Harnessing Machine Learning Techniques for Intelligent Disease Prediction

Malik Imran Asghar[1, *], Fahad Ahmed[2] and Shan Khan[3]

[1] Department of Computer Science, University of Innsbruck, Innsbruck, 6020, Austria

[2] Faculty of Engineering, University of Central Punjab, Lahore, 54000, Pakistan

[3] Faculty of Information Technology and Computer Science, University of Central Punjab, Lahore, 54000, Pakistan

[*]Corresponding Author: Malik Imran Asghar. Email: Malik.Asghar@student.uibk.ac.at

Abstract: Anemia, a prevalent medical illness characterized by a deficiency in red blood cells or hemoglobin, remains a substantial global health issue. The significance of timely identification and intervention in minimizing adverse health effects cannot be stressed. This research study provides a comprehensive analysis focused on the prediction of anemia utilizing three distinct machine learning (ML) algorithms: Logistic Regression (LR), K-Nearest Neighbors (KNN), and Naive Bayes (NB). The results of this research indicate that the logistic regression algorithm exhibits exceptional predictive capabilities, achieving a notable accuracy rate of 98.95%. The empirical results support the LR technique's superior performance in comparison to the KNN and NB algorithms within the specific context of anemia prediction. This work makes substantial contributions to comprehending the potential advantages of data-driven approaches in enhancing the timely identification of anemia. These findings' benefits are significant in expediting medical procedures and ultimately improving the overall quality of patient care.

Keywords: Anemia; Artificial intelligence (AI); Machine learning (ML); Logistic regression (LR); K-nearest neighbors (KNN); Naive bayes (NB)

1 Introduction

Anemia is distinguished by a reduced quantity or concentration of red blood cells or haemoglobin in the bloodstream [1]. Hemoglobin, a protein found in red blood cells, facilitates the transportation of oxygen from the lungs to other organs inside the human body. Anemia can lead to symptoms such as fatigue, weakness, and cognitive impairment, ultimately leading to a decline in overall quality of life. Severe instances can lead to cardiovascular stress, respiratory difficulties, and an elevated susceptibility to mortality. There are many anemia variants, each characterized by its distinct etiology. Anemia can manifest as either an acute or chronic condition with varying severity.

Depending on the underlying cause, anemia may be treated with iron supplements, vitamin supplements, blood transfusions, or medications to treat the underlying cause [2]. Several clinical procedures, involving medical history [3], physical examination [4], and laboratory tests [5], can detect anemia. Clinical methods are invasive, costly, time-consuming, and has less accuracy.

Several machine learning (ML) algorithms have proven effective in making accurate predictions across diverse domains, including healthcare, weather forecasting, stock price prediction, and product recommendation [15-19,24]. ML is preferred over clinical methods in predicting anemia because it is non- invasive, cost-effective, time-saving, and accurate also blockchain may play vital role in case of providing secure mechanism [20, 21]. This study endeavours to predict the occurrence of anemia through the utilization of ML algorithms, specifically Logistic Regression (LR), K-Nearest Neighbors (KNN), and Naive Bayes (NB).

Section 2 briefly reviews existing related work. Section 3 presents the proposed methodolog

Section 4 presents the simulation and results, while section 5 offers a conclusion and future work.

2 Related Work

Comparative studies have been conducted to evaluate the efficacy of different ML algorithms in detecting anemia, and the results have shown that ML algorithms are effective in detecting anemia. Bevilacqua et al. [6] developed a method involving the acquisition of eye conjunctiva images, from which they estimated blood haemoglobin levels and subsequently predicted, with an accuracy rate of 84.4%, whether the patient was afflicted with anemia or not, employing the SVM algorithm.

Tamir et al. [7] employed a SVM technique to identify anemia from images of the eye conjunctiva, achieving an accuracy rate of 78.9%. Jahidur Rahman Khan et al. [8] developed a ML system utilizing linear discriminant analysis (LDA), CART, KNN, SVM, random forest (RF) and LR algorithms to identify anemia in 600 children, with the RF model achieving the highest accuracy of 68.53%.

