Application Of Machine Learning Techniques, Big Data Analytics In Health Care Sector – A Literature Survey

Dr. M. Sughasiny
Assistant Professor, Department of Computer Science
Srimad Andavan Arts and Science College (Autonomous)
Trichy, Tamilnadu, India
drsugha.2008@gmail.com

J. Rajeshwari
Research Scholar, Department of Computer Science
Srimad Andavan Arts and Science College (Autonomous)
Trichy, Tamilnadu, India
rajeejnr.cs@gmail.com

***Abstract*—*The triumphant utilization of data mining in extremely evident areas like trade, commerce, and e-business has directed to its application in another industry. The medical conditions are still knowledge rich but information low. There is an abundance of information feasible inside the medical practices. Still, there is a shortage of essential investigation mechanisms to recognize hidden trends and relationships in data. Many researchers have applied Data Mining methods for the prognosis and diagnosis of several diseases. Machine Learning methods have broadly utilized in the prognostication of different diseases at the beginning stages. The current decade has observed an abnormal development in the variety and volume of electronic data associated with the development and research, patient self-tracking, and health records together suggested to as Big Data. This paper presents a comprehensive literature survey on the importance of Feature Selection methods, Supervised Machine Learning methods, Unsupervised Machine Learning methods and big data for the healthcare industry.*

Keywords—Data Mining; Feature Selection; Big Data; Supervised Machine Learning; Unsupervised Machine Learning; Healthcare Industry

I. Introduction

Data mining is described as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database" by Fayyad [1]. Healthcare databases have a vast volume of data, but there is a scarcity of sufficient analysis tools to discover the in-depth knowledge. Appropriate computer-based information and/or decision making systems can support physicians in their work to recommend less expensive therapeutically similar choices. Efficient and reliable implementation of a computerized system needs a similar comparison of several techniques available. Disease Prediction plays a vital role in data mining. Data Mining is used intensively in the area of medicine to prognosticate diseases such as skin cancer, heart disease, lung cancer, breast cancer etc. In this paper, it has been present an overview of the modern research being carried out applying the data mining techniques, machine learning, big data for the diagnosis and prognosis of various diseases, to point out crucial issues and summarizing the methods in a set of accomplished lessons.

II. A Tsunami of Information in Healthcare Industry

fifteen minutes. That is how long the doctor has to examine patient, assess the patient record accusation, analyze an answer and see the patient out the entryway – ideally on the pathway back to wellbeing. This is not much time when it examines the wealth of knowledge that he/she has to examine. A patient report, the medical analysis related to the complaint, the answers about the condition that it provides, the primary examination ("say aaah") that is actioned. So how will the doctor deal when challenged with the tsunami of healthcare information that will happen when it is regular for the patient history to carry data about the genome, the microbiome (bugs in the body) and its fitness regime?

An e-health record is quickly becoming the most compelling tool in the medical toolkit. Every data will be put away in the cloud. It should be because the extent of the e-document carrying the whole patient record is considered to be as much as 6TB. That is a fourth of the entire of Wikipedia (24Tbs)!

A data file that large is required to enable the practice of precision medicine. This is a new revolution in healthcare. It is the capacity to target healthcare treatment specifically for a person. Notwithstanding enhancing wellbeing outcomes, precision medicine will spare imperative wellbeing dollars because it is empowered by one of a kind data bits of knowledge that lead to more targeted treatments. From the human anatomy, the following types of data are extracted.

- ***Social Data:*** Personal circumstances, such as living situation and income.
- ***Device Data:*** Information collected from apps that measure fitness and sleeping, electronic inhalers etc.
- ***Metabolome:*** Chemicals which are created, modified and broken by bodily methods such as enzymatic responses.

- ***Transcriptome:*** Messages generated from DNA to make the stencil (mRNA) of proteins.□

- ***Clin***ical Data: Medical documents of the patient.

- Genome: The Total set of genes 'written' in DNA

- ***of t***he patient.□

- ***Exposome:*** Result of the outside environment, such as tobacco smoke and pollution etc.□

- ***Proteomo:*** An arrangement of proteins, including chemicals, which are the building blocks of the body.

Microbiome: Combined name for 100 trillion microscopic flies living inside us. □

- ***Epigenetic (Methylome):*** The arrangement of nucleic and methylation changes in a human genome.

- ***Imaging:*** Such as x-rays, Medical images, ultrasound, scans. □

III. Key Ideas correlated with Machine Learning(ML)□

Artificial Intelligence(AI): Intelligence exhibited by machines. In the zone of Computer Science(CS), the perfect 'wise' machine is a delicate, normal cause that recognizes its condition and applies drives that expand it is indications of achievement at some objective.□

Big Data: Huge volume and complicated data sets that might be analyzed computationally to distribute patterns, examples, and associations – structured, unstructured and semi-structured data can be dug for learning. Holding a Huge volume dataset is required to utilize ML and AI effectively.

