G-Tech : Jurnal Teknologi Terapan
Vol. 9, No. 1, 2025, p. 436-444
E-ISSN: 2623-064X | P-ISSN: 2580-8737

Classification of Anaemia Status Using The K-Nearest Neighbor Algorithm

Zulfa Faradila[1], Ahmad Homaidi [2]✉, Jarot Dwi Prasetyo [3]

[1, 3] Computer Science, Faculty of Science and Technology, Universitas Ibrahimy, Indonesia
[2] Information Technology, Faculty of Science and Technology, Universitas Ibrahimy, Indonesia

✉ **Corresponding Author :** Ahmad Homaidi (e-mail: ahmadhomaidi@ibrahimy.ac.id)

Article Information

Article History

Received : January 02, 2025
Revised : January 08, 2025
Published : January 16, 2025

***Keywords*:**

Anaemia;
Classification;
Data Mining;
K-nearest neighbor.

ABSTRACT

Early detection and accurate diagnosis of anemia are crucial for public health management, with conventional methods like complete blood count often being costly and unavailable in remote areas. The use of machine learning techniques, specifically the k-nearest neighbor (KNN) algorithm, shows promise in classifying medical conditions including anemia with competitive accuracy compared to traditional methods. The implementation of KNN not only offers accuracy but also time and cost efficiency, providing reliable results quickly for medical professionals in the field. The algorithm's application involves determining the appropriate k-value for optimal accuracy, calculating distances using Euclidean distance, and voting for class prediction based on nearest neighbors. The analysis showcases the model's efficiency in predicting anemia status with an accuracy of 94.72% and promising precision and recall rates.

INTRODUCTION

Anaemia is a medical condition characterised by low levels of haemoglobin in the blood, which can lead to fatigue, weakness and impaired organ function. According to the World Health Organization (World Health Organization, 2021), anaemia is a global health problem that affects more than 1.62 billion people, mainly in developing countries. Treating anaemia is critical, not only to improve the quality of life of individuals, but also to support the overall productivity of society (Balarajan et al., 2011).

Early detection and accurate determination of anaemia status is crucial for public health management. Conventional methods for diagnosing anaemia, such as complete blood tests, are often expensive and not always available in remote areas. Therefore, there is an urgent need for faster and more efficient alternative methods, such as the use of machine learning techniques (Anggraini et al., 2023; Balarajan et al., 2011; DeZern & Churpek, 2021; Dimauro et al., 2023; Hevessy et al., 2024; Jansen, 2019; Válka & Čermák, 2018).

One of the promising algorithms in data classification is k-nearest neighbor (KNN). KNN is a non-parametric method that does not assume a particular data distribution, so it can be applied to various types of datasets. Several studies have shown that KNN can be used to diagnose various medical conditions, including anaemia, with competitive accuracy over traditional methods (Appiahene et al., 2023; Asare, Appiahene, Arthur, et al., 2023; Asare, Appiahene, Donkoh, et al., 2023; DeZern & Churpek, 2021; Hevessy et al., 2024; Jansen, 2019; Ramzan et al., 2024; Válka & Čermák, 2018) .

The use of KNN in anaemia status determination not only offers advantages in terms of accuracy, but also in terms of time and cost efficiency. By using available data, KNN can provide

fast and reliable results, making it a valuable tool for medical personnel in the field (Dhakal et al., 2023; Sherin et al., 2024). In addition, this research aims to explore the influence of certain parameters, such as age, gender, and nutritional status, on KNN classification results (Sahni et al., 2022).

Through this research, it is expected to make a significant contribution to the development of more efficient and accurate anaemia diagnosis tools, as well as present empirical evidence regarding the effectiveness of the KNN algorithm in the context of public health (Patel et al., 2021; Zhang et al., 2022). As such, this research not only focuses on the technical aspects of the KNN algorithm, but also considers its practical implications in improving anaemia diagnosis, which in turn can contribute to the overall improvement of public health.

