different types of data, high speed of streaming data, uncertain and incomplete data [5]. The three main types of machine learning are supervised, unsupervised, and reinforcement learning. The following Table 1 illustrates the different types of machine learning with categories and examples.

Table 1. Different machine learning techniques.

SL.No.	Learning Types	Categories	Examples
1	Supervised learning	Learning problems	Support Vector Machine (SVM), Na ïve Bayes (NB), Neural Networks (NN)
2	Unsupervised learning	Learning problems	K-means, Gaussian mixture model, Dirichlet process mixture model
3	Reinforcement learning	Learning problems	Q-learning, R-learning TD learning
4	Semi-supervised learning	Hybrid learning problems	speech analysis, internet content classification, protein sequence classification
5	Self-supervised learning	Hybrid learning problems	generative adversarial networks (GANs), autoencoders
6	Multi-instance learning	Hybrid learning problems	medical image classification, molecule activity
7	Inductive learning	Statistical inference	weather prediction
8	Transductive learning	Statistical inference	K-Nearest Neighbors (KNN)
9	Online learning	Learning techniques	gradient descent
10	Transfer learning	Learning techniques	image classification
11	Ensemble learning	Learning techniques	stacking, bagging
12	Deep learning	Learning techniques	automatic speech recognition, medical image analysis

First, supervised learning describes a class of problem that involves using a model to learn a mapping between input examples and the target variable, so it requires training with labeled data which has inputs and desired outputs. Second, unsupervised learning describes a class of problems that involves using a model to describe or extract relationships in data, so it does not require labeled training data and the environment only provides inputs without desired targets Third, reinforcement learning describes a class of problems where an agent operates in an environment and must learn to operate using feedback. Some other learnings which is much needed for solving the big data problems are:

- Representation learning
- Active learning

- Deep learning
- Transfer Learning
- Distributed and parallel learning
- Kernel-based learning

Among all of these deep learning places an important role in the field of healthcare. Deep learning is a type of machine learning, and deep learning solves the problems that were unsolvable with machine learning. Deep learning uses Neural Networks to increase computational work and provide accurate results. Deep learning is assisting medical professionals and researchers to discover the hidden opportunities in data and to serve the healthcare industry better.

1.5. Deep learning for healthcare big data

DNNs (Deep Neural Networks) is the state-of-the-art in machine learning and big data analytics, being used in a large number of applications, ranging from defense and surveillance to human-computer interaction and question answering systems. DNN architecture comes in many different forms, which can be grouped into three general families. They are Feed-forward Neural Network, Convolution Neural Networks (CNN), and Recurrent Neural Networks (RNN) [6]. Deep learning in healthcare provides doctors the analysis of any disease accurately and helps them treat them better, thus resulting in better medical decisions. Deep learning technologies can be applied to hospital management information systems to achieve: lower cost, fewer hospital stays, and its length, control of insurance fraud, change detection in disease patterns, high-quality healthcare, and better efficiency of medical resource allocation. In the following paragraphs, several application examples are presented according to the different nature of biomedical information: biomedical images, biomedical time signals, and other biomedical data like those from laboratory results, genomics, and wearable devices.

1.5.1. Drug discovery

Deep learning in healthcare helps in the discovery of medicines and their development. The technology analyzes the patient's medical history and provides the best treatment for them. Moreover, this technology is gaining insights from patient symptoms and tests.

1.5.2. Medical imaging

Medical imaging techniques such as MRI scans, CT scans, ECG, are used to diagnose dreadful diseases such as heart disease, cancer, brain tumor. Hence, deep learning helps doctors to analyze the disease better and provide patients with the best treatment.

1.5.3. Insurance fraud

Deep learning is used to analyze medical insurance fraud claims. With predictive analytics, it can predict fraud claims that are likely to happen in the future. Moreover, deep learning helps the insurance industry to send out discounts and offers to their target patients.

1.5.4. Alzheimer’s disease

Alzheimer’s is one of the significant challenges the medical industry faces. A deep learning technique is used to detect Alzheimer’s disease at an early stage.

1.5.5. Genome

A deep learning technique is used to understand a genome and help patients get an idea about the disease that might affect them. Deep learning has a promising future in genomics, and also the insurance industry. Cells cope uses deep learning techniques and helps parents to monitor the health of their children through a smart device in real-time, thus minimizing frequent visits to the doctor. Deep learning in healthcare can provide doctors and patients with astonishing applications, which will help doctors to make better medical treatments.