Augmented Reality: A live or direct perspective of the physical world, supplemented by computer-produced tangible info (such as sound, graphics or GPS, video data).

Computing Techniques: Various approaches and techniques for unraveling intricacies utilizing strategies that are mathematical or can be built and incorporated into a computer.

Data Mining: Can be measured a superset of a wide range of techniques to extract bits of knowledge from data. Might include conventional statistical strategies and machine learning.

Computer Learning(CL): Different name for Machine Learning – clarifies the process of how computers are 'learning' by human information and preparing. □

Deep Learning(DL): Deep learning (a term printed by Geoffrey Hinton in the year 2006) connects advances in computing force and one of a kind sorts of Neural Networks (NN) to learn complex examples in an immense volume of data. It is a piece of ML.□

Internet of Things (IoT): A change in machine-to-machine (m2m) communication and development of the Internet in which everyday objects have network connectivity. It will take into consideration constant incite connection and data sending/receiving progressively.

Machine Learning (ML): The objective of ML is to comprehend the composition of data with the goal that legitimate and accurate prediction can be done based on the characteristics of that data. □

Operational Intelligence (OI): A section of ongoing, compelling business analytics – delivers entrance into ordinary business processes. Gifts perception of IT frameworks and technology structure inside the business - permits educated decisions.□

Precision Medicine: A medical model that recommends the customization of healthcare – with products, medications, medical decisions, and practices being custom-made to each patient. □

Predictive Analytics: A field of data mining that dealings with extracting data from the dataset and utilizing the data to predict standards of conduct. It is utilized to construct predictions about unknown future issues.□

IV. Literature survey

The aim of this section is used to highlight what has been done so far in the field of healthcare by using Feature Selection techniques, Machine Learning techniques, and Big Data analytics. This literature survey helps to improve the research methodology, focus on the research problems in the healthcare domain.

A. Literature Survey on the Feature Selection

Table 1 gives the detailed literature survey on the importance of feature selection in the healthcare diligence by using Data Mining techniques.

TABLE 1: Literature Survey on the Feature Selection methods in different healthcare fields

Description	**Methods used**	**Dataset and its sample size**
[1]This paper examined the cancer classification by using SVM-based wrapper feature selection method □	Support Vector Machine, Correlation-based Feature Selection	Leukemia dataset and its sample size are 172.
[2]This paper proposes a meta-heuristic strategy utilizing stochastic local search (SLS) converged with arbitrary timberland (RF) where the arrangement is to characterize the most pertinent proteins and qualities prompting the better classification of Acute Myeloid Leukemia (AML) patients.	Stochastic Local Search, Random Forest	AML dataset and its sample size are 271 features.
[3]This paper proposed a	C4.5 Decision Tree,	Waveform

novel C4.5 calculation based on wrapper include selection technique, to help astute clinical decision-making in the healthcare fields. □	Bagging algorithm, Naïve Bayes DT, K-Nearest Neighbor, Bayes Net□	Dataset, Sick Dataset, Letter dataset, Sonar dataset, Adult dataset, Electrocardiography dataset
[4]This paper proposed a novel half and half component selection techniques expand on Symmetrical Uncertainty and Genetic calculation	Symmetrical Uncertainty (SU), Genetic Algorithm(GA)□	Ionosphere dataset, Soybean dataset, Diabetes, Segment challenge, Vote, Dermatology, Lung Cancer, Wine, Hepatitis, Vehicle and the total sample size is 257 features.□
[5]This paper proposed a component selection technique based on PSO and Quantum PSO with Elitist reproducing.	Particle Swarm Optimization, Quantum PSO	Cancer datasets like Lung cancer (192 samples), Colon cancer (202 samples), Blood cancer (66 samples) and Cervical cancer (156 samples)
[6]This paper, proposed a novel component selection and area associated prevent words extraction from unstructured with class unevenness in discharge outline notes.	Logistic Regression Model	Congestive Heart Failure (CHF) Admission dataset
[7]This paper carried out the challenging tasks of choosing critical highlights from the huge arrangement of accessible highlights and coronary illness analyze.	Differential Evolution algorithm, Feed-forward Neural Network, Fuzzy Analytical Hierarchy Process (AHP)	Heart Disease (HD) Dataset from UCI repository
[8]This paper investigated the capability of predicting treatment success for patients in emotional wellness care.	Feature Similarity analysis, Random Forest, K-Nearest Neighbor	780 exclusive patients dataset collected from DSM-IV
[9]This paper displayed an altered cuckoo search technique with unpleasant sets is mimicked to manage high dimensionality data by highlight selection	Cuckoo Search algorithm, Rough set theory	Breast cancer, Hepatitis, Iris, Dermatology, Pima Indians, Lenses, Lung cancer
[10]In this paper, a new meta-learning architecture has proposed to recommend feature selection algorithms□	Consistency-based Filter, Correlation-based Feature Selection, Infogain, ReliefF, Alternating Decision Tree, CART, J48, kNN, MLP, NB, SVM	150 UCI repository datasets
[11]This paper presented to predict effective Drug-Drug Interaction.	Incremental Feature Selection, Minimum Redundancy Maximum Relevancy, Random Forest□	DrugBank and 36,615 pairs
[12]This paper, displayed a novel component selection approach called artificial honey bee colony calculation based on a new multi-objective, coordinated with the non-ruled arranging procedure and genetic administrators□	Artificial Bee Colony, Linear Forward selection, Greedy Stepwise backward selection	12 Benchmark datasets
[13]This paper displayed a hybrid selection mechanism by linking Bayesian network and symmetrical uncertainty □	Symmetrical Uncertainty and Bayesian Network	KHNANES Dataset

