Classification Algorithms for performing Predictive Analysis in Healthcare using Machine Learning Techniques

Afshan Fatima[1], Saurabh Pal[2], Venkateswara Rao Ch[3]
[1]Research Scholar, CSE Department, Vbs Purvanchal University
[2]Professor, CSE Department, Vbs Purvanchal University
[3]Associate Professor, Department of CSE, Siddhartha Institute of Engineering& Technology
[1]afshafatima0889@gmail.com
[2]drsaurabhpalvbspu@gmail.com
[3]chvenkatsrh@gmail.com

Abstract: Machine Learning and Artificial Intelligence have garnered significant attention from researchers in the field of healthcare and medical sciences. The ever-increasing volume, velocity, and variety of healthcare data necessitate the development of an efficient machine learning tool to improve prediction accuracy in the healthcare domain. The primary objective of this paper is to identify the most optimal and appropriate algorithm for disease prediction and diagnosis, and to explore the application of machine learning in healthcare systems. Additionally, this paper provides a comprehensive overview of data science concepts, ranging from data mining techniques to machine learning classification algorithms. The abstract succinctly encapsulates the key aspects of the entire paper, adhering to a predetermined sequence.

Keywords: Artificial Intelligence, Machine Learning, Data Mining, Diagnosis, Classification, Diabetes, Healthcare.

1. Introduction

Due to immense increasing of applications in the field of “Machine Learning (ML)” in the realm of healthcare afford us a glimpse into a future where the amalgamation of data, analysis, and innovation seamlessly collaborate to provide invaluable assistance to countless patients is unrecognized to them. It is not far-fetched to imagine a scenario where ML driven applications, integrated with real-time patient data sourced from diverse healthcare systems across multiple nations, proliferate, thus amplifying the effectiveness of novel treatment options that were once beyond reach.

The domain of healthcare encompasses intricate processes encompassing diagnosis, treatment, and prevention of ailments. The medical industry in numerous countries is undergoing a rapid transformation [1]. The healthcare sector is endowed with numerous amounts of data, including “electronic medical records, administrative reports, and other noteworthy discoveries”, thereby rendering it a data-rich environment.

Health informatics, a research-intensive discipline, has emerged as the principal consumer of public funds. Owing to the advent of computers and novel algorithms, the healthcare landscape has witnessed a proliferation of computational tools that can no longer be ignored. Consequently, healthcare and computing have converged to give rise to health informatics, which is poised to engender enhanced efficiency and efficacy in the healthcare system, while simultaneously ameliorating the quality of healthcare and curtailing costs.

The “Machine learning (ML) and Deep Learning (DL)” algorithms possess an immense potential to elevate the precision of forecasting in the realm of healthcare issues by exceeding the boundaries being established by previous researchers.

2. LITERATURE REVIEW

Traditionally, the field of data mining has been regarded as a statistical learning approach, focusing on the extraction of valuable insights and robust features for data analysis. The primary objective is to construct prediction or clustering models based on the available data [1]. However, this process is not without its challenges, particularly in terms of pre-processing complex data and leveraging domain knowledge expertise [2].

Fortunately, recent advancements in machine learning technologies have opened up new avenues for obtaining end-to-end learning models from intricate data structures. In this enlightening research article, esteemed researchers delve into the existing literature, exploring the application of machine learning technologies in the realm of healthcare [3].

When compared to conventional prediction algorithms, machine learning algorithms have showcased remarkable performance in terms of prediction accuracy. In fact, these algorithms exhibit maximum accuracy and boast a convergence speed that surpasses that of other disease risk prediction methods [4]. The impact of machine learning extends beyond the healthcare domain, as it has made significant contributions to various disciplines, including vision and natural language processing [5].