2. Literature survey

2.1. Diabetes prediction using machine learning

Diabetes a chronic disease that occurs when the pancreas no longer able to make insulin, or when the body cannot make good use of the insulin it produces. Insulin is a hormone made by the pancreas that acts as a key to let glucose from the food we eat pass from the bloodstream into the cells in the body to produce energy. All carbohydrate foods are broken down into glucose in the blood. Insulin helps glucose get into the cells. There are three main types of diabetes-type1, type2 and gestational. Among these type 2 diabetes is more common. It occurs in adults and accounts for around 90% of all diabetes cases. When we have type 2 diabetes, our body does not make good use of the insulin that it produces. The cornerstone of type 2 diabetes treatment is a healthy lifestyle, including increased physical activity and a healthy diet. However, over time most people with type 2 diabetes will require oral drugs and/or insulin to keep their blood glucose levels under control.

Kai Hwang et al. proposed a Convolution Neural Network-based multimodal disease risk prediction (CNN-MDRP) algorithm which is applicable for big data [7]. The disease risk model is obtained by the combination of structured and unstructured features and the accuracy is analyzed. It gave better accuracy than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm. The data used in the work were taken from the hospital in China which includes EHR, medical image data, and gene data. The data focus on inpatient department data mainly composed of structured and unstructured text data. The stochastic gradient descent algorithm specially used for big data applications is used to train the parameter in the CNN-MDRP algorithm. The transformation of using sophisticated technologies by healthcare provides to gain insights from clinical datasets and make informed decisions had changed by big data analytics with the help of the Hadoop framework. Effective healthcare management can be achieved by providing effective data-driven services to people by predicting their needs. Big data analytics in healthcare is defined as the ability to acquire, store, process, and analyze a large volume of health data in various forms, and deliver meaningful information to users, which allow them to discover business values and insights promptly on time. The various big data analytical techniques include data mining, machine learning, statistical analysis, and visualization.

Ya Zhang and Tao Zheng proposed a semi-automated framework based on machine learning using an EHR database which is big data [8]. The data was collected from 15 local EHR systems were automatically deposited into the centralized repository every 24 hours in China. The framework was based on a supervised learning algorithm. The raw EHR is often unstructured and sparse so to properly structure, feature engineering was needed. Some 16 features were extracted and constructed to be used in the machine learning framework. The three machine learning algorithms Random Forest, Logistic Regression, and Ada Boost were used and obtained better accuracies. The algorithms also optimize the filtering criteria to improve recall at the same time keeping low false-positive rates.

Gang Luo proposed an automatically analyzing machine learning prediction results. Predictive modeling is a process that uses data mining and probability to forecast outcomes [9]. Each model is made up of several predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation, or it may be a complex Neural Network, mapped out by sophisticated software. As additional data becomes available, the statistical analysis model is validated or revised. Predictive analytics can support population health management, financial success, and better outcomes across the value-based care sequence. Instead of simply presenting information about past events to a user, predictive analytics estimates the likelihood of a future outcome based on patterns in the historical data. The electronic medical record data set from the Practice Fusion diabetes classification competition containing patient records from all 50 states in the United States were utilized in this work and illustrated the method of predicting type 2 diabetes diagnosis within the next year. The prediction was done using two models, one for prediction and another for the explanation. The first model is used only for making predictions and aims at maximizing accuracy without being concerned about interpretability. It can be any machine learning model and arbitrarily complex. The second model is a rule-based associative classifier used only for explaining the first model's results without being concerned about its accuracy.

Ya Zhang et al. proposed a machine learning-based framework to identify type2 diabetes using EHR [10]. This work utilized 3 years of EHR data. The data was stored in the central repository, which has been managed by the District Bureau of Health in Changning, Shanghai since 2008. The EHR data generated from 10 local EHR systems are automatically deposited into the centralized repository hourly. The machine learning models within the framework, including K-Nearest-Neighbors, Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression. Also, the identification performance was higher than the state-of-the-art algorithm.

Dilip Singh Sisodia et al. proposed the Prediction of Diabetes using Classification Algorithms [11]. The motivation of this work was to design a model that can prognosticate the likelihood of diabetes in patients with maximum accuracy. Three machine learning classification algorithms such as SVM, Naive Bayes, and Decision Tree were used. The data were taken from the Pima Indians Diabetes Database (PIDD) which is sourced from the UCI machine learning repository. Among all the three machine learning algorithms that were implemented Naïve Bayes gave better accuracy.

