International Journal of Early Childhood Special Education (INT-JECS)
DOI:10.9756/INTJECSE/V14I7.118 ISSN: 1308-5581 Vol 14, Issue 07 2022

INVOLVEMENT OF MACHINE LEARNING ALGORITHMS IN HEALTH CARE DATA DIAGONOSIS

Vidya Gopinath
Research Scholar
Department of Computer science and Engineering
JAIN (Deemed-To-Be University), Bangalore

Dr.MK Jayanthi Kannan
Professor
Department of Computer science and Engineering
Faculty of Engineering and Technology,
JAIN (Deemed-To-Be University), Bangalore

Abstract

In the current health care scenario, there are many illness conditions which need to be recognised at their early stages to start correct treatments. If not, they could be chronic and deadly. Because of that there is a demand of analysing complex medical data, medical representations, medical reports, and at a marginal time but with higher accuracy. There are even some incidents where certain deformities cannot be directly recognized by individuals.

In medical field for computational decision making, machine learning approaches are being used in these types of circumstances where a crucial data analysis needs to be conduct on medical data to release hidden relationships or chronic abnormalities which are not visible to Individuals. Implementing algorithms to carry out such tasks itself is difficult, but what makes it even more challenging is to increase the correctness of the algorithm while reduce the required time for the algorithm to execute. In the early days, processing of large amount of medical data was a predominant task which resulted in machine learning being adjusted in the biological domain.

Alteration in machine learning algorithms are being performed and tested daily to improve the correctness of the algorithms in analysing and presenting more precise Information. Machine learning has been involved in the healthcare field right from information extraction that involves medical documents until the diagnosis or prognostic of a disease. Throughout this study, various machine learning algorithms and procedure that are being used for decision making in the healthcare sector will be talk about along with the involvement of machine learning in medical healthcare applications in the current scenario. When giving review towards the broad view by combining the findings, it is noticeable that computational biology and biomedicine-based decision making in healthcare have now become dependent on machine learning algorithms, and thus they cannot be segregate from the field of artificial intelligence.

1. Introduction

Artificial Intelligence includes certain techniques and approaches and like machine learning algorithms and machine reasoning. In this analysis, the key emphasis will be towards ML or machine learning techniques as it is being applied using diverse algorithms and techniques in various industries, healthcare activities and analysis.

Using machine learning techniques to resolve various clinical problems is critical and named as one of the revolutionary clinical decision-making processes. Clinical decision making combined with machine learning implies that the system will observe a given individual by gathering and rendering data related to the health of that individual and suggest the best actions that need to be executed to sustain or improve the individual's well-being. In this context of machine learning, the scheme needs to study the context of the problem and the quality of data that is provided. These algorithms are, usually, not very robust and concrete at the start, but by performing repetitive tasks, the algorithm improves and becomes strong with greater number of previous practices.

There are two discrete methods that any clinical decision-making solution infers by associating previous data that is included in the dataset. This is one of the intuitive or fastest approach which uses original clinical pattern recognition and is generally used in many medical emergencies. However, these have a higher likelihood of being flawed and provide partial perception. In today's times, machine learning applications have contributed immensely to the healthcare sector across the world to improve its quality and will continue to do so in the future as well.

As an example, considering the current scenario that has arisen with COVID-19 pandemic, it can be seen that every one of the above mentioned points need to be measured in healthcare and such tasks need to be carried out within a significantly small time duration. In short, the best method is to use ML based decision making in the healthcare industry in such times. This is one of the primary the reason for the demand "emergency machine learning" in the present world.

Therefore, it is crucial to give thought when implementing such systems and balance them with the advantage that they are created with more efficient healthcare systems by the high and accurate computational power of computing at a considerably lower cost. In addition, algorithms in artificial intelligence have the ability to carry out computerized predictive analysis by ordering, filtering and finding for patterns from large datasets from different sources to provide quick and informed decisions. In the current context, due to this discussed matter, most authorities do not allow directly applying these procedures to make the final decisions, but rather use them as an assistance for diagnosis.

