INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 5)
Available online at: https://www.ijariit.com

A disease prediction system based on Machine Learning algorithms and ensembles using real-time data gathered through wearable medical sensors

Sameer Kulkarni
ksameerpm@gmail.com
Independent Researcher

Gargi Hartalkar
gargihartalkar01@gmail.com
Independent Researcher

Bhumika Mahajan
mahajanbhumi11@gmail.com
Independent Researcher

Pranav Jawaji
jawajipranav@gmail.com
Independent Researcher

ABSTRACT

Even with an Annual expenditure $8.2 billion (F.Y. 2018-19), Indian healthcare system is far from being affordable and accessible. While the economic development has been gaining momentum since the last decade, India needs major reforms in existing healthcare system. Technology has an important role to play in streamlining the health infrastructure. A Disease prediction system has been proposed to address various deficiencies and provide affordable, readily-available and cost-effective healthcare. The proposed system can effectively predict life threatening diseases like Diabetes and Heart Disease. The system has been implemented by using Machine Learning Algorithms and enhanced to be more accurate than the existing systems. The system is backed by robust Machine Learning Algorithms and ensembles that will predict presence/absence of the diseases accurately. Accurate analysis of medical data benefits early disease detection, patient care and community services. To overcome the difficulty of incomplete data, the system uses Wearable Medical Sensors to gather Real-time data and make prediction based on the collected data. Presently, we demonstrate the system for two diseases but it can be scaled for tackling more diseases. The Disease Prediction system will not only reduce the burden on existing healthcare and diagnosis system but also provide personalized medication.

Keywords: *Clinical Decision Support System, Machine Learning Algorithms, Wearable Medical Sensors*

1. INTRODUCTION

Enhanced by advanced technology in healthcare, the medi- cal management has significantly progressed. In spite of many strides in healthcare, there is much more to achieve. Many studies show heart diseases and diabetes have emerged as one of the most fatal diseases in world. Of the 56.9 million deaths worldwide in 2016, more than half (54%) were due to the top 10 causes. Ischemic heart disease and stroke are the world's biggest killers, accounting for a combined 15.2 million deaths in 2016. These diseases have remained the leading causes of death globally in the last 15 years. Chronic obstructive pulmonary disease claimed 3.0 million lives in 2016, while lung cancer (along with trachea and bronchus cancers) caused 1.7 million deaths. Diabetes killed 1.6 million people in 2016, up from less than 1 million in 2000. Deaths due to dementias more than doubled between 2000 and 2016, making it the 5th leading cause of global deaths in 2016 compared to 14th in 2000. Therefore, the main objective of this paper is to report on a research project where we took advantage of those available technological advancements to develop prediction models for heart disease and Diabetes Type II.

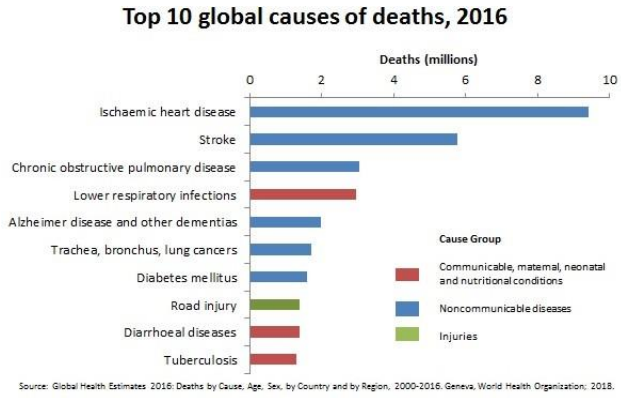

Fig. 1: Top Ten Global Causes of Deaths, 2016

The proposed system will have multi-tier structure, backed by robust machine learning, that enables diseases to be tracked individually by a disease diagnosis module. We will demonstrate the feasibility of such a system through disease diagnosis modules aimed at two disease categories. We will show that the system is scalable using more disease categories. Our aim is to Predict Heart Disease and Diabetes using Machine Learning Algorithms(MLA) and to use combination of various Machine Learning Algorithms(MLA) and ensemble techniques to enhance the accuracy of existing system. The

system being developed can effectively predict life threatening and unidentifiable diseases like Type II Diabetes and Cardiovascular Diseases. The system can be implemented using combination of Machine Learning Algorithms to achieve more accuracy. The said system will be cost effective and will prove to be a boon for people deprived of medical facilities. This clinical decision support system (CDSS) will not only reduce the burden on existing healthcare and diagnosis system but also provide personalized medication.