2.2. Heart disease and sepsis prediction using machine learning

Heart failure (HF) is a chronic, progressive condition in which the heart muscle is unable to pump enough blood to meet the body's needs for blood and oxygen. The heart can't keep up with its workload. Heart failure mortality is similar to or even higher than that due to various cancers. It is usually associated with disease progression, though sudden death has also been reported as a frequent cause of mortality. Electronic Health Record (EHR) contain patient diagnostic records, physician records, and records of hospital departments. For heart failure, we can obtain bulk unstructured data from the EHR time series. By analyzing and mining these time-based EHR, we can determine the links between diagnostic events and predict when a patient will be diagnosed. However, EHR data are highly convoluted, given the structure and span of information captured (spanning provider behavior, care usage, treatment pathways, and patient disease case) and uneven sampling frequency.

Jyotishman Pathak et al. proposed the performance of SHFM (Seattle Heart Failure Model) using EHR at Mayo Clinic and desired to develop a risk prediction model using machine learning techniques that apply routine clinical care data [12]. The Seattle Heart Failure Model (SHFM) is one of the most popular models to calculate HF survival risk that uses different clinical variables to predict HF prognosis and also integrates the impact of HF therapy on patient outcomes. The data used in this work was taken from a cohort of 119,749 Mayo Clinic patients between 1993–2013 with research authority to access EHR data, they found 5044 patients with a diagnosis of HF after applying some specific criteria and excluding the number of patients due to incomplete data. The results offer that (i) HF survival models built on EHRs are more accurate than the SHFM, (ii) integrating co-morbidities into the HF survival analysis prediction models improve the accuracy of the models, and (iii) there are potential hidden interactions between diagnoses history of the patient, co-morbidities, and survival risk. The models were built using multiple different machine learning algorithms and the results showed that Logistic Regression and Random Forest return more accurate classifiers.

Andy Schuetz et al. proposed Recurrent neural network (RNN) models using gated recurrent units (GRUs) were adapted to detect relations among time-stamped events (e.g., disease diagnosis, medication orders, procedure orders, etc..) with a 12 to 18-month observation window of cases and controls [13]. The RNN provides a nonlinear growth in model generalization and more scalability than many of the traditional methods. Data used in this work were from a health system's EHR on 3884 incident HF cases and 28,903 controls, identified as primary care patients, between May 16, 2000, and May 23, 2013. To represent clinical events in EHR data as computable event sequences, the one-hot vector format was adopted, often used for NLP (natural language processing) tasks. Model performance metrics were compared to regularized Logistic Regression, Neural Network, Support Vector Machine, and K-Nearest Neighbor classifier approaches. RNN models are naturally appropriate to temporal sequenced data, and several variants have been developed for sequenced features.

Shulong Zhang et al., proposed a predictive model frame-work for HF diagnosis using LSTM (long short term memory) methods [14]. One-hot encoding and word vectors are used to model the diagnosis events and predicted HF events using the basic principles of an LSTM network model. LSTMs have an edge over conventional feed-forward Neural Networks and RNN in many ways. This is because of their property of selectively remembering patterns for long durations of time. LSTM adds three gates to the basics of the original RNN network, an input gate, a forget gate, and an output gate. This work used the Electronic Health Record (EHR) data from real-world datasets related to congestive heart disease to experiment. First, the records of patients who had HF disease for more than four years were extracted. The dataset consists of two parts: dataset A and dataset B. Dataset A contains the diagnostic records of 5000 patients who have been diagnosed with HF. The records mainly include recording times, diagnosis events, and diagnosis times. Dataset B contains the diagnostic records for 15,000 patients who have not been diagnosed with heart HF. The records mainly include recording times and diagnostic events. Compared to popular methods such as

Logistic Regression (LR), Random Forest (RF), and AdaBoost, the LSTM method express admirable performance in the prediction of the heart failure diagnosis.

Sepsis is a possibly life-threatening condition caused by the body's response to an infection. The body normally releases chemicals into the bloodstream to fight infection. Sepsis occurs when the body’s response to these chemicals is out of balance, triggering changes that can batter multiple organ systems. If sepsis progresses to septic shock, blood pressure drops adequately. This may lead to death. Sepsis and its combined syndromes are among the principal causes of worldwide morbidity and mortality and are responsible for laying an enormous cost hinder on the healthcare system.

