J. Biol. Today's World. 2018 Jan; 8 (2): 1-10
Received: 27 June 2019 • Accepted: 25 July 2019

Journal of
Biology and Today's World
http://journals.lexispublisher.com/jbtw
doi: 10.15412/J.JBTW.01070206

Research

Machine Learning in Healthcare Data Analysis: A Survey

Arwinder Dhillon*, Ashima Singh

Thapar Institute of Engineering and Technology, Patiala, India

*Correspondence should be addressed to Arwinder Dhillon, Thapar Institute of Engineering and Technology, Patiala, India; E-mail: arwindhillon999@gmail.com

Journal of Biology and Today's World is published by Lexis Publisher; Journal p-ISSN 2476-5376; Journal e-ISSN 2322-3308.

Abstract

In recent years, healthcare data analysis is becoming one of the most promising research areas. Healthcare includes data in various types such as clinical data, Omics data, and Sensor data. Clinical data includes electronic health records which store patient records collected during ongoing treatment. Omics data is one of the high dimensional data comprising genome, transcriptome and proteome data types. Sensor data is collected from various wearable and wireless sensor devices. To handle this raw data manually is very difficult. For analysis of data, machine learning is emerged as a significant tool. Machine learning uses various statistical techniques and advanced algorithms to predict the results of healthcare data more precisely. In machine learning different types of algorithms like supervised, unsupervised and reinforcement are used for analysis. In this paper, different types of machine learning algorithms are described. Then use of machine learning algorithms for analyzing various healthcare data are surveyed..

KEYWORDS: Healthcare, Machine Learning, Clinical Data, Sensor Data, Omics Data.

INTRODUCTION

Health care is a wide term that concerns to a system that involves improvement of medical services in order to serve the medical demands of the people. In healthcare, efforts are made by patients, physicians, vendors, health companies and IT companies for maintaining and restoring health records. Over the past decade, Indian health care is known as one of the fast-rising industry in the world. Healthcare analysis is handling various types of diseases including cancer, diabetes, strokes and so on using machine learning. Cancer is one of the deadliest diseases. Different types of cancer are present in this human word including lung cancer, breast cancer, prostate cancer, stomach cancer and so on. Around 12% cases of lung cancer come every year in which 10% cases died from it. Similarly, for breast cancer, 11% cases come in which 9% dies from breast cancer. This happens in each type of cancer. Cancer prevalence of year 2018 is taken and is shown in **Figure 1** below. It describes the total cases and death cases for each cancer type. For handling cancer in health care analysis, there is a need to generate correct and quality data. In this competitive world, healthcare must need to use the data in such a manner that there will always be rise in quality of health care and decline in the cost needed for the treatment purpose [1]. From past years, heath care research with Machine Learning (ML) has been increasing steadily. Due to variety of medical data including clinical data, omics data or EHR data, it is difficult for humans to infer the data and to make decisions. Accordingly, ML has been proposed in health care for better understanding of data and for better decision-making process [2].

METHODS

Machine learning

Machine Learning was originated by Samuel in 1950 to

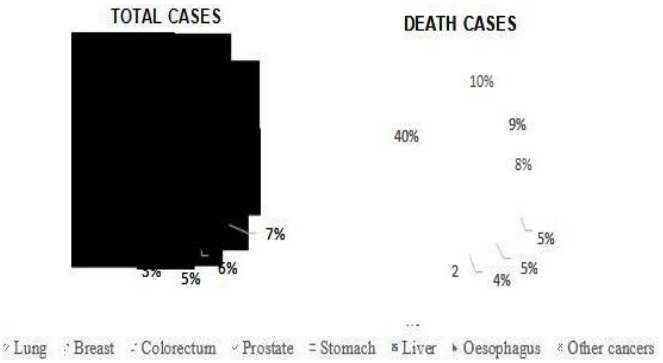

Figure 1: Evolution of the effective thermal conductivity of some monolithic, silica aerogels.

play strategic games like chess. It is the mechanism of making machines to learn automatically without being explicitly programmed. The main focus of Machine Learning is to develop a computer program which can access the data and use this data for learning purpose. It is the ability of machine to make use of statistical techniques and advanced algorithms to make more powerful prediction and making the data driven system more powerful by replacing the rule-based system. The main component of machine learning is data which is the backbone for any model. The more relevant data is the more accurate predictions are. After data, we need to select the algorithm based on the problem for more accurate predictions. Machine Learning can be used in many fields such as finance, retail, health care and social data [3].

Types of machine learning algorithms

Machine learning can be used for different purpose. Machine learning algorithms are basically classified into three categories based on their objective which varies from each other. It includes supervised learning, unsupervised learning and reinforcement learning.

