Advanced Performance of Multimodal Disease Risk Prediction in Machine Learning using Healthcare Bigdata

M. Janani[1], A. Maria Stella[2]
[1]Research Scholar, Department of Science, JJ College of Arts & Science, Pudukottai, India
[2]Assistant Professor, Department of Computer Application, JJ College of Arts & Science, Pudukottai, India

***Abstract*: Prediction of future diseases from historical data of medical patients is a topic that has gained increasing interest given the growing availability of such data in computerized format. In recent years, there has been numerous studies constructing disease network with diverse sources of data. However, it is still demanding to achieve a force- emotional pick-and-place task, unless an appreciable number of sensors monitor the operation. The predictive capacities of deficiency in conventional machine learning object-sorting schemes are defined. Furthermore, various regions performance particular attributes of convinced local diseases, which may diminish the indicator of disease blast. In this paper, authorize ML algorithms for effective prediction of stable sickness interruption in disease- large societies. To damage the difficulty of unsatisfactory data, help an essential aspect model to reproduce the lost data. That investigation on a provincial consistent sick of analytical stoppage. Introduce an improved convolutional neural network (CNN)-based advanced multimodal disease risk prediction algorithm using arranged and unarranged from clinics. Correspond with many common prediction algorithms, the indicating performance of our enhanced recommend algorithm reaches 97.4% with a concurrence speed and accuracy. In the end, this expected a techniques to criterion various algorithms and deciding the end model.**

***Keywords*: Big data analytics, machine learning, healthcare**

1. Introduction

Machine learning (ML) is the collection of exponentially growing techniques which are used to find some useful information, patterns and knowledge from already given data. This useful information helps to advance existing research and improve productivity. The applications of machine learning (ML) are uncountable; it is used almost in every aspect of life [2]. Some major applications of machine learning (ML) are Healthcare, Market Analysis, Finance, Education, Manufacture Engineering, Corporation Surveillance and Agriculture. The machine learning (ML) techniques can broadly distinguish into two types predictive & descriptive. The predictive family can further classify into classification, regression and time series analysis. In this paper have limited our research scope to apply data-mining techniques for crops disease and loss prediction which belongs to classification (predictive) family. Machine learning (ML) techniques help in agriculture for the exclusion of manual jobs and for decision making which enable to decrease production cost and improves productivity.

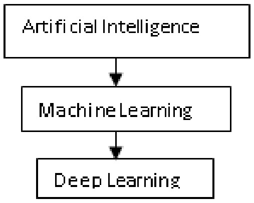

Fig. 1. Structure of machine learning

The increasing volume of healthcare data contained in Electronic Health Records (EHRs) has caused many to consider the possibility of designing automated clinical support and disease detection systems based on patient history and risk factors. A number of past studies have attempted to use patient laboratory tests diagnoses and medications as means of predicting disease onset. Such models have also been used to identify potentially unknown risk factors often while simultaneously improving sensitivity and specificity of detection. A number of recent studies have been successful in predicting disease via various methods, including support vector machines logistic regression random forests neural networks and time series modeling techniques. ML to interrogate latest decision to make prognosis about future.

It is known that many countries gather electronic records of patients' medical checkups from over the years. These records usually include symptoms, examination results and the history of diseases of each patient. In the case of Japan, all active workers have to undergo every year a medical checkup to monitor their health, data that is also electronically recorded. This results in large databases of health information that can potentially be used in many ways to train classification or prediction models.

The problem of disease risk prediction has been studied through different approaches including statistical analysis, if-then rules, data mining and machine learning. In recent years, with the rise of new machine learning techniques and the growing amounts of medical data available, the problem has gained renewed interest from the data science community.

Many popular supervised machine learning (ML) algorithms-such as Naive Bayesian (NB), K-nearest Neighbor (KNN), Decision Tree (DT) algorithm, Support Vector Machine (SVM) using for disease prediction.

A. *Predictive analytics in health care*

Predictive Analytics is supporting different segments of health care life sciences and providers. It aims in diagnosing the diseases accurately, enhancement of patient care, resource optimization and also improves clinical outcomes. Predictive Analytics helps organizations to prepare for the health care by optimizing the cost. The accomplishment of predictive analytics in this industry is likely to provide proficient outcome by improving the service quality. Predictive Analytics have the future to transform the health care industry.

2. Classification techniques

Here describe four different classifiers using different classification techniques. These classification techniques are generally used for classification can be also used for Disease Prediction.

