SURVEY ON BIG DATA ANALYTICS IN HEALTH CARE

P.Saranya[1],Dr.P.Asha[2]
Research Scholar[1]Assistant Professor[2],
Department of Computer Science and Engineering[1,2]
Sathyabama Institute of Science and Technology,Chennai[1,2]
asha.cse@sathyabamauniversity.ac.in,saranya.pattusamy@gmail.com

Abstract- Massive amount of data in different forms need to be handled in any healthcare applications. Type of data, size of data, data security and other features has more significance in handling the data. The term big data refers to data with certain characteristics, volume, velocity, value, veracity and variability. Such big data need to be stored, processed, and analyzed for required results. Medical data has more complexity in predicting the results from it, which will have more significance in patient's treatment. Because of its significance, there is need of developing efficient and better performing algorithms, techniques and tools to analyze medical big data. Whereas, the traditional algorithms are not capable for analyzing such complex data. Machine learning algorithms well fit for these kinds of data and analytics. In this survey paper, we discussed about characteristic of big data, features of big data, how to represent big data, different types of machine learning algorithms used in big data analytics. We discussed about big data analytics in major healthcare areas like EHR maintenance, disease diagnose, prediction of emergency condition of patients, etc., .Also stated different machine algorithms usage in disease diagnose and patient's data analysis and discussed about importance of various machine learning algorithms. Here, we have highlighted the areas where big data analytics have been applied in healthcare sectors. It describes the characteristics and features of big data, importance of big data analytics in healthcare sectors, various machine learning algorithms used in big data analytics and their efficiency.

Keywords: Big data, Health care, disease prediction, SVM, CNN

I INRODUCTION

Digitization of medical data reforms the dimensions of data as well as increases the size of data and significance of data analytics. Big data has its own characteristics volume, velocity, veracity, variability, value. Traditional methods to analyses the data are not fit to big data analytics. Big data analytics has its own tools and techniques. Medical data are sensitive data and it has different forms. They are Electronic health records, Administrative information, Claims data, Patient - Disease information, Health surveys, Clinical trials. Imaging has its own significance in various fields of medical research and clinical practice. When the patients have abnormalities and the symptom is clearly exhibited, healthcare schemes which are in practice can treat them. Early diagnosis of intense diseases helps to treat the patient, which reduces the risk. Otherwise it leads to chronic disorders, even death of patient. 59 percent of annual death caused by delayed diagnosis of intense diseases [1].

Types of Features-Big data
Structured Features Linked Data Multisource and Multiview Data Streaming Data and Features Scalability Stability

Fig.no.1. Different features of Big data

Fig.1.Gives different features of big data. Big data processes these feature that make big data different from nominal data.

Fig.2. specifies big data representation and reduction.

Some challenges for the traditional feature selection task. That said, its characteristics also bring new opportunities.[2]

Fig.3 listed various types of algorithms used for performing classification, prediction and clustering in big data.

Data Organization And Processing	Graph Representation Ontology Representation Fuzzy Representation
Data Cleaning And Reduction	Principal Component Analysis (PCA). Kernel Principal Component Analysis (KPCA). Singular Value Decomposition (SVD). Independent Components Analysis (ICA). Linear Discriminant Analysis (LDA). Non-Negative Matrix Factorization (NMF). Canonical Correlation Analysis (CCA). Locally Linear Embedding (LLE). Laplacian Eigenmaps
Data Integration And Processing	Cloud Computing Tensor Networks
Big Data Applications	Media And Entertainment Banking And Finances Insurance Transportation Education Manufacturing Healthcare Energy And Utilities

Fig.no.2 Data Representation and Reduction

II MACHINE LEARNING ALGORITHMS

Classification - Process of Grouping the Data and labeling them. For Example, Banking Customers Are Classified Into Depositors And Borrowers.

Prediction- It is the process of predicting the future based on the current and past data.For example, based on the past transactions made by a customer, the future financial needs of the customers is predicted.

Clustering-Sets are divided into number of subsets or example; borrowers are divided into personal loan borrowers, vehicle loan borrowers, education loan borrowers, etc.