Supervised learning*:* Supervised learning involves training the model on the labeled data and uses this trained model to make predictions on the new data. It involves splitting of data into two sets including training set and testing set. First the model is trained on training set and afterwards the performance is tested on the testing set. The performance of the model can be evaluated using performance metrics [4]. Supervised learning can be classification problem or regression problem. In supervised classification, the labeled value is a discrete value. The algorithms in this are used to classify to which class or category the problem belongs. On the other side, the models are used to predict the outcome based on continuous (numeric) data is supervised regression learning [4]. For the classification of raw data, first the data is selected and then preprocessing is performed in which all NA values are removed. Then the data is normalized using z-score or min max normalization. Once the normalization is performed feature selection procedure is applied to select the best features. After the features are selected, some supervised ML algorithms includes K Nearest neighbor, Decision trees, Support Vector Machines, Naïve Bays Classifier, Neural Network and Ensemble methods [3] are used for classification of raw data as shown in **Figure 2.**

Unsupervised learning*:* Unsupervised Learning also involves training of the data except for the fact that the labeled value or target value is not known. In this, machine try to cluster the similar type of the data by finding the hidden pattern. Rather than making prediction, the main aim of unsupervised learning is to discover the patterns. The performance of the model in unsupervised learning cannot be evaluated as the label value is absent or unknown. The algorithms involved in unsupervised learning are K-mean clustering, Association Rule Mining, Topic Modeling and Dimensionality Reduction Techniques [3].

Semi-supervised learning: As supervised learning works on labeled data and unsupervised learning on unlabeled data, then a lot of information is lost from labeled data which can be obtained from unlabeled data. So, in this case semi-supervised learning comes to mind. It is a mixture of supervised and unsupervised learning in which it takes both the unlabeled and labeled data. Labeled data should be of shorter length as compared to unlabeled data. The idea behind semi-supervised learning is that there is a considerable change in performance when both labeled and unlabeled data is used in conjunction. The training set used is of shorter length. It is normally used to detect outliers.

Reinforcement learning: Reinforcement Learning works by developing a system which improves its performance by taking feedback from the environment and taking possible steps to improve them. It is an act of learning from environment by interacting with it without any help from humans. It is an iterative process.

The different types of machine learning algorithms and their applications are shown in **Figure 2** above.

Related surveys

As healthcare is emerging now days, researchers are focusing on types of data used for prediction. For Example, Ajay et al. focus on clinical and genomic data and used machine learning algorithms to analyze them. But other data types are also present to work upon including sensor and Omics data. The prime motive of our survey is to include all types of data and analyze them using machine learning. It is described in the **Table 1** as follows.

Paper organization

Section 2 presents different type of data used by authors

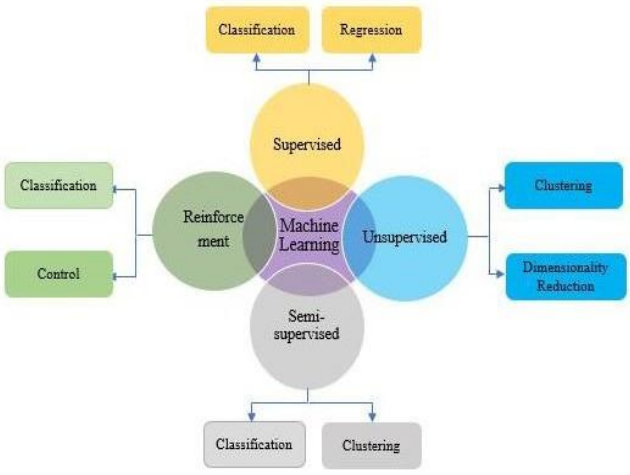

Figure 2: Types of Machine Learning algorithm.

for diagnosis and prevention of certain kind of disease and their work done for achieving it. Section 3 shows conclusion achieved from the related surveys.

RESULTS AND DISCUSSION

Healthcare analysis using ML

As in healthcare sector, there is enormous information about the patient health. So it is impossible for humans to process it. Consequently, ML provides a technique to recognize patterns from the massive data and use algorithms to predict future outcome of the patients. ML in healthcare helps users to perceive understanding about the potency of existing programs and identify the treatment that provides best result for patients according to their condition.

Types of healthcare data

Different types of data have come into view in healthcare now days including clinical data, sensor data, Omics data and so on. This type of data includes different mining methods to extract the more relevant features and then different algorithms needs to be trained for better future prediction.

Clinical data: Clinical data is the data which is collected during the ongoing treatment of the patient including the Electronic Health Record (EHR) data which is comprised of laboratory tests, radiology images, allergies and so on **(Figure 3)**. The work on clinical is applied by following authors.

Wengert et al. [5] proposed ML algorithms for early prediction of pathological complete response (pcr) to neoadjuvant chemotherapy and survival outcome of breast cancer patients using Multiparametric Magnetic Resonance Imaging (mpMRI) data. Samples of 38 women with breast cancer were taken and eight classifiers including linear support vector machine, linear discriminant analysis, logistic regression, random forests, stochastic gradient descent, adaptive boosting and Extreme Gradient Boosting (XGBoost) were applied to rank the features for pcr including residual cancer burden (RCB), Recurrence Free Survival (RFS) and disease-specific survival DSS. Area under Curve value was extracted for each feature of pcr. From the experimental results, XGBoost produces the best result with higher accuracy for RCB and DSS and logistic regression for RSS as compared to other classifiers. Dagli et al. [6] defined multilevel perception model for two year survival prediction of non-small cell lung cancer patients. Samples of 559 patients were taken and attributes were ranked with RelifF feature selection method. From results, Multilayer Neural Network was found as the best prediction model with area under curve value of 0.75. Kayal et al. [7] proposed new improved classification approach for survival prediction of Hepatocellular Carcinoma (HCC) patients. Samples of 165 patients were taken from which authors defined that out of 49 risk factors, 15 risk factors were responsible for HCC. The outcome of the experiment proved that the accuracy obtained by Deep Neural Network is significantly higher than Cox models (SVM) and Unsupervised model (KNN).

