Machine Learning With Health Care: A perspective

*Meenakshi Malik, Rekha Khatana , Aditi Kaushik

1,2&3Assistant professor, School of Computer Science and Engineering, Starex University, Gurugram, India,

Corresponding Author Email: meenakshimalik16@gmail.com

Abstract: ML "machine learning" is an ever-expanding research field with plenty of possibilities for study and implementation. Mr. James Collin stated at MIT that ML is the technology defining this decade, even though it has had a meagre effect on healthcare. Several fresh businesses in the ML industry are applying themselves earnestly on healthcare. Even google has jumped in to the race and it has designed a ML application for identification of cancer tumour on mammograms. To identify skin cancer, Stanford uses a Deep Learning algorithm. Around one trillion GB of data is getting generated per year by the USA health system. Various academic experts and scientists have worked out various characteristics and several factors of risk involved in chronic illness. Additional data stands for more learning for machine, but for higher precision, these many features need a huge quantity samples. So if machines can harvest clinically greater risk oriented feature it would definitely be better. Precision is improved when the data in the form of exploratory data analysis and feature engineering is pre-processed. The multi-class classification might be capable of evaluating a patient's different disease risk levels. In health care, correct identification of the percentage of diseased individuals "sensitivity" is a primary concern rather than correct identification of the percentage of healthy individuals "specificity". Our paper introduces one of ML's super challenger and emergent application i.e. Healthcare. The careful and sympathetic relation with care providers will always be necessary for the patients. ML cannot remove this, but will become instrumental for health professionals in improving and strengthening on-going care. Our paper discusses the several established models of ML along with its applications in healthcare system. We also bring up directions for imparting more efficiency to ML model. In addition to this we have also discussed ML’s case study for Breast cancer.

Key words: medical diagnosis, healthcare, Machine Learning, ML techniques

1. Introduction

AI "Artificial intelligence" can be simply described as a science and engineering stream that manages to integrate computer science and robust dataset to solve problems. It requires insights of computation by creating manmade tools and devices that perform demonstrate required type of behaviour [1]. ML and AI are integral part of computer science which are having close relation and connection with each other. Both of these are the trendiest technologies which are being utilised in the creation of smart systems, devices and gadgets. In clinical applications and diagnosis field also ML tools and techniques are being used over several years. The core aim was to enhance capability of healthcare field by improving the capabilities, performance and consistency of medical professionals. Enhanced precision and presence of huge ML data and techniques quickly ramped choice and shortlisting of ML for medical diagnostic troubleshooting. In further execution for medical usage, ML software contributed in great manner. This is why we are now in the fourth decade of AI having wide range of ML applications covering areas from medication to clinical diagnosis and surgery. The key benefits of ML are precision, cost-effectiveness and simulacrum based on constant and consistent
learning. For instance; Various datasets from multiple online sources (UCI ML repository, Kaggle or data.gov.in) for the same ailment classification having distinct numbers and name for the functions esp. datasets of heart related diseases [2–5]. Our paper presents and attempts for providing a better understanding of the ML techniques in healthcare and clinical diagnostic applications continually evolving in recent years.

2. ML Techniques

ML is one of the main branch and most expanding sub-field of AI. Software of AI tracks a range of human brain energy features, including logic, pattern to learn, plan, understanding, sense, and the ability to transfer and use data and objects [6]. AI utilizes several methods in these areas to imitate human intelligence. ML methodologies collect scientific findings as inputs and predict data structures. Statistical evidence and scientific findings include sexuality, age, medical exam, imagery, expression of genes etc. Majorly following four ML techniques may be utilized to integrate huge data sets.

The same have been shown below as Fig. 1: (a) “Supervised Learning” (b) “Unsupervised Learning” (c) Semi Supervised Learning” (d) “Reinforcement Learning”

![Types of ML techniques](image)

**Fig. 1** Types of ML techniques

2.1. “Supervised machine learning”:

