Machine Learning Models for Classification of Anemia from CBC Results: Random Forest, SVM, and Logistic Regression

Muhammad Rafli Aditya[1], Teguh Sutanto[2], Haldi Budiman[3], M.Rezqy Noor Ridha[4], Usman Syapotro[5], Noor Azijah[6]

[1,2,3,4,5,6]Faculty of Information Technology Islamic University of Kalimantan Muhammad Arsyad Al-Banjari, Indonesia

Email: muhrafliaditya061@gmail.com[1], teguhsutanto17@gmail.com[2], haldibudiman@uniska-bjm.ac.id[3], rezqyridha@gmail.com[4], 05usman.syapotro.study@gmail.com[5], noorazijah444@gmail.com[6]

Abstract

In an effort to increase diagnostic efficiency and accuracy, this work investigates the application of machine learning models Random Forest, SVM, and Logistic Regression for the categorization of anemia. Hematocrit and hemoglobin levels were included in the dataset, which was divided into training and testing sets. Using CatBoost, Random Forest outperformed SVM (82.1%) and Logistic Regression (75.1%) with the greatest accuracy (99.2%). SVM and Logistic Regression work well with simpler data, while Random Forest performs best with intricate medical datasets, which makes it perfect for applications involving the detection of anemia.

Keywords

Anemia, Classification, Random Forest, Logistic Regression, Support Vector Machine

Introduction

Anemia is a condition where the number of red blood cells or hemoglobin (Hb) level in the blood is lower than the normal value for a group of people according to age and sex. In healthy people, red blood cells contain hemoglobin, which is responsible for carrying oxygen and other nutrients such as vitamins and minerals to the brain and body tissues. Normal Hb levels in men and women are different. Hb levels for anemic men are less than 13.5 g/dl, while Hb levels in women are less than 12 g/dl. Anemia can cause several clinical symptoms (Rati Astuti, 2023).

(Raj, 2019) employed machine learning techniques in his earlier study to identify anemia up to 95% of the time using the Random Forest algorithm. In the meanwhile, study by (Yeruva et al., 2021) found that sickle cell anemia could be identified with 76% accuracy using the Support Vector Machine (SVM) method. While SVM's accuracy is not as high as that of Random Forest and Multi-Layer Perceptron (MLP), it is still a promising technique for diagnosing sickle cell and thalassemia red blood cell abnormalities.

Submission: 20 October 2024; **Acceptance:** 21 November 2024

Although previous research has explored anemia prediction using various models, this study offers a novel approach by incorporating the strengths of CatBoost, a stacking framework that has not yet been explored for anemia prediction. This framework aims to address the limitations of individual models and leverage their combined strengths to achieve more accurate anemia predictions.

The paper's remainder covers the methodology, results, discussion, and conclusion. The methodology focuses on applying machine learning models, including Random Forest, SVM, and Logistic Regression, with CatBoost used for boosting. These models excel in handling complex data, improving diagnostic accuracy and speed. SVM and Logistic Regression are less effective but useful for simpler tasks. Future research should incorporate more diverse data and advanced models like CNN or RNN to enhance performance.

Methodology

Random Forest

Random Forest is an advancement of the Decision Tree method that utilizes multiple decision trees, trained on individual samples with random attribute selection. The advantages include improved accuracy with missing data, resistance to outliers, efficient data storage, and feature selection capabilities that enhance performance on large and complex datasets. (Supriyadi et al., 2020).

SVM (Support Vector Mechine)

SVM is a system that employs a hypothesis space in the form of linear functions in a high-dimensional feature space, which is then implemented by a learning algorithm based on optimization theory that incorporates the learning bias from statistical learning. The concept of SVM classification is to determine a line that functions to separate two classes of data. The idea behind SVM is to maximize the separation distance between the data classes. SVM operates on high-dimensional datasets using kernel methods. SVM only uses a few data points that contribute (Support Vectors) to form the model used in the classification process. (N. A. Gifran, R. Magdalena, 2019).

Logistic Regression

Logistic regression is a statistical analysis approach that describes the connection between a response variable (dependent variable) with two or more categories and one or more categorical or interval-scaled explanatory factors (Hendayana, 2015).

Cat Boost Stacking

CatBoost, or categorical boosting, is an algorithm developed by Yandex. CatBoost is one of the algorithms that implements gradient boosting, using binary decision trees as base predictors. CatBoost can handle categorical features, and both ordered features and overfitting are addressed by Bayesian estimators. The CatBoost algorithm uses Prediction Values Change (PVC) or Loss Function Change (LFC) to determine the ranking of features in the developed model. PVC is the default method used in the CatBoost machine learning model. LFC is generally used to rank a specific model among various models. (Purbolingga et al., 2023)

Results and Discussion

Table 1 Comparison of results

Algorithm	Splitting Dataset	Accuracy
Logistic Regression	80% Train, 20% Test	0.7509
Random Forest Using Staking CatBoost	**80% Train, 20% Test**	**0.9922**
Random Forest	80% Train, 20% Test	0.9883
SVM	80% Train, 20% Test	0.8210

This study examines the categorization of anemia using Complete Blood Count (CBC) data using three machine learning models: Random Forest, Support Vector Machine (SVM), and Logistic Regression. The accuracy and efficiency of each model vary, offering important information about which versions are most suited for use in medical settings.

Based on the findings, Random Forest is the best model; it uses CatBoost to achieve an accuracy of 99.2%. Using ensemble decision tree approaches, our model handled complicated datasets with ease and produced a dependable result. Random Forest is a great option for medical data because of its ability to minimize overfitting, as medical data frequently has complex and variable properties.

