An Ensemble Machine Learning Based Approach for Health Risk Prediciton

Amit Yadav
Department of Computer Science and Engineering
IET, DAVV
Indore, India
amit.sunoracle@gmail.com

Amit Kumar Mittal
Department of Computer Science and Engineering
IET, DAVV
Indore, India
amittal@ietdavv.edu.in

***Abstract*—Chronic health risks have risen among young individuals due to several factors such as sedentary lifestyle, poor eating habits, sleep irregularities, environmental pollution, workplace stress etc. The problem seems to be more menacing in the near future, with the exacerbation of lifestyle conditions and unforeseen breakout of pandemics such as COVID-19. One possible solution is thus to design health risk prediction systems which can evaluated some critical features of parameters of the individual and then be able to predict possible health risks. As the data shows large divergences in nature with non-correlated patterns, hence choice of machine learning based methods becomes inevitable to design systems which can analyze the critical factors or features of the data and predict possible risks. This paper presents an ensemble approach for health risk prediction based on the steepest descent algorithm and decision trees. It is observed that the proposed work attains a classification accuracy of 93.72%. A simple graphic user interface has also been created for the ease of use and interaction and for prototype testing.**

Keywords— Information and Communications Technology (ICT), Automated Health Risk Prediction, Ensemble Classifiers, Deep Neural Networks, Decision Trees, Classification Error, Accuracy.

I. INTRODUCTION

The sudden and unforeseen outbreak of Covid-19 has forced people worldwide to re-evaluate their lifestyle and health condition. With increase in the sedentary lifestyle of people around the globe, different health risks are affecting people worldwide [1]. While life expectancy has increased, but increasing health risks can be seen throughout the world. The majority of the population are pre-occupied in sedentary and non-active vocations neglecting the health markers which has seen an earlier precedence of health risks in people. The major reasons happen to be [2]:

1) Sedentary Lifestyle
2) Lack of Physical Exercise.
3) Poor Food Choices.
4) Environmental Pollution.
5) Climate Change
6) Stress in everyday life etc.

Hence, an urgent need to address the health risks has become imperative. However, the cost of healthcare medications is also continuing to rise. It is the government's job to have an efficient, cost-effective medical system which can cater to the needs of all socio-economic sections of the society [3]-[4]. However, medical care sees a continuous upsurge in costs and sudden pandemics put additional burden on the existing medical infrastructure. Thus, in practical scenarios, rendering personalized medical care to individuals belonging to different strata of the society is challenging as different individuals have different requirements and different prevailing health conditions and the medical infrastructure is already strained to the limit. Hence, it is necessary to look for alternatives which can address the present issues. One such avenue, which has been garnering a lot of interest among researchers, is Information and Communications Technology (ICT) in healthcare [5]. The objective of this paper is to identify the major health risks globally and design a recommendation system which can indicate possible health risks in the future. For this purpose, the major health risks causing mortalities or serious ailments is analyzed and depicted in figure 1.

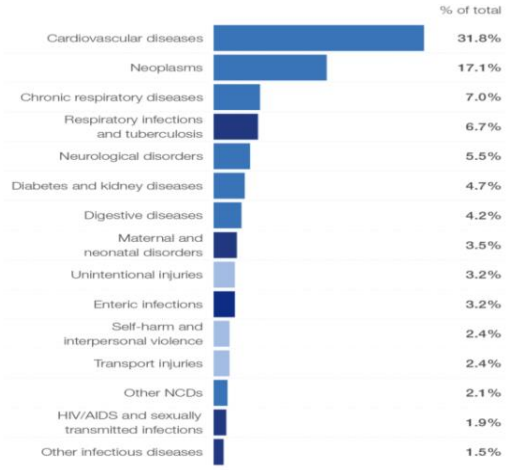

Fig. 1. Global Health Risk Analysis: 2019

(Source: World Economic Forum, [6])

The analysis of figure 1 helps us in deciding upon the data collection pertaining to symptoms, range of medical parameters and positivity rate of the major causes of health ailments. The data collection and labelling would subsequently lead to the design of an effective automated tool for health risk prediction [7]-[8].

The paper is organized as follows: Section I introduces the necessity of designing an automated health risk prediction system, the motivation and objectives behind the work. Section II presents the existing methods for designing such recommendation systems. Section III presents the proposed methodology for designing the system. Section IV presents

the simulation results. Section V presents conclusion and directions for future work.

II. EXISTING METHODS

The medical system's digitization has led to the influx of massive amounts of medical data [9]-[10]. This data can help medical care institutions to improve the efficiency of the health system, enhance the quality of healthcare and minimize healthcare costs. Moreover, the data available can also be utilized for research in non-conventional and upcoming fields such as machine learning and deep learning to come up with tools which can aid medical care [11]-[12]. The advent of cheap smart phones and internet services, Healthcare Information Systems (HIS) and Electronic Medical Records (EMRs) have become easily accessible far and wide [13]. However, accessing medical data, analyzing medical data and coming up an accurate predication system is complex due to the following reasons [14]:

1) Medical records may be extremely unstrcutured in nature with large divergences in data files such as text, audio, images, video etc.

2) Medical data, in general is ethnically diverse and its is clahhenging to extract intelligeble features (parameters) which can cater to a large diaspora of patients globally.

3) Medical data may be distributed or sparse. Data might have huge amounts of missing values owing to different human considerations.

Several techniques have been investigated to find patterns in medical data and thus create a mapping among the cases, symptoms and recorded test values [15]. Machine learning has been commonly used in numerous healthcare systems, such as medical imaging risk identification, diagnosis of illness, and prediction of health status from electronic health records [16]-[17]. Machine learning offers a way to automatically identify trends and predict results [18]-[19]. There are several current experiments on various types of electronic medical data on data mining and data analytics. Machine learning based classifiers are typically much more accurate and faster compared to the conventional classifiers [20]. They render more robustness to the system as they are adaptive and can change their characteristics based on the updates in the dataset [21]. The common classifiers which have been used for the classification of glaucoma cases are:

Regression Models:

In this approach, the relationship between the independent and dependent variable is found utilizing the values of the independent and dependent variables. The most common type of regression model can be thought of as the linear regression model which is mathematically expressed as [22]:

$$y = \theta_1 + \theta_2 x \quad (1)$$

Here,
x represents the state vector of input variables
y represents the state vector of output variable or variables.
Θ1 and Θ2 are the co-efficients which try to fit the regression learning models output vector to the input vector.
Often when the data vector has large number of features with complex dependencies, linear regression models fail to fit the input and output mapping. In such cases, non-linear regression models, often termed as polynomial regression is used. Mathematically, a non-linear or higher order polynomial regression models is described as:

$$y = \theta_0 + \theta_1 x^3 + \theta_2 x^2 + \theta_3 x \quad (2)$$

Here,
x is the independent variable
y is the dependent variable
$\theta_1, \theta_2 \ldots . \theta_n$ are the co-efficients of the regression model.
Typically, as the number of features keep increasing, higher order regression models tend to fit the inputs and targets better. A typical example is depicted in figure 2

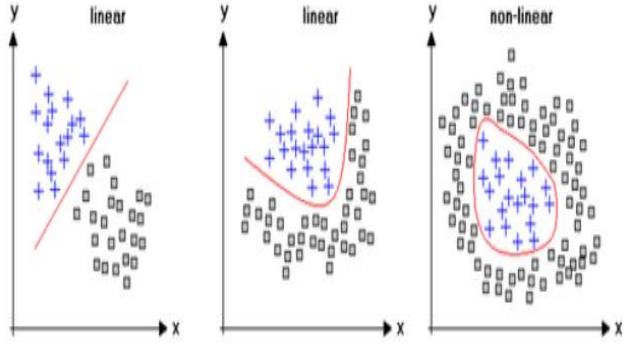

Fig. 2. Linear and Non-Linear Regression fitting.

Support Vector Regression:

This technique works on the principle of the hyper-plane which tries to separate the data in terms of 'n' dimensions where the order of the hyperplane is (n-1). Mathematically, if the data points or the data vector 'X' is m dimensional and there is a possibility to split the data into categories based on 'n' features, then a hyperplane of the order 'n-1' is employed as the separating plane [232. The name plane is a misnomer since planes corresponds to 2 dimensions only but in this case the hyper-plane can be of higher dimensions and is not necessarily a 2-dimensional plane. A typical illustration of the hyperplane used for SVM based classification is depicted in figure 3.

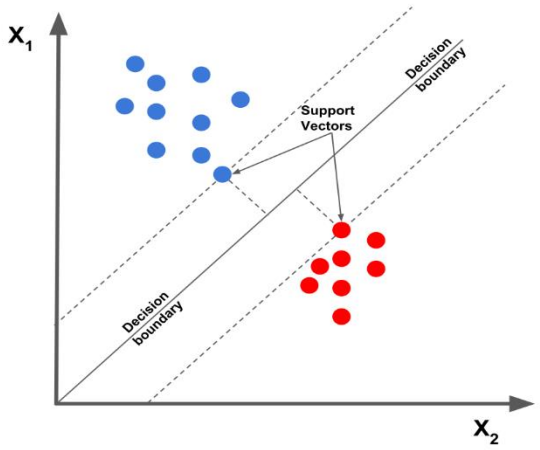

Fig. 3. Separation of data classes using SVM.

The selection of the hyperplane H is done on the basis of the maximum value or separation in the Euclidean distance d given by:

$$d = \sqrt{x_1^2 + \cdots \ldots \ldots x_n^2} \quad (3)$$

Here,
x represents the separation of a sample space variables or features of the data vector,
n is the total number of such variables
d is the Euclidean distance

The (n-1) dimensional hyperplane classifies the data into categories based on the maximum separation. For a classification into one of **'m'** categories, the hyperplane lies at the maximum separation of the data vector **'X'.** The categorization of a new sample **'z'** is done based on the inequality:

$$d_x^z = Min(d_{C1}^z, d_{C2}^z \ldots d_{C2=m}^z) \quad (4)$$

Here,
d_x^z is the minimum separation of a new data sample from **'m'** separate categories
$d_{C1}^z, d_{C2}^z \ldots d_{C2=m}^z$ are the Euclidean distances of the new data sample **'z'** from m separate data categories.

Neural Networks:

Owing to the need of non-linearity in the separation of data classes, one of the most powerful classifiers which have become popular is the artificial neural network (ANN) [23]. The neural networks are capable to implement non-linear classification along with steep learning rates. The neural network tries to emulate the human brain's functioning based on the fact that it can process parallel data streams and can learn and adapt as the data changes. This is done through the updates in the weights and activation functions. The mathematical model of the neural network is depicted in figure 4 [24].

