DESIGN OF MACHINE LEARNING TOOLS FOR BIG DATA ANALYTICS

Mrs A VAIDEGHY1, Dr C THIYAGARAJAN2

*1Assistant Professor, Department of Computer Technology, PSG College of Arts & Science, Coimbatore, Tamilnadu, INDIA.*

*2Associate Professor, Department of Computer Technology, PSG College of Arts & Science, Coimbatore, Tamilnadu, INDIA.*

1vaideghy@psgcas.ac.in

2thiyagarajan@psgcas.ac.in

***Abstract— Big data has become more prevalent in healthcare in recent years, owing to three primary factors: the vast amount of data available, rising healthcare expenditures, and a focus on personalised care. Big data processing in healthcare refers to the generation, collection, analysis, and storage of clinical data that is too large or complicated to be inferred by traditional data processing methods. The Internet of Things (IoT), Electronic Medical Record/Electronic Health Record (EMR/EHR), which contains a patient's medical history, diagnoses, medications, treatment plans, allergies, laboratory and test results, genomic sequencing, Medical Imaging, Insurance Providers, and other clinical data, are examples of big data sources for healthcare. This paper discusses numerous machine learning methods that were used on diverse healthcare data sets. Furthermore, the difficulties of processing and handling*** ***big data, and their applications. The scope of the paper is to elaborate on the application of machine learning algorithms and the need for handling and utilizing big data from a different perspective.***

***Keywords***— **Big Data, Machine Learning, Healthcare, Genomic Sequencing, Data Processing.**

Introduction

Machine Learning is a subfield of artificial intelligence that refers to the ability of IT systems to find solutions to problems on their own by recognising patterns in databases. Machine Learning enables IT systems to recognise patterns and build appropriate solution concepts based on current algorithms and data sets. As a result, artificial knowledge is formed in machine learning based on experience. To learn from data sets, statistical and mathematical methods are utilised in machine learning. There are two types of approaches: symbolic approaches and sub-symbolic approaches. Sub-symbolic systems are artificial neural networks, whereas symbolic systems are propositional systems in which the knowledge content, i.e., the induced rules and examples, is openly recorded. These operate on the basis of the human brain, in which the knowledge contents are implicitly represented. The critical issues of machine learning for big data are large scale data, different types of data, high speed of streaming data, uncertain and incomplete data [1]. The three main types of machine learning are supervised, unsupervised, and reinforcement learning.

LITERATURE SURVEY

This article compares the assessments made by many writers in tackling a clinical challenge in the domain of machine learning.

Machine Learning, as a part of AI, teaches the system from previously collected data to recognise patterns and make judgements with minimal human intervention. Support Vector Machine (SVM), logistic regression, clustering, and other such techniques are examples.

Kai Hwang et al. [2] suggested a big data-applicable Convolution Neural Network-based multimodal disease risk prediction (CNN-MDRP) algorithm. The accuracy of the disease risk model is determined by combining structured and unstructured variables.

Ya Zhang and Tao Zheng[3] proposed a semi-automated method based on machine learning that makes use of a large data EHR database. A supervised learning algorithm served as the foundation for the framework. Because raw EHR is frequently unstructured and sparse, feature engineering was required to correctly structure it. A total of 16 features were extracted and built to be used in the machine learning framework. The three machine learning methods Random Forest, Logistic Regression, and AdaBoost were applied, and the results were improved. The algorithms additionally optimise the filtering criteria to boost recall while minimising false-positives.

Jyotishman Pathak et al.,[4] proposed the performance of the SHFM (Seattle Heart Failure Model) using EHR at Mayo Clinic, with the goal of developing a risk prediction model using machine learning techniques that use routine clinical care data.

Andy Schuetz et al.,[5] suggested recurrent neural network (RNN) models based on gated recurrent units (GRUs) to find relationships between time-stamped events (e.g., disease diagnosis, medicine prescriptions, procedure orders, etc.) during a 12 to 18-month observation window of patients and controls.

Shulong Zhang et al. [6] proposed an LSTM (long short term memory) prediction model framework for HF diagnosis.

Jana Hoffman et al. [7] proposed a machine-learning classification system that uses multivariable combinations of easily obtained patient data (vitals, peripheral capillary oxygen saturation, Glasgow Coma Score, and age), to predict sepsis using the retrospective Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-Ⅲ dataset, restricted to intensive care unit (ICU) patients aged 15 years or more.

Arjun K. Venkatesh et al. [8] proposed a machine learning approach to predict in-hospital mortality of ED(emergency department) patients with sepsis which is a big data-driven approach.

Susan E. Clare et al. [9] proposed a novel concept-based filter and a prediction model to detect local recurrences using EHR in breast cancer patients.

Thiyagarajan C., Anandha Kumar K., Bharathi K., [10] has described the various machine learning approaches in detecting diabetes levels.