Jana Hoffman et al. proposed a machine-learning classification system that uses multivariable combinations of easily obtained patient data (vitals, peripheral capillary oxygen saturation, Glasgow Coma Score, and age), to predict sepsis using the retrospective Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III dataset, restricted to intensive care unit (ICU) patients aged 15 years or more [15]. The sepsis prediction model utilizes the Insight algorithm which uses only the EHR-entered components of the MIMIC-III set and does not require real-time waveform data or the analysis of free-text notes. A variety of data from the MIMIC-III dataset were collected to define sepsis onset and calculate the *InSight* score, as well as other scores such as MEWS(modified early warning score) and SOFA (sequential (sepsis-related) organ failure assessment)for comparison. All data are extracted from the MIMIC-III set using custom PostgreSQL (PostgreSQL Global Development Group) queries. The training and testing process for the *InSight* prediction system consists of 4 stages: Data partitioning, feature construction, classifier training, and classifier testing *InSight* produced superior classification performance compared with the alternative scores as measured by area under the receiver operating characteristic curves (AUROC) and area under precision-recall curves (APR).

Arjun K. Venkatesh et al. proposed a machine learning approach to predict in-hospital mortality of ED(emergency department) patients with sepsis which a big data-driven approach [16]. Data were obtained from four EDs over 12-month (October 2013 to October 2014). All EDs were part of a single health care system: the first ED was an urban, academic, Level I trauma center with an annual census of over 85,000 patients; the second ED was an urban community-based, academic Level II trauma center with an annual census of over 70,000 patients; the third ED was a community-based center with an annual census over 75,000 patients; and the fourth ED was a suburban, free-standing ED with annual census over 30,000 patients. A Random Forest (RF) model was designed using over 500 clinical variables from data available within the EHR of four hospitals to predict in-hospital mortality. The machine learning prediction model was then compared to a classification and regression tree (CART) model, Logistic Regression model, and previously developed prediction tools on the validation data set using the area under the receiver operating characteristic curve (AUC) and chi-square statistics.

2.3. Cancer prediction using machine learning

Cancer refers to any one of a large number of diseases characterized by the development of abnormal cells that divide uncontrollably and have the ability to infiltrate and destroy normal body tissue. Cancer often can spread all over our bodies. Cancer is the second-leading cause of death in the world. But survival rates are improving for many types of cancer, thanks to improvements in cancer screening and cancer treatment. Cancer is not just one disease. There are many types of cancer. It’s not just one disease. Cancer can start in the lungs, the breast, the colon, or even in the blood. Cancers are alike in some ways, but they are different in the ways they grow and spread. Some cancers grow and spread fast, others grow more slowly. They also respond to treatment in different ways. Some types of cancer are best treated with surgery; others respond better to drugs called chemotherapy.

Susan E. Clare et al. proposed a novel concept-based filter and a prediction model to detect local recurrences using EHR in breast cancer patients [17]. Machine learning and NLP are used to identify these recurrences. Patients diagnosed with breast cancer at Northwestern Memorial Hospital between 2001 and 2015 were identified by (International Classification Diseases) ICD9 codes were used in this work. 50 progress notes were reviewed and extract partial sentences were extracted that indicate breast cancer local recurrence. Then using MetaMap these partial sentences were processed to obtain a positive set of concepts called features. MetaMap is a highly configurable application developed by the Lister Hill National Center for Biomedical Communications at the National Library of Medicine (NLM) to map biomedical text to the UMLS (Unified Medical Language System) Metathesaurus or, equivalently, to identify Metathesaurus concepts referred to in English text. These features combined with the number of pathology reports recorded for each patient are used to train a Support Vector Machine to identify local recurrences. Compared to other baseline classifiers this model achieved the best AUC.

Shruti Garg et al. proposed a machine learning model by utilizing the hyperparameters for Breast cancer prediction [18]. This work applied six machine learning algorithms such as K-Nearest Neighbour (KNN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and deep learning using ANN. Deep learning used Neural Networks along with Adam Gradient Descent cost function. Adam is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure to update network weights iterative based on training data. The Adam learning combines adaptive gradient algorithm (AdaGrad) and root mean square propagation (RMSProp). The data used were taken from Wisconsin Breast Cancer Dataset (WBCD), which is already classified as malignant and benign. The dataset consists of 30 features computed using fine-needle aspiration (FNA) of the breast mass. The accuracy attained by Deep learning is higher than other machine learning algorithms used.

