



Data Mining Techniques for Predictive Classification of Anemia Disease Subtypes

Johan Setiawan¹, Dita Amalia², Iwan Prasetiawan³

^{1,2,3}Department of Information Systems, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia

¹johan@umn.ac.id, ²dita.amalia@student.umn.ac.id, ³iwan.prasetiawan@lecturer.umn.ac.id

Abstract

Anemia, characterised by insufficient red blood cells or reduced haemoglobin, hinders oxygen transport in the body. Understanding anemia's diverse types is vital to tailor effective prevention and treatment. This research explores data mining's role in predicting and classifying anemia types, emphasising Complete Blood Count (CBC) and demographic data. Data mining is key to building models aiding healthcare professionals in anemia diagnosis and treatment. Employing the Cross-Industry Standard Process for Data Mining (CRISP-DM), with its six phases, facilitates this endeavour. Our study compared Naïve Bayes, J48 Decision Tree, and Random Forest algorithms using RapidMiner's tools, evaluating accuracy, mean recall, and mean precision. J48 Decision Tree outperformed the others, highlighting algorithm choice's significance in anemia classification models. Moreover, our analysis identified renal disease-related and chronic anemia as the most prevalent types, with higher occurrences among females. Recognising gender disparities in anemia's prevalence informs tailored healthcare decisions. Understanding demographic factors in specific anemia types is crucial for effective care strategies.

Keywords: anemia; data mining; J48 decision tree; naïve bayes; random forest

How to Cite: J. Setiawan, D. Amalia, and I. Prasetiawan, "Data Mining Techniques for Predictive Classification of Anemia Disease Subtypes", J. RESTI (Rekayasa Sist. Teknol. Inf.), vol. 8, no. 1, pp. 10 - 17, Jan. 2024.

DOI: <https://doi.org/10.29207/resti.v8i1.5445>

1. Introduction

In the current era of information, the vast availability of health data presents exceptional opportunities for delving into profound insights related to diseases that impact the population. A crucial challenge in the healthcare domain is the precise and effective classification of subtypes within the spectrum of anemia. Anemia, a condition that affects over one-third of the global population, is a significant public health concern associated with increased mortality, sickness, reduced workplace productivity, and impaired brain development [1]. Notably, anemia is prevalent in more than 20% of individuals, with a substantial portion (30% to 50%) resulting from iron deficiency, especially among children, adolescents, and adults [2]. Young women, in particular, are susceptible to anemia due to menstrual iron loss and dietary restrictions aimed at achieving a slender physique [1]. Pregnant women also experience a heightened risk of anemia, with implications for both

maternal and fetal health, including complications like premature birth and fetal mortality [3].

In Indonesia, the number of pregnant women climbed from 37.1% in 2013 to 48.9% in 2018, while the proportion of anemia increased for the age groups of 15-24 and 25-34 years [4]. Furthermore, anemia can exacerbate the severity of conditions like COVID-19, as it disrupts haemoglobin and hinders oxygen transport by red blood cells [5]. Despite limited reductions in overall anemia prevalence, from 40% to 33% between 1990 and 2016 [1], the need for timely and accurate prediction in medical science remains imperative for prevention and intervention efforts [6].

Predictive analytics plays a pivotal role in addressing anemia, enabling early detection and tailored treatments to alleviate its impact on individuals and healthcare systems. Data mining, a technique increasingly utilized in medicine, offers opportunities for healthcare professionals to provide cost-effective, quality care to patients [7]. Data mining is currently a

technique utilized to create applications in the field of medicine. Data mining is used in the healthcare industry to assist healthcare professionals in providing appropriate care, which benefits patients by lowering costs and ensuring quality care [8]. Health-related research projects also make use of data mining techniques, particularly categorization and prediction [9].

Research has been done in the past, particularly in comparative data mining predictions of anemia that produce the best accuracy in the J48 Decision Tree method [10] [11][12]. In contrast, research [6] claims that Naïve Bayes has the best accuracy for predicting anemia using data from Complete Blood Counts.