RESEARCH METHODS

Good research is certainly carried out with clear and definite stages, as well as this research. This research uses several stages as shown in Figure 1 below;

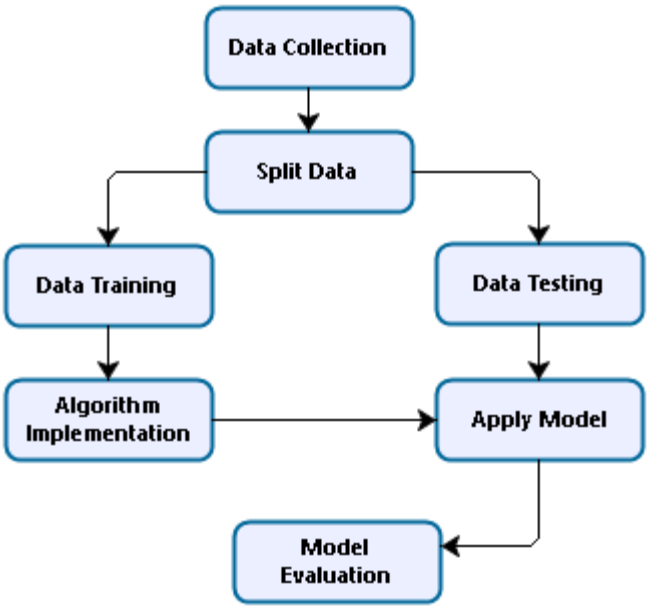

Figure 1. Research stages

Data Collection

Data collection is a crucial first step in this research. The data used in this study is taken from the anaemia dataset available on Kaggle, which can be accessed via the following link: https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset. This dataset contains information on various anaemia-related parameters, including haemoglobin, gender, mean corpuscular haemoglobin, mean corpuscular haemoglobin concentration, mean cell volume and outcome. The dataset includes more than a thousand data entries that allow for more comprehensive analyses (Rao, 2023).

The data in this dataset includes information from a wide range of individuals, allowing for a thorough analysis of anaemia prevalence across different demographic groups. According to available data, the worldwide prevalence of anaemia stands at around 24.8% in the general population, with higher rates in certain groups such as pregnant women and children. Data collection also involves a cleaning and validation process to ensure that the data used in the analysis is accurate and reliable. This is important to avoid biases that could affect the results of the analysis. For example, missing or inconsistent data should be addressed by imputation or deletion methods, depending on the context and proportion of missing data (Little & Rubin, 2019). Thus, the quality of the data collected will improve the accuracy of the model built. For the gender attribute, a record of "0" indicates male gender and "1" indicates female gender. As for the result,

a record of "0" indicates no anaemia, and "1" indicates anaemia. With proper data collection and strict procedures, we can proceed to the next step in this research, which is the separation of data for training and testing the model.

Tabel 1. Anemia dataset

No	Gender	Hemoglobin	MCH	MCHC	MCV	Result
1	1	14.9	22.7	29.1	83.7	0
2	0	15.9	25.4	28.3	72	0
3	0	9	21.5	29.6	71.2	1
4	0	14.9	16	31.4	87.5	0
5	1	14.7	22	28.2	99.5	0
6	0	11.6	22.3	30.9	74.5	1
7	1	12.7	19.5	28.9	82.9	1
8	1	12.7	28.5	28.2	92.3	1
9	0	14.1	29.7	30.5	75.2	0
10	1	14.9	25.8	31.3	82.9	0
.....	.....	.....	.....	.....	.....	.....
1413	1	12.1	18.9	31.2	76	1
1414	0	16	23.6	29.4	95.6	0
1415	0	13.2	25.6	31.1	100.4	0
1416	0	13	26.2	31.3	74.9	0
1417	1	13.2	20.1	28.8	91.2	1
1418	0	10.6	25.4	28.2	82.9	1
1419	1	12.1	28.3	30.4	86.9	1
1420	1	13.1	17.7	28.1	80.7	1
1421	0	14.3	16.2	29.5	95.2	0
1422	0	11.8	21.2	28.4	98.1	1

Split Data

Once the data collection process is complete, the next step is to separate the data into two sets, the training set and the testing set. This separation process is important to ensure that the model built can be tested on data that has never been seen before, so as to measure its ability to accurately predict anaemia status. In this study, the general proportion used is 80% for the training set and 20% for the testing set. The 80%-20% split is considered balanced because it provides enough data for training, but still enough data for evaluation. If too much data is used for training (e.g. 90% or more), it is possible for the model to be too "trained" with the data and become overfit. Conversely, if too much data is used for testing (e.g. 50% for testing), the amount of training data will be very limited and the model may not be able to capture sufficiently complex patterns.

