Upgrading the Performance of Machine Learning Based Chronic Disease Prediction Systems using Stacked Generalization Technique

Ekta Maini¹, Bondu Venkateswarlu², Dheeraj Marwaha³ and Baljeet Maini⁴

¹Department of Computer Science, Dayananda Sagar University, Bengaluru, India
²Department of Computer Science, Dayananda Sagar University, Bengaluru, India
³Senior Technical Architect, Leoforce LLC., Hyderabad, India
⁴Baljeet Maini, Professor, Teerthanker Mahaveer University, Moradabad, India

E-mail address: ekta.marwaha@gmail.com², bonduvenkat-cse@dsu.edu.in², dheeraj.marwaha@leoforce.com⁴,
mainibaljeet@gmail.com⁴

Abstract: During the past few years, mortality due to chronic diseases has increased manifolds globally. Low- and middle-income group countries are worst affected as the healthcare facilities are neither affordable nor accessible easily. The challenges of unaffordable and inaccessible healthcare can be faced gracefully by using machine learning based prediction models. These techniques are used to learn patterns from the medical datasets and build decision support systems for diagnosis of diseases in early stages and hence prevent high mortality. This research work is aimed at upgradation of the performance of chronic disease prediction models using Stacked Generalization approach. In this work, stacked generalization ensembling approach has been applied over five base classifiers namely Logistic Regression (LR), K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Naïve Bayes (NB) and Decision Tree (DT) with 10-fold cross validation. Experimental results highlight the effectiveness of Stacked Generalization method in enhancing accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of prediction system by reducing variance error to avoid overfitting. The prediction models developed in this study can efficiently be used in primary health care centres to diagnose five chronic diseases namely cardiovascular diseases, diabetes, breast cancer, hepatitis, and chronic kidney disease.

Keywords: Cardiovascular diseases, breast cancer, diabetes, hepatitis, chronic kidney disease, stacked generalization, machine learning

1. INTRODUCTION

An illness which continues for a long duration is termed as chronic disease. Some chronic illnesses generally found amongst most of the world population in recent years include cardiovascular diseases, diabetes, cancer, arthritis, asthma, hepatitis, acquired immunodeficiency syndrome, and kidney diseases. Cardiovascular diseases (CVDs) are the greatest reason of mortality globally [1]. CVDs are associated with the conditions of heart like stroke, coronary heart disease, cardiomyopathy, valvular cardiac disorder, peripheral artery diseases etc. According to World Health Organization, CVDs claim 3 million lives annually globally. A variety of factors increase the risk of heart disease like high body mass index, age, hypertension, diabetes, stress, and sedentary lifestyle. Early diagnosis of these diseases shall enable a person to make modifications in lifestyle habits and hence may reduce the mortality rate. In addition to CVDs, past few decades have witnessed significant increase in cancer cases. The most common cancer among women worldwide is breast cancer [2]. Mutation and uncontrolled reproduction of cells of mammary tissue results formation of tumors commonly referred to as breast cancer. The malignant tumors spread to other parts of the body via blood or lymph. If the cancer is not detected in time, it may lead to death. Mortality associated with breast cancer is reported to be around two million annually. It is reported that high mortality rate is accountable to the fact that early signs and symptoms of cancer are usually ignored by the patients and this disease is usually diagnosed in advanced stages resulting in less chances of recovery. Diabetes, hepatitis, and chronic kidney diseases have also increased substantially worldwide in the recent years. Elevated amounts of blood glucose levels for a prolonged time due to metabolic
disorders results in diabetes. Many complications can occur if diabetes is not treated in time. Diabetes is proven to be the most important reason of diabetic retinopathy, kidney malfunctioning, CVDs, and lower limb amputation. Diabetic patients were found to be at higher risk of COVID -19 as compared to non-diabetics. High mortality is associated with diabetes (around 3 million) in many countries like India where healthcare facilities are neither affordable nor accessible easily. Early diagnosis of diabetes is essential in these countries [3]. Inflammation of the liver results in hepatitis. Though some patients are asymptomatic, others may tend to develop some symptoms like yellowing of the skin, loss of appetite, fatigue, pain in abdomen and diarrhoea. Hepatitis becomes chronic if it lasts more than six months [4]. If left untreated, hepatitis may lead to liver failure and in worst case liver cancer. Chronic kidney disease occurs due to weakening of renal organs because of various factors like abnormal excretion of albumin or improper functioning of kidneys. This disease is diagnosed using glomerular filtration rate which indicates how effectively wastes are being filtered from the blood by the kidneys [5].