Throughout the paper various machine learning techniques and algorithms and their usage in the field of computational decision making for healthcare will be considered along with the approaches that are used to enhance the efficiency of the algorithms with the target to highlight the importance of scalable machine learning algorithms in healthcare sectors. The aim of this paper is to discuss the contribution of machine learning algorithms in the healthcare industry to perform computational decision making right from the early phase where machine learning was announced to computational biology till the peak it stands today which is the introduction of precision medicine in biomedicine. This paper is categorised in different segments. The discussion part will comprise of an assessment and comparison of machine learning algorithms with respect to the applications of such algorithms in healthcare which will be followed by recognising different mechanisms used to improve the accuracy

of the algorithms, along with additional information on the involvement of scalable machine learning algorithms for cognitive decision making in healthcare. The latter part of the paper consists of the conclusion.

2. Machine Learning Methods and Algorithms

Machine learning can be presented as a scientific discipline that emphases how computers study data and continuously improve themselves. It is largely grounded on statistics and even probability. However, it is more powerful than the typical statistical practices when it comes to decision making. Information collected from a dataset that is being consumed by the algorithm is called features. The correctness of the predictions made by the model is reliant on the quality of the features given to the algorithm. It is the responsibility of a machine leaching developer to detect the subset of features that could be suitable for the purpose, thereby increasing the accuracy of the model, which is not an easy task. Continuous trials should be performed to identify the said feature subset for the algorithm. During such a performance measurement, it is also critical to lower the bias and to increase the variance in this testing phase. A good machine learning algorithm must improve the bias-variance trade-off. The assessment of the final machine learning algorithm performance is carried out based on the validation dataset in the validation period. Initially, it would be better to have knowledge about various approaches taken in machine learning along with numerous algorithms that are being used excessively for classification and clustering purposes in machine learning.

2.1. Supervised Learning

In supervised learning, a training set is supplied with suitable objectives during this approach. Regression and classification are the two categories found in supervised learning. With the use of classification methods, the trained scheme allocates inputs into classes. In regression, the sources are continuous instead of discrete. The root-mean-squared error is used to assess regression predictions, while correctness is being used to evaluate classification predictions. Supervised learning has the aim to predict a known output based on a shared dataset. Tasks carried out by supervised learning can be performed by trained personnel most of the times. Supervised learning emphases on classification which includes selection among subgroups to best describe a new instance of data and its prediction, which involves approximating an unknown parameter. Often this is used to estimate and model any risks while finding relationships which are not readily noticeable to humans. Below are some of the supervised learning algorithms which are extensively used in the field of computational biomedicine and biology.

2.1.1. K-Nearest Neighbour (KNN)

KNN is a widespread supervised classification algorithm that is used in several fields such as intrusion detection, pattern recognition, and so on. KNN is a simple algorithm that is much easy to understand. In addition, the precision is high in KNN, but some of the issues about this algorithm are that it is computationally quite expensive, and it has a greater memory requirement as both training and testing data need to be kept. Prediction a new instance is attained by finding the

most similar cases at first and then briefing the yielded variable according to these similar cases. Mean value is for regression, and mode value is for classification. To regulate similar instances, the distance measure is calculated. Euclidean distance is one of the most widespread approaches that is used to compute the distance. The training dataset should contain vectors that is in a multidimensional feature space, with each containing a class label.

2.1.2. Support Vector Machine (SVM)

SVM is an example of supervised machine learning algorithm that is not only used to report mainly classification issues but also for regression problems. In this algorithm, firstly, the information items are plotted as points in an n-dimensional space with feature value being the actual coordinate. It, then, identifies the hyperplane that splits the datapoints into two different classes. Now, the marginal distance between the instances and decision hyperplane that are close to the edge can be increased. SVM can map points with other dimensions using nonlinear relationships and this makes SVM ahead in comparison with other algorithms. SVM is also called nonprobabilistic binary classifier as it divides datapoints into two classes. SVM has greater accuracy in comparison with several other algorithms. This algorithm can solve both nonlinear and linear problems, but nonlinear SVM is better preferred compared to linear SVM due to its better performance.