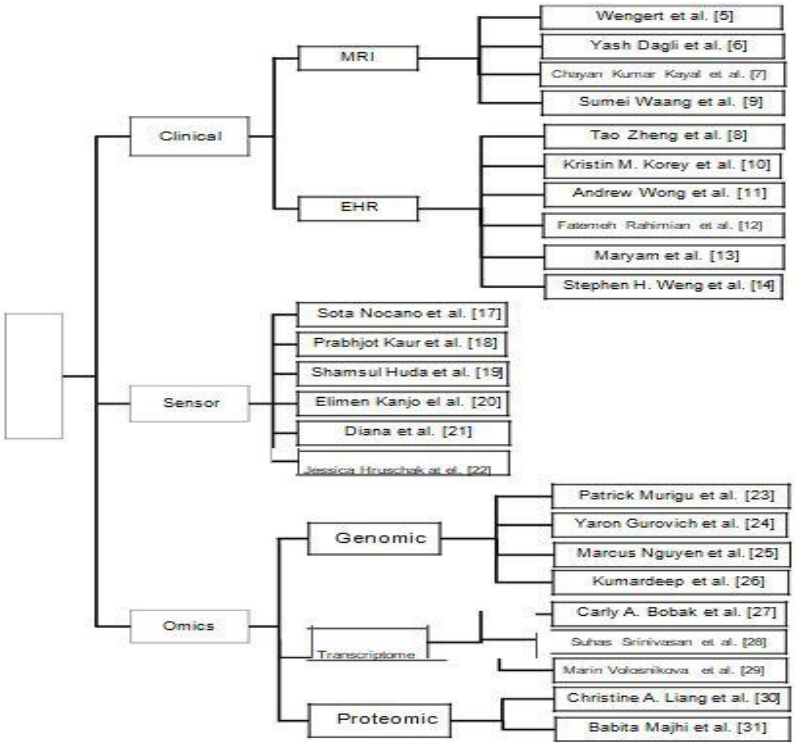

Figure 3: Types of dataset.

Table 1: Comparison with other surveys.

Authors	Clinical	Omics			Sensor
		Genomic	Transcriptomic	Proteomic	
Ajay Kumar et al.	✓	✓	_	_	_
This Survey	✓	✓	✓	✓	✓

Zheng et al. [8] proposed a framework to identify Type-2 Diabetes Mellitus (T2DM) patients using Electronic Health Record (EHR) data. A total of 300 patient samples were taken and 114 features were extracted on which different machine learning algorithms including k-Nearest Neighbor (kNN), Random Forest (RF), Decision Tree (DT), naïve bayes, Support Vector Machine (SVM) and logistic regression were applied. From results, SVM produces the best result with accuracy of 96%. Sumei et al. [9] developed computer assistant classification method by combining convolution MRI and profusion MRI data for diagnosis of different type of brain tumor and for grading of gliomas. Samples of 102 brain tumor patient were taken and Support vector machine recursive feature elimination (SVM-RFE), k-nearest neighbor and linear discriminant analysis were applied to them. The result showed that SVM RFE produced the best result with accuracy of 85% for classification of tumor and 88% for grading the gliomas. Kristin et al. [10] defined different machine learning algorithms including penalized logistic regression, random forest models, and extreme gradient boosted decision trees for identification of high-risk surgical patients. Authors trained the algorithms on Pythia data containing electronic health records having 194 clinical features including patient demographics, smoking status, medications, comorbidities, procedure information, and proxies for surgical patients. The experimental results show that the best result was produced by penalized logistic regression model with AUC value of 0.924. Andrew et al. [11] investigated five machine learning algorithms comprising penalized logistic regression, gradient boosting machine, artificial neural network with a single hidden layer, linear support vector machine and random forest for delirium risk prediction based on electronic health record data. A total of 18223 patient samples were taken and experiment was performed. From results, it was proved that gradient boosting algorithm produced the best result with AUC value of 0.855. Fatemeh et al. [12] proposed machine learning models for first emergency admission prediction based on EHR data. Authors applied Cox model on a sample of 4.6 million patient samples for prediction of risk for first emergency admission and then random forest and gradient boosting algorithm were used. Authors identified that gbm model performed best with AUC value of 0.779. Maryam et al. [13] investigated Seattle heart failure model for the prediction of heart failure by using EHR data. Samples of 5044 patient were taken and features were extracted to calculate the survival score. Authors first calculated the survival score of heart patients with Cox proportional regression model who survived for one, two or five years and then the patient who died after five years were excluded and different machine learning models comprising random forest, logistic regression, support vector regression, decision tree and ada boost were applied on the remaining patients. From experiment results, logistic regression performed best with 11% improvement in AUV curve value. Stephen H.

