Proceedings of the 4th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/63/2024.17942

Disease Prediction Models Based on Medical Big Data

Laura Liu

The Affiliated International School (YHV) of Shenzhen University, Shenzhen, China

1768346058@qq.com

Abstract. The advent of big data technology has heralded a transformative era in healthcare, with significant implications for disease prediction. This review article delves into the integration of medical big data in predictive modeling, highlighting the pivotal role of data preprocessing, feature engineering, and machine learning algorithms. We explore the escalating research interest, as evidenced by an upward trend in academic publications from 2010 to 2023. The paper underscores the advantages of big data analytics in healthcare, leading to more accurate and personalized disease predictions. Furthermore, we discuss the importance of interdisciplinary collaboration between data scientists, clinicians, and bioinformaticians in enhancing predictive modeling.

Keywords: Disease prediction models, Big Data, Data Preprocessing.

1. Introduction

The digital revolution has ushered in a new era of healthcare, with big data technology at its forefront. One of the most promising applications of this technological shift is the use of medical big data for disease prediction. This vast reservoir of information encompasses patient demographics, medical histories, diagnoses, treatment options, and outcomes, offering unprecedented opportunities for healthcare advancement, particularly in disease prediction and prevention [1].

Big data analytics in healthcare employs sophisticated statistical and machine learning techniques to analyze extensive health datasets [2]. Such models integrate a wide array of factors, including demographic information, historical health data, lifestyle factors, and genetic predispositions, to assess an individual's risk of developing specific diseases such as diabetes, cardiovascular conditions, or cancer [3].

The incorporation of big data analytics into disease prediction offers several significant advantages. It allows for the integration of a comprehensive set of variables, including genetic information, lifestyle factors, and real-time health monitoring data. Moreover, machine learning algorithms can process these complex datasets to elucidate intricate relationships between risk factors and disease outcomes [4]. This approach has the potential to transform healthcare from a reactive to a preventive model. By identifying at-risk individuals early, healthcare providers can implement targeted interventions and lifestyle modifications, potentially reducing the incidence and severity of diseases [5].

This review article explores the current state of medical big data in disease prediction, highlighting key approaches, challenges, and future directions in this rapidly evolving field. As we continue to refine our analytical techniques and address related challenges, the promise of improved patient outcomes and a more efficient healthcare system becomes increasingly attainable.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/63/2024.17942

2. Literature Survey

Figure 1 presents the annual publication count of scholarly articles retrieved with the search terms "Disease Prediction" and "Medical Big Data" on Google Scholar. The graph demonstrates a consistent upward trajectory in the quantity of relevant academic literature from 2010 through to 2023. The annual count experienced a significant increase, starting from 23,100 publications in 2010 and reaching a zenith of 89,800 in 2022. It is noteworthy that despite a minor decline in the number of papers in 2023 compared to the previous year, the disparity is negligible. This upward trend underscores the escalating research interest and academic focus on the development and implementation of Disease Prediction Models within the realm of Medical Big Data.

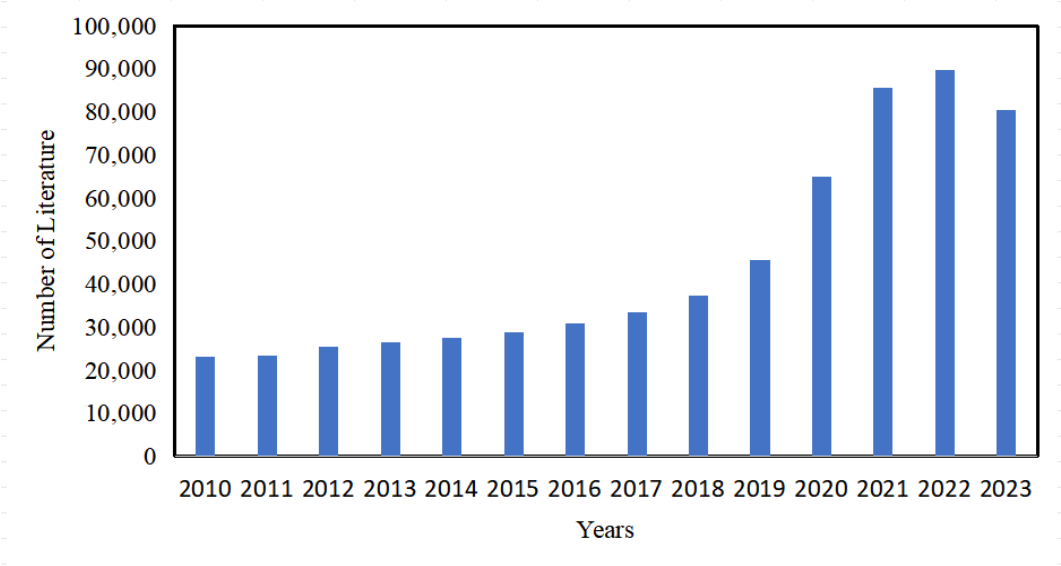

Figure 1. The number of papers searched using “Disease Prediction” and “Medical Big Data” per year

3. Integration of Medical Big Data in Disease Prediction Models

3.1. Data Preprocessing and Feature Engineering

In data preprocessing, the quality and availability of data are generally ensured by processing missing data and dimension reduction. When processing missing data, it can be filled by statistical values (such as mean, median, mode) [6]. Missing values can be predicted using machine learning models, such as random forests [7].