The rest of the paper is organized as follows. Section 2 provides background material required for understanding the rest of the paper. Section 3 presents the system architecture. Section 4 gives Analysis of Machine Learning Algorithms and Ensemble Methods. Section 5 presents Algorithmic Design. Section 6 discusses the limitations of this paper and future research opportunities. Finally, Section 7 concludes the paper.

2. BACKGROUND

In this section we will discuss the background knowledge required to understand the rest of the paper.

2.1 Clinical Decision Support System

A clinical decision support system (CDSS) is a health information technology system that is designed to provide physicians and other health professionals with clinical decision support (CDS), that is, assistance with clinical decision-making tasks. A working definition has been proposed by Robert Hayward of the Centre for Health Evidence: "Clinical decision support systems link health observations with health knowledge to influence health choices by clinicians for improved health care". CDSSs constitute a major topic in artificial intelligence in medicine. It is widely used in hospitals and clinics all over the world, supported by well-developed commercial platforms, such as Safety Surveillor, TheraDoc, Sentri7, QC PathFinder and Med Mined. In CDSS, a case specific advice is generated with the help of Electronic Health Record(EHR). In order to generate such an advice an encoding technique provided by World Health Organization(WHO), known as the International Statistical Classification of Diseases and Related Health Problems (ICD) coding system. But, the effectiveness of CDSS is in dispute due to many reasons.

- The CDSS is integrated into the clinical workflow rather than as a separate log-in or screen.
- The CDSS only provides decision support at the time and location of care (hospitals and clinics) rather than prior to or after the patient encounter.
- There is a large gap between the proposed and actual benefits of CDSS.
- The cost-effectiveness of CDSS is yet to be demon- strated.
- *Machine Learning Algorithms and Ensemble Methods*

2.1.1 Machine Learning Algorithms: MLAs enable computer to learn through building of analytical models. Machine learning algorithms can be divided into 3 broad categories: supervised learning, unsupervised learning, and reinforcement learning.

- Supervised learning: is useful in cases where a property (label) is available for a certain dataset (training set), but is missing and needs to be predicted for other instances.
- Unsupervised learning: is useful in cases where the challenge is to discover implicit relationships in a given unlabeled dataset (items are not pre-assigned).
- Reinforcement learning: falls between these 2 extremes there is some form of feedback available for each predictive step or action, but no precise label or error message. Referring these categories here, we use supervised MLA. There are various types of supervised Machine Learning Algorithms such as decision tree, Support Vector Machine(SVM), etc.

2.1.2 Ensemble Methods: Ensemble Methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. Numerous types of ensemble methods are available, some of them are: Bagging, Random tree, etc. Each algorithm and ensemble method performs optimally for a specific application category. Moreover, a proper combination of a supervised MLA and an ensemble technique can successfully generate stipulated outcomes. Considering these facts, in this paper, we primarily focus on five supervised MLAs and five Ensemble Methods. They are shown in the table below.

Name	Abberivation	Description
Naive Bayes	NB	Bayes theorem based probabilistic learner
Bayesian Network	BN	Network driven,conditional tables at node
J48	J48	Pruned or unpruned decision tree
Support Vector Machine	SVM	Support vector based linear seperator
Best-First decision tree	BFTree	Tree with binary splits on features
Random Tree	RT	Bagging on tree sampling instances
Random Forest	RF	Bagging on tree sampling instances
BAGGing	BAG	Training with sampled subsets
Decorate	DEC	Voting through diversified base learners
Voting	VOT	Voting through diversified base learners

Fig. 2: MLA and EM

2.2 Wearable Medical Sensors

Sensors and wearables allow continuous physiological monitoring with reduced manual intervention and at low cost. Evidently, Wearable Medical Sensors can be used to overcome the limitations of CDSS as mentioned above, to provide continuous health decision support. Due to rapid advancements in low-power sensing, computing, and communication, battery-powered WMSs are becoming increasingly ubiquitous. According to a report from Business Insiders, more than 56 million wearable sensors were sold worldwide in 2015. The list of collectable physiological signals includes, but is not limited to, heart rate (HR), body temperature (BT), respiration rate (RESP), blood pressure (BP), electroencephalogram (ECG), electrocardiogram (EEG), Galvanic skin response (GSR), oxygen saturation (SpO2), blood glucose (BG), and body mass index (BMI). For the purpose of our project we are using AD8232 ECG sensor module.