Weng et al. [14] defined machine learning algorithms including random forest, logistic regression, gradient boosting machines and neural networks on samples of 378,256 patients for the prediction of cardiovascular risk. After data was prepared and features were extracted. Authors applied the different machine learning algorithms and identified that neural network performed best with AUC value of 0.72 as shown in **Table 2**.

Sensor data: Data elements produced by sensors including time series signals which is an ordered sequence of pairs is sensor data. These data elements are processed by computing devices and can be simple numerical or categorical value or can be more complex data. The work on sensor data is applied by following authors. Luca et al. [15] proposed machine learning algorithms to detect Parkinson's disease (PD) by using data streams collected from wearable sensors.

Experiment was performed on 20 individuals and movement of individual was recorded by 6 wearable sensors. Total 13 tasks were performed by individuals and experiment was conducted on one day and was repeated 2 weeks later. From this a total of 41,802 data clips were used. After the data was trained using convolutional neural networks and random forest classifier for the detection of bradykinesia and tremor. Results proved that random forest classifier performed better with AUROC value of 0.73 for the detection of bradykinesia and 0.79 for the detection of tremor.

David et al. [16] defined machine learning classification algorithms for detection of risk of developmental arrays (AD) and Typical Development (TD) in infants. Long day inertial movement of infants were recorded using Opal sensors fixed on the ankle of the infant and data was divided into two sets, 0 to 6 months and 6 to 12 months. A total of 19 movement features including movement count, duration, average acceleration and peak acceleration from two sets were extracted using univariate feature selection methods which were Recursive Feature Elimination (RFE), and stepwise feature selection. Authors used three machine learning algorithms support vector

Table 2: Summarized clinical data analysis using Machine Learning.

	Type of health care data	ML algorithms	Performance parameter	Results	Future scope
Wengert et al.[5]	Clinical (mpMRI)	Support vector machine,linear discriminant analysis,logostic regression,random forests,stochastic gradient descent,adaptive boosting,extreme gradient boosting (XG Boost)	Area under curve	XG Boost produced the best result with AUC value of 0.94 for RCB and 0.92 for DSSwith AUC value of 0.83	The dataset used in this way very small. In this some features are extracted which effect the imaging features as well as the prediction of RCB. DSS. So this can be covered in future.
Yash Dagali etal. [6]	Clinical	Multilayer Neural Network,Logistic Regression,Single Perception neural network	Area under curve,95% confidence interval,Misclassification rate,True positive rate,false positive rate,accuracy and precision	Multilayer Neural Network produced the best result with AUC value of 0.75,confidence value of 0.693-0.806,true positive rate of 0.68,false positive rate of, accuracy of 0.76 and precision value of 0.72	
Chayan Kuma Karvey et al.	Clinical	Deep Neural Network, Support Vector Machine, K-Nearest Neighbor	Accuracy, Precision, Recall, Fmeasure	Deep neural network produced higher accuracy of 78% and precision,Recall and Fmeasure value of 83.58, 81.25 and 80%	In future,focus should be on appropriate feature selection method for efficient survival prediction
Tao Zheng et al.[8]	Clinical (EHR)	Support vector machine, k-nearest neighbor, logistic regression,random forest decision, tree, naïve bayes	Accuracy, Sensitivity, Specificity, Precision, Area under curve	SVM produced best result with accuracy 96%, sensitivity 95%, specificity 96%, precision 91% and AUC value of 0.96.	Dataset used was very small. Large sample can be used for better prediction
Sumei Waang et al.[9]	Clinical (MRI)	Support vector machine recursive feature elimination,Linear nearest neighbor	Accuracy, Sensitivity, Specificity,No of retained features, entropy, standard deviation based on t-test	SVM RFE produced best result for both classification of tumor and grading of gliomas with accuracy, sensitivity and specificity value of 85%, 87%, 79% and 88%, 85% and 96%. The Nf value is 20, entropy and sd is 0.82 and 0.92	Dataset used was very small. Large sample can be used for better prediction
Kristin M. Korey et al. [10]	Clinical (HER)	Penalized logistic regression, random forest models,and extreme gradient boosted decision trees basis function networks	Accuracy, Sensitivity, Specificity, Area under curve, threshold, positive predictive value	Penalized logistic regression produced best result with accuracy,sensitivity, specificity,AUC, threshold and ppv value of 95%, 76%, 76%, 0.924, 0.174 and 0.390	The effect of Pythia risk calculator needed to be evaluated for better results.
Andrew Wong et al. [11]	Clinical (EHR)	Penalized logistic regression,Gradient boosting machine, Artificial neural network with a single hidden layer,Linear support vector machine and random forest	Sensitivity, Specificity, Area under curve	Gradient boosting machine produced best result with sensitivity,specificity and AUC value of 59.7%, 23.1% and 0.855	
Fatemeh Rahimian et al. [12]	Clinical (EHR)	Cox model, Gradient boosting, Random forest	Area under curve, confidence interval	Gradient boosting machine produced best result with AUC and 95% CI value of 0.779 and 0.847	Better techniques can be used for risk prediction
Maryam et al. [13]		Cox model,Gradient boosting, Random forest Support vector regression, Decision tree, Ada boost,logistic regression	Area under curve	Logistic regression performed best with an improvement of 11% in AUC value.	
StephenH Weng et al. [14]	Clinical (EHR)	Random forest, Logistic regression,Gradient boosting machines and Neural networks	Area under curve, confidence interval, positive predictive and negative predictive value	Neural networks produced best result with AUC,CI, PPV and NPV value of 0.728, 0.75-0.76, 18.4% and 95.70%	Complexity may lead to overfitting. Best models needed for better results

