International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Oct.-2021)

http://dx.doi.org/10.12785/ijcds/100192

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

Ekta Maini[1], Bondu Venkateswarlu[2], Dheeraj Marwaha[3] and Baljeet Maini[4]

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

Received 30 Jul. 2020, Revised 22 Jul. 2021, Accepted 05 Aug. 2021, Published 28 Oct. 2021

Abstract: During the past few years, there has been a tremendous increase in mortality due to chronic diseases. Mortality is highest in rural areas of low- and middle-income group countries as there is a lack of affordable and easily accessible healthcare facilities. World Health Organization (WHO) emphasizes on early detection of these diseases to reduce the mortality. In this era of artificial intelligence, machine learning techniques offer promising ways to improve healthcare facilities. Machine learning based prediction tools can be used for early detection of diseases to prevent mortality. This research work has been carried out to build high performance screening tools to diagnose five chronic ailments in early stages. Stacked generalization ensembling approach with ten-fold cross validation has been applied over five conventional classifiers. Results obtained in this research work demonstrate that stacked generalization method enhances the performance considerably by reducing variance error. This research work is highly beneficial for primary healthcare centre in rural areas to diagnose diseases in early stages and hence prevent mortality.

Keywords: Chronic Diseases, Affordable Healthcare, Mmachine Learning, Stacked Generalization, Early Diagnosis

1. INTRODUCTION

An illness which continues for a long duration is termed as chronic disease. Chronic diseases like cardiovascular diseases, diabetes, cancer, arthritis, asthma, hepatitis, acquired immunodeficiency syndrome, chronic kidney diseases etc. are spread across the world. This research work is primarily focused on five chronic diseases namely breast cancer, cardiovascular diseases, hepatitis, diabetes and chronic kidney disease. A brief description of these diseases is presented below.

Cardiovascular diseases (CVDs) are morbidity conditions like stroke, coronary heart disease, peripheral artery diseases etc. [1]. Breast cancer among other malignancies is also a major non- communicable morbidity now amongst women [2]. Surveys carried out by WHO reveal that around 3 million people die annually due to cardiovascular diseases while mortality associated with breast cancer is around 2 million [3]. Other chronic diseases like diabetes, hepatitis, and chronic kidney diseases have also increased substantially in the recent years. Conditions like diabetes need to be diagnosed at the earliest as it may further lead to diseases like diabetic retinopathy, kidney malfunctioning, CVDs, and lower limb amputation [4]. Hepatitis of viral origin (type B, C) is a chronic progressive damage of liver. Chronic kidney diseases are also increasing directly as well as complications of diabetes, raised blood pressure and medicines of various types. [5].

An enormous burden of these ailments in public healthcare sector accounts to high cost of treatment of these chronic diseases. The situation is pathetically serious in low- and middle-income group countries which lack in affordable and easily accessible healthcare facilities [6]. As a result, early presentation in hospitals is at very dismal rates. The diseases are usually diagnosed in advanced stages, causing expensive treatment needs and lesser chances of recovery. Mortality thus is increasing due to delayed presentation in hospitals. Early diagnosis of diseases is an efficient approach to avert these diseases or their complications. 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].

E-mail: ekta.marwaha@gmail.com[1], bonduvenkat-cse@dsu.edu.in[2], dheeraj.marwaha@leoforce.com[3], mainibaljeet@gmail.com[4]

Machine learning algorithms have a tremendous potential to explore, visualize and learn patterns from structured and unstructured data. The algorithms can generate meaningful insights from the healthcare and help to build prediction system to discover the threat of illness in a patient.

Extensive review of literature suggests that algorithms like logistic regression and Naïve Bayes have been used by researchers to build disease prediction models. However, literature also confirms that there is a scope of improvement in the performance of such models. Limited work has been done to improve the performance of prediction models further. This work is an effort to create high performance disease prediction models for five chronic diseases. Stacked generalization ensembling method applied over five traditional classifiers has been employed in the present study. It has been observed that Stacked Generalization resulted in significant improvement of performance of prediction models. 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 paper is designed as follows: Related work conducted by scientists so far has been discussed in Section 2. The description of the datasets used in the study has been presented in section 3. Methodology of the proposed work has been discussed in Section 4. Outcomes obtained in the study have been discussed in Section 5 followed by the conclusion.