2.2.1 AD8232 ECG sensor module: This sensor is a cost-effective board used to measure the electrical activity of the heart. This electrical activity can be charted as an ECG or Electrocardiogram and output as an analog reading. ECGs can be extremely noisy, the AD8232 Single Lead Heart Rate Monitor acts as an op amp to help obtain a clear signal from the PR and QT Intervals easily. The AD8232 is an integrated signal conditioning block for ECG and other bio-potential measurement applications. It is designed to extract, amplify, and filter small bio-potential signals in the presence of noisy conditions, such as those created by motion or remote electrode placement. The AD8232 module breaks out nine connections from the IC that you can solder pins, wires, or other connectors to. SDN, LO+, LO-, OUTPUT, 3.3V, GND provide essential pins for operating this monitor with an Arduino or other development board. Also provided on this board are RA (Right Arm), LA (Left Arm), and RL (Right Leg) pins to attach and use your own custom sensors. Additionally, there is an LED indicator light that will pulsate to the rhythm of a heartbeat. Features:

- Operating Voltage - 3.3V

- Analog Output
- Leads-Off Detection
- Shutdown Pin
- LED Indicator
- 3.5mm Jack for Biomedical Pad Connection or Use 3 pin header

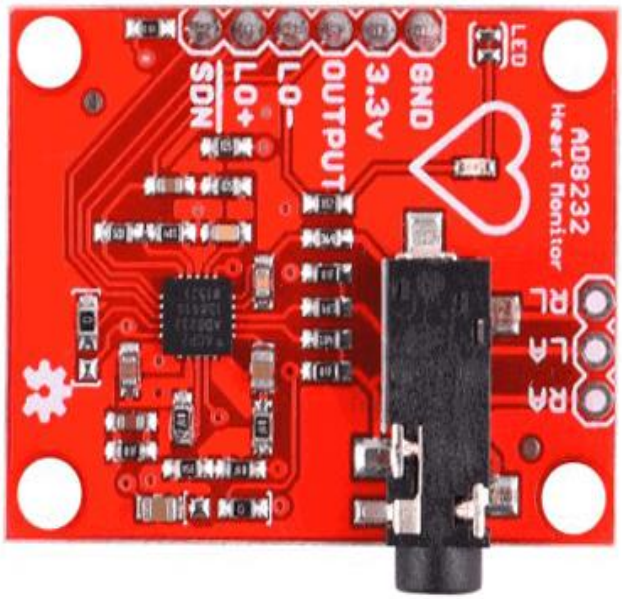

Fig. 3. AD8232 ECG Sensor Module

3. SYSTEM ARCHITECTURE

The system architecture can be broadly categorized into six categories being input module, user interface, knowledge base, ensemble techniques/methods, post diagnostic decision support module and output module.

3.1 Input module

Keeping the user aid in mind, diverse ways are made avail- able to provide the input values. One way is to directly feed the input values using the user interface. Other option is to provide the input using voice recognition i.e. Google Speech to Text API. Another alternative is to use the WMS made available (only for ECG).

3.2 User Interface

A deployed website takes input from the input module as per the dataset attributes and medical standards. The units in which the values should be considered are suggested here for user assistance.

3.3 Knowledge Base

The knowledge base can be further represented into various steps. The steps are as follows:

1) Dataset
2) Preprocessing
3) Training Dataset
4) Machine Learning Algorithms
5) Testing Dataset
6) Testing Dataset

The module assists with daily health monitoring. It incorpo- rates decision modules trained using clinical domain knowl- edge, and transmits information across the system. This helps individuals, even those without professional medical training, track their diseases. It also incorporates various Machine Learning Algorithms (MLAs).

