An Ensemble Approach for Forecasting Critical Health Risks

[1]Muskan Gangrade, [2]Dr. Sachin Patel
[1]*Mtech scholar, Sagar Institute of Research & Technology, SAGE University, Indore*
[2]*Associate Professor, Sagar Institute of Research & Technology, SAGE University, Indore*

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. 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% which is comparatively higher than baseline techniques.

General Terms
Automated Health Risk Assessment, Machine Learning, Ensemble Learning.

Keywords
Health Risk Prediction, Ensemble Classifier, Classification Error, Accuracy.

1. INTRODUCTION

With increase in the sedentary lifestyle of people around the globe, different health risks are affecting people worldwide. 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

No.	Cause	Estimated number of deaths (in millions)	Percent of all deaths
1	Ischaemic heart disease	7.25	12.8
2	Cerebrovascular disease	6.15	10.8
3	Lower respiratory infections	3.46	6.1
4	Chronic obstructive pulmonary disease	3.28	5.8
5	Diarrhoeal diseases	2.46	4.3
6	HIV/AIDS	1.78	3.1
7	Trachea, bronchus, lung cancers	1.39	2.4
8	Tuberculosis	1.34	2.4
9	Diabetes mellitus	1.26	2.2
10	Road traffic accidents	1.21	2.1
11	Hypertensive heart disease	1.15	2.0
12	Prematurity and low birth weight	1.00	1.8

Fig 1: Global Health Risk Analysis: 2022
(Source: WHO, [1])

By presenting patient-centered medications, this can be accomplished. By implementing predictive analytics in reality, further expenses spent on medical systems can be prevented. It helps to eliminate huge amounts of money wasted on unnecessary medicine and health treatments by making proper use of the significant amount of complex data produced by medical systems. Health activity (diet, exercise and sleep) is generally recognized as having a significant effect on the state of human health. Such relationships between health activity and predictor of health condition (blood pressure (BP) and glucose level) are commonly researched in inpatient configurations through clinical studies [3].

2. NEED FOR AUTOMATED HEALTH RISK ASSESSMEMT

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 [4]. Machine learning offers a way to automatically identify trends and predict results. There are several current experiments on various types of electronic medical data on data mining and data analytics. To assess the performance of the implemented algorithm various machine learning algorithms such as Decision Tree, Support Vector Machine (SVM), and Naive Bayes are utilized. The desired outcome is based on the most frequently used metrics: accuracy, accuracy, recall, micro-average F1, macro-average F1.

The medical system's digitization has led to a massive amount of medical data. These data help medical care institutions to improve the efficiency of the health system, enhance the quality of healthcare and minimize healthcare costs. Big data helps businesses make better decisions to produce high revenues, improve performance, and gain comparative advantages. According to the enormous value derived from big data, the creation of Big Data systems has been actively encouraged in recent years. Multiple companies from various areas have been increasingly digitized, obtaining knowledge and information from enormous quantities of big data. With the development of Healthcare Information Systems (HIS), Electronic Medical Records (EMRs), and mobile and smart phones, Healthcare has also undergone this technological transition. Big Data applications have the ability to shift the enterprise to provide effective, reliable care. It can promote decision-making, aid initial infection diagnosis, and anticipate disease course In addition, big data can assist healthcare facilities in management, price-effectiveness, and analysis customization. By giving recommendations, big data could minimize medical resources [5]. Following are the major issues for health recommendation system.

- In general, Recommendation System (RS) depends on asset popularity and can often be deceptive in HRS
- A large amount of medical content, such as text file, audio speech, and email messages, can contain unstructured data. Usually, unorganized content includes a personal touch to properly read record and analyze.
- Medical records are usually extremely complicated and ethnically diverse. The main difficulty of evaluating large-scale medical data is to create a suitable range of features from a variety of complex features without person intervention.
- Data analysis is a customer modeling problem that causes racial, gender and sexual-oriented ethical concerns.
- Generally, medical data is distributed or sparse. Data might have huge amounts of missing values owing to different human considerations.

3. EXISTING METHODS

Off late machine learning based classifiers are being used for the classification problems. Machine learning based classifiers are typically much more accurate and faster compared to the conventional classifiers. They render more robustness to the system as they are adaptive and can change their characteristics based on the updates in the dataset. 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:

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

Here,

x represents the state vector of input variables

y represents the state vector of output variable or variables.