2. Related Work

Conventional procedures like Logistic regression, Naïve Bayes and k- nearest neighbor are reported to 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 83.4% 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 85.88% [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 85% 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 83% [13]. Research work carried out by Alehegn et al achieved appreciable good results in diagnosing diabetes on Pima diabetes dataset available online by University of California, Irvine (UCI) [14]. The importance of hyperparameter tuning for optimum performance has been discussed. Usage of machine learning techniques like Bayesian networks 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 each 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 research work is an attempt to address both these issues. In this paper, stacked generalization technique has been applied to enhance the performance of prediction models for multiple chronic diseases.

3. Chronic Disease Datasets Description

This research work has been carried out on five chronic diseases. The medical datasets for these diseases 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 [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 key medical parameters which are used to diagnose diabetes in a patient. 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 hepatitis.

iv) Breast Cancer: Wisconsin breast cancer dataset available from the UCI repository has been used [20]. The dataset has 699 records. 9 input medical attributes are used to detect breast cancer.

v) Chronic Kidney Disease: UCI Chronic kidney disease dataset 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 I.

TABLE I DESCRIPTION OF DATASETS

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

4. METHODOLOGY

There is a series of steps to be pursued for developing 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) Handing missing values: Missing data values may result in errors. The missing values should be handled carefully. Either these values may be dropped or may be replaced with their mean/median values. Median of data was computed to replace missing values.

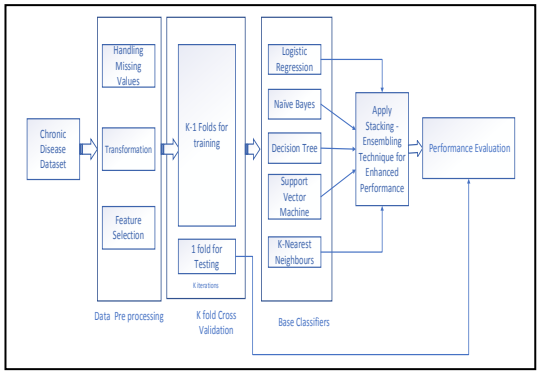

Fig. 1 Workflow Diagram of Proposed Work

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 procedure was reiterated 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 medical 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.

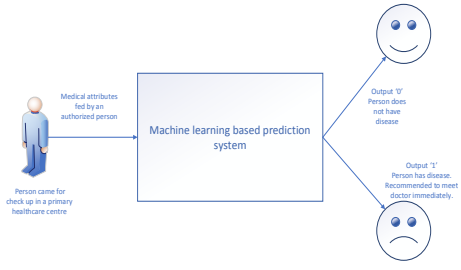

Fig. 2 Machine learning based chronic disease prediction system

Many machine learning procedures have demonstrated their excellence in identifying unseen patterns in the data. Some of the well-established machine learning techniques used in this study are:

i) Logistic Regression: It is a commonly used robust classifier 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:

$$log - odds = \log_e(\frac{p}{1-p}) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots \alpha_n x_n \quad (1)$$

where, x1, x2…x_n are the various input attributes and α_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.

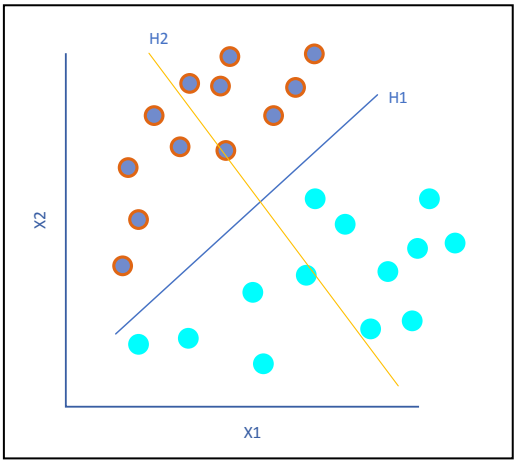

Fig. 3 Concept of Hyperplanes in Support Vector Machine

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:

$$p(m|N) = \frac{p(N|y)p(m)}{p(N)} \quad (2)$$

Where, m is the output class variable (Disease present or absent) and N represents n input medical attributes of a person namely n1, n2…. n_n. Equation 2 can be rewritten as:

$$p(m|n_1 \cdots n_n) = \frac{p(m)\prod_{i=1}^{n} p(n_i|m)}{p(n)p(n_2)p(n)\cdots p(n)} \quad (3)$$

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

$$p(m|n_1 \cdots n_n) \propto p(m)\prod_{i=1}^{n} p(n_i|m) \quad (4)$$

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:

$$m = argmax_m\, p(m) \prod_{i=1}^{n} p(n_i|m) \quad (5)$$

iv) K- nearest neighbor: k-NN is a non-parametric lazy classifier. To determine the category of a 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

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 resulting 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.[ref] 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 method 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.