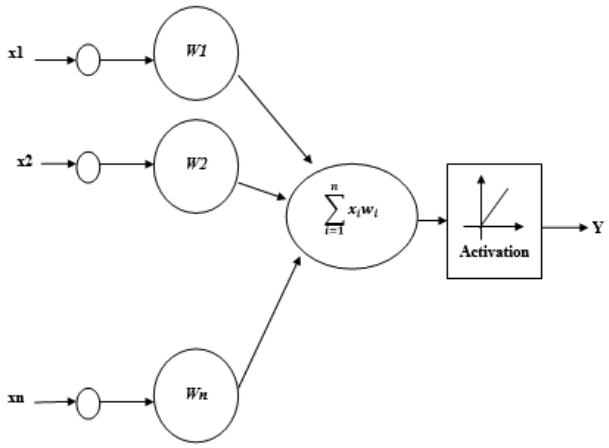

Fig. 4. Mathematical Model of Single Neuron.

The mathematical equivalent of an artificial neuron is depicted in figure 4 where the output can be given by:

$$y = f(\sum_{i=1}^{n} x_i w_i + b) \quad (5)$$

Here,
x denote the parallel inputs
y represents the output
w represents the bias
f represents the activation function

The neural network is a connection of such artificial neurons which are connected or stacked with each other as layers. The neural networks can be used for both regression and classification problems based on the type of data that is fed to them. Typically the neural networks have 3 major conceptual layers which are the input layer, hidden layer and output layer. The parallel inputs are fed to the input layer whose output is fed to the hidden layer. The hidden layer is responsible for analysing the data, and the output of the hidden layer goes to the output layer [25]. The number of hidden layers depends on the nature of the dataset and problem under consideration. If the neural network has multiple hidden layers, then such a neural network is termed as a deep neural network. The training algorithm for such a deep neural network is often termed as deep learning which is a subset of machine learning. Typically, the multiple hidden layers are responsible for computation of different levels of features of the data.

Decision Trees:

Decision trees are a set of probabilistic multivariate classifiers which recursive splitting is employed to classify a new data sample [26]. The splitting operation starts at the root node and terminates till no further splits are possible in the terminal nodes. The decision trees are depicted in figure 5, which exhibits the recursive classification method.

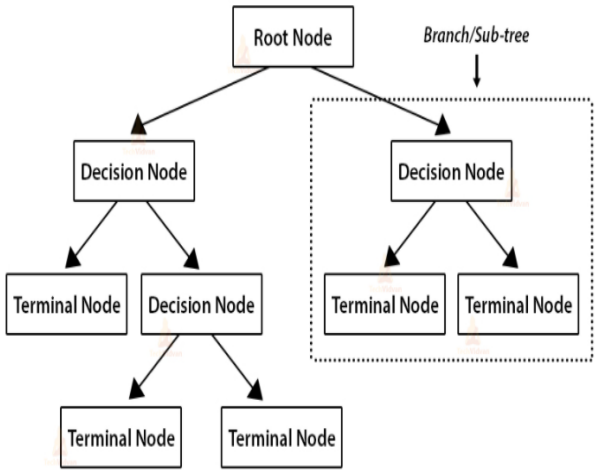

Fig. 5. A Typical Decision Tree Model

The proposed system is an amalgamation of the decision trees and the deep neural networks and is elaborately discussed in the next section.

III. Proposed Comparator Design

The proposed methodology presents an ensemble of the neural networks and decision trees to extract the attributes of both classifying paradigms which are:

1) Pattern recognition
2) Probabilistic Classification

The pattern recognition is performed using the gradient decent or scaled conjugate gradient. To update θ1 and θ2 values in order to reduce Cost function (minimizing MSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values

and then iteratively updating the values, reaching minimum cost. The main aim is to minimize the cost function J [27].

If the descent vector is given by g, then

$$g = f(J, w) \tag{6}$$

Here,
f stands for a function of.
w are the network weights.
J represents the Jacobian Matrix which is essentially the matrix containing the second order differentials of errors of the networks with respect to the weights. The Jacobian matrix is defined as:

$$J = \begin{bmatrix} \frac{\partial^2 e_1}{\partial w_1^2} & \cdots & \frac{\partial^2 e_1}{\partial w_m^2} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 e_n}{\partial w_1^2} & \cdots & \frac{\partial^2 e_n}{\partial w_m^2} \end{bmatrix} \tag{7}$$

Here,
The error e is computed as:

$$e = (y_i - y'_i) \tag{8}$$

Here,
y_i corresponds to the target variable (actual output).
y'_i represents the predicted output.

The ensemble approach presented in this paper comprises of a deep neural network based on the concept of steepest descent and the decision tress with regularization [28]. The steepest descent searches for a gradient matrix which can result in the steepest descent of the cost function or objective function during pattern recognition, and is typically chosen as the mean squared error (mse) defined as [29]:

$$mse = \frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2 \tag{9}$$

Here,
n are the number of samples in the error matrix.

The steepest descent tries to leverage a search vector which can render the steepest descent of all the possible error matrices corresponding to weight updates. The concept is depicted in figure 6.

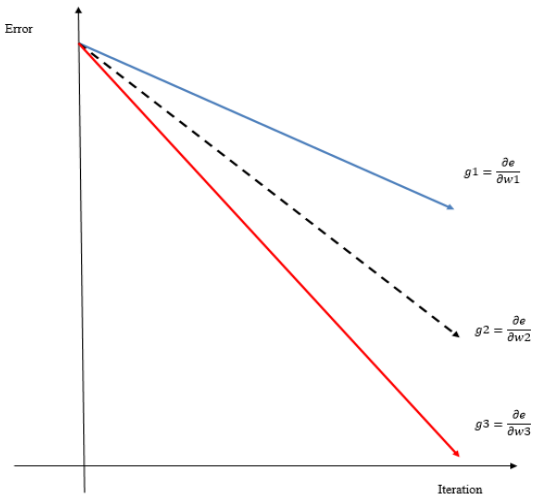

Fig. 6. Concept of Steepest Descent

From figure 6, it can be observed that the three gradients corresponding to three separate weight updates would render three different gradients given by g_1, g_2 and g_3 respectively. However, the gradient g_3 shows the quickest fall in the objective function which is the error. Thus at the beginning of each iteration, the search vector is to be evaluated which would point towards the direction of the steepest descent. Mathematically,

$$S_0 = -g_0 \tag{10}$$

Here,
S_0 corresponds to the initial values of the search vector at initialization of the weights. The weights are updated in each iteration as:

$$w_{k+1} = w_k + \mu_k g_k \tag{11}$$

Here,
w_{k+1} corresponds to the weights of the subsequent iteration.
w_k corresponds to the weights of the present iteration.
μ_k corresponds to the combination co-efficient.
k corresponds to the iteration number.

The search vector S for the iteration number $'k'$ is computed as:

$$S_k = -g_k + \beta_k S_{k-1} \tag{12}$$

The term β_k is computed as:

$$\beta_K = \frac{(|g_{k+1}|^2 - g_{k+1}{}^T g_k)}{g_k{}^T g_k} \tag{13}$$

Here,
$\frac{\partial e}{\partial w}$ represents the gradient of the error with respect to the weights and is represented as g_k.

It can be observed here that the search vector in each iteration takes into account the errors, thus the gradient of the previous iteration. This implies that the outputs of the previous iterations are considered as variables in the present iteration, thereby implementing "back-propagation", which happens to be one of the most effective learning techniques for deep neural networks (DeepNets) [30]. Subsequent to the pattern recognition, the final classification is based on the Decision Trees which is a probabilistic classifier.

The classification inaccuracy is computed based on the Gini's Index defined in equation 14 [31].

$$G = P(C) * [1 - P(C)] \tag{14}$$

Here,
G corresponds to Gini's Index.
$P(C)$ corresponds to the probability of a data sample corresponding to a defined or labelled category C.
$1 - P(C)$ corresponds to the probability of a data sample not belonging to a defined or labelled category C.

Binary trees are generated at each node corresponding to splits at each decision node and the overall split index is computed using equation 15.

$$G_O = \sum_{i=1}^{k} G_{N,i} w_i + G_{M,i} w_i \tag{15}$$

Here,
G_O corresponds to the overall Gini's index.
$G_{N,i}$ corresponds to the Tree's Index for left partitioning sub-tree.
$G_{M,i}$ corresponds to the Tree's Index for right partitioning sub-tree.
w_i corresponds to the weights of each split i.

The performance of the classifier is computed based on the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values which are used to compute the accuracy of the classifier, mathematically expressed in equation 16 [32].

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

IV. Simulation Results

The system is designed on Matrix Laboratory (MATLAB) 2020a. The system configurations are: Available RAM 16GB, processor intel i7, with a clock speed of 2.4GHz. The dataset has bene obtained from the UCI machine learning repository: https://archive.ics.uci.edu/ml/datasets/diabetes [33]

The parameters or features used in the system design are:

Age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, resting blood pressure, resting electrocardiogram (ECG), maximum heart rate, exercise induced angina (chest pain or discomfort), ECG parameters such as RR interval, QRS complex, slope of ST segment and the target is a vector containing three cases:

1) Reversibel health ailment.
2) Irreversible health ailment.
3) No health ailment or Noraml condition.

The iterations to convergence and the ensemble method used in depicted in figure 7.

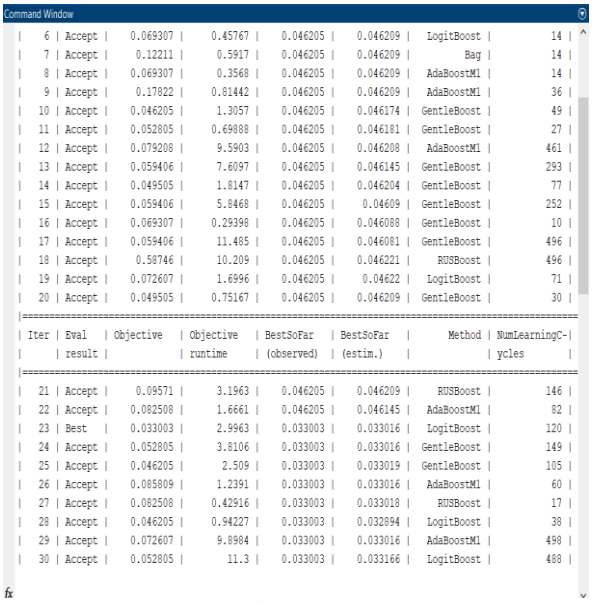

Fig. 7. Iterations to Convergence

The iterations to convergence and ensemble method is depicted in figure 7. The training stops for regression analysis in 2 cases:

1) Objective function stabilizes for validation check counts.
2) Maximum pre-defined iterations are over.

It can be observed that the system converges in 30 iterations.

