

Healthcare Data Using Machine Learning Models

Vivekanandan G E

Research Scholar, Computer Science Engineering, Monad University (U.P) India

Dr Amit Singhal

Research Guide, Computer Science , Engineering Monad University (U.P.) India

ABSTRACT

The extensive use of computer-based technology in the health care industry has resulted in the accumulation of electronic data. The challenges faced by medical practitioners in the study of symptoms and early sickness detection stem from the substantial amount of data that has to be processed. Supervised machine learning (ML) algorithms have exhibited significant potential in surpassing traditional approaches for illness detection and aiding healthcare practitioners in promptly detecting high-risk conditions. The primary aim of this research is to discern trends in the diagnosis of illnesses through the use of several supervised machine learning models. This will be achieved by assessing performance metrics associated with these models.

Key Word: healthcare, machine learning

I. INTRODUCTION

The field of healthcare has evolved into a highly lucrative enterprise in the modern era. The healthcare business generates and makes use of a significant quantity of data, which may be parsed to provide a patient with information on a specific illness they may be suffering from. This healthcare information will be utilised in the future to provide patients with treatment that is both effective and of the highest possible quality. This is another area that might benefit from some refinement through the use of the informative data gained from healthcare sciences. However, because the data is available in such a large quantity, one of the most difficult challenges is to extract the information from the data. As a result, several data mining and machine learning approaches are utilised. The anticipated outcome of this research is to be able to diagnose the illness in advance. This will allow the potential threat to human life to be averted at an earlier stage, which will not only save lives but also bring down the overall cost of providing treatment.

The non-manual style of medical therapy is superior in terms of both enhancing and comprehending the state of human health, and as a result, it need to be implemented in India as well. The primary goal is to implement the idea of machine learning into the medical field in order to enhance the way in which patients are cared for. Learning machines have already made it considerably simpler to recognize and anticipate a wide variety of illnesses. The predictive analysis of the disease, which is carried out with the assistance of a great number of machine learning algorithms, enables us to forecast the disease and assists in treating patients in an efficient manner. The prediction of disease using machine learning makes use of the patient's medical history and data as well.

This is accomplished by employing a variety of concepts and methods, including data mining, machine learning, and other algorithms. Many works have also utilised data mining techniques to pathological data in order to forecast certain diseases. This has been done in a number of different ways. These strategies were devised with the intention of predicting the recurrence of specific illnesses in advance. Additionally, several systems attempted to foresee the disease while simultaneously regulating it. Deep learning has recently been applied to a variety of subfields within machine learning, which has resulted in a paradigm shift towards machine learning models that can learn and comprehend the hierarchical representations of raw data with the use of basic preprocessing. More consideration is being given to the process of disease forecasting as a result of the emergence of a concept known as big data technology.

MACHINE LEARNING OVER BIG HEALTHCARE DATA PREDICTS DISEASE

According to a survey by McKinsey, fifty percent of Americans have one or more chronic illnesses, and seventy-five percent of the cost of medical care in the United States goes towards the treatment of chronic disorders. The prevalence of chronic diseases is on the rise in conjunction with the general increase in living conditions. The cost of treating chronic diseases in the United States has been estimated at an annual average of 2.7 trillion dollars. This sum accounts for 18 percent of the total yearly GDP of the United States. The issue of chronic illnesses as they relate to healthcare is also a very serious one in a great number of other nations. According to Chinese research on nutrition and chronic illnesses published in 2015, in China, chronic diseases account for

86.6% of all fatalities. This statistic indicates that chronic diseases are the leading cause of mortality in China. As a result, carrying out risk assessments for those who have chronic conditions is very necessary.

