Non Communicable Disease Prediction System Using Machine Learning

Priyanka Kadu, Amar Buchade

Abstract: In day to day life, routine checkup has become a tedious task as patients need to visit a clinic and this task consumes more time and money. It is convenient to have a system which tells the user or patient the probability of the disease that might happen in the future. Machine learning helps to predict the disease proneness for the given attributes that are collected on a periodic interval for the patient. This eventually can become an advantage to the health care system to play a vital role in the patient's life. Some methodologies are existing which works on the precise attribute values to estimate the disease proneness. Most of the time it becomes difficult to provide the precise measuring parameters due to the involvement of the pathology laboratory. To provide an effective solution, the proposed system uses abstract parameters, viz.: step count, hours of sleep, weight, BMI (Body Mass Index) and calories burned. In the proposed work KNN (k-nearest neighbor) and HMM (Hidden Markov Model) are used to estimate the probability of disease proneness. To enhance the performance of system fuzzy classification is used. The system is tested for estimating the disease proneness for six diseases. It is observed that the accuracy of estimation is above 80%.

Index Terms: Hidden Markov model, Health prediction, k- nearest neighbor clustering, Shannon Information Gain, Fuzzy Classification.

1 INTRODUCTION

The advancements in the medical and the healthcare industry have been advancing at an unprecedented rate with various activities such as accurate detection and analysis of various diseases and patients with the help of extensive medical data. This particular data has been responsible for the average increase in the life expectancy of human beings. As data is one of the most important aspects of bio-medical care community, it proves to be a valuable asset for the accurate analysis of the medical condition. For this purpose, there is a need for very high-quality medical data as the amount of quality data available is directly proportional to the accuracy of the analysis. The data is also very helpful in scenarios where there is an outbreak or an epidemic of a disease. The prediction of diseases can be accurately determined with the help of a good quantity of quality data. The prediction of various diseases and their epidemics can be highly useful as it can help the doctors and helps the hospitals stock up the proper medication in advance to help ameliorate the effects of an epidemic. In many medical communities, data is collected for different health-related issues. These data can be utilized using various techniques and algorithms to gain useful information. Many times, collected data does not give insight on information or it is not in understandable format and humans cannot get useful information from it. Thus, there is a need for different techniques to explore information. There are various techniques are available like machine learning, data mining which helps in exploring data and gain insight on its. Thus these techniques may help to predict various diseases. It's been always a tedious task for patients to visit a doctor for the routine checkups, as this task consumes more time and money. Many times, people often misjudged the symptoms and bear severe consequences. Many diseases attack a person so instantly that is hard gets any time to get treated with. So it is convenient to have a system which tells the user or patient the probability of the disease he/she might prone in the future. Scope of machine learning in the prediction of diseases has been broadening significantly. Machine learning techniques improve accuracy in disease prediction which helps in avoiding patients who don't require treatment at an initial stage rather than mistreating them which would put a physical and financial burden on the patients. This research paper dedicates section 2 for analysis of past work as literature survey, section 3 deeply elaborates the proposed technique and whereas section 4 evaluates the performance of the system and finally section 5 concludes the paper with traces of future enhancement.

2 LITERATURE REVIEW

This section of the literature survey eventually reveals some facts based on thoughtful analysis of many authors work as follows. Mahmood Hussain Kadhem and Ahmed M. Zeki [1] presented the study which shows the powerful analysis capability and importance of the data mining (DM) classification algorithms. The dataset collected by a medical expert has been processed using the CRISP-DM for proper identification of the factors affecting the prediction of the acute inflammations of the urinary system disease and for dataset preprocessing and analysis. In order to choose the best classification algorithm to be used for building the prediction model, a comparative analysis has been conducted on the OneR, Ridor, and J48 algorithms in terms of accuracy, precision, and time needed for the model construction. Prof. Dhomse Kanchan B. and Mr. Mahale Kishor M. [2] presented heart disease prediction by using SVM, Decision tree and Naive Bayes. These algorithms applied to the dataset with or without using PCA. They used PCA to minimize the attributes number. In their result, they found that SVM outperforms Decision Tree and Naive Bayes if the size of the dataset reduces. GUI system software may be constructed using SVM and dataset to expect the possibility of cardiovascular disease in a patient and for diabetes facts prediction. They used WEKA tools to implement their algorithms and to analyze their algorithm accuracy. These algorithms compare classifier accuracy to every other on the premise of correctly classified