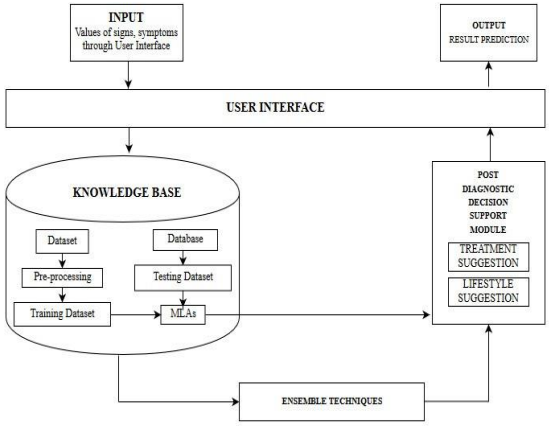

Fig. 4: System Architecture

3.4 Post Diagnostic Decision Support Module

This module decides the disease occurrence based on the prediction percentage computed in the previous module. If the percentage is above a specified threshold relevant treatment is suggested else lifestyle changes are suggested.

3.5 Output Module

The combined results and observations from the post diagnostic decision support module are represented here. Predominantly, the actual prediction results are embodied here.

4. ANALYSIS OF MACHINE LEARNING ALGORITHMS ANDENSEMBLE METHODS

The tabular analysis of MLAs and Ensemble Methods specified in section 2B is as follows;

Table 1: Diabetes without feature selection

Algorithm	Base Learner	Meta Learner		
		Bagging	*Decorate*	*Voting*
NB	75.91	76.3	75.78	-
BN	75.26	76.56	76.04	-
J48	73.83	74.09	75.13	-
SVM	76.95	76.95	75.78	-
BFTree	74.48	75.13	74.22	-
RT	68.36	75.13	72.79	-
RF	76.04	76.17	76.82	-
NB and BN	-	-	-	77.21
NB and SVM	-	-	-	76.95
NB and BFT	-	-	-	75.34
BN and SVM	-	-	-	76.95
BN and BFT	-	-	-	74.74
SVM and BFT	-	-	-	75.91

Table 2: Diabetes: With Feature Selection

Algorithm	Base Learner	Meta Learner		
		Bagging	*Decorate*	*Voting*
NB	75.78	76.3	76.04	-
BN	74.74	75.78	76.56	-
J48	73.17	75.26	76.56	-
SVM	77.6	76.82	76.95	-
BFTree	74.74	75.78	74.22	-
RT	70.83	73.83	73.83	-
RF	77.08	77.08	74.87	-
NB and BN	-	-	-	76.82
NB and SVM	-	-	-	77.6
NB and BFT	-	-	-	75.78
BN and SVM	-	-	-	77.6
BN and BFT	-	-	-	75
SVM and BFT	-	-	-	77.31

The algorithmic analysis of the proposed prediction model is given above. For the sake of simplicity and to reduce complexity, the algorithms which produce the maximum results are considered for building the prediction model. The algorithms selected are highlighted in the tables above.

1) For Diabetes: Support Vector Machine (With feature selection)
2) For Heart Disease: Naive Bayes (Without feature selection)

Table 3: Heart Disease:Without Feature Selection

Algorithm	Base Learner	Meta Learner		
		Bagging	***Decorate***	***Voting***
NB	79.21	79.43	79	-
BN	77.89	78.56	79	-
J48	73.08	73.09	72.43	-
SVM	77.68	77.9	77.68	-
BFTree	73.08	75.93	75.71	-
RT	66.73	74.4	73.96	-
RF	75.71	76.15	76.15	-
NB and BN	-	-	-	77.9
NB and SVM	-	-	-	77.68
NB and BFT	-	-	-	76.81
BN and SVM	-	-	-	77.68
BN and BFT	-	-	-	74.84
SVM and BFT	-	-	-	76.80

