2021 International Conference on e-Health and Bioengineering (EHB) | 978-1-6654-4000-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/EHB52898.2021.9657673

Machine Learning Techniques for Medical Image Processing

Baidaa Mutasher Rashed[1,2], Nirvana Popescu[1]
[1,2]Computer Science Department, University Politechnica of Bucharest, Bucharest, Romania,
[2]University of Thi-Qar, Iraq

Abstract— **In last decade, machine learning (ML) techniques increased the capability to automatically learn the experience without being explicitly programmed. Different machine learning methods are used for a number of tasks such as image processing, predictive analytics, data mining, and are used for classification, regression, clustering, and dimensionality reduction. ML is widely utilized in medical imaging research field. This paper introduces a survey on ML used in medical image processing and it focuses on two main types (supervised and unsupervised learning) to importance them in medical image processing with explains the foremost important algorithms of machine learning, discussing the most important advantages and drawbacks of applying Machine Learning techniques in medical imaging. In addition, some common algorithms were applied like k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees, Logistic Regression and Random Forest on medical dataset in order to check the efficiency of algorithms.**

Keywords*— *Medical images; Machine learning techniques; Accuracies.

I. Introduction

Machine learning (ML) is a subset of artificial intelligence (AI) that permits computer systems to learn directly from examples, data, and experiment. ML can diagnose and predict accurate and faster the danger of diseases and stop them in time [1]. ML algorithms allow the possibility to be invested deeply in every fields of medical science, from drug discovery to clinical deciding, significantly changing the way medicine is practiced [2]. ML is employed in various industries that need future prediction, identification of patterns and autonomous deciding and it is widely utilized in healthcare, finance, banking, manufacturing and transportation sectors. ML techniques can be categorized as supervised learning, unsupervised and Reinforcement Learning [3]. A consistent set of ML algorithms has been developed and some aspects are mandatory to be analyzed when deciding how to select the right algorithm for a specific problem: the type of algorithm, parametrization, memory size, overfitting tendency, time of learning, time of predicting [4].

II. Medical Images

An image is a collection of measurements in two-dimensional (2-D) or third-dimensional (3-D) space. In medicinal images, these measurements or photo intensities are frequently radiation absorption in X-ray imaging, acoustic stress in ultrasound, or RF (radiofrequency) sign amplitude in MRI (magnetic resonance imaging). If one measurement is considered at every domain inside the image, then the image is named a scalar image. Otherwise, the image is a vector or multi-channel image. Images may be obtained in the continuous space like on X-ray image, or in discrete domain like in MRI. Medical images play an essential part in helping health care providers to access patients for detection and handling of the disease [5]. Their processing represents a group of procedures to get clinically meaningful information from various imaging modalities for diagnosis. There are four imaging modalities: Projectional Imaging (X-rays), computerized tomography (CT), Magnetic Resonance (MR) and Ultrasound Imaging [6].

III. Machine Learning Techniques

ML Techniques are mainly divided into three classes depending on their purpose: supervised learning, unsupervised learning, and reinforcement learning. In this paper, we briefly discuss every kind of learning technique with the scope of establishing the most relevant algorithms for every type; these algorithms can be applied to almost any data problem, as it can be seen in Fig 1.

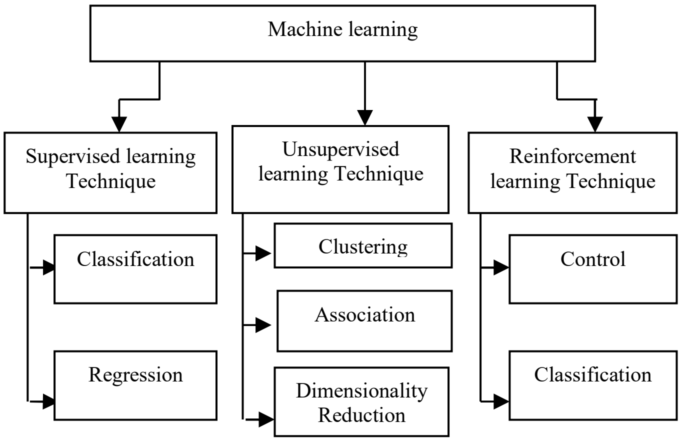

Fig. 1. Machine learning techniques

A. Supervised Learning technique

This category consists of those algorithms that learn a mapping from input data to output from a group of training examples. It contains classification and regression algorithms [7]. The most used algorithms for supervised learning technique are :

B. 1. K- Nearest Neighbors Algorithm

The KNN classification is performed by using a training data group which comprises both the input and the objective variables therefore by comparing the test data that comprise only the input variables to that reference group; the space of the unknown to the K nearest neighbors defines its category assignment by either averaging the category numbers of the K nearest reference points or by gaining by a plurality vote for them. KNN are often utilized in medical diagnosis of various diseases showing alike symptoms [8].