A. *Decision Tree Classifier*

Decision tree classifier is a simple and popularly used

algorithm to classify data. Decision Tree represents a tree like structure with internal nodes representing the test conditions and leaf nodes as the class labels. This classification approach poses carefully crafted questions about the attributes of the test data set. Each time an answer is received another follow up question is asked until it can correctly classify the class of the test data. This classifier handles over-fitting by using post pruning approaches.

B. *Support Vector Machine Classifier*

This algorithm works on a simple strategy of separating hyper planes. Given training data, the algorithm categorizes the test data into an optimal hyper plane. The data points are plotted in a n-dimension vector space (n depends upon the features of the data points). SVM algorithm is used for binary classification and regression tasks but in our case, it has a 3-class sentiment analysis making it multiclass SVM classification. It adopts the pair wise classification technique where each pair of classes will have one SVM classifier trained to separate the classes. The overall accuracy of this classifier will be accuracies of every SVM classification included [2]. Then on performing classification it finds a hyper plane that differentiates the 3 classes very well.

C. *K- Nearest Neighbor Classifier*

KNN Classifier is an instance-based learner used for both classification and regression tasks. This algorithm does not use the training data to make any generalizations. It is based on feature similarity. A test sample is classified based on a majority vote of its neighbors; the class assigned to the test sample is the most common class among k nearest neighbors [3]. This classifier is a lazy learner because nothing is done with the training data until the model tries to classify the test data. It has taken the k value to be 3 which gave us the most accurate result. The k value must not be too large that it includes the noise points or points that belong to the neighboring class.

D. *Gaussian Naïve Bayes Classifier*

Naïve Bayes is a popular text classifier. This classifier is highly scalable. This algorithm makes use if the Bayes Theorem of conditional probability. Since it dealing with continuous values it makes use of the Gaussian distribution. Gaussian NB is easier to work with as it only needs to compute mean and standard deviation from the training data. This classifier passes each tweet and calculates the product of the probabilities of every feature present in the tweet for each class label i.e. positive, negative and neutral. The class label is assigned to the tweet based on the sentiment label that has biggest sentiment product. The equation for normal distribution is described as

$$P(x_i|y) = \frac{1}{\sqrt{2\,\pi\sigma_y^2}} \exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2})$$

3. Proposed work

In Proposed Work, propose a new Improved Convolutional neural network based advanced multimodal disease risk prediction (CNN-EMDRP) algorithm using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data Analytics. The development of big data with analytics technology, more consideration has been paid to disease prediction from the aspect of big data analysis; various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification, rather than the previously selected characteristics. However, those existing work mostly considered structured data.

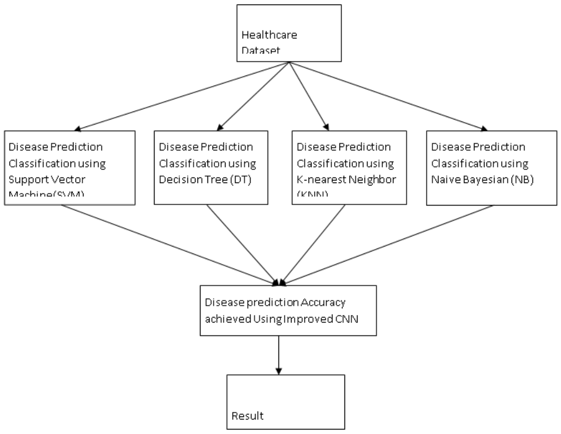
Healthcare Dataset
Disease Prediction Classification using Support Vector Machine(SVM)
Disease Prediction Classification using Decision Tree (DT)
Disease Prediction Classification using K-nearest Neighbor (KNN)
Disease Prediction Classification using Naive Bayesian (NB)
Disease prediction Accuracy achieved Using Improved CNN
Result

Fig. 2. Proposed System architecture

To predict the risk of cerebral infarction disease. This is because these three machine learning methods are widely used.

For T-data, propose CNN-based Advanced Multi modal disease risk prediction (CNN-EMDRP) algorithm to predict the risk of cerebral infarction disease. For S&T data, predict the risk of cerebral infarction disease by the use of CNN -

MDRP algorithm, which is denoted by CNN-MDRP(S&T-data) for the sake of simplicity. In the following section, the details about CNN-EMDRP (T data and S&T data) will be given.