Table 4: Heart Disease: With Feature Selection

Algorithm	Base Learner	Meta Learner		
		Bagging	***Decorate***	***Voting***
NB	78.77	78.77	78.99	-
BN	78.12	77.68	77.89	-
J48	76.8	75.27	76.38	-
SVM	75.71	76.15	78.55	-
BFTree	72.86	76.36	-	-
RT	70.46	75.05	74.18	-
RF	76.14	77.24	78.34	-
NB and BN	-	-	-	78.99
NB and SVM	-	-	-	75.71
NB and BFT	-	-	-	76.36
BN and SVM	-	-	-	75.71
BN and BFT	-	-	-	75.71
SVM and BFT	-	-	-	75.22

5. ALGORITHM WORKING FLOW

5.1 Algorithmic approach

Step 1: Take signs and symptoms as input through User Interface.

Step 2: Input this data to the prediction model to predict the presence / absence of disease.

Step 3: Display the output whether the signs and symptoms correspond to any disease.

Step 4: If prediction is positive for presence of disease then, suggest a suitable treatment. Else suggest a suitable style of living.

5.2 Work Flow

The detailed working flow of the system is divided into sections namely:

1) Module 1: Diabetes prediction using SVM.
2) Module 2: Heart disease prediction using Naive Bayes.

a) Working flow of diabetes prediction module: The work- ing flow of diabetes prediction module is as follows.

b) More about Support Vector Machine (SVM): Support Vector Machine (SVM) is a Supervised Machine Learning Technique i.e. it works on Labeled training data. Support Vector Machine has its applications in the field of Classification as well as Regression. But in this system purpose is limited to classification only. In this algorithm, each data item is plotted as a point in an n-dimensional space. Where 'n' is the number of features in training data. After performing this initial stage, classification is performed by finding an optimal hyperplane. Hyperplane is decision boundary which separates data into number of target classes. In this system Support Vector Machine with Linear Kernel is implemented in Python which gives higher accuracy with Diabetes dataset among other Learning Techniques.

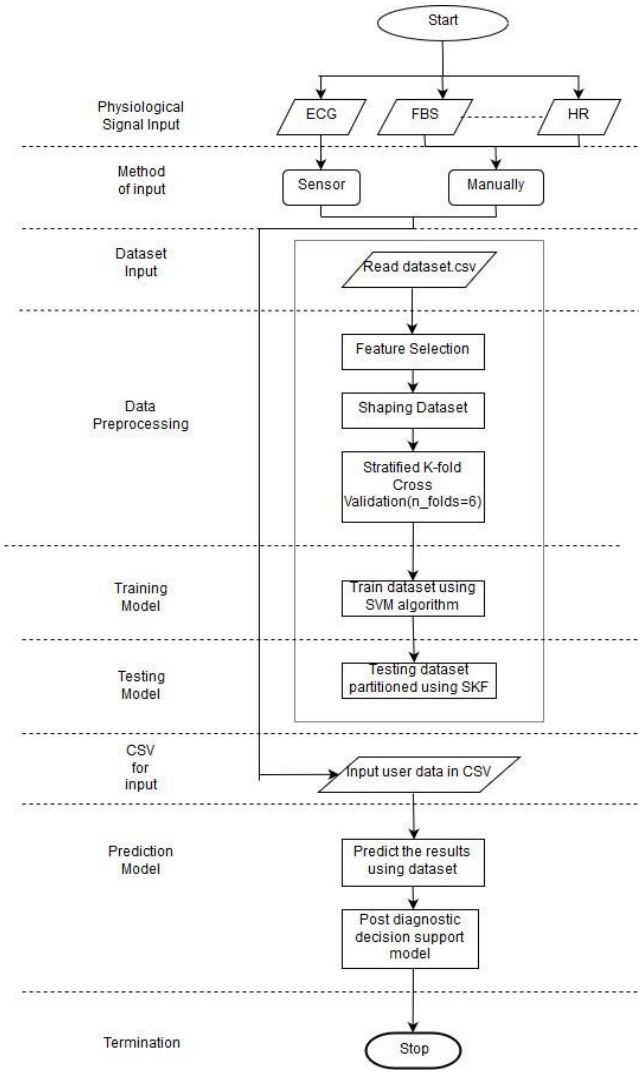

Fig. 5: Workflow (diabetes prediction module)

1) Working flow of heart disease prediction module: The working flow of heart disease prediction module is as shown in the flow diagram.

