Analysis of AI techniques for healthcare data with implementation of a classification model using support vector machine

P G Jaiswal¹, M Gaikwad² and N Gaikwad³

¹Assistant Professor, Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India
²Research Scholar, Department of Computer Engineering and Information Technology, VJTI, Mumbai, India.
³B.E Final Year Student Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India

E-mail: pgi.ycce@gmail.com, 9866@ycce.in

Abstract. Artificial intelligence (AI) is imposed to impersonate human cognitive functions. AI Techniques are most popular across healthcare. The motive behind implementing an AI system is to make the system more fast and efficient. Now, AI can assist medical physician for fast and accurate diagnosis of diseases. When the time of deployment of the AI system will come then, systems need to be ‘trained’ for a huge amount of data will be generated from different clinical performance data. Now a day’s data is available in a structured, unstructured and, semi-structured format. For supporting, retrieving results and knowledge from this data, its analysis using different AI techniques are available. This includes machine learning methods for structured data and unsupervised learning for unstructured data which is useful for retrieving features when the outcome for some subjects is missing. In this paper different conventional machine learning techniques used in healthcare, domains is analyzed using different data types. Also, a comparison of different methods used in Artificial intelligence fiction in the healthcare domain is explored. A flow from clinical data creation, through NLP data enhancement and Machine learning data analysis for making clinical diagnosis decisions and its predictions are discussed and implementation using a support vector machine (SVM) on the healthcare dataset consisting of the patient questionnaire is done.

Keywords: Artificial intelligence, supervised learning, unsupervised learning, support vector machine, healthcare, machine learning, natural language processing, classification.

1. Overview of Medical Artificial Intelligence (AI) Research
Currently, artificial intelligence techniques are most popular across the healthcare domain for disease diagnosis and prediction. The motive behind implementing an AI system is to make the system more fast and efficient. Now, AI can assist medical physician for fast and accurate diagnosis of diseases. When the time of deployment of the AI system will come then systems need to be ‘trained’ as a huge amount of data will be derived from different clinical actions such as screening, images, patient history, and diagnosis treatment, and so on. The following given graph as shown in figure 1.1 represents the type of data considered for medical study using AI techniques, where the different data types are measured in the artificial intelligence literature. In the given graph comparison is given which is obtained through searching the diagnosis techniques in the artificial intelligence literature from the year 2013 to date and it is observed that the data of Diagnostic imaging is highly used for a
research.

![Comparison of different data types used in Artificial intelligence literature in healthcare domain](image)

**Figure 1.1.** Comparison of different data types used in Artificial intelligence literature in healthcare domain

After performing analysis of different data sources, it is concluded that we can differentiate them with image, inherited and electrophysiological (EP) data; the reason for this conclusion is that it contains a bulky section of unstructured description texts such as medical or scientific notes, which cannot be openly or directly analyzed. Because before processing unstructured data needs to convert into a machine-understandable format. For that, there is a need for the implementation of mathematical AI devices.

### 1.1 Overview of AIDevice

It is observed that most of the data which is available in healthcare is unstructured data and which will lead to the conversion of unstructured data into machine understandable format and it is possible with the use of AI device. Mainly AI devices are of the following two types:

#### 1.1.1 Machine learning Techniques

Machine learning techniques will help to an analysis of structured data ex. images, genetic, and EP data. To help in this regard various clustering and classification techniques can be used for predicting and inferring the probability of various disease outcomes based on available data.

Working with Machine learning techniques it is categories into supervised learning and unsupervised learning, additional this when data is insufficient and wants to use it then can go with transfer learning also. Unsupervised learning is widely used for feature extraction and supervised learning is used for predictive modeling. Add on to this recent semi-supervised learning has also been proposed as an amalgam of supervised and unsupervised learning. Unsupervised learning is applicable when the outcome is unknown for certain subjects. Following figure 1.2 represents the graphical illustration of unsupervised, supervised and semi-supervised learning.
1.1.2 Natural Language Processing

Natural language processing methods can help extract data from random data such as clinical notes, medical journal, random patient history, and treatment plan, etc. Natural language processes will aim to translate texts into systematic data, and this structured data will be analyzed using machine learning methods.

2. Proposed System

The flowchart presented in Figure 2.1 represents a roadmap from clinical data construction, for enrichment of NLP data and machine learning data analysis by creating an effective clinical decision.

3. Classical Machine Learning Techniques

At this stage, the review of AI techniques with all classical machine learning techniques is explored with supervised and unsupervised techniques. The old ML technique creates data analysis algorithms to extract features from the data. In this proposed method of incorporation into the machine learning algorithm is a ‘feature’ of the patient and sometimes medical results for use or reproduction. Patient train coverers or include basic information such as age, gender, weight, responsible parameters,
disease history, diagnostic images, EP tests, physical examination results, etc. the previous section. Supervised reading is widely used to exclude the feature and supervised reading is used for predictive modeling. In addition to these several recently monitored readings have been suggested as a combination of supervised and unsupervised learning. Unsupervised learning is effective if the result is unknown in certain subjects. Clustering and Principal component analyses (PCA) are the two most important unsupervised learning methods.

In this study, clustering techniques are very helpful in clustering patient labels by increasing and reducing patient similarities within and between groups. Also, it classifies topics with 'similar' features into collections, without using result details. Many clustering algorithms are found in other popular clustering algorithms including K-means clustering, Hierarchical clustering, and Gaussian clustering.