The accumulation of electronic health records (EHR) is becoming an increasingly easy process as a result of the expansion of medical data. In addition, the authors first introduced a bioinspired high-performance heterogeneous vehicular telematics paradigm. This paradigm was designed in such a way that the gathering of mobile users' health-related real-time large data could be accomplished with the deployment of advanced heterogeneous vehicular networks. A healthcare system that uses smart clothes to do sustained health monitoring was suggested by Chen et al. After doing extensive research on heterogeneous systems, Qiu et al. came up with the best findings for cost reduction on tree and simple route scenarios for heterogeneous systems. The electronic health record (EHR) stores the statistical information, test results, and illness history of patients. This provides us with the ability to find prospective data-centric solutions that might minimise the expenses of medical casework. studies both a data coherence protocol for the PHR (Personal Health Record)-based distributed system as well as an efficient flow estimation algorithm for the telehealth cloud system were devised and suggested by the authors. Bates presented an ideal big data sharing strategy to manage the complicated data set involved in telehealth using cloud approaches, as well as six uses of big data in the area of healthcare. One of the uses is to detect high-risk patients, which may be used to lower medical costs, since high-risk patients often need costly medical treatment. This is one of the applications that can be used. In addition to this, the first publication to propose a healthcare cyber-physical system was the one that pioneeringly introduced the idea of prediction-based healthcare applications, one of which was health risk assessment. Predictions made using classic disease risk models often need the use of an algorithm for machine learning (such as logistic regression and regression analysis, among other examples), and more specifically, an algorithm for supervised learning, which requires the use of training data labelled with certain categories in order to "train" the model. Patients in the test set might be assigned to the high-risk or low-risk category, depending on the severity of their condition. These models have been extensively researched due to their usefulness in therapeutic settings. Nevertheless, these strategies are flawed and suffer from the qualities listed below.

HEALTHCARE MACHINE LEARNING CLASSIFIERS

There has been a slow but steady increase in the number of life support machines and systems available to patients during the last several decades. As a consequence, this expansion leads to an increase in the standard human lifespan. However, modern health care systems confront a number of hurdles and problems, some of which include providing patients with false information, protecting the privacy of their data, not having enough reliable data, not having enough medical information, and not having enough classifiers for predictions. In order to address these problems, a great variety of illness diagnostic and prediction systems have been created. Some examples of these systems are the expert system, the clinical prediction system, the decision support system, and the personal health record system. The purpose of this system is to give assistance to doctors and medical professionals so that they may accurately diagnose ailments. Finding the symptoms of an illness more precisely might be seen of as being the same thing as making a diagnosis of that condition.

Once the symptoms have been accurately diagnosed, the condition will be simple to treat from that point on. However, it has been brought to our attention that these medical systems need a significant amount of computing power and resources. It has also been noted that the computing demands of medical systems are significant. Diabetes is a non-communicable illness that affects individuals all over the globe, and it is expected to impact 300 million people by the year 2030. A lot of academics try to solve the problem of missing values in medical data by doing one of two things: either they identify the missing value and remove the related data instances from the dataset or they use some default techniques to fill in the missing values (such the mean, the median, or the neighbour, for example). However, none of these approaches is sufficient to deliver the best possible outcomes. In addition, there are certain data that include outliers, which brings down the overall effectiveness of the classifier. A select group of academics also concentrate on the identification of outliers in medical datasets, although to this day, the topic has not been thoroughly researched. This study also takes into consideration two well-known data issues: i) missing value imputation, and ii) outlier. Both of these issues are referred to as outliers.

OBJECTIVES

1. To develop and evaluate machine learning models for the classification and disease prediction of healthcare data.
2. To assess the performance and accuracy of different machine learning algorithms in healthcare data analysis.

II. RESEARCH METHODOLOGY

In this section, we will cover the methodology of the hybrid feature selection (entropy-notion) strategy that was recommended for use with medical datasets. The attributes that have a strong correlation cannot be a part of the feature subset, according to a piece of research that was conducted by Hall in the year 2000 and published in the year 2000. In this study, it was also said that the information gain that the traits have is proportionate to the degree to which they may be regarded independent from one another. This was discussed in relation to the study's findings on the relationship between information gain and independence. As a consequence of this, it is predicted that the outputs will be of a higher quality when compared to data that has not been inspected.

DATA ANALYSIS

This section will begin by providing a description of the experiments that were carried out as part of the current study utilising a total of 14 real-world benchmark medical datasets that were collected from the data repository at the University of California, Irvine (UCI). The sizes of the datasets differ; for instance, one of them has a very small sample size but a very high dimension, another has a high dimension but a relatively large sample size, and yet another has a medium sample size but fewer features. Table 1 is a summary of the datasets that may be accessed. Following the completion of the tabularization of the data, the next step is to do the statistical analysis.