2) More about Naive Bayes (NB): Naive Bayes is also a Supervised Machine Learning Technique, but it is probabilistic in nature. This classifier is based on Bayes' Theorem where it assumes independence among predictors. Naive Bayes classi- fier assumes that the existence of particular feature in a class is not related to existence of others. In this system Gaussian Naive Bayes is implemented which is extension of simple Naive Bayes. Gaussian Naive Bayes assumes Gaussian or Normal Distribution among features. This algorithm is used because it gives higher and stable accuracy with Heart Disease Dataset.

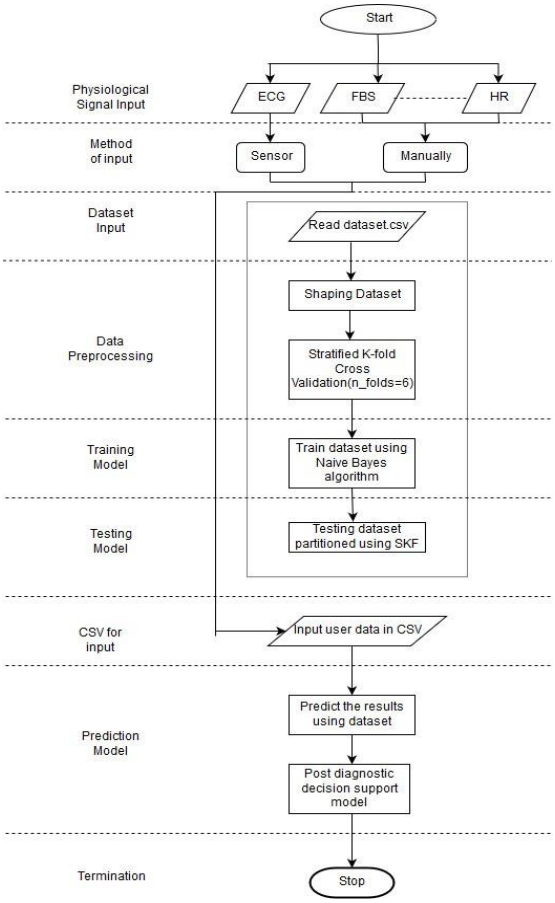

Fig. 6: Workflow (heart disease prediction module)

6. RESULTS OBTAINED

The comparative representation of analyzed results and obtained results is given graphically as follows;

6.1 Results Obtained - Stage II

1) Diabetes Prediction using Support Vector Machine:
a) Analyzed Prediction Accuracy: 77.6 %
b) Computed Prediction Accuracy: 78.74 %
2) Heart Disease prediction using Naive Bayes:
a) Analyzed Prediction Accuracy: 79.21 %
b) Computed Prediction Accuracy: 78.33 %

The system is having a limited scope for only two diseases which are Diabetes Type II and Heart Disease. The prediction model can be extended to predict various other life threatening diseases.