This study predicts the classification of forms of anemia using data mining techniques and the CRISP-DM framework based on prior research references. Referring to earlier studies [11][6] This study took advantage of a larger sample size of the anaemic patients, it employed additional data in the form of demographic information. The number of criteria used to categorize anemia in this study differs from those used in earlier studies [12]. Naïve Bayes, J48 Decision Tree, and Random Forest are the only supervised learning algorithms used in the research [10], [13], which compare them.

This study, building on prior research, aims to classify different types of anemia and develop a highly accurate algorithmic model using data mining techniques.

The paper is organized into four sections, beginning with an overview of anemia's global significance and its specific context in Indonesia (Section 1). The proposed research methodology is detailed in Section 2, while Section 3 presents research findings and in-depth discussions. Finally, Section 4 provides a comprehensive conclusion to this study on predictive analytics in the context of anemia. comprehensively.

2. Research Methods

Predicting anemia and categorizing the many types of anemia are the main points of this research's overall explanation of its purpose. A deficiency of red blood cells in the body causes anemia, a disease. Anemia can be detected with a complete blood count (CBC). A test called a complete blood count (CBC) is used to assess the red, white, and platelet counts in the blood [18]. This examination can find many diseases and problems as well as determine general wellness. Aplastic anemia, chronic anemia, iron deficiency anemia, thalassemia, and anemia due to renal illness are the five different forms of anemia [14], [15]

This research approach combines classification techniques with data mining approaches [14] and uses 15 variables from the Complete Blood Count data collected from the National Health and Nutrition Examination Survey between 2017 and 2018. The use of 2017-2018 data in this study is justified by the

stable nature of anemia-related parameters, the reliability of NHANES data collection, and the richness of the CBC dataset. This temporal focus facilitates a meaningful exploration of anemia subtypes without compromising the validity and applicability of the research to current conditions.

The Naive Bayes, J48 Decision Tree, and Random Forest algorithms are utilized to compare the accuracy results. RapidMiner will be used in this study as a data mining tool [14].

2.1 Data Collection

Table 1. CBC NHANES Data Attribute

No	Attributes	Representation
1	SEQN	Respondent sequence number
2	RIDAGEYR	Age
3	RIAGENDR	Gender
4	RIDETH1	Race/Hispanic origin
5	DMDCITZN	Citizen Status
6	LBXMCVSI	Mean cell volume (fL)
7	LBXHCT	Hematocrit (%)
8	LBXHGB	Hemoglobin (g/dL)
9	LBXMC	Mean Cell Hemoglobin Concentration (g/dL)
10	LBXRDW	Red cell distribution width (%)
11	LBXWBCSI	White blood cell count
12	LBXRBCSI	Red blood cell count
13	BMXWT	Weight (kg)
14	BMXHT	Standing Height (cm)
15	Anemia_prediction	Anemia Prediction

Table 1 lists the properties from the dataset that contain the anemia-related variables or indicators. The Centers for Disease Control and Prevention (CDC) provided the data for this study. They have a regular program called the National Health and Nutrition Examination Survey (NHANES), which is intended to evaluate the nutritional status of adults and children in the United States. Complete blood count (CBC) data from 2017–2018 make up the dataset.

2.2 Research Variables

Based on the variables in the dataset, the impact of anemia disease prediction results is determined. The dependent variable and the independent variable are the two categories into which variables fall [19]. Age, sex, weight, height, race/Hispanic origin, citizen status, mean cell volume, hematocrit, haemoglobin, mean cell haemoglobin concentration, red cell distribution width, white blood cell count, and red blood cell count are the independent factors. The dependent variable is anemia prediction.