B. Literature Survey on the Supervised Machine Learning Techniques

The supervised ML techniques incorporate the Classification and Regression for training the network to get the appropriate result.

TABLE II: Literature Survey on Supervised Machine Learning techniques in the health care domains □□

Description	Methods used	Dataset and its sample size
[14]This paper introduced a concurrent model based on Machine Learning(ML) has proposed for supporting outpatient physicians in performing analyze □	SVM and Neural Network (NN)□	Medical Data collected from (Class II-Grade A) Hospital, Wuhan city, China, from Jan 2013 to Aug 2015□
[15]This paper introduced different researchers musings that describe their approach to sufficiently exhibit the arrangements concerning the forecast of different cardiovascular medical problems at various levels □	Naïve Bayes algorithm, J48	UCI Cardiovascular disease dataset
[16]This paper proposed a KGRNN for the investigation and finding of sort II diabetes □	K-Means clustering, ANN	Pima Indian diabetes dataset
[17]This paper introduced an efficiently recognize passionate circumstances by examining the highlights of EEG called electroencephalography signals, which have produced from EEG sensors that noninvasively assess the electrical activity of neurons in the human mind, and choose the ideal incorporation of these highlights for	one-way ANOVA, SVM, KNN, LDA, NB, Random Forest, Deep Learning, four ensemble methods(bagging, boosting, stacking and voting)	scalp EEG data of 21 healthy subjects

recognition		
[18]This paper examined with the construction of classifiers that can be intelligible and in addition strong in performance for the quality dataset of AD utilizing a decision tree□	Decision Tree, Chi-Squared, Information Gain, Gain Ratio, J48, C4.5	Ensemble gene, AlzGene, GenCard and NCBI
[19]This paper built up a new knowledge-based framework for classification of bosom cancer malady utilizing clustering, commotion expulsion, and classification techniques	Expectation Maximization, Classification and Regression Tree, Fuzzy Logic, Principal Component Analysis	Wisconsin Diagnostic Breast Cancer and Mammographic mass datasets
[20]In this paper, a new technique has proposed for the automatic finding of typical and Coronary Artery Disease conditions utilizing Heart Rate Variability (HRV) flag extracted from an electrocardiogram (ECG)	Principal component analysis, Support Vector Machine	86 lengthy ECG recordings of 80 human subjects
[21]Multiple Kernel Learning with Adaptive Neuro-Fuzzy Inference System (MKL with ANFIS) based profound learning strategy is proposed in this paper for coronary illness determination.	Deep Learning, Multiple Kernel Learning, Adaptive Neuro-Fuzzy Inference System, Support Vector Machine, Least Square, LDA, GDA	Heart Disease Dataset
[22]This paper intended to analyze and compare the accuracy of four diverse machine learning calculations with receiver working characteristic (ROC) curve for predicting and diagnosing coronary illness by the 14 qualities from UCI Cardiac Datasets	Supervised ML, Unsupervised ML, and Reinforcement Learning	UCI Cardiac Datasets and 14 attributes
[23]This paper took the upside of points of interest of an incremental machine learning technique, Incremental help vector machine, to build up a new strategy for Unified Parkinson's Disease Rating Scale (UPDRS) prediction□	Support Vector Machine, Non-Linear Iterative Partial Least Squares, Self Organizing Map	Parkinson's Disease dataset
[24]This paper proposed a framework to outline physiological measures to subjective self-announced torment scores utilizing machine learning techniques	Multinomial Logistic Regression, KNN, SVM, and RF	40 in-patient participants with their clinical data recorded on admission at Duke University Hospital, from June 2015 to April 2017.
[25]The principle point of this paper is to explore different data mining and machine learning techniques utilized in the investigation of rheumatoid joint pain prediction based on clinical and genetic factors	Adaboost, SVM, ANN	rheumatoid arthritis disease dataset
[26]This paper exhibited a framework which will help in decreasing the dynamic visits to the center in addition help in the early assurance of risky sicknesses	Neural Network(NN) with Multi-Layer Perceptron(MLP)	UCI repository Heart Disease dataset
[27]Prevention and diagnosis of NAFLD is an ongoing area of interest in the healthcare community. Screening is complicated by the fact that the accuracy of noninvasive testing lacks specificity and sensitivity to make and stage the diagnosis	Machine Learning method	Dataset from the Canadian primary care sentinel surveillance network database
[28]This paper exhibited a plan of a CDSS to help patients with Low Back Pain LBP in their self-referral to essential care	Supervised Machine Learning methods (Decision Tree, Random Forest, and Boosted Tree)	1288 fictive cases of LBP, 63 physiotherapists, and GPs on referral advice during a vignette study
[29]This paper proposed a new knowledge-based framework for illnesses prediction utilizing clustering, commotion expulsion, and prediction techniques	Expectation Maximization clustering, PCA, CART and Fuzzy Logic	Pima Indian Diabetes, Mesothelioma, WDBC, StatLog, Cleveland and Parkinson's telemonitoring datasets
[30]This paper exhibited a current predictive model in medicine and healthcare have critically assessed	Supervised ML, Unsupervised ML	Various healthcare datasets
[31]The paper meant to audit the current writing on the utility of ML techniques in the gauge of subjects with bipolar confusion	Support Vector Machine, Pattern recognition, Unsupervised ML	Bipolar disorder dataset
[32]This paper exhibited ML techniques such as ANNs are important devices for looking at and assessing substantial and complex datasets. ANNs have still to be utilized for risk factor examination in orthopedic medical procedure □	Artificial Neural Network	The American College of Surgeons National Surgical Quality Improvement Program (ACSNSQIP) database
[33]This paper utilized decision tree concentrate to build up an instrument to scale and measure the risk of NMSC in Liver Transplant (LT), recipients	Cox Regression analysis, Decision Tree,	non-melanoma skin cancer (NMSC) dataset
[34]This paper exhibited a specialized skin ailment processing model is characterized by Dermatology Disease□	Bayesian Network	Dermatology Dataset
[35]This paper displays the research center investigation of data given by the UCI machine learning (ML) storehouse. Weka open source ML device given by Waikato University uncovers the hidden fact behind the	J48, Naïve Bayes, ID3 classification methods	Heart Disease dataset