Barstugan M., Ozkaya U., and Ozturk S.,[[11]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7309744/%22%20%5Cl%20%22b0080) has adopted the machine learning methods for  Coronavirus detection in early stage with the aid of CT Scan images. The proposed method has been adopted for 53 cases obtained from 150 numbered abdominal images CT scans and 150 numbered images of chest CT images. In this study, Images were cropped and the required features have been extracted using various feature extraction techniques. The classifier, Support Vector Machine classifies the features extracted. The finest outcome was obtained from the feature extraction techniques of GLSZM consisting of an accuracy rate of 99.68%.

Naseer Qreshi K, Din S., & Jeon G [12] has proposed a precise, vibrant, smart M-Health system. Their work invokes a predictive model that works in the machine learning domain supporting collection of data, pre-processing the data, partitioning the data, learning the algorithm and making decisions based on the trained datasets.

Lalmuanawma, S., Hussain, J., & Chhakchhuak, L [13] highlights the applications of various machine learning approaches for the outbreak situation of Coronavirus. Their work reveals various measures to address the pandemic.

Shafaf N., & Malek H [14] encompasses the uses of machine learning approaches in emergency medicine. Their study compiles and evaluates the approaches in existing studies to recent years in the field of artificial intelligence related to emergency medicine.

Christopher Tack[15] has labeled the vital applications of machine learning in musculoskeletal medicine.

Schwartz, J. M., Moy, A. J., Rossetti, S. C., Elhadad, N., & Cato, K. D[16] describes the support rendered to the clinicians through machine learning systems in analyzing electronic medical record data and carrying their diagnosis and treatment process.

Garg A., & Mago V[17] suggests the various machine learning approaches supportive to the medical field.

[Hang Lai](https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6#auth-Hang-Lai), [Huaxiong Huang](https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6%22%20%5Cl%20%22auth-Huaxiong-Huang), [Karim Keshavjee](https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6#auth-Karim-Keshavjee), [Aziz Guergachi](https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6#auth-Aziz-Guergachi) & [Xin Gao](https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6%22%20%5Cl%20%22auth-Xin-Gao) [18] have made a comparison study on the algorithms such as GBM and Logistic Regression models and finally has arrived at the result that, they outperforms the methods such as random forest and decision tree models.

Above works from distinct authors suggest numerous ideas in identifying the disease using various machine learning techniques. Though the research done is large in number, the lack of appropriate methods for identifying the disease encourages us to crack out new methods.

MACHINE LEARNING APPROACHES

*SUPPORT VECTOR MACHINE (SVM)*

Support Vector Machine is a machine learning algorithm that belongs to supervised learning and suitable for dealing with problems related to classification and regression. The algorithm generates the best decision boundary called hyper plane for segregation of n-dimensional space into classes to put new data points in the correct category. Hyper plane is created by choosing the extreme vector called the support vector.

*RANDOM FOREST CLASSIFIER*

Random Forest Classifier is a machine learning algorithm that supports supervised learning favoring problems related to classification and regression problems. In solving complex problems, this algorithm implies the feature of ensemble learning to combine multiple classifiers. Also this approach improves the performance of the system. This classifier improves the prediction accuracy of a particular dataset by computing the average of decision trees on various subsets for the given dataset.

*K-NEAREST NEIGHBOUR CLASSIFIER*

A non-parametric and most suited machine learning algorithm for classification problems related to supervised learning. The algorithm works by making a similarity assumption amongst new and available cases, categorizing by locating the new case that is most similar to the available categories. Based on similarity, the algorithm classifies the new data point from the available data. The algorithm is otherwise called a lazy learner algorithm as it does not learn from the training set immediately. Instead, during classification, from the dataset stored, it implements action. During the training phase, the algorithm just stores the dataset. At the time of facing new data, data will be classified and categorized based on the similarity with the new data.

*GRADIENT BOOST ALGORITHM (GBM)*

Gradient Boosting Machine algorithm generates final predictions by combining predictions from multiple decision trees. Decision trees are constructed from the weak learners. Every node in the decision tree takes features of different subsets in order to select the best split. This feature allows capturing different signals from the data. Furthermore, every new tree counts the errors committed by the previous trees and builds a successive decision tree. This act builds the trees sequentially.

*LOGISTIC REGRESSION MODEL*

This algorithm also supports machine learning concepts belonging to the supervised learning category. This model predicts the output of a categorical dependent variable obtained from the independent variables.

*GREY WOLF OPTIMIZATION (GWO)*

This is a population-based, meta-heuristics algorithm that pretends the leadership hierarchy and the hunting mechanism of grey wolves in nature. Grey wolves are measured as apex predators bagging the top of the food chain. Grey wolves prefer to live in groups (packs), each group containing 5-12 individuals on average. In the group, each individual follows a firm social dominance hierarchy.