Alkhawaldeh et al. proposed a Multi-Layer Perceptron (MLP) model for Breast cancer prediction [19]. The tuned MLP model was divided into two categories, one checked the reduction of the extracted feature size, while the other inspected the enhancement of the classification power. The dataset was tabbed from WDBC which includes 569 instances and 32 attributes. The Tuned MLP is based on reducing the size of the extracted feature. The different search methods were used with different attribute evaluators and then voted of the best relevant features where it reduced features from 31 attributes to only 4 attributes of the WDBC dataset. The proposed Tuned MLP confirms the use of the Grid Search method to find the optimal hyperparameters of a model which results in the most 'accurate' predictions as compared to the basic MLP.

Seenivasagam et al. proposed a machine learning method to classify lung cancer stages over big data [20]. The machine learning algorithms are combined with Apache Spark design for active classification to handle the big data. Apache Spark is a lightning-fast cluster computing designed for fast computation. It was built on top of Hadoop Map Reduce and it extends the Map Reduce model to accurately use more types of computations which adds Interactive Queries and Stream Processing. The data used were sputum color images collected from the microscope lab. Before handled in big data apache spark framework, images at 330 × 330 processes using the map-reduce framework.

Since these images contain all types of cells, the work target on dividing cells related to the lungs such as eosinophills, bronchial mucus, squamous carcinoma, and so on. The method applies T-BMSVM (threshold-based Support Vector Machine with binary and multi-class) to classify the stages of lung cancer with high accuracy.

Oludayo O. Olugbara et al. proposed an NSCLC (non-small cell lung cancer) prediction model to predict lung cancer [21]. The Artificial Neural Network (ANN) ensemble with a histogram of oriented gradient (HOG) genomic features are used to predict the class of the genes. HOG, or Histogram of Oriented Gradients, is a feature descriptor that is often used to extract features from image data and able to provide the edge direction as well. This is done by extracting the gradient and orientation of the edges. The normal nucleotides of the three genes in this study were extracts from the collaborative consensus coding sequence (CCDS) archive in the National Centre for Biotechnology Information (NCBI) repository. Mutation datasets for each of the genes were collected from the Integrated Genomic Database of non-small cell lung cancer (IGDB.NSCLC), which is an online corpus dedicated to the archiving of NSCLC genetic defects. The MLP-ANN (Multilayer Perceptron-Artificial Neural Network) ensemble classifier attained better accuracy than other classifiers.

Tabitha Peter et al. proposed a machine learning method for disease classification with lung cancer screening image data [22]. A collection of linear, nonlinear, and ensemble predictive classifying models, along with several feature selection methods, was used to classify the binary outcome of malignant or benign status. Elastic net and Support Vector Machine combined with either a linear combination or correlation feature selection method. The data used were taken from 200 CT scans of the lungs of patients at the University of Iowa Hospital. Pathology and radiology reports were reviewed to identify an analysis set of patients who met eligibility criteria of having (a) a solitary lung nodule (5–30 mm) and (b) a malignant nodule confirmed on histopathology or a benign nodule confirmed on histopathology or by size stability for at least 24 months. The 416 radiomic features which were available for this investigation quantified nodule characteristics from CT images acquired from a variety of scanner protocols through the University of Iowa Hospital. Radiomics is a method that extracts large amounts of features from radiographic medical images using data-characterization algorithms. The use of radiomic biomarkers with machine learning methods is a promising diagnostic tool for tumor classification.

Wazir Muhammad et al. proposed a neural network model to predict stratify ovarian cancer risk using personal health data [23]. The risk classification was high, medium, and, low. Ovarian cancer is a type of cancer that begins in the ovaries. The female reproductive system contains two ovaries, one on each side of the uterus. Ovarian cancer often goes undetected until it has spread within the pelvis and abdomen. At this late stage, it is more difficult to treat. Early-stage ovarian cancer, in which the disease is confined to the ovary, is more likely to be treated successfully. The neural network was designed with 0 to 3 hidden layers with 4–12 neurons per hidden layer. The data used were from two different sources: the National Health Interview Survey (NHIS) and Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO). AS a result, depending on the economic cost and harms/benefit trade-off, less conservative boundaries could be selected, resulting in more people with cancer classified as high risk. The following Table 2 illustrates the accuracy and AUC measures of various machine learning techniques.

Table 2. Accuracy/AUC of machine learning algorithms using healthcare big data.

