REVIEW ON PREDICTING DISEASE SEVERITY – LEARNING ALGORITHMS AS CLASSIFIERS FOR DATA WAREHOUSE ENVIRONMENTS

Miss SakshiHooda*, Dr. Suman Mann

**(Research Scholar, IPU, New Delhi. Email id- sakshi.hooda@gmail.com)*

*** (Associate Professor, MSIT, New Delhi. Email-ID-sumanmann2007@gmail.com*

Abstract

When research fields such as biological and medical arenas are considered, computer science-based applications have been utilized on an increasing basis. In the process, data has accumulated in significant amounts. In particular, most of the big data that has resulted has accrued from experimental and medical results that are reported. The eventuality is that in the healthcare platforms, data mining techniques have been employed towards timely predictions of conditions with which patients present, a process achieved through data analysis. Also, the data analyses have strived to discern correlations among disease-related variables or parameters in otherwise voluminous information. It is also notable that data mining has received increasing adoption in the healthcare field because of the perceived ability to provide room for informed decision-making, reliable treatment decisions, and the detection of potential fraudulence when it comes to health care service payments. While these interplays point to a promising trend regarding the utilization of data mining and disease severity prediction, the eventual bid data has proved to be so complex that the use of traditional methods of data processing and analyzing is deemed less effective. In addition, given data warehouse, the use of traditional methods towards data extraction makes it difficult to discern some of the hidden intersections, pointing to the importance of a new approach for data classification and disease severity prediction. The aim of this review paper was to demonstrate how the use of classifiers in the form of machine learning algorithms could be utilized in the prediction of the severity of diseases. Indeed, the context of the research entailed warehouse environments. To have an ideal kernel established towards data classification (that would then enable additional diagnostic tests), some of the machine learning algorithms that were investigated included SVM (support vector machine) and ANN (artificial neural network). From the review outcomes, it was established that when machine learning algorithms are used to obtain trend similarities in disease severity through predictive analytics, similar severity magnitudes of the given diseases could be projected in terms of their probabilities of and trends in occurrence. Also, the review outcomes demonstrated that disease severity prediction through machine learning algorithms is an insightful practice because it allows for the prediction of patients' conditions, as well as the prevision of room for optimal decision-making regarding treatment modalities to be adopted by healthcare firms.

Keywords: Data warehouse, disease severity, machine learning, ANN, SVM

1 Introduction

An increasing amount of data continues to result within the healthcare industry. This data arises from different diagnostic procedures that are conducted. To have the manual process of extracting data avoided, technological incorporations have been utilized [1]. Here, the continuous use of technology is motivated by the growing trend in the development and implementation of e-healthcare in which electronic health records are maintained [2, 3]. Through such technological incorporations, it is worth indicating that significant amounts of time are saved when it comes to health condition detection and severity prediction [3]. To have health conditions diagnosed, it is worth indicating that health records ought to be analyzed in data warehouse contexts [4, 5]. In the latter data systems, patients' records tend to be voluminous, calling for contemporary data mining methods to have hidden patterns and intersections established [5].

When it comes to the concept of machine learning through different models or algorithms, it can be seen that it has gained increasing usage in the healthcare sector because of the need for health condition detection, allowing for informed decision-making concerning treatment modalities [6]. Specifically, some of the machine-learning algorithms that have been used include the decision tree approach, ANN, SVM, logistic regression, and Naïve Bayes, as well as principal component analysis, K-means, and KNN

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

(K-Nearest Neighbors) [7-10]. In this review paper, the central purpose is to evaluate the attribute of machine learning and discern the accuracy and timeliness of previously proposed models relative to disease severity prediction. Important to note is that the study focused on previous scholarly investigations whose experimental practices and simulation data analyses were conducted and the results reported from the perspective of data warehouse settings.