The splitting of the data was done randomly to avoid any bias that might arise if the data were selected systematically. By using stratified sampling method, we can ensure that the proportion of each class (anaemic and non-anemic) is maintained in both sets. This is very important as this dataset may have class imbalance, where the number of anaemic individuals may be much less than the non-anemic ones.

The training set is used to train the KNN model, while the testing set is used to test how well the model predicts anaemia status based on data that has never been seen before. By using cross-validation techniques, researchers can further evaluate the performance of the model and reduce the risk of bias that may arise from unequal data splits (Hastie et al., 2001; Jerome Friedman et al., 2013).

Data segregation also allows researchers to analyse the features that contribute most to anaemia prediction. For example, using feature analysis, researchers can better understand how haemoglobin levels and red blood cell counts relate to anaemia status. This knowledge is not only useful for this study but can also provide valuable insights for health practitioners in identifying and treating anaemia in their population.

Algorithm Implementation

The application of the K-Nearest Neighbor (KNN) algorithm is an important stage in this research. KNN is a machine learning algorithm used for classification and regression, which works by finding the 'nearest neighbour' of a new data point in the feature space. In the context of this research, KNN will be used to classify anaemia status based on the variables that have been collected.

In the application of KNN, choosing the right k value is key to achieving optimal accuracy. Researchers can experiment with various k values to determine the value that gives the best results. For example, previous studies have shown that smaller k values can provide higher accuracy, but also risk noise in the data. Therefore, choosing a balanced k value is crucial in this context.

Once the K value is determined, the next step is to calculate the distance between the new data point and all the data points in the training set. A commonly used method to calculate the distance is the Euclidean distance, which measures the distance between two points in a multidimensional space.

$$ED = \sqrt{(x_1 + y_1)^2 + (x_1 + y_1)^2 + \cdots + (x_n + y_n)^2} \quad (1)$$

After the distance is calculated, the KNN algorithm will select the K nearest neighbours and vote to determine the class of the new data point. If the majority of the nearest neighbours belong to the anaemia class, then the new data point will be classified as anaemic, and vice versa. This process allows the model to learn from the training data and provide accurate predictions on the test data.

$$Y = mode\ (y_1, y_2, \ \ldots\ , y_n) \quad (2)$$

With proper application of the KNN algorithm, we can expect significant results in anaemia status classification. Next, we will evaluate the model to measure its performance using confusion matrix.

Model Evaluation

After the application of the K-Nearest Neighbor algorithm, the evaluation stage becomes very important to assess the performance of the model that has been built. One of the tools used in evaluation is the confusion matrix, which provides a visual representation of the model's performance in classification. The confusion matrix shows the number of correct and incorrect predictions for each class, in this case, anaemia status (anaemia, and not anaemia, and mild anaemia) (Sokolova & Lapalme, 2009).

Confusion matrix consists of four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive is the number of correct predictions for the anaemia class, True Negative is the number of correct predictions for the non-anaemia class, False Positive is the number of incorrect predictions for the anaemia class, and False Negative is the number of incorrect predictions for the non-anaemia class. By calculating these metrics, we can determine the accuracy, precision, recall, and F1-score of the model.

$$Accuration = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

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

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (5)$$

$$\text{F1} - \text{Score} = 2\ \text{x}\frac{\text{Presisi x Recall}}{\text{Presisi + Recall}} = \frac{2\ \text{x TP}}{2\ \text{x TP+FP+FN}} \quad (6)$$

Accuracy is calculated by dividing the number of correct predictions by the total number of predictions. Precision measures the proportion of correct positive predictions, while recall measures the proportion of actual positives successfully predicted. F1-score is the harmonic mean of precision and recall, providing a more comprehensive picture of the model's performance in situations where there is class imbalance.