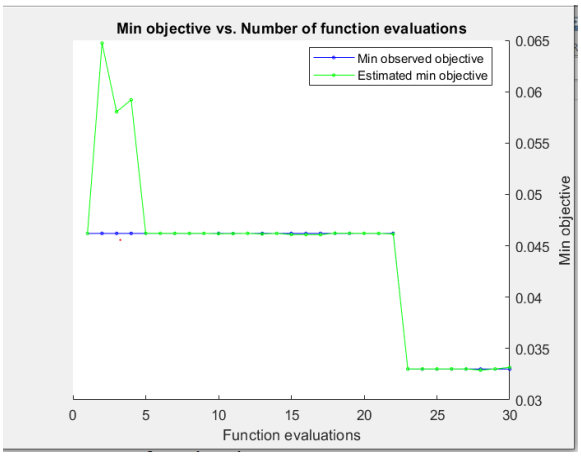

Fig. 8. Variation of Objective Function with respect to function evaluations.

The variation of the objective function with respect to the number of function evaluations is depicted in figure 8. It can be observed that as the function evaluations increases, the objective function shows a plummeting nature stabilizes eventually.

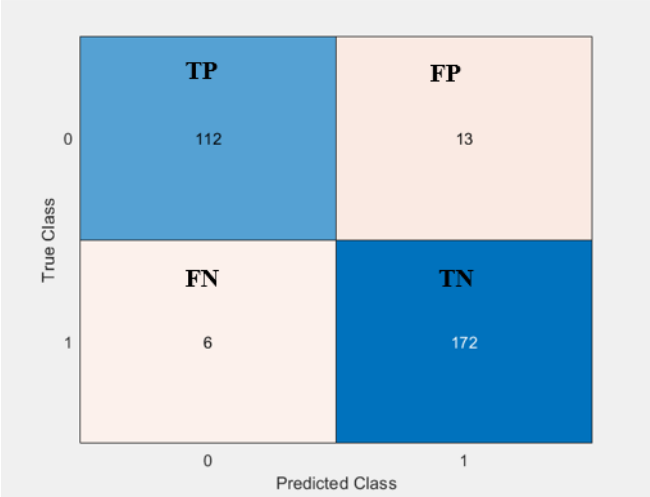

Fig. 9. Confusion matrix corresponding to classification.

Figure 9 depicts the confusion matrix corresponding to the predictive classification of the model. A total of 303 samples are tested with a training to testing division ration of 75:23 adopted in this experiment. The accuracy of classification is computed as:

$$\boldsymbol{Ac} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} = \frac{172+112}{172+112+13+6} = \mathbf{93.72\%}$$

For the ease of use of any end user, or for building a prototype model for the proposed system, a simple graphical user interface (GUI) is creased in MATLAB with the inputs to be fetched from the values entered by the user. The GUI designed in depicted in figure 10.

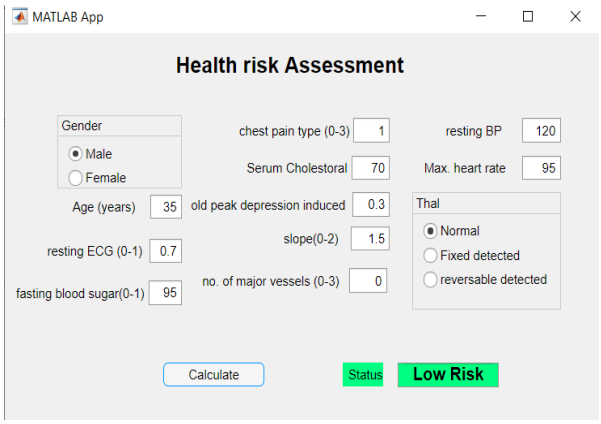

Fig. 10. Graphical User Interface (Prototype) designed.

The designed GUI prototype can be used for prototype testing and can be later extended to web or mobile applications and deployed. The summary of the obtained results for the proposed system is tabulated in table I.

TABLE I

SUMMARY OF RESULTS

S.No.	Parameter	Value
1.	Number of attributes (features)	13
2.	Categories of Target Vector	3
3.	Classifier Type	Ensemble
4.	Regression Analysis	Steepest Descent
5.	Number of Hidden Layers	5
6.	Classification Layer	Decision Trees
7.	Iterations to Convergence	30
8.	Classification Accuracy	93.72%

The results show that the proposed system attains a classification accuracy of 93.72% for used dataset. The prototype can be extended for a much larger disease dataset with a multitude of parameters (not necessarily numeric in nature). A comparison with existing work is cited in table II.

TABLE II

COMPARISON WITH EXISTING WORK

S.No.	Authors and Reference	Performance
1.	J.Archenaa, & E.A.Mary Anita [2]	Highest Accuracy of 89% achived.
2.	P. Chiang and S. Dey [3]	Mean Absolute Error (MAE) of 14.5% achieved giving an accuracy of 85.5%
3.	X. Li and J. Li [5]	Accuracy of 86.7% achieved
4.	**Proposed Approach**	**Accuracy of 93.72% achieved.**

It can be observed that the proposed approach outperforms existing techniques in terms of classification accuracy.

V. CONCLUSION

It can be concluded that the necessity of automated tools for health risk estimation is necessary keeping in mind the lifestyle changes risks at earlier ages. This paper presents an ensemble learning based approach for health risk estimation. In this classifier design, the training data which is labelled is applied to the algorithm for pattern analysis which assumes the data events in classes to be true. Based on the analyzed patterns, the new data sample's probability to belong to a specific category is evaluated. The performance of the system has been evaluated in terms of the true positive, true negative, false positive and false negative rates, Based on these metrics, the accuracy of the system has bene evaluated. The experimental results show that the proposed system attains a classification accuracy of 93.7%. The number of iterations are also less which are 30 for convergence. The designed GUI can also be used for prototype testing. Thus, the proposed system effectively predicts health risks based on medical record datasets which relatively high accuracy. Future research directions may include using clustering to find regular patterns in disease categories. Data optimization and dimensional reduction techniques such as principal component analysis (PCA) can also be tried out in conjugation with the existing system.

REFERENCES

[1] V Ilakkuvan, A Johnson, AC Villanti, WD Evans, "Patterns of social media use and their relationship to health risks among young adults", Journal of Adolescent Health, Elsevier, 2019, vol. 64, no. 2, pp. 158-164. https://doi.org/10.1016/j.jadohealth.2018.06.025

[2] J.Archenaa, & E.A.Mary Anita. (2017). Health Recommender System using Big data analytics. Journal of Management Science and Business Intelligence, vol.2, no.2, pp. 17–24. http://doi.org/10.5281/zenodo.835606

[3] P. Chiang and S. Dey, "Personalized Effect of Health Behavior on Blood Pressure: Machine Learning Based Prediction and Recommendation," 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom), 2018, pp. 1-6, doi: 10.1109/HealthCom.2018.8531109.

[4] E. Sezgin and S. Özkan, "A systematic literature review on Health Recommender Systems," 2013 E-Health and Bioengineering Conference (EHB), 2013, pp. 1-4, doi: 10.1109/EHB.2013.6707249.

[5] X. Li and J. Li, "Health Risk Prediction Using Big Medical Data - a Collaborative Filtering-Enhanced Deep Learning Approach," 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom), 2018, pp. 1-7, doi: 10.1109/HealthCom.2018.8531143.

[6] World Economic Forum Report, https://reports.weforum.org/global-risks-report-2020/false-positive/

[7] N. S. Rajliwall, R. Davey and G. Chetty, "Machine Learning Based Models for Cardiovascular Risk Prediction," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), 2018, pp. 142-148, doi: 10.1109/iCMLDE.2018.00034.

[8] AC Dimopoulos, M Nikolaidou, FF Caballero, "Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk", BMC Med Res Methodol 18, Springer 2018, vol.18, no. 179. https://doi.org/10.1186/s12874-018-0644-1.

[9] A Maxwell, R Li, B Yang, H Weng, A Ou, H Hong, "Deep learning architectures for multi-label classification of intelligent health risk prediction", BMC Bioinformatics Springer 2017, , vol.18, no. 523, https://doi.org/10.1186/s12859-017-1898-z

[10] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[11] B. Nithya and V. Ilango, "Predictive analytics in health care using machine learning tools and techniques," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), 2017, pp. 492-499, doi: 10.1109/ICCONS.2017.8250771.

[12] EG Ross, NH Shah, RL Dalman, KT Nead, "The use of machine learning for the identification of peripheral artery disease and future mortality risk", Journal of Vascular Surgery, Elsevier 2016,vol. 64, no. 5, pp. 1515-1522.

[13] D. LaFreniere, F. Zulkernine, D. Barber and K. Martin, "Using machine learning to predict hypertension from a clinical dataset," 2016 IEEE Symposium Series on Computational Intelligence (SSCI), 2016, pp. 1-7, doi: 10.1109/SSCI.2016.7849886.

[14] D Tay, CL Poh, RI Kitney," A novel neural-inspired learning algorithm with application to clinical risk prediction", Journal of Biomedical Informatics, Elsevier 2015, vol. 54, pp. 305-314

[15] K. Sowjanya, A. Singhal and C. Choudhary, "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices," 2015 IEEE International Advance Computing Conference (IACC), 2015, pp. 397-402, doi: 10.1109/IADCC.2015.7154738.

[16] LM Hlaváč, D Krajcarz, IM Hlaváčová, S Spadło, "Precision comparison of analytical and statistical-regression models for AWJ cutting", Precision Engineering, Elsevier 2017,vol. 50, pp. 148-159

[17] C Bergmeir, RJ Hyndman, B Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction", Computational Statistics & Data Analysis, Elsevier 2018, vol.120, pp. 70-83.

[18] S Bandaru, AHC Ng, K Deb, "Data mining methods for knowledge discovery in multi-objective optimization: Part A-Survey", Expert Systems with Applications, Elsevier 2017, vol. 70, no.15 pp.139-159

[19] A. Karpatne, I. Ebert-Uphoff, S. Ravela, H. A. Babaie and V. Kumar, "Machine Learning for the Geosciences: Challenges and Opportunities," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 8, pp. 1544-1554, 1 Aug. 2019, doi: 10.1109/TKDE.2018.2861006.

[20] V. Sze, Y. Chen, T. Yang and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," in Proceedings of the IEEE, vol. 105, no. 12, pp. 2295-2329, Dec. 2017, doi: 10.1109/JPROC.2017.2761740.

[21] W. Zhou, J. Li, M. Zhang, Y. Wang and F. Shah, "Deep Learning Modeling for Top-N Recommendation With Interests Exploring," in IEEE Access, vol. 6, pp. 51440-51455, 2018, doi: 10.1109/ACCESS.2018.2869924.

[22] Machine Learning Notes: Stanford University: http://cs229.stanford.edu/materials.html.

[23] Haykin S, "Neural Networks and Learning Machines", 3rd Edition, Pearson Publications.

[24] Hagan M, "Neural Network Design", 2nd Edition, Cengage Publication.