2.3 Research Flow

Figure 1 shows the CRISP-DM framework, which comprises 6 phases, which is used in this study as a model stage [20]. CRISP-DM, also known as the Cross-Industry Standard Process Framework for Data Mining, was created in 1996 by analysts from a variety of businesses, including Daimler Chrysler, SPSS, and

NCR. The CRISP-DM standardizes data mining techniques as a generic approach to problem-solving for a business or research unit [16].

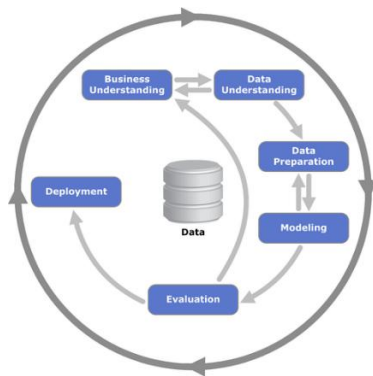


Figure 1. CRISP-DM framework
 Source: Wikipedia

The main goal of model formation in the initial part of this study is to categorize different types of anemia by contrasting three supervised learning algorithms and seeking the best outcomes. Using demographic information, the anemia sufferers identified by the best algorithm will be examined. The NHANES program, run by one of the health agencies in the United States, specifically the CDC, collected the Complete Blood Count data that was used. RapidMiner is a data mining program used in this investigation.

Data from three NHANES datasets, including laboratory data, demographic data, and examination data, will be used in the second stage, which is data understanding. The choice of attributes is modified to account for both naming and research requirements. In the following stage, which involves data preparation with the SEQN ID property, data will be combined.

Data preparation, which will be broken down into three stages (data purification, parameter setting, and split data), will be the third stage. If one of the attributes has a missing value, all of the selected attributes will be destroyed in a single row during the data cleansing. The anemia prediction attribute will be utilized as a label type in the parameter setting section. The classification of each form of anemia is the parameter that should be employed [11]. In the split data, the data will be split into training data (70%) and testing data (30%). The data can be processed to the data modelling stage if the data preparation step is determined to be complete.

Data modelling, which will contrast the Naive Bayes algorithm, Decision Tree, and Random Forest, is the fourth stage. The choice of this approach was made utilizing prior research on data mining-based anemia prediction. Figure 2 illustrates how this model will employ RapidMiner capabilities in conjunction with a flowchart.

The performance of each created model is measured in the fifth stage's evaluation using accuracy, mean recall, and mean precision. The prediction data for anemia patients will be examined with demographic data at the

next stage, namely deployment if the best model performance results are available.

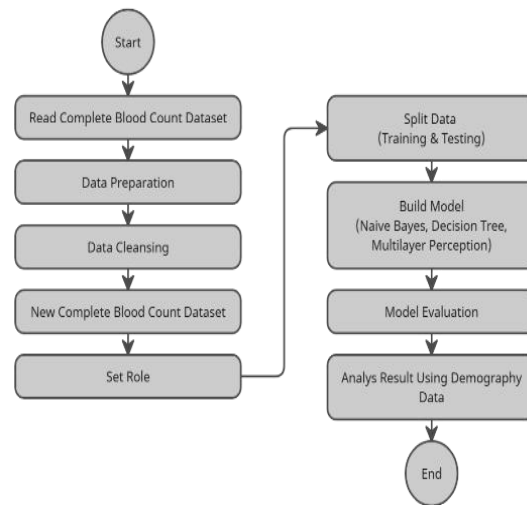


Figure 2. Flowchart Modeling Phase

The deployment of CRISP-DM's final step will make use of data visualization methods. It examines demographic information like gender, average age, weight, height, race/origin category, and citizen category to predict statistics for anemia sufferers from the best model.

3. Results and Discussions

3.1 Business Understanding Phase

This study employs data mining techniques and the CRISP-DM framework to predict the classification [17], [18] of the different forms of anemia. The Naive Bayes, J48 Decision Tree, and Random Forest algorithms will be contrasted. Data from the 2017–2018 NHANES (National Health and Nutrition Examination Survey) health program were used in this investigation. To create a more precise and reliable classification, the data source year is chosen depending on the availability of the most recent data from the NHANES. Data for 2019-2020 have not yet (2021) been released by the relevant organization.