An enormous burden on healthcare sector accounts to high cost of treatment of these chronic diseases. The situation is pathetically serious in low- and middle-income group countries where the patients ignore the early signs and symptoms of the diseases due to lack of affordable and easily accessible healthcare facilities [6]. This results in diagnosis of these chronic diseases in advanced stages, subsequently leading to high costs of treatment and less chances of recovery eventually leading to mortality. Since most of the risk factors of these diseases are related to lifestyle, early diagnosis of diseases can help in prevention of these diseases. Machine learning techniques are being explored by the researchers to build disease prediction systems which can act as cost effective technological aid to detect diseases in early stages [7]. Machine learning algorithms have an immense has tremendous potential to explore, visualize and learn patterns from structured and unstructured data. The algorithms can generate meaningful insights from the healthcare care and help to build prediction system to discover the threat of illness in a patient.

Conventional machine learning techniques like Naïve Bayes, logistic regression, support vector machine etc. have been used by researchers to build disease prediction models. Available literature signifies the need of improvement in the performance of such models. Very limited work has been done to improve the performance of prediction models. In this research work, stacked generalization approach has been implemented on five base classifiers to develop efficient prediction systems to diagnose five chronic diseases namely cardiovascular diseases, diabetes, hepatitis, cancer, and chronic kidney diseases in early stages. The base machine learning classifiers used in this work are logistic regression, naïve Bayes, k NN, support vector machine and decision tree. Accuracy, sensitivity, specificity along with positive predictive value and negative predictive value have been evaluated for all models. It was observed that Stacked Generalization resulted in significant improvement of performance. Leveraging these high-performance ensemble-based machine learning models in primary healthcare centers can provide cost effective solution to detect chronic diseases in early stages and hence save millions of lives.

The remaining paper is structure as follows: Work correlated to this subject has been discussed in Section 2. The description of the datasets used in the study has been presented in section 3. Section 4 discusses the methodology of the proposed work. Outcomes obtained in the study have been discussed in Section 5 followed by the conclusion.

2. RELATED WORK

During the last few years, researchers have shown keen interest in leveraging supervised learning algorithms for building disease prediction systems. Conventional procedures like Logistic regression, support vector machine, k nearest neighbors and naïve Bayes have been extensively employed by scientists to create prediction models for various diseases. Research work carried out by Maini E et al attained an accuracy of 86.4% in predicting heart diseases [8]. Majority voting rule used in the study carried out by Khalid Raza improved the performance of heart disease prediction considerably to 88.88% [9]. Research work carried out by S. Vijayarani et al signify that Naïve Bayes and radial basis were efficient in early diagnosis of chronic kidney diseases [10]. An accuracy of 76.3% was observed in this study. Accuracy of predicting chronic kidney disease was increased significantly to 95% in the work done by Olayinka Ayodele Jongbo et al using ensemble techniques namely random subspace and bagging [11]. For the early diagnosis of diabetes, Quan Zou et al performed feature selection using principal component analysis followed by neural network to attain an accuracy of 77% [12]. A pilot study carried out by Tao Zheng et al to diagnose diabetes from the electronic health record of population of Shanghai by feature extraction and machine learning algorithms using 4-fold cross validation achieved an accuracy of 88% [13]. Research work carried out by Alehegn et al achieved appreciable good results in diagnosing diabetes on UCI Pima diabetes dataset [14]. This work highlights the importance of hyperparameter tuning of machine learning algorithms to achieve optimum performance. Usage of machine learning techniques like Bayesian networks and CART for diagnosis of diseases has been discussed in [15]. Work carried out by Maini E et al discusses the
significance of feature selection in disease prediction system [16].
Two main knowledge gaps have been identified. Firstly,
all the research works carried out so far are focused on
one disease and no research work has been carried out to
build prediction systems for multiple diseases. Secondly,
all the above-mentioned research works highlight the
need to carry out further work to improve the
performance. This paper aims at building prediction
models for multiple chronic diseases and enhancing the
performance using Stacked Generalization technique.