[25] K. D. Humbird, J. L. Peterson and R. G. Mcclarren, "Deep Neural Network Initialization With Decision Trees," in IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 5, pp. 1286-1295, May 2019, doi: 10.1109/TNNLS.2018.2869694.

[26] A. Segatori, F. Marcelloni and W. Pedrycz, "On Distributed Fuzzy Decision Trees for Big Data," in IEEE Transactions on Fuzzy Systems, vol. 26, no. 1, pp. 174-192, Feb. 2018, doi: 10.1109/TFUZZ.2016.2646746.

[27] A Wanto, M Zarlis, D Hartama, "Analysis of Artificial Neural Network Backpropagation Using Conjugate Gradient Fletcher Reeves in the Predicting Process", IOP Science, vol. 930, pp.1-7

[28] G Yuan, Z Wei, Y Yang, "The global convergence of the Polak–Ribière–Polyak conjugate gradient algorithm under inexact line search for nonconvex functions", Journal of Computational and Applied Mathematics, Elsevier, 2019, vol. 362, 15 pp.262-275.

[29] D Di Serafino, V Ruggiero, G Toraldo, "On the steplength selection in gradient methods for unconstrained optimization", Applied Mathematics and Computation, Elsevier 2018, vol. 318, pp. 176-195.

[30] H. Su, G. Li, D. Yu and F. Seide, "Error back propagation for sequence training of Context-Dependent Deep NetworkS for conversational speech transcription," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6664-6668.

[31] BK Das, HS Dutta, "GFNB: Gini index–based Fuzzy Naive Bayes and blast cell segmentation for leukemia detection using multi-cell blood smear images", Journal of Medical & Biological Engineering & Computing, Springer 2020, vol.58, pp. 2789–2803.

[32] JP Kandhasamy, S Balamurali, "Performance analysis of classifier models to predict diabetes mellitus", Procedia Computer Science, Elsevier, 2015, vol.47, pp. 45-51.

[33] https://archive.ics.uci.edu/ml/datasets/diabetes

Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 24 (2021), No. 5, pp. 1179–1187
DOI : 10.1080/09720529.2021.1936900

Taylor & Francis
Taylor & Francis Group

Hybrid symmetric cryptography approach for secure communication in web application

Vivek Kapoor *
Rati Gupta †
Department of Information Technology
Institute of Engineering & Technology
Devi Ahilya Vishwavidyalaya
Indore
Madhya Pradesh
India

Abstract

Cryptography is a science that studies methods so that data or messages remain safe when sent, from the sender to the recipient without experiencing interference from unauthorized parties. The cryptographic system used is a combination of symmetrical cryptographic algorithms, and asymmetric cryptographic algorithms also called hybrid cryptosystems. This research work tries to explore the need of integration of different encryption algorithms and enhance the strength of encryption and complexity. This work observes that integrity check algorithm can also be used to generate key of asymmetric key cryptography and they can help to enhance power of confidentiality. In this paper, we present a new hybrid by using Blowfish, RC6 and RSA algorithms. Within this, the data is encrypted through symmetric algorithm using RC6 and blowfish algorithm and key is encrypted using an asymmetric algorithm called RSA. The complete solution was implemented using Java technology and it was evaluated based on computation time.

Subject Classification: *68P25 (Data Encryption).*

Keywords: *Cryptography, Hybrid cryptography, RSA, RC6, SHA-1.*

1. Introduction

The reach of communication network and internet is increasing day by day. In computer network independent nodes or machines are

*E-mail: vkapoor@ietdavv.edu.in (Corresponding Author)
†E-mail: rati.gupta01@gmail.com

inter connected and communicate with each other by a set of mutually acceptable rules, which are known as protocols. Data often travels in these unsecure medium in an unprotected way. It is desired to have a secure communication and protection of data with all security features to be addressed [4, 5, 6, 7] . With an increase in processing power and technology, there is a variety of attack which can be attempted to steal, modify and fabricate the data travelling through this unsecure medium. In present scenario traditional procedures i.e single encryption process are not adequate to address the problem which we are facing [8, 9, 10]. Our experimental work in this paper uses two types of cryptography i.e symmetric and asymmetric along with secure hash algorithm (SHA) to achieve three main security features such as confidentiality, integrity and authentication in a single shot. Combination of these two types of cryptographic procedures is popularly known as hybrid cryptography scheme that dissuades an attacker to steal any meaningful information from the data that travels in unsecure medium and achieve best of both the world features. Our proposed system helps us to achieve these security features mentioned above.

In our work we uses two symmetric key algorithm i.e Blowfish and RC6 for encryption of plain text. Data is divided into two chunks i.e odd and even, even chunks are encrypted using RC6 algorithms and odd chunks are encrypted using Blowfish algorithm. In addition to it hash is applied to the plain text to get message digest which act as a key for these above symmetric key operations. For the security of this symmetric key, public key cryptography is roped in to encrypt this key which in turn reduces the chances of stealing of the key. Symmetric key algorithms used here are fast in there working, less computational power is needed for encryption and decryption, less amount of volatile memory is needed. Cipher text obtained is of compact in size and secure in nature.

This paper is written as follows. Section 2 is related to the examination of most related work. Section 3 addresses to apply the huge amount of man hours and intelligent thinking to do this work. Section 4 proposes the solution and process to solve this problem. In section 5 experimental results are discussed. Finally, in section 6 we draw conclusions and future work.

2. Literature Review

Rahul Yadav et al. [1, 6] in his work provides an efficient hybrid encryption/ decryption scheme to enhanced the security features in the

cyber space. Proposed scheme is implemented and results are shown. Here hybrid cryptographic system uses Data Encryption Standard (DES) as symmetric key operation and RSA algorithm as asymmetric key operation along with secure hash algorithm (SHA) to provide security feature. Results in terms of space and time complexity is determined and compared with the results obtained just by using RSA cryptography.

P.Chinnasamy et al. [2] in his proposed work provided security to the electronic medical data to protect it from unauthorized access by using hybrid cryptography. This Electronic Health Records (EHR) includes various types of x ray images, reports etc and patient current and sensitive health information. Security system should be able to chech authentication of the person accessing it along with the integrity and confidentiality features should be addressed too. Popularly cloud based services are roped in to store these type of data. Hybrid cryptography is the only solution for this type of problem. P.Chinnasamy proposed model is less time consuming and moe secure when compared with single encryption scheme such as AES and Blowfish.

M.Harini et al. [3] in his novel hybrid cryptographic scheme uses combination of AES, RSA and MD5 algorithms. Results shows that enhanced security features. Proposed system is capable to thwart all possible basic attacks that is being done on single cryptographic mechanism.

3. Rationale

Problem Proper security should be provided to all type of data which is consider to be sensitive, while travelling through unsecure channel. Confidentiality is an important issue. Another issue is trust. This feature is generally provided in traditional systems by roping a third party. Inclusion of third party raises security concerns. Another issue which comes into picture is of key exchange and key stealing. To address all these issues and to prevent any malicious activity to take place we proposed a hybrid cryptography mechanism as stated below.

Here in our system efficiency of public key cryptography is combined with efficiency of symmetric key cryptography. Public key cryptosystem is simple in their working as problem of key exchange is solved. But internal working of encryption and decryption is complicated and computationally intensive as compare with symmetric key algorithm. For encryption/decryption of large data, public key cryptography in advised at all. So to achieve best of both the worlds feature hybrid crypto systems

are used. Here data is encrypted by using more efficient symmetric key cryptography as data is of large size. Then symmetric key is encrypted by using public key cryptography which is of very small size. Thus in this way more efficient symmetric key system is combined with more convenient asymmetric key cryptography, which solves the problem of key exchange also. It is also found that cipher text obtained in public key cryptography is of more size, which adds more overhead on the network bandwidth when transmitted.

Thus best of both the worlds features are obtained by using hybrid cryptography which encapsulates the cipher text and encrypted symmetric key in form of a digital envelope.

4. Methodology

Many of the cryptographic techniques is used when dealing with the issues of security of the confidential information. Now-a-days, data is growing with great speed because of the use of social networking and e-commerce sites. With every single minute data is growing and its management is also important for its security. Many of the cryptographic techniques are there which deals with the security of tera bytes and peta bytes of data. Data is in multiple format like image, text, audio, video etc. Traditional cryptographic technique does not assure about accurate and secure security of data.

Following steps are suggested to provide safety during storage, encryption process is followed as:

Step 1: Plain text is taken and is passed to check integrity using SHA-1. Hash value thus generated is used as a key for blowfish and RC6 algorithm.

Step 2: After integrity calculation data is divided into chunks.

Step 3: BLOWFISH and RC6 algorithm is used to encrypt chunks. Odd chunks is encrypted using blowfish algorithm and even chunks is encrypted using RC6 algorithm thus produce cipher text.

Step 4: Afterwards, key is encrypted using RSA algorithm.

Step 5: Cipher text and ciphered key are combined to generated combined message and this combine message is forwarded directly to the network.

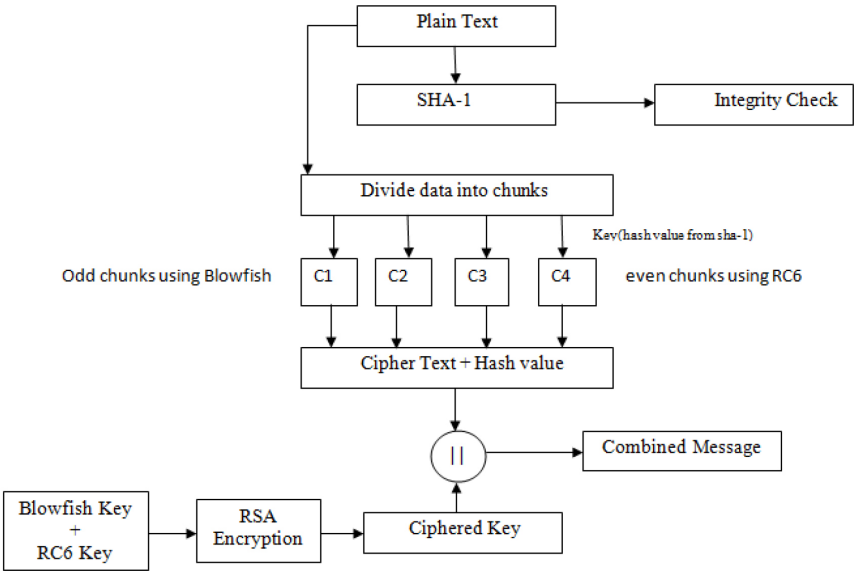

Figure 1

Encryption process at sender's end

Decryption process at receiver end is followed as:

Step 1: Receiver receives the combined message and separates the ciphered key and ciphered text.

Step 2: Decryption is performed on ciphered key using receiver's private key i.e RSA algorithm to generate key.

Step 3: Similarly, for ciphered text decryption is performed using the generated key and it will convert cipher text into plain text.