3.2 Data Understanding Phase

The dataset is divided into three categories: demographics, examination, and laboratory. There are 8160 columns and 46 attributes in demographic data. There are 8160 columns and 3 attributes in the examination data. Data from the lab consists of 7844 and 22 characteristics. The SEQN attribute, or respondent sequence number, is the ID for every survey respondent and is present in these three data types. Based on the new NHANES data, there are 69 total attributes and 7844 columns.

3.3 Data Preparation Phase

Data preparation according to requirements is the first step in creating a model. The recommended attributes to be used in modelling are chosen using the select

contributing to the broader understanding of predictive analytics in the domain of this study.

3.5. Evaluation Phase

Based on performance criteria (accuracy, recall, and precision) for each algorithm, the three models of Naive Bayes, J48 Decision Tree, and Random Forest were evaluated in terms of their performance. To compare accuracy, ROC (Receiver Operating Characteristic) data also corroborates the confusion matrix's conclusion. ACD (Chronic Anemia), ARD (Anemia of Renal Disease), IDA (Iron Deficiency Anemia), THAL (Thalassemia), and APA (Aplastic Anemia) are the five categories of anemia types. Here are three confusion matrix findings, one for each algorithm [16].

Table 2. Confusion Matrix Naive Bayes

Predict	True				
	ACD	ARD	IDA	THAL	APA
ACD	65	6	2	0	0
ARD	11	79	1	1	2
IDA	4	2	7	2	0
THAL	0	1	0	0	0
APA	0	0	0	0	0

Based on Table 2's Confusion Matrix and using Naive Bayes, the model classifies 151 patients as TP (True Positive), out of which 65 patients have chronic anemia, 79 have renal disease-related anemia, 7 have iron deficiency anemia, and there are no cases of thalassemia or aplastic anemia.

Table 3. Confusion Matrix J48 Decision Tree

Predict	True				
	ACD	ARD	IDA	THAL	APA
ACD	80	0	0	0	0
ARD	0	92	0	0	2
IDA	0	0	10	0	0
THAL	0	0	0	3	0
APA	0	0	0	0	0

The model classifies TP (True Positive) with a total of 185 patients based on Table 3's Confusion Matrix J48 Decision Tree, with a description of the classification of anemia, Chronic Anemia, 80 patients, Anemia of Renal Disease, 92 patients, Iron Deficiency Anemia, 10, Thalassemia, 3, and No Patients with Aplastic Anemia.

Table 4. Confusion Matrix Random Forest

Predict	True				
	ACD	ARD	IDA	THAL	APA
ACD	79	0	1	0	0
ARD	1	92	0	1	2
IDA	0	0	9	2	0
THAL	0	0	0	0	0
APA	0	0	0	0	0

The model classified TP (True Positive) with a total of 180 patients based on Table 4 of the Random Forest Confusion Matrix, with an explanation of the

classification of anemia, Chronic Anemia in 79 patients, Anemia of Renal Disease in 92 patients, Iron Deficiency Anemia 9 patients, and None Patients with Thalassemia and Aplastic Anemia.

Table 5. Comparison of Algorithm Performance

Algorithm	Accuracy	Recall	Precision
Naive Bayes	92.69%	55.18%	43.29%
J48 Decision Tree	99.89%	71.67%	71.33%
Random Forest	99.61%	64.79%	62.48%

According to Table 5, the J48 Decision Tree algorithm has the most accuracy with a percentage of 99.89%, while Naive Bayes has the lowest accuracy with a percentage of 92.69%. The findings of this accuracy are noteworthy when compared to studies by [19] - [21] but they differ from those of a study by [6], which found that Naive Bayes had the best accuracy.

