Analyzing the Application of Machine Learning in Anemia Prediction

Yuxi Li

Faculty of innovation Engineering, Macau University of Science and Technology, Chongqing, 402167, China

Abstract. This paper explores the applications of machine learning in the prediction of anemia, highlighting its potential to revolutionize clinical diagnosis and management. Anemia, a prevalent condition affecting millions globally, is often underdiagnosed due to traditional diagnostic methods that rely on clinical judgment and standard laboratory tests. Machine learning techniques provide innovative solutions by analyzing complex datasets that incorporate questionnaire, clinical features, demographic information, and laboratory results, thereby enhancing the accuracy of anemia predictions. This paper examines decision trees, random forests, support vector machines, and neural networks, emphasizing their efficacy in identifying patterns and risk factors associated with anemia. Obstacles such as data quality, feature selection, and model interpretability continue to hinder clinical adoption. The review identifies future research directions aimed at improving model generalizability and interpretability, ensuring that these technologies can be effectively integrated into healthcare practice. This paper advocates for the systematic adoption of machine learning methodologies in anemia management, positing that such innovations are crucial for advancing public health and optimizing resource allocation in clinical settings.

1 Introduction

Anemia is a widespread public health concern characterized by a deficiency of red blood cells or hemoglobin in the bloodstream. This condition can lead to a variety of symptoms, including fatigue, headaches, and palpitations. In severe cases, anemia can elevate the risk of mortality. According to a 2023 World Health Organization survey, nearly 40% of children aged 6 to 59 months, about 37% of pregnant women, and approximately 30% of women aged 15 to 49 years globally are affected by anemia [1]. Therefore, it is important to diagnose and forecast anemia on time.

Nevertheless, Traditional diagnostic methods for anemia primarily rely on laboratory tests, such as complete blood counts and blood chemistry analyses. However, these methods often have limitations, as various factors can influence their results and may lead to delays in treatment.

Machine learning allows computers to learn without being explicitly programmed, and also based on historical dataset to make delicate forecast. Recently, it has been widely used

Corresponding author: 1230005529@student.must.edu.mo

to handle data efficiently and forecast in medical field, especially in early disease prediction and risk assessment. For instance, it is difficult to make model by logic regression. Such as the familial hypercholesterolemia, an arterial thrombotic disorder, and human immunodeficiency virus. However, other types of machine learning models like neural networks which allow transformations of input features to predict outcomes better. It extremely helps researchers process vast amounts of clinical data and uncover potential disease-related features. This offers new approaches and ideas for the early diagnosis and personalized treatment of anemia. In the research which explores the diagnostic procedures between traditional doctors and machine learning in ishaemic heart disease. They use the naive and the semi-naive Bayes and Assistant-R as the skills. Consequently, researchers indicated that step-by-step calculations for post-test probability could significantly enhance the accuracy of machine learning models. These algorithms have shown a 6% improvement in correctly classifying cases as positive or negative. When the probability of diagnosing or ruling out the disease surpasses 0.90, the approach for diagnosing ischemic conditions becomes highly reliable. Additionally, the naive Bayesian classifier has proven to be more effective than traditional methods, improving the accuracy of positive classifications by 17% and negative classifications by nearly 37% [2].

The objective of this research has three points. First of all, the application of machine learning in anemia prediction will be explored by analyzing relevant literature and specific case studies. Secondly, assessing the accuracy and practicality of machine learning models in predicting anemia. Last but not least, discussing their potential and challenges in clinical practice.

2 Machine Learning in Anemia Prediction

Anemia is a worldwide public health issue. It has multiple pathogenesis, including iron deficiency anemia, megaloblastic anemia, regenerative anemia and so on. And among these types, the most common one is iron deficiency anemia. For this, 2% of the mature men are diagnosed, the female of that who are non-Hispanic whites are from 9% to 12%, and also the prevalence of non-Hispanic white Black and Mexican American women are about 20% in general [3]. Moreover, megaloblastic anemia is causing by the factors such as lacking of iron, malabsorption and blood loss. By contract, megaloblastic anemia is often associated with vitamin B12 or folate deficiency, while aplastic anemia is associated with bone marrow dysfunction. Anemia can result from both nutritional and non-nutritional factors or a combination of the two. Nutritional causes include inadequate intake or poor absorption of essential micronutrients. Non-nutritional causes encompass zoonotic diseases such as malaria, helminth infections, or schistosomiasis, as well as chronic inflammatory conditions and genetic hemoglobin disorders. Among nutritional causes, iron deficiency is the most prevalent. Data indicates that iron deficiency affects a quarter to half of children aged 6 to 59 months and women aged 15 to 49 years who have anemia. However, in populations with a high burden of infection, the proportion of anemia cases due to iron deficiency may be lower [4].