- *Priyanka Kadu, Computer Engineering, Pune Institute of Computer Technology, Pune, India. priyankakadu2394@gmail.com*
- *Amar Buchade, Computer Engineering, Pune Institute of Computer Technology, Pune, India. amar.buchade@gmail.com*

instances, time is taken to build model, mean absolute error and ROC place. They can conclude that most ROC area means method great predictions performance as compared to different algorithms. Nabil Alshurafa et al. [3] introduced an enhanced RHM (Remote Health Monitoring) system, called as Wanda-CVD. The Wanda-CVD RHM system is an advanced version of a previous RHM system developed by the authors named as Wanda that targets patients at risk of CVD through Wireless Coaching. Wireless Coaching is in the form of automated messages encourage the user to take definite actions, such as measuring their blood pressure or reminding them to increase exercise intensity. There are several key components in the complete design of the Wanda-CVD system. The first component is the Android-based Smartphone application designed as a means to collect data from the user while displaying clinician feedback. Anis Davoudi et al. [4] proposed a method to predict delirium by using Electronic Health Records. They studied seven ML (machine learning) models: generalized additive models, logistic regression, support vector machines, naïve Bayes, random forests, extreme gradient boosting, and neural networks. Among these models, random forests and generalized additive models were the best performing delirium, prediction models. The data that they used for these models are available at the point of access to preoperative care and do not require specialized assessments or self-report information. They included complex variables, like residency ZIP codes and attending doctor. ZIP codes can act as a surrogate of neighborhood socioeconomic characteristics, which has been shown to be associated with numerous disease and health behaviors. The performance of attending doctors can potentially be a factor of postoperative outcomes as well. They found that factors like alcohol, drug, age, socioeconomic status, the severity of the medical problem and underlying medical issue can affect the risk of delirium. Vineet Kumar Singh et al. [5] introduced two different types of classification technique that they used to predict some of the basic human motion using a portable sensor. Feature extraction and classification are the main methods that they used to develop their proposed system. The main goal of their study was to devise an efficient method to differentiate, recognize and assign a class. The use of statistical tools aided them in processing and displaying their results in a clear-cut way. They also compare their proposed system with two classification techniques. They not only identified the human motion but also differentiate between movements that are closely related. This model can be used as a safety measure and for keeping track of the elderly and infants. B. Nithya and Dr. V. Ilango [6] presented the studies of various prediction technique and tools for machine learning. Machine Learning has given medical providers new tools to work with, novel ways to practice medicine. It also confirms that machine learning tools and techniques are decisive in health care province and exclusively used in the diagnosis and predictions of various types of cancers. There are a lot of open problems and future challenges in dealing with massive amounts of heterogeneous, distributed, diverse, highly dynamic data sets and increasingly large amounts of unstructured and non-standardized information with respect to varied types of cancers. Some of the most important challenges in clinical practice and biomedical research include the need to develop and apply novel tools for the effective