The results of the confusion matrix and other performance metrics provide valuable insight into the strengths and weaknesses of the KNN model in anaemia status classification. For example, if the model shows a high False Negative rate, this could indicate that the model is likely to fail to detect anaemia in individuals who actually suffer from the condition. This is important to note, as early detection of anaemia can have a significant impact on an individual's health management.

Using a comprehensive evaluation, this study aims to ensure that the developed KNN model is not only accurate but also effective in clinical practice. The results of this evaluation will form the basis for further improvements in the model and can also provide insights for future research into the use of the algorithm in the detection of other medical conditions.

RESULTS AND DISCUSSION

Dataset visualisation is a crucial first step in data analysis, including anaemia status determination. Scatter plot visualisations can be used to illustrate the relationship between haemoglobin levels and other variables, such as age or gender. Using tools such as RapidMiner, we can easily create these visualisations and perform further analysis to understand the patterns present in the dataset. The operator used in visualising this data is Read CSV, because the dataset obtained is stored in .csv format. Good visualisation not only helps in understanding the data, but can also help in communicating the findings to wider stakeholders (Musa et al., 2018).

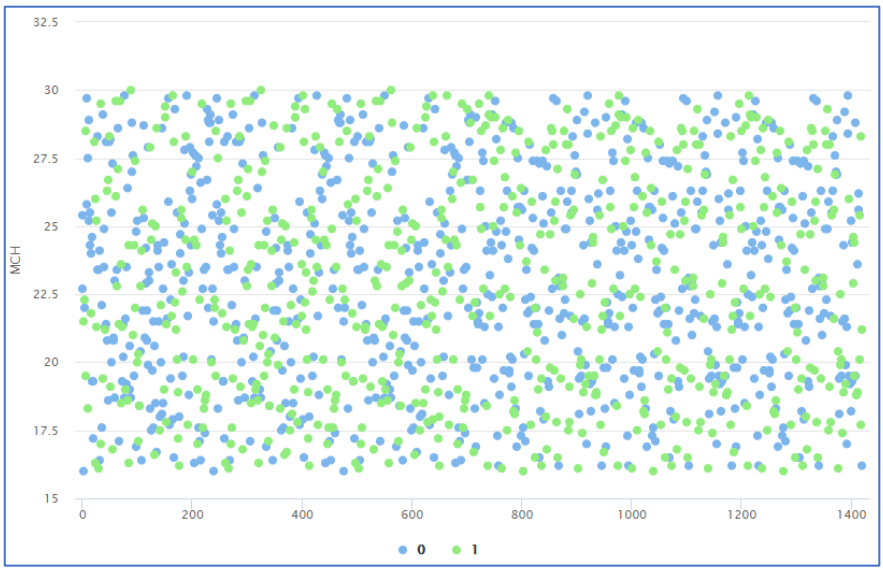

Figure 2. Data visualisation

The data visualization in Figure 2 above explains the distribution of anemia and non-anemia status based on mean corpuscular hemoglobin. Mean corpuscular hemoglobin is the average amount in each of your red blood cells of a protein called hemoglobin, which carries oxygen throughout the body. The use of visualisation also allows the identification of potential problems in the data, such as missing values or inconsistent data. For example, if there are many missing values in the haemoglobin level column, this may affect the accuracy of the KNN model to be built. Therefore, it is important to perform data cleaning before proceeding to the next stage (Ilyas & Rekatsinas, 2022).

In this study, dataset visualisation not only serves as a tool for initial exploration, but also as a basis for decision-making in modelling. With a better understanding of the data, we can choose the right parameters for the KNN algorithm and improve the accuracy of anaemia status prediction.

The implementation of the K-Nearest Neighbor (KNN) algorithm in anaemia status determination involves several key steps. Firstly, we need to separate the dataset into two parts: training data and testing data. The training data will be used to train the model, while the testing data will be used to measure the performance of the model after being trained. This is done in a ratio of 80:20, depending on the size of the dataset available (Bharadwaj et al., 2021).