3. CHRONIC DISEASE DATASETS DESCRIPTION

This research work has been carried out on five
chronic diseases namely cardiovascular diseases, diabetes,
hepatitis, cancer, and chronic kidney diseases. The
medical datasets have been collected from UCI repository
for carrying out this work. A comprehensive description
of these datasets is as follows:

i) Cardiovascular Diseases: Cleveland heart disease
dataset was worked upon in this investigation study. It is
aimed to determine if a patient has heart disease or not
[17]. This dataset has 13 input attributes, based on which
occurrence of heart disease is predicted. The dataset has
303 records.

ii) Diabetes: UCI Pima Diabetes Dataset Repository
has been worked upon in this work [18]. This dataset has
8 input independent medical attributes which are used to
predict if a patient is diabetic or not. There are 768
records in this dataset.

iii) Hepatitis: UCI hepatitis dataset has been used to
build prediction model for hepatitis [19]. This dataset has
155 records. A total of 19 input attributes are available to
diagnose the presence or absence of hepatitis.

iv) Breast Cancer: Work has been carried out on
Wisconsin breast cancer dataset available from the UCI
repository [20]. The dataset has 699 records. 9 input
medical attributes are used to identify the presence/abscence of breast cancer.

v) Chronic Kidney Disease: Chronic kidney disease
dataset collected from UCI has been studied [21]. This
dataset has 400 records. 24 input attributes are used to
check if a patient is infected.

Highlights of these datasets are provided in Table 1

<table>
<thead>
<tr>
<th>Disease</th>
<th>Dataset</th>
<th>Input features</th>
<th>No. of records</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart disease</td>
<td>UCI Cleveland heart disease</td>
<td>13</td>
<td>303</td>
<td>Cleveland, USA</td>
</tr>
<tr>
<td>Diabetes</td>
<td>PIMA dataset</td>
<td>8</td>
<td>768</td>
<td>USA</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>UCI hepatitis dataset</td>
<td>19</td>
<td>155</td>
<td>Yugoslavia</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>Wisconsin breast cancer dataset</td>
<td>9</td>
<td>699</td>
<td>Wisconsin, USA</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>Chronic kidney dataset</td>
<td>24</td>
<td>400</td>
<td>India</td>
</tr>
</tbody>
</table>

4. METHODOLOGY

This section describes the various steps followed in
the research work. There is a sequence of steps to be
followed while building a machine learning model. Fig.1
illustrates the complete methodology of the proposed
work. A summary of pre-processing methods, various
base classifiers, Stacked Generalization method and
evaluation criteria has been presented below:

A. Data pre-processing

Data is pre-processed before building the prediction
model [22]. As shown in Fig.1, pre-processing involves
three main tasks: handling missing values, data
transformation and feature selection. The details are
provided below:

a) Handling missing values: There are many options
to cope with the issues of missing values. Some options
are dropping the missing values, dropping a complete
row, dropping a complete column, or filling the missing
values with mean, median or mode. The datasets of
chronic diseases gathered from UCI repository had
missing values. In this work, the missing values were
replaced with the median values.
b) Data Transformation: The attributes of the chronic disease datasets have different ranges. For example, the age of patients ranges from a minimum of 20 years to a maximum of 90 years, while the total cholesterol levels vary from 134mg/dL to 330 mg/Dl. Normalization is an important data transformation technique where data values are scaled to a similar range. In this project work, z-score normalization was carried out to scale the data.

c) Feature Selection: Presence of redundant or irrelevant features tends to reduce the performance of the prediction systems. Feature selection was carried out identify the most important input attributes for the chronic diseases. Recursive feature elimination was implemented in this work.

B. k-fold cross validation

For building each of the chronic disease prediction system ten - fold cross validation method was applied. Each dataset was segregated randomly into 10 equal parts. Training of the models was carried out on nine subsets while remaining one subset was left out for validation purpose. The process was repeated ten times for each dataset. Each of ten subgroups was applied precisely one time to assess the performance.