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

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

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

Complete scheme of stacked generalization is shown in Fig. 4.

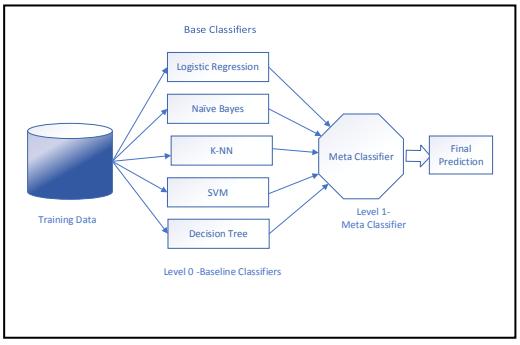

Fig. 4 Phases of Stacked Generalization Ensemble Technique

Algorithm for stacked generalization technique is provided in Table II.

TABLE II. STACKED GENERALIZATION TECHNIQUE

Let M = [m_1, m_2, m_3, … m_n] be the given dataset

B = {b_1, b_2, b_3, …b_n}, the set of base learners

Xtrain = training set

Xtest = test set

K = meta model classifier

Len = n(M)

for i =1 to Len do

S(i) = Model trained using B(i) on Xtrain

Next i

S= S UK

Output = Xtest classified by S

In this study, five classifiers discussed earlier have been used as base learners. Logistic regression has been used as meta-model.

E. Performance Criteria

Performance of the disease prediction was assessed using five metrics. Confusion matrix used to evaluate the performance has been discussed in Table III. A brief description of these performance metrics is provided below:

TABLE III. CONFUSION MATRIX

		Predicted Outcome	
		Person predicted to have disease	Person predicted to be healthy
Actual Outcomes	Person has disease	True Positive (A)	False Negative(D)
	Person is healthy	False Positive (C)	True Negative (B)

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

$$Accuracy = \frac{A+B}{A+B+C+D} \quad (6)$$

ii) Sensitivity: It measures the percentage of people who are correctly predicted to have disease among all those who have the disease. Mathematically,

$$Sensitivity = \frac{A}{A+D} \quad (7)$$

iii) Specificity: It is the proportion of people who are accurately forecasted to be healthy among those who are healthy.

$$Specificity = \frac{B}{B+C} \quad (8)$$

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

$$PPV = \frac{A}{A+C} \quad (9)$$

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

$$NPV = \frac{B}{B+D} \quad (10)$$

5. Results And Discussions

Details of the experimental results of this study have been provided in this section. Details of the datasets used in this study have been provided earlier in the paper. Brief description of base classifiers and stacked generalization method have also been discussed already. Logistic regression was used as level 1 meta model for stacking. Table IV illustrates the results obtained on Cleveland cardiovascular disease dataset. Sensitivity and PPV achieved by Support vector machine was observed to be the best. An accuracy of 86.9% with a PPV and NPV of

87.4% and 89.8% respectively has been achieved using Stacked Generalization technique. These results are better than the results reported in the literature so far [9]. This significant increase in performance has been depicted in Fig 5. It is evident from the table that stacking enhanced the performance of cardiovascular disease prediction system considerably.

TABLE IV. PERFORMANCE OF CARDIOVASCULAR PREDICTION SYSTEM

Performance Metrics	Conventional Algorithms					Ensemble
	LR	k-NN	SVM	NB	DT	Stacking
Accuracy	78.7%	80.8%	84.8%	80.8%	80.8%	86.9%
Sensitivity	80.3%	81.9%	86.6%	89.3%	81.8%	87.8%
Specificity	85.1%	86.3%	89.2%	81.3%	82.8%	82.9%
PPV	82.1%	81.8%	78.2%	89.2%	85.6%	87.4%
NPV	80.0%	84.8%	82.1%	82.1%	86.2%	89.8%