2 Results

One of the novel techniques that have been proposed has been that which seeks to store big data via data warehouse environment-based image implementation [11]. Here, data warehouse systems in the healthcare environment have been used. Relative to the computational outcomes, it has been reported that the proposed model comes with superiority when it comes to reliability and adaptability, with a particular focus on health big data organization and routing for the prediction of disease severity. In another investigation, the ANN model has been used in a machine learning-based proposed novel approach [12]. The aim of the investigation has been to steer the prediction of patients' survival chance and health condition recurrence, especially in persons diagnosed with breast cancer. In the findings, computational outcomes have demonstrated that the proposed machine learning approach was ideal relative to cancer severity prediction, outperforming the remainders of statistical-based techniques. For an additional experimental simulation [13], the focus was on mining concepts and their effectiveness or criticality relative to cardiovascular disease symptom prediction. In the study, there was a comparison of the outcomes with those reported after implementing other frameworks, with the insights gained suggesting that the PLS-LDA exhibited superiority regarding time efficiency and accuracy, especially after proving to exhibit the least and otherwise negligible error. When it comes to diabetes mellitus disease severity prediction, a study was also conducted using the bioinformatics concept and a machine learning tool [14]. From the results, it was affirmed that the proposed technique was innovative and that upon being implemented on big data, it exhibited superiority in terms of performance efficiency and high accuracy. As such, it can be seen that some of these previously proposed and implemented novel techniques are highly informative and contributory to the field of healthcare because they allow for early interventions, informed decision-making regarding disease diagnosis and treatment, and disease control [15]. However, a common denominator is that in these previous investigations, the majorities of data mining technique implementations have experienced barriers in terms of the complexity of data collection and extraction in warehouse environments.

In another study [16], there was the development of a prediction algorithm via perceptual layer selection from the given image data warehouse. Here, the data warehouse was that which constituted information about brain disease severity. Thus, the selected data warehouse was that which hosted online website-based input data. Through the ETL (extraction transformation loading) technique, there was the processing of the data, whereby selected feature sets' extractions allowed for the training exercise. With ANN's variant also utilized, it allowed for Euclidian distance model-based mathematical organization towards the understanding of the degree to which disease severity prediction could be realized via machine learning implementation, especially by targeting both the behaviors of diseased organs or tissues and healthy ones. As documented in the literature [17], the similarity measurement approach based on the Euclidian distance allows for the selected feature sets' optimization, with image data warehouses allowing further for the storage of the resultant trained data.

In additional investigation [18], the Tunable-Q Wavelet Transform (TQWT) criterion has been proposed and implemented. Here, the TQWT framework is that which has relied on heart signal detection and the eventual role of heart disease diagnosis. In the research, with the LS-SVM (least squares support vector machine) model employed, findings have demonstrated TQWT-based accuracy improvements to about 96.8%, as well as 100% sensitivity and 93.7% specificity. Similarly, the data mining algorithm has been implemented in a scenario aimed at the evaluation of the occurrence of coronary heart disease [19]. In particular, the latter study relied on the linear time-variant technique for the purpose of disease assessment. Similar to the work conducted in [19], which sought to examine heart disease identification via the utilization of a network, it was observed that the feature selection method that was utilized proved superior and reliable relative to the process of diagnosing ischemic heart disease. In particular, a value of 89.4% was reported as the precision rate.

In another study [20], a model geared towards accurate predictions of Parkinson's disease (PK) was proposed, especially concerning the health condition's initial stages. With a three-year period used to collect and analyze the outcomes, it was reported that the regression model that was adopted yielded a 99.3% accuracy level when compared to threshold values reported previously in the experimental data. Thus, the proposed model's performance superiority was confirmed, especially due to its capacity to measure the GABA amount. In another scholarly study [21], the focus was on the implementation of an ANN-based novel algorithm and how it could aid in learning disability analysis. In children, it is important to note that the learning disability evolves as a neurological disorder [21]. Hence, the latter study's motivation lay in the establishment of an ideal technique through which children's learning experiences might be created. For the results, there were further comparative analyses to discern the algorithms' states of simplicity and efficiency. In the results, it was documented that, given the experimental data trends, the incorporation of ANN in data mining approaches translates into better outcomes when compared to other classifiers, including J48, SVM, and Naïve Bayes [21].

The OLAP query method has also been implemented [1]. Here, the processing of OLAP and ETL has been done using a selected classifier, a process that aided in disease severity prediction. Also, the choice of OLAP has been informed by a situation in which the experimental study involved a multi-dimensional analytical query. With the data processing stage entailing ETL technique implementation, three main steps have been utilized. Initially, there has been the data extraction exercise, whereby the data was read from the warehouse environment's database. The next step has been data transformation, in which data would be converted from its previous form to a new state before being channeled to other databases. The third procedure has involved the loading process, whereby the transformed data was written into the targeted databases.