Step 4: Hash value of whole file will be calculated to measure integrity factor of content using SHA-1.

Step 5: This generates hash value and validate the message.

Step 6: If true then the digest message will be accepted.

Step 7: If false then the message is rejected and the message is not the original message.

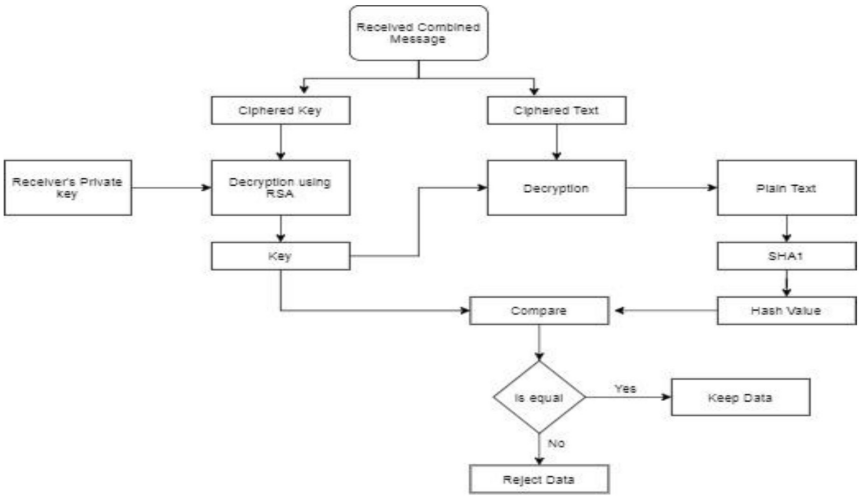

Figure 2

Decryption process at receiver's end

5. Experimental Analysis

The complete work has been implemented using Java technology to realize the significance of proposed solution. A cloud based Java application has been developed and evaluated on basis of milliseconds for every stage of process. A snapshot to represent the results is shown below

File Statistics :-

File ID	File Name (KB)	File Size (KB))	Encrypted File Size (KB)	SHA1 Hash Size (Byte)	Number of Chunks
C1-F10	500.txt	499.23	682.54	32	999

Key Statistics :-

Key Size (Bit)
1024

Time Statistics :-

RC6 Crypto Time (ms)	BlowFish Crypto Time (ms)	SHA1 Generation Time (ms)	File Chunking Time (ms)	Total Process Time (ms)
29.148099999999978	66.6264	8.5327	1836.7615	1941.0687

Figure 3

Result in Table Format for .txt File

For deep analysis, a comparison has been made based on total computation time. Different memory size input files have been considered as input to identify the time and effort. Two types of input have been considered.

(1)Text File (2) PDF File

This work considers 1KB, 10KB, 100KB, 200KB, 300KB, 400KB; 500KB file size as input and evaluate computation time for individual stage, RC6, Blowfish, SHA-1 and Total Computation Time. A comparison table for total computation time has been shown below

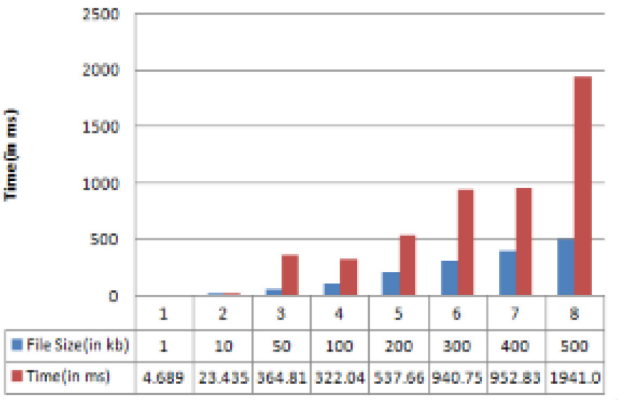

Figure 4

Performance Comparison of Text File (Encryption)

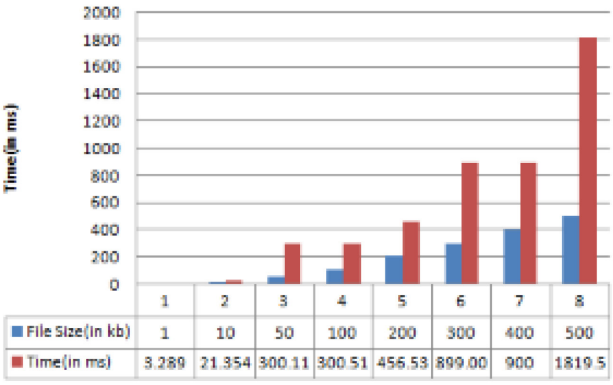

Figure 5

Performance Comparison of .txt File (Decryption)

6. Conclusion and Future work

The hybrid encryption/decryption process is the combination of two different cryptographic systems to enhance the security with less complexity is proved in our work. This excellent scheme uses combination of symmetric and asymmetric key cryptography to achieve best of both the worlds' features. The present demand of having compact size cipher text is also achieved.

Here the integrity feature along with confidentiality is appropriately addressed This mechanism is suitable to work for different types of data. It can be successfully applied to any cloud based storage system which stores sensitive information. Our proposed system fulfill the sensitivity feature of integrity and confidentiality at its best as shown in results. Conclusion is that hybrid cryptosystems developed by us is more efficient and requires less resourses.

References

[1] Dr. Vivek Kapoor, Rahul Yadav, "A Hybrid Cryptography Technique for Improving Network Security ", *International Journal of Computer Applications* Volume 141 – No.11, May 2018

[2] P.Chinnasamy, P.Deepalakshmi " Design of Secure Storage for Health-care Cloud using Hybrid Cryptography" Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)

[3] M. Harini, K. Pushpa Gowri, C. Pavithra, M. Pradhiba Selvarani "A Novel Security Mechanism Using Hybrid Cryptography Algorithms"

[4] Dian Rachmawati, Mohammad Andri Budiman, Muhammad Ishan Wardhono "Hybrid Cryptosystem for Image Security by Using Hill Cipher 4x4 and ElGamal Elliptic Curve Algorithm" 2018 *IEEE International Conference on Communication, Networks Satellite* (Comnetsat)

[5] Dr. Mahmood Zaki Abdullah, Zinah Jamal Khaleefah "Design and implement of a hybrid cryptography textual system" ICET2017, Antalya, Turkey.

[6] Dr. Vivek Kapoor, Rahul Yadav, "A Hybrid Cryptography Technique to Support Cyber Security Infrastructure", *International Journal of Advanced Research in Computer Engineering & Technology* (IJARCET) Volume 4 Issue 11, November 2015

[7] Punam V. Maitri, Aruna Verma "Secure File storage in Cloud Computing using Hybrid Cryptography Algorithm", IEEE WiSPNET 2016 conference

[8] Jayraj Gondaliya, Jinisha Savani "Hybrid Security RSA Algorithm in Application of Web Service", 2018 1st International Conference on Data Intelligence and Security

[9] Chitra Biswas, Udayan Das Gupta "An Efficient Algorithm for Confidentiality, Integrity and Authentication Using Hybrid Cryptography and Steganography", 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE),7-9 February, 2019.

[10] Supriya S. Sonawane, N. M. Shahane, "High Capacity Data Embedding Technique for Separable Encrypted Data Embedding in Encrypted Image".

VARDHAMAN COLLEGE OF ENGINEERING

(AUTONOMOUS)

Affiliated to JNTUH, Approved by AICTE, Accredited by NAAC with A++ Grade, ISO 9001:2015 Certified

Kacharam, Shamshabad, Hyderabad - 501218, Telangana, India

ICAMS2021

4th International Congress on Advances in Mechanical Sciences

CERTIFICATE OF APPRECIATION

This is to certify that

Mr. Vijay Kumar Karma

has presented a paper entitled

"Analysis of Effects of No of Teeth and Pitch Error on Transmission Error using Interference Volume Method "

in ICAMS-2021 organized by the Department of Mechanical Engineering held during November 26th & 27th 2021.

Convener

Conference Chairman

Patron

PDEU PANDIT DEENDAYAL ENERGY UNIVERSITY
Formerly Pandit Deendayal Petroleum University

IEEE GUJARAT SECTION

CERTIFICATE OF PARTICIPATION

1st IEEE International Conference on Artificial Intelligence and Machine Vision (AIMV) - 2021

This is to certify that **BABITA PATHIK** has presented a paper titled **Analysis of Effort Estimation for Test Suite using Control Graph** in the 1st IEEE International Conference on Artificial Intelligence and Machine Vision (AIMV) - 2021, organized by the Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar, during 24th - 26th September, 2021.

Dr. Rajeev Kumar Gupta
Publication Chair, AIMV-2021

Dr. Samir B. Patel
General Chair, AIMV-2021

Dr. Santosh Kumar Bharti
General Chair, AIMV-2021

IEEE Student Branch @PDPU

Department of Science & Technology Government of Gujarat

MTT-S IEEE MICROWAVE THEORY & TECHNIQUES SOCIETY

SOT SCHOOL OF TECHNOLOGY

An Ensemble Machine Learning Based Approach for Health Risk Prediciton

Amit Yadav
Department of Computer Science and Engineering
IET, DAVV
Indore, India
amit.sunoracle@gmail.com

Amit Kumar Mittal
Department of Computer Science and Engineering
IET, DAVV
Indore, India
amittal@ietdavv.edu.in

***Abstract*—Chronic health risks have risen among young individuals due to several factors such as sedentary lifestyle, poor eating habits, sleep irregularities, environmental pollution, workplace stress etc. The problem seems to be more menacing in the near future, with the exacerbation of lifestyle conditions and unforeseen breakout of pandemics such as COVID-19. One possible solution is thus to design health risk prediction systems which can evaluated some critical features of parameters of the individual and then be able to predict possible health risks. As the data shows large divergences in nature with non-correlated patterns, hence choice of machine learning based methods becomes inevitable to design systems which can analyze the critical factors or features of the data and predict possible risks. This paper presents an ensemble approach for health risk prediction based on the steepest descent algorithm and decision trees. It is observed that the proposed work attains a classification accuracy of 93.72%. A simple graphic user interface has also been created for the ease of use and interaction and for prototype testing.**

Keywords— Information and Communications Technology (ICT), Automated Health Risk Prediction, Ensemble Classifiers, Deep Neural Networks, Decision Trees, Classification Error, Accuracy.

I. INTRODUCTION

The sudden and unforeseen outbreak of Covid-19 has forced people worldwide to re-evaluate their lifestyle and health condition. With increase in the sedentary lifestyle of people around the globe, different health risks are affecting people worldwide [1]. While life expectancy has increased, but increasing health risks can be seen throughout the world. The majority of the population are pre-occupied in sedentary and non-active vocations neglecting the health markers which has seen an earlier precedence of health risks in people. The major reasons happen to be [2]:

1) Sedentary Lifestyle
2) Lack of Physical Exercise.
3) Poor Food Choices.
4) Environmental Pollution.
5) Climate Change
6) Stress in everyday life etc.