2.1.3. Decision Trees (DTs)

Decision Tree is a supervised algorithm that has a tree like model through which decisions, any possible consequences, and their results are being measured. Each node transports a question, and each branch signifies a result. Class labels are the leaf nodes. When a sample data reaches the leaf node, the label of the matching node will be allocated to the sample. This approach is best suited when the problem is easy and when the sample dataset is small. Even though this algorithm is easy to realize, it has certain concerns such as biased outcomes and overfitting problem when working with extreme datasets. But Decision Tree is capable of plotting both nonlinear and linear relationships.

2.1.4. Classification and Regression Trees (CARTs)

This is a predictive approach where the output value is forecasted based on the current values in the constructed tree. CART model is represented as a binary tree in which each root signifies a single input and a split point on that variable. The leaf nodes contain an output which is then used to make predictions.

2.1.5. Logistic Regression (LR)

This is a widespread mathematical modelling process which is then used for epidemiologic datasets in machine learning. It initially calculates using the logistic function. Coefficients for logistic regression model is learned and then it finally makes forecasts using that this logistic regression model. This is a generalized linear model and contains two parts, link function and linear part. The linear part is accountable for performing calculations of the classification model, and the link function is used for getting the output of the calculation. Logistic Regression is a supervised algorithm which needs a suggestion or hypothesis

and also a cost function. Optimizing the cost function is also important.

2.1.6. Random Forest Algorithm (RFA)

RFA is one of the trending machine learning techniques that is capable to carryout both classification and regression. It is in fact a supervised learning algorithm in which the base procedure is recursion. A group of decision trees are created in this algorithm and the bagging technique is used for the training purpose. Random Forest Algorithm is insensitive to most noise and it can also be used for most imbalanced datasets. In addition, the problem of over fitting is also not protuberant in RFA.

2.1.7. Naive Bayes (NB)

This is an example of a classification algorithm that is used for multiclass and binary problems. The classifiers are a group of classifying algorithms that are based on the Bayes theorem. However, they all obey a common principle that each pair of features that are being classified must be self-governing with each other. This is in fact like SVM's, with the difference that this process takes advantage from statistical approaches. When there is a new input using this input, the probabilistic value will be considered among the classes regarding any given input and the data will be categorized with the class which has the maximum probabilistic value for that given input.

2.1.8. Artificial Neural Network (ANN)

This is a supervised machine learning method which is well branded for image classification complications. In machine learning, artificial neurons are measured to be the elementary concept of ANN and it is comparable to an organic neural network. ANN contains 3 layers, and all the nodes in each layer is linked with all the nodes in other layers. By increasing the number of hidden layers, a deeper neural system can be created. There are three types of functions in neural networks. The search function will classify the changes that would decrease the error function. The error function will regulate how good or bad the output was for an agreed set of inputs. The update function will govern how the changes will be made as per the search function. This iterative process would finally improve the overall performance of the algorithm.

2.2. Unsupervised Learning

When there is no clear understanding of the data that are involved with the method, it is highly impossible to label the data and deliver them as the training dataset. In such cases, the machine learning algorithms themselves can be used to sense differences and similarities between the data objects. This is known as the unsupervised method of machine learning. By this approach, existing patterns will be recognized, and the data will be bunched according to the identified patterns. As an example, when grouping people according to their environment, genetics, and medical history, few relationships that were not visible before might get recognized by unsupervised machine learning algorithms. Mean shift, K-means, affinity propagation, Gaussian mixture modelling, density-based spatial clustering of applications with noise (DBSCAN), Markov random fields, and fuzzy C-means systems, iterative self-organizing data (ISODATA) are some examples for unsupervised algorithms.

Grouping or clustering is an approach in unsupervised learning, and it can be used for separating inputs into clusters. But these bunches are not identified primarily but are grouped based on similarities. In clustering, the root methods are separated as per the different structures that they carry..

2.2.1. Partition Clustering

Here, the objects are separated and may change clusters based on the difference. It is useful in bioinformatics when the number of clusters is decided such as for a minor gene expression dataset. The drawback of this method is that the user needs to physically enter the number of clusters as the input. This approach is, however, very commonly used in bioinformatics. COOLCAT, Fuzzy k-means, clustering large applications based on randomized search (CLARANS) and clustering large applications (CLARA) are some examples of partition clustering algorithms.