integration, analysis, and interpretation of complex biomedical data with the aim to identify the testable hypothesis and build accurate models to diagnose and predict various types of cancers and their recurrences. Abderrahmane Ed-daoudy and Khalil Maalmi [7] proposed a real-time health status prediction system. The proposed system is a data processing, monitoring application combining socket streams and Spark Streaming. This system will process real-time data sent by connected devices as socket streams and store that data for real-time analytics. The use of big data tools especially spark, significantly enhanced the performance and the efficiency of the proposed health status prediction system, mainly in terms of system development time, complexity of programs and processing time, in comparison with traditional analytics tools, which requires a variety of skills, intensive and more expensive programs and significant amount of time and money. Fahad P K and Pallavi M S [8] proposed CNN-MDR with the help of back propagation algorithm based on CNN. They used different hospitals data available in both structured and unstructured way. They achieved a 96.4% prediction rate. Po-Han Chiang and Sujit Dey [9] proposed a data-driven model to investigate the individual effect of health behavior on BP using wearable devices and BP monitors. Their machine learning model can provide not only a daily prediction of SBP and DBP but also importance score of health behavior factors on an individual's daily BP. By extracting the time-series related data and integrating the RF-based feature selection technique, they enhance the prediction performance of the original RF model. The experiment result shows that their technique outperforms other existing techniques in terms of MSE and MAE. They showed a significant change in BP after users changed their most significant health behavior features suggested by their model, which validates the personalized recommendation on health behavior. Raid M. Khalil and Adel Al-Jumaily [10] presented the technique to predict the depression associated with type 2 diabetes. The author selected four different machine learning techniques because of the possibility of obtaining accurate results using the available data size. The techniques that used include the support vector machine K-Mean, (SVM), F-C mean, and Probabilistic Neural Network (PNN). It is clear from the results that the SVM classifier generates more precise results than the others. Abdul Mahatir Najar et al. [11] presented the ELM architecture to predict the Dengue Hemorrhagic Fever (DHF) using machine learning. The risk level prediction of DHF required variable like weather condition as input and predict DHF cases as output. The results showed that ELM can be a very promising model for the risk level of DHF prediction. Based on their approach, the best performance is ELM network using binary activation function with 50 hidden neurons where the MAE is 0.08698 and MAPE is 3.00536. Nitten S. Rajliwall et al. [12] proposed a predictive modeling framework which used a clinical static and low-velocity data from electronic health records and clinical notes, and live streaming data captured with biosensor enabled wearable. They used the “Physical Activity_PAQ_B” and “Cardiovascular Fitness_CVX_B-Model” dataset for model creation. “Framingham Heart Study dataset” which is the second dataset, is obtained from the publicly accessible section of Framingham Heart institute dataset. Researched framework based on the neutron networks and supervised machine learning algorithms processing including grouping and filtering based on year born, sex and degree earned allows commerce with tasks of streaming high-volume data from biosensor watches. Researched assessment of the planned predictive framework, based on static and low-velocity