Ɵ1 and Ɵ2 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:

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

Here,

x is the independent variable

y is the dependent variable

$\theta_1, \theta_2 \ldots\ldots \theta_n$ 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

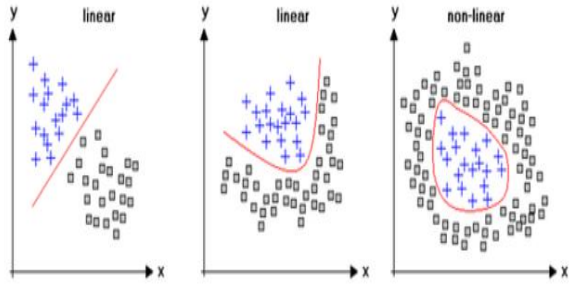

Fig 2: Linear and Non-Linear Regression fitting.

Support Vector Machine (SVM): 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. 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.

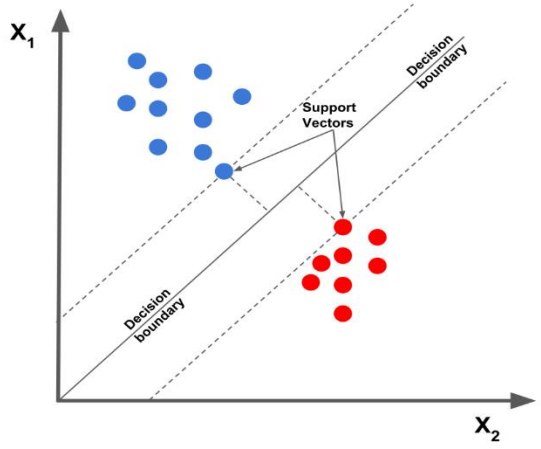

Fig 3: Separation of data classes using SVM.

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

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

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:

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

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). 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.

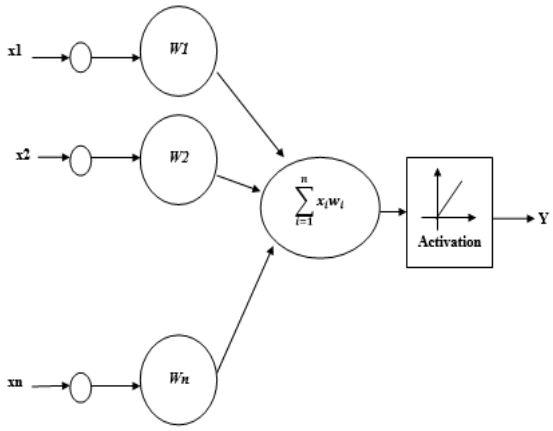

Fig 4: Mathematical Model of Single Neuron.

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. 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: The decision trees are another class of multivariate classifiers.The tree tries to estimate event outcomes based on probabilities, where the target or output variable is dependent on several input or governing variables. The decision tree is obtained by recursively splitting the source data set (know as root node) into subsequent branches termed as the children.

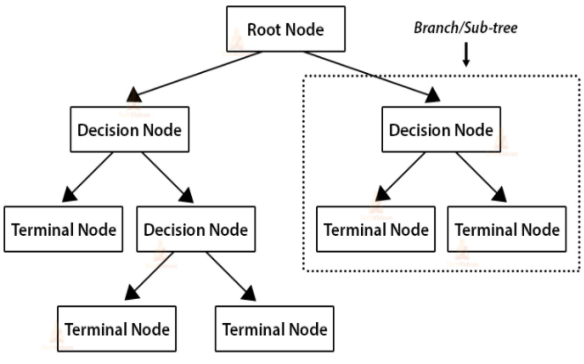

Fig 5: A Typical Decision Tree Model

The decision tree structure is depicted in figure 5 which exhibits the recursive classification method.

4. PROPOSED METHODOLOGY

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. If the descent vector is given by 'g', then

$$g = f(J, w) \quad (6)$$

Here,
F stands for a function of.
W are the network weights

The gradient descent based approach if often employed to train a neural model such that, the increase in iterations continuously decrease a cost function defined in the above section. Moreover, as discussed in the previous approaches, there are several techniques and mechanisms to train a neural network out of which one of the most effective techniques is the back propagation based approach. The scaled conjugate gradient tries to find the steepest descent vector prior to weight update in each iteration and is mathematically given by:

$$A_0 = -g_0 \quad (7)$$

Here,
A is the initial search vector for steepest gradient search
g is the actual gradient

$$w_{k+1} = w_k + \mu_k g_k \quad (8)$$

Here,
w_{k+1} is the weight of the next iteration
w_k is the weight of the present iteration
μ_k is the combination co-efficient
For any iteration k, the search vector is given by:

$$A_k = -g_k + \beta_k A_{k-1} \quad (9)$$

And

$$\beta_K = \frac{(|g_{k+1}|^2 - g_{k+1}{}^T g_k)}{g_k{}^T g_k} \quad (10)$$

Here,

The customary g represents $\frac{\partial e}{\partial w}$

The ensemble also uses the decision trees which is effective for multi-class decisions. For a multi-class classification, the conditional probability of the sentiment can be also seen as an overlapping event with the classification occurring with the class with maximum conditional probability. The probability of inaccurate classifications based on the Gini's Index is given by:

$$G = Prob(C) * [1 - Prob(C)] \quad (11)$$

Here,

G denotes the Gini's Index.