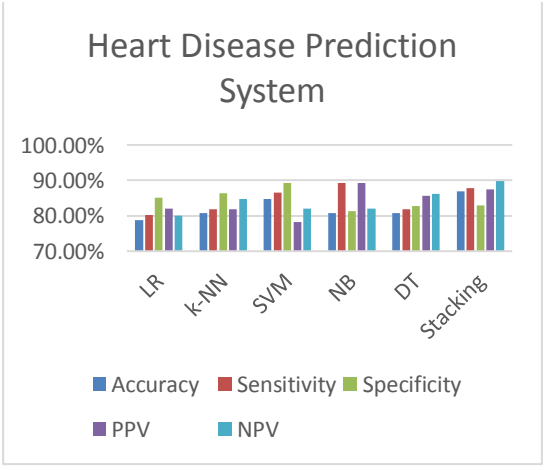

Fig. 5 Performance of Heart Disease Prediction System

The performance of diabetes prediction system developed using the beforementioned techniques is tabulated in Table V. Sensitivity of 86.6% is achieved using Naïve Bayes classifier which is significantly improved to 88.8% using Stacked generalization technique. Results obtained using Stacked generalization techniques suggest that there was a modest enhancement in the performance. Staking technique has surpassed the scores attained earlier by the researchers as reported in the literature. Results using stacking technique are graphically represented in Fig.6

TABLE V. PERFORMANCE OF DIABETES PREDICTION SYSTEM

Performance Metrics	Conventional Algorithms					Ensemble
	LR	k-NN	SVM	NB	DT	Stacking
Accuracy	89.7%	81.3%	78.8%	84.8%	83.8%	90.9%
Sensitivity	83.4%	80.5%	79.6%	86.6%	84.6%	88.8%
Specificity	88.6%	84.2%	80.2%	81.4%	80.4%	84.9%
PPV	85.4%	87.8%	87.3%	84.2%	82.7%	89.4%
NPV	78.4%	78.3%	85.3%	83.1%	86.2%	88.8%

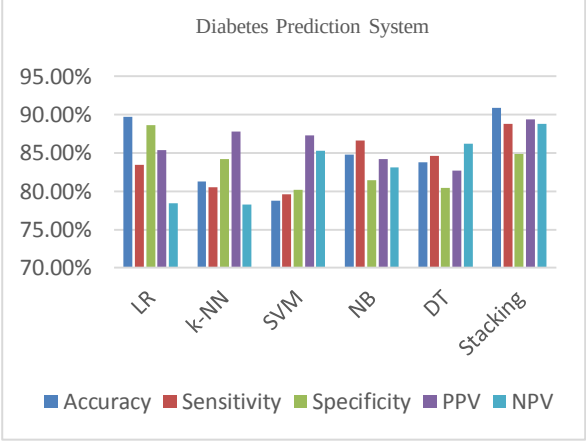

Fig. 6 Performance of Diabetes Prediction System

The performance characteristics of hepatitis prediction systems are shown in Table VI. It is obvious from these results that hepatitis prediction system attained an accuracy of 90% and sensitivity of 93.2% by stacking technique. Results of hepatitis prediction system are graphically shown in Fig.7. These results are superior to the results as reported in the literature till now.

TABLE VI. PERFORMANCE OF HEPATITIS PREDICTION SYSTEM

Performance Metrics	Conventional Algorithms					Ensemble
	LR	k-NN	SVM	NB	DT	Stacking
Accuracy	80.7%	85.3%	88.4%	80.8%	88.8%	90.0%
Sensitivity	82.3%	88.5%	89.6%	81.3%	84.6%	93.2%
Specificity	81.0%	87.2%	86.2%	80.2%	89.4%	90.9%
PPV	83.6%	87.8%	88.3%	82.2%	80.7%	89.4%
NPV	80.5%	87.3%	84.4%	80.1%	83.2%	88.8%

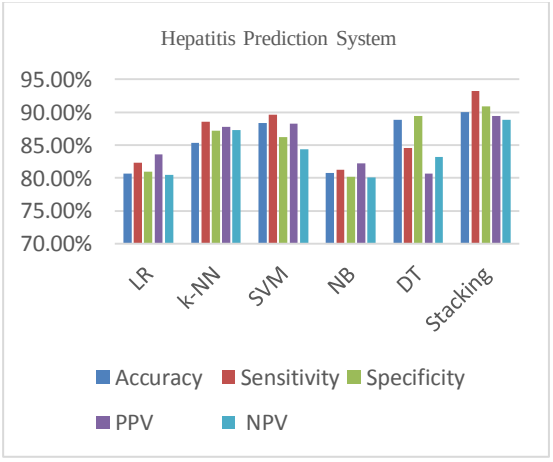

Fig.7 Performance of Hepatitis Prediction System

Using the stacked generalization technique, an accuracy of 88% was attained for breast cancer prediction system Specificity and sensitivity are 84.4% and 80.3% respectively. It is clear from Table VII and Fig.8 that stacked generalization upgrades the performance of the prediction system. Such high-performance results have not been published in technical papers till now. These results signify the importance of stacking techniques in upgrading the performance of prediction system.