Noor et al. [9] introduced an innovative methodology that integrates the utilization of palpebral conjunctiva images acquired by a mobile phone camera. Utilizing the image processing capabilities of MATLAB, this approach facilitates the extraction of the relative proportions of red, green, and blue pixels from the images, hence permitting the prediction of haemoglobin levels. This work aimed to examine the effectiveness of three classifiers: Linear SVM, Coarse Decision Tree (DT), and Cosine KNN. The DT classifier exhibited the best accuracy in detecting anemia, with an outstanding accuracy rate of 82.61%.

Dimauro et al. [10] utilized the KNN algorithm to predict anemia, and when tested on images of both anemic and non-anaemic patients, their approach demonstrated an impressive accuracy rate of 90.26 %.

3 Proposed Methodology

The proposed model's flowchart is shown in Figure 1. It makes use of the Anemia dataset that was taken from Kaggle's respiratory [11].The gender, haemoglobin levels, MCHC (mean corpuscular haemoglobin concentration), MCV (mean corpuscular volume), MCH (mean corpuscular haemoglobin), and a binary 'results' feature are some of the important attributes included in this anemia dataset. Individuals are classified as either "0," indicating that they are not anemic, or "1," indicating that they are. This dataset's primary goal is to predict whether a patient is likely to have anemia

After data is acquired, it goes through a thorough preparation step to ensure it is accurate and complete. Furthermore, the dataset is carefully checked for any missing numbers; if there are any, the proper steps are taken to fix them. After the preprocessing step, the dataset is divided into training and testing sets. About 80% of data is used to train the models, while the remaining 20% is utilized to evaluate the models' performance. Three different ML methods are used on the training dataset to help with predictive modeling. With the data and target variable given, these algorithms are used to train predictive models.

Once the models have been trained well, the testing data is used to make predictions based on the models. The models are tested during this step to see how well they can predict and classify individuals as anemic or not anemic based on the available information.

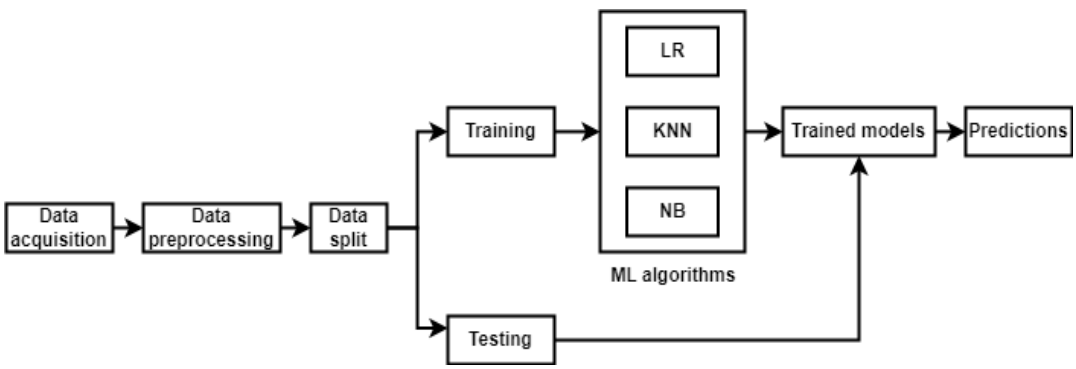

Figure 1. Flowchart of the proposed model

3.1 LR

LR is an essential technique in artificial intelligence (AI) and ML, as it is implemented to train ML models to perform complex data processing tasks autonomously [12]. Using one or more predictor variables, LR is a statistical method for binary classification that estimates the likelihood of a binary outcome. It is called "logistic" because the logistic function is employed to model the correlation between the predictors and the binary response.

3.2 KNN

The KNN algorithm is a supervised machine learning (ML) algorithm that may be utilized for Classification and regression tasks. The discussion method is a non-parametric approach that effectively retains and uses all existing data, enabling the classification of new data points based on their resemblance to the stored data [13]. Determining the number of neighbors to consider while classifying a unique data point is a critical parameter in the KNN algorithm. The KNN algorithm computes the similarity of data points using distance measures, like the Euclidean distance [14].

$$d(a,b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \quad (1)$$

n= n-space

a,b= two points in Euclidean n-space

A_i, b_i= Euclidean vectors, starting from the origin of the space (initial point)

3.3 NB

The NB algorithm is widely known and utilized in ML due to its ease of execution and Effectiveness [22]. The NB algorithm is an ML technique that employs probabilistic principles to classify various tasks [23]. The NB algorithm is a simple approach that operates under the assumption of feature independence, meaning that the value of a particular feature is considered to be unrelated to the value of any other feature, given the class variable. The method relies on the principles of Bayes' theorem and is employed to compute the probability of a hypothesis in light of the available empirical data.