data, from diverse widely available cardiovascular organization studies i.e. Framingham Heart Study and National Centre for Health statistics in Centre for Disease Control, had shown good results. For the initial dataset i.e. "National Centre for Health statistics in Centre for Disease Control", ideal features selection methods are picked based on information theory ranking lead to improved performance. For the other dataset i.e. FHS dataset, a grouping based on subdivision using filtered variables method resulted in the best stats.

3 PROPOSED METHODOLOGY

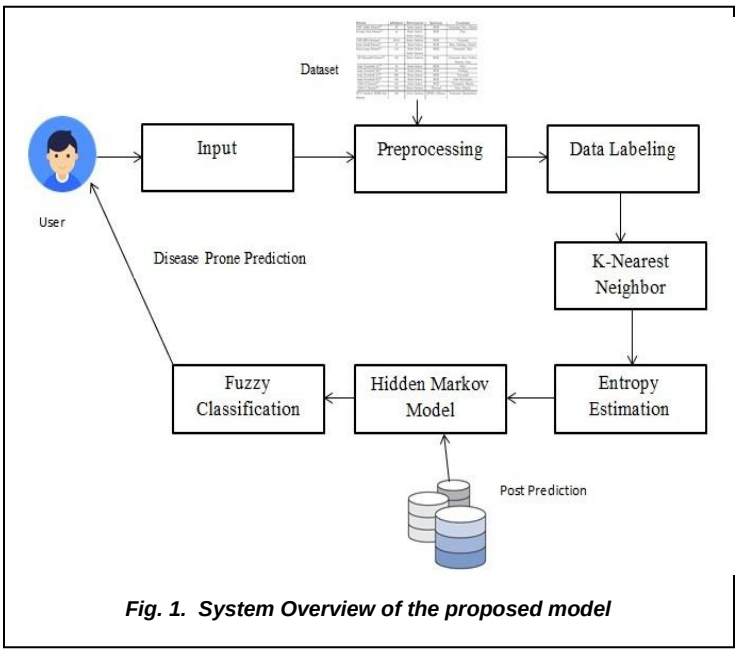

Fig. 1. System Overview of the proposed model

The proposed model of disease proneness detection is depicted in the figure 1 and the steps that involved in the process are deeply elaborated in the given mentioned steps. *Step 1: Dataset Collection and storage*- The proposed model of disease proneness prediction is deployed using the abstract health parameter dataset. This dataset is obtained from the open source dataset repository from Kaggle [13]. This dataset contains attributes like user ID, Step count, mood, calories burned, hours of sleep, bool_of_active, weight_kg, and BMI. This dataset is downloaded in the form of a workbook and the proposed model uses the JXL API to read this workbook. The dataset in the workbook is stored in the database and an interactive user interface is provided to the user to add more data into the database based on this user ID. *Step 2: User Input and Preprocessing*- The proposed model provides an interactive user interface through which the user can login into the system to perform some operations like add data into the database and viewing the data from database. Along with these operations users are provided to see the disease proneness possibilities for his/ her stored data. Once the user has selected the option of disease proneness detection then that user's complete data is extracted in the form of a double dimension list from the database, which is subjected to preprocessing process. In preprocessing step, some of the attributes are selected from each of the row like step count, calories burned, hours of sleep, weight and BMI. These attributes are added in a list and then into an another list labeled as the preprocessed list. *Step 3: K Nearest Neighbor Clustering*- This is the step where the first machine learning model is deployed on the preprocessed double dimension list. The model of K- nearest neighbor consisting of 5 major steps as described below. *[A] Distance Evaluation* - In this step of KNN the preprocessed list is subject to evaluate the distance, where each row is tending to find the distance with all other remaining rows using the equation 1. The mean of all these distances is called as the Row Distance RD, which are appended at the end of each row of the preprocessed list. The average of all the row distances RD is referred as the average distance of the preprocessed list AVG_D.

$$E_D = \sqrt{(x1-x2)^2 + (y1-y2)^2}$$ --------(1) Where, E_D = Euclidean Distance x1, y1, x2, y2 - attributes *[B] Sorting* - The preprocessed list where each row is appended with the row distance R_D is subjected to sort in ascending order using the bubble sort technique with respect to the row distance R_D. *[C] Data point selection* - In this section of implementation 5 random numbers are selected from the sorted list which are in the range of 1 to 100. Then these numbers are normalized in the range of 1 to preprocessed list to call them as the data points. *[D] Centroid Estimation* - Specific rows are selected for the estimated data points as their positions. Then the R_D of each row is considered and stored in a list to call them centroids of the clusters. *[E] Boundary Estimation and Cluster formation* - Once the centroids are estimated then for each of the centroid a boundary is defined by using the Average distance of the preprocessed list. The boundary estimation can be defined using the equation 2

$$f(B) = \int_{i=1}^{n} (Ci - AVGD) \rightarrow (Ci + AVGD)$$ ------(2)