The very first methodology is a supervised ML and it is related the latter known as unsupervised ML whereas the last one is semi supervised learning which enhances learning. Known classified dataset are always there in supervised ML. It is utilized mainly to predict the relationship between interest trait and result. We act as teachers interactively by training the data on the computer and feeding it. The data which is trained, contains the input means the predictor and the output of product, the results of which are used for testing the patterns. It builds relationships and dependencies with input and specific values of output. Classification and regression problems are the main category under supervised machine learning problems. Decision trees, the Naïve Bayes, Nearest Neighbour, Linear Regression, SVM “Support Vector Machines” [7] and neural networks are few of the most common algorithms of this type. Out of these SVM [8] may be termed as one of the simplest understandable ML algorithms. Supervised algorithm being used in clinical diagnosis more frequently and it produces more accurate results. SVM is devised for binary (two class) problems and may be used to classify data sets in multi-classes. Neural networks [7] are also one of the alternative ML methods which works best if the line between the two classes is not straight. In medical diagnosis, the SVM is widely employed. A neural network denotes a trinity of layers: the main is the input layer, covered with a second layers made of plenty of layers, lastly showing an output layer (Fig. 2).
The inputs of each node generate an output which is loaded to the next network node. The trait values are entered into the network as initial input. The result defines the traits classification. Ramesh et al [10] revealed usage of neural networks as the frequent diagnostics tools. Another study also showed that the most commonly used AI technique are "fuzzy logic-neural networks". It may be used rigorously in field of genetic, radiology, cardiology etc. The advanced version of the classical neural network technique is "Deep learning". In the past few years, revival of the same has emerged dramatically increased power of computation and availability of novel data sets.

Deeper learning can analyse further challenging data sets particularly non-linear ones. Due to the growing data size and its ramifications, deep learning has become very prominent in recent times [12]. It uses more hidden layers and has more complex data with different structures in comparison to classical neural network [13]. In clinical application, some of the prevalent and important networks are "network of neural recurrent", supported by the "network of neural deepness", "network of neural convolution (CNN)" and "network of deep belief" that has become the most common method of developing deep learning algorithms. Lately, with popular software such as Microdot's CNTK, Google's Tensor Flow, CNN has been efficient in diagnosing diseases.

2.2. “Unsupervised machine learning”:

The chief categorisation of unsupervised learning technique consists of "Association Rule Learning Algorithms" and "Clustering Algorithms". In order to identify most significant data classification characteristics, popularly utilised unsupervised ML techniques are "Association Rules", "K-means clustering" and "Principal Component Analysis (PCA)" [14, 15]. Different clusters or groupings can be utilized in "K-means" to find the one where the points in each group are closest to one another. Whereas in the PCA procedure, the data is arranged into a centred matrix. Independent rows and columns portray the data model and the feature value respectively. We subtract the mean of the column from the values in each column in centred matrix. It is done to make mean of each column as zero.
2.3. “Semi-supervised machine learning”

This is a blend or fusion learning amongst supervised and unsupervised that is suited to data and situations in which the result is not complete for certain topics[16]. We have labels for all observations in datasets under supervised learning and we have no labels for them under unsupervised learning. Semi-supervised learning stands amidst both these learnings. The labelling costs are now high as they involve the knowledge of human experts. Some models have thus been developed, with presence of labels in few observations and unlabelled data in plenty of observations. These types of semi-supervised models are the perfect factors for building of model. Despite unlabelled data, the pattern is explored through hold of data info of group parameters.

2.4. ”Reinforcement learning”

It is like a human child who learns to walk and speak. This technology has an agent that learns primarily from environment. Learning is carried out through the simulation of positive and negative rewards. If the actions are carried out by the agent as expected, they would be considered as positive reward else the results will be called negative reward. An example of reality is the formation of a robot to grab and stack the blocks at a site. The environment is therefore the blocks, grid and buckets. Robot's hand acts as agent. The robot is now programmed to collect the highest block, move it into a bucket and stack it. If robot slips the block in the middle of journey after completion of first simulation, it is concluded that robot does not do the job correctly. It sends out a false signal or alarm. Repeated picking and dropping actions are done until the robot hand does not learn to take action without a negative reward. Temporal difference, Q-learning, deep adversarial networks are the most famous reinforcement learning algorithms.

3. ML’s case study in Breast Cancer

3.1. Generalized Work flow

ML support in Diagnosis for Breast Cancer is the most important and revolutionary step which is proving a support system in reducing the expiry rate of woman due to breast cancer, nowadays. A reliable futurist prototype for preliminary diagnosis can reduce women's weakening explicitly, the previous neutral estimate indicates whether the patients are owed to a cancerous/ malignant or non cancerous i.e. benevolent. In the area of malignant exploration, it is typically significant to retain trained data in an appropriate set of data which subsequently scrub up and revision of irrelevant or unacceptably unacceptable information aids to demonstrate additional cancer analysis interpreters.