datasets on applying the administered mathematical demonstrated calculation		
[36]In this work, the dataset is right off the bat classified utilizing diverse calculations, and after that, it is resolved what classification calculation performs better to predict lumbar spine pathologies□	Naïve Bayes, J48, Random Forest, Decision Table, SVM, MLP	The dataset is from the outpatient department of Joshi Neuro Trauma Centre, Jalandhar and Johal Multispecialty Hospital, Jalandhar for seven months from 1/1/2016 to 31/7/2016

C. *Literature Survey on Unsupervised Machine Learning Techniques*

Table 3 gives the unsupervised ML techniques incorporate the Clustering and Association Rule Mining method for getting the appropriate result.

TABLE III: Literature Survey on Unsupervised Machine Learning techniques in the healthcare domains.□

Description	Methods used	Dataset and its sample size
[37]This paper introduced a half and half technique that combines k-harmonic means and covering k-means calculations	K-Means clustering, K-Harmonic clustering	Medical Datasets
[38]This paper introduces the objective function of proposed strong fuzzy clustering techniques by incorporating Laplacian kernel-induced distance, Canberra distance, possibilistic enrollments, and fuzzy participations	Fuzzy C-Means	Breast Cancer database
[39]This paper recommended the advancement of a calculation that can incorporate high-dimensionality data to accomplish comparative outcomes is critical	K-Means clustering	The micronucleus (MN) Mode of Action (MoA) signatures of 20 chemicals
[40]This paper combined the Clustering, Association Rules, and Neural Networks for the appraisal of heart-occasion related risk factors, focusing on the reduction of CVD risk	K-Means Clustering, Association Rules, and Neural Network,	Heart Disease dataset
[41]This paper focused on the new technique based on a hybrid model for combining fuzzy segment strategy, and greatest likelihood gauges clustering calculation for diagnosing medical maladies.	Maximum likelihood estimates clustering, Fuzzy Partition Method	Online News Popularity, Iris Dataset, miRNA dataset
[42]This paper built up a solitary pass dynamic rate association control mining calculation	Association Rule Mining	cardiovascular disease, breast cancer, and hepatitis dataset
[43]This paper meant to discover the answers to analyze the illness by breaking down examples found in the dataset through Data Mining.□	Association rule mining, Artificial Neural Network,	UCI repository dataset
[44]This paper presented data mining all in all by condensing mainstream data mining calculations and their applications exhibited in genuine healthcare settings	SVM, K-Means clustering, Apriori Association Rule Mining algorithm	The dataset contains 2,637 de-identified health reports from 696 healthy participants with 906 measurement variables
[45]This paper examined the calculations and instruments used for the utilization of affiliation lead mining.	Association Rule Mining	Healthcare dataset
[46]The paper planned to lead a deliberate audit of the utilization of machine learning, data mining strategies and devices in the field of diabetes look into as for a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic foundation and Environment, and e) Healthcare and Management with the main class seeming, by all accounts, to be the most well known□	Association Rule Mining, Support Vector Machine	Clinical dataset
[47]The objective of this paper is to find illness co-event and arrangement designs from extensive scale tumor determination narratives in New York State	Apriori algorithm	Statewide Planning and Research Cooperative System dataset
[48]This paper displayed an algorithmic look strategy for numerous biomarkers which may foresee or demonstrate Alzheimer's ailment (AD) and different sorts of dementia.	Association rule mining	CAMD database and 5821 patients records
[49]The objective of this paper is to investigate visit malady co-event and successful examples of disease patients in New York State utilizing SPARCS data	Association rule learning	Cancer patients records

D. *Literature Survey on the Big Data analytics for Healthcare domain*

Table 4 depicts the literature survey on the Big Data analytics in the healthcare domain.