MACHINE LEARNING PARADIGMS FOR BIG DATA

There are different paradigms used for big data. It follows as: Deep learning- It is same as neural networks, used to present data. Many hidden layers are working for data transformation. Online learning-Data set should exist, when process is executing. Local learning-Only subsets of data set are considered for processing. Transfer learning- Based on the knowledge

III BIG DATA IN HEALTHCARE

Big data Analytics in maintaining Healthcare data: Big data in healthcare Electronic health record[4](EHR) stores electronic health information about individual patients. The EHR stores range of electronic health data, including demographics, medical history, medication and allergies, immunization status, laboratory test results, and personal stats like age and weight Based on IOT, sanjeevani EHR used to store the health information of patients and it helps to handle patient data with high availability, security

CLASSIFICATION	Decision tree Linear Classifiers: Logistic Regression, Naive Bayes Classifier. Support Vector Machines. Decision Trees. Boosted Trees. Random Forest. Neural Networks. Nearest Neighbor
PREDICTION	Logistic regression, Linear Regression .Decision Tree. SVM. Naive Bayes. KNN K-Means. Random Forest

CLUSTERING	K-means Mean-Shift Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM) Agglomerative Hierarchical Clustering

Fig.no.3 Different algorithms used for different techniques

from unauthorized use, efficient with large amount of data.

Modern healthcare systems[12] enabled with P4 capabilities is proposed to address healthcare related problems.P4 termed from predictive,preventive,personalized and participatory capabilities.Nano sensors are used to predict the emergency need of medicines to the patients, so that timely action can be taken.Physical Nano sensors for measuring physical characteristics(force, displacement etc), Chemical Nano-sensors for finding type of molecule and bio sensors to identify physiological changes under chemical reactions.

UbeHealth[6] is to overcome the issues in existing networked healthcare domain.ISPDSL-II from 2013 and Waikato-VIII from 2018 are the datasets used with this system. Previous to UbeHealth some healthcare systems are implemented, but they lack in latency, bandwidth, reliability, security and energy efficiency. UbeHealth with edge computing, deep learning, IoT and high performance computing overcome these issues and have achieved better efficiency in some other Qos parameters as well. It has scope for improvement in the features security, privacy with little more reliability.

Based on patient discharge data in the database, disease co-occurrence network is designed.It is to reduce the health care expenditure and for care monitoring.With data mining methods and network analysis, patient data at the time admission (PoA) is analyzed, healthcare is improvised in predicting the required future treatment.

Big data Analytics in predicting and diagnosing diseased: Personalized diabetes diagnosis [5] provides suggestions to patients. It is a 5 G smart diabetes testbed that has wearable like smart clothing.Patient can personalize it according to their requirements, comfortable to wear, sustainable and cost effective. Integration of SVM, ANN and decision tree algorithms is used here for validate the performance of the system.

To reduce the false positive ratio in disease diagnosis (considered heart attack), fuzzy rule based big data analytics-healthcare as a service is proposed [7].Efficient fuzzy based classification with expectation maximization and Cloud-based repository is used for classifying the medical data. This system achieves better performance, accuracy and reduces false positive rate.

Precision diseases are predicted from multi-omic data and EHR data[9].From the biological samples, molecular profiles(genomic, transcriptomic, epigenomic, proteomic, metabolomics) are identified for predicting the diseases. In biomedical data analytics, minimum redundancy maximum relevance method is used for filtering the features.To eliminate the unnecessary features, SVM algorithm is used.By earlier prediction of diseases increases the efficiency of healthcare systems.

An emotion-aware connected healthcare big data framework enabled with IoT techniques [10] is developed to identify the patient medical status.Speech and image are analyzed, based on the emotion of the patient, their state is identified.With Fourier transformation,SVM classifier is used for speech processing and video processing.Accuracy of 99.87% is achieved with this framework.Health data visualization tool [11] is proposed to monitor the health status of the patient.Based on the color circles,the medical status of a person are identified,which user friendly.

Machine Learning Algorithms in medical diagnosis: Machine decision diagnosis auxiliary algorithm [13] is used to diagnosis non-small cell lung cancer with large dataset.It has accuracy of 77% in predicting the disease. Both images and diagnostic parameters are considered by algorithm. An online contextual learning algorithm [14] is used reduces the false positive rate in taking diagnosis decision breast cancer screening. Contextual learning used clinical context to reduce false positives as well. FIsher criterion and genetic optimization, called FIG is used for better recognition of diseases like lung cancer. It has computation efficiency and effectiveness[15].Multiple streams of 2D convolution networks[16], which are multi-view, are used for detecting false positive among the scan. When used with lung CT, 3D CNN perform better. This system has sensitivity of 85% to 90% in detecting false positive. High order back propagation algorithm. Gamification is identified for predicting breast cancer with better efficiency.It is best for handling medical image database.3D Rieszwavelets[19] are used, that perform better in enhancing the inter and intra variations in lung CT.It has accuracy of 81.3 - 82.7% in predicting the recurrence of the disease.This system is capable of predicting the NSCLC,which is most affecting disease.Machine learning algorithm can be used for voice pathology detection.With existing CNN algorithm, transfer learning[20] is used, which increases the accuracy up to 97.5% in Saarbrucken voice disorder database.