Authors	Year	Dataset	Machine learning techniques	Accuracy(%)/AUC
Yixue Hao et al. [7]	2017	Real-life hospital data	Convolution Neural Networks (CNNs)	94.8%
Ya Zhang et al. [8]	2017	EHR	Random forest (RF) Logistic regression (LR) Ada Boost	0.897 0.882 0.90
Gang Luo [9]	2017	EHR	Associative Classifier	0.884
Tao Zheng et al. [10]	2017	EHR	K-Nearest Neighbour (KNN) Logistic regression (LR) Naive Bayes (NB)	0.91 0.92 0.98
Deepti Sisodia et al. [11]	2018	Pima Indians Diabetes Database	Naive Bayes (NB)	76.30%
Maryan Panahiazar et al. [12]	2015	EHR	Logistic regression (LR)	0.81
Jimeng Sun et al. [13]	2016	EHR	Recurrent Neural Networks (RNNs)	0.883
Bo Jin et al. [14]	2018	EHR	Long Short Term Memory (LSTM)	0.6484
Thomas Desautels et al. [15]	2016	EHR	Elastic net regularization	0.88
Andrew Taylor et al. [16]	2015	Real-life hospital data	Random Forest (RF)	0.860
Zexian Zeng et al. [17]	2018	HER	Support Vector Machine (SVM)	0.93
Puja Gupta et al. [18]	2020	Wisconsin Breast Cancer Dataset (WBCD)	Deep Learning-Artificial Neural Network (DL-ANN)	98.24%
Bassan AI shargabi et al. [19]	2019	Wisconsin Diagnosis Breast Cancer Dataset (WDBC)	Multilayer Perceptron (MLP)	97.70%
Sujitha et al. [20]	2020	Sputum cell images	Support Vector Machine (SVM)	86%

Continued on next page

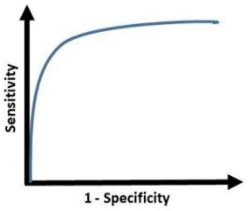
Authors	Year	Dataset	Machine learning techniques	Accuracy(%)/AUC
Emmanuel Adetita et al. [21]	2015	Collaborative consensus coding sequence (CCDS)	Multilayer Perceptron-Artificial Neural Network (MLP-ANN)	95.90%
Tabitha Peter et al. [22]	2019	Imaging biomarkers from CT scans of lung nodules	Elastic net	0.747
Ying Liang et al. [23]	2019	National Health Interview Survey (NHIS) and Prostate, Lung, Colorectal, and Ovarian Cancer (PLCO) Screening Trial	Neural Network (NN)	0.71

2.4. Big data classification and measures used to assess machine learning models

Classification is one of the data mining techniques that classify unstructured big data into the structured class and groups and it aids the user for knowledge discovery and future ideas. Classification provides creative decision making. There are two phases in classification, first is the learning process phase in which huge training data sets are supplied and inquiry takes place then rules and patterns are created. Then the execution of the second phase starts that is evaluation or test of data sets and archives the accuracy of a classification pattern. There was an approach that creates a binary search tree (BST) to be used following by the KNN to speed up the big data classification [24]. This approach is based on finding the furthest pair of points (diameter) in a data set and used to sort the examples of the training data set into a BST. At each node of the BST, the furthest-pair is found and the examples located at that particular node are further sorted based on their distances to these local furthest points. The created BST is then searched for a test example to the leaf and used to classify the test example using the KNN classifier. Yi-Hung Huang et al. proposed an Improved Cat Swarm Optimization (ICSO) algorithm for big data classification [25]. Feature selection was done using this ICSO. Feature selection is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset, but a dimensionality reduction method does so by creating new combinations of attributes, whereas feature selection methods include and exclude attributes present in the data without changing them. Two methods were used to improve Cat Swarm Optimization (CSO). In the first step of seeking mode, rather than randomly generating N copies of the existing cat (all of which are candidate solutions), a crossover operation was used to generate candidate solutions. In the second step of seeking mode, to change the position of the cats the original method was replaced. Using term frequency-inverse document frequency (TF-IDF) with ICSO for feature selection is more accurate than using TF-IDF alone, in the case of big data text classification.