TABLE VII. PERFORMANCE OF BREAST CANCER PREDICTION SYSTEM

Performance Metrics	Conventional Algorithms					Ensemble
	LR	k-NN	SVM	NB	DT	Stacking
Accuracy	85.4%	85.8%	84.7%	84.3%	81.7%	88%
Sensitivity	86.6%	80.6%	81.3%	86.5%	80.3%	87.3%
Specificity	83.2%	82.4%	87.0%	82.2%	84.4%	88.9%
PPV	85.3%	80.7%	83.6%	84.8%	83.6%	86.9%
NPV	82.4%	81.2%	80.5%	83.3%	80.5%	85.4%

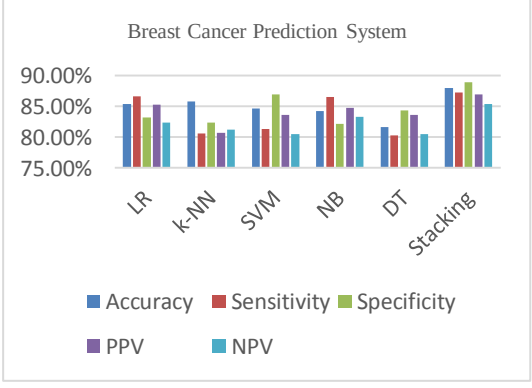

Fig.8 Performance of Breast Cancer Prediction System

Performance analysis of Chronic kidney disease prediction system are represented in Table VIII. It is evident from the results that chronic kidney disease prediction attained an accuracy of 88.3%. Specificity and sensitivity were observed to 89.2% and 87.4%. These results have been graphically represented in Fig.9

TABLE VIII. PERFORMANCE OF CHRONIC KIDNEY DISEASE PREDICTION SYSTEM

Performance Metrics	Conventional Algorithms					Ensemble
	LR	k-NN	SVM	NB	DT	Stacking
Accuracy	84.3%	86.4%	81.3%	83.7%	83.7%	88.3%
Sensitivity	81.5%	85.6%	84.5%	82.3%	82.3%	87.4%
Specificity	82.2%	84.5%	80.2%	88.0%	86.0%	89.2%
PPV	83.8%	83.4%	80.8%	83.6%	82.6%	85.3%
NPV	86.7%	80.5%	82.3%	80.5%	80.5%	87.3%

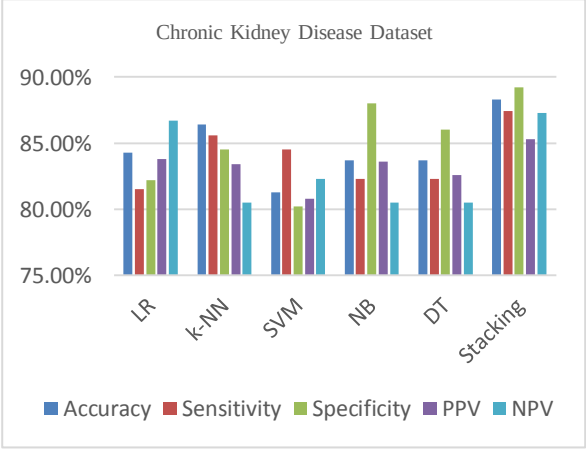

Fig.9 Performance of Chronic Kidney Disease Prediction System

The experimental results prove that stacked generalization technique significantly enhanced the performance of the disease prediction models built using the traditional machine learning algorithms. This work unambiguously demonstrates the impact of stacked generalization technique in upgrading the performance of chronic disease prediction systems.

6. CONCLUSION

Stacked generalization based high performance prediction models for five chronic diseases have been developed in this research work. The experimental results clearly validate the potential of stacked generalization technique in enhancing the performance of the conventional classifiers. Stacked generalization technique-based disease prediction systems developed in this study perform substantially well and can be used in

primary healthcare centers of rural areas for early diagnosis of diseases. These models shall aid in prevention of diseases and reduce mortality. These models provide effective technological aid in realizing two main goals of Health 4.0-namely affordability and easy accessibility. Authors intend to extend this work for building prediction models for other diseases as well.

