International Research Journal of Education and Technology

Peer Reviewed Journal, ISSN 2581-7795

An Overview of ML Based Critical Health Risk Prediction System

Etesh[1], Prof. Virendra Verma[2]
M.Tech Scholar[1], Asst. Professor[2]
Department of Computer science & Engineering[1,2]
LNCT(Bhopal) Indore Campus, Indore, India[1,2]

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. The choice of classifier here is important as the data often shows overlapping nature. An overview of the aspect and different methodologies adopted in this regard are presented in this paper.

Keywords: Health Risk Assessment, Machine Learning, Error Performance, Accuracy Estimation.

I. 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 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.

Figure.1 Global Health Risk (Source: Statista

https://cdn.statcdn.com/Infographic/images/normal/23755.jpeg)

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].

II. PREVIOUS WORK

Li et al. [6] showed that the massive amount of medical data accumulated from patients and healthcare providers has become a vast reservoir of knowledge source that may enable promising applications such as risk predictive modeling, clinical decision support, disease or safety surveillance. However, discovering knowledge from the big medical data can be very complex because of the nature of this type of data: they normally contain large amount of unstructured data; they may have lots of missing values; they can be highly complex and heterogeneous. To address these challenges, authors have proposed a Collaborative Filtering-Enhanced Deep Learning approach. In particular, we estimate missing values based on patients' similarity, i.e., we predict one patient's missing features based on the values of similar patients. This is implemented with the Collaborative Topic Regression method, which tightly couples topic model and probability matrix factorization and is able to utilize the rich information hidden in the data. Then a deep neural network-based method is applied for the prediction of health risks. This method can help us handle complex and multi-modality data. Extensive experiments on a real-world dataset have been performed and the results show improvements of the proposed algorithm over the state-of-the-art methods.

Rajilwall et al. [7] proposed a machine learning primarily based prognostic modelling framework, which may run in static/low speed, massive information from electronic health records, furthermore as extreme velocity, streaming massive information settings captured from wearables, like fitness bands and biosensor watches. Authors describe a scalable algorithm called Neuron network, which is used to achieve highly accurate results in fuzzy data. Authors have presented the outcomes of the proposed framework implementation for static and low-velocity/volume settings from the EHR & clinical DBs, with the experimental authentication of the planned framework, for 2 openly accessible CVD data sets which are “NHANES” dataset, and the “Framingham Heart Study” dataset), shown promising outcomes, in terms of performance of different modelling algorithms for the disease status prediction.

Dimopoulos et al. [8] used of Cardiovascular Disease (CVD) risk estimation scores in primary prevention has long been established. However, their performance still remains a matter of concern. The aim of this study was to explore the potential of using ML methodologies on CVD prediction, especially compared to established risk tool. Depending on the classifier and the training dataset the outcome varied in

efficiency but was comparable between the two methodological approaches. In particular, the system showed accuracy 85%, specificity 20%, sensitivity 97%, positive predictive value 87%, and negative predictive value 58%, whereas for the machine learning methodologies, accuracy ranged from 65 to 84%, specificity from 46 to 56%, sensitivity from 67 to 89%, positive predictive value from 89 to 91%, and negative predictive value from 24 to 45%; random forest gave the best results, while the k-NN gave the poorest results.

Maxwell et al. [9] showed that multi-label classification of data remains to be a challenging problem for medical records because of the complexity of the data, it is sometimes difficult to infer information about classes that are not mutually exclusive. For medical data, patients could have symptoms of multiple different diseases at the same time and it is important to develop tools that help to identify problems early. Intelligent health risk prediction models built with deep learning architectures offer a powerful tool for physicians to identify patterns in patient data that indicate risks associated with certain types of chronic diseases. The results suggest that Deep Neural Networks (DNN), a DL architecture, when applied to multi-label classification of chronic diseases, produced accuracy that was comparable to that of common methods such as Support Vector Machines. We have implemented DNNs to handle both problem transformation and algorithm adaption type multi-label methods and compare both to see which is preferable.

Chen et al. [10] showed that with big data growth in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care, and community services. However, the analysis accuracy is reduced when the quality of medical data is incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. In this paper, we streamline machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities. To overcome the difficulty of incomplete data, authors use a latent factor model to reconstruct the missing data. We experiment on a regional chronic disease of cerebral infarction. The authors propose a new convolutional neural network (CNN)-based multimodal disease risk prediction algorithm using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared with several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed, which is faster than that of the CNN-based unimodal disease risk prediction algorithm.

