B-PRED: AN INTELLIGENT AND ADAPTABLE MEDICAL DIAGNOSIS SYSTEM BASED ON BAGGING MACHINE LEARNING

Soreen Ameen Fattah, Hussein Attya Lafta, Sura Z. Alrashid

Abstract— Advancement in medical information systems has facilitated the development of automated diagnosis systems. Several Artificial Intelligence (AI) techniques have been implemented and studied in modern researches to come up with the most suitable and accurate medical diagnosis system. Bagging is one of these techniques, and it has been proven by several researches to be a powerful and convenient tool for such systems. In this research; bagging algorithm is used to produce a diagnosis system for two of the most common diseases: diabetes and heart diseases, where this algorithm used verified datasets of attributes that are combined with the same attributes values submitted by the patient through a dedicated interface. Testing the system and comparing it to other prediction systems proved its efficiency and accurate prediction rates.

Index Terms— Machine Learning, Ensemble Learning, Bagging, Medical Diagnosis, Diabetes, Heart Diseases, dataset splitting

1 INTRODUCTION

CHANGES in global population and demography, advancement in the medical field were noted, and the need for an easy and fast access to medical consultation and advisement has also developed. In response to the advancement in the healthcare field that was accompanied with advancement in the information technology fields too, integration and collaboration of these two fields has forced itself into existence in the research world. Remote access to illness diagnosis and medical advisement was one of the variant research fields and software applications that emerged. This integration involved combining patient’s physiological data acquisition equipment with real-time and automated health care analyses. Information technology researchers have employed the power of Artificial Intelligence (AI) into healthcare applications, for its techniques’ ability to produce diagnosis that is based on different criteria and on data that comes from multiple sources. thanks to the internet; these medical data sources are mostly freely available, and developers can use them in developing their applications with confidence that these data are verified and revised by workers in the healthcare domain. One of the techniques that are widely used in medical diagnosis systems in literature are the ensemble machine learning techniques [1]. An ensemble is a combination of several learning models that would results in having an improved machine learning performance. This technique is mostly useful when having very diverse datasets (Like medical records of diverse population attributes) with different data types (text, numbers…) that come in various data structures Bagging, as one of the ensemble methods, is a technique that is mostly used to aggregate different decisions and predictions coming from different machine learning algorithms’ runs to come up with a more accurate prediction about the problem the system is tackling [2]. The work elaborated in this research shows that using an ensemble technique (bagging in specific) in producing and accurate medical diagnosis produces accurate diagnosis, especially when depending on textual data provided by patients. This system (named B-Pred as an acronym for Bagging Prediction) also allows physicians to submit diagnosis to unclassified symptoms (submitted by patients) if the ensemble learning system was unable to produce one. The next section presents recent researches of using machine learning approaches in medical diagnosis, followed by elaborated explanation of the technical aspects of the diagnosis system. Afterwards; test results and analysis of findings is presented, along with screenshots of implementation interfaces. The manuscript is concluded with highlighting the main techniques used and the main advantages of adopting the system based on analysis of findings.

2 RELATED WORK

The diagnosis of several diseases doesn’t only depend on symptoms provided by the patients, but also overall health condition, life style, historical medical data. Artificial intelligence was introduced as a tool to help in conducting diagnosis through the collection of such data and build relationships to reach a diagnosis. Heart diseases had been classified as the leading cause of death in the world [3], and hence huge amounts of researches took these diseases in building an automated diagnosis system, like the work by [3] in which the authors used the decision trees which was used (by other researchers like [4], and [5], in medical data mining wasn’t up to expectations, thus the bagging algorithm is introduced as a new helping diagnosis tool since bagging combines a series of learned models to create an improved composite model there for increases accuracy, and its open any future improvements. The authors in [6] considered that, in the field of machine-aided diagnosis, accuracy is critical. And in order to get enhancement in the accuracy of the base classifiers, two main methods were suggested: the ensemble data-mining method (EDMM) and the learning algorithms having a combination of multiple base models. As a result, a drastic enhancement was observed in the accuracy of the results. An important point worth mentioning here is that it is recommended to use the proper ensemble classifier for each dataset, specific cases didn’t need any ensemble classifier. Using variations to ensemble classifiers has been widely adopted in recent literature, the work in [7] this study the use of ensemble learning was investigated in order to improve the
classifiers where the main powerful and popular representatives are bagging [8], boosting [9] and random subspace [10]. Those have proven their efficiencies in many practical classification problems. The study of [11] focused on valvular heart disorders to evaluate the performance of the investigated ensemble methodology. The results suggested the feasibility of the ensemble classification methods for these kinds of heart diseases.