TABLE IV: Literature Survey on Application of Big Data Analytics in the Healthcare domains.

Description	Methods used	Dataset and its sample size
[50]In this paper, LDA is utilized to lessen the element and SVM model with a weighted bit work strategy to group more highlights from the information ECG flag	Linear Discriminant Analysis, Support Vector Machine, Multi-layer perceptron, Principal Component Analysis (PCA)□	MHEALTH dataset and number of attributes is 23

[51]There is expanded enthusiasm for conveying big data innovation in the healthcare industry to oversee monstrous accumulations of heterogeneous health datasets, for example, electronic health records and sensor data, which are expanding in volume and variety because of the commoditization of computerized gadgets, for example, cell phones and remote sensors.	Cloud, IoT, Big Data	Healthcare datasets
[52]This paper displayed different diagnostic roads that exist in the patient-driven healthcare framework from the point of view of different partners□	Big Data analytics, Machine Learning	-
[53]This paper investigates the big data execution cases, looked to see how big data examination abilities change authoritative practices, along these lines creating potential advantages□	Big Data analytics	-
[54]This paper gives an understanding of how we can reveal extra an incentive from the data produced by healthcare and government□	Big Data, Hadoop, Map Reduce□	-
[55]To address the potential advantages of big data investigation, this paper analyzed the chronicled advancement, engineering plan and segment functionalities of big data examination□	Big data - analytics architecture, capabilities□	-
[56]The goal of this paper is to build up a structure to upgrade health expectation with the reconsidered combination hub and deep learning paradigms□	Deep Learning, Bayesian functions, Neural Network	Electronic Health Records□
[57]This paper proposed the Internet of Things (IoT) architecture to store and process scalable sensor data (big data) for healthcare applications. Proposed architecture consists of two main sub-architecture, namely, MetaFog-Redirection (MF-R) and Grouping & Choosing□ (GC) architecture	Internet of Things, Big data analytics, MF-R, GC architecture□	Cleveland Heart Disease Dataset
[58]This paper proposed a big data-based learning administration framework to build up the clinical choices. The proposed information framework is produced based on a variety of databases, for example, Electronic Health Record (EHR), Medical Imaging Data, Unstructured Clinical Notes and Genetic Data.□	Big Data analytics	Big Genomics Data
[59]This paper utilized a Bayesian hidden Markov model (HMM) with Gaussian Mixture (GM) Clustering technique to model the DNA duplicate number variation over the genome.□	Big Data, a Bayesian network, HMM, GM, Clustering□□	DNA Genome data
[60]This paper proposed a big data investigation empowered business esteem display in which we utilize the resource-based theory (RBT) and limit structure view to delineate how big data examination capacities can be created and what potential additions can be gotten by these abilities in the healthcare ventures.□	Big data analytics	Genome data
[61]This paper bunches the prior examinations on the Floating Catchment Area theories, a transcendent class of methodologies that contain healthcare accessibility, and presents a structure that conceptualizes receptiveness figuring.□	Big data analytics	Geographic Information Data
[62]This paper proposed a model that uses keen home big data as methods for learning and investigating human action designs for healthcare applications. □	Big Data, Association Rule Mining,	The dataset utilized in this study is a collection of smart meters data from five houses in the UK□
[63]The intention of this paper is application-oriented architecture for big data systems, which is based on a study of published big data architectures for specific	Big data, Machine Learning algorithms	-

use cases. This paper also provides an overview of the state-of-the-art machine learning algorithms for processing big data in healthcare and other applications.		
[64]This work aims at developing a real-time remote health status prediction system built around open source Big Data processing engine, the Apache Spark, deployed in the cloud which focuses on applying machine learning model on streaming Big Data.	Big data machine learning	Heart disease dataset
[65]This paper characterized the traits of disseminated data networks and frameworks the data and scientific foundation expected to fabricate and keep up a fruitful network	Machine Learning algorithm, Big Data	Electronic Health Records
[66]This examination will give researchers in the healthcare informatics network with all-encompassing learning of healthcare big data inquire about and also look into hotspots and future research bearings	Data Mining, Machine Learning, Big Data	Clinical dataset
[67]The most recent decade has seen a remarkable increment in the volume and variety of electronic data identified with innovative work, wellbeing records, and patient self-following by and large alluded to as Big Data	Big data analytics	-

V. Research Analyzation

It is additionally essential to understand that in the present world a patient's restorative information does not only one live inside the breaking points of a healthcare supplier. The medicinal protection scope and pharmaceuticals enterprises additionally hold data about particular cases and the highlights of endorsed medicates individually. Regularly, patient-produced information from IoT techniques, for example, wellness trackers, blood pressure screens, and measuring scales are likewise giving basic data about the everyday way of life attributes of a person. Bits of knowledge got from such information created by the connecting among EMR information, lab information, essential information, prescription data, manifestations and their total.