C. Classification using the base classifiers

After the pre-processing of data, conventional classifiers were applied on the training data to develop prediction model capable of identifying the risk of disease in a patient. Basically, development of a disease prediction model is a classification task. Based on the medial attributes of a person, the prediction model should identify if the disease exists or not. The scheme is represented in Fig. 2. If the output of the system is 1, it indicates that the patient has disease but if the output is 0, it signifies a healthy person who does not have disease.

Various machine learning algorithms have proved their excellence in understanding hidden patterns in the data and have been used in this study to diagnose five chronic diseases. Some of the well-established machine learning techniques used in this study are:

i) Logistic Regression: It is one of the most robust classifiers used for classification in machine learning. Presence or absence of disease is determined by calculating the probabilities (p) after applying logistic function. Mathematically, it can be written as:

$$\log\text{odds} = \log\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1x_1 + \alpha_2x_2 + ... + \alpha_nx_n \tag{1}$$

where, $x_1$, $x_2$,...,$x_n$ are the various input attributes and $\alpha_i$ are the model parameters.

A cut off value is chosen. Probabilities greater than the cut off value generate 1 as output while the probability less than the cut off values results in 0. Logistic regression provides deep insights of the data. It enables a clear understanding of the relationship between an output attribute and input attributes. This method enables us to get a clear idea about the prominent risk factors of a disease.

ii) Support Vector machine: The classification decision boundary is linear in case of support vector machine. The training dataset utilized in our research work comprise of data records belonging to two classes: positive cases of diseases and negative cases of diseases. The objective is to detect existence or non-existence of illness in a new dataset. For a n-dimensional datapoint, n-1 hyperplanes exist which can classify the data. A hyperplane which separates the datapoints with maximum margin is selected. Generalization error of support vector machine is low if the margin of separation is high. The concept of support vector machine is explained in Fig. 3. Evidently, hyperplane H1 separates the datapoints belonging to two classes better than hyperplane H2.
iii) Naïve Bayes: This is a kind of probabilistic classifier. The algorithm holds its roots in Bayes theorem with an assumption that there is no dependence amongst the individual attributes. All features contribute equally and independently in the prediction. The mathematically equation can be written as:

\[ p(m|N) = \frac{p(N|y)p(m)}{p(N)} \]  

Where, \( m \) is the output class variable (Disease present or absent) and \( N \) represents \( n \) input medical attributes of a person namely \( n_1, n_2, \ldots, n_n \). Equation 2 can be rewritten as:

\[ p(m|n_1 \ldots n_n) = \frac{p(m)\prod_{i=1}^{n} p(n_i|m)}{p(n_1)p(n_2)p(n_3)\ldots p(n_n)} \]  

Since the denominator in equation 3 is constant, we can infer that,

\[ p(m|n_1 \ldots n_n) \propto p(m)\prod_{i=1}^{n} p(n_i|m) \]  

In other words, for all possible values of output \( y \), the probability of given set of inputs is calculated. The output with maximum probability is finally picked up. The result can be summarized as:

\[ m = \arg\max_{m} p(m)\prod_{i=1}^{n} p(n_i|m) \]  

iv) K- nearest neighbor: k-NN belongs to the family of non-parametric methods used for classification. K-NN is a lazy classifier. To decide the class of a test data point, labels of \( k \) (1,3,5, etc.) nearest neighbors of the data point are identified. Highly popular class of k- nearest neighbors is allocated to the test data point. K-NN has been conceptualized in Fig. 4. The class of test object shall be B if \( k=3 \). If 12 nearest neighbors are considered, then the class of test object shall be A.

v) Decision Tree: This classifier resembles a graph similar in construction to a tree. Classification is done by sorting the tree from root to the leaf node. Selection of the attributes is carried out based on conditions like entropy and information gain for further splitting. Depth of the tree should be chosen wisely to avoid overfitting.

D. Stacked Generalization Ensemble Technique

The most promising way to enhance accuracy, specificity and predictive values of a classifier is ensembling. In this technique, various weak base learners are combined to generate a strong learner with enhanced performance. The bias and variance errors of base learners are greatly reduced by ensembling which results in better performance. Various types of ensemble methods like bagging and boosting are built using homogeneous base learners and have been employed by researchers for building high performance systems. In this paper we use a special type of ensemble method called stacked generalization. Stacked generalization is also referred to as stacking. It is an ensembling methods which combines heterogeneous base classifiers.