machine, logistic regression and adaboost for prediction and the outcome of the result proved that SVM performed best for 0-6 month infants with accuracy of 90% and adaboost for 6-12 month infant with accuracy of 83%. Sota et al. [17] proposed shoe-type pressure sensor and single inertial measurement unit attached to the trunk for detection of assistance motion with different foot. A total of 8 Flexiforce sensors were attached to the sole. 5 people were asked to perform the experiment and were asked to perform in two variations including short step and long step. Features were extracted from the data

obtained and were trained using classification method. Experimental results show that proposed system performed best with accuracy of 90%. Prabhjot et al. [18] investigated hybrid approach comprising Bayesian network and heuristic technique in neural network for stress detection using mobile phone sensing mechanism by measuring the Blood Pressure Management (BPM) and Heart Rate (HR) value. Data was collected using sensors embedded in mobile phones and hybrid approach was applied to detect stress using BPM values and HR values as shown in **Table 3**. From result, hybrid approach performed well with accuracy of 92.86% for BPM and 85.71% for HR. Shamsul et al. [19] proposed Deep-belief network for recognition of human activity using data from body sensors. Sensor data was collected and important feature were extracted using Kernel Principle Component Analysis (KPCA) and Linear Discriminant Analysis (LDA). Then, the model was trained using deep-belief network with 40 hidden layers. From results, it is cleared that deep belief network performed best for activity detection with an accuracy of 97.5%. Elimen et al. [20] defined hybrid approach comprising Convolutional Neural Network and Long Short-term Memory Recurrent Neural Network (CNN-LSTM) for emotion detection using smart phones and wearable sensor devices data. Sample of 40 female patients were taken from which 550,432 sensor data values were collected comprising of on-body data, environmental data and self-report emotion level data captured using mobile phone app. Then data was preprocessed and trained using hybrid CNN-LSTM model for emotion detection. The outcome of the result proved that the proposed hybrid approach performed best with accuracy of 95%. Diana et al. [21] investigated four machine learning classifier including decision trees, ensemble, logistic regression and Deepnets for the detection of fall in elderly people using 3D-axis accelerometer fitted in 6lowPAN wearable device. The accelerometer reading was collected and feature was extracted with sliding window technique. Fall was detected using machine learning classifiers and the outcome of result proved that ensemble algorithm performs best with accuracy of 94%. Jessica et al. [22] proposed 90 second fear induction task to measure the motion of participant using a wearable sensor for the detection of anxiety and depression

Table 3: Summarized sensor data analysis using Machine Learning.

	Type of health care data	ML algorithms	Performance parameters	Results	Future scope
Luca Lonini et al [15]	Wearable sensors data from sensors attached to arms, hands and thighs	Random Forest, Convolutional neural network	Area under ROC curve, confidence interval, Standard Deviation	Random forest classifier performed better with AUROC, CI and SD value of 0.73, o.68-0.77 and 21.7% for the detection of bradykinesia and 0.79, 0.74-0.84 and 11.5% for the detection of tremor.	Due to limited number of individuals, it was difficult to detect dyskinesia. So this can be a research area in future.
David Goodfellow et al [16]	Wearable sensors data collected from sensors attached to ankle of infrant	Support vector machine, Logistic regression, Adaboost	Accuracy precision Recall, F1 score	SVM performed best with accuracy, precision, recall and F1 score value of 90%, 92%, 90% and 90% for 0-6 month infant and adaboost performed best for 6-12 month infant with accuracy, precision, recall and F1 score value of 83%, 83%, 84%, 83%.	Future studies were required to determine the robustness of classifier and for better prediction of typical development
Sota Nocano et al. [17]	Shoe type pressure sensor data	Classification methods	Accuracy	Proposed system performed best with 80% accuracy.	For better results, appropriate foot position is necessary. So this can be done in future.
Prabhjot Kaur et al. [18]	Wireless sensor using mobile phone	Hybrid approach including Bayesian network and heuristic technique in neural network	Accuracy	Hybrid approach performed well with accuracy of 92.86% for BPM and 85.71% for HR.	Detection of stress using real time sensor data can be a topic for future study.
Shamsul Huda et al. [19]	Body Sensors	Support vector machine, Deep belief network	Accuracy	Deep belief network performed best with an accuracy of 97.5%	Parallelism was not considered in this approach. So this can be a future research.
Elimen Kanjo et al. [20]	Physiological, environmental and location data using mobile phones and wearable devices	Convolutional neural network, Long short-term memory recurrent neural network	Accuracy, Precision, Recall, F1 score, RMSE Error rate	Hybrid CNN-LSTM performed best with Accuracy, Precision, Recall, F1 score, and RMSE value of 92%, 95%, 94%, 94% and 29%.	
Diana et al. [21]	Data collected using 3D axis accelerometer fitted in 6lowPAN wearable device	Logistic regression, Ensemble, Deepnet, Decision tree	Accuracy, Precision, Sensitivity, specificity, Gain	Ensemble performed best with accuracy, Precision, Sensitivity, specificity and gain value of 92%, 92%, 92% and 67% avg gain.	For more benefits, integration of more sensors and developments of new services related to health can be considered as research topic.
Jessica Hruschak et al. [22]	Data collected from wearable sensor	K nearest neighbour	Accuracy	KNN performed best with 75% accuracy.	Future studies should consider additional feature selection techniques to reduce the number of features.