VI. Conclusion

Nowadays healthcare industries are running from a volume-based business into value-based business, which needs overwork from doctors and nurses to be extra productive and effective. This will increase healthcare practice, changing unique lifestyle and driving them into longer life, prevent diseases, infections and illnesses. Through this survey on various research articles, a new framework will be revealed for the predicting the severity of the disease by using Machine Learning techniques, Big Data analytics, and Data Science.

REFERENCE

[1] Abinash, M. J., and V. Vasudevan. "A Study on Wrapper-Based Feature Selection Algorithm for Leukemia Dataset." *Intelligent Engineering Informatics*. Springer, Singapore, 2018. 311-321.

[2] Chebouba, Lokmane, Dalila Boughaci, and Carito Guziolowski. "Proteomics Versus Clinical Data and Stochastic Local Search Based Feature Selection for Acute Myeloid Leukemia Patients' Classification." *Journal of medical systems*42.7 (2018): 129.

[3] Lee, Shin-Jye, et al. "A novel bagging C4. 5 algorithm based on wrapper feature selection for supporting wise clinical decision making." Journal of biomedical informatics 78 (2018): 144-155.

[4] Venkataraman, Sivakumar, and Rajalakshmi Selvaraj. "Optimal and Novel Hybrid Feature Selection Framework for Effective Data Classification." *Advances in Systems, Control and Automation*. Springer, Singapore, 2018. 499-514.

[5] Chaudhari, Poonam, and Himanshu Agarwal. "Improving Feature Selection Using Elite Breeding QPSO on Gene Data set for Cancer Classification." *Intelligent Engineering Informatics*. Springer, Singapore, 2018. 209-219.

[6] Sundararaman, Arun, Srinivasan Valady Ramanathan, and Ramprasad Thati. "Novel Approach to Predict Hospital Readmissions Using Feature Selection from Unstructured Data with Class Imbalance." *Big Data Research* (2018).

[7] Vivekanandan, T, and Narayana I. N. C. Sriman. "Optimal Feature Selection Using a Modified Differential Evolution Algorithm and Its Effectiveness for Prediction of Heart Disease." *Computers in Biology and Medicine*. 90 (2017): 125-136.

[8] van, Breda W, Vincent Bremer, Dennis Becker, Mark Hoogendoorn, Burkhardt Funk, Jeroen Ruwaard, and Heleen Riper. "Predicting Therapy Success for Treatment As Usual and Blended Treatment in the Domain of Depression." *Internet Interventions*. 12 (2018): 100-104.

[9] Aziz, Mohamed A. E, and Aboul E. Hassanien. "Modified Cuckoo Search Algorithm with Rough Sets for Feature Selection." *Neural Computing and Applications*. 29.4 (2018): 925-934.

[10] Parmezan, Antonio Rafael Sabino, Huei Diana Lee, and Feng Chung Wu. "Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework." *Expert Systems with Applications* 75 (2017): 1-24.

[11] Liu, Lili, et al. "Analysis and prediction of drug–drug interaction by minimum redundancy maximum relevance and incremental feature selection." *Journal of Biomolecular Structure and Dynamics* 35.2 (2017): 312-329.

[12] Hancer, Emrah, et al. "Pareto front feature selection based on artificial bee colony optimization." Information Sciences 422 (2018): 462-479.

[13] Park, Hyun Woo, et al. "A Hybrid Feature Selection Method to Classification and Its Application in Hypertension Diagnosis." *International Conference on Information Technology in Bio- and Medical Informatics*. Springer, Cham, 2017.

[14] Hu, Ying, et al. "Simultaneously aided diagnosis model for outpatient departments via healthcare big data analytics." *Multimedia Tools and Applications* 77.3 (2018): 3729-3743.

[15] Bhatt, Anurag, Sanjay Kumar Dubey, and Ashutosh Kumar Bhatt. "Analytical Study on Cardiovascular Health Issues Prediction Using Decision Model-Based Predictive Analytic Techniques." *Soft Computing: Theories and Applications*. Springer, Singapore, 2018. 289-299.

[16] Ndaba, Moeketsi, Anban W. Pillay, and Absalom E. Ezugwu. "An Improved Generalized Regression Neural Network for Type II Diabetes Classification." *International Conference on Computational Science and Its Applications*. Springer, Cham, 2018.