When dealing with large data sets, reducing the data dimension can improve computational efficiency and reduce the computational resources, time, and memory required when making model predictions. Dimensionality reduction techniques such as principal component analysis (PCA) can map data at high latitudes into two or three dimensions for easy data visualization [8]. Dimensionality reduction techniques can help identify and select the most influential features, leading to more concise predictive models [9].

3.2. Model Training and Validation

In the development of predictive models within the realm of big data analytics, the metrics employed for the assessment of model training and validation are crucial. These metrics are pivotal in ascertaining the precision and the capacity of the model to generalize findings across diverse datasets. The key performance indicators (KPIs) that are typically utilized include the accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (ROC-AUC) [10]. Within the domain of medical diagnostics, emphasis is often placed on the recall and accuracy of the predictive model. Recall denotes the proportion of actual positive instances that are accurately detected by the model. Accuracy is

DOI: 10.54254/2753-8818/63/2024.17942

the most elementary and overt indicator of model performance, representing the ratio of total correct predictions made by the model to the total number of predictions [11].

4. Applications and Case Studies

4.1. Applications in Specific Diseases

In the field of contemporary public health, the convergence of big data in different fields, such as medical, transportation and environmental science, provides a more comprehensive and informative foundation for disease prevention [12]. The analytical power of big data enables the construction of predictive models for disease risk assessment by reviewing a wide range of data sets, including electronic medical records, test results, and behavioral patterns. Such models are designed to forecast the potential for developing chronic conditions, such as diabetes and cardiovascular diseases, within individuals or cohorts over a specified temporal frame [13,14]. This foresight helps to implement pre-emptive measures and interventions, thereby improving the prospects of avoiding such diseases.

People use machine learning algorithms such as support vector machines, random forests, and neural networks, these algorithms are able to handle a large number of variables and complex data structures, thus improving the accuracy of predictive models [15]. Machine learning algorithms can also simplify models through dimensionality reduction to assist doctors in making clear decisions in diagnosing and treating patients.

4.2. Case Studies

Eswari et al. (2015) used the predictive analysis algorithm in Hadoop/Map Reduce environment to predict the diabetes types prevalent, complications associated with it and the type of treatment to be provided [16]. Mir & Dhage (2018) aimed at building a classifier model using WEKA tool to predict diabetes disease by employing Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm [17]. Sarwar et al. (2018) discussed the predictive analytics in healthcare, Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes [18]. Hemalatha & Poorani (2021) aimed to develop predictive models to analyze the datasets relevant to heart disease based on random forest, Support Vector Machine, Bayesian prediction and Multilayer Perceptron [19].

5. Future Directions

5.1. Advancement in Predictive Modeling

In the realm of future advancements, the potential of federated machine learning and distributed data processing in healthcare merits extensive exploration. Federated machine learning enables multiple hospitals to utilize data and construct machine learning models while adhering to stringent requirements for user privacy protection, data security, and governmental regulations [20]. By engaging in collaborative model training without the need to share patient data, healthcare institutions can enhance the generalizability and precision of predictive models [21]. This approach not only upholds patient confidentiality but also empowers the development of personalized medicine.

Furthermore, distributed systems facilitate real-time data analysis, which is crucial for time-sensitive clinical decision-making and the surveillance of disease outbreaks [22]. The implementation of such systems bolsters the accessibility and flexibility of data, empowering healthcare providers with more effective tools for managing and analyzing patient information.

5.2. Expanding Applications and Personalization

By employing a variety of predictive algorithms, hospitals can leverage individual genetic data and lifestyle metrics, derived from personal health assessments, to conduct personalized disease predictions. This approach enhances the predictive accuracy of traditional statistical methods. The proliferation of wearable devices and mobile health applications has facilitated the collection of individual lifestyle data,

DOI: 10.54254/2753-8818/63/2024.17942

enabling more convenient predictions of disease likelihood. The application of Artificial Intelligence (AI) and big data technology in the field of precision medicine is also of significant importance. AI algorithms are capable of processing and analyzing vast and complex medical data, thereby identifying disease patterns, predicting disease progression, and proposing personalized treatment recommendations [23].

Furthermore, interdisciplinary collaboration is crucial for the application of big data in disease prediction. This collaboration is divided into three parts: data scientists, clinical physicians, and bioinformaticians. Data scientists are responsible for developing and optimizing various algorithms, clinical physicians provide medical knowledge and patient care experience, while bioinformaticians specialize in the analysis and interpretation of biomedical data. This interdisciplinary collaborative model not only promotes the exchange and integration of knowledge but also accelerates the translation from laboratory to clinical practice, making a substantial contribution to the advancement of disease prediction [24].