Figure 3 depicts a ML system's general process flow. It begins with the user's raw data collection. Data features need to be extracted which can be added to the ML algorithm in order to make the data understandable. Thus, the data are purified and pre-processed in the next step. Irrelevant features are deleted in feature engineering and new features are created that will help train the ML model. This dataset with feature selection is split into a test and train datasets, where test dataset is reserved for the purpose of evaluation and prediction. For validation purposes, the train data set is further divided. ML model is trained on train datasets and validation is done with validation datasets. Once the model is finalized, the user can make prediction for the test datasets. Improvisation on model is done with further data or better ML algorithms by customer feedback and predictions.
Fig 3 Generalized work flow of Machine learning

3.1.1. “Data Acquisition” or precise and early diagnosis of any illness and disorder, biomarkers are essential. Health technology progress has made it less complicated to obtain biomarkers today. The biomarker collected data becomes tools for driving ML healthcare applications. The data sets can be selected on the basis of factors like

- Data size and type
- Data acquisition cost and ease
- Continuation in data updation
- Amount of clinical attention needed
- Patient’s physical visit to clinics

3.1.2. “Data Pre-processing” Data collected from whatever channel, such as sensor, must be in machine-understandable form. The first step in the removal of unnecessary data and outliers is data cleaning. Significant data for specific problems must be identified. For instance, it is possible to remove fraudulent or missed data field. Consequently, the data transformation, i.e. selection, scaling and normalisation of the data to be done. Seldom is the size of data reduced to save computing effort for huge datasets. (Fig. 3).

3.1.3. “Feature Engineering” Functional engineering can make or break an ML model the most important and crucial step. This is where knowledge of the domain can help as shown in figure 4. Feature engineering allows ML algorithms to interpret data. The processes of feature engineering include the creation of new features, removal of unused functionalities, correlation of features, categorical one-hot codification, detection of outliers, clustering operations, function transformation. The selection of features greatly influences model performance. The choice is done utilizing various combination of features, feature importance and model evaluations. In order to store a data subset for a final analysis of the model, feature engineered data is split into test and train dataset before training in the ML model. Training datasets are furthermore divided with Validation set for CV "cross validation" purpose. As selection of an apt ML model is not completely simple, general practise would be to train datasets through multiple ML algorithms.
There are a lot of impressive open source ML libraries available with different algorithms. Linear and Logistic Regression, SVM "Support Vector Machine", KNN "K-Nearest Neighbours", Tree Regressors, Naïve Bayes, Neural Networks and Boosted Tree Recorder are among the most popular libraries and algorithms. The next step is to validation and comparison of few chosen algorithms with techniques like K-fold CV method. In case of insufficient accuracy or overfitting, every or all process is repeated and models are retrained for more data acquisition, data pre-processing and feature engineering.

3.1.4. “Evaluation and Prediction”

Only one metric like accuracy cannot support in evaluation of complete model efficiency effectively in a classification problem. Hence for every kind of skin lesion disease, the accuracy, precision, recall, F1 score and support are measured. The confusion matrix is also plotted to verify the performance of our model in each class.

4. ML’s Applications in Medical Diagnosis

The increasing demands for precise and perfect knowledge in healthcare is forcing us to predict information, carry new tests and put novelty efforts hand in hand for the assistance of countless patients suffering from 'n' number of diseases. ML has made it possible which seemed tough or almost impossible earlier. As a specific area of AI, ML is the reflection of human being intelligence by computer system. It manages to combine both statistics and computer science; where computer science addresses identification of problems and their solutions and statistics takes care of hypothesis, data modelling and reliability measurement [17]. Machine Learning is a framework that learns from previous experiences for future performance improvements. Also development of programmes is majorly focused in which data is trained to learn by own selves. The following are few healthcare ML applications:

- Improved Radiotherapy
- Drug Discovery of drug and its Manufacturing
- Maintaining Smart Health Record
- Outbreak Predictions
• Identification of ailments
• Medical Imaging Diagnosis
• Medicine Personalization
• Clinical Trials and Researches
• Behavioural Modification based on ML
• Data Collections through crowdsourcing

The health domain is concerned with large data, like sensitive and detailed patient information, making it difficult for individuals to analyse and maintain these massive records. In order to overcome these issues, ML has shown the way forward by automatic discovery of patterns as well as providing predictions on data. This allows medical analysts to provide accurate medicine through personal care. We can handle large quantities of medical data efficiently by means of ML systems [18]. It provides numerous data classification and prediction algorithms, which is the main goal of the analysis of medical data. Hence it supports us to decide more accurately and more precisely.

5. Research Directions
All of the above-mentioned challenges offer several research possibilities to improve the accuracy, precision, and adaptability of computation models. Thus, some research directions are mentioned below by taking all of them into account:

• Integration: Various parameters from different aspects and sources should be integrated for better results in detection, analysis and measurement of biomedical information. Data from various sources could describe various aspects of the problem. Comprehensive insights into a particular disease can be obtained by means of integrative analysis.