[17] Mehmood, Raja Majid, Ruoyu Du, and Hyo Jong Lee. "Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors." *cities* 4 (2017): 5.

[18] Kumar, Ashwani, and Tiratha Raj Singh. "A new decision tree to solve the puzzle of Alzheimer’s disease pathogenesis through standard diagnosis scoring system." *Interdisciplinary Sciences: Computational Life Sciences* 9.1 (2017): 107-115.

[19] Nilashi, Mehrbakhsh, et al. "A knowledge-based system for breast cancer classification using fuzzy logic method." *Telematics and Informatics* 34.4 (2017): 133-144.

[20] Dolatabadi, Azam Davari, Siamak Esmael Zadeh Khadem, and Babak Mohammadzadeh Asl. "Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM." *Computer methods and programs in biomedicine* 138 (2017): 117-126.

[21] Manogaran, Gunasekaran, R. Varatharajan, and M. K. Priyan. "Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system." *Multimedia tools and applications* 77.4 (2018): 4379-4399.

[22] Kannan, R., and V. Vasanthi. "Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease." *Soft Computing and Medical Bioinformatics*. Springer, Singapore, 2019. 63-72.

[23] Nilashi, Mehrbakhsh, et al. "A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques." *Biocybernetics and Biomedical Engineering* 38.1 (2018): 1-15.

[24] Yang, Fan, et al. "Improving pain management in patients with sickle cell disease from physiological measures using machine learning techniques." *Smart Health* (2018).

[25] Shanmugam, S., and J. Preethi. "Design of Rheumatoid Arthritis Predictor Model Using Machine Learning Algorithms." *Cognitive Science and Artificial Intelligence*. Springer, Singapore, 2018. 67-77.

[26] Yadav, Bharti, Shilpi Sharma, and Ashima Kalra. "Supervised Learning Technique for Prediction of Diseases." *Intelligent Communication, Control and Devices*. Springer, Singapore, 2018. 357-369.

[27] Perveen, Sajida, et al. "A Systematic Machine Learning Based Approach for the Diagnosis of Non-Alcoholic Fatty Liver Disease Risk and Progression." *Scientific reports* 8.1 (2018): 2112.

[28] Nijeweme-d’Hollosy, Wendy Oude, et al. "Evaluation of three machine learning models for self-referral decision support on low back pain in primary care." *International journal of medical informatics* 110 (2018): 31-41.

[29] Nilashi, Mehrbakhsh, et al. "An analytical method for diseases prediction using machine learning techniques." *Computers & Chemical Engineering* 106 (2017): 212-223.

[30] Alanazi, Hamdan O., Abdul Hanan Abdullah, and Kashif Naseer Qureshi. "A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care." *Journal of medical systems* 41.4 (2017): 69.

[31] Librenza-Garcia, Diego, et al. "The impact of machine learning techniques in the study of bipolar disorder: a systematic review." *Neuroscience & Biobehavioral Reviews* 80 (2017): 538-554.

[32] Kim, Jun S., et al. "Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion." *Spine*43.12 (2018): 853-860.

[33] Tanaka, Tomohiro, and Michael D. Voigt. "Decision tree analysis to stratify risk of de novo non-melanoma skin cancer following liver transplantation." *Journal of cancer research and clinical oncology* 144.3 (2018): 607-615.

[34] Rani, Sangeeta. "A Dual Phase Probabilistic Model for Dermatology Classification." *Computer Communication, Networking and Internet Security*. Springer, Singapore, 2017. 443-450.

[35] Bhatt, Anurag, et al. "Data Mining Approach to Predict and Analyze the Cardiovascular Disease." *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*. Springer, Singapore, 2017.

[36] Bedi, Rajni, and Ajay Shiv Sharma. "Classification Algorithms for Prediction of Lumbar Spine Pathologies." *Advanced Informatics for Computing Research*. Springer, Singapore, 2017. 42-50.

[37] Khanmohammadi, Sina, Naiier Adibeig, and Samaneh Shanehbandy. "An improved overlapping k-means clustering method for medical applications." *Expert Systems with Applications* 67 (2017): 12-18.

[38] Kannan, S. R., et al. "Effective Kernel-Based Fuzzy Clustering Systems in Analyzing Cancer Database." *Data-Enabled Discovery and Applications* 2.1 (2018): 5.

[39] Huang, Z. H., et al. "Development of a data-processing method based on Bayesian k-means clustering to discriminate aneugens and clastogens in a high-content micronucleus assay." *Human & experimental toxicology* 37.3 (2018): 285-294.

[40] Pasanisi, Stefania, and Roberto Paiano. "A Hybrid Information Mining Approach for Knowledge Discovery in Cardiovascular Disease (CVD)." *Information* 9.4 (2018): 90.

[41] Simić, Svetlana, et al. "A Hybrid Clustering Approach for Diagnosing Medical Diseases." *International Conference on Hybrid Artificial Intelligence Systems*. Springer, Cham, 2018.