Yilin Bel et al. proposed an improved KNN algorithm and compare it with the traditional KNN algorithm [26]. The classification is performed in the query instance neighborhood of the conventional KNN classifier, and weights are assigned to each class. The algorithm considers the class distribution around the query instance to ensure that the assigned weight does not negatively affect the outliers. The improved KNN algorithm based on cluster denoising and density cropping remove the shortcoming in the traditional KNN in processing larger data sets. Yihong Li et al. proposed a Radial Basis Function (RBF) Neural Network classification algorithm based on manifold analysis and nearest neighbor propagation (AP) algorithm for classifying health big data [27]. The health big data has the characteristics of redundancy, polymorphism, and incompleteness, so the traditional supervised learning algorithm must carry on the preprocessing and feature selection before training the classifier. The manifold analysis is to analyze the data from the perspective of observation space. It is to cluster the data set based on the idea of the neighborhood and the threshold of the gap between classes and then delete the clusters with fewer samples in the class, which will eliminate the influence of some isolated points on the AP algorithm to a certain extent. The data used were 3231 cases of coronary heart disease, 9628 cases of diabetes, 5628 cases of bronchial tuberculosis data set provided by the city health and Family Planning Bureau, and some municipal hospitals. The classification accuracy of these three data sets is more than 85%. Machine learning techniques applied to medical and health big data can develop actionable insights, from improving upon patient risk score systems to predicting the onset of disease, to establishing hospital operations [28]. To evaluate the performance of machine learning techniques there is a need for common performance measures, to find how good the developed model will function with the applied ML algorithms. The following Table 3 details the common performance measures used to evaluate ML models.

Table 3. Performance measures of machine learning.

SL.No.	Performance measure	Definition	Formula or plot
1	Accuracy	The number of correct classifications made by a model (true positives and true negatives) divided by the total number of predictions made	$A=\frac{TP+TN}{TP+FP+TN+FN}$
2	Calibration	A measure of how closely predicted probabilities for an outcome match the observed outcome in test data, e.g., the Brier score	$\text{Brier score}=1/N\sum_{i=1}^{N}(p_i-o_i)$
3	Discrimination	A measure of how well a model discriminates between randomly selected true positive cases and true negative cases, usually measured as the area under the receiver operator curve(AUC)	

Continued on next page

SL.No.	Performance measure	Definition	Formula or plot
4	Negative predictive value	The total number of correct negative classifications made (true negatives) divided by the total number of negative classifications made (true negatives and false negatives)	$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$
5	Precision (also called positive predictive value)	The total number of correct positive classifications made (true positives) divided by the total number of positive classifications made (true positives and false positives)	$\mathrm{P\ or\ PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$
6	Recall (also called sensitivity or the true positive rate)	The total number of correct positive classifications made (true positives) divided by the number of positive class members in the data (true positives and false negatives)	$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
7	Specificity (also called the true negative rate)	The total number of correct negative classifications made (true negatives) divided by the number of negative class members in the data (true negatives and false positives)	$\mathrm{S} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$

3. Conclusions and future scope

The role of big data in healthcare is one where we can build better health profiles and better predictive models around individual patients so that we can better diagnose and treat disease. One of the main limitations with healthcare today and in the pharmaceutical industry is the understanding of the biology of disease. Big data comes into play around aggregating more and more information around multiple scales for what constitutes a disease from the DNA, proteins, and metabolites to cells, tissues, organs, organisms, and ecosystems. Those are the scales of the biology that we need to be modeling by integrating big data. In this paper, we discussed the applications, processing, and handling of big data using various machine learning techniques. Also, the measures used to evaluate the performances of the machine learning models are based on big data. Machine Learning also helps in effective decision-making by applying different techniques to predict diseases and timely diagnoses which can affect the health of a patient in a positive way. The information can be predicted in advance and diseases can be prevented at an early stage. Machine learning allows building models by using various algorithms to help and predict variables, keeping the accuracy perfect. The rapid enactment of EHR has created a wealth of new data about patients, which is a bonanza for improving the understanding of human health. In the future, we will see the rapid, broad implementation and use of big data using machine learning across the healthcare industry and the healthcare organization.