Lung cancer patient survival prediction[21] has been done using convolutional auto encoder.Kaplan-Meier analysis and cox proportional hazards divide the patients groups into two, high risk and low risk groups.LASSO-Cox model used for feature selection with Cox model building.High level significance attributes for tumor detection like calcification, sphericity, texture are considered by deep multi-view CNN[22].The performance measure in terms of accuracy has been improvised.

Multi task CNN, Mask R CNN,CNN Ensemble approach, DeepWalk method. etc., are used in different ways in handling tumor patients data like prediction, segmentation, Classification, etc. Deep learning methods[26][27] are used in Neuroradiology,Medical image classification and segmentation,Genomic sequencing and analysis of gene expression,to predict protein structure,to focus on MRI scans.

Machine Learning in Lung Cancer Prediction:

Lung cancer is deadly disease that has higher death rate, it increases the need for predicting the diseases. Hybrid support vector machine, K means, Deep learning,Supervised learning and fusion model are some algorithms used widely for increased accuracy and sensitivity.

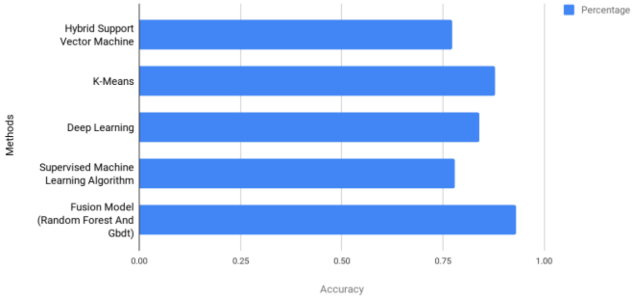

Fig.no.5 Comparison of various techniques

Multiclass SVM

In Multiclass SVM, efficiency is improvised by adding an additional layer in classification level. Multi Class SVM classification of lung cancer images improvises the percentage of accuracy in prediction. Multi class SVM provides 70 percentage of training and 30 percentage of testing on the data. deep learning is to unsupervised learning and can handle unlabelled or unstructured data in neural network. Multi class SVM is implemented with automated Deep Learning using K-means clustering. sensitivity rate and specificity has been up to the scratch. Filtered Images are processed for feature selection and entailed features are selected. Network has been designed to extract the features that are required for classification. Network has to be trained with preconfigured layers, with these training, pre processed images stored in the network to process the feature. After these required steps, Multiclass SVM is applied to obtain the better accuracy level.

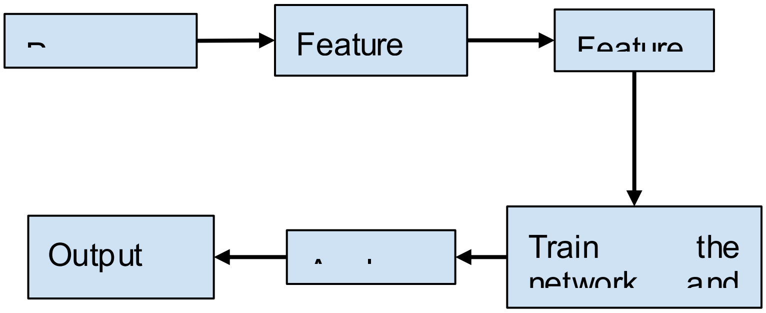

Fig.no.4 Proposed Model for lung cancer image classification

IV DISCUSSIONS AND CONCLUSION

Big data analytics have vast applications over healthcare sector. Big data features and data representations are specified. Big data has various characteristics that have to be analyzed using better algorithm, where traditional algorithms are not capable of. Such kind of better performing algorithms are discussed. Use of machine learning algorithms like CNN, SVM, etc., and their efficiencies in predicting various diseases have been narrated. Different methodologies like DeepWalk, Ensemble methods, etc., used in maintaining healthcare data, to perform efficient segmentation, classification, to analyze scans and for different purposes. We concluded with characteristics and features of big data, importance of big data analytics in healthcare sectors, various machine learning algorithms used in big data analytics and their efficiency.

REFERENCES

1. XiaokangWang, LaurenceT. Yang, Huazhong Liu,M. Jamal Deen, "A Big Data-as-a-Service Framework: State-of-the-Art and Perspectives" in IEEE Transactions on Big Data ,Volume 4 , Issue: 3 , Sept. 1 2018,P 325 – 340.