4 Simulation and Results

The simulation and subsequent analysis of results in this study were conducted utilizing Google Colab and the Python programming language. Several performance metrics are employed to assess the proposed model's performance, including accuracy, misclassification rate, precision, sensitivity, specificity, and F1 score (equations 2-7).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} * 100 \quad (2)$$

$$\text{Misclassification rate} = \frac{FP + FN}{TP + FP + FN + TN} * 100 \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} * 100 \quad (4)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} * 100 \quad (5)$$

$$\text{Specificity} = \frac{TN}{TN + FP} * 100 \quad (6)$$

$$\text{F1 score} = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} \quad (7)$$

True positive, false positive, false negative and true negative are denoted by the letters TP, FP, FN, and TN. Figure 2 displays the confusion matrix of three ML algorithms: LR, KNN, and NB.

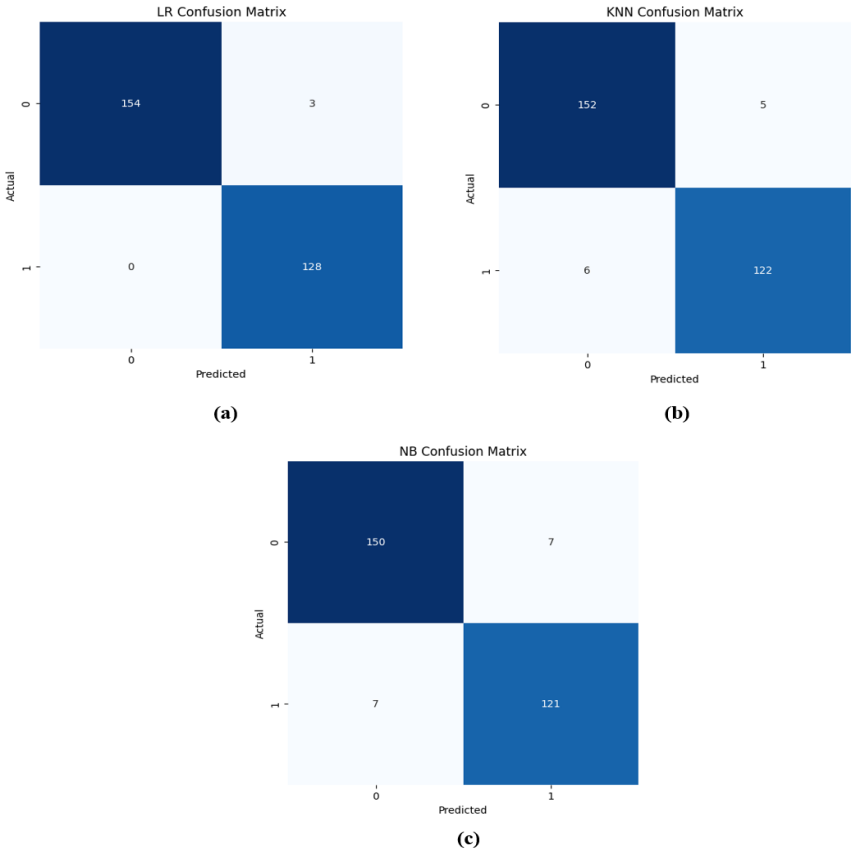

Figure 2. Confusion matrix of ML algorithms

Table 3 presents the comparative analysis with three machine learning algorithms – LR, KNN, and NB. It was identified that the LR algorithm achieves overall good performance compared to other models.

Table 1: Comparative analysis of ML algorithms

Models	Accuracy (%)	Misclassification rate (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1 score
LR	98.95	1.05	98.09	100	97.71	0.99
KNN	96.14	3.86	96.82	96.20	96.06	0.97
NB	95.09	4.91	95.54	95.54	94.53	0.96

As shown in Figure 3, the accuracy comparison is taken between various algorithms. In figure 3, the X-axis indicates the algorithms, the Y-axis indicates the accuracy (%), and the LR algorithm provides better accuracy with 98.95% compared with other algorithms.