Nowadays, machine learning technology makes huge contribution to medical fields, especially in the area for diagnosing and forecasting diseases. By analyzing multiple datasets, such as the results from the lab, patient medical history and questionnaire, the machine learning model could discover potentially complex mechanisms in the data. As for random forests, the decision tree classifier gives rise to the Random Forest (RF) method. This set of tree predictors combines the findings of all the trees in collection and makes a prediction based on a majority vote. This can visually display the decision-making process through a tree structure, which is easy to understand and explain.

A decision tree is a tree in which each leaf node represents a decision and each branch node represents a choice among numerous options. It is a type of ensemble learning that significantly improves the accuracy and robust of prediction by building multiple decision trees and voting. It's widely used in a variety of disciplines [5].

According to Rahul, ML is categorized into supervised and unsupervised learning. Supervised learning is commonly applied in risk assessment, whereas unsupervised learning is utilized in cases of heterogeneous conditions, such as myocarditis. This process begins by grouping individuals with similar undefined conditions, followed by myocardial biopsies and the application of immunostaining and other techniques to examine the cellular components of each sample. Moreover, as for unsupervised learning. Furthermore, some disasters like anemia and cancer could be defined and predicted earlier using algorithms like support vector machines (SVM), random forests (RF), and deep learning [6]. These measures of machine learning not only improve the accuracy of diagnosis, but also provide more delicate therapeutic schedule and reducing the costs. However, there are still some tasks remaining for machine learning. First and foremost, it may be hard to get enough useful data to training the machine. Next, it's also a problem that programming effective algorithms to support the machine. Last but not least, the need for the interaction between specialized doctors and the machine learning technology is also necessary. Therefore, future research needs to focus more on how to optimize machine learning algorithm to better integrate them into clinical practice and truly realize delicate medicine by using data.

3 Progress in machine learning research of anemia

The machine learning for medical care over the past century has improved dramatically, especially the forecasting application of anemia takes many people's attractions. Recently, researchers use the methods of machine learning to deep analyze and predict anemia, which makes significant progress. With the help of machine learning, a variety of problems can be easily handled, such as discovering the relationships between two variables, classifying concepts by unique standards, predicting with initial features, as well as identifying similar patterns items. As for prediction of anemia, machine learning has already been widely applied to analyze various causes which are related to anemia. For example, the age, gender, dietary habit for patients. And also, Popular machine learning methods are used for anemia forecast. Additionally, popular machine learning methods, like k-Nearest Neighbor (k-NN), Convolutional Neural Networks (CNN) are included for forecasting anemia[7].

SVM, a widely-used supervised learning algorithm, is particularly effective in classification and regression tasks. Its primary goal is to identify the optimal hyperplane that maximizes the margin between different classes, thereby efficiently separating data points from various categories. A study by Tamir explored the use of the SVM computational model in detecting anemia through eye conjunctiva images. With a sample of 19 images, the study achieved an accuracy of 78.9%, correctly identifying 15 out of the 19 cases with known hemoglobin levels. Furthermore, the incorporation of image processing algorithms and computer vision techniques in anemia detection using LS-SVM models demonstrated an accuracy of up to 85%, with a sensitivity of 92% and a specificity of 70%, based on a set of 77 tested images (21 non-anemic and 56 anemic)[8]. In conclusion, the LSSVM reveals advanced properties in the realm of anemia forecast. But it also shows that while ensuring high sensitivity, the specificity is not fully optimized. As a result, this algorithm should be improved deeper in the reality application, in order to strengthen the accuracy and reliability. The further research could consider combining other machine learning technologies or optimizing existing models to enhance specificity, hence enhancing its usefulness in clinical Settings. Therefore, although LS-SVM technology has certain advantages in anemia detection, but it still needs more optimization in practice.

Furthermore, the introduction of deep learning opens up a new direction for anemia prediction. In particular, CNN are used in medical image analysis. Through a deep learning on blood images and bone marrow images, anemia-related pathological changes can be detected. This approach not only improves the identification of anemia types, but can also be used to monitor disease progression. Study leading by Delgado-Rivera obtain a result that 77.58% of the sensitivity is different from which in the laboratory, utilizing segmented images of the detection of eyes conjunctiva. Meanwhile, Dimauro used the images of the conjunctiva of the eyes to detect anemia, proposing capturing of images with the use of k-NN algorithm. This successfully gets to a level of 90.26% of accuracy, including images of non-anemic of 84 patients and images of anemic of 29 patients [9]. To sum up, all this research illustrates that the potential application prospect for deep learning. Although the sensitivity and accuracy are different in those reports from different lab, the overall trend shows that these methods have the abilities to improve the efficiency and accuracy of anemia identification, providing a more reliable tool for clinical diagnosis as well.