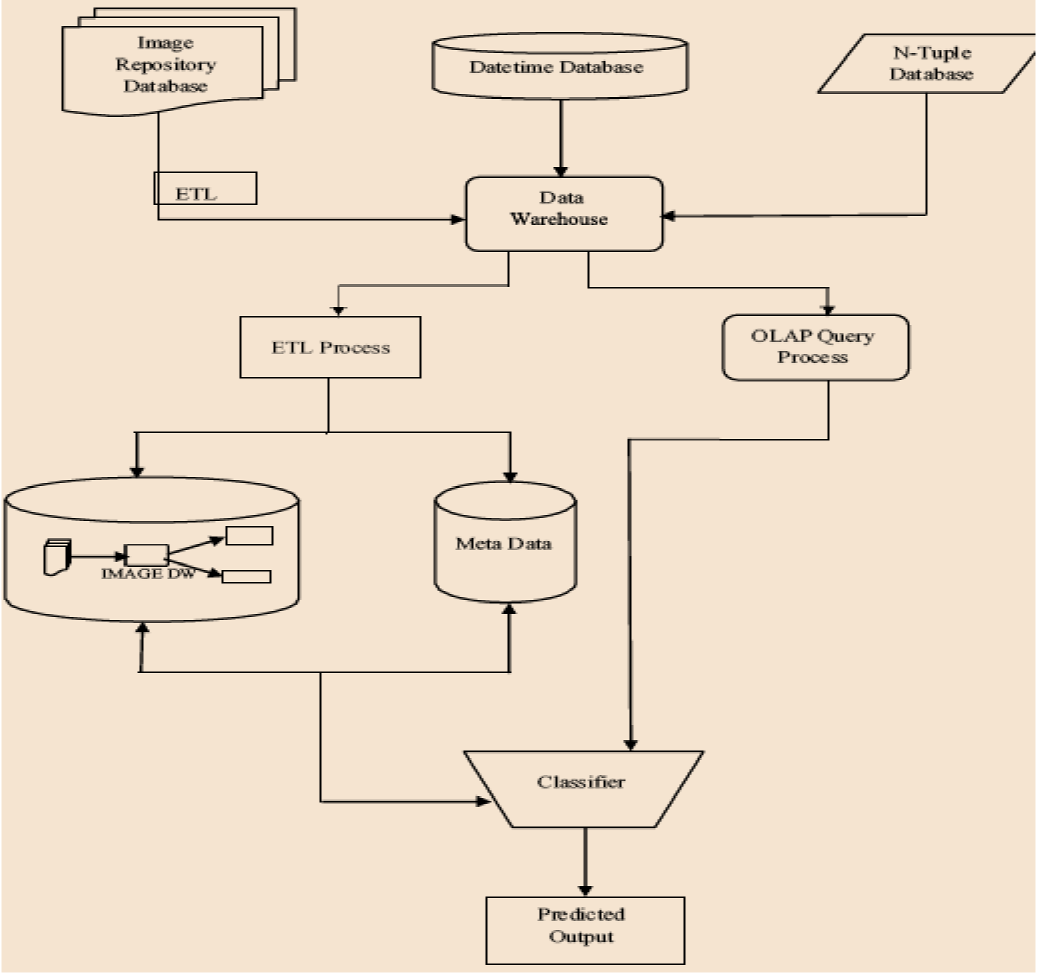

Figure 1: A summary of the proposed system's process implementation

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Source: Hooda and Mann (2020) [1]

From the block diagram above (Fig. 1), four main steps are deduced. The initial step entails the extraction of the input data in the selected data warehouse context. During the second step, there is the ETL process-based feature extraction. Regarding the third step, it involves OLAP process implementation, culminating into the fourth step that is seen to involve classifier models and the provision of the output outcomes.