among young children. Samples of 64 children were taken and they were subjected to 20 second potential threat phase. Data was collected after 20 second threat phase and features were extracted from sensor data. Authors then subjected the data to k-nearest neighbor model and proved that the proposed model produced best result with an accuracy of 75%.

Omics data: Omics data is collection of huge amount of complex and high dimensional data consisting of genomic, transcriptomic and proteomics data. Handling this type data required various techniques including machine learning algorithms.

Genomic data: Genomic data is collection of gene expression, copy number variation, sequence number and DNA data and is used in bioinformatics. The work on genomic data is applied by following authors. Patrick et al. [23] proposed machine learning algorithms for improving hazard characterization in microbial risk assessment. Because of high dimensionality of genomics data, authors defined ML based predictive risk modelling for risk assessment. Dataset related to DNA isolation and sequencing were collected and feature extraction was performed to extract the relevant features. Machine Learning classifiers including random forest, support vector machine, logic boost were applied and results were evaluated. From results, it was proved that logic boost performed best with an accuracy of 75%. Yaron et al. [24] proposed DeepGestalt, a deep learning framework for identification of facial phenotypes of genetic disorders. Samples of 17000 patients with 200 syndromes were taken. Features were extracted and DeepGestalt was applied in which face detection was done using deep convolutional neural network (DCNN) and then image is normalized and cropped into different segments which is then converted to grey scale. After Gestalt model was trained and predict the syndrome with 91% accuracy. Marcus et al. [25] investigated machine learning algorithm XGBoost for prediction of minimum antimicrobial concentration among patients. Samples of 5278 non-typhoidal Salmonella genomes were collected. Shortread sequenced data was collected for each strain with genome assembled service and XGBoost was applied which used gradient boosting ensemble method to reduce the error. The outcome of the result proved that XGBoost produced best result with accuracy of 95%. Kumardeep et al. [26] defined deep learning model and six machine learning algorithms comprising random forest, support vector machine, linear discriminant analysis, prediction analysis for microarrays, recursive partitioning and regression trees and generalized boosting model for prediction of estrogen receptor status in breast cancer patients based on metabolomics data. Samples of 271 patients were taken in which 204 patients are with positive estrogen receptor and 67 with negative receptor. K-nearest neighbor was used for normalization of data. The normalized data was trained using machine learning and deep learning algorithm. From experimental results, it was proved that deep learning algorithm performed best with AUC value of 0.93.

Transcriptomic data: Transcriptomic data is a collection of multiple mRNA transcripts data within a biological sample. These samples are analyzed and extracted to generate different datasets. The work on transcriptomic data is applied by following authors. Carly et al. [27] proposed a framework to integrate multiple gene expression datasets to identify gene signatures for the diagnosis of tuberculosis. Samples of 1164 patients were taken by integrating 4 datasets. Features were extracted and machine learning algorithms including random forest, support vector machine with polynomial kernel and Partial least square discriminant analysis applied and results were evaluated. From results, it was proved that random forest performed best with an accuracy of 95%. Suhas et al. [28] proposed a hybrid approach comprising a deep unsupervised single cell clustering which integrates the feature generated by deep learning model for profiling of single-cell RNAsequencing data. Samples were taken and features were extracted. Model was trained and the proposed model performed the best result with accuracy of 96%. Marin et al. [29] investigated machine learning algorithm for tracking age related changes of human muscle skeleton on transcriptomic data. Gene-expression profiles of donor were analyzed to compare signatures of old and young donors. Machine learning algorithm comprising neural network was applied on signature data which built a biomarker for aging. The outcome of the result proved that proposed technique produced best result with accuracy of 80%.