In this work, will introduce machine learning and deep learning algorithms used in this work briefly. For S-data, use four conventional ML algorithms,

- Naive Bayesian (NB)
- K-nearest Neighbor (KNN)
- Decision Tree (DT) algorithm
- Support Vector Machine(SVM)

A. Improved CNN-based advanced multimodal disease risk prediction (CNN-EMDRP)

In this work also utilize softmax classifier. Here introduce to train the improved CNN-EMDRP algorithm, the particular training process explained below. However, those existing work mostly considered structured data. For unstructured data, for example, using convolutional neural network (CNN) to extract text characteristics automatically has already attracted wide attention and also achieved very good results.

The achievement of machine learning techniques is consistent in terms of some performance measure indices. A confusion matrix for actual and predicted class is formed comprising of TP, FP, TN, and FN to evaluate the parameter.

The significance of the terms is given below

TP = True Positive (Correctly Identified)

TN = True Negative (Incorrectly Identified)

FP = False Positive (Correctly Rejected)

FN = False Negative (Incorrectly Rejected)

The performance of the proposed system is measured by the following formulas:

$$\text{Accuracy (Acc)} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$\text{Sensitivity (Sen)} = \frac{(TP)}{(TP+FN)}$$

$$\text{Specificity (Spec)} = \frac{(TN)}{(TN+FP)}$$

$$\text{False Discovery Rate (FDR)} = \frac{(FP)}{(FP+TP)}$$

$$\text{False Omission Rate (FOR)} = \frac{(FN)}{(FN+TN)}$$

True Positive Rate (TPR) and false positive rate (FPR are defined as follows:

$$\text{TPR} = \frac{(FN)}{(FN+TN)}, \quad \text{TFR} = \frac{(FN)}{(FN+TN)}$$

Proposed algorithm of Improved CNN-based advanced multimodal disease risk prediction (CNN-EMDRP) for structured and unstructured data.

4. Implementation and results analysis

In this paper, proposed patient relationship frameworks with other state-of-the-art metric learning techniques, and show that our proposed methods can automatically exceed the measure methods. This proposed Advanced Multimodal using Kernel Support Vector Machine, Naive Bayesian, Decision Tree, K-Nearest Neighbors which is implemented in a high configuration computer. The computer configuration was Intel Core i7 with 8GB RAM.

In this Work have used Pycharm which is an open-source software developed in Python for ML library. An IDE named as Spyder is used to run the program.

Table 1
Run Time Comparison

	CNN-MDRP (S&T-data)	CNN- Advanced MDRP (S&T-data)
Data Center	178.5s	178.2s
Personal Computer	1646.4s	1637.2s

Table 2
Confusion matrix both CNN-MDRP and improved CNN-EMDRP

	TP	TN	FP	FN
CNN-MDRP	404	214	8	3
Improved CNN-EMDRP	409	218	1	1

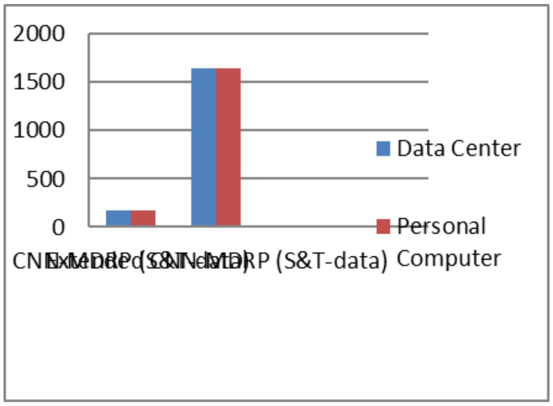

Fig. 3. Running time comparison of existing and proposed method

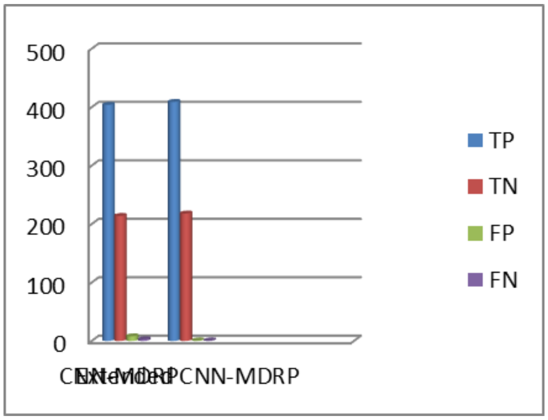

Fig. 4. Comparison of CNN-MDRP and Improved CNN-EMDRP

Table 3
Performance measure indices

Parameters	S&T-DATA	
	CNN-MDRP	Advanced CNN-MDRP
Accuracy (%)	94.8	97.4
Sensitivity (%)	94.1	97.3
Specificity (%)	94.6	97.1
False Discovery Rate (%)	0.72	0.24
False Omission Rate (%)	0.67	0.26

Therefore, in this work, leverage not only the structured data but also the text data of patients based on the proposed Improved CNN-EMDPR algorithm. Find that by combining these two data, the accuracy rate can reach 97.40%, so as to better evaluate the risk of cerebral infarction disease.