The Principal component analyses (PCA) is mainly useful for dimension reduction, without trailing too much information. PCA is the basic step before performing clustering or classification on any dataset. On the other hand, supervised learning considered traits with the subject outcome. After that, it will go through a training process to find out the finest output connected with the inputs that are close to the outcome on average. Compared with unsupervised learning with supervised learning, supervised learning provides more clinically significant results. Different available techniques are given below.

a) Linear regression  
b) Logistic regression  
c) Naïve Bayes  
d) Decision tree  
e) Nearest Neighbor  
f) Random forest,  
g) Support vector machine (SVM)  
h) Neural network  

Figure 3.1 shows the recognition of different supervised learning techniques in medical healthcare applications. It is noticed that SVM and Neural networks are the most useful and popular techniques through a literature survey.
4. Implementation using Support vector Machine (SVM)

After analyzing all the algorithms, we have implemented our system using a Support Vector Machine. A support vector machine (SVM) enables a proficient approach of extracting the features and set of rules to perform classification. It is preferred here because of its high-performance accuracy and competency in processing high dimensional data.

To understand the working of SVM here real-time data is collected from various dental organizations and hospitals which is useful for predicting the severity of tooth disease. While collecting the data questionnaire is framed which covered all the responsible parameters required for the finding the cause of disease. It also covers the overall history of patient which covers attribute such as weight, age, eating habits routine, medical history, etc.

![Figure 4.1 Multilevel classification using SVM](image)

4.1 Data pre-processing

Following pre-processing steps are performed for cleaning of the data before applying to the Machine learning model for classification. Various steps involved in pre-processing of the data are as below:

- Missing Value Imputation
- Conversion of categorical variables into numerical variable with the help of dummy variable. (133 rows and 40 columns).
- Standardization of data (Zero Mean and variance of one)
- Test - Train split by taking 80% for training and 20% for testing a model.

![Figure 4.2 Sample Dataset](image)
Figure 4.3 bar Plot for Target Class

Figure 4.4 Support Vector Classifier score for different kernels

Figure 4.5 Dataset Information
In the figure 4.3, 4.4, 4.5, and 4.6 basic steps on dataset are given, where the standardization of data is done. For training the classification model the ratio for test and train split is considered as 80 percent for training and 20 percent for testing. The results after applying SVM are given below in different formats where pre-processing, class balancing, selection of kernel for SVM. In this work, the performance of the classification model is measured with the help of a confusion matrix, which will check the accuracy of classification results.

It is observed from the plot shown in figure 4.4, that the linear kernel performed the best for this dataset and achieved a score of 65.90%.
Figure 4.8. Confusion matrix of SVM.

The confusion matrix is shown in figure 4.5 which is used to measure the accuracy of the model concerning the actual and predicted class of given data set which is trained using SVM. Table 4.1 shows the summary of results using SVM on the healthcare dataset that is based on the questionnaire or history of the patient.

Table 4.1 Results Summary using SVM.

| Classification model      | Accuracy |
|---------------------------|----------|
| Support vector machine    | 73.21%   |

5. Result and Discussion

The main reason for this review of the AI strategy for health care has presented a variety of health information that AI has analyzed and evaluated the major pathogens used by AI. Therefore, details have been discussed about the two main categories of AI devices that include machine learning and natural language processing, with a more detailed discussion of conventional machine learning technique. After that, the analysis of different classification and clustering techniques used in the medical healthcare system is done and it is found that SVM and neural networks are more popular. Then the implementation of SVM for multi-class classification is done using a patient history dataset where the disease is classified into four different classes. And the Accuracy gained by using SVM is 73.21% percent which is promising for further future work.

6. References

[1] Neill DB. Using artificial intelligence to improve hospital inpatientcare. IEEE Intell Syst 2013;28:92–5.
[2] Goodfellow I, Bengio Y, Courville A. Deep Learning. First Edition: The MIT Press,2016
[3] Anand, A. and Shakti, D., 2015. Prediction of diabetes based on personal lifestyle indicators. In Next generation computingtechnologies (NGCT), 2015 1st international conference on (pp.673–676). IEEE
[4] Kanchan, B.D. and Kishor, M.M., 2016. Study of machine learning algorithms for special disease prediction using principal of component analysis. In Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016 International Conference on (pp. 5–10).IEEE.
[5] Hossain, R., Mahmud, S.H., Hossin, M.A., Noori, S.R.H. and Jahan, H., 2018. PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques. Procedia Computer Science, 132, pp.1068–1076.
[6] Patel VL, Shortliffe EH, Stefanelli M, et al. The coming of age of artificial intelligence in medicine. ArtifIntell Med 2009;46:5–17. Jha S, Topol EJ. Adapting to Artificial Intelligence: radiologists and pathologists as information specialists. JAMA2016;316:2353–4.
[7] Pearson T. How to replicate Watson hardware and systems design for your own use in your basement. 2011 https://www.ibm. com/developerworks/ community/ blogs/ InsideSystemStorage/ entry/ ibm_ watson_ how_ to_ build_your_ own_ watson_ jr_ in_ your_ basement? lang=en (accessed 1 Jun2017).