Hence, an urgent need to address the health risks has become imperative. However, the cost of healthcare medications is also continuing to rise. It is the government's job to have an efficient, cost-effective medical system which can cater to the needs of all socio-economic sections of the society [3]-[4]. However, medical care sees a continuous upsurge in costs and sudden pandemics put additional burden on the existing medical infrastructure. Thus, in practical scenarios, rendering personalized medical care to individuals belonging to different strata of the society is challenging as different individuals have different requirements and different prevailing health conditions and the medical infrastructure is already strained to the limit. Hence, it is necessary to look for alternatives which can address the present issues. One such avenue, which has been garnering a lot of interest among researchers, is Information and Communications Technology (ICT) in healthcare [5]. The objective of this paper is to identify the major health risks globally and design a recommendation system which can indicate possible health risks in the future. For this purpose, the major health risks causing mortalities or serious ailments is analyzed and depicted in figure 1.

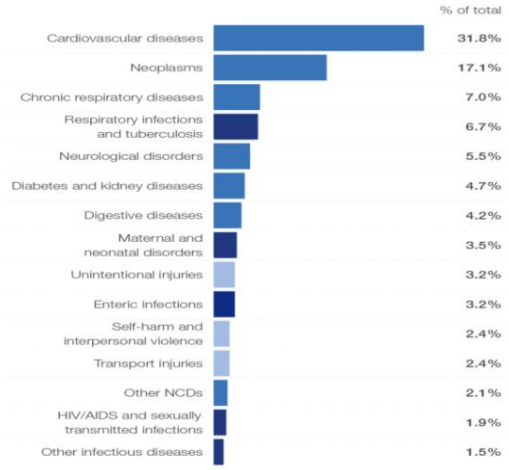

Fig. 1. Global Health Risk Analysis: 2019

(Source: World Economic Forum, [6])

The analysis of figure 1 helps us in deciding upon the data collection pertaining to symptoms, range of medical parameters and positivity rate of the major causes of health ailments. The data collection and labelling would subsequently lead to the design of an effective automated tool for health risk prediction [7]-[8].

The paper is organized as follows: Section I introduces the necessity of designing an automated health risk prediction system, the motivation and objectives behind the work. Section II presents the existing methods for designing such recommendation systems. Section III presents the proposed methodology for designing the system. Section IV presents

the simulation results. Section V presents conclusion and directions for future work.

II. EXISTING METHODS

The medical system's digitization has led to the influx of massive amounts of medical data [9]-[10]. This data can help medical care institutions to improve the efficiency of the health system, enhance the quality of healthcare and minimize healthcare costs. Moreover, the data available can also be utilized for research in non-conventional and upcoming fields such as machine learning and deep learning to come up with tools which can aid medical care [11]-[12]. The advent of cheap smart phones and internet services, Healthcare Information Systems (HIS) and Electronic Medical Records (EMRs) have become easily accessible far and wide [13]. However, accessing medical data, analyzing medical data and coming up an accurate predication system is complex due to the following reasons [14]:

1) Medical records may be extremely unstrcutured in nature with large divergences in data files such as text, audio, images, video etc.

2) Medical data, in general is ethnically diverse and its is clahhenging to extract intelligeble features (parameters) which can cater to a large diaspora of patients globally.

3) Medical data may be distributed or sparse. Data might have huge amounts of missing values owing to different human considerations.

Several techniques have been investigated to find patterns in medical data and thus create a mapping among the cases, symptoms and recorded test values [15]. Machine learning has been commonly used in numerous healthcare systems, such as medical imaging risk identification, diagnosis of illness, and prediction of health status from electronic health records [16]-[17]. Machine learning offers a way to automatically identify trends and predict results [18]-[19]. There are several current experiments on various types of electronic medical data on data mining and data analytics. Machine learning based classifiers are typically much more accurate and faster compared to the conventional classifiers [20]. They render more robustness to the system as they are adaptive and can change their characteristics based on the updates in the dataset [21]. The common classifiers which have been used for the classification of glaucoma cases are:

Regression Models:

In this approach, the relationship between the independent and dependent variable is found utilizing the values of the independent and dependent variables. The most common type of regression model can be thought of as the linear regression model which is mathematically expressed as [22]:

$$y = \theta_1 + \theta_2 x \quad (1)$$

Here,
x represents the state vector of input variables
y represents the state vector of output variable or variables.
Θ1 and Θ2 are the co-efficients which try to fit the regression learning models output vector to the input vector.
Often when the data vector has large number of features with complex dependencies, linear regression models fail to fit the input and output mapping. In such cases, non-linear regression models, often termed as polynomial regression is used. Mathematically, a non-linear or higher order polynomial regression models is described as:

$$y = \theta_0 + \theta_1 x^3 + \theta_2 x^2 + \theta_3 x \quad (2)$$

Here,
x is the independent variable
y is the dependent variable
$\theta_1, \theta_2 \ldots . \theta_n$ are the co-efficients of the regression model.
Typically, as the number of features keep increasing, higher order regression models tend to fit the inputs and targets better. A typical example is depicted in figure 2

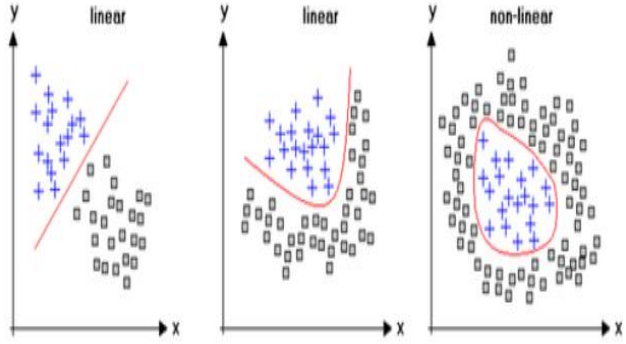

Fig. 2. Linear and Non-Linear Regression fitting.

Support Vector Regression:

This technique works on the principle of the hyper-plane which tries to separate the data in terms of 'n' dimensions where the order of the hyperplane is (n-1). Mathematically, if the data points or the data vector 'X' is m dimensional and there is a possibility to split the data into categories based on 'n' features, then a hyperplane of the order 'n-1' is employed as the separating plane [232. The name plane is a misnomer since planes corresponds to 2 dimensions only but in this case the hyper-plane can be of higher dimensions and is not necessarily a 2-dimensional plane. A typical illustration of the hyperplane used for SVM based classification is depicted in figure 3.

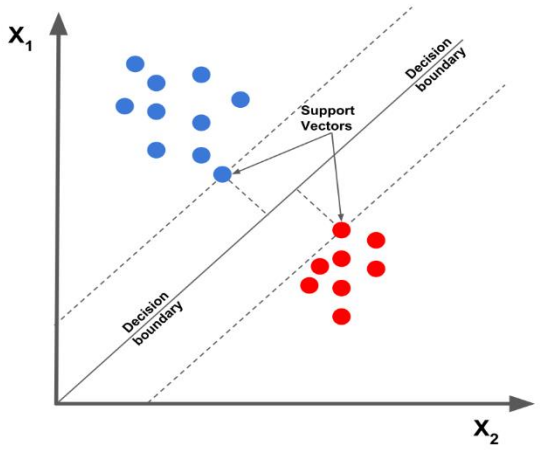

Fig. 3. Separation of data classes using SVM.

The selection of the hyperplane H is done on the basis of the maximum value or separation in the Euclidean distance d given by:

$$d = \sqrt{x_1^2 + \cdots \ldots \ldots x_n^2} \quad (3)$$

Here,
x represents the separation of a sample space variables or features of the data vector,
n is the total number of such variables
d is the Euclidean distance

The (n-1) dimensional hyperplane classifies the data into categories based on the maximum separation. For a classification into one of **'m'** categories, the hyperplane lies at the maximum separation of the data vector **'X'.** The categorization of a new sample **'z'** is done based on the inequality:

$$d_x^z = Min(d_{C1}^z, d_{C2}^z \ldots d_{C2=m}^z) \quad (4)$$

Here,
d_x^z is the minimum separation of a new data sample from **'m'** separate categories
$d_{C1}^z, d_{C2}^z \ldots d_{C2=m}^z$ are the Euclidean distances of the new data sample **'z'** from m separate data categories.

Neural Networks:

Owing to the need of non-linearity in the separation of data classes, one of the most powerful classifiers which have become popular is the artificial neural network (ANN) [23]. The neural networks are capable to implement non-linear classification along with steep learning rates. The neural network tries to emulate the human brain's functioning based on the fact that it can process parallel data streams and can learn and adapt as the data changes. This is done through the updates in the weights and activation functions. The mathematical model of the neural network is depicted in figure 4 [24].

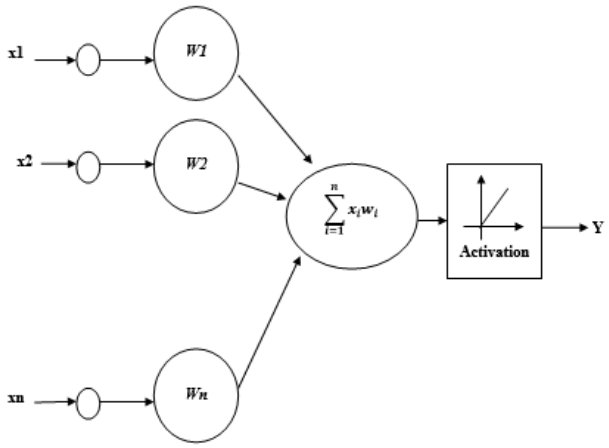

Fig. 4. Mathematical Model of Single Neuron.

The mathematical equivalent of an artificial neuron is depicted in figure 4 where the output can be given by:

$$y = f(\sum_{i=1}^{n} x_i w_i + b) \quad (5)$$

Here,
x denote the parallel inputs
y represents the output
w represents the bias
f represents the activation function

The neural network is a connection of such artificial neurons which are connected or stacked with each other as layers. The neural networks can be used for both regression and classification problems based on the type of data that is fed to them. Typically the neural networks have 3 major conceptual layers which are the input layer, hidden layer and output layer. The parallel inputs are fed to the input layer whose output is fed to the hidden layer. The hidden layer is responsible for analysing the data, and the output of the hidden layer goes to the output layer [25]. The number of hidden layers depends on the nature of the dataset and problem under consideration. If the neural network has multiple hidden layers, then such a neural network is termed as a deep neural network. The training algorithm for such a deep neural network is often termed as deep learning which is a subset of machine learning. Typically, the multiple hidden layers are responsible for computation of different levels of features of the data.

Decision Trees:

Decision trees are a set of probabilistic multivariate classifiers which recursive splitting is employed to classify a new data sample [26]. The splitting operation starts at the root node and terminates till no further splits are possible in the terminal nodes. The decision trees are depicted in figure 5, which exhibits the recursive classification method.