5. Conclusion

In this paper focused on building a multi-model classifier which aims at predicting the diseases. This proposed models that predict the disease based on multimodal (Support Vector Machine, Naive Bayesian, K-nearest Neighbor, Decision Tree Algorithm) classifier algorithms with CNN. The proposed model will be very helpful for the medical industry as well as general people. The classifier obtained by supervised machine learning techniques will be very supportive in the field of medical disorders and proper diagnosing. In this Research, proposed a Improved convolutional neural network based Advanced multimodal disease risk prediction (CNN-EMDRP) algorithm using structured and unstructured data from hospital. Proposed algorithm reach 97.4% prediction accuracy with a convergence speed. It’s faster than existing algorithm.

References

[1] P. Groves, B. Kayyali, D. Knott, and S. van Kuiken, The‘Big Data’ Revolution in Healthcare: Accelerating Value and Innovation. USA: Center for US Health System Reform Business Technology Office, 2016.

[2] M. Chen, S. Mao, and Y. Liu, ‘‘Big data: A survey,’’ Mobile Netw. Appl., vol. 19, no. 2, pp. 171–209, Apr. 2014.

[3] P. B. Jensen, L. J. Jensen, and S. Brunak, ‘‘Mining electronic health records: Towards better research applications and clinical care,’’ Nature Rev. Genet., vol. 13, no. 6, pp. 395–405, 2012.

[4] D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, ‘‘A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics,’’ IEEE Trans. Intell. Transp. Syst., vol. 16, no. 6, pp. 3033–3049, Dec. 2015.

[5] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, ‘‘Wearable 2.0: Enable human-cloud integration in next generation healthcare system,’’ IEEE Commun., vol. 55, no. 1, pp. 54–61, Jan. 2017.

[6] M. Chen, Y. Ma, J. Song, C. Lai, and B. Hu, ‘‘Smart clothing: Connecting human with clouds and big data for sustainable health monitoring,’’ ACM/Springer Mobile Netw. Appl., vol. 21, no. 5, pp. 825–845, 2016.

[7] A. Mucherino, J. Papajorgji, and P. M. Pardalos, "K-nearest neighbor classification," in Data Mining in Agriculture, 2009. New York: Springer, 2009, vol. 4, pp.83-106.

[8] P. A. Paul and G. P. Munkvold, "A model-based approach to pre-planting risk assessment for gray leaf spot of maize," Phytopathology journal, vol 94, pp. 1350-1357, 2004.

[9] A. Petrakova, M. Affenzeller and G. Merkurjeva, "Heterogeneous versus homogeneous machine learning ensembles," Information Technology and Management Science, vol. 21, pp. 135-140, 2015.

[10] Y. H. Kim et al., "Crop pests prediction method using regression and machine learning technology: Survey," in International Conference on Agriculture and Biosystem Engeneering, 2008. Berlin: IERI Procedia, 2009, pp. 901-908.

[11] M. Watts and S. Worner, "Predicting the distribution of fungal crop diseases from abiotic and biotic factors using multi-layer perceptrons," in International Conference on Neural Information Processing, 2008. Berlin: Springer, 2009, pp. 901-908.

[12] Y. Chtioui, S. Panigrahi and L. Francl, "A generalized regression neural network and its application for leaf wetness prediction to forecast plant disease," in Chemometrics and Intelligent Laboratory Systems, 1999. USA: Elsevier, 1999, vol. 48, pp. 47-58.

[13] D. Moshou et. al., "Crop health condition monitoring based on the identification of biotic and abiotic stresses by using hierarchical self-organizing classifiers," in Precision agriculture'15, 2015. Wageningen Academic Publishers, 2015, vol. 8, pp. 619-626.

[14] R. Pahlavan, M. Omid and A. Akram, "Energy inputeoutput analysis and application of artificial neural networks for predicting greenhouse basil production," Energy, vol. 37, pp. 171-176, 2011.

[15] R. Pascanu, C. Gulcehre, K. Cho and Y. Bengio, "How to construct deep recurrent neural networks," in ICLR, 2014.