SVM, with an accuracy of 82.1%, comes in second. Even while SVM doesn't perform as well as Random Forest, it might still be helpful when dealing data that can be divided linearly. It works by identifying the hyperplane that optimizes the margin between classes; nevertheless, it is less adaptable when dealing with non-linear or more complicated data.

Among the models, the performance of Logistic Regression is the lowest, with an accuracy of 75.1%. Logistic regression is a straightforward binary classification model that is frequently used as a benchmark for comparison. It is nevertheless applicable to simpler or linear classification problems, even if it is less appropriate for large, complicated datasets.

The use of machine learning, particularly Random Forest, shows great potential in improving the efficiency and accuracy of anemia diagnosis. Random Forest significantly accelerates the diagnostic process compared to manual methods, reduces the likelihood of human error, and enables more precise detection of anemia. With an accuracy of 99.2% when utilizing CatBoost, Random Forest proves ideal for medical applications that require rapid and accurate predictions.

While, SVM and Logistic Regression, although not as powerful as Random Forest, still have their place in certain classifications, especially for simpler datasets or when model interpretability is crucial. For example, SVM is effective in scenarios where linear separation suffices for identifying anemia, such as in cases of mild anemia, where the relationship between clinical parameters is simpler.

This research is currently limited to CBC data (e.g., hemoglobin and hematocrit), which, while critical, may not encompass all the information needed for more complex diagnoses. Future studies could integrate additional clinical data such as genetic information or patient health history to enhance prediction accuracy. Additionally, exploring other advanced machine learning models could further improve diagnostic outcomes, potentially leading to even more accurate and efficient medical applications.

Conclusion

The study's overall findings demonstrate that anemia may be accurately identified using CBC data and machine learning algorithms, especially Random Forest. When using CatBoost, Random Forest obtained the greatest accuracy of 99.2%, demonstrating its remarkable efficacy in handling complicated datasets and producing accurate predictions.However, while Random Forest outperforms SVM and Logistic Regression in handling more complicated data, they are still useful models for simpler or more linear classification tasks (accuracy of 82.1% and 75.1%, respectively). The benefits of applying machine learning techniques to enhance the precision and efficacy of anemia diagnosis are demonstrated in this work.

Recommendation

Future research should incorporate a more diverse dataset, including additional clinical parameters such as genetic data and health history, to improve the accuracy of anemia diagnoses. Exploring advanced models like CNN or RNN, and employing ensemble techniques to combine different models, could further enhance performance. Additionally, focusing on developing more interpretable models for medical professionals and integrating real-time patient data would support continuous monitoring and diagnosis.

References

Cholil, S. R., Handayani, T., Prathivi, R., & Ardianita, T. (2021). Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *6*(2). https://doi.org/10.31294/ijcit.v6i2.10438

Hendayana, R. (2015). Penerapan Metode Regresi Logistik Dalam Menganalisis Adopsi Teknologi Pertanian. *Informatika Pertanian*, *22*(1), 1. https://doi.org/10.21082/ip.v22n1.2013.p1-9

N. A. Gifran, R. Magdalena, and R. Y. N. F. (2019). Deteksi Anemia Melalui Citra Sel Darah Menggunakan Metode Discrete Wavelet Transform (Dwt) Dan Klasifikasi Support Vector Machine (Svm) Anemia. *EProceedings of Engineering, Vol. 6, No. 2*, *6*(2), 3760–3767.

Pardede, D., Hayadi, B. H., & Iskandar. (2022). Kajian Literatur Multi Layer Perceptron Seberapa Baik Performa Algoritma Ini. *Journal of Ict Aplications and System*, *1*(1), 23–35. https://doi.org/10.56313/jictas.v1i1.127

Purbolingga, Y., Marta, D., Rahmawatia, A., & Wajhi, B. (2023). Perbandingan Algoritma CatBoost dan XGBoost dalam Klasifikasi Penyakit Jantung. *Jurnal APTEK Vol. 15 No 2 (2023) 126-133*, *15*(2), 126–133. http://journal.upp.ac.id/index.php/aptek/article/download/1930/1163/4970

Raj, A. (2019). A Review on Machine Learning Algorithms. *International Journal for Research in Applied Science and Engineering Technology*, *7*(6), 792–796. https://doi.org/10.22214/ijraset.2019.6138

Rati Astuti, E. (2023). JAMBURA JOURNAL OF HEALTH SCIENCE AND RESEARCH LITERATURE REVIEW: FAKTOR-FAKTOR PENYEBAB ANEMIA PADA REMAJA PUTRI LITERATURE REVIEW: FACTORS CAUSES ANEMIA IN ADOLESCENT WOMEN the license CC BY-SA 4.0. *Jambura Journal of Health Science and Research*, *5*(2), 550–561. https://ejurnal.ung.ac.id/index.php/jjhsr/index

Supriyadi, R., Gata, W., Maulidah, N., & Fauzi, A. (2020). Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah. *E-Bisnis : Jurnal Ilmiah Ekonomi Dan Bisnis*, *13*(2), 67–75. https://doi.org/10.51903/e-bisnis.v13i2.247

Yeruva, S., Sharada Varalakshmi, M., Pavan Gowtham, B., Hari Chandana, Y., & Krishna Prasad, P. E. S. N. (2021). Identification of sickle cell anemia using deep neural networks. *Emerging Science Journal*, *5*(2), 200–210. https://doi.org/10.28991/esj-2021-01270