2. Jundong Li,Huan Liu, "Challenges of Feature Selection for Big Data Analytics", in IEEE Intelligent Systems, Volume 32,Issue 2 ,fo 2017,P 9-15

3. MinChen,Yixue Hao, Kai Hwang,Lu Wang, LinWang,"Disease Prediction by Machine Learning Over Big Data from Healthcare Communities", in IEEE Access, Special Section on Healthcare Big Data, Volume 5,April 2018, DOI 10.1109/ACCESS.2017.2694446,P 8869 - 8879

4. Chandrasekar Vuppalapati, Anitha Ilapakurti, Santosh Kedari,"The Role of Big Data in Creating Sense EHR, An Integrated Approach to Create Next Generation Mobile Sensor and Wearable Data Driven Electronic Health Record", in 2016 IEEE Second International Conference on Big Data Computing Service and Applications (Bigdataservice),DOI:10.1109/Bigdataservice.2016.18

5. MinChen,Jun Yang,Jiehan Zhou, Yixue Hao, Jing Zhang,Chan-Hyun Youn,"5G-Smart Diabetes:Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds", in Advances in next Generation Networking Technologies for Smart Healthcare Volume: 56 Issue: 4 April 2018,P 16 - 23

6. Ubehealth: A Personalized Ubiquitous Cloud and Edge-Enabled Networked Healthcare System for Smart Cities" in *Big* Data Learning and Discovery in Volume: 6 ,June 2018,DOI: 10.1109/ACCESS.2018.2846609T P 32258 - 32285

7. Anish Jindal ,Neeraj Kumar , Athanasios V. Vasilakos, Joel J. P. C. Rodrigues, "Providing Healthcare-as-a-Service using Fuzzy Rule Based Big Data Analytics in Cloud Computing" in IEEE Journal of Biomedical and Health informatics,Volume: 22 Issue: 5 ,Jan 2018,DOI: 10.1109/JBHI.2018.2799198, P 1605 - 1618

8. Karthik Srinivasan , Faiz Currim ,Sudha Ram, "Predicting High-Cost Patients at Point of Admission using Network Science ", in IEEE Journal of Biomedical and Health Informatics Volume: 22 Issue: 6 Nov 2018,DOI: 10.1109/JBHI.2017.2783049.P 1970 - 1977

9. Po-Yen Wu,Chih-Wen Cheng,Chanchala D. Kaddi, Janani Venugopalan,Ryan Hoffman,May D. Wang, " –Omic and Electronic Health Record Big Data Analytics for Precision Medicine" in IEEE Transactions on Biomedical Engineering, Volume: 64 Issue: 2 ,Oct 2016,DOI: 10.1109/TBME.2016.2573285,P 263 - 273

10. Shamim Hossain, Ghulam Muhammad "Emotion-Aware Connected Healthcare Big Data Towards 5GM",in IEEE Internet of Things Journal, VOL. 5, NO. 4, AUGUST 2018 P 2399-2406

11. Antonino Galletta,Lorenzo Carnevale ,Alessia Bramanti, Maria Fazio,"An Innovative Methodology for Big Data Visualization for Telemedicine", in IEEE Transactions on Industrial Informatics ,Volume 15 Issue 1 Jan 2019 P 490 - 497

12. Ali Rizwan ,, Ahmed Zoha, Rui Zhang ., Wasim Ahmad,Kamran Arshad,Najah Abu Ali ,Akram Alomainy ,Muhammad Ali Imran Qammer H. Abbasi ,"A Review on the Role of Nano-Communication in Future Healthcare Systems: A Big Data Analytics Perspective",Volume: 6 July 2018,DOI: 10.1109/ACCESS.2018.2859340, P 41903 - 41920

13.Jiawu,Yanlintan,Zhigangchen,Mingzhao,Decision Based on Big Data Research for Non-Small Cell Lung Cancer in Medical Artificial System in Developing Country", in Computer Methods and Programs in Biomedicine, Volume 159, June 2018, Pages 87-101

14. Linqi Song, William Hsu, Jie Xu, and Mihaela Van Der Schaar, "Using Contextual Learning to Improve Diagnostic Accuracy: Application in Breast Cancer Screening" in Volume: 20 Issue: 3 Mar 2105, P 902 - 914