Stacking is distinct from its other ensembling counterparts like bagging and boosting. In bagging, the base learners are typically same i.e. decision trees, and these are fit on samples of training dataset. However, in stacked generalization technique, the base learners are different i.e. logistic regression, k-NN, SVM etc. which are fit on same dataset. In contrast to boosting, in stacked generalization, a solo model is utilized to discover how to best blend the forecasts from the contributing classifiers rather than a series of prototypes which alter the predictions of previous classifiers.

http://journals.uob.edu.bh
Learning is carried out in parallel and a meta classifier is used to combine the predictions made by weak base learners. Two essential stages of stacked generalization are:

i) Level -0 learning: Predictions are made using base classifiers.

ii) Level-1 learning: A metamodel is learnt to combine the predictions of base learners.

Complete scheme of stacked generalization is shown in Fig. 5.

In this study, five classifiers discussed earlier namely logistic regression, support vector machine, k-NN, naïve Bayes and decision trees have been used as base learners. Logistic regression has been used as meta-model.

### E. Performance Criteria

In this research work, performance of the disease prediction was assessed using five metrics. Accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. These values are obtained from confusion matrix as shown in Table 3. A brief description of these performance metrics is provided below:

#### TABLE 3. CONFUSION MATRIX

<table>
<thead>
<tr>
<th>Actual Outcomes</th>
<th>Predicted Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease-No</td>
<td>True Negative (B)</td>
</tr>
<tr>
<td>Disease-Yes</td>
<td>True Positive (A)</td>
</tr>
<tr>
<td>False Positive (C)</td>
<td></td>
</tr>
<tr>
<td>False Negative (D)</td>
<td></td>
</tr>
</tbody>
</table>

i) Accuracy: This score represents the fraction of number of correct predictions out of total predictions made by the system. Mathematically,

\[
\text{Accuracy} = \frac{A+B}{A+B+C+D} \tag{6}
\]

ii) Sensitivity: It implies the degree to which actual positives are not ignored. It measures the percentage of people who are predicted to have disease among those who have the disease. Mathematically,

\[
\text{Sensitivity} = \frac{A}{A+D} \tag{7}
\]

iii) Specificity: It evaluates the ratio of healthy people who are accurately categorized healthy.

\[
\text{Specificity} = \frac{B}{B+C} \tag{8}
\]

iv) Positive Predictive value (PPV): It represents the ratio of true positives to the number of people predicted to have disease. It is also called as precision.

\[
\text{PPV} = \frac{A}{A+C} \tag{9}
\]

v) Negative Predictive value (NPV): It refers to the proportion of true negatives to the number of people predicted to be healthy.

\[
\text{NPV} = \frac{B}{B+D} \tag{10}
\]
5. RESULTS AND DISCUSSIONS

This segment offers us the details of the experimental results obtained in this study. Datasets for five chronic diseases discussed earlier i.e. cardiovascular diseases, diabetes, hepatitis, breast cancer and chronic kidney diseases have been collected from UCI repository. Five state of art classifiers namely logistic regression, naive Bayes, k-NN, support vector machine and decision trees have been used to build the disease prediction models. Later, stacked generalization ensemble technique was applied to enhance the performance of the system. Logistic regression was used as level 1 meta model for stacking. The performance of the models was evaluated in terms of accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Table 4 illustrates the results obtained on Cleveland cardiovascular disease dataset. Sensitivity and PPV achieved by Support vector machine was observed to be the best. It is evident from the table that stacking enhanced the performance of heart disease prediction system considerably. Accuracy of the system increased considerably from 78.7% to 86.9 % using stacked generalization method. There was a significant increase of 7% in sensitivity as compared to logistic regression technique.

<table>
<thead>
<tr>
<th>TABLE 4. PERFORMANCE OF CARDIOVASCULAR PREDICTION SYSTEM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>PPV</td>
</tr>
<tr>
<td>NPV</td>
</tr>
</tbody>
</table>

The performance of diabetes prediction system developed using the beforemenioned techniques is tabulated in Table 5. Though the performance of conventional algorithms was good, yet results obtained using Stacked generalization techniques suggest that there was a modest enhancement in the performance.