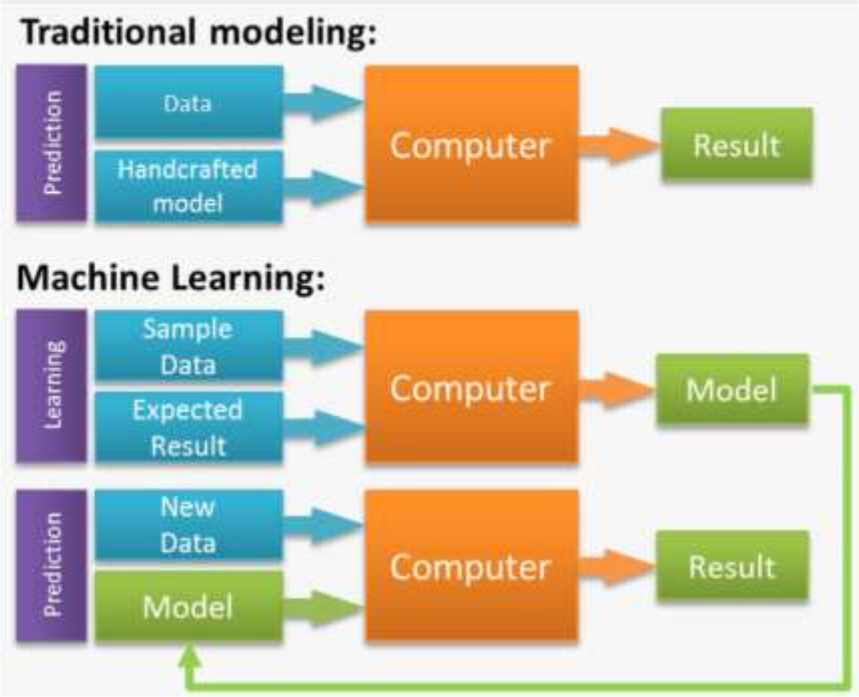

Figure 1. Traditional vs. machine learning approach

In the conventional methodology of data analysis, the initial step entails presenting the machine with a pre-defined model. However, in the realm of ML, the process takes a contradictory route. Instead, it commences with the data itself, allowing the extraction of a model that can subsequently be employed to process novel data [6]. And some of the data mining techniques are:

2.1 Association:

In the realm of association, the puzzling art of data mining unveils intricate patterns, delicately woven together by the interconnection of items within a singular operation. This puzzling technique, also known as the relation technique, finds its purpose in the depths of market basket analysis, where it uncovers the hidden relationships between products. Moreover, its enigmatic powers extend beyond the realm of commerce, delving into the mysterious associations between diseases and their symptoms, unearthing the interlinkages that bind them [7].

2.2 Classification:

Classification, a timeless technique rooted in the art of ML which confers upon us the power to categorize items, arranging them meticulously into predefined clusters or groups. Using a tapestry of mathematical techniques such as “decision trees, linear programming, neural networks” and the mystical realm of statistics, we embark on a journey to distinguish the true nature of each item and assign it to its rightful place. Within the realm of healthcare the classification algorithms grace its own uniqueness through “K-Nearest Neighbor [8], Decision Tree [9], Support Vector Machine, Neural Networks and Bayesian Methods” which performs probabilistic enlightenment.

Within the vast domain of healthcare, “breast cancer” stands as a formidable adversary, a threat that afflicts countless women. Many researches were conducted tha employees the mystical powers of the Weka tool to analyze a breast cancer dataset. The performance of various classifiers was meticulously evaluated, employing the sacred 10-fold cross-validation method. As the results unfolded, data mining revealed its potential to confer invaluable benefits upon the blood bank sector. In this realm, the J48 algorithm and the revered WEKA tool emerged as trusted companions, guiding researchers through the warren of discovery. The classification rules, woven with precision and finesse, showcased their prowess in identifying potential blood donors, achieving an awe-inspiring accuracy rate of 89.9%.

2.3 Clusterization:

Within the realm of data analysis, clusterization serves as a potent technique for crafting coherent and consequential clusters of objects that possess akin properties or characteristics. The most implemented technique will clarify various classes and the objects that belong to each class. In contrast, classification techniques involve the allocation of objects to pre-established classes. Clusterization represents a ubiquitous descriptive undertaking, wherein the aim is to ascertain a limited assortment of categories, classes, or clusters that aptly encapsulate and explicate the data at hand [8]. Authors in reference [9] have ingeniously employed the clusterization methodology, specifically the vector quantization approach, to foretell the likelihood of readmissions in the domain of intensive medicine. The algorithm harnessed for the vector quantization method is none other than the esteemed k-means algorithm.