$Prob(C)$ denotes probability of a data sample to belong to class 'C'.

$1 - Prob(C)$ denotes the complement of probability of a data sample to belong to class 'C'.

The continued or recursive splitting generates binary trees at every decision node and the final Gini Index is computed as the weighted sum of all the individual splits. Thus the total split index is given by:

$$G_{tot} = \sum_{i=1}^{k} G_{L,i} w_i + G_{R,i} w_i \quad (12)$$

Here,

G_{tot} denotes total Gini Index

$G_{L,i}$ denotes Left Partitioned Tree's Index

$G_{R,i}$ denotes Right Partitioned Tree's Index

w_i denotes the weights of the partitioning.

The performance metrics of the classifiers are generally 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 and sensitivity of the classifier, mathematically expressed as:

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

Sensitivity: It is mathematically defined as:

$$Se = \frac{TP}{TP+FN} \quad (14)$$

$$Recall = \frac{TP}{TP+FN} \quad (15)$$

$$Precisiosn = \frac{TP}{TP+FP} \quad (16)$$

$$F - Measure = \frac{2.Precision.Recall}{Precision+Recall} \quad (17)$$

The aim of any designed approach is to attain high values of accuracy of classification along with other associated parameters. The computation complexity of the system often evaluated in terms of the number of training iterations and execution time is also a critically important metric which decides the practical utility of any algorithm on hardware constrained devices.

5. EXPERIMENTAL RESULTS

The system has been designed on MATAB 2020a. The results obtained are presented subsequently.

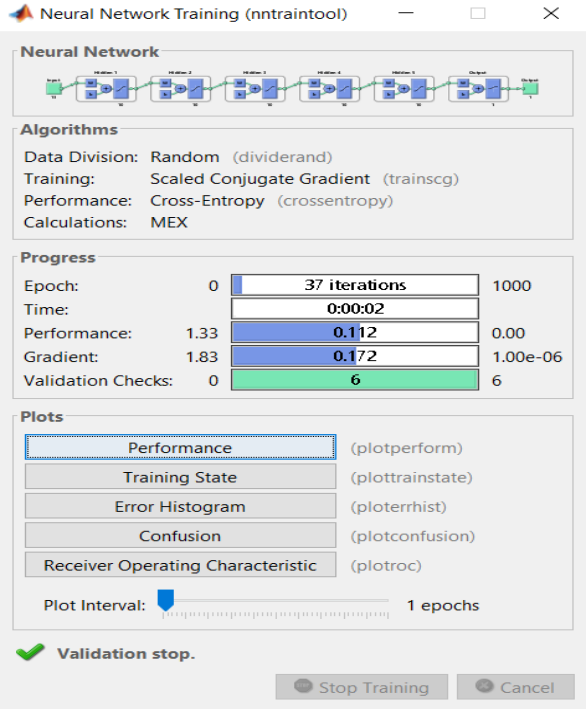

Fig 6: The Regression Analysis Model

Figure 6 depicts the regression analysis model

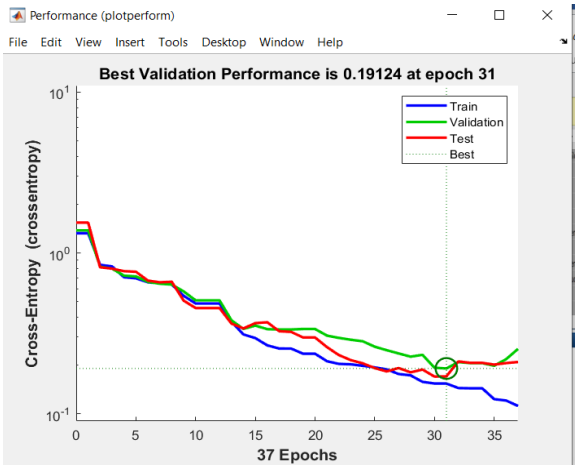

Fig 7: Variation of MSE

Figure 7 shows the variation of the cost function (MSE) in this case with the increase in the number of iterations (epochs). The training stops for regression analysis in 2 cases:

1) Objective function stabilizes for validation check counts.
2) Maximum epochs are over.