In this study [1], the number of inputs stood at 115. These inputs were obtained from a source in the form of an online website. Also, the MATLAB software was used in the simulation exercise. Additionally, it is worth remembering that the findings that were obtained after implementing the proposed model of machine learning-based disease severity prediction were compared with those associated with the ANN and SVM algorithms. Here, the motivation behind the conducting of the comparative analysis lay in the quest to discern a model that could be deemed the most effective when it comes to predicting the severity of health conditions. Therefore, there was the severity prediction exercised based on the appropriate classification of brain-linked disease, including schizophrenia and bipolar disorder. Some of the variables that were used to allow for the discernment of differences in model performances included the error rate, the elapsed time during the severity prediction process, the recall parameter, model accuracy, model precision, and condition specificity.

To determine the accuracy of the models before discerning one that would be deemed the best, there was the determination of the truly classified samples' ratio relative to the total samples – in terms of false-negative (FN), false positive (FP), true negative (TN), and true positive (TP) [1, 2]. Regarding the specificity parameter, it was established in terms of the TN ratio to the total value of combining FP with TN samples. When it came to the precision parameter, it was obtained by determining the TP ratio to the total value of combined FP and TP samples. In relation to the false positive rate (FPR), its calculation involved the determination of the FP ratio to the total value of the FP and TN. Regarding the false negative rate (FNR), its determination involved the ratio between the FN and the total value of the FN and TP. Relative to the classifier, then, the OLAP query- and ETL-generated input arguments would then be established and the outcomes correlated with the sample data, allowing for the determination of the kernel deemed ideal in relation to the role of the SVM classifier. With the SVM classifier receiving the input data, as well as the ANN, disease severity prediction was determined before allowing for the performance evaluation exercise and also the comparison of the results with those yielded by the proposed model. The interplays above have been summarized as follows [1]:

$$FNR= (FN)/ (TP+FN) \quad (1)$$

$$FPR= FP/ (TN+FP) \quad (2)$$

$$P= TP/ (TP+FP) \quad (3)$$

$$S.P = TN/ (TN+FP) \quad (4)$$

$$ACU = (TP+TN)/ (TP+TN+FP+FN) \quad (5)$$

Regarding the sensitivity and specificity parameters as statistical evaluations, the variables paved the way for the establishment of classification tests' performances. Here, the objective lay in the determination of the state of performance of classification functions that the study employed. Notably, sensitivity entails a state of the TP rate, the recall, and the probabilities of detection. In the current investigation thus, the sensitivity variable or TP rate allowed for the determination of real positives' proportions determined correctly. For instance, the latter scenario was achieved when the study sought to discern the number of sick persons who were most likely to be identified accurately – to be worth diagnosing for a brain disease. Regarding the specificity variable or the TN rate, it allowed for the determination of the number of real negatives that were likely to be identified with accuracy or precision. A specific illustration depicting this exercise was that which involved the determination of the number of healthy persons who were likely to be identified (with precision) as those without brain disease. Hence, the TN rate concentrated on the variable of specificity, with a particular focus on healthy individuals who were identified with precision – not to be exhibiting a brain condition. On the other hand, the TP rate concerned a sensitivity variable that was out to discern sick individuals (with precision) as those who exhibited a brain disease. In the findings, it was documented that the proposed SVM algorithm-based classifier exhibited efficiency relative to disease severity prediction, as well as the least error rate and quick response time [1].

3 Discussion

From the review outcomes, it is evident that mixed outcomes accrue from different scholarly studies conducted from the data warehouse environments,

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

especially by targeting the effectiveness of different machine learning algorithms – when applied to different health conditions. Here, this study has established that in data mining, especially in the healthcare sector, the crucial dilemma lies in the presence of heterogeneous and massive or voluminous medical data. Indeed, this state has been found to exhibit a significant impact on patient treatment and disease diagnosis, as well as the discernment of the severity of health conditions – in the wake of increasing data warehouse environments due to e-health applications that yield electronic health records. Apart from the complexities with which data collection and extraction processes are associated in data warehouse settings, the review outcomes suggest that machine learning techniques have gained increased scholarly attention in relation to their capacity to allow for the prediction of severities with which various health conditions are associated. Indeed, an emerging theme is that the proposed machine learning algorithms exhibit varying but promising degrees of superiority regarding their capacity to allow for the least error rates and quick response times when it comes to disease identification and severity prediction. However, an area that remains dilemmatic involves the manner in which the simulation of the various proposed classifiers could be conducted based on different algorithms and through the incorporation of additional numbers of performance parameters while seeking to ensure more accurate outcomes concerning disease severity prediction.