Table 6 Comparison of Research Findings with Previous Studies

Researcher	Algorithm accuracy (%)			Data set	Attribute	Tools
	NB	J48	RF			
Dita (author)	92.69	99.89	99.61	6,019	15/5	Rapid-Miner
Manal [22]	68.75	93.75	-	41	7/5	WEKA
Meena [12]	-	97	-	9,265	13/1	R Programming WEKA
Manish [23]	96	95.5	95	200	18/5	WEKA
Nurul [24]	70.03	94.27	-	898	13/-	WEKA

This discrepancy may be brought about by variations in the quantity of datasets and technologies employed, as seen by the comparison of research findings with earlier studies in Table 6. In contrast to [12] which used just one anemia parameter variable, namely haemoglobin, this study uses five anemia parameter variables: haemoglobin, hematocrit, mean cell volume, mean cell haemoglobin, and red cell distribution. However, the [25] study did not include anemia parameter factors since it was more concerned with haematological remarks in the form of different blood illnesses than it was with anemia classification.

Demographic information was used to examine the data for each patient based on the J48 Decision Tree model's predictions. The following five data visualizations are based on patient data and include demographic information such as gender, average age, height and weight, race, and country of origin.

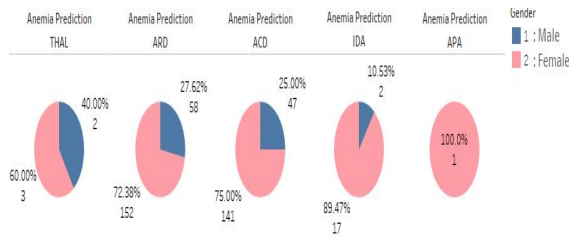


Figure 5. Visualization of Anemia Patients by Gender

According to Figure 5, women predominately experience all types of anemia. The prevalence of the THAL type of anemia is 60% in women and 40% in males. ARD-type anemia affects 72.38% of females and 27.62% of males. The prevalence of this form of ACD anemia is 25% in men and 75% in women. IDA-type anemia affects 89.47% of females and 10.53% of males. Women are entirely affected by the APA form of anemia.

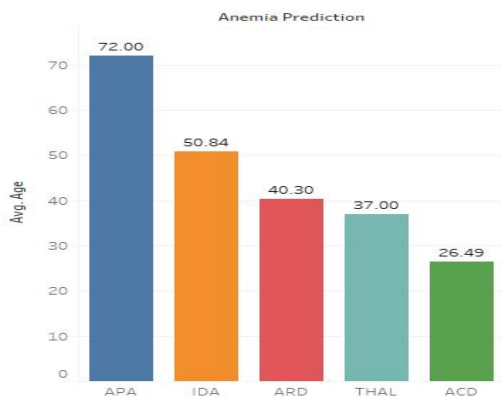


Figure 6. Visualization of Anemia Patients Based on Average Age

The average age of anemia patients differs for each kind of anemia, according to Figure 6. The APA kind of anemia has the oldest average age (72 years), whereas the ACD type of anemia has the youngest average age (26.49 years).

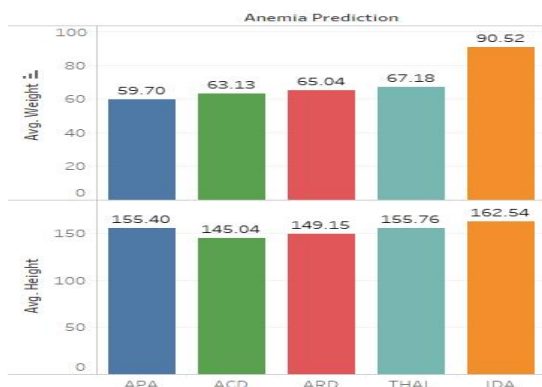


Figure 7. Visualization of Anemia Patients Based on Height and Weight

Each categorisation's average height and weight are different, as shown in Figure 7. The range of the average weight is 59.70-90.52, while the range of the average height is 155.40-162.54. The APA kind of

anemia had the lowest average weight, whereas the ACD type had the lowest average height.