Table 1 the UCI data set that was chosen (the original).

Problem name	Number of non-target attributes	Number of classes	Number of examples
Breast cancer Wisconsin	10	2	699
Dermatology	34	6	366
Ecoli	8	8	336
Heart (Hungarian)	13	5	294
Heart (Swiss)	13	5	123
Heart (Cleveland)	13	5	303
Hepatitis	19	2	155
Liver disorder	6	2	345
Lung cancer	56	3	32
Lymphography	18	4	148
New- thyroid	5	3	215
Pima Indian diabetes	8	2	768
Primary tumor	17	22	339
Sick	29	2	3772

If you take a look at Table 1, you will notice that some of the datasets, such as Breast cancer, Dermatology, Primary tumour, and Sick, are examples of relatively large datasets with a greater number of characteristics. On the other hand, some of the datasets, such as Heart (Swiss), Hepatitis, and Lung cancer, are examples of high dimensional small size datasets. The datasets that can still be accessed are those that are not too large and include only a select few variables and attributes. Table 1 has these various datasets for your perusal.

Try out the newly added feature-selection method.

Using 14 distinct datasets, the method of feature selection that has been built and put into use is evaluated. The attributes of the datasets that have been chosen for presentation are outlined in an overview that may be found in table 2. The table contains information on each of the following topics: the name of each dataset, which is represented as DN; the number of occurrences; the number of features; the number of selected features; and the selected features.

The fundamental terms that are used in this article are provided below for the aim of offering more information regarding the work that was completed in this study pertaining to feature reduction.

- Original Features: Features that are known to be contained in the original datasets that have been submitted to the UCI repository are what are referred to as "original Features." When it comes to determining the cause of sickness, the decisions that need to be taken are those that are made by trained experts (physicians).
- chosen Features: - Features are known as selected features when they are picked from each original dataset by decreasing noisy (irrelevant/redundant) features using a suitable approach for feature reduction. The process of choosing features from each dataset is known as "selecting features."
- Independent Features: The qualities of a feature that are known as "independent features" are the characteristics of a feature that do not depend on any other characteristics included within the feature. Because each of these characteristics is contingent on the target feature, their overall contribution to the decision-making process is substantial because of the interdependence among them.

A succinct explanation is also provided below with reference to the feature number that is shown in Table 2. Let's take a look at a dataset on a disease called the Heart (Swiss) dataset so that we can better understand the situation. The specialists working at the repository at the University of California, Irvine have decided to consider

13 non-target traits to be features, out of a total of 76 unprocessed characteristics. In the experiment that is now being conducted, these thirteen traits are being treated just as though they were the initial characteristics. The following is a description of each of them, which have been given the new names A1, A2, A3,..., A13 (or, to shorten, 1, 2, 3,..., 13).

1. Age (A1): the number of years that have passed since the person was born. 2. Sex (A2): the individual's gender, with 1 being a male and 0 representing a female.

2. Cp (A3): a type of chest pain (there are four different types of chest pain). [This is the typical case of angina. Angina of a non-typical kind is value 2. Value 3: Pain Not Associated with the Heart. Asymptomatic, with a value of 4].

3. The patient's blood pressure when they were at rest when they were admitted to the hospital, measured in millimetres of mercury.

4. Chol (A5): the quantity of cholesterol that is found in the serum, represented as mg/dl, fractional basic units.

5. Fbs (A6): blood sugar while fasting more than 120 mg/dl [1 = true; 0 = false] 7. Fbs (A7): blood sugar after eating more than 120 mg/dl 8.

6. results on an electrocardiogram taken when the patient is at rest (A7): there are three distinct types of results like this. [Value 0: normal; Value 1: having ST-T wave abnormality (T-wave inversions and/or ST elevation or depression of > 0.05 mV); Value 2: showing probable or definite left ventricular hypertrophy according to Estes' criteria] Value 0: normal; Value 1: having ST-T wave abnormality (T-wave inversions and/or ST elevation or depression of > 0.05 mV); Value 2: indicating likely or Value 0 indicates that the ST-T wave is normal, whereas Value 1 indicates that there is an anomaly in the ST-T wave (T-wave inversions and/or ST elevation or depression of > 0.05 mV).