In the research of anemia protection, the resource and select process is the fundamental but essential part. The common ways for collecting data are about questionnaire, public health database, and laboratory result. The questionnaire provides widely different situation happening to people in different areas, and the laboratory result gives formal and delicate information. What is more, the huge database discovered by public health database will significantly help recognize the relation between features of different group of people and their health condition

Data preprocessing is the main step to make sure the effectiveness of machine learning models. First of all, as the lack value might be missed, the common ways include mean filling, interpolation, or predictive filling using a specific model. These methods could help to remain the integrity of dataset. Furthermore, the detection and solution of outliers are also important. Through methods such as box plots can identify and eliminate values that significantly deviate from the normal range, then reducing their adverse impact on model training. Nevertheless, it also shouldn't be ignored the data standardization and normalization. These could almost eliminate the influence of different feature dimensions, making model training more stable and improving the convergence speed of the algorithm.

In this study for machine learning in anemia forecast, model construction and evaluation are one of the core stages. Common algorithms of machine learning are about Logistic regression, decision trees, random forests and neural networks. And among them, random forests and decision tree are the most useful methods. The common evaluation for model performance include accuracy, recall, F1 score and Area under ROC Curve (AUC). This index can reflect the performance of models in different aspects, such as identifying positive samples and the overall classification effect. Through using these evaluating method flexibly, it could understand the ability of forecasting and also provide reliable evidence for clinical application. For example, in research called 'An anemia screening tool based on deep learning with conjunctiva images', they use AUC, positive predictive value, PPV and negative predictive value, NPV to test the diagnostic efficiency of this model [10].

4 Conclusion

This research introduces a widespread public health issue called anemia and the reasons causing this disease and also compared machine learning forecast of anemia with the traditional medical therapies, with multiple analysis of the drawbacks in traditional ways. And also, it explores the widely use of machine learning in anemia forecast, by integrating various data sources, such as clinical characteristics. Machine learning models can improve the accuracy and efficiency of anemia prediction. Using flexible algorithms, such as include decision trees, random forests, and neural network, are particularly effective at mining latent

patterns in complex data sets to gain insights into the factors that influence anemia. Although there are still challenges such as data quality, feature selection and model interpretability, which will cause the problems that the machine learning models are suitable for reality application. Future development directions should focus on optimizing the universality and interpretability of the model. Overall, machine learning provides new ideas for the early identification, prevention and treatment of anemia, promotes the progress of personalized medicine, and is expected to play an important role in the future public health field. As the technology works well, it will significantly help to public health. Hence more accurate predictive models can lead to better management strategies and improved health outcomes for populations at risk, ultimately helping to build a healthier society.

Reference

1. World Health Organization, Anemia, 1 May 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/anemia
2. I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, Artif. Intell. Med. 23(1), 89-109 (2001).
3. American Family Physician, Iron Deficiency Anemia, 2007; 75(5): 671-678. [Online]. Available: https://familydoctor.org/familydoctor/en/diseases-conditions/anemia.html
4. G.A. Stevens, C.J. Paciorek, et al., National, regional, and global estimates of anemia by severity in women and children for 2000–19: a pooled analysis of population-representative data, Lancet Haematol. 9(4), e249-e260 (2022).
5. P. Verma, V. Chopra, A Review on Machine Learning Algorithms for Anemia Disease Prediction, May 2022.
6. R.C. Deo, MD, PhD, Machine Learning in Medicine, Circulation 132(17), Nov 20, 2015.
7. P. Cristiano, A survey of machine learning in medical diagnosis, J. Biomed. Informat. 55, 1-12 (2015).
8. J.W. Asare, P. Appiahene, E.T. Donkoh, et al., Detection of anaemia using medical images: A comparative study of machine learning algorithms – A systematic literature review, Informatics Med. Unlocked 40, 2023.
9. G. Delgado-Rivera, A. Roman-Gonzalez, A. Alva-Mantari, et al., Method for the automatic segmentation of the palpebral conjunctiva using image processing, in Proceedings of the IEEE International Conference on Automation and the XXIII Congress of the Chilean Association of Automatic Control (ICA-ACCA), 2018, pp. 1-4.
10. H. Xiaoyan, L. Haoyang, et al., An anemia screening tool based on deep learning with conjunctiva images, J. Med. Imaging 2023.