[42] Borah, Anindita, and Bhabesh Nath. "Identifying risk factors for adverse diseases using dynamic rare association rule mining." *Expert Systems with Applications* 113 (2018): 233-263.

[43] Singh, Pankaj Pratap, et al. "Classification of Diabetic Patient Data Using Machine Learning Techniques." *Ambient Communications and Computer Systems*. Springer, Singapore, 2018. 427-436.

[44] Cheng, Chih-Wen, and May D. Wang. "Healthcare Data Mining, Association Rule Mining, and Applications." *Health Informatics Data Analysis*. Springer, Cham, 2017. 201-210.

[45] Altaf, Wasif, Muhammad Shahbaz, and Aziz Guergachi. "Applications of association rule mining in health informatics: a survey." *Artificial Intelligence Review* 47.3 (2017): 313-340.

[46] Kavakiotis, Ioannis, et al. "Machine learning and data mining methods in diabetes research." *Computational and structural biotechnology journal* 15 (2017): 104-116.

[47] Wang, Yu, Wei Hou, and Fusheng Wang. "Mining co-occurrence and sequence patterns from cancer diagnoses in New York State." *PloS one* 13.4 (2018): e0194407.

[48] Szalkai, Balázs, Vince K. Grolmusz, and Vince I. Grolmusz. "Identifying combinatorial biomarkers by association rule mining in the CAMD Alzheimer's database." *Archives of gerontology and geriatrics* 73 (2017): 300-307.

[49] Wang, Yu, and Fusheng Wang. "Association Rule Learning and Frequent Sequence Mining of Cancer Diagnoses in New York State." *VLDB Workshop on Data Management and Analytics for Medicine and Healthcare*. Springer, Cham, 2017.

[50] Varatharajan, R., Gunasekaran Manogaran, and M. K. Priyan. "A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing." *Multimedia Tools and Applications* 77.8 (2018): 10195-10215.

[51] Shafqat, Sarah, et al. "Big data analytics enhanced healthcare systems: a review." *The Journal of Supercomputing* (2018): 1-46.

[52] Palanisamy, Venketesh, and Ramkumar Thirunavukarasu. "Implications of Big Data Analytics in developing Healthcare Frameworks–A review." *Journal of King Saud University-Computer and Information Sciences* (2017).

[53] Wang, Yichuan, et al. "An integrated big data analytics-enabled transformation model: Application to health care." *Information & Management* 55.1 (2018): 64-79.

[54] Archenaa, J., and EA Mary Anita. "A survey of big data analytics in healthcare and government." *Procedia Computer Science* 50 (2015): 408-413.

[55] Wang, Yichuan, LeeAnn Kung, and Terry Anthony Byrd. "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations." *Technological Forecasting and Social Change* 126 (2018): 3-13.

[56] Zhong, Hongye, and Jitian Xiao. "Enhancing health risk prediction with deep learning on big data and revised fusion node paradigm." *Scientific Programming* 2017 (2017).

[57] Manogaran, Gunasekaran, et al. "Big data analytics in healthcare Internet of Things." *Innovative healthcare systems for the 21st century*. Springer, Cham, 2017. 263-284.

[58] Manogaran, Gunasekaran, et al. "Big data knowledge system in healthcare." *Internet of things and big data technologies for next generation healthcare*. Springer, Cham, 2017. 133-157.

[59] Manogaran, Gunasekaran, et al. "Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering." *Wireless personal communications* (2017): 1-18.

[60] Wang, Yichuan, and Nick Hajli. "Exploring the path to big data analytics success in healthcare." *Journal of Business Research* 70 (2017): 287-299.

[61] Plachkinova, Miloslava, et al. "A conceptual framework for quality healthcare accessibility: a scalable approach for big data technologies." *Information Systems Frontiers* 20.2 (2018): 289-302.

[62] Yassine, Abdulsalam, Shailendra Singh, and Atif Alamri. "Mining human activity patterns from smart home big data for health care applications." *IEEE Access* 5 (2017): 13131-13141.

[63] Manogaran, Gunasekaran, and Daphne Lopez. "A survey of big data architectures and machine learning algorithms in healthcare." *International Journal of Biomedical Engineering and Technology* 25.2-4 (2017): 182-211.

[64] Nair, Lekha R., Sujala D. Shetty, and Siddhanth D. Shetty. "Applying spark based machine learning model on streaming big data for health status prediction." *Computers & Electrical Engineering* 65 (2018): 393-399.

[65] Popovic, Jennifer R. "Distributed data networks: a blueprint for Big Data sharing and healthcare analytics." *Annals of the New York Academy of Sciences* 1387.1 (2017): 105-111.

[66] Gu, Dongxiao, et al. "Visualizing the knowledge structure and evolution of big data research in healthcare informatics." *International journal of medical informatics* 98 (2017): 22-32.

[67] Adam, Nabil R., Robert Wieder, and Debopriya Ghosh. "Data science, learning, and applications to biomedical and health sciences." *Annals of the New York Academy of Sciences*1387.1 (2017): 5-11.