7. Thalach, often known as A8, is the greatest potential heart rate that may be obtained.

8. Exercise-induced angina is referred to as Exang (A9). A score of 1 shows that it is present, while a score of 0 indicates that it is not.

9. Oldpeak (A10): A fall in ST elevation as compared to the resting condition, brought on by activity.

10. Slope (A11), also known as the slope of the peak exercise ST (T-wave) segment, with values [Value 1: the slope is ascending, Value 2: it is horizontal, and Value 3: it is falling]. [Value 1: the slope is ascending, Value 2: it is horizontal, and Value 3: it is descending].

11. Ca (A12): the number of main blood arteries that are coloured by fluoroscopy (ranging from 0-3) in the patient.

12. Thal (A13): Thalassemia is a blood disease that can be handed down through generations [3 = normal; 6 = fixed defect; 7 = reversible defect]. [3 = normal; 6 = fixed defect; 7 = reversible defect].

As a consequence of this, the hybrid approach to feature selection that was applied in this process resulted in the selection of the characteristics listed below: 9 (A9), 10 (A10), and 11 (A11).

The significance of the method that has been described can be seen in the performance of the classifiers, particularly J48 (Java version of C4.5) [133], naive Bayes [134], and JRIP (Java version of RIPPER [135]), over the dataset (both before and after selecting their features), as shown in Tables 3 and Table 4, respectively. These tables contain the results of the classifiers' analysis of the dataset.

The evaluation of the suggested technique for the selection of hybrid features makes use of the aspects that are listed below.

- The three states of the learners that were put to use were naive Bayes, C4.5, and RIPPER.
- A 10-fold cross-validation conducted over the course of ten distinct runs (both before and after feature selection) to ascertain the average proportion of accurate classifications as well as the standard deviation
- A reduced number of features overall as a direct result of the approach that was selected

Table 2 datasets with a certain property

Name of dataset (DN)	Number of instances (NI)	Number of features in the original dataset	Number of selected features	selected features
Breast cancer Wisconsin	699	10	6	1,3,4,5,6,9
Dermatology	366	34	21	2,3,4,5,6,7,9,13,14,15,16,19,20,21,22,26,27,28,29,30,33
Ecoli	336	8	6	1,2,3,5,6,7
Heart (Hungarian)	294	13	6	2,3,6,9,10,11
Heart (Swiss)	123	13	3	9,10,11
Heart (Cleveland)	303	13	8	2,3,8,9,10,11,12,13
Hepatitis	155	19	9	1,2,6,11,12,14,17,18,19
Liver disorder	345	6	2	3,5
Lung cancer	32	56	17	1,3,5,9,13,14,15,20,21,25,26,38,41,45,48,50,56

Lymphography	148	18	10	1,2,7,8,9,11,13,15,16,18
New- thyroid	215	5	5	1,2,3,4,5
Pima Indian diabetes	768	8	4	2,6,7,8
Primary tumor	339	17	12	1,2,3,4,5,7,9,10,13,15,16,17
Sick	3772	29	8	1,8,10,14,15,20,24,29

Table 3 10 fold cross validation with 100 iterations (before feature selection) evaluating the performance of the J48, JRIP, and naïve bayes classifiers on the dataset.