The accuracy of diabetes prediction system is as high as 90.9 %. PPV and NPV values are significantly high at 83.4% and 85.8% respectively. Sensitivity and specificity were also recorded to 88.8% and 89.9% respectively.

<table>
<thead>
<tr>
<th>TABLE 5. PERFORMANCE OF DIABETES PREDICTION SYSTEM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
</tbody>
</table>

The performance characteristics of hepatitis, breast cancer and chronic kidney disease prediction systems are shown in Table 6,7 and 8, respectively. It is evident from Table 6 that stacked generalization technique-based hepatitis prediction system performs appreciably well. The accuracy is 90%. Stacked generalization resulted in a prediction system with sensitivity and specificity of 93.2 % and 90.9% which is significantly better than conventional algorithms.

<table>
<thead>
<tr>
<th>TABLE 6. PERFORMANCE OF HEPATITIS PREDICTION SYSTEM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>PPV</td>
</tr>
<tr>
<td>NPV</td>
</tr>
</tbody>
</table>

From Table 7, we can infer that Stacked generalization method increase the accuracy of prediction system as compared to conventional methods. All the performance metrics achieved a value more than 85% indicating the potential use as screening tool for detecting breast cancer in early stages. Accuracy is observed to be 88% while PPV and NPV are noted to be 86.9% and 85.4%respectively.

<table>
<thead>
<tr>
<th>TABLE 7. PERFORMANCE OF BREAST CANCER PREDICTION SYSTEM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>PPV</td>
</tr>
<tr>
<td>NPV</td>
</tr>
</tbody>
</table>

Stacked Generalization performed remarkably well for detection of chronic kidney disease. An accuracy of 88.3 % is achieved which is considerably better than logistic regression, Support vector machine and decision tree. Not only accuracy, Sensitivity and specificity were also improved to a great extent by stacking. PPV and NPV were evaluated as 85.3% and 87.3% respectively.
The experimental results prove that stacked generalization technique significantly enhanced the performance of the disease prediction models built using the traditional machine learning algorithms. These models can be used to detect the diseases in early stages. Use of such systems shall help in reducing mortality. The patients can be screened for various illnesses even in the rural areas where they do not have easy access to medical infrastructure.

6. Conclusion

In this research work, machine learning based prediction systems were developed for five chronic diseases. In addition to conventional techniques, stacked generalization technique was explored. The experimental results clearly indicate that stacked generalization technique is an effective way to enhance the performance of the classifiers. The disease prediction system developed in this study achieved a remarkable performance and can be efficiently used to diagnose the chronic diseases in early stages. These models shall serve as technological aid in reducing the mortality rate owing to chronic diseases. Application of these chronic disease prediction systems in primary healthcare centers shall prove to be a great step towards improving the healthcare facilities. It is intended to extend this work for building prediction models for other diseases as well.

References

Author Ekta Maini completed her B.Tech degree from Kurukshetra University, India in 2002 and ME from Panjab University in 2004. She is a gold medalist in BTech as well as ME. Presently she is pursuing PhD in Dayananda Sagar University. Her interest areas include Deep Learning and Machine Learning. She has participated in many national and international conferences and published more than 15 papers in refereed journals.

Bondu Venkateswarlu received his Ph.D. degree in 2016. He is currently working as an Associate Professor in the Department of Computer Science and Engineering of Dayananda Sagar University, Bangalore, India. His current research interests include Data Mining, Soft Computing Techniques & Software Engineering. He has published many papers in refereed journals.

Dheeraj Marwaha received his MTech degree in Computer Science in 2010 from Manipal University, India. He has a rich industrial experience of 15 years in software companies like Microsoft, Siemens and Philips. Presently is working as Senior Technical Architect in Leoforce, Hyderabad, India.

Baljeet Maini completed MBBS in 1999 and MD(Pediatrics) in 2002 from Delhi, India. He has authored more than 20 papers in various journals and conferences. His interest includes applying technology to make healthcare affordable in India.