Where AVG_D - Average Euclidean distance n- Number of centroids C_i - Instance Centroid As the cluster boundaries are defined, then each of the clusters is formed based on the R_D of each row which lies in between the estimated boundaries. The rows which have not belonged to any of the boundaries are gathered as the outliers and added as the another cluster at the end. *Step 4: Entropy Estimation* - The formed clusters are subjected to estimate the entropy using the Shannon information gain theory. The estimation of the entropy helps to identify the most important clusters which are having the possibilities of the disease proneness. For this process every cluster is considered to estimate the count for the different attributes of each row in the cluster. Some protocols are set to estimate the count for a cluster like if the step count is less than or equal to 250, if calories burned is less than or equal to 50, number of hours of sleep is less than or equal to 5 and if the BMI is less than 18 or more than 25. Once the count is estimated for the defined protocol then if the count is more than 2 then that row count R_C is tending to increase. After this process the information gain is estimated using the equation 3. Where the clusters whose gain value are greater than or equal 0.5 is considered as the cluster with the probability of having some disease elements and they are selected for the next process. IG(E) = - (P / T) log (P / T) - (N / T) log (N / T) --------(3) Where P= Row count R_C T= Cluster Elements Size N= T-P IG(E) = Information Gain for the given cluster *Step 5: Hidden Markov Model* - The clusters obtained from the past step are subjected to Hidden Markov model to estimate the Hidden probability of the disease proneness. In this process the mean and standard deviation of each cluster is evaluated based on the row distances as mentioned in the equation 4 and 5. The equation 6 denotes the range of the row distance

that indicates the quality of the data responsible for disease proneness. Then, based on the minimum, maximum and in between the range values of the mean and standard deviation the row of each cluster is segregated and stored in a list called probability cluster list.

$$\mu = \frac{(\sum_{i=0}^{n} RDi)}{n} \quad \text{---------(4)}$$

$$\delta = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(RDi-\mu)^2} \quad \text{----------(5)}$$

$f(R_R) = (\mu - \delta) \rightarrow (\mu + \delta)$ ----------(6) Where, μ = Mean δ = Standard Deviation RD = Row Distance n = Number of rows This probability cluster is fed to Baum Welch model estimate the most probability rows that are eventually have the fact of disease proneness. The evaluation of the Hidden factors can be shown in the below algorithm 1.

Algorithm 1: BAUM WELCH

// Input: Probability Cluster List P_{CL}
// Output: Hidden Factor List H_{FL}
Function: hidden Factor Estimation (P_{CL})
Step 0: Start
Step 1: H_{FL} = Ø
Step 2: ***for*** i=0 TO Size of P_{CL}
Step 3: count=0
Step 4: $S_G = P_{CL[i]}$
Step 5: ***for*** j=0 TO Size of S_G
Step 6: Row= S_{Gj}
Step 7: S_C= Row[0] [Step Count]
Step 8: C_B= Row[1] [Calorie Burned]
Step 9: H_S= Row[2] [Hours of Sleep]
Step 10: B_{MI}= Row[3] [BMI]
Step 11: IF S_C <=100 THEN count++
Step 12: IF C_B <=10 THEN count++
Step 13: IF H_S <=4 THEN count++
Step 14: IF B_{MI} <18 || B_{MI}> 25 THEN count++
Step 15: IF count>=2
Step 16: $H_{FL=}$ H_{FL}+Row
Step 17: **End *for***
Step 18: **End *for***
Step 19: return H_{FL}
Step 20: Stop

Step 6 - Fuzzy Classification- In this last step the evaluated hidden factors for step count, Calorie burned, hours of sleep and BMI are counted against the set protocol. Then a fuzzy crisp value is evaluated based on the algorithm 2.

ALGORITHM 2: FUZZY CRISP SET FORMATION

//Input : Probability set P_S
//Output: F_C- Fuzzy Crisp Set
1: Start
2: R1=0, R2=0
3: D= P_S SIZE
4: ***For*** i=1 to 5
5: MIN=R1
6: R2=R2+D
7: R1=R2
8: T= Ø
9: ADD MIN to T
10: ADD R2 to T
11: ADD T to F_C
12: ***End for***
13: **return** FC
14: stop

Each of the count for the measured parameters is analyzed for their fuzzy crisp value positions. Based on the obtained positions different disease proneness is predicted. The proposed model predicts the disease like Diabetes, Insomnia, Obesity, Anxiety, Hypertension and Cardiovascular. If the measured count doesn't fall in any of the fuzzy crisp ranges, then the person is declared as the Healthy by the system.