3 THE B-PRED LEARNING SYSTEM

The predication system that is developed in this research will be based on Bagging ensemble machine learning technique, where an online medical training set is used to build a prediction model. An interface for the users (patients and physicians) is developed in C# programming language to facilitate submitting symptoms and interacting with the system. The system is composed mainly of two parts: one for training from available medical datasets (heart diseases and diabetes in specific), and another for receiving the symptoms and provide diagnosis (and tune the system’s prediction model in case of wrong diagnosis). A clarification on the workflow between the two parts is illustrated in figure 1.

The training module, which is developed using R-language as a backend of the system, embeds bagging algorithm to build a prediction model from a training dataset. This bagging algorithm will classify training datasets, and produce predications based on received symptoms. The training works over two folds: producing a diagnosis based on training set’s prediction model using Bagging algorithm, and measuring the accuracy of the used dataset in the process of diagnosis production, through constant enhancements on the prediction rules. The training process is illustrated in figure 2

On the other side, the patient enters symptoms into a test set through a C# interface program (frontend of the system). These test instances will be sent to the prediction system to produce a medical diagnosis. The predicted diagnosis is sent to the patient for confirmation; if the patient confirms that the system’s diagnosis is the same as the one received from the physician, then these set of symptoms are added to the prediction system to help in future diagnosis, but if the produced diagnosis doesn’t match the physician’s diagnosis, the physician is asked (through another C# program interface) to submit the correct diagnosis, and all of these data instances are added to the prediction system for future diagnosis.

3.1 Training

Indirect classification in machine learning systems relies on having a priori dataset combined with classification rules. [12] In medical applications in specific, this type of classification is best since membership functions for classification deterministically known classifying function [13]. Bagging (Bootstrap Aggregation) is a technique that is mostly used to aggregate different decisions and predictions coming from different machine learning algorithms’ runs to come up with a more accurate prediction about the problem the system is tackling. The famous Classification and Regression Trees
(CART), which is known for producing estimations with high variance, are used in each estimation iteration in order to produce a better estimation and reduce variance in the final result. This technique is a widely used ensemble technique that finds the most accurate prediction from multiple predictions [1], and according to [15] bagging provides the highest predicted values accuracy in big data, due to the accumulation of accuracy measures in the training sets calculated separately. Figure 3 shows a simplification of how Bagging algorithm works.

---

**Fig. 3.** Bagging algorithm

---

This technique is a widely used ensemble technique that finds the most accurate prediction from multiple predictions [1], and according to [15] bagging provides the highest predicted values accuracy in big data, due to the accumulation of accuracy measures in the training sets calculated separately. Figure 3 shows a simplification of how Bagging algorithm works.

---

**R-Language** has been widely adopted in machine learning systems for its ability to deal with large and various dataset to produce a prediction. It has the “bagging” ensemble learning algorithm embedded within its compiler’s libraries, which facilitates the prediction procedure, especially the library (iPred) developed in the project by [14]. This library is composed of several classes which are used in the ensemble prediction process; the ‘ipredbag’ class for example, has several arguments and functions, like ‘bagging’, to produce various responses.

Several datasets could be called within the bagging algorithm, which makes the system easily adaptable for multiple medical cases, and the enhancement proposed in the B-Pred system, assures reaching good accuracy of predications, once the training dataset are called. The data set is separated into two parts; the majority of the data are used in the training process (about 80% of the dataset), while the other part is used for the testing process at the user’s end. A pseudo code of the training module is presented in the following textbox:

---

```
// (initialize parameters)
define D = ∅ // the ensemble
define L = the number of classifiers // the number of classifiers to train 
for k = 1 to L // determine range
take a bootstrap sample S_k from I // Bootstrapping
build a classifier D_k using S_k // as the training set
D = D U D_k // add classifier to current ensemble
return D // (classification phase)

run D_1, .., D_L on the input(x) // examine input
take a majority vote on each of subset of the data (bootstap samples). Classification takes place afterwards; and assigns labels to each of the discovered classes, upon which diagnosis and decision making will be made.
```

3.2 Producing diagnosis

The patient (system user) submits the symptoms into the system through the C# interface that is integrated into a webpage. The 20% remaining part of the dataset, that wasn’t used for training, is fed into the prediction system along with the attributes (symptoms) submitted by the user. Bagging algorithm is used again to produce a diagnosis given the provided attributes, and the learnt model. The pseudo code of the diagnosis module (user’s frontend) is shown below:

---

```
read variables ← user input // user enters data on interface
send Instance to test(x) file // data is exporting .csv file
DS ← Training set(y) // read training data from dataset
DS TEST ← read test(x) // read test data from test dataset
DSB ← DS // create a copy of the defined dataset
DS MODEL ← bagging(DSB) // train data using bagging algorithm
Result ← predict(DS_MODEL, DS_TEST) // use model to predict test data
Print Result to output(z) file // export result to text file
read Result to output(z) // read file from text file
if Result = True // examine if result is true
    add instance to Training
else // if result is wrong
    read correct answer ← user input // get correct answer
    update Result // add it to the instance
add instance to Training
```

Prediction is produced using the test data, and the produced diagnosis is mapped with these test datasets. The produced prediction is mapped against the real diagnosis (available with the test dataset) to verify the prediction. Once the prediction is produced, it is displayed on the user’s interface to verify that
the system has reached the correct diagnosis. If this diagnosis matches the diagnosis of the physician, then the predicted diagnosis is added as a training dataset instance, otherwise; the user is asked to submit the correct diagnosis (or the physician) to add it to the training set's instances to enhance future predictions.

4 SYSTEM IMPLEMENTATION

B_Pred. was trained on two datasets that are related to heart diseases and diabetes. As mentioned in the system's description; more datasets for other diseases can be added to the training and testing processes to widen the range of users and available diagnosis. The dataset used is available on the internet, on the Center for Machine Learning and Intelligent Systems website, (namely the Cleveland Dataset and the Diabetes Dataset) where the attributes that indicate a certain disease are set upon physician's advice. Each of these datasets are used to train the system produce a judgment of whether the user has that disease (diabetes or heart conditions) or not. The user chooses the condition he seeks diagnosis about as the first step in using the system through the interface shown in figure 4:

Attributes that are submitted by the user include who seek diagnosis on heart diseases: cp (chest pain type which varies between: typical angina; atypical angina; non-angina pain and asymptomatic), trestbps (resting blood pressure), chol (serum cholesterol), fbs (fasting blood sugar), restecg (resting electrocardiographic), thalach (maximum heart rate), exang (exercise induced angina), oldpeak (ST depression induced by exercise relative to rest), slope (the slope of the peak exercise ST segment), ca (number of major vessels colored by fluoroscopy), thal (Thalassemia) in addition to the age and gender of the patient. These are displayed in an interface on the patient’s side as in the one shown in figure 5.

Diabetes medical condition prompts the user to enter valued for attributes: nucleus of the solitary tract (NTS), proliferator activated receptor gamma coactivator 1-alpha (PGC), diastolic blood pressure (DBP), fasting plasma insulin (INS), Body Mass Indicator (BMI), TSFT and DPF tests along with the patient’s age. These are submitted through the interface in figure 6.
After the system is trained on the training dataset (which is 80% of the whole dataset), it combines the data submitted by the user with the testing dataset (the remaining part of the dataset) to produce a diagnosis. This diagnosis is displayed to the user along with the confidence level about it. The user is asked to verify if the diagnosis is correct or (whether he was diagnosed with diabetes or has a heart condition or not). The interface through which the user submits this piece of information is shown in figure 7.

The same interface is displayed for diabetes diagnosis that shows the level of confidence (accuracy) of the system’s prediction, of whether the patient is diabetic or not. This interface is shown in figure 8:

The result is either positive (the attributes submitted indicates that the user has the disease) or negative. If the diagnosis doesn’t match the one issued by the physician, the user can submit the correct diagnosis through the input text box in the dialog box.

5. EVALUATION OF OUTCOMES
The confusion matrix is a very effective method used to evaluate an automated system’s performance against the performance of a human expert. It consists of 4 main values: one for the True Positive TP count (the number of times system’s outcome for positive cases was positive), one for the True Negative TN count (the number of times system’s outcome for negative cases was negative), one for the False Negative FN (the number of times system’s outcome for positive cases was negative), and another for False Positive FP count (the number of times system’s outcome for negative cases was positive). Once this matrix is set, several accuracy measurements can be drawn from the values in it. Accuracy for instance is a measure of counts for when the system produced correct predictions. It is done using equation 1:

\[
\text{Accuracy} = \frac{\text{TN}+\text{TP}}{\text{TP}+\text{FN}+\text{TN}+\text{FP}}
\]  

(1)

Precision and recall are other measures that shows how good the prediction is, they are computed using equations 2 and 3 respectively:

\[
\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}
\]

(2)

\[
\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}
\]

(3)

As begging divides, the training data set into smaller subsets, the number of these subsets is critical to the produced accuracy of predications. By computing the accuracy, precision and recall for each number of bags used in the system’s implementation, the best number of bags was between 20 to 25 bags. These data are illustrated in table 1 below:

Cross-validation was applied to the training sets to make a judgment on when will the subsets of training data generalize to an independent data set. The data set was first split to three equal thirds (each part has 33% of the data, training was done on 66% of the data and the remaining 34% was used for testing). Then the data set was split into 80-20% portions where 80% was used for training and the 20% was used for testing.