Proteomic data: Proteomic data is a collection of proteins expressed in the form cell, tissue or an organism. It is the representation of actual functional molecules in the cell. The work on proteomic data is applied by following authors. Christine et al. [30] proposed deep learning algorithms for the analysis of FLT3-ITD in acute leukemia patients. Samples of 191 patients with protein data were taken which have serum level of 231 patients. Deep learning with stacked auto-encoders was used and dimensionality reduction reduces the proteins from 291 to 20. From results, it was proved that the proposed model performed best with accuracy of 97% as shown in **Table 4**. Babita et al. [31] proposed a hybrid space for the prediction of protein structure class. A hybrid approach including SkipGram based word2hovac and Atchleys space II, III, IV for electron ion interaction were applied for amino acid sequence representation [32]. For feature extraction of time and frequency domain, Stockwell transformation was applied. It was applied on six datasets including small sized samples comprising 498, 277 and 204 and large sized samples comprising PDB25, 640 and FC699. Deep recurrent neural network was used for classification. The result proved that proposed approach performed best with accuracies of 95.9%, 94.9%, 85.36%, 84.2%, 94.3% and 93.1% for both small sized and large sized datasets.

Table 4: Summarized Omics data analysis using Machine Learning.

	Type of health care data	ML algorithms	Performance parameters	Results	Future scope
Patrick Murigu et al.	Genomic data (Sequencing and DNA isolation)	Support vector, machine, Random forest, Logic boost	Accuracy, sensitivity, specificity, confidence interval, positive and negative predictive value	Logic Boos produced the best result with accuracy, sensitivity, specificity, PPV, NPV and CI value of 75%, 60%, 86%,0.25, 0.95 and 0.60-.80.	
Yaron Gurovich et al. [24]	Genomic	Deep Gestalt	Accuracy	DeepGestalt correctly identified syndrome with 91% accuracy.	Dataset used was very work small work on large dates was required for better predictions
Marcus Nguyen et al. [25]	Whole Genome Sequence data	XGBoost	Accuracy, quartile bound	XGBoost produced the best Result accuracy and quartile bound value of 95% and 89%	In future, subtle Genomic changes must identified that result in different MIC.
Kumardeep Chaudhary et al.[26]	Genomics (metabolomics data)	random forest, support vector machine, linear discriminant analysis, prediction analysis for microarrays, recursive partitioning and regression tree sand generalized boosting model, deep learning model	Area under curve	Deep learning produced best result with AUC value of 0.93.	Systematic identification of DL methods was not there. This can be a research topic in future.
Carly A. Bobak et al. [27]	transcriptomic (gene signature data from multiple datasets)	random forest, support vector machine with polynomial kernel and Partial least square discriminant analysis	Accuracy, Sensitivity, Specificity, Area under curve	RF produced best result with accuracy, sensitivity, specificity and AUC value of 95%, 89%, 97%, 97%.	Additional dataset may be included for better prediction
Suhas Srinivasan et al. [28]	Single-cellRNA sequencing data	Deep unsupervised single cell clustering	Accuracy	Proposed technique produced best result with accuracy of 95.	
Marin Volosnikovaet al.[29]	transcriptomic data	Random Forest,Support vector machine, Elastic net, Deep feature selection	Accuracy, pearson correlation, coefficient of determination mean average error	Deep feature selection produced best result with accuracy of 95,0.96,0.92 and 5.6.	Proposed technique can be applied to further disease prognosis

Christine A. Liang et al.[30]	Protein Data	Deep neural network with stacked auto-encoders	Accuracy, Sensitivity, Specificity	Proposed technique produced best result with accuracy, sensitivity and specificity value of 97%, 90% and 98%.	This technique can be used in further research to determine chemotherapy response.
Babita Majhi et al.[31]	Small sized and large sized Protein datasets	Recurrent neural network	Accuracy, precision, recall, f1 score	Proposed approach performed best with accuracies of 95.9%, 94.9%, 85.36%,84.2%,94.3%and 93.1% for both small sized and large sized datasets.	Another factors of Atchleys may be considered for better prediction. This is work of future studies.

CONCLUSION

A different type of data is present in healthcare. To analyze this variety of data various Machine learning algorithms such as supervised, unsupervised and reinforced algorithms are used to improve prediction which can be analyzed using various performance parameters like accuracy, sensitivity, specificity, precision, F1 score, and Area under Curve. In this paper, machine learning algorithms are defined and use of machine learning algorithms for analyzing different types of healthcare data like clinical, omics and sensor data is done. From the survey, it is concluded that for analyzing different types of data in healthcare, various machine learning algorithms and feature extraction techniques are proposed by various authors for survival prediction of cancer patients.

ACKNOWLEDGEMENT

I am thankful to Dr. Ashima Singh for her assistance in preparing article

AUTHORS CONTRIBUTION

Arwinder Dhillon designed the study, contributed to the study, contributed in analyzing the data and also wrote the paper. Dr. Ashima Singh read, help in required changes and approved the final manuscript.

CONFLICT OF INTEREST

The authors declare no potential conflicts of interests with respect to the authorship and/or publication of this paper.