4 RESULTS AND DISCUSSIONS

The proposed system for disease proneness prediction is developed and executed under windows environment in an Intel Pentium corei5 processor with 6GB RAM. The Java Programming language is used for the development with NetBeans as the standard IDE and the model uses MySQL as the standard database to store the dataset values. To measure the efficiency of the proposed model some, experiments are conducted as described below. To measure the effectiveness of the proposed model, accuracy is considered as the measuring parameter. Accuracy can be stated with the below mentioned equation 7. Accuracy = (TP + TN)/N ----------(7) Where TP- True Positive, that indicates the properly predicted Disease in the given number of trails. TN- True Negative, that indicates the properly predicted negative outcome of the disease for the given number of trails. N- Number of instances. The proposed model is set to measure the accuracy by conducting 10 sets of experiments, where each experiment consists of 10 trails. Based on the obtained output accuracy is estimated, which is shown in the table 1.

Experiment Number	Number of Trails	TP	TN	Accuracy
1	10	6	3	90
2	10	7	1	80
3	10	6	2	80
4	10	7	1	80
5	10	8	1	90
6	10	5	2	70
7	10	5	2	70
8	10	7	2	90
9	10	6	2	80
10	10	7	2	90

Table 1: Accuracy measurement for disease prediction

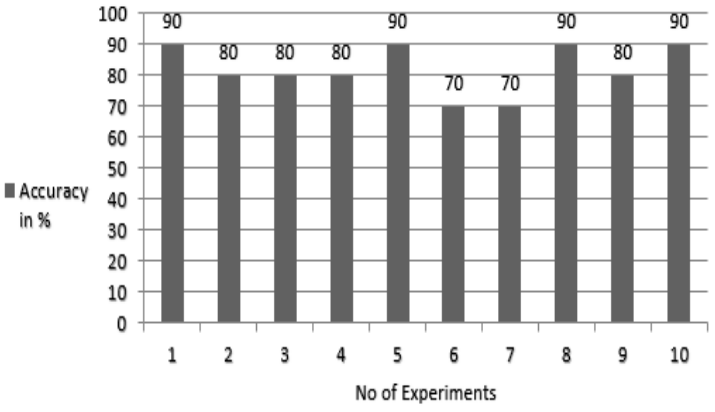

Fig. 2: Accuracy Measurement Results

Table 1 indicates the data for the conducted experiments to measure accuracy. Table 1 depicts the average obtained accuracy of the proposed model as 82%.

RMSE:The effectivity of the proposed model has been evaluated effectively with the help of a few experiments that are conducted to ascertain the error level of the system. For this purpose, RMSE or root mean square error technique has been utilized, this technique has been widely used to analyze the error between two continuous and corresponding entities. The two entities in the presented technique are the predicted disease proneness and the actual disease proneness. These two entities are utilized in equation 8 to calculate the RMSE value and depicted in table 2.

$$RMSE_{fo} = [\frac{\sum_{i=1}^{N}(z_{fi} - z_{oi})^2}{N}]^{\frac{1}{2}}$$

---------(8) Where, $\sum$ - Summation $(Z_{fi} - Z_{oi})^2$ - Differences Squared for the Disease Proneness Detection. N - Number of Conducted Experiments.

Execution Time:

Experiment No	KNN	Entropy Estimation	HMM	Fuzzy Classification
1	47ms	0ms	16ms	468ms
2	62ms	0ms	0ms	344ms
3	62ms	0ms	16ms	532ms
4	78ms	0ms	0ms	421ms
5	62ms	0ms	16ms	531ms

Table 2: Process Required Time for Each Module

Table 2 shows that the processing time is taken by each module.