4 Conclusion

In summary, this review paper sought to demonstrate the manner in which the utilization of classifiers in terms of machine learning models could be implemented in disease severity prediction. The target context of the study entailed warehouse environments. Some of the selected machine learning algorithms that were investigated and insights gained from previous scholarly studies included SVM, ANN, PLS-LDA, and TQWT. Some of the diseases to which the selected algorithms have been applied as classifiers relative to health condition severity prediction include Parkinson's disease, breast cancer, and heart disease. From the results, the paper has established that the use of machine learning algorithms as classifiers towards trend similarity determination and disease severity prediction via predictive analytics is a promising path because it paves the way for the future projections of health condition probabilities and trends in occurrence. Regarding future research thus, there is a need for more scholarly investigations to focus on how the proposed machine learning algorithms could be implemented while incorporating additional numbers of performance parameters.

References

[1] Hooda S and Mann S. Examining the Effectiveness of Machine Learning Algorithms as Classifiers for Predicting Disease Severity in Data Warehouse Environments. *Revista Argentina de Clinical Psicologica* 2020; 14(4), 133-151

[2] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, 15, 104-116

[3] Hooda S. (2020). A Focus on the ICU's Mortality Prediction Using a CNN-LSTM Model. *International Journal of Psychosocial Rehabilitation*, Vol. 24(6), 8045-8050

[4] Hooda, S. and Mann, S., 2020. A Focus on the ICU's Mortality Prediction Using a CNN-LSTM Model. *International Journal of Psychosocial Rehabilitation*, Vol. 24, Issue. 6, pp. 8045-8050

[5] Qureshi, M. A., & Mir, I. A. (2017). Comparative Study of Existing Techniques for Heart Diseases Prediction Using Data Mining Approach. *Asian Journal of Computer Science and Information Technology*, 50-56

[6] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, 2(2), 250-255

[7] Teixeira, J. W., Annibal, L. P., Felipe, J. C., Ciferri, R. R., & de Aguiar Ciferri, C. D. (2015). A similarity-based data warehousing environment for medical images. *Computers in Biology and Medicine*, 66, 190-208

[8] Zia, U. A., & Khan, N. (2017). An Analysis of Big Data Approaches in Healthcare Sector. *International Journal of Technical Research & Science*, 2(4), 254-264

[9] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, 9, 515

[10] Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters* 2019

[11] Huang S, Yang J, Fong S, Zhao Q. Mining Prognosis Index of Brain Metastases Using Artificial Intelligence. *Cancers* 2019; 11(8): 1140

[12] Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications*, 10(6), 261-268

[13] Anitha, S., & Sridevi, N. (2019). Heart disease prediction using data mining techniques. *Journal of Analysis and Computation*, 8(2), 48-55

[14] Annepu, D., & Gowtham, G. (2019). Cardiovascular disease prediction using machine learning techniques. *International Research Journal of Engineering and Technology*, 6 (4), 3963-3971

[15] Benjamin, H., David, F., & Belcy, S. A. (2018). Heart disease prediction using data mining techniques. *ICTACT Journal of Soft Computing*, 9 (1), 1824-1830

[16] Dulhare, U. N. (2018). Prediction system for heart disease using naïve bayes and particle swarm optimization. *Biomedical Research*, 29 (12), 2646-2649

[17] Jagtap, A., Malewadkar, P., Baswat, O., & Rambade, H. (2019). Heart disease prediction using machine learning. In*ternational Journal of Research in Engineering, Science and Management*, 2 (2), 352-355

[18] Khourdifi, Y., & Bahaj, M. (2018). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering and Systems*, 12 (1)

[19] Nashif, S., Raiban, M., Islam, M., & Imam, M. H. (2018). Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. *World Journal of Engineering and Technology*, 6, 854-873

[20] Rammal, H., & Emam, A. Z. (2018). Toward robust heart failure prediction models using big data techniques. In *Proceedings of the Tenth International Conference on e-Health, Telemedicine and Social Medicine*, 85-91

[21] Anita, S., & Priya, P. A. (2017). Estimation of Parkinson's Disease Risk by Statistical Model. IIOAB Journal, 8(3), 42-48