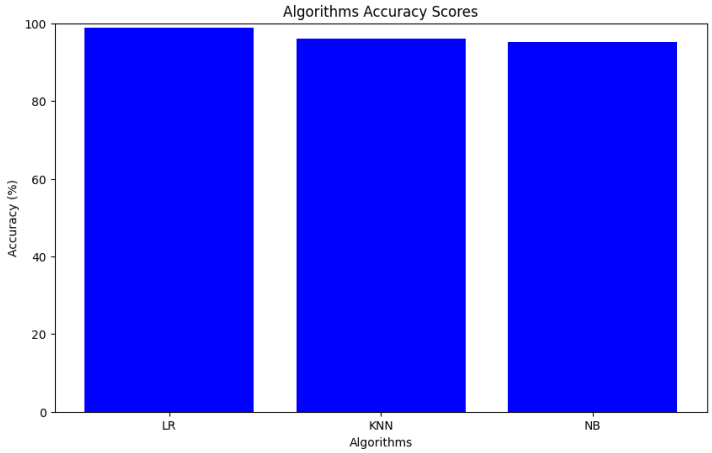

Figure 3. Accuracy comparison of ML algorithms

Table 2 shows the proposed model comparison with previous state of the art methodologies. The proposed model has better accuracy as compared to the previous methodologies.

Table 2: Comparison with state of the art methodologies

Author	Year	Algorithms	Accuracy (%)	Misclassification rate (%)
Bevilacqua et al. [6]	2016	SVM	84.4	15.6
Tamir et al. [7]	2018	SVM	78.9	21.1
Jahidur Rahman Khan et al. [8]	2019	LDA CART KNN SVM RF LR	63.15 62.35 61.95 62.75 68.53 62.75	36.85 37.65 38.05 37.25 31.47 37.25
Noor et al. [9]	2019	SVM DT KNN	73.91 82.61 73.91	26.09 17.39 26.09
Dimauro et al. [10]	2020	KNN	90.26	9.74
Proposed Model	2023	LR KNN NB	98.95 96.14 95.09	1.05 3.86 4.91

5 Conclusion and Future Work

In this article, an extensive evaluation of the performance of three distinct ML algorithms has been conducted to predict the occurrence of anemia disease. The experimental findings, as derived from a representative sample dataset analysis, have indicated that the LR classification algorithm consistently exhibits superior predictive accuracy compared to the KNN and NB algorithms. This noteworthy observation underscores the potential of LR as a formidable tool in anemia prediction.

In the future, the field of anemia prediction can benefit from integrating various deep learning methodologies, particularly those tailored to the analysis of image-based datasets. Incorporating such advanced techniques further enhances the accuracy and robustness of anemia prediction systems, thus improving healthcare outcomes.

References

[1] R. Provenzano, E. V. Lerma, and L. Szczech, *Management of anemia*. 2018. doi: 10.1007/978-3-030-91483-7_27.

[2] "Anemia: Causes, Symptoms, Diagnosis, Treatments." https://www.webmd.com/a-to-z-guides/understanding-anemia-basics (accessed Sep. 15, 2023).

[3] M. E. Conrad, *Marcel e. conrad*. Butterworth Publishers, a division of Reed Publishing, 1990.

[4] "How Is Anemia Diagnosed? | Hematology-Oncology Associates of CNY." https://www.hoacny.com/patient-resources/blood-disorders/anemia/how-anemia-diagnosed (accessed Sep. 15, 2023).

[5] "Anemia Testing - Testing.com." https://www.testing.com/anemia-testing/ (accessed Sep. 15, 2023).

[6] V. Bevilacqua *et al.*, "A novel approach to evaluate blood parameters using computer vision techniques," in *2016 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2016 - Proceedings*, 2016, pp. 1–6. doi: 10.1109/MeMeA.2016.7533760.