After dividing the data, the next step is data normalisation. Normalisation is very important in KNN because this algorithm relies on the distance between data points to determine the nearest neighbours. If one variable has a much larger range of values than the others, it can dominate the distance calculation. Therefore, normalisation techniques such as Min-Max Scaling or Z-score Scaling are often used to ensure that all attributes are on the same scale (Noor et al., 2019).

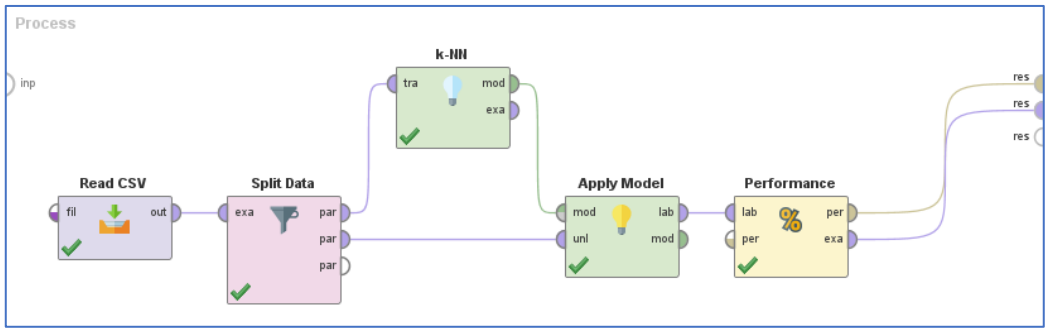

Figure 3. KNN implementation using RapidMiner

Figure 3 as above explains the stages of the implementation process using RapidMiner. Some of the operators used include Read CSV which is used to read the dataset downloaded from kaggle, the split data operator is used to divide the dataset into training data and testing data, k-NN is used to apply the algorithm to the training data, Apply Model is used to form a model and test with testing data, and finally the Performance operator is used to measure the accuracy of the model formed in the previous process.

In RapidMiner, we can easily set the K parameter, which determines the number of nearest neighbours to be considered in decision making. Choosing the right K value is very important, as too small a K value can make the model too sensitive to noise, while too large a K value can cause the model to be too general (Sun et al., 2018). During the training process, the KNN model will calculate the distance between the testing data and all the training data, then select the K nearest neighbours based on the distance. After that, the model will determine the anaemia status based on the majority of the labels of the nearest neighbours. The value in this study is 5 and the distance calculation is done using Euclidean Distance (Alexandropoulos et al., 2019).

Once the model is trained, an evaluation of the model's performance is performed using test data. A commonly used evaluation method in this context is the confusion matrix, which provides an overview of the number of correct and incorrect predictions. In addition, we can also calculate metrics such as accuracy, precision, recall, and F1-score to gain a better understanding of the performance of the KNN model in determining anaemia status.

Tabel 2. Confusion Matrix

	true 0	**true 1**	**class precision**
pred. 0	158	13	92.40%
pred. 1	2	111	98.23%
class recall	98.75%	89.52%	

From the confusion matrix table, we can get information that True Positives (TP): 158, True Negatives (TN): 111, False Positives (FP): 13, and False Negatives (FN): 2. So that from the above, an accuracy of 94.72% can be generated as calculated by the following formula;

$$Accuracy = \frac{158+111}{158+111+13+2} = \frac{269}{284} = 94.72\ \%$$

$$Presision = \frac{158}{158+13} = \frac{158}{171} = 92.40\ \%$$

$$Recall = \frac{158}{158+2} = \frac{158}{160} = 98.75\ \%$$

$$F1 - Score = \frac{2\ x\ 158}{2\ x\ 158+13+2} = \frac{316}{331} = 95.47\%$$

The results of the implementation of the KNN algorithm in determining anaemia status show that this model can provide good accuracy. From the evaluation conducted, the KNN model with the optimal K=5 value showed an accuracy of 94.72%, with precision and recall of 92.40% and 98.75%, respectively. The value of k=5 is chosen so that the distance between neighbors is in the middle, not too small and not too large, while for the selection of odd K values to avoid the same number of neighbors. The test results using the k=7 value resulted in an accuracy rate of 91.55%, while the K=3 value showed an accuracy rate of 95.42%. The difference in the accuracy of both is influenced by the value of K determined, therefore in this study the value of k = 5 is chosen as the middle value, so that the value of the neighbors is not too large and not too small. This shows that the model is able to classify anaemia status well, although there is still room for improvement.