2.2.2. Graph-Based Clustering

Markov cluster algorithm (MCL), super-paramagnetic clustering (SPC), restricted neighbourhood search cluster (RNSC) and molecular complex detection (MCODE) are some examples for graph-based clustering algorithm.

2.2.3. Hierarchical Clustering

Here the objects are separated into a tree of nodes and these nodes are then considered as clusters. There are child and parent nodes. A node will have just one parent, and each of the node can have zero or even more child nodes. This approach is most popular in bioinformatics as clusters can be directed at various levels of granularity. The cons are that they are mostly slow, and errors made when integrating clusters cannot be corrected even though it affects the result, also if large clusters are combined, then some interesting local cluster structure may be gone. This approach is used to signify protein sequence family associations and also could be used to display gene relations reflecting their gene resemblance. Robust clustering using links (ROCK), chameleon, spectral and scalable information bottleneck (LIMBO) are some examples for hierarchical clustering algorithms.

2.2.4. Density-Based Clustering

This method uses the local density principle, and also the clusters are subspaces during which the objects are compressed and are also detached by subspaces of low density. It is often used in bioinformatics in order to discover the densest subspaces in interactome systems, especially involving cliques. The ability to find clusters of arbitrary shapes and time efficiency are the rewards of using this method. Few of these procedures accept user limitations, but it is not the amount of clusters. Clustering in quest (CLIQUE), Ordering points to identify the clustering structure (OPTICS), clustering categorical data using summaries (CACTUS) and density based clustering (DENCLUE) are some examples for density-based clustering algorithms.

2.2.5. Model-Based Clustering

It is presumed that objects match a model which is often belonging to a statistical distribution. The model can be made user specified using a constraint, and can even

be changed in the course. This approach can even be found in bioinformatics to integrate background knowledge into gene interactomes, expressions and sequences. Time to process large datasets, which most of the times, is slow is a major drawback of this procedure. When defining the models, if the user expectations are false, then the results will also be erroneous. COBWEB, SVM-based clustering, and AutoClass are some model-based clustering algorithms.

2.3. Semi supervised Learning

For this method, a partial training set of data is provided. This type of training is implemented when some mislaid results could be targeted by some training data. Semi supervised learning procedures are trained on both unlabelled and labelled data. Owing to this reason, it exhibits the features of both unsupervised and supervised machine learning algorithms.

2.4. Evolutionary Learning

This technique is mainly used in the biology field in order to learn about biological organisms and forecast their survival rate. Using this method, the level of accuracy of a result can also be predicted.

2.5. Active Learning

The system gets the training labels only for a restricted set of incidences. By using it, the optimality of the substances can be improved to gain labels for the essential goal. The benefit in this approach is that the procedure not only continuously studies, but also gets the actualities which were self-learnt accepted either by querying a user or a data source in an interactive way. It is something alike to budget functions in a group and is a modern machine learning method for decision making.

2.6. Deep Learning

Deep learning is an innovative phase of machine learning which grows around neural networks for predicting and learning data. Using this method, complex generalized systems can be realised which are able to accept any type of issue and give forecasts regarding it.

2.7. Reinforcement Learning

The training data are provided only as a reply to the program's actions in a self-motivated state. It has a nonstop learning process from the environment in an iterative manner.

After debating about several machine learning methods, it would be better to list down a few instances on the applications of ML in the field of biomedicine so that this review will be motivating. In neuroscience, machine learning classifiers are used to study structural and functional dynamics of the brain. Machine learning approaches are used in cancer forecast and prognosis. ANN has been used in categorizing diverse subtypes of psychogenic no epileptic seizures. SVM classifiers are used to spot prostate cancer. Hierarchical clustering has been used in study of Alzheimer's disease. With the knowledge collected on various machine learning methods and machine learning procedures which are mostly linked with computational biology and biomedicine.