2.4 Predictions:

Within the realm of data mining, the technique of prediction attentively unravels the complex connections that exist between independent variables and the interplay between dependent and independent variables. A creditable effort by [10] researched into the realm of prognosticating immunize-able diseases. This astute team devised a data mining model, aptly named the Mathematical Model (MM), to forecast the occurrence of diseases that afflict children aged 0 to 5 years. This model was thoughtfully tailored and implemented in six meticulously selected localities within the confines of Osun State, located in Nigeria [11].

2.5 Sequential Patterns:

In the aspect of data mining, the techniques adapted through sequential patterns analysis eagerly accomplishments to extract resemblances, recurring events, and prevailing trends entrenched within transactional data over an extended period of time.

2.6 Decision Tree

The decision tree, a widely utilized technique in data mining, garners popularity due to its user-friendly model. Comprising a root node, branches, and leaf nodes, it forms a hierarchical structure. Notably, decision trees have proven to be exceptional predictors in breast cancer diagnosis and prognosis. With an impressive accuracy rate of 93.62% on benchmark datasets and SEER data set, they outperform other methods [11].

In a research study the power of data mining in uncovering valuable patterns within vast datasets was highlighted whose ability is to extract observable patterns who has shown to enhance the decision-making process in the pharmaceutical industry [12]. Furthermore, the decision trees exhibited remarkable efficacy in predicting patients without heart disease, surpassing the performance of alternative models with an 89% accuracy rate [14].

3. MACHINE LEARNING CLASSIFICATION ALGORITHMS

Following are the main five classifications of algorithm used for analysis in heathcare .

3.1 Categorization:

Categorization is a methodology utilized to organize information into a preferred and distinct number of groupings, where we can allocate designations to each group. The realm of Machine Learning encompasses various applications of Categorization, such as Healthcare Diagnosis, Speech Recognition, Handwriting Recognition, Biometric Identification, and Document Classification. Within the realm of classifiers, we encounter two categories: Binary classifiers and Multi-Class classifiers. Binary classifiers are employed when the classification involves two distinct classes or possesses two potential outcomes. On the other hand, Multi-Class classifiers are utilized when the classification involves more than two distinct classes.

3.2.1 Naive Bayes or Naive Bayes Classifier: Naive Bayes is a probabilistic classifier that draws inspiration from the Bayes theorem. It operates on the simplistic assumption that attributes exhibit conditional independence. The primary advantage of the Naive Bayes algorithm lies in its ability to generate accurate predictions with minimal training data required to estimate the essential parameters. Furthermore, Naive Bayes classifiers boast exceptional speed when compared to more intricate methodologies. Among the various classifiers, the Naive Bayes Classifier emerges as the most efficient, demonstrating the highest rate of correct predictions, with a remarkable accuracy rate of 86.53% for patients diagnosed with heart disease. It is closely followed by Neural Networks and Decision Trees.

3.2.2 Support Vector Machine: Is commonly referred to as SVM, is a powerful algorithm that represents training data as points in space, effectively separating them into distinct categories. The goal is to create a clear gap between these categories, maximizing the width of the gap. Subsequently, new scenarios are mapped into the same space and categorized based on which side of the gap they fall on. SVM utilizes three key parameters: the type of kernel, the gamma value, and the C value. SVM offers several advantages, including its efficiency in high-dimensional spaces and its ability to utilize a subset of training points in the decision-making process. Moreover, it is memory efficient.

3.2.3 K-Nearest Neighbors (KNN): The algorithm classifies an item by considering the majority vote of its neighboring items within the input parameter space. The object in question is assigned to the class that is most prevalent among its "k" nearest neighbors, where "k" is an integer value. This classification is determined through a simple majority vote, taking into account the classifications of the k closest neighboring points. KNN boasts several advantages. First and foremost, it is relatively simple to implement, making it accessible even to those with limited experience. Additionally, it exhibits robustness in the presence of noisy training data, allowing for reliable classification outcomes. Moreover, KNN proves to be particularly effective when dealing with large volumes of training data.