*BAT ALGORITHM (BA)*

BA is a recent, swarm based meta-heuristic algorithm for global optimization built on the echolocation trait of microbats. The algorithm can be applied for problems related to missing data estimation.

*FIREFLY ALGORITHM (FA)*

FA is a recent, meta-heuristic algorithm built on the trait of fireflies and their patterns of flashing. The algorithm can be applied for problems related to missing data estimation.

PERFORMACE MEASURES IN MACHINE LEARNING

Accuracy - The number of correct classifications made by a model (true positives and true negatives) divided by the total number of predictions made.

Calibration - A measure of how closely predicted probabilities for an outcome match the observed outcome in test data, e.g., the Brier score.

Discrimination - A measure of how well a model discriminates between randomly selected true positive cases and true negative cases, usually measured as the area under the receiver operator curve.

Negative predictive value - The total number of correct negative classifications made (true negatives) divided by the total number of negative classifications made (true negatives and false negatives)

Precision (also called positive predictive value) - The total number of correct positive classifications made (true positives) divided by the total number of positive classifications made (true positives and false positives).

Recall (also called sensitivity or the true positive rate) - The total number of correct positive classifications made (true positives) divided by the number of positive class members in the data (true positives and false negatives).

Specificity (also called the true negative rate) - The total number of correct negative classifications made (true negatives) divided by the number of negative class members in the data (true negatives and false positives)

Conclusions AND FUTURE SCOPE

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

References

1) Xu Y, Qiu J, Wu Q, et al. (2016) “A survey of machine learning for big data processing”, EURASIP J Adv Signal Proc 2016: 67

2) Hao Y, Hwang MCK, Wang L, et al. (2017) Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5: 8869-8879.

3) Zhang Y and Zheng T, (2017) A big data application of machine learning-based framework to identify type 2 diabetes through electronic health records, In: International Conference on Knowledge Management in Organizations, Springer, 451-458.

4) Pereira N, Taslimitehrani V, Pathak J, et al. (2015) Using EHRs and machine learning for heart failure survival analysis. Stud Health Technol Inf 216: 40.

5) Stewart WF, Sun J, Choi E, et al. (2017) Using recurrent neural network models for early detection of heart failure onset. J Am Med Inf Assoc 24: 361-370.

6) Liu Z, Zhang S, Jin B, et al. (2018) Predicting the risk of heart failure with EHR sequential data modeling. IEEE Access 6: 9256-9261.

7) Calvert J, Hoffman J, Jay M, et al. (2016) Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. JMIR Med Inf 4: e28.

8) Hall MK, Pare JR, Venkatesh AK, et al. (2015) Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. Acad Emerg Med 23: 269-278.

9) Neapolitan R, Zexian SE, Roy A, et al. (2018) Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinfor 19: 65-74.

10) C. Thiyagarajan, K. Anandha Kumar, A. Bharathi “A Survey on Diabetes Mellitus Prediction using Machine Learning Techniques,” International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 3 (2016) pp 1810-1814

11) M. Barstugan, U. Ozkaya, and S. Ozturk, “Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods,” 5, pp. 1–10, 2020.

12) Naseer Qureshi K, Din S, Jeon G, Piccialli F. An accurate and dynamic predictive model for a smart M-Health system using machine learning. Inf. Sci. 2020;538:486–502. doi: 10.1016/j.ins.2020.06.025.

13) Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons Fractals. 2020;139:110059. doi: 10.1016/j.chaos.2020.110059.

14) Shafaf, N., Malek, H.: Applications of Machine Learning Approaches in Emergency Medicine; a Review Article. Arch. Acad. Emerg. Med. 7, (2019).

15) Tack C. Artificial intelligence and machine learning | applications in musculoskeletal physiotherapy. Musculoskelet. Sci. Pract. 2019;39:164–169. doi: 10.1016/j.msksp.2018.11.012.

16) Schwartz JM, Moy AJ, Rossetti SC, Elhadad N, Cato KD. Clinician involvement in research on machine learning–based predictive clinical decision support for the hospital setting: A scoping review. J. Am. Med. Inform. Assoc. 2021;28:653–663. doi: 10.1093/jamia/ocaa296.

17) Garg A, Mago V. Role of machine learning in medical research: A survey. Comput. Sci. Rev. 2021;40:100370. doi: 10.1016/j.cosrev.2021.100370.

18) [Hang Lai](https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6#auth-Hang-Lai), [Huaxiong Huang](https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6%22%20%5Cl%20%22auth-Huaxiong-Huang), [Karim Keshavjee](https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6#auth-Karim-Keshavjee), [Aziz Guergachi](https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6#auth-Aziz-Guergachi) & [Xin Gao](https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6%22%20%5Cl%20%22auth-Xin-Gao), “Predictive    models for diabetes mellitus using machine learning techniques”, BMC Endocrine Disorders, Article Number. 101, 2019.