3. Machine Learning in Disease Detection and Prediction

Different approaches have been applied to sense or predict an ailment at its preliminary stages so that the action for it would be less complex. It also would rise the probability of the patient being cured. As a result of these methods, different types of diseases have been noticed but with varied accuracy levels depending on features such as the feature set, used algorithm, training dataset, etc. In this segment, a few nominated ailments will be conversed as instances, along with the importance of diagnosing a disease at the earliest. These ML approaches implemented to detect the disease, and the features that were measured to make forecasts. A descriptive contrast of the machine learning methods which have been applied will be conducted in the discussion section of the paper, which will be followed by proposals to further improve them.

3.1. Cancer

The human body has the right amount of cells of each type. Cancer begins with unexpected changes in the cell structure. Signals which are being made by cells determine the division and control of cells. When these signals become defective, cells multiply too much that forms a lump called tumour. Today, thermography is more dependable as it is nonionizing and non-invasive. With the emergence of technology, it has been producing positive and efficient results which have made it greater over other technologies. From these thermographic images, using feature extraction methods and machine learning practices, the occurrence of cancer cells can be detected. Speeded up robust feature (SURF) and Scale invariant feature transform (SIFT) techniques can be used to extract features from imageries. Using PCA or principal component analysis, the features could be further filtered to make improved interpretations.

3.1.1. Breast Cancer

Breast cancer is one type of cancer that is typically seen in women and is an important cause for women's death. However, this can be reduced by initial detection of cancerous cells using mammogram, magnetic resonance imaging (MRI), biopsy and ultrasound. Breast cancer is analysed by classifying the tumour. Tumours can either be malignant or benign. Malignant tumours are more harmful than the benign tumours, and it is not an easy job for physicians to differentiate among these tumours. This makes ML algorithms significant as they can automatically study and improve from the skills without being explicitly programmed.

During the previous years, many machine learning methods were established for breast cancer finding and classification. Their process could be analysed in three phases: pre-processing, feature extraction, and classification. The feature extraction phase is vital as it helps in refining between malignant and benign tumours. Then, the image properties such as coarseness, smoothness, regularity, and depth are mined using segmentation.

3.2. Diabetes

This is a chronic disease, and it needs to be recognized at the early stages for correct medication. Diabetes is caused when the sugar proportion in blood surges. This makes life complicated for the patients due to several reasons. Diabetes can be grouped under three types, such as, diabetes 1, diabetes 2, and gestation diabetes.

DA or Discriminant analysis is a process in which the class tag of an input is determined by a sequence of equations that are attained by input features. Usually, DA uses two likely objectives which are finding a connected equation for categorising test samples and explanation of the predictive equation to better comprehend the relationship among features. When a patient is pregnant, the weight, glucose concentration, blood pressure, diabetes pedigree function (DPF), the ratio of insulin in blood, patient age and skinfold thickness are few features that can be considered for the sorting.

3.3. Heart Diseases

Heart diseases are severe events that are triggered by blockage in the heart arteries. Chronic heart disease is the increase of plaque inside the coronary arteries. These developments slowly sometimes lead to a heart attack. Indications of a heart disease may also include weakness of physical body, shortness of breath, fatigue, and swollen feet with related signs, and so on.

In cardiology, the tasks that precision medicine has achieved include diagnostics and therapeutics in numerous subfields. Personalized treatment options in amending heart rhythms, interventional cardiology, some gender differences affecting the result of cardiovascular diseases, and many works done in genomics can be emphasised as areas in which tasks have been achieved by precision medicine in the field of cardiology.

3.5. Parkinson's Disease (PD)

Parkinson's disease is a progressive and chronic movement ailment. It has no permanent cure, no causes and limited treatment choices. It is found that PD happens due to reduced creation of dopamine, which is a chemical that controls coordination and movement. Rigidity, tremors, postural instability, and slowness of movement are some of indicators of PD. Abnormal writhing movement is a significant symptom of this ailment. Some scientists have applied machine learning procedures on computer vision and video recordings to distinguish healthy controls from the PD patients. Few scholars have also used voice samples to distinguish healthy controls from PD patients. This disease belongs to the neurodegenerative disease group that may indirectly or directly affect the brain cells which will then result in speech, affecting movement, and other cognitive areas.