6. Conclusion

The integration of medical big data into disease prediction models represents a significant leap forward in the healthcare sector. This review has highlighted the transformative potential of big data analytics, which enables a more nuanced understanding of disease risks through the analysis of extensive and diverse datasets. The application of machine learning algorithms has facilitated the development of predictive models that not only enhance the accuracy of disease prediction but also support the shift towards personalized and preventive healthcare.

The increasing volume of academic literature on this subject reflects a growing consensus on the importance of big data in healthcare. Despite the challenges related to data privacy, security, and the potential for algorithmic bias, the benefits of leveraging big data in disease prediction are manifold. Interdisciplinary collaboration has emerged as a cornerstone in the advancement of predictive analytics. The synergistic efforts of data scientists, clinicians, and bioinformaticians are essential in refining algorithms, interpreting complex biological data, and ensuring that predictive models are clinically relevant and ethically sound.

Looking ahead, the future of disease prediction lies in the continued advancement of predictive modeling techniques and the expansion of big data applications. As we address the current challenges and harness the power of emerging technologies, the vision of a more proactive and efficient healthcare system becomes increasingly attainable.

References

[1] Kumar, S., & Singh, M. (2018). Big data analytics for healthcare industry: impact, applications, and tools. *Big data mining and analytics*, *2*(1), 48-57.

[2] Yoo, C., Ramirez, L., & Liuzzi, J. (2014). Big data analysis using modern statistical and machine learning methods in medicine. *International neurourology journal*, *18*(2), 50.

[3] Damen, J. A., Hooft, L., Schuit, E., Debray, T. P., Collins, G. S., Tzoulaki, I., ... & Moons, K. G. (2016). Prediction models for cardiovascular disease risk in the general population: systematic review. *bmj*, *353*.

[4] Lin, R., Ye, Z., Wang, H., & Wu, B. (2018). Chronic diseases and health monitoring big data: A survey. *IEEE reviews in biomedical engineering*, *11*, 275-288.

[5] Sahoo, P. K., Mohapatra, S. K., & Wu, S. L. (2016). Analyzing healthcare big data with prediction for future health condition. *IEEE Access*, *4*, 9786-9799.

[6] Abouelmehdi, K., Beni-Hssane, A., Khaloufi, H., & Saadi, M. (2017). Big data security and privacy in healthcare: A Review. *Procedia Computer Science*, *113*, 73-80.

[7] Patil, H. K., & Seshadri, R. (2014, June). Big data security and privacy issues in healthcare. In *2014 IEEE international congress on big data* (pp. 762-765). IEEE.

[8] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), 37-52.

DOI: 10.54254/2753-8818/63/2024.17942

[9] Ozsahin, D. U., Mustapha, M. T., Mubarak, A. S., Ameen, Z. S., & Uzun, B. (2022, August). Impact of outliers and dimensionality reduction on the performance of predictive models for medical disease diagnosis. In 2022 International Conference on Artificial Intelligence in Everything (AIE) (pp. 79-86). IEEE.

[10] Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018, March). Prediction of cardiovascular disease using machine learning algorithms. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* (pp. 1-7). IEEE.

[11] Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific reports, 12(1), 5979.

[12] Velmovitsky, P. E., Bevilacqua, T., Alencar, P., Cowan, D., & Morita, P. P. (2021). Convergence of precision medicine and public health into precision public health: toward a big data perspective. Frontiers in Public Health, 9, 561873.

[13] Sonar, P., & JayaMalini, K. (2019, March). Diabetes prediction using different machine learning approaches. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 367-371). IEEE.

[14] Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018, March). Prediction of cardiovascular disease using machine learning algorithms. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (pp. 1-7). IEEE.

[15] Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC research notes, 4, 1-14.

[16] Eswari, T., Sampath, P., & Lavanya, S. J. P. C. S. (2015). Predictive methodology for diabetic data analysis in big data. Procedia Computer Science, 50, 203-208.

[17] Mir, A., & Dhage, S. N. (2018, August). Diabetes disease prediction using machine learning on big data of healthcare. In 2018 fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1-6). IEEE.

[18] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) (pp. 1-6). IEEE.

[19] Hemalatha, D., & Poorani, S. (2021). Machine learning techniques for heart disease prediction. Journal of Cardiovascular Disease Research, 12(1), 93-96.

[20] Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), 305-311.

[21] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., ... & Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific reports, 10(1), 12598.

[22] Joyce, J., Lomow, G., Slind, K., & Unger, B. (1987). Monitoring distributed systems. ACM Transactions on Computer Systems (TOCS), 5(2), 121-150.

[23] Schork, N. J. (2019). Artificial intelligence and personalized medicine. Precision medicine in Cancer therapy, 265-283.

[24] Chao, H. T., Liu, L., & Bellen, H. J. (2017, October). Building dialogues between clinical and biomedical research through cross-species collaborations. In Seminars in cell & developmental biology (Vol. 70, pp. 49-57). Academic Press.