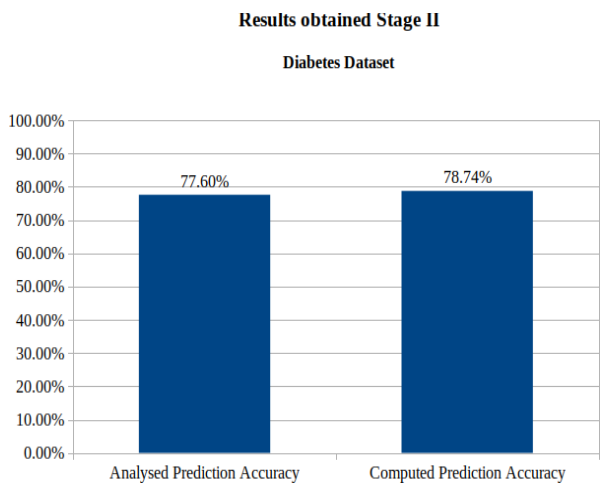

Fig. 7: Comparison of Accuracy - Diabetes

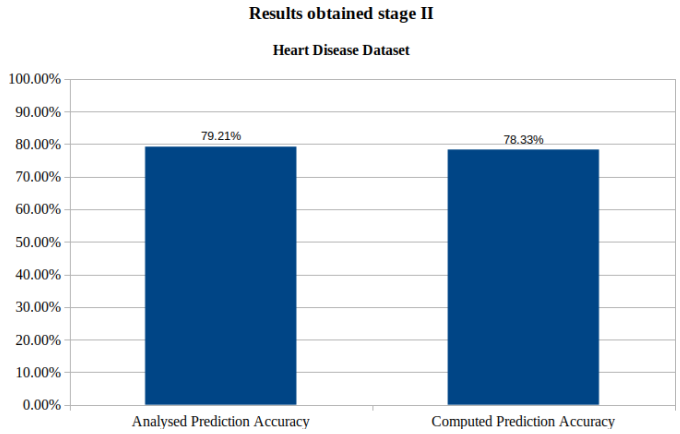

Fig. 8: Comparison of Accuracy - Heart Disease

7. CONCLUSIONS

The proposed system can effectively predict life threatening diseases like Type II Diabetes and Cardiovascular Diseases. The system is tested using combination of Machine Learning algorithms and the best combination is applied to achieve more accuracy.

The said system will be cost effective and will prove to be a boon for people deprived of medical facilities. This prediction system will not only reduce the burden on existing health care and diagnosis system but also provide personalized medication.

8. FUTURE WORK

The system is implemented by taking numerical inputs for each symptom from user. Real-time sensor data can be used to produce patient specific results which increases usability. Along with the real-time sensor data, a Voice Assistant can be deployed to input signs and symptoms to enhance user-experience.

9. REFERENCES

[1] Hongxu Yin, Student Memeber, IEEE, Niraj K.Jha, Fellow, IEEE, A Health DecisionSupport System for Disease Diagnosis Based on Wearable Medical Sensors and MachineLearning Ensembles., *In Proceedings of IEEE Transactions on Multi-Scale ComputingSystems*

[2] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang* "Disease Prediction by Machine Learning over Big Data from Healthcare Communities" DOI 10.1109/ACCESS.2017.2694446, IEEE Access

[3] (2016) UCI machine learning repository. [Online]. Available:http://archive.ics.uci.edu/ml

[4] Hongyan Liu, Hongpu Hu*, Quan Chen, Fan Yu,Yang Liu "Application of the Clinical Decision Support Systems in the Management of Chronic Diseases" The 2016 3rd International Conference on Systems and Informatics (ICSAI 2016)

[5] Arvin Agah, CRC Press "Medical Applications of Artificial Intelligence" Published: December 18, 2017

[6] M. M. Baig and H. Gholamhosseini, Smart health monitoring systems: An overview of design and modeling, J. Medical Systems, vol. 37, no. 2, pp. 114, 2013

[7] J. A. Salomon, H. Wang, M. K. Freeman, T. Vos, A. D. Flaxman, A. D. Lopez, and C. J. Murray, Healthy life expectancy for 187 countries, 19902010: A systematic analysis for the global burden disease study 2010, The Lancet, vol. 380, no. 9859, pp. 21442162, 2013

[8] (2016) Health care costs 101. [Online]. Available: https: //www.chcf.org/publications/2016/12/health-care-costs-

101
[9] D. L. Hunt, R. B. Haynes, S. E. Hanna, and K. Smith, Effects of computer- based clinical decision support systems on physician performance and patient outcomes: A systematic review, J. American Medical Association, vol. 280, no. 15, pp. 13391346, 1998
[10] P. Melville and R. J. Mooney, Constructing diverse classifier ensembles using artificial training examples, in Proc. Int. Joint Conf. Artificial Intelligence, vol. 3, 2003, pp. 505510
[11] L. Pollonini, N. O. Rajan, S. Xu, S. Madala, and C. C. Dacso, A novel handheld device for use in remote patient monitoring of heart failure patients Design and preliminary validation on healthy subjects, J. Medical Systems, vol. 36, no. 2, pp. 653659, 2012
[12] (2016) Health watch medical grade smart clothing technology. [Online]. Available: http://www.personal-healthwatch.com/ technology