6. SYSTEM’S PERFORMANCE ANALYSIS
The B-Pred. system performance and results accuracy was compared against two of the most commonly used machine learning algorithms for prediction: Neural Network (NN), Decision Trees (DT), and Naïve Bayes Classifier (NB). The diagnosis was done using the same dataset with the same procedure of dividing the data into 3 parts, 2 for training and the remaining part for testing (as stated above over two trials, one is 66-34% and the other on 80-20%). The diagnosis was produced on the first run using the training subset of the data alone, once for each of the diseases. A second run of the system was done after adding the remaining part of the dataset and using them in the diagnosis procedure. Bagging showed outstanding performance against the other algorithms. Table 2 shows test metrics values for each of the tested prediction systems on heart disease diagnosis, when using 66% of the data in the training process.
After updating the dataset from user’s feedback, bagging algorithm showed major enhancement in test metrics over the other techniques. These values are shown in table 3:

Table 4 shows the same test metrics for the same techniques and also on heart disease diagnosis, but with 80% of the data set for training with the initial data set.

When the data was updated by user’s feedback, predication using bagging showed its superiority again, with the 80% training data, with the updates, as shown in table 5

The chart in figure 9 shows averaged accuracy of prediction before and after using the test data subset for the heart disease dataset.

![Accuracy Comparison](image)

As noted from the chart, bagging algorithm produced highest accuracy values when updating the datasets in the training stage. Enhancement in performance for the bagging algorithm was by 11% when using the test data set with the submitted attributes from the user. The diabetes dataset was used in the same way, and bagging algorithm also proved to be superior in producing the highest accuracy values among all other machine learning algorithms. The chart in figure 10 plots these results in both cases where the prediction is done using only the training dataset, and when prediction is done with the testing subset of data.

Superiority of bagging algorithm in prediction is also depicted in this dataset, with an enhancement in performance after using the test data set in diagnosis.

7. CONCLUSION

In this research; a description of the use of bagging algorithm for medical diagnosis was introduced. This algorithm is proving over and over its superiority in providing the most accurate medical diagnosis to prominent diseases that have a variety of symptoms. The system that was named B-Pred. was designed and implemented, and then tested using freely available, yet reliable, medical datasets from a credible internet resource, which was verified by physicians. A strong point about this system is that when changing the datasets, the system can produce diagnosis to other diseases with the same proficiency, which can be used later to predict other related health issues when learning is advanced, and more test cases are used. The main idea behind B-Pred. was to separate the dataset into two parts; where the bigger part was used in the training process, in which bagging algorithm was used to classify the data, and the remaining part was used in the diagnosis production process, also using bagging, where this part of the dataset was combined with the data submitted by the user to validate the produced diagnosis, and provide the user with the accurate diagnosis. Comparing B-Pred. with other known machine learning algorithms (namely: Neural Networks, Decision Trees, and Naive Bayes Classifier) showed that bagging algorithm produced way far more accurate diagnosis than these algorithms. The adaptability of the system allows it to incorporate various datasets to make it produce diagnosis for almost any type of disease.

<table>
<thead>
<tr>
<th>Metric</th>
<th>BG</th>
<th>DT</th>
<th>NB</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.95</td>
<td>0.90</td>
<td>0.85</td>
<td>0.63</td>
</tr>
<tr>
<td>Precision</td>
<td>0.96</td>
<td>0.91</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>Recall</td>
<td>0.94</td>
<td>0.90</td>
<td>0.84</td>
<td>0.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>BG</th>
<th>DT</th>
<th>NB</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.86</td>
<td>0.78</td>
<td>0.83</td>
<td>0.70</td>
</tr>
<tr>
<td>Precision</td>
<td>0.85</td>
<td>0.75</td>
<td>0.87</td>
<td>0.61</td>
</tr>
<tr>
<td>Recall</td>
<td>0.85</td>
<td>0.78</td>
<td>0.74</td>
<td>0.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>BG</th>
<th>DT</th>
<th>NB</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.97</td>
<td>0.90</td>
<td>0.84</td>
<td>0.65</td>
</tr>
<tr>
<td>Precision</td>
<td>0.95</td>
<td>0.92</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td>Recall</td>
<td>1.00</td>
<td>0.88</td>
<td>0.85</td>
<td>0.36</td>
</tr>
</tbody>
</table>
REFERENCES