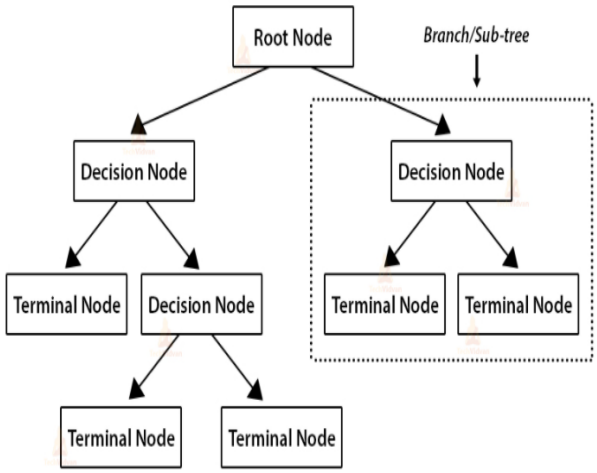

Fig. 5. A Typical Decision Tree Model

The proposed system is an amalgamation of the decision trees and the deep neural networks and is elaborately discussed in the next section.

III. Proposed Comparator Design

The proposed methodology presents an ensemble of the neural networks and decision trees to extract the attributes of both classifying paradigms which are:

1) Pattern recognition
2) Probabilistic Classification

The pattern recognition is performed using the gradient decent or scaled conjugate gradient. To update θ1 and θ2 values in order to reduce Cost function (minimizing MSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values

and then iteratively updating the values, reaching minimum cost. The main aim is to minimize the cost function J [27].

If the descent vector is given by g, then

$$g = f(J, w) \tag{6}$$

Here,
f stands for a function of.
w are the network weights.
J represents the Jacobian Matrix which is essentially the matrix containing the second order differentials of errors of the networks with respect to the weights. The Jacobian matrix is defined as:

$$J = \begin{bmatrix} \frac{\partial^2 e_1}{\partial w_1^2} & \cdots & \frac{\partial^2 e_1}{\partial w_m^2} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 e_n}{\partial w_1^2} & \cdots & \frac{\partial^2 e_n}{\partial w_m^2} \end{bmatrix} \tag{7}$$

Here,
The error e is computed as:

$$e = (y_i - y'_i) \tag{8}$$

Here,
y_i corresponds to the target variable (actual output).
y'_i represents the predicted output.

The ensemble approach presented in this paper comprises of a deep neural network based on the concept of steepest descent and the decision tress with regularization [28]. The steepest descent searches for a gradient matrix which can result in the steepest descent of the cost function or objective function during pattern recognition, and is typically chosen as the mean squared error (mse) defined as [29]:

$$mse = \frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2 \tag{9}$$

Here,
n are the number of samples in the error matrix.

The steepest descent tries to leverage a search vector which can render the steepest descent of all the possible error matrices corresponding to weight updates. The concept is depicted in figure 6.

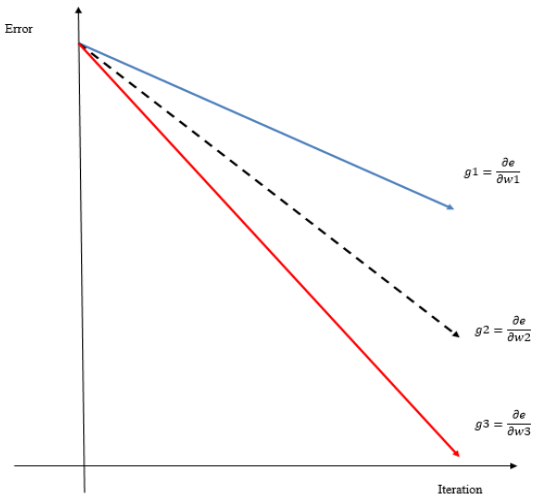

Fig. 6. Concept of Steepest Descent

From figure 6, it can be observed that the three gradients corresponding to three separate weight updates would render three different gradients given by g_1, g_2 and g_3 respectively. However, the gradient g_3 shows the quickest fall in the objective function which is the error. Thus at the beginning of each iteration, the search vector is to be evaluated which would point towards the direction of the steepest descent. Mathematically,

$$S_0 = -g_0 \tag{10}$$

Here,
S_0 corresponds to the initial values of the search vector at initialization of the weights. The weights are updated in each iteration as:

$$w_{k+1} = w_k + \mu_k g_k \tag{11}$$

Here,
w_{k+1} corresponds to the weights of the subsequent iteration.
w_k corresponds to the weights of the present iteration.
μ_k corresponds to the combination co-efficient.
k corresponds to the iteration number.

The search vector S for the iteration number $'k'$ is computed as:

$$S_k = -g_k + \beta_k S_{k-1} \tag{12}$$

The term β_k is computed as:

$$\beta_K = \frac{(|g_{k+1}|^2 - g_{k+1}{}^T g_k)}{g_k{}^T g_k} \tag{13}$$

Here,
$\frac{\partial e}{\partial w}$ represents the gradient of the error with respect to the weights and is represented as g_k.

It can be observed here that the search vector in each iteration takes into account the errors, thus the gradient of the previous iteration. This implies that the outputs of the previous iterations are considered as variables in the present iteration, thereby implementing "back-propagation", which happens to be one of the most effective learning techniques for deep neural networks (DeepNets) [30]. Subsequent to the pattern recognition, the final classification is based on the Decision Trees which is a probabilistic classifier.

The classification inaccuracy is computed based on the Gini's Index defined in equation 14 [31].

$$G = P(C) * [1 - P(C)] \tag{14}$$

Here,
G corresponds to Gini's Index.
$P(C)$ corresponds to the probability of a data sample corresponding to a defined or labelled category C.
$1 - P(C)$ corresponds to the probability of a data sample not belonging to a defined or labelled category C.

Binary trees are generated at each node corresponding to splits at each decision node and the overall split index is computed using equation 15.

$$G_O = \sum_{i=1}^{k} G_{N,i} w_i + G_{M,i} w_i \tag{15}$$

Here,
G_O corresponds to the overall Gini's index.
$G_{N,i}$ corresponds to the Tree's Index for left partitioning sub-tree.
$G_{M,i}$ corresponds to the Tree's Index for right partitioning sub-tree.
w_i corresponds to the weights of each split i.

The performance of the classifier is computed based on the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values which are used to compute the accuracy of the classifier, mathematically expressed in equation 16 [32].

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

IV. Simulation Results

The system is designed on Matrix Laboratory (MATLAB) 2020a. The system configurations are: Available RAM 16GB, processor intel i7, with a clock speed of 2.4GHz. The dataset has bene obtained from the UCI machine learning repository: https://archive.ics.uci.edu/ml/datasets/diabetes [33]

The parameters or features used in the system design are:

Age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, resting blood pressure, resting electrocardiogram (ECG), maximum heart rate, exercise induced angina (chest pain or discomfort), ECG parameters such as RR interval, QRS complex, slope of ST segment and the target is a vector containing three cases:

1) Reversibel health ailment.
2) Irreversible health ailment.
3) No health ailment or Noraml condition.

The iterations to convergence and the ensemble method used in depicted in figure 7.

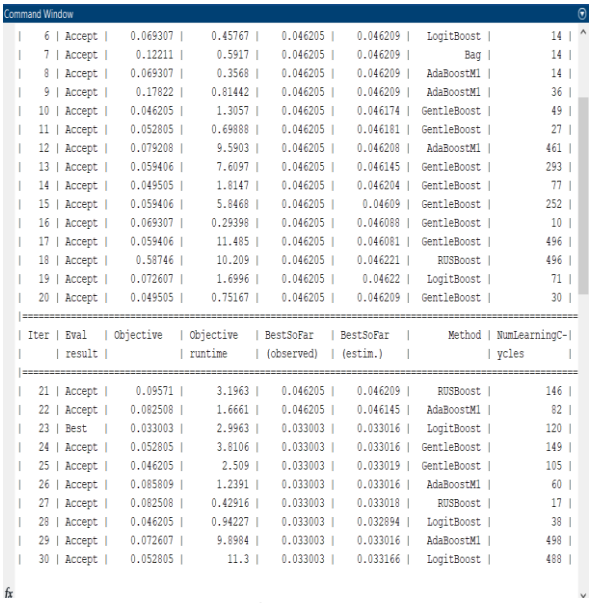

Fig. 7. Iterations to Convergence

The iterations to convergence and ensemble method is depicted in figure 7. The training stops for regression analysis in 2 cases:

1) Objective function stabilizes for validation check counts.
2) Maximum pre-defined iterations are over.

It can be observed that the system converges in 30 iterations.

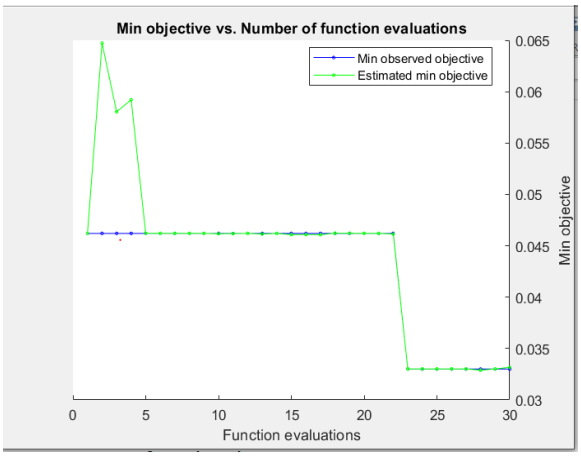

Fig. 8. Variation of Objective Function with respect to function evaluations.

The variation of the objective function with respect to the number of function evaluations is depicted in figure 8. It can be observed that as the function evaluations increases, the objective function shows a plummeting nature stabilizes eventually.

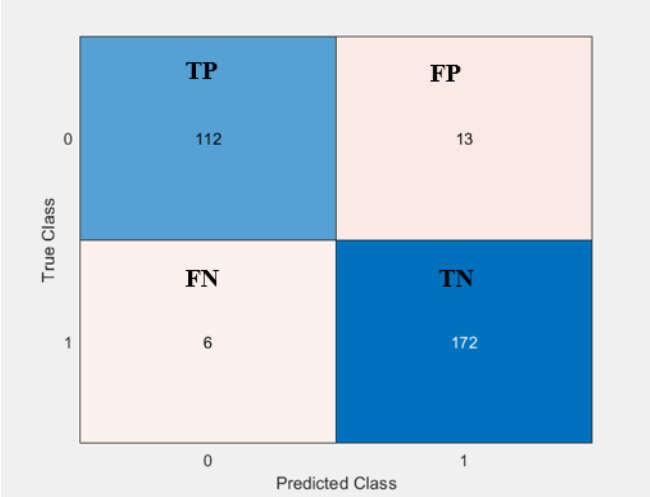

Fig. 9. Confusion matrix corresponding to classification.