C. 2. Naive Bayes Algorithm

It is based on Bayes theorem and supports the probabilistic type of Bayes theorem. It is used for solving classification problems and utilized for predict of cancer retrogression or advancement after radiotherapy [9].

D. 3. Decision Trees Algorithm

It belongs to the classical algorithm content. Its employed rule is that when processing data information, it begins from the basis node of the gathering instance and stretches the location where the nodes gather to form it complete. It used to solve predictions and regression problems. It is often utilized in applications such as tumors diagnosis [10].

E. 4. Linear Regression Algorithm

It is the simplest form of regression where the attempt is made to fit a straight line to the dataset and it is possible when the relationship between the variables of dataset is linear. In linear regression we add the inputs multiplied by some constants to obtain the output. It is used in predictions such as Predicting Data from Medical Images [11].

F. 5. Logistic Regression Algorithm

It is used to solve binary classification problem and to predict the probability of an outcome having only two values. It works properly when the dataset can be split linearly. It is used for several classification problems such as diabetes prediction, cancers diagnosis [12]. Logistic Regression deals with prediction of target variable which is categorical whereas Linear Regression deals with prediction of values of continuous variable [11].

G. 6. Support Vector Machines (SVM) Algorithm

It is the most current used supervised ML algorithm. It implies effective classifiers which are utilized for classifying the binary dataset for two types with the assist of hyperplanes. SVMs revolve round the concept of a margin for each part of a hyperplane which divides two data kind. It can be used for solving classification and regression problems; in medicine, it has been successfully used in data classification for cancer diagnosis [13].

H. 7. Random Forest Algorithm

It is one of the most important utilized algorithms, due to its clarity and diversity. It is commonly utilized for solving regression and classification problems. Unlike a normal tree, random forest splits each node using the best among a subset of predictors randomly chosen at that node [14].

I. 8. Gradient Boosting Algorithm

It is used when we deal with a great number of data to create a prediction model with high prediction power from weakly prediction models. It combines various weakly predictors to a construct robust predictor. It is used to solve regression and classification problems [15].

J. 9. Artificial Neural Networks Algorithm

ANNs are nonlinear information processing structures, constructed from interconnected basic processing elements known as neurons, stimulated in the same way as the biological nervous systems. It was used for recognizing the faces of persons and if there is a big likeness between the image in the database and the input image, a good correspond is provided. The applications of ANNs comprise recognition of patterns, predicating, clinical diagnosis and still wider [16].

K. Unsupervised Learning technique

It processes the data without labels and is used to detect patterns. It contains clustering and association rule learning algorithms [17]. The most common algorithms for supervised learning technique are:

L. 1. k-means Clustering Algorithm

It is utilized for solving the problem of clustering and classification and for medical image segmentation. It is well known as K-Means as it generates 'K' discrete clusters. In this technique, the objects which possess identical characters are positioned in same cluster. The purpose of this algorithm is to division the n observations in the data into k clusters in which every observation belongs to the cluster with the closest mean [18].

M. 2. Hierarchical Clustering Algorithm

It is an algorithm of cluster analysis that seeks to construct a hierarchy of clusters; it may usually be beneficial to build a hierarchy of concepts through construct a group of nested clusters which may be organized to compose a tree structure. It is used to aid in diagnose based on image data visualization for specific diseases and for other medical image processing modalities [19].

3. A priori Algorithm

It is an algorithm used in sorting information where the sorting information is useful with any data management process. It proceeds simply through detecting the repeated individual items in the database and then extending them to large object sets. This algorithm is used to discover the attributes that show up at one time and in medical image classifications [20].

4. Principal Component Analysis

PCA is used for decreasing the dimensions of the data to make the computations quicker and less complicated [21]. It is also used for features extraction. The key concept in PCA is to detect a subset of variables from a large group, according to which the main variables have the perfect correlations with what is known as principal components. It can be used for many applications, like compression the images and analysis the medical ones [22].

N. *Reinforcement learning technique*

This technique implies a certain type of learning that makes decisions according to which actions to take such that the result is more positive. The learner has no knowledge about what actions to take until a certain situation is given. The action which is taken by the learner may affect the situations and their actions in the future. Reinforcement learning solely relies on two criteria: trial and error search and delayed result. On the whole, reinforcement learning offers a learning method which expands data collection according to statistics and dynamic learning. *Its representative learning methods include Q learning algo*rithm and temporal difference learning algorithm [23].