In addition, the analysis results show that the use of the Euclidean distance metric is appropriate for the dataset used in this study, which has a relatively normal data distribution. In the context of real-world applications, the results of this KNN model can be used to assist medical personnel in conducting an initial check of a patient's anaemia status. By using tools such as RapidMiner, this process can be done quickly and efficiently, thus allowing medical personnel to make better and faster decisions in patient handling.

Overall, the results from this study show that the KNN algorithm is an effective tool in anaemia status determination. However, it should be kept in mind that the accuracy of the model can be affected by various factors, including data quality and proper parameter selection. Therefore, further research is needed to develop better and more accurate models for anaemia detection.

CONCLUSION

The application of the K-Nearest Neighbor algorithm for anaemia status determination shows promising results with a very high accuracy of 94.72%. By utilising a dataset containing information from already diagnosed patients, the algorithm can provide accurate predictions of the anaemia status of new patients. KNN's ability to make quick predictions can help medical personnel make better decisions. For example, if a patient presents with suspicious symptoms, using KNN can speed up the diagnosis process and start the necessary treatment earlier. Going forward, further research is needed to explore the integration of KNN with other health technologies, such as artificial intelligence-based health information systems, to improve anaemia detection and management in the community. However, there are some challenges that need to be addressed, such as computation time and selection of the optimal K value. Further research could focus on developing optimisation techniques to improve the efficiency of KNN. Recommendations for future research include exploring the use of ensemble techniques that combine KNN with other algorithms to improve classification accuracy. In addition, the use of larger and more diverse datasets can help in testing the reliability of the model in various conditions.

ACKNOWLEDGEMENTS

The author would like to thank all parties who have assisted in the completion of this research, especially to the head of the computer science study program who has continuously provided support throughout the process of completing this research.

REFERENCES

Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. *Knowledge Engineering Review*, *34*, 1–33. https://doi.org/10.1017/S026988891800036X

Anggraini, P., Kurniawan, T. M., & Lestarini, I. A. (2023). Diagnosis and Management of Anemia in The Elderly. *Jurnal Biologi Tropis*, *23*(4), 150–153. https://doi.org/10.29303/jbt.v23i4.5530

Appiahene, P., Dogbe, S. S. D., Kobina, E. E. Y., Dartey, P. S., Afrifa, S., Donkoh, E. T., & Asare, J. W. (2023). Application of ensemble models approach in anemia detection using images of the palpable palm. *Medicine in Novel Technology and Devices*, *20*, 1–11. https://doi.org/10.1016/j.medntd.2023.100269

Asare, J. W., Appiahene, P., Arthur, E. J., Korankye, S., Afrifa, S., & Donkoh, E. T. (2023). Detection of anemia using conjunctiva images: A smartphone application approach. *Medicine in Novel Technology and Devices*, *18*, 1–12. https://doi.org/10.1016/j.medntd.2023.100237

Asare, J. W., Appiahene, P., Donkoh, E. T., & Dimauro, G. (2023). Iron deficiency anemia detection using machine learning models: A comparative study of fingernails, palm and conjunctiva of the eye images. *Engineering Reports*, *5*(11), 1–21. https://doi.org/10.1002/eng2.12667

Balarajan, Y., Ramakrishnan, U., Özaltin, E., Shankar, A. H., & Subramanian, S. V. (2011). Anaemia in low-income and middle-income countries. *The Lancet*, *378*(9809), 2123–2135. https://doi.org/10.1016/S0140-6736(10)62304-5

Bharadwaj, Prakash, K. B., & Kanagachidambaresan, G. R. (2021). Pattern Recognition and Machine Learning. In *EAI/Springer Innovations in Communication and Computing* (pp. 105–144). https://doi.org/10.1007/978-3-030-57077-4_11