Nithya et al. [11] proposed that machine Learning (ML) is the fastest rising arena in computer science, and health informatics is of extreme challenge. The aim of Machine Learning is to develop algorithms which can learn and progress over time and can be used for predictions. Machine Learning practices are widely used in various fields and primarily health care industry has been benefitted a lot through machine learning prediction techniques. It offers a variety of alerting and risk management decision support tools, targeted at improving patients' safety and healthcare quality. With the need to reduce healthcare costs and the movement towards personalized healthcare, the healthcare industry faces challenges in the essential areas like, electronic record management, data integration, and computer aided diagnoses and disease

predictions. Machine Learning offers a wide range of tools, techniques, and frameworks to address these challenges. This paper depicts the study on various prediction techniques and tools for Machine Learning in practice. A glimpse on the applications of Machine Learning in various domains are also discussed here by highlighting on its prominence role in health care industry.

Ross et al. [12] proposed that effectiveness of precision medicine is beginning to be realized in some areas of medicine. In Oncology, genetic profiling is now being used to identify patients for whom tailored chemotherapy regimens—directed against the individual's personal cancer mutation—can be used to significantly improve outcomes relative to traditional therapy. Rather than the current empirical approach to treatment, there is hope that with a deeper understanding of biology and pharmacogenomics, we may one day be able to guarantee that every patient receives the right dose of the right medicine at the right time. The proposed machine-learned models outperformed stepwise logistic regression models both for the identification of patients with PAD (area under the curve, 0.87 vs 0.76, respectively; P = .03) and for the prediction of future mortality (area under the curve, 0.76 vs 0.65, respectively; P = .10). Both machine-learned models were markedly better calibrated than the stepwise logistic regression models, thus providing more accurate disease and mortality risk estimates.

LaFreniere et al. [13] proposed that artificial neural network is a powerful machine learning technique that allows prediction of the presence of the disease in susceptible populations while removing the potential for human error. In this paper, authors identify the important risk factors based on patients' current health conditions, medical records, and demographics. These factors are then used to predict the presence of hypertension in an individual. These risk factors are also indicative of the probability of a person developing hypertension in the future and can, therefore, be used as an early warning system. Authors design a neural network model for predicting hypertension with about 82% accuracy. This is good performance given our chosen risk factors as inputs and the large integrated data used for the study. The proposed network model utilizes very large sample sizes (185,371 patients and 193,656 controls) from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) data set.

Tay et al. [14] proposed a novel learning algorithm – a key factor that influences the performance of machine learning-based prediction models – and utilities it to develop CVD risk prediction tool. This novel neural-inspired algorithm, called the Artificial Neural Cell System for classification (ANCSc), is inspired by mechanisms that develop the brain and empowering it with capabilities such as information processing/storage and recall, decision making and initiating actions on external environment. Specifically, we exploit on 3 natural neural mechanisms responsible for developing and enriching the brain – namely neurogenesis, neuroplasticity via nurturing and apoptosis – when implementing ANCSc algorithm. Benchmark testing was conducted using the Honolulu Heart Program (HHP) dataset and results are juxtaposed with 2 other algorithms – i.e. Support Vector Machine (SVM) and Evolutionary Data-Conscious Artificial Immune Recognition System (EDC-AIRS). Empirical experiments indicate that ANCSc algorithm (statistically) outperforms both SVM and EDC-AIRS algorithms. Key clinical markers identified by ANCSc algorithm include risk factors related to diet/lifestyle,

pulmonary function, personal/family/medical history, blood data, blood pressure, and electrocardiography. These clinical markers, in general, are also found to be clinically significant providing a promising avenue for identifying potential cardiovascular risk factors to be evaluated in clinical trials.

Sowjanya et al. [15] showed that deficiency of knowledge about diabetes causes untimely death among the population at large. Therefore, acquiring a proficiency that should spread awareness about diabetes may affect the people in India. In this work, a mobile/android application based solution to overcome the deficiency of awareness about diabetes has been shown. The application uses novel machine learning techniques to predict diabetes levels for the users. At the same time, the system also provides knowledge about diabetes and some suggestions on the disease. A comparative analysis of four machine learning (ML) algorithms were performed. The Decision Tree (DT) classifier outperforms amongst the 4 ML algorithms. Hence, DT classifier is used to design the machinery for the mobile application for diabetes prediction using real world dataset collected from a reputed hospital in the Chhattisgarh state of India.

III. EXISTING MODELS

The major challenge with design of health risk prediction systems are:

1) The data is extremely complex and uncorrelated in nature.
2) The number of variables being large makes it extremely challenging to carry out regression analysis.
3) The outcomes are often individual dependent not exhibiting alignment to fixed patterns.

Mostly, evolutionary algorithms are used in the domain to design models for health risk prediction**.** Evolutionary algorithms try to mimic the human attributes of thinking which are:

1) Parallel data processing
2) Self-Organization
3) Learning from experiences

The major approaches employed in the domain of health risk prediction are:

Some of the commonly used techniques are discussed below:

1) Statistical Regression: These techniques are based on the time series approach based on the fitting problem that accurately fits the data set at hand. The approach generally uses the auto-regressive models and means statistical measures. They can be further classified as [16]:

a) Linear

b) Non-Linear

Mathematically:

Let the time series data set be expressed as:

$$Y = \{Y1, Y2 \dots \dots \dots \dots \dots \dots Yt) \quad (1)$$

Here,

Y represents the data set

t represents the number of samples

Let the lags in the data be expressed as the consecutive differences.