TABLE I. COMPARISON OF MACHINE LEARNING STYLES

Machine Learning Algorithm	**Goal**	**Advantages**	**Disadvantages**
1.K-Nearest Neighbors	Regression and classification	•It is an easy method to implement. •It is flexible classification •Building the model is inexpensive.	•With the increase in training set size the algorithm becomes intensive in computationally. •It handles massive statistics units and subsequently steeply-priced calculation.
2.Naive Bayes	Classification	• It is a convenient to put in force and rapid method to predict the category of the dataset. • It is a rapid, noticeably scalable and offers precise performance.	•If the specific variable belongs to a class that wasn't accompanied up in the training set, then the model will supply it a likelihood of zero that will prevent it from any prediction. • It is intensive in computationally particularly for models that including numerous variables.
3.Decision Trees	Regression and Classification tree analysis	•It has the ability to fill lacking values in attributes with the highest possibly value. • It has high performance.	• It may be unstable. • The control of the size of tree is very difficult.
4.Linear Regression	Regression	• It is perfect for learning on the data evaluation process. • It is simple to understand and simple to avert over fitting by regularization.	•It is inappropriate if we need to use with non-linear relationships. •Dealing with complex models is complicated.
5.Logistic Regression	Classification	• Simplicity of implementation. • It is not affected with the resource of small noise within side the facts and multicollinearity.	• Not suitable for foresee the value of a binary variable, Boolean values only. • Incapability to solve non-linear problem.
6.Suppport Vector Machines	Classification	•It can deal with structured and semi structured data. •It may be expanded by high dimensional data.	• Its performance decreases with big data due to the increased time of training •When the set of data is noisy it does not work perfectly.
7.Random Forest	Regression and classification	•It is flexible algorithm. •It automates lacking values current in the data. • No need to normalize data as it uses a rule-based approach.	•It requires a lot of resources and computational power as it builds many trees to integrate their outputs. •It requires much time for training.
8.Gradient Boosting	Regression and classification	• Often affords predictive accuracy. • Lots of flexibility. • No data pre-processing required.	•Time and computation expensive. •the implement in real time platform is very difficult.
9.Artificial Neural Networks	Classification or predictions	•It is capable of taking sample of data as a substitute than the complete dataset to give the output result. •Increase current data analysis techniques owing to their developed predictive capabilities.	•The suitable network structure is done via experience and error so There is no particular rule for defining the structure of ANN. •It works with numerical data so the problems need to be translated to numerical values before being inserted to ANN.
10.K-means Clustering	Clustering	• When variables are huge, it is computationally more efficient. • It is easy in implementation and interpretation of the results.	• Estimation of K value is difficult. • The user requires defining the number of clusters in advance.
11.Hierarchical Clustering	Cluster analysis	• No apriori details on the number of clusters wanted. • Easy in implement and outcomes.	• Algorithm can never undo what was completed previously. • It is not appropriate for huge data sets as it is computationally expensive.
12.Apriori Algorithm	Regression	• It is easy to execute and simple. • It is used to mine all repeated item sets in database.	•It suffers from some weak points in spite of being clear and simple. • It is inefficient when memory capacity is limited with big number of transactions.
13.Principal Component Analysis	Dimension reduction of data sets	•Enhances the Performance of Algorithm. •Decreases Overfitting.	•The standardization of data should be done before applying PCA •Information loss.

IV. Discussion and Results

In this work, in order to prove the accuracy of some discussed algorithms, an open access brain tumors dataset [24] has been used. This dataset consists of 150 magnetic resonance images used for training (84 images belong to the *benign* class and 66 images belong to the *malignant* class) and 50 resonance images used for testing (16 images are classified as "benign" and 34 images are classified as "malignant"). Five methods have been tested from the previously discussed methods selected from the most common techniques. The results are presented in Table II.

TABLE II. Comparison For Classification Accuracy For Five ML Algorithms

Method	**Accuracy%**
K-Nearest Neighbor	100%
Decision Trees	99.429%
Logistic Regression	94.246%
Support Vector Machines (SVM)	97.287%
Random Forest	90.128%

Table II indicates that K-Nearest Neighbors algorithm gave the best outcome in contrast with the others, followed by the Decision Trees algorithm which got the accuracy of 99.429%. SVM algorithm proved an accuracy of 97.287% followed by Logistic Regression algorithm (94.246%). Random Forest algorithm had an accuracy of 90.128%.

V. Conclusions

Machine learning techniques show up an increasing effective in image-based diagnosis and illness prognosis.Currently machine learning algorithms are used in most of the cases for classification problems. The paper offered an overview of a consistent number of Machine Learning algorithms utilized for medical image, defining them briefly and offering a critical comparison of their advantages and disadvantages, analyzing their way in solving clustering, regression and classification issues. Thus, ML techniques have been discussed in the context of three categories: supervised learning, unsupervised learning, reinforcement learning. We discussed the cases when we have lower amount of data and plainly labelled data for training, and we should select Supervised Learning. Unsupervised Learning would commonly provide better performance and outcomes for big data sets. If you have a vast data set without any available problems, reinforcement learning algorithms would be more appropriate.