DeZern, A. E., & Churpek, J. E. (2021). Approach to the diagnosis of aplastic anemia. *Blood Advances*, *5*(12), 2660–2671. https://doi.org/10.1182/bloodadvances.2021004345

Dhakal, P., Khanal, S., & Bista, R. (2023). Prediction of Anemia using Machine Learning Algorithms. *International Journal of Computer Science and Information Technology*, *15*(1), 15–30. https://doi.org/10.5121/ijcsit.2023.15102

Dimauro, G., Griseta, M. E., Camporeale, M. G., Clemente, F., Guarini, A., & Maglietta, R. (2023). An intelligent non-invasive system for automated diagnosis of anemia exploiting a novel dataset. *Artificial Intelligence in Medicine*, *136*, 1–11. https://doi.org/10.1016/j.artmed.2022.102477

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning : Data Mining, Inference, and Prediction. In *The Mathematical Intelligencer* (Issue 2). Springer.

Hevessy, Z., Toth, G., Antal-Szalmas, P., Tokes-Fuzesi, M., Kappelmayer, J., Karai, B., & Ajzner, E. (2024). Algorithm of differential diagnosis of anemia involving laboratory medicine specialists to advance diagnostic excellence. *Clinical Chemistry and Laboratory Medicine*, *62*(3). https://doi.org/10.1515/cclm-2023-0807

Ilyas, I. F., & Rekatsinas, T. (2022). Machine Learning and Data Cleaning: Which Serves the Other? *Journal of Data and Information Quality*, *14*(3). https://doi.org/10.1145/3506712

Jansen, V. (2019). Diagnosis of anemia—A synoptic overview and practical approach. *Transfusion and Apheresis Science*, *58*(4), 375–385. https://doi.org/10.1016/j.transci.2019.06.012

Jerome Friedman, by, Hastie, T., Tibshirani John Weatherwax, R. L., & Epstein, D. (2013). A Solution Manual and Notes for: The Elements of Statistical Learning. *Jerome Friedman, by, Hastie, T., Tibshirani John Weatherwax, R. L., & Epstein, D. (2013). A Solution Manual and Notes for: The Elements of Statistical Learning. Http://Doi.Org/10.1007/B94608*.

Little, R. J. A., & Rubin, D. B. (2019). Statistical analysis with missing data. In *Statistical Analysis with Missing Data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Musa, O., Malik, A., Buna, I., Lasena, M., Rizal Isnanto, R., & Suryono, S. (2018). Application Of K-Nearest Neighbor (K-NN) Algorithm On Lung Disease Diagnosis Expert System. *International Journal of Scientific & Engineering Research*, *9*(12), 479–483. http://www.ijser.org

Noor, N. Bin, Anwar, M. S., & Dey, M. (2019). Comparative Study between Decision Tree, SVM and KNN to Predict Anaemic Condition. *IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health*, 24–28. https://doi.org/10.1109/BECITHCON48839.2019.9063188

Ramzan, M., Sheng, J., Saeed, M. U., Wang, B., & Duraihem, F. Z. (2024). Revolutionizing anemia detection: integrative machine learning models and advanced attention mechanisms. *Visual Computing for Industry, Biomedicine, and Art*, *7*(1), 18. https://doi.org/10.1186/s42492-024-00169-4

Rao, B. (2023). *Anemia Dataset*. Kaggle. https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset

Sherin, K., Victoria, A. P. A., Harini, S., & Jensen, J. C. (2024). Automated Diagnosis of Anemia Signs using Machine Learning. *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 1–6. https://doi.org/10.1109/ICRITO61523.2024.10522270

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4). https://doi.org/10.1016/j.ipm.2009.03.002

Sun, J., Du, W., & Shi, N. (2018). A Survey of kNN Algorithm. *Information Engineering and Applied Computing*, 1–10.

Válka, J., & Čermák, J. (2018). Differential diagnosis of anemia. *Vnitrni Lekarstvi*, *64*(5). https://doi.org/10.36290/VNL.2018.067

World Health Organization. (2021, September 30). *The global prevalence of anaemia in 2011*. https://www.who.int/publications/i/item/9789241564960