Analysis

Machine Learning Algorithm	Application Area/Disease prediction	Accuracy
Density-Based Clustering	Lung Cancer	87.6%
Support Vector Machine	Heart Disease	64.4%
K-Nearest Neighbour	Heart Disease	87%
Decision Trees	Heart Disease	85%
Random Forest Algorithm	Heart Disease	90.16%
Logistic Regression	Liver Disease	75%
Naive Bayes	Cancer	98.2%
Artificial Neural Network	Diabetes Disease	87.3%
Hierarchical Clustering	Dengue	76.19%
Partition Clustering	Heart Disease	93%
Graph-Based Clustering	Covid-19 Transmission	74%
Evolutionary Learning	Heart Disease	60%

Table1: Comparison of various Algorithms

Discussion

In this section, the key concern would be given to highlight crucial and important facts of the topics debated throughout the paper. To start with, it is better to give concern towards the performance of ML procedures. In order to identify the best performing algorithms, the classifier log loss and classifier accuracy are the two factors that can be measured. The classifier accuracy needs to be high while the classifier log loss needs to be low for a process to be recognised as a well-performing process. Thus, while selecting an appropriate algorithm to address a certain concern, the above-mentioned factors are measured to select the process out of many diverse existing algorithms that would suit our purpose.

It is not likely to highlight which algorithm is better than the other. One of the main reason is that it depends on the area that the training and the tests are being performed, the dataset that is involved in the training and also the testing procedures, the level of pre-processing that needs to be made on the dataset, the selected feature set for the procedure or the feature selection algorithms that is being used on the dataset, the magnitude of the dataset and the data types in the dataset, the performance level and the size of the machine, etc. This makes it important to select the proper process that would be perfect for the condition.

A lot of studies have been performed using DT, LR, and ANN to identify kidney diseases, and out of these ANN has outperformed both LR and DT by a giant margin. According to former studies, the accuracy of the results gained with regards to the lung cancer analysis is in the order of SVM, ANN, and DT. In addition, various other methods also have been used and they have also exhibited a reasonable level of precision. These are GA, Hopfield neural network, BPNN and LDA. Another benefit by using SVM classification is that it is possible to compute even the phase of the lung cancer. Regression can be used to prototype and predict the association between the dependent and the independent variables. Regression can be characterised as logistic regression and linear regression. In linear regression, the dependent variable should be continuous,

Currently, medical image cataloguing is generally based on pattern recognition approaches. It has been found that classification-based neural networks have superior performance than other supervised machine learning procedures. DNN can capture valuable information while discarding the interfering noise after learning from training models. Currently, DNN mainly includes CNN, auto encoder (AE) and deep belief network (DBN). CNN performs convolution processes on both vertical and horizontal directions. This is a good method for image data as it is applicable in both directions. But for biomedical time series with numerous channels, this is not appropriate as only horizontal direction is applicable and the vertical direction is autonomous. For this cause, multichannel CNN can also be applied to obtain better classification performance.

Reference

[1] A. Bharat, N. Pooja, and R. A. Reddy, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” in *Proceedings of the 3rd International Conference on Circuits, Control, Communication and Computing*, Bangalore, India, July 2018.

[2] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, “Machine learning for medical

imaging," *RadioGraphics*, vol. 37, no. 2, pp. 505–515, 2017.

[3] P. Radhika, R. Nair, and G. Veena, "A comparative study of lung cancer detection using machine learning algorithms," in *Proceedings of the IEEE International Conference on Electrical, Computer and Communication Technologies*, Coimbatore, India, November 2019.

[4] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, "Breast cancer detection using K-nearest neighbor machine learning algorithm," in *Proceedings of the 9th International Conference on Developments in eSystems Engineering*, Liverpool, UK, September 2016.

[5] M. R. Ahmed, S. M. Hasan Mahmud, M. A. Hossin, H. Jahan, and S. R. Haider Noori, "A cloud based four-tier architecture for early detection of heart disease with machine learning algorithms," in *Proceedings of the IEEE 4th International Conference on Computer and Communications*, Chengdu, China, April 2018.

[6] J. Latif, C. Xiao, A. Imran, and S. Tu, "Medical imaging using machine learning and deep learning algorithms: a review," in *Proceedings of the 2nd International Conference on Computing, Mathematics and Engineering Technologies*, Sukkur, Pakistan, March 2019.