REFERENCES

[1] Noncommunicable diseases country profiles 2018 [Internet]. World Health Organization. 2019 [cited 17 December 2019]. Available from: https://www.who.int/nmh/publications/ncd-profiles-2018/en/

[2] Breast cancer. (2020). Retrieved 27 June 2020, from https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[3] Diabetes. (2020). Retrieved 27 June 2020, from https://www.who.int/news-room/fact-sheets/detail/diabetes

[4] Global hepatitis report, 2017. (2020). Retrieved 27 June 2020, from https://www.who.int/hepatitis/publications/global-hepatitis-report2017/en/

[5] Bikbov B, Perico N, Remuzzi G (23 May 2018). "Disparities in Chronic Kidney Disease Prevalence among Males and Females in 195 Countries: Analysis of the Global Burden of Disease 2016 Study". Nephron. 139 (4): 313–318. doi:10.1159/000489897. PMID 29791905.

[6] Ekta Maini, Bondu Venkateswarlu. Artificial Intelligence-Futuristic Pediatric Healthcare. Indian Pediatrics 2019 ;56: 796

[7] Maini E, Venkateswarlu B, Gupta A. Applying Machine Learning Algorithms to Develop a Universal Cardiovascular Disease Prediction System. International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. 2018;:627-632.

[8] Maini E, Venkateswarlu B, Gupta A. Determination of Significant Features for Building an Efficient Heart Disease. Prediction System.IJRTE.2019;8(2):4500-6

[9] Raza, K. (2019). Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. U-Healthcare Monitoring Systems, 179–196. doi:10.1016/b978-0-12-815370-3.00008-6

[10] S, V., & S, D. (2015). Data Mining Classification Algorithms for Kidney Disease Prediction. International Journal on Cybernetics & Informatics, 4(4), 13–25. doi:10.5121/ijci.2015.4402

[11] Jongbo, O. A., Adetunmbi, A. O., Ogunrinde, R. B., & Badeji-Ajisafe, B. (2020). Development of an Ensemble Approach to Chronic Kidney Disease Diagnosis. Scientific African, e00456. doi:10.1016/j.sciaf.2020.e00456

[12] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in Genetics, 9. doi:10.3389/fgene.2018.00515

[13] Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., … Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International Journal of Medical Informatics, 97, 120–127. doi:10.1016/j.ijmedinf.2016.09.014

[14] Alehegn, Minyechil & Joshi, Rahul & Mulay, Preeti. (2018). Analysis and prediction of diabetes mellitus using machine learning algorithm. International Journal of Pure and Applied Mathematics. 118. 871-878.

[15] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

[16] Maini, E. (2019). Role of Feature Selection in Building High Performance Heart Disease Prediction Systems. Adbu Journal of Engineering Technology, Retrieved from http://journals.dbuniversity.ac.in/ojs/index.php/AJET/article/view/687/pdf_95

[17] UCI Machine Learning Repository: Heart Disease Data Set. (2020). Retrieved 27 June 2020, from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[18] UCI Machine Learning Repository: Diabetes Data Set. (2020). Retrieved 27 June 2020, from https://archive.ics.uci.edu/ml/datasets/Diabetes

[19] UCI Machine Learning Repository: Hepatitis Data Set. (2020). Retrieved 27 June 2020, from https://archive.ics.uci.edu/ml/datasets/Hepatitis

[20] UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. (2020). Retrieved 27 June 2020, from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[21] UCI Machine Learning Repository: Chronic_Kidney_Disease Data Set. (2020). Retrieved 27 June 2020, from http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

[22] Maini, E. (2019). Implementing an End to End Solution for Data Science Project Cycle-A Complete Roadmap for Data Aspirants. International Journal Of Modern Electronics And Communication Engineering (IJMECE), 7(3). Retrieved from http://www.ijmece.org/viewissues.php?month=5&year=2019

Author Ekta Maini received her B.Tech degree from Kurukshetra University, India in 2002 and ME from Panjab University in 2004. She is gold medalist is 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 conference and published more than 15 papers in refereed journals.

Bondu Venkateswarlu was awarded Ph.D. degree in 2016. He is presently working as an Associate Professor in the Department of Computer Science and Engineering of Dayananda Sagar University, Bengaluru. India. His research interests include Data Mining, Soft Computing Techniques & Software Engineering.

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 (Paediatrics) 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.