3.2.4 DECISION TREE: The Decision Tree algorithm employs a tree-like model to make decisions. It divides the sample into distinct subsets, known as leaves, based on the most significant distinguishing factors in the input variables. To determine which factor or predictor to choose, the decision tree algorithm evaluates all features and performs a binary split on categorical data. It selects the factor with the least cost, or highest accuracy, and repeats this process recursively until it successfully splits the data into all leaves or reaches the maximum depth. The key advantage of the decision tree is its simplicity in understanding and visualizing, requiring minimal data preparation, and its ability to handle both numerical and categorical data.

3.2.5 RANDOM FOREST: Random Forest (RF) is a collective model that grows multiple trees and classifies objects based on the combined votes of all the trees. The RF classifier acts as a meta-estimator by fitting several decision trees on different subsets of the dataset and leveraging averaging to enhance the predictive accuracy of the model while controlling overfitting. The size of each subset is always the same as the original input sample size, but the samples are drawn with replacement. This approach allows random forest to handle large datasets with high dimensionality, provide insights into variable importance, facilitate data exploration, and handle missing data while maintaining accuracy. One major advantage of the random forest is its ability to reduce overfitting, and it generally outperforms decision trees in terms of accuracy in most cases.

Stephan Dreiseitl et al undertook an evaluation of the discriminative prowess of various machine learning algorithms, including K-nearest neighbors, artificial neural networks (ANNs), logistic regression, support vector machines (SVMs), and decision trees. Their objective was to classify pigmented skin lesions (nevi), dysplastic nevi, or melanoma [15]. Meanwhile, N. G. Maity et al conducted a case study analysis on machine learning and showcased the implementation of Bayesian Inference in diagnosing Alzheimer's disease, leveraging cognitive test results and demographic data. Additionally, they focused on the programmed classification of cell images to ascertain the progression and severity of breast cancer using ANN [16].

The realm of medical applications for machine learning is vast. It encompasses diverse areas such as disease identification and diagnosis, drug discovery and manufacturing, medical imaging diagnosis, personalized medicine, and outbreak prediction. Machine learning's integration into the healthcare industry assumes a pivotal role in predicting and diagnosing diseases. Through thorough analysis of datasets, valuable and concealed knowledge can be unearthed.

4. CONCLUSION

In this comprehensive study, the researcher has meticulously examined numerous research papers pertaining to the utilization of machine learning methodologies in the realm of healthcare applications. The prevalent machine learning classification algorithms, namely Decision Tree and Support Vector Machine, have emerged as the favored choices among researchers for their predictive healthcare investigations, primarily due to their unparalleled accuracy. It is worth emphasizing that the scope of machine learning and artificial intelligence in the healthcare and medical sector is virtually boundless.

Machine learning techniques have proven to be invaluable in streamlining administrative procedures, facilitating disease diagnosis, accurately prognosticating ailments, devising customized treatment regimens, and ultimately aiding in the thorough comprehension and management of various medical conditions. Looking ahead, it is highly plausible that more sophisticated machine learning techniques will be developed, with a specific focus on early disease detection, precise diagnosis, and comprehensive prognosis. As evidenced by the analyzed body of work, it is strongly suggested by researchers that the integration of AI and machine learning methodologies has the potential to revolutionize the translation of vast healthcare datasets into tangible enhancements in human well-being.

References:

[1] Whiting, David & Guariguata, Leonor & Weil, Clara & Shaw, Jonathan. (2011). IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. Diabetes research and clinical practice. 94. 311-21. 10.1016/j.diabres.2011.10.029.

[2] Sherwani, S. I., Khan, H. A., Ekhzaimy, A., Masood, A. & Sakharkar, M. K. Significance of hba1c test in diagnosis and prognosis of diabetic patients. Biomarker Insights11, BMI–S38440 (2016).

[3] NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4*4 million participants. Lancet 2016; http://dx.doi.org/10.1016/S0140-6736(16)00618-8.

[4] Seuring T, Archangelidi O, Suhrcke M. The economic costs of type 2 diabetes: A global systematic review. PharmacoEconomics. 2015; 33(8): 811–31.

[5] IDF Diabetes Atlas, 6th ed. Brussels, International Diabetes Federation; 2013.

[6] Nada Lavrac, "Selected techniques for data mining in medicine" , Artificial Intelligence in Medicine 16 (1999) 3–23

[7] Frawley W, Piatetsky-Shapiro G, Matheus C. Knowledge discovery in databases: an overview. In Piatetsky-Shapiro G, Frawley W, editors. Knowledge discovery in databases. Menlo Park, CA: The AAAI Press, 1991.