Problem name	Number of non-target attributes	J48 (acc.±s.d.)	JRIP (acc.±s.d.)	naive bayes (acc.±s.d.)
Breast cancer Wisconsin	10	73.28±6.05	71.45±6.45	74.70±7.74
Dermatology	34	90.11±3.34	86.61±4.89	91.01±2.41
Ecoli	8	82.13±5.73	81.41±6.30	84.51±5.46
Heart (Hungarian)	13	78.22±7.95	79.57±6.64	80.95±6.27
Heart (Swiss)	13	36.45±13.73	38.08±9.36	36.38±13.13
Heart (Cleveland)	13	76.94±6.59	78.95±6.77	81.34±7.20
Hepatitis	19	77.22±9.57	76.13±9.04	81.81±9.70
Liver disorder	6	62.84±7.40	68.57±7.55	54.89±8.83
Lung cancer	56	68.25±21.50	73.92±19.15	76.42±41.12
Lymphography	18	74.84±11.05	75.11±41.37	82.13±8.89
New- thyroid	5	9.11±4.32	94.01±4.51	94.1±4.16
Pima Indian diabetes	8	73.89±5.27	74.18±4.14	75.7±5.32
Primary tumor	17	41.139±6.94	38.34±5.51	47.71±6.46
Sick	29	93.13±0.55	93.29±4.68	92.8±84.36

Table 4 Performance of J48, JRIP and naive Bayes classifiers on the data sets using 10-fold cross validation over 10 runs (after feature-selection)

Problem name	Number of non-target attributes	J48 (acc.±s.d.)	JRIP (acc.±s.d.)	naive bayes (acc.±s.d.)
Breast cancer Wisconsin	6(10)	78.71±1.39	76.97±2.03	80.13±1.23
Dermatology	21(34)	90.88±1.20	88.15±2.18	92.14±1.23
Ecoli	6(8)	82.49±2.46	80.91±1.82	85.61±1.19
Heart (Hungarian)	6(13)	78.84±1.19	80.06±1.81	80.7±1.27
Heart (Swiss)	3(13)	36.43±13.73	38.08±9.36	36.38±13.13
Heart (Cleveland)	8(13)	76.94±6.59	78.95±6.77	81.34±7.20
Hepatitis	9(19)	76.2±9.57	76.13±9.04	81.81±9.70
Liver disorder	2(6)	62.84±7.40	68.57±7.55	54.89±8.83
Lung cancer	17(56)	68.25±21.50	73.92±19.15	76.42±41.12
Lymphography	10(18)	74.84±11.05	75.11±41.37	82.13±8.89
New- thyroid	5(5)	9.11±4.32	94.01±4.51	94.1±4.16
Pima Indian diabetes	14(8)	73.89±5.27	74.18±4.14	75.7±5.32
Primary tumor	12(17)	41.139±6.94	38.34±5.51	47.71±6.46
Sick	8(29)	93.13±0.55	93.29±4.68	92.8±84.36

Performance study of the proposed feature-selection method

The following is a depiction of the selected learners' dataset accuracy results, which can be seen in Tables 3 and 4 respectively.

Eliminating superfluous characteristics almost always results in an increase in the competent classifier's accuracy (expressed as a percentage). The performance of the naive Bayes classifier is superior the majority of the time, which suggests that the technique that is proposed for feature-filtration successfully decreases the amount of features. The naive Bayes classifier is an effective tool for assessing datasets that contain a wide range of characteristics, and the approach that is advised chooses which of these qualities to employ. Those traits that give very little or no more information have been omitted from the filtered list. As a direct consequence of this, datasets may be necessary in order to reduce the significance of dependent traits.

Reduced feature datasets have a classifier accuracy standard deviation (sd) that is lower than that of the original datasets. This means that reduced feature datasets are more reliable. This is due to the fact that the standard deviation in the original datasets is greater.

The amount of time needed for learning the hybrid model will decrease as a result of the reduction in the number of features, and the same is true for the number of induced rules.

Dimension reduction, which is also known as noise reduction, works well in some datasets. However, the metrics of other learners either do not change at all, perform extremely poorly, or perform poorly enough to be acceptable. It's probable that this is due to the different ways in which students take in knowledge, which need the inclusion of additional material throughout the course.

The process of selecting and extracting features helps to increase the accuracy of classification. This is accomplished by compressing the original data set into a data vector with fewer dimensions during the process of feature selection and extraction. A method for selecting hybrid features that is based on entropy is going to be shown in the next chapter. This technique combines the characteristics of the dataset subsets in order to produce the most useful subset feasible. The MDDS uses a wide variety of techniques for the data collection and segmentation processes. The performance of the model was evaluated utilising C4.4 (decision tree), JRIP (sequential covering), and naive Bayes classifiers on the datasets that could be available in the UCI data repository. Each dataset has produced findings that meet expectations.