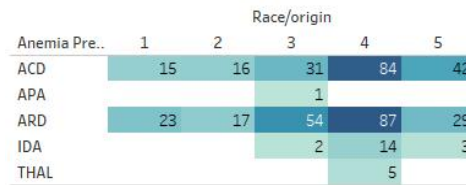


Figure 8. Visualization of Anemia Patients Based on Race/origin

According to Figure 8, there are five categories of race or origin: 1 (Mexican American), 2 (Other Hispanic), 3 (Non-Hispanic White), 4 (Non-Hispanic Black), and 5 (Other Race - Including Multi-Racial). Category 4 (Non-Hispanic Black) has the most anemia patients out of all the racial/ethnic groups, with 190 anemia patients.

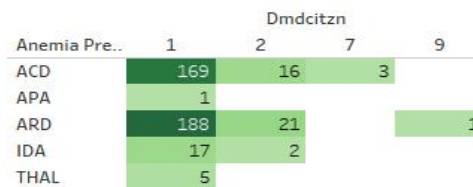


Figure 9. Visualization of Anemia Patients Based on Citizen

Figure 9 shows the classification of citizens into four groups: 1 (Citizen by birth or naturalization), 2 (Not a US citizen), 7 (Refused), and 9 (Don't Know). With 380 anemia patients, category 1 (citizen by birth or naturalization) has the highest number of citizens.

4. Conclusions

This study predicts the classification of the kind of anemia using data mining techniques used in the healthcare industry. Using information from the Centers for Illness Control and Prevention (CDC) from the regular National Health and Nutrition Examination Survey (NHANES), anemia is an illness that may be identified depending on the forms of anemia. The following conclusions can be drawn from the study's findings:

The Naive Bayes, J48 Decision Tree, and Random Forest data mining algorithm models are used to carry out data mining processes. Six stages make up the CRISP-DM framework's cross-industry standard approach for data mining: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. One unique ID attribute, five integer attributes, and ten real attributes make up the attribute data type of the NHANES dataset that was used.

The J48 Decision Tree, which achieved the greatest percentage value of the three performance evaluation factors, is the data mining method that produces the most ideal outcomes, according to the researcher's analysis of three performance assessment variables, namely accuracy (99.89%), recall (71.67%), and precision (71.33%).

The data mining process used reveals fresh information about the classification of the many kinds of anemia, specifically that patients with anemia are overwhelmingly female. Based on averages for age, weight, and height, it generates forecasts for anemia patients that differ depending on the kind of anemia. Patients from the non-Hispanic Black race/origin category have the most patients, followed by patients who are citizens by birth or naturalization in the citizen category.

The research findings offer new information for reducing anemia sufferers through health programs targeted at the non-Hispanic Black female sex population and citizens by birth or naturalization.

Limitation:

While this study provides valuable insights into the prediction and classification of anemia types using data mining techniques and the CRISP-DM framework, it is essential to acknowledge certain limitations. Firstly, the data used in this research is based on the National Health and Nutrition Examination Survey (NHANES) from 2017 to 2018. Given that healthcare data continuously evolves, the use of more recent data could potentially yield different results. Future research needs to incorporate the most up-to-date datasets available to ensure the accuracy and relevance of predictions. Secondly, this study primarily focuses on the application of three specific data mining algorithms (Naïve Bayes, J48 Decision Tree, and Random Forest) to predict anemia types. Other advanced machine learning algorithms and techniques may also contribute to improved prediction accuracy, and future research should explore these alternatives. Additionally, while demographic factors were included to enhance the prediction model, other potential contributing variables, such as dietary habits, genetic predisposition, or environmental factors, were not extensively considered in this study. Future research should explore a broader range of variables to refine the prediction models further.