5 Conclusion

The proposed model uses the abstract health parameters like a number of hours sleep, calories burned, no of steps walked every day, BMI and weight in kg as the major attributes. The proposed model successfully clusters the data based on the K-nearest neighbor algorithm. These clusters are evaluated for the disease proneness factor and selected for the hidden factor estimation process by Shannon information gain theory. HMM along with the fuzzy classification successfully predicts the proneness of some disease for the abstract user attributes using the clusters evaluated by the Shannon information gain theory. The proposed model successfully predicts some diseases like diabetes, insomnia, obesity, anxiety, hypertension and cardiovascular. The obtained accuracy of the proposed system is above 80%. In the future the proposed model can be deployed for various types of the diseases like cancer based on the more abstract parameters.

References

[1] M. H. Kadhem, A. M. Zeki, “Prediction of Urinary System Disease Diagnosis: A Comparative Study of Three Decision Tree Algorithms,” International Conference on Computer Assisted System in Health, 2014, pp. 58-61.

[2] K. B. Dhomse, K. Mahale, M. H. Kadhem, A. M. Zeki, “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis,” International Conference on Global Trends in Signal Processing, Information Computing and Communication, 2016, pp. 5-10.

[3] N. Alshurafa, C. Sideris, M. Pourhomayoun, H. Kalantarian, M. Sarrafzadeh, J. Eastwood, “Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data,” IEEE Journal of Biomedical and Health Informatics, vol. 21, Issue 2, pp. 507-514, 2017.

[4] A. Davoudi, A. Ebadi, P. Rashidi, T. O. Baslanti, A. Bihorac, A. C. Bursian, “Delirium Prediction using Machine Learning Models on Preoperative Electronic Health Records Data,” IEEE 17th International Conference on Bioinformatics and Bioengineering, 2017, pp. 568-573.

[5] V. K. Singh, S. Bhatia, A. Verma, S. Shaily, S.Kalaivani, “Human Motion Prediction,” International Conference on Communication and Electronics Systems, 2017, pp. 31-35.

[6] B. Nithya, V. Ilango, “Predictive Analytics in Health Care Using Machine Learning Tools and Techniques,” International Conference on Intelligent Computing and Control Systems, 2017, pp. 492-499.

[7] A. Daoudy, K. Maalmi, “Application of Machine Learning Model on Streaming Health Data Event in Real-Time to Predict Health Status Using Spark,” 2018 International Symposium on Advanced Electrical and Communication Technologies, 2018, pp. 1-4.

[8] P. K. Fahad, M. S. Pallavi, “Prediction of Human Health using Machine Learning and Big Data,” International Conference on Communication and Signal Processing, 2018, pp.1-8.

[9] P. Chiang, S. Dey, “Personalized Effect of Health Behavior on Blood Pressure: Machine Learning Based Prediction and Recommendation,” International Conference on e-Health Networking, Applications and Services, 2018, pp. 1-6.

[10] R. M. Khalil, A. Jumaily, “Machine Learning Based Prediction of Depression among Type 2 Diabetic Patients,” 12th International Conference on Intelligent Systems and Knowledge Engineering, 2017, pp. 1-5.

[11] A. M. Najar, M. I. Irawan, D. Adzkiya, “Extreme Learning Machine Method for Dengue Hemorrhagic Fever Outbreak Risk Level Prediction,” International Conference On Smart Computing And Electronic Enterprise, 2018, pp. 1-5.

[12] N. S. Rajliwall, R. Davey, G. Chetty, “Machine Learning Based Models For Cardiovascular Risk Prediction,” International Conference on Machine Learning and Data Engineering, 2018, pp. 142-148.

[13] Fitness Trends Dataset, Kaggle, Sept. 2018. [Online]. Available: https://www.kaggle.com/aroojanwarkhan/fitness-data-trends