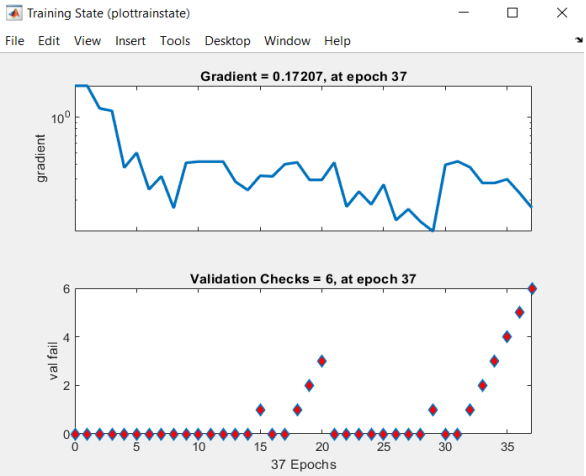

Fig 8: Training States

Figure 8 depicts the variation of the training states as a function of iterations.

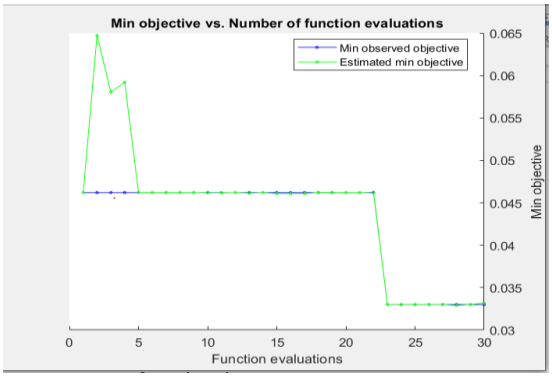

Fig 9: Function Evaluations.

The ensemble tree using which the iterations to convergence are achieved are depicted in figure 9. The iterations to convergence and the learning rate are depicted subsequently.

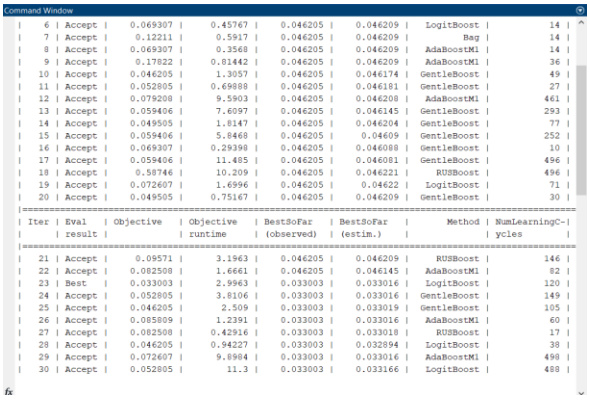

Fig 10: Iterations to Convergence

The iterations to convergence and ensemble method is depicted in figure 10. It can be seen that eh system converges in 30 iterations.

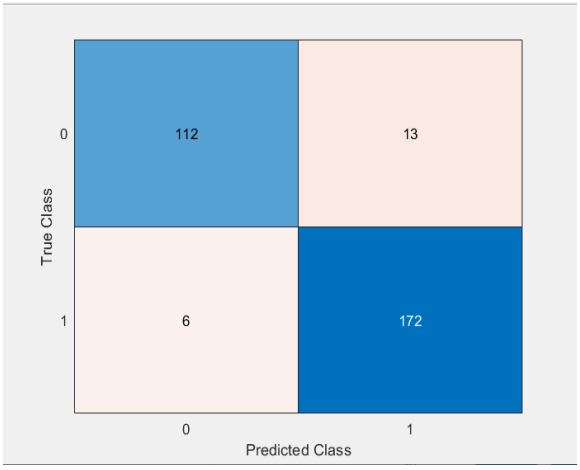

Fig 11: Confusion Matrix

The confusion matrix is depicted in figure 11. The accuracy is thus computed as:

$$Ac = \frac{172 + 112}{172 + 112 + 13 + 6} = 93.72\%$$

6. 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% which is comparatively higher than the existing system [1] which attains a classification accuracy of 87.7% for the same dataset. The number of iterations are also less which are 30 for convergence. Thus, the proposed system effectively predicts health risks based on medical record datasets which relatively high accuracy.

REFERENCE

[1] https://www.who.int/news-room/questions-and-answers/item/what-is-the-deadliest-disease-in-the-world.

[2] 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

[3] 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

[4] 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.

[5] 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.

[6] 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.

[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] D Kumar, KN Rai, "Numerical simulation of time fractional dual-phase-lag model of heat transfer within skin tissue during thermal therapy",

Journal of Thermal Biology, Elsevier 2017, vol. 67, pp. 49-58

[19] M. Chen, U. Challita, W. Saad, C. Yin and M. Debbah, "Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial," in IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3039-3071, Fourthquarter 2019, doi: 10.1109/COMST.2019.2926625.

[20] I. H. Laradji, R. Pardinas, P. Rodriguez and D. Vazquez, "Looc: Localize Overlapping Objects with Count Supervision," 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 2316-2320, doi: 10.1109/ICIP40778.2020.9191122.

[21] 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

[22] 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.

[23] 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.

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

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

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