[7] A. Tamir *et al.*, "Detection of anemia from image of the anterior conjunctiva of the eye by image processing and thresholding," *5th IEEE Region 10 Humanitarian Technology Conference 2017, R10-HTC 2017*, vol. 2018-Janua, pp. 697–701, 2018, doi: 10.1109/R10-HTC.2017.8289053.

[8] J. R. Khan, S. Chowdhury, H. Islam, and E. Raheem, "Machine Learning Algorithms To Predict The Childhood Anemia In Bangladesh," *Journal of Data Science*, vol. 17, no. 1, pp. 195–218, 2019, doi: 10.6339/jds.201901_17(1).0009.

[9] N. Bin Noor, M. S. Anwar, and M. Dey, "Comparative Study between Decision Tree, SVM and KNN to Predict Anaemic Condition," *BECITHCON 2019 - 2019 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health*, pp. 24–28, 2019, doi: 10.1109/BECITHCON48839.2019.9063188.

[10] G. Dimauro *et al.*, "Estimate of Anemia with New Non-Invasive Systems—A Moment of

Reflection,"
Electronics, vol. 9, no. 5, p. 786, 2020, doi: 10.3390/electronics9050780.

[11] "Anemia Dataset | Kaggle." https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset (accessed Sep.15, 2023).

[12] "What is Logistic Regression? - Logistic Regression Model Explained - AWS."https://aws.amazon.com/what-is/logistic-regression/ (accessed Sep. 15, 2023).

[13] "K-Nearest Neighbor (KNN) Algorithm for Machine Learning - Javatpoint." https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (accessed Sep. 15, 2023).

[14] "K-Nearest Neighbor (KNN) Algorithm - GeeksforGeeks." https://www.geeksforgeeks.org/k-nearest-neighbours/ (accessed Sep. 15, 2023).

[15] Ahmed, F., Asif, M. and Saleem, M., 2023. Identification and Prediction of Brain Tumor Using VGG-16 Empowered with Explainable Artificial Intelligence. International Journal of Computational and Innovative Sciences, 2(2), pp.24-33.

[16] Saleem, M., Khan, M.S., Issa, G.F., Khadim, A., Asif, M., Akram, A.S. and Nair, H.K., 2023, March. Smart Spaces: Occupancy Detection using Adaptive Back-Propagation Neural Network. In 2023 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-6). IEEE.

[17] Athar, A., Asif, R.N., Saleem, M., Munir, S., Al Nasar, M.R. and Momani, A.M., 2023, March. Improving Pneumonia Detection in chest X-rays using Transfer Learning Approach (AlexNet) and Adversarial Training. In 2023 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-7). IEEE. [1]

[18] Sajjad, G., Khan, M.B.S., Ghazal, T.M., Saleem, M., Khan, M.F. and Wannous, M., 2023, March. An Early Diagnosis of Brain Tumor Using Fused Transfer Learning. In 2023 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-5). IEEE.

[19] Saleem, M., Abbas, S., Ghazal, T.M., Khan, M.A., Sahawneh, N. and Ahmad, M., 2022. Smart cities: Fusion-based intelligent traffic congestion control system for vehicular networks using machine learning techniques. Egyptian Informatics Journal, 23(3), pp.417-426.

[20] Saleem, M., Khadim, A., Fatima, M., Khan, M.A., Nair, H.K. and Asif, M., 2022, October. ASSMA-SLM: Autonomous System for Smart Motor-Vehicles integrating Artificial and Soft Learning Mechanisms. In 2022 International Conference on Cyber Resilience (ICCR) (pp. 1-6). IEEE.

[21] Malik, J.A. and Saleem, M., 2022. Blockchain and Cyber-Physical System for Security Engineering in the Smart Industry. In Security Engineering for Embedded and Cyber-Physical Systems (pp. 51-70). CRC press.

[22] "Naïve Bayes Algorithm: Everything You Need to Know - KDnuggets."https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html (accessed Sep. 15, 2023).

[23] "What are Naive Bayes classifiers? | IBM." https://www.ibm.com/topics/naive-bayes (accessed Sep. 15,2023).

[24] Saeed, S., Suayyid, S. A., Al-Ghamdi, M. S., Al-Muhaisen, H., & Almuhaideb, A. M. (2023). A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience. Sensors, 23(16), 7273.