REFERENCES

1. Raheja K, Dubey A, Chawda R. Data analysis and its importance in health care. Int Computer Trends and Technology J. 2018;48:176-180.
2. Bisaso KR, Anguzu GT, Karungi SA, Kiragga A, Castelnuovo B. A survey of machine learning applications in HIV clinical research and care. Comput Biol Med. 2017;91:366-371.
3. Alpaydin E. Introduction to Machine Learning. MIT press; 2009.
4. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: Applications and ethics. Behav Sci Law. 2019;37(3):214-222
5. Tahmassebi A, Wengert GJ, Helbich TH, Bago-Horvath Z, Alaei S, Bartsch R, et al. Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. Invest Radiol. 2019;54(2):110-117.
6. Dagli Y, Choksi S, Roy S. Prediction of two year survival among patients of non-small cell lung cancer. ICCMIA. 2019;31:169-177.
7. Kayal CK, Bagchi S, Dhar D, Maitra T, Chatterjee S. Hepatocellular carcinoma survival prediction using deep neural network. Proceedings of International Ethical Hacking Conference. 2019;349-358.
8. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017;97:120-127.
9. Zacharaki EI, Wang S, Chawla S, Soo Yoo D, Wolf R, Melhem ER, et al. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. Magn Reson Med. 2009;62(6):1609-1618.
10. Corey KM, Kashyap S, Lorenzi E, Lagoo-Deenadayalan SA, Heller K, Whalen K, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study. PLoS Med. 2018;15(11):e1002701.
11. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. JAMA Netw Open. 2018;1(4):e181018.
12. Rahimian F, Salimi-Khorshidi G, Payberah AH, Tran J, Solares RA, Raimondi F, et al. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. PLoS Med. 2018;15(11): e1002695.
13. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and machine learning for heart failure survival analysis. Stud Health Technol Inform. 2015; 216:240.
14. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PloS One. 2017;12(4): e0174944.
15. Lonini L, Dai A, Shawen N, Simuni T, Poon C, Shimanovich L, et al. Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. Npj Digit Med. 2018;1:64.
16. Goodfellow D, Zhi R, Funke R, Pulido JC, Mataric M, Smith BA. Predicting Infant Motor Development Status using Day Long Movement Data from Wearable Sensors. arXiv preprint. 2018; arXiv:1807.02617.
17. Kitagawa K, Uezono T, Nagasaki T, Nakano S, Wada C. Classification Method of Assistance Motions for Standing-up with Different Foot Anteroposterior Positions using Wearable Sensors. 2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT). 2018;1-3.
18. Kaur P, Malhotra S. Improved SLReduct Framework for Stress Detection Using Mobile Phone-Sensing Mechanism in Wireless Sensor Network. Advanced Computing and Intelligent Engineering. 2019;499-507.
19. Hassan MM, Huda S, Uddin MZ, Almogren A, Alrubaian M. Human activity recognition from body sensor data using deep learning. J of med syst. 2018;42(6):99.
20. Kanjo E, Younis EM, Ang CS. Deep learning analysis of mobile physiological,

environmental and location sensor data for emotion detection. Information Fusion. 2019;49:46-56.

21. Yacchirema D, de Puga JS, Palau C, Esteve M. Fall detection system for elderly people using IoT and Big Data. Procedia computer science. 2018;130:603-610.
22. McGinnis RS, McGinnis EW, Hruschak J, Lopez-Duran NL, Fitzgerald K, Rosenblum KL, et al. Wearable sensors and machine learning diagnose anxiety and depression in young children. 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). 2018;410-413.
23. Njage PM, Leekitcharoenphon P, Hald T. Improving hazard characterization in microbial risk assessment using next generation sequencing data and machine learning: Predicting clinical outcomes in shigatoxigenic Escherichia coli. Int J Food Microbiol. 2019;292:72-82.
24. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019;25(1):60.
25. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, et al. Using machine learning to predict antimicrobial minimum inhibitory concentrations and associated genomic features for nontyphoidal Salmonella. J Clin Microbiol. 2018;57(2).
26. Alakwaa FM, Chaudhary K, Garmire LX. Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. J Proteome Res. 2017;17(1):337-347.
27. Bobak CA, Titus AJ, Hill JE. Comparison of common machine learning models for classification of tuberculosis using transcriptional biomarkers from integrated datasets. Applied Soft Computing. 2019;74:264-273.
28. Srinivasan S, Johnson NT, Korkin D. A hybrid deep clustering approach for robust cell type profiling using single-cell RNA-seq data. BioRxiv. 2019:511626.
29. Mamoshina P, Volosnikova M, Ozerov IV, Putin E, Skibina E, Cortese F, et al. Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. Front Genet. 2018;9:242.
30. Liang CA, Chen L, Wahed A, Nguyen AN. Proteomics analysis of FLT3-ITD mutation in acute myeloid leukemia using deep learning neural network. Ann Clin Lab Sci. 2019;49(1):119-126.
31. Panda B, Majhi B. A novel improved prediction of protein structural class using deep recurrent neural network. Evolutionary Intelligence. 2018;1-8.
32. Ajay K, Sushil R, Tiwari A. cancer survival analysis using Machine Learning. 2019;26-28.