A comparison of their efficiency has been also introduced based on five algorithms selected from the most common ones. These methods have brought very good results.

References

[1] S. Patil, S. Bhosale, "Machine Learning Applications in Medical Image Analysis", JETIR (Journal of Emerging Technologies and Innovative, February 2019

[2] J. L. Wang , J. Rao,Tchoyoson Lim, "Deep Learning Applications in Medical Image Analysis", Digital Object Identifier 10.1109/ACCESS.2017.2788044, March 13, 2018

[3] J. Goo Lee, S. Jun ,Y.Won Cho, H. Lee, G. Bae Kim, J. Beom, , N.Kim "Deep Learning in Medical Imaging: General Overview", KJR (Korean Journal Radiol) 18(4), Jul/Aug 2017.

[4] V.Jatana, "Machine Learning Algorithms", DOI: 10.13140/RG.2.2.20559.92329, June 2019.

[5] D. Patil, Ms. Sonal G. Deore, "Medical Image Segmentation: A Review" , IJCSMC (International Journal of Computer Science and Mobile Computing), January 2013.

[6] E. Miranda, M. Aryuni, E. Irwansyah,"A Survey of Medical Image ClassificationTechniques", ICIMTech (International Conference on Information Management and Technology), 2016.

[7] M. Bruijne, "Machine learning approaches in medical image analysis: From detection to diagnosis", Medical Image Analysis 33 94–97, 2016.

[8] R. J. Ramteke, K. Monali Y. "Automatic Medical Image Classification and Abnormality Detection Using K-Nearest Neighbour" , International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-6 December 2012.

[9] S.B.Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques ", Informatica, 31, 2007.

[10] R. Vijaya, K. Reddy, U. Ravi Babu, " A Review on Classification Techniques in Machine Learning", IJARSE (International Journal of advance research in science and engineering) ,March 2018.

[11] S. Ray, "A Quick Review of Machine Learning Algorithms", International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019.

[12] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions ", SN Computer Science 2:160, 2021.

[13] F.Y. Osisanwo, J.E.T. Akinsola, O. Awodele, J.O. Hinmikaiye, O. Olakanmi, J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison '', International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3 June 2017.

[14] H. Kumar Gianey, R.Choudhary, "Comprehensive Review on Supervised Machine Learning Algorithms", DOI: 10.1109/MLDS.2017.11, December 2017.

[15] A. Abdi, "Three types of Machine Learning Algorithms",DOI: 10.13140/RG.2.2.26209.10088, November 2016.

[16] S.N. Deepa, B. Aruna Devi, "A survey on artificial intelligence approaches for medical image classification", Indian Journal of Science and Technology, Vol. 4 No. 11 ISSN: 0974- 6846, Nov 2011.

[17] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A. Arindra Adiyoso Setio, F. Ciompi,M. Ghafoorian, J.A.W.M. van d. Laak, B.Ginneken, Clara I.S´anchez, "A Survey on Deep Learning in Medical Image Analysis", arXiv:1702.05747v2 [cs.CV] 4 Jun 2017.

[18] Meenakshi, "Machine Learning Algorithms and their Real-life Applications: A Survey ", International Conference on Innovative Computing and Communication (ICICC), 2020.

[19] X.Ma, S.Dhavala, "Hierarchical Clustering with Prior Knowledge", arXiv:1806.03432v3 [stat.ML] 25 Aug 2018.

[20] E. Heni , L. Kurniawati, T.Haryanti 3, N. Mutiah, A. Kurniawan, and B. Said Renhoran, "Data Mining Technique to Determine the Pattern of Fruits Sales & Supplies Using Apriori Algorithm", ICAISD (International Conference on Advance Information Scientific Development), doi:10.1088/1742-6596/1641/1/012070 (2020).

[21] A.Dey, "Machine Learning Algorithms: A Review", IJCSIT (International Journal of Computer Science and Information Technologies), Vol. 7 (3) , 1174-1179 , 2016.

[22] A. Taloba , D. A. Eisa, S. Ismail, "A Comparative Study on using Principle Component Analysis with Different Text Classifiers", arXiv:1807.03283v1 [cs.IR] 4 Jul 2018.

[23] Wei Jin, "Research on Machine Learning and Its Algorithms and Development", JCSP, J. Phys.: Conf. Ser. 1544 012003, 2020.

[24] OASIS dataset, http://www.oasis-brains.org/.