• Security: The safety and privacy of patient data is very essential and crucial. Sun et al.[19] said that very slight changes to lab records can totally alter the prediction of a predictor pattern for the patient's EHR. In consideration of adverse attacks it is very important for researchers to develop defence mechanisms.

• Transparency: Computational models are very important for some areas of the medical field. Most of the models currently presented are black box models. In order to achieve a better precision, accuracy and adaptability, these Black Boxes should be explored more.

• Database creation: The database is not sufficiently available or does not take into consideration all the parameters and aspects, or any particular region or medical institute for many diseases like aids, dengue, mental ailments, flu, heart disease, etc. Hence many databases with various parameters are still required.

6. Conclusion
The roots of machine learning are firmly grown in history. If technology is to improve quality of care in future, the ability of analytics and machine learning must increase the electronic details available to doctors. Medicine machine learning has made headlines recently. The priority should be on how machine learning can be used to increase care for patients. The different models of ML and its applications in the health system are discussed in our paper. We also proposed guidelines to give the ML model greater efficiency. In addition to this we have also discussed the ML's case study for Brest cancer.
References

[1] S.C. Shapiro, Artificial intelligence, in Encyclopedia of Artificial Intelligence, vol. 1, 2nd edn., ed. by S.C. Shapiro (Wiley, New York, 1992)

[2] C.B. Gokulnath, S.P. Shantharajah, An Optimized Feature Selection Based on Genetic Approach and Support Vector Machine for Heart Disease (Springer Nature, Iran, 2018)

[3] E.R.Q. Fernanded, A.C.P.L.F. de Carvalho, X. Yao, Ensemble of classifiers based on multiobjective genetic sampling for imbalanced data. IEEE Trans. Knowl. Data Eng. 14(8) (2015)

[4] F. Babič, J. Olejár, Z. Vantová, J. Paralić, Predictive and descriptive analysis for heart disease diagnosis, in FedCSIS, vol. 11 pp. 155–163, IEEE Catalog Number: CFP1785N-ART c 2017, Slovakia, https://doi.org/10.15439/2017f219. ISSN 2300-5963

[5] R. Pari, M. Sandhya, S. Sankar, A Multitier Stacked Ensemble Algorithm for Improving Classification Accuracy (IEEE, 2018)

[6] A. Agah, Introduction to medical applications of artificial intelligence, in Medical Applications of Artificial Intelligence (CRC Press, 2013), pp. 19–26

[7] Meenakshi, M. (2020, May). Machine learning algorithms and their real-life applications: A survey. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC)

[8] C. Cortes, V. Vapnik, Support-vector networks. Mach. Learn. 20(3), 273–297 (1995)

[9] Meenakshi Malik, Rainu Nandal, Surjeet Dalal, Vivek Jalglan, Dac-Nhuong Le,” Driving Pattern Profiling and Classification Using Deep Learning”, Intelligent Automation & Soft Computing, Vol.28, No.3, 2021, pp.887-906

[10] J.H. John, Neural network and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. U.S.A. 79, 2554–2558 (1982)

[11] A.N. Ramesh, C. Kambhampati, J.R. Monson, P.J. Drew, Artificial intelligence in medicine. Ann. R. Coll. Surg. Engl. 86(5), 334 (2004)

[12] A. Yardimci, A survey on use of soft computing methods in medicine, in Proceedings of the 17th International Conference on Artificial Neural Networks, Porto, Portugal (2007), pp. 69–79

[13] D. Ravi, C. Wong, F. Deligianni, M. Berhelot, J. Andreu-Perez, B. Lo, G.Z. Yang, Deep learning for health informatics. IEEE J. Biomed. Health Inform. 21(1), 4–21 (2016)

[14] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning (MIT Press, Cambridge, 2016)

[15] S. Lloyd, Least squares quantization in PCM. IEEE Trans. Inf. Theory 28(2), 129–137 (1982)

[16] C. Amuthadevi, D.S. Vijayan, Varatharajan Ramachandran, “Development of air quality monitoring (AQM) models using different machine learning approaches”, Journal of Ambient Intelligence and Humanized Computing, https://doi.org/10.1007/s12652-020-02724-2

[17] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurrol. 2(4), 230–243 (2017)
[18] https://www.computerworld.com/article/3479833/machine-learning-essential-to-healthcaresays-orion-health.htm

[19] M. Sun, F. Tang, J. Yi, F. Wang, J. Zhou, Identify susceptible locations in medical records via adversarial attacks on deep predictive models, in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2018), pp. 793–801