15. Xiabi Liu, Ling Ma, Li Song, Yanfeng Zhao, Xinming Zhao, Chunwu Zhou, "Recognizing Common CT Imaging Signs of Lung Diseases Through a New Feature Selection Method Based on Fisher Criterion and Genetic Optimization", in IEEE Journal of Biomedical and Health Informatics, Volume: 19 Issue: 2 ,Jun 2014,P 635 - 647

16. Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, ColinJacobs, Sarah J. Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I. Sánchez, and Bram Van Ginneken,""Pulmonary Nodule Detection in CT Images: False Positive Reduction using Multi-View Convolutional Networks", in IEEE Transactions on Medical Imaging,March 2016 Volume 35 Issue 5 P 1160 - 1169

17. Zhennan Yan, Yiqiang Zhan, Zhigang Peng, Shu Liao, Yoshihisa Shinagawa, Shaoting Zhang, Dimitris N. Metaxas, Xiang Sean Zhou, "Multi-Instance Deep Learning: Discover Discriminative Local Anatomies for Bodypart Recognition" in IEEE Transactions on Medical Imaging May 2016 Volume 35 Issue 5 P 1332 - 1343

18. Shadi Albarqouni, Christoph Baur, Felix Achilles, "Aggnet: Deep Learning from Crowds for Mitosis Detection in Breast Cancer Histology Images", in IEEE Transactions on Medical Imaging May 2016 Volume 35 Issue 5 1313 - 1321

19. Pol Cirujeda, YashinDicente Cid, Henning Müller, Daniel Rubin, Todd A. Aguilera, Billy W. Loo, Jr.,Maximilian Diehn, Xavier Binefa,Adrien Depeursinge, "A 3-D Riesz-Covariance Texture Model for Prediction of Nodule Recurrence in Lung CT ",in IEEE Transactions on Medical Imaging Dec 2016 Volume 35 Issue 12 P 2620-2630

20. Musaed Alhusseinand Ghulam Muhammad, "Voice Pathology Detection using Deep Learning on Mobile Healthcare Framework ",in IEEE Access Mobile Multimedia for Healthcare Volume, Jul 2018 6 41034 - 41041

21. Shuo Wang, Zhenyu Liu, Xi Chen, Yongbei Zhu, Hongyu Zhou, Zhenchao Tang, Wei Wei, Di Dong,Meiyun Wang, Jie Tian, "Unsupervised Deep Learning Features for Lung Cancer Overall Survival Analysis" in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) July 2018, 10.1109/EMBC.2018.8512833

22. Sarfaraz Hussein, Robert Gillies , KunlinCao , Qi Song , Ulas Bagcil, "Tumornet: Lung Nodule Characterization using Multi-View Convolutional Neural Network with Gaussian Process", in IEEE International Symposium on Biomedical Imaging (ISBI) 2017.DOI:10.1109/EMBC.2018.8512833
23. Naji Khosravan, Ulas Bagci, "Semi-Supervised Multi-Task Learning for Lung Cancer Diagnosis",in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 10.1109/EMBC.2018.8512294 DOI: 10.1109/IJCNN.2018.8489345inQi,
24. Rahul Paula , Lawrence Hall , Dmitry Goldgof , Matthew Schabath , and Robert Gillies,, "Predicting Nodule Malignancy using a CNN Ensemble Approach", in 2018 International Joint Conference on Neural Networks (IJCNN)
25 Guanghui Li , Jiawei Luo , Qiu Xiao , Cheng Liang , Pingjian Ding , Buwen Cao, "Predicting Microrna-Disease Associations using Network Topological Similarity Based on Deepwalk",in IEEE Access, Volume 5,DOI: 10.1109/ACCESS.2017.2766758
26. X G. Zaharchuk, X E. Gong, X M. Wintermark, X D. Rubin, and X C.P. Langlotz, "Deep Learning in Neuroradiology", in American Journal of Neuroradiology October 2018, 39 (10) 1776-1784; DOI: Https://Doi.Org/10.3174/Ajnr.A5543
27. Chensi Cao , Feng Liu , Hai Tan , Deshou Song , Wenjie Shu , Weizhong Li, Yiming Zhou , Xiaochen Bo, Zhi Xie,"Deep Learning and its Applications in Biomedicine", in Genomics, Proteomics & Bioinformatics Volume 16, Issue 1, February 2018, Pages 17-32. Https://Doi.Org/10.1016/J.Gpb.2017.07.003
28. Alexander Selvikv, Lundervolda, , Arvid Lundervolda, "An Overview of Deep Learning in Medical Imaging Focusing on MRI", in Zeitschrift Für Medizinische Physik, Dec 2018,Https://Doi.Org/10.1016/J.Zemedi.2018.11.002