Future Research:

Building upon the findings of this study, future research avenues can address the following areas:

Incorporating Real-Time Data: Future studies should incorporate real-time or more recent data sources to enhance the accuracy and relevance of predictions. This will ensure that the models remain effective in dynamic healthcare environments.

Exploring Advanced Machine Learning Techniques: Investigating more advanced machine learning techniques, such as deep learning and ensemble methods, can lead to more robust prediction models for anemia classification. These methods may capture complex relationships within the data more effectively.

Including Additional Variables: Expanding the range of variables considered in anemia prediction models,

including lifestyle factors, genetic markers, and environmental conditions, can provide a more comprehensive understanding of anemia risk factors and contribute to more accurate predictions.

Validation and External Testing: Future research should focus on validating the developed models on external datasets to assess their generalizability and performance in different populations and healthcare settings.

Clinical Implementation: Exploring the practical implementation of the developed prediction models within clinical settings can lead to improved anemia diagnosis and personalized treatment strategies, ultimately benefiting patient care and outcomes.

In summary, this study lays a solid foundation for anemia prediction using data mining techniques, but there is ample room for further research and refinement to enhance the accuracy and practical applicability of the models in clinical practice.

Acknowledgement

The accomplishment of this research has been achievable due to the steadfast backing and priceless aid extended by Universitas Multimedia Nusantara.

References

- [1] C. M. Chaparro and P. S. Suchdev, "Anemia epidemiology, pathophysiology, and aetiology in low- and middle-income countries," *Ann. N. Y. Acad. Sci.*, vol. 1450, no. 1, pp. 15–31, 2019, doi: 10.1111/nyas.14092.
- [2] L. I. Vázquez, E. Valera, M. Villalobos, M. Tous, and V. Arija, "Prevalence of anemia in children from latin america and the Caribbean and effectiveness of nutritional interventions: Systematic review and meta-analysis," *Nutrients*, vol. 11, no. 1. 2019, doi 10.3390/nu11010183.
- [3] R. Suryanarayana, M. Chandrappa, A. Santharam, S. Prathima, and S. Sheela, "Prospective study on the prevalence of anemia of pregnant women and its outcome: A community-based study," *J. Fam. Med. Prim. Care*, vol. 6, no. 4, 2017, doi: 10.4103/jfmpe.jfmpe_33_17.
- [4] R. 2018, "Kemenkes RI. Hasil Riset Kesehatan Dasar Tahun 2018. Kementerian Kesehatan RI," *J. Kesehat. Sainatika Meditory*, vol. 1, no. 1, 2018.
- [5] W. Liu and H. Li, "COVID-19: attacks the I-beta chain of haemoglobin and captures the porphyrin to inhibit human heme metabolism," *ChemRxiv*, vol. 12, no. 1, p. 31, 2020.
- [6] M. Jaiswal, A. Srivastava, and T. J. Siddiqui, "Machine learning algorithms for anemia disease prediction," in *Lecture Notes in Electrical Engineering*, 2019, vol. 524, pp. 463–469, doi: 10.1007/978-981-13-2685-1_44.
- [7] M. G. Tuck, F. Alemi, J. F. Shortle, S. Avramovic, and C. Hesdorffer, "A comprehensive index for predicting risk of anemia from patients' Diagnoses," *Big Data*, vol. 5, no. 1, 2017, doi: 10.1089/big.2016.0073.
- [8] F. Akter, M. A. Hossain, G. M. Daiyan, and M. M. Hossain, "Classification of Hematological Data Using Data Mining Technique to Predict Diseases," *J. Comput. Commun.*, vol. 06, no. 04, 2018, doi: 10.4236/jcc.2018.64007.
- [9] A. K. Ramotra, A. Mahajan, R. Kumar, and V. Mansotra, "Comparative Analysis of Data Mining Classification Techniques for Prediction of Heart Disease Using the Weka and SPSS Modeler Tools," in *Smart Innovation, Systems and Technologies*, 2020, vol. 165, doi: 10.1007/978-981-15-0077-0_10.
- [10] R. Abd Rahman, I. B. Idris, Z. M. Isa, R. A. Rahman, and Z. A. Mahdy, "The Prevalence and Risk Factors of Iron