III. CONCLUSION

The primary objective of this research endeavour is to improve the healthcare infrastructure by an examination of diseases, addressing the limitations of existing disease management methods, and creating a Medical Disease Diagnosis System (MDDS). The research also highlights the contemporary patterns in the advancement of the healthcare system, the examination of risk factors linked to specific cardiovascular disorders, and the development of a machine learning model to tackle overlooked occurrences of preterm delivery in the field of obstetrics. The growing ubiquity of computing technology has resulted in the incorporation of data as an essential element inside the digital domain. The phenomenon of digitization is leading to the production of a substantial volume of data. There is a need to create an intelligent model that utilises machine learning techniques to categorise the large amount of data included inside our databases. The rate of human data analysis is considerably slower when compared to the magnitude of data storage. The process of classification has more importance in the domain of healthcare data due to its role in enabling the identification, diagnosis, and treatment of patients through the use of insights obtained from categorised information. The healthcare data possesses attributes that can be described as voluminous, multidimensional, and diversified in nature. As a result, there is an increasing demand for the application of data mining methodologies in order to facilitate categorization and prediction tasks.

REFERENCES

- [1]. Acharjya, D. P. (2020). A Hybrid Scheme for Heart Disease Diagnosis Using Rough Set and Cuckoo Search Technique. *Journal of Medical Systems*, 44(1), 27.
- [2]. Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Plawiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer methods and programs in biomedicine*, 179, 104992.
- [3]. Narayan, S., & Sathiyanamoorthy, E. (2019). A novel recommender system based on FFT with machine learning for predicting and identifying heart diseases. *Neural Computing and Applications*, 31(1), 93-102.
- [4]. Bucholc, M., Ding, X., Wang, H., Glass, D. H., Wang, H., Prasad, G., ... & Finn, D. P. (2019). A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. *Expert systems with applications*, 130, 157-171.
- [5]. Nilashi, M., Ahmadi, H., Shahmoradi, L., Ibrahim, O., & Akbari, E. (2019). A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. *Journal of infection and public health*, 12(1), 13-20.
- [6]. Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D., & Lyu, C. (2019). A reliable method for colorectal cancer prediction based on feature selection and support vector machine. *Medical & biological engineering & computing*, 57(4), 901-912.
- [7]. Nilashi, M., Ibrahim, O., Samad, S., Ahmadi, H., Shahmoradi, L., & Akbari, E. (2019). An analytical method for measuring the Parkinson's disease progression: A case on a Parkinson's telemonitoring dataset. *Measurement*, 136, 545-557.
- [8]. Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An Automated Diagnostic System for Heart Disease Prediction Based on χ^2 Statistical Model and Optimally Configured Deep Neural Network. *IEEE Access*, 7, 34938-34945.
- [9]. Kanimozhhi, U., Ganapathy, S., Manjula, D., & Kannan, A. (2019). An intelligent risk prediction system for breast cancer using fuzzy temporal rules. *National Academy Science Letters*, 42(3), 227-232.
- [10]. Gokulnath, C. B., & Shanthyrajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. *Cluster Computing*, 22(6), 14777-14787.
- [11]. Kirk, R. A., & Kirk, D. A. (2019, July). Introducing a Decision Making Framework to Help Users Detect, Evaluate, Assess, and Recommend (DEAR) Action Within Complex Sociotechnical Environments. In *International Conference on Human Interface and the Management of Information* (pp. 223-239). Springer, Cham.
- [12]. Devarajan, M., & Ravi, L. (2018). Intelligent cyber-physical system for an efficient detection of Parkinson disease using fog computing. *Multimedia Tools and Applications*, 1-25.
- [13]. Peker, M. (2019). A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM. *Journal of medical systems*, 40(5), 116.
- [14]. Huda, S., Yearwood, J., Jelinek, H. F., Hassan, M. M., Fortino, G., & Buckland, M. (2019). A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. *IEEE access*, 4, 9145-9154.
- [15]. Aličković, E., & Subasi, A. (2019). Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Computing and Applications*, 28(4), 753-763.