As big data analytics becomes more common, concerns such as safeguarding security, establishing standards, guaranteeing privacy, and governance, and constantly improving the tools and technologies will amass attention. Big data analytics and applications in healthcare are at an amorphous stage of development, but rapid advances in platforms and tools can accelerate their evolving process.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. Manogaran G, Lopez D, Thota C, et al. (2017) Big data analytics in healthcare internet of things, In: *Innovative Healthcare Systems for the 21st Century*, Springer, Cham, 263–284.
2. Lopez D and Manogaran G, (2017) A survey of big data architectures and machine learning algorithms in healthcare. *Int J Biomed Eng Technol* 25: 182.
3. Khamlichi KY, Chaoui NEH and Khennou F, (2018) Improving the use of big data analytics within electronic health records: A case study based open HER. *Procedia Compute Sci* 127: 60–68.
4. Sharma M, Kaur P and Mittal M, (2018) Big data and machine learning-based secure healthcare framework. *Procedia Compute Sci* 132: 1049–1059.
5. Xu Y, Qiu J, Wu Q, et al. (2016) A survey of machine learning for big data processing. *EURASIP J Adv Signal Proc* 2016: 67.
6. Gerrero-Curieses A, Munoz-Romero S, Bote-Curiel L, et al. (2019) Deep learning and big data in healthcare: A double review for critical beginners. *Appl Sci* 9: 2331.
7. Hao Y, Hwang MCK, Wang L, et al. (2017) Disease prediction by machine learning over big data from healthcare communities. *IEEE* Access, 5: 8869–8879.
8. Zhang Y and Zheng T, (2017) A big data application of machine learning-based framework to identify type 2 diabetes through electronic health records, In: *International Conference on Knowledge Management in Organizations*, Springer, 451–458.
9. Luo G, (2016) Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. *Health Inf Sci Syst* 4: 2.
10. You M, Yang G, Chen Y, et al. (2017) A machine learning-based framework to identify type 2 diabetes through electronic health records. *Int J Med Inf* 97: 120–127.
11. Sisodia DS and Sisodia D, (2018) Prediction of diabetes using classification algorithms. *Procedia Comp Sci* 132: 1578–1585.
12. Pereira N, Taslimitehrani V, Pathak J, et al. (2015) Using EHRs and machine learning for heart failure survival analysis. *Stud Health Technol Inf* 216: 40.
13. Stewart WF, Sun J, Choi E, et al. (2017) Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inf Assoc* 24: 361–370.
14. Liu Z, Zhang S, Jin B, et al. (2018) Predicting the risk of heart failure with EHR sequential data modeling. *IEEE Access* 6: 9256–9261.
15. Calvert J, Hoffman J, Jay M, et al. (2016) Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Med Inf* 4: e28.

16. Hall MK, Pare JR, Venkatesh AK, et al. (2015) Prediction of in hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 23: 269–278.
17. Neapolitan R, Zexian SE, Roy A, et al. (2018) Using natural language processing and machine learning to identify breast cancer local recurrence. *BMC Bioinfor* 19: 65–74.
18. Garg S and Gupta P, (2020) Breast cancer prediction using varying parameters of machine learning models. *Procedia Comput Sci* 171: 593–601.
19. Alkhawaldeh RS, Al-Shami F and Al-Shargabi B, (2019) Enhancing multi-layer perception for breast cancer prediction. *Int J Adv Sci Tech* 130: 11–20.
20. Seenivasagam V and Suijitha R, (2020) Classification of lung cancer stages with machine learning over big data healthcare framework. *J Ambient Intell Humanized Comput* 2020.
21. Olugbara OO and Adetiba E, (2015) Lung cancer prediction using neural network ensemble with histogram of oriented gradient genomic features. *Sci World J* 2015: 786013.
22. Peter T, Delzell DAP, Smith M, et al. (2019) Machine learning and feature selection methods for disease classification with application to lung cancer screening image data. *Front Oncol*, 9: 1393.
23. Nartowt BJ, Hart GR, Deng J, et al. (2019) Stratifying ovarian cancer risk using personal health data. *Front Big Data* 2: 24.
24. Hassanat ABA, (2018) Furthest-pair-based binary search tree for speeding big data classification using K-nearest neighbors. *Big Data* 6: 225–235.
25. Hung JC, Lin KC, Zhang KY, et al. (2016) Feature selection based on an improved cat swarm optimization algorithm for big data classification. *J Supercomput* 72: 3210–3221.
26. Bei Y and Xing W, (2019) Medical health big data classification based on KNN classification algorithm. *IEEE Access*, 8: 28808–28819.
27. Li Y and Jiang C, (2019) Medical health big data classification based on KNN classification algorithm. *IEEE Access*, 7: 176782–176789.
28. Shah NH and Callahan A, (2017) Machine learning in healthcare, In: *Key Advances in Clinical Informatics*, Academic Press, 279–291.