The first lag is given by:

$$\Delta Y_1 = Y_{t-1} \quad (2)$$

Similarly, the j^{th} lag is given by:

$$\Delta Y_j = Y_{t-j} \quad (3)$$

2) Correlation based fitting of time series data: The correlation based approaches try to fit the data based on the correlation among the individual lags. Mathematically it can be given by [18]:

$$A_t = corr(Y_t, Y_{t-1}) \quad (4)$$

Here,

Corr represents the auto-correlation (which is also called the serial correlation)

Y_t is the t^{th} lagged value

Y_{t-1} is the $(t-1)^{st}$ lagged value

The mathematical expression for the correlation is given by

$$corr(Y_t, Y_{t-1}) = \frac{conv(Y_t, Y_{t-1})}{\sqrt{varY_t, varY_{t-1}}} \quad (5)$$

Here,

Conv represents convolution given by:

$$conv\{x(t), h(t)\} = \int_{t=1}^{\infty} x(\vartheta)h(t-\vartheta)d\vartheta \quad (6)$$

μ_t is the time dependent combination-coefficient

4) Artificial Neural Networks (ANN) and Deep Neural Networks (DNNs): In this approach, the time series data is fed to a neural network resembling the working of the human based brain architecture with a self-organizing memory technique [19].

The approach uses the ANN and works by training and testing the datasets required for the same. The general rule of the thumb is that 70% of the data is used for training and 30% is used for testing. The neural network can work on the fundamental properties or attributes of the human brain i.e. parallel structure and adaptive self-organizing learning ability. Mathematically, the neural network is governed by the following expression:

$$Y = f(\sum_{i=1}^{n} X_i . W_i + \theta_i) \quad (7)$$

Here,

X_i represents the parallel data streams

W_i represents the weights

θ represents the bias

f represents the activation function.

The second point is critically important owing to the fact that the data in time series problems such as sales forecasting may follow a highly non-correlative pattern and pattern recognition in such a data set can be difficult. Mathematically:

$$x = f(t)$$

Here,

x is the function

t is the time variable.

The relation f is often difficult to find being highly random in nature.

The neural network tries to find the relation f given the data set (D) for a functional dependence of x(t).

The data is fed to the neural network as training data and then the neural network is tested on the grounds of future data prediction. The actual outputs (targets) are then compared with the predicted data (output) to find the errors in prediction. Such a training-testing rule is associated for neural network. Deep Neural Networks are the neural networks with multiple hidden layers and are generally used for training complex datasets.

IV. EVALUATION PARAMETERS

The performance metrics of the machine learning based classifier is generally done based on:

The parameters which can be used to evaluate the performance of the ANN design for time series models is given by:

1) Mean Absolute Error (MAE)
2) Mean Absolute Percentage Error (MAPE) and
3) Mean square error (MSE)

The above mentioned errors are mathematically expressed as:

$$MAE = \frac{1}{N}\sum_{t=1}^{N} |V_t - \hat{V}_t| \quad (8)$$

Or

$$MAE = \frac{1}{N}\sum_{t=1}^{N} |e_t| \quad (9)$$

$$MAPE = \frac{100}{N}\sum_{t=1}^{N} \frac{|V_t - \hat{V}_t|}{V_t} \quad (10)$$

The mean square error (MSE) is given by:

$$MSE = \frac{1}{N}\sum_{t=1}^{N} e_t^2 \quad (11)$$

Here,

N is the number of predicted samples
V is the predicted value
$\hat{V}_t$ is the actual value
e is the error value

Conclusion: It can be concluded form the previous discussions that a vast amount of clinical data scattered across different sites on the Internet hinders users from finding helpful information for their well-being improvement. Besides, the overload of medical information (e.g., on drugs, medical tests, and treatment suggestions) have brought many difficulties to medical professionals in making patient-oriented decisions. These issues raise the need to apply recommender systems in the healthcare domain to help both, end-users and medical professionals, make more efficient and accurate health-related decisions. In this article, we provide a systematic overview of existing research on healthcare recommender systems.

References:

[1] World Health Risk Tabulation: Statista, https://cdn.statcdn.com/Infographic/images/normal/23755.jpeg

[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," IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom), 2021, pp. 1-7

[7] N. S. Rajliwall, R. Davey and G. Chetty, "Machine Learning Based Models for Cardiovascular Risk Prediction," 2020 International Conference on Machine Learning and Data Engineering (iCMLDE), 2019, 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] 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.