Figure 9 depicts the confusion matrix corresponding to the predictive classification of the model. A total of 303 samples are tested with a training to testing division ration of 75:23 adopted in this experiment. The accuracy of classification is computed as:

$$\boldsymbol{Ac} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} = \frac{172+112}{172+112+13+6} = \mathbf{93.72\%}$$

For the ease of use of any end user, or for building a prototype model for the proposed system, a simple graphical user interface (GUI) is creased in MATLAB with the inputs to be fetched from the values entered by the user. The GUI designed in depicted in figure 10.

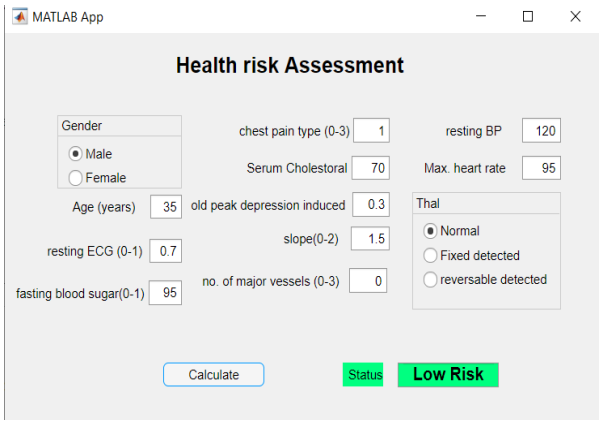

Fig. 10. Graphical User Interface (Prototype) designed.

The designed GUI prototype can be used for prototype testing and can be later extended to web or mobile applications and deployed. The summary of the obtained results for the proposed system is tabulated in table I.

TABLE I

SUMMARY OF RESULTS

S.No.	Parameter	Value
1.	Number of attributes (features)	13
2.	Categories of Target Vector	3
3.	Classifier Type	Ensemble
4.	Regression Analysis	Steepest Descent
5.	Number of Hidden Layers	5
6.	Classification Layer	Decision Trees
7.	Iterations to Convergence	30
8.	Classification Accuracy	93.72%

The results show that the proposed system attains a classification accuracy of 93.72% for used dataset. The prototype can be extended for a much larger disease dataset with a multitude of parameters (not necessarily numeric in nature). A comparison with existing work is cited in table II.

TABLE II

COMPARISON WITH EXISTING WORK

S.No.	Authors and Reference	Performance
1.	J.Archenaa, & E.A.Mary Anita [2]	Highest Accuracy of 89% achived.
2.	P. Chiang and S. Dey [3]	Mean Absolute Error (MAE) of 14.5% achieved giving an accuracy of 85.5%
3.	X. Li and J. Li [5]	Accuracy of 86.7% achieved
4.	**Proposed Approach**	**Accuracy of 93.72% achieved.**

It can be observed that the proposed approach outperforms existing techniques in terms of classification accuracy.

V. CONCLUSION

It can be concluded that the necessity of automated tools for health risk estimation is necessary keeping in mind the lifestyle changes risks at earlier ages. This paper presents an ensemble learning based approach for health risk estimation. In this classifier design, the training data which is labelled is applied to the algorithm for pattern analysis which assumes the data events in classes to be true. Based on the analyzed patterns, the new data sample's probability to belong to a specific category is evaluated. The performance of the system has been evaluated in terms of the true positive, true negative, false positive and false negative rates, Based on these metrics, the accuracy of the system has bene evaluated. The experimental results show that the proposed system attains a classification accuracy of 93.7%. The number of iterations are also less which are 30 for convergence. The designed GUI can also be used for prototype testing. Thus, the proposed system effectively predicts health risks based on medical record datasets which relatively high accuracy. Future research directions may include using clustering to find regular patterns in disease categories. Data optimization and dimensional reduction techniques such as principal component analysis (PCA) can also be tried out in conjugation with the existing system.

REFERENCES

[1] V Ilakkuvan, A Johnson, AC Villanti, WD Evans, "Patterns of social media use and their relationship to health risks among young adults", Journal of Adolescent Health, Elsevier, 2019, vol. 64, no. 2, pp. 158-164. https://doi.org/10.1016/j.jadohealth.2018.06.025

[2] J.Archenaa, & E.A.Mary Anita. (2017). Health Recommender System using Big data analytics. Journal of Management Science and Business Intelligence, vol.2, no.2, pp. 17–24. http://doi.org/10.5281/zenodo.835606

[3] P. Chiang and S. Dey, "Personalized Effect of Health Behavior on Blood Pressure: Machine Learning Based Prediction and Recommendation," 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom), 2018, pp. 1-6, doi: 10.1109/HealthCom.2018.8531109.

[4] E. Sezgin and S. Özkan, "A systematic literature review on Health Recommender Systems," 2013 E-Health and Bioengineering Conference (EHB), 2013, pp. 1-4, doi: 10.1109/EHB.2013.6707249.

[5] X. Li and J. Li, "Health Risk Prediction Using Big Medical Data - a Collaborative Filtering-Enhanced Deep Learning Approach," 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom), 2018, pp. 1-7, doi: 10.1109/HealthCom.2018.8531143.

[6] World Economic Forum Report, https://reports.weforum.org/global-risks-report-2020/false-positive/

[7] N. S. Rajliwall, R. Davey and G. Chetty, "Machine Learning Based Models for Cardiovascular Risk Prediction," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), 2018, pp. 142-148, doi: 10.1109/iCMLDE.2018.00034.

[8] AC Dimopoulos, M Nikolaidou, FF Caballero, "Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk", BMC Med Res Methodol 18, Springer 2018, vol.18, no. 179. https://doi.org/10.1186/s12874-018-0644-1.

[9] A Maxwell, R Li, B Yang, H Weng, A Ou, H Hong, "Deep learning architectures for multi-label classification of intelligent health risk prediction", BMC Bioinformatics Springer 2017, , vol.18, no. 523, https://doi.org/10.1186/s12859-017-1898-z

[10] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[11] B. Nithya and V. Ilango, "Predictive analytics in health care using machine learning tools and techniques," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), 2017, pp. 492-499, doi: 10.1109/ICCONS.2017.8250771.

[12] EG Ross, NH Shah, RL Dalman, KT Nead, "The use of machine learning for the identification of peripheral artery disease and future mortality risk", Journal of Vascular Surgery, Elsevier 2016,vol. 64, no. 5, pp. 1515-1522.

[13] D. LaFreniere, F. Zulkernine, D. Barber and K. Martin, "Using machine learning to predict hypertension from a clinical dataset," 2016 IEEE Symposium Series on Computational Intelligence (SSCI), 2016, pp. 1-7, doi: 10.1109/SSCI.2016.7849886.

[14] D Tay, CL Poh, RI Kitney," A novel neural-inspired learning algorithm with application to clinical risk prediction", Journal of Biomedical Informatics, Elsevier 2015, vol. 54, pp. 305-314

[15] K. Sowjanya, A. Singhal and C. Choudhary, "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices," 2015 IEEE International Advance Computing Conference (IACC), 2015, pp. 397-402, doi: 10.1109/IADCC.2015.7154738.

[16] LM Hlaváč, D Krajcarz, IM Hlaváčová, S Spadło, "Precision comparison of analytical and statistical-regression models for AWJ cutting", Precision Engineering, Elsevier 2017,vol. 50, pp. 148-159

[17] C Bergmeir, RJ Hyndman, B Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction", Computational Statistics & Data Analysis, Elsevier 2018, vol.120, pp. 70-83.

[18] S Bandaru, AHC Ng, K Deb, "Data mining methods for knowledge discovery in multi-objective optimization: Part A-Survey", Expert Systems with Applications, Elsevier 2017, vol. 70, no.15 pp.139-159

[19] A. Karpatne, I. Ebert-Uphoff, S. Ravela, H. A. Babaie and V. Kumar, "Machine Learning for the Geosciences: Challenges and Opportunities," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 8, pp. 1544-1554, 1 Aug. 2019, doi: 10.1109/TKDE.2018.2861006.

[20] V. Sze, Y. Chen, T. Yang and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," in Proceedings of the IEEE, vol. 105, no. 12, pp. 2295-2329, Dec. 2017, doi: 10.1109/JPROC.2017.2761740.

[21] W. Zhou, J. Li, M. Zhang, Y. Wang and F. Shah, "Deep Learning Modeling for Top-N Recommendation With Interests Exploring," in IEEE Access, vol. 6, pp. 51440-51455, 2018, doi: 10.1109/ACCESS.2018.2869924.

[22] Machine Learning Notes: Stanford University: http://cs229.stanford.edu/materials.html.

[23] Haykin S, "Neural Networks and Learning Machines", 3rd Edition, Pearson Publications.

[24] Hagan M, "Neural Network Design", 2nd Edition, Cengage Publication.

[25] K. D. Humbird, J. L. Peterson and R. G. Mcclarren, "Deep Neural Network Initialization With Decision Trees," in IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 5, pp. 1286-1295, May 2019, doi: 10.1109/TNNLS.2018.2869694.

[26] A. Segatori, F. Marcelloni and W. Pedrycz, "On Distributed Fuzzy Decision Trees for Big Data," in IEEE Transactions on Fuzzy Systems, vol. 26, no. 1, pp. 174-192, Feb. 2018, doi: 10.1109/TFUZZ.2016.2646746.

[27] A Wanto, M Zarlis, D Hartama, "Analysis of Artificial Neural Network Backpropagation Using Conjugate Gradient Fletcher Reeves in the Predicting Process", IOP Science, vol. 930, pp.1-7

[28] G Yuan, Z Wei, Y Yang, "The global convergence of the Polak–Ribière–Polyak conjugate gradient algorithm under inexact line search for nonconvex functions", Journal of Computational and Applied Mathematics, Elsevier, 2019, vol. 362, 15 pp.262-275.

[29] D Di Serafino, V Ruggiero, G Toraldo, "On the steplength selection in gradient methods for unconstrained optimization", Applied Mathematics and Computation, Elsevier 2018, vol. 318, pp. 176-195.

[30] H. Su, G. Li, D. Yu and F. Seide, "Error back propagation for sequence training of Context-Dependent Deep NetworkS for conversational speech transcription," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6664-6668.

[31] BK Das, HS Dutta, "GFNB: Gini index–based Fuzzy Naive Bayes and blast cell segmentation for leukemia detection using multi-cell blood smear images", Journal of Medical & Biological Engineering & Computing, Springer 2020, vol.58, pp. 2789–2803.

[32] JP Kandhasamy, S Balamurali, "Performance analysis of classifier models to predict diabetes mellitus", Procedia Computer Science, Elsevier, 2015, vol.47, pp. 45-51.

[33] https://archive.ics.uci.edu/ml/datasets/diabetes