[8] HianChyeKoh and Gerald Tan,―Data Mining Applications in Healthcare, journal of Healthcare Information Management – Vol 19, No 2.

[9] Mohammed Ali Shaik and Dhanraj Verma, (2020), Enhanced ANN training model to smooth and time series forecast, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022038, doi.org/10.1088/1757-899X/981/2/022038

[10] Kincade, K. (1998). Data mining: digging for healthcare gold. Insurance & Technology, 23(2), IM2-IM7.

[11] Milley, A. (2000). Healthcare and data mining. Health Management Technology, 21(8), 44-47

[12] Mohammed Ali Shaik, Dhanraj Verma, P Praveen, K Ranganath and Bonthala Prabhanjan Yadav, (2020), RNN based prediction of spatiotemporal data mining, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022027, doi.org/10.1088/1757-899X/981/2/022027

[13] Christy, T. (1997). Analytical tools help health firms fight fraud. Insurance & Technology, 22(3), 22-26

[14] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers is an imprint of Elsevier., 500 Sansome Street, Suite 400, San Francisco, CA 94111, ISBN 13: 978-1-55860-901-3

[15] S.Yamini , Dr.V.Khanaa , Dr.Krishna Mohantha - A State of the Art Review on Various Data Mining Techniques, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Issue 3, March 2016

[16] Mohammed Ali Shaik and Dhanraj Verma, (2020), Deep learning time series to forecast COVID-19 active cases in INDIA: A comparative study, 2020 IOP Conf. Ser.:Mater.Sci.Eng. 981 022041, doi.org/10.1088/1757-899X/981/2/022041

[17] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Mag., pp. 37–54, 1996.

[18] J.-J. Yang, J. Li, J. Mulder, Y. Wang, S. Chen, H. Wu, Q. Wang, and H. Pan, "Emerging information technologies for enhanced healthcare," Comput. Ind., vol. 69, pp. 3–11, 2015.

[19] N. Wickramasinghe, S. K. Sharma, and J. N. D. Gupta, "Knowledge Management in Healthcare," vol. 63, pp. 5–18, 2005

[20] . Mohammed Ali Shaik, "Time Series Forecasting using Vector quantization", International Journal of Advanced Science and Technology (IJAST), ISSN:2005-4238, Volume-29, Issue-4 (2020), Pp.169-175.

[21] Shortliffe, EH.,Perrault, LE., (Eds.). Medical informatics: Computer applications in health care and biomedicine (2nd Edition). New York: Springer, 2000.

[22] Denis Rothman, "Artificial Intelligence by Example"", Ingram short title (2018),1788990544,50-250

[23] Nick Bostrom,"Superintelligence: Paths, Dangers, Strategies", Oxford University Press, 2014, ISBN 0199678111, 9780199678112

[24] Shai Shalev-Shwartz, Shai Ben-David "Understanding Machine Learning", Cambridge University Press,United States of America, ISBN 978-1-107-05713-5

[25] Y. LeCun, Y. Bengio, G. Hinton, Deep learning Nature, 521 (7553) (2015), pp. 436-444

[26] Joshi SR, Parikh RM. India - diabetes capital of the world: now heading towards hypertension. J Assoc Physicians India. 2007;55:323–4

[27] Kumar A, Goel MK, Jain RB, Khanna P, Chaudhary V. India towards diabetes control: Key issues. Australas Med J. 2013;6(10):524–31.

[28] Mohammed Ali Shaik, "A Survey on Text Classification methods through Machine Learning Methods", International Journal of Control and Automation (IJCA), ISSN:2005-4297, Volume12, Issue-6

[29] Kaveeshwar SA, Cornwall J. The current state of diabetes mellitus in India. Australas Med J. 2014;7:45–8"

[30] Global report on diabetes. WHO Library Cataloguing-in-Publication Data, ISBN 978 92 4 156525 7.

[31] Dilip Kumar Choubey, Sanchita Paul and Joy Bhattachrjee, "Soft Computing Approaches for Diabetes Disease Diagnosis: A Survey", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 9, Number 21 (2014) pp. 11715-11726