- Deficiency Anemia Among Pregnant Women in Malaysia: A Systematic Review,” *Frontiers in Nutrition*, vol. 9, 2022, doi: 10.3389/fnut.2022.847693.
- [11] P. Verma and V. Chopra, “MACHINE LEARNING ALGORITHMS FOR ANEMIA DISEASE PREDICTION-A REVIEW,” 2304.
- [12] K. Meena, D. K. Tayal, V. Gupta, and A. Fatima, “Using classification techniques for statistical analysis of Anemia,” *Artif. Intell. Med.*, vol. 94, 2019, doi: 10.1016/j.artmed.2019.02.005.
- [13] S. Priya, G. Tripathi, D. B. Singh, P. Jain, and A. Kumar, “Machine learning approaches and their applications in drug discovery and design,” *Chemical Biology and Drug Design*, vol. 100, no. 1, 2022, doi: 10.1111/cbdd.14057.
- [14] D. Papakyriakou and I. S. Barbounakis, “Data Mining Methods: A Review,” *Int. J. Comput. Appl.*, vol. 183, no. 48, pp. 5–19, 2022, doi: 10.5120/ijca2022921884.
- [15] S. Anwar Lashari, R. Ibrahim, N. Senan, and N. S. A. M. Taujuddin, “Application of Data Mining Techniques for Medical Data Classification: A Review,” in *MATEC Web of Conferences*, 2018, vol. 150, doi: 10.1051/mateconf/201815006003.
- [16] D. M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” no. January 2008, 2020.
- [17] R. S. Oetama, Y. Heryadi, Lukas, and W. Suparta, “Improving Candle Direction Classification in Forex Market Using Support Vector Machine with Hyperparameters Tuning,” in *2022 7th International Conference on Informatics and Computing, ICIC 2022*, 2022, doi: 10.1109/ICIC56845.2022.10006993.
- [18] D. A. Kristiyanti, R. Aulianita, D. A. Putri, L. A. Utami, F. Agustini, and Z. I. Alfianti, “Sentiment Classification Twitter of LRT, MRT, and Transjakarta Transportation using Support Vector Machine,” in *2022 International Conference of Science and Information Technology in Smart Administration, ICSINTESA 2022*, 2022, doi: 10.1109/ICSINTESA56431.2022.10041651.
- [19] S. Nazari Nezhad, M. H. Zahedi, and E. Farahani, “Detecting diseases in medical prescriptions using data mining methods,” *BioData Min.*, vol. 15, no. 1, 2022, doi: 10.1186/s13040-022-00314-w.
- [20] “Complete Blood Count (CBC),” 2019, 2019. .
- [21] S. A. King et al., “Search and Selection Procedures of Literature Reviews in Behavior Analysis,” *Perspect. Behav. Sci.*, vol. 43, no. 4, 2020, doi: 10.1007/s40614-020-00265-9.
- [22] M. Abdullah and S. Al-Asmari, “Anemia types prediction based on data mining classification algorithms,” in *Communication, Management and Information Technology - Proceedings of the International Conference on Communication, Management and Information Technology, ICCMIT 2016*, 2017.
- [23] M. Jaiswal, A. Srivastava, and T. J. Siddiqui, “Machine learning algorithms for anemia disease prediction,” *Lect. Notes Electr. Eng.*, vol. 524, no. April, pp. 463–469, 2019, doi: 10.1007/978-981-13-2685-1_44.
- [24] M. N. Amin and A. Habib, “Comparison of Different Classification Techniques Using WEKA for Hematological Data,” *Am. J. Eng. Res.*, no. 43, 2015.
- [25] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM process model,” *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199