Data Mining Techniques for Predictive Classification of Anemia Disease Subtypes

by Johan Setiawan

Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 1 (2024) 10 - 17 e-ISSN: 2580-0760

Data Mining Techniques for Predictive Classification of Anemia Disease Subtypes

Johan Setiawan[1], Dita Amalia[2], Iwan Prasetiawan[3]
[1,2,3]Department of Information Systems, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia
[1]johan@umn.ac.id, [2]dita.amalia@student.umn.ac.id, [3]iwan.prasetiawan@lecturer.umn.ac.id

Abstract

Anemia, characterized by insufficient red blood cells or reduced hemoglobin, hinders oxygen transport in the body. Understanding the various types of anemia is vital to tailor effective prevention and treatment. This research explores data mining's role in predicting and classifying anemia types, emphasizing Complete Blood Count (CBC) and demographic data. Data mining is key to building models that aid healthcare professionals in the diagnosis and treatment of anemia. 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. The J48 Decision Tree outperformed the others, highlighting the importance of algorithm choice in anemia classification models. Furthermore, our analysis identified renal disease-related and chronic anemia as the most prevalent types, with a higher incidence among women. Recognizing gender disparities in the prevalence of anemia informs personalized healthcare decisions. Understanding demographic factors in specific types of anemia 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 to deepen our understanding of the disease that affects 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 more than a third of the global population, is a significant public health concern associated with increased mortality, illness, reduced productivity in the workplace, and impaired brain development [1].

In particular, 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 diet restrictions that aim to achieve a slim physique [1]. Pregnant women also experience an increased risk of anemia, with implications for maternal and fetal health, including complications such as premature birth and fetal mortality [3].

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

Predictive analytics plays an important role in the treatment of anemia, enabling early detection and customized treatments to alleviate its impact on individuals and healthcare systems. Data mining, a technique increasingly used in medicine, offers opportunities for healthcare professionals to provide cost-effective, quality care to patients [7].

Received: 17-09-2023 | Accepted: 25-12-2024 | Published Online: 15-01-2024

Data mining is a technique utilized to create applications in the field of medicine. Data mining is used in the healthcare industry to help healthcare professionals provide adequate care, which benefits patients by reducing 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 comparison of 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 Nave Bayes has the best accuracy to predict anemia using data from Complete Blood Counts.

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

This study, based on previous research, aims to classify different types of anemia and develop a highly accurate algorithmic model using data mining techniques. The article is organized into four sections, beginning with an overview of the global importance of anemia 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 complete blood count (CBC) is used to evaluate red, white and platelet counts in the blood [18]. This examination can find many diseases and problems as well as determine general health. Aplastic anemia, chronic anemia, iron deficiency anemia, thalassemia, and renal disease-related anemia 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 total blood count data collected from the National Health and Nutrition Examination Survey between 2017 and 2018. The use of data from 2017-2018 in this study is justified by the stable nature of anemia-related parameters, the reliability of the collection of NHANES data, 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 of the dataset that contain the variables or indicators related to anemia. 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. The complete blood count (CBC) data for 2017-2018 comprise the data set.

2.2 Research Variables

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

2.3 Research Flow

Figure 1 shows the CRISP-DM framework, which comprises 6 phases and 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. CRISP-DM standardizes data mining techniques as a generic approach to problem solving for a business or research unit [16].

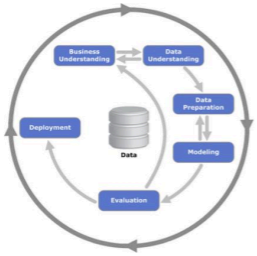

Figure 1. CRISP-DM framework
Source: Wikipedia

The main goal of the formation of models in the initial part of this study is to categorize different types of anemia by contrasting three supervised learning algorithms and looking for the best results. Using demographic information, patients with anemia 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 were 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 the naming and research requirements. In the following stage, which involves data preparation with the SEQN ID property, the 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 data cleansing. The anemia prediction attribute will be used as a label type in the parameter settings section. The classification of each form of anemia is the parameter to use [11]. In the split data, the data will be split into training data (70%) and testing data (30%). Data can be processed to the data modeling stage if the data preparation step is determined to be complete.

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

The performance of each created model is measured in the fifth stage of the evaluation using accuracy, mean recall, and mean precision. Prediction data for patients with anemia will be examined with demographic data in the next stage, namely deployment, if the best results of the performance of the model are available.

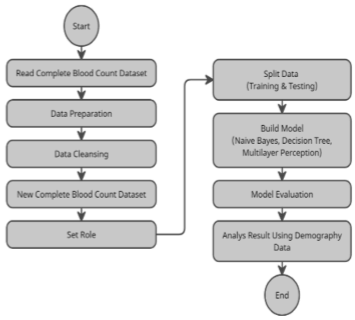

Figure 2. Flowchart Modeling Phase

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

3. Results and Discussions

3.1 Business Understanding Phase

This study uses 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 based on the availability of the most recent NHANES data. 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 the demographic data. There are 8160 columns and 3 attributes in the examination data. The laboratory data consists of 7844 and 22 characteristics. The SEQN attribute, or the response sequence number, is the ID of 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

The preparation of the data according to the requirements is the first step in creating a model. The recommended attributes to be used in modeling are chosen using the select attributes operator. For the

classification of different types of anemia and other characteristics that support the classification, 15 attributes were selected.

The next phase of data preparation is to use the filter example operator to remove missing values. With one ID attribute and 14 more numeric attributes, the total row in the cleaning stage becomes 6019. Anemia_prediction is a new attribute that was created using the generate attribute operator. According to previous investigations, expressions for anemia categorization parameters hemoglobin (LBXHGB), hematocrit (LBXHCT), mean cell volume (LBXMCVSI), and mean cell hemoglobin (LBXMC) were created, and red cell distribution (LBXRDW).

The aim of forecasting a property's classification as a label must be categories. The category type of an attribute can be changed using the set-role operator in RapidMiner. The label category type is the target role in the anemia_prediction attribute. The final step in the preparation phase is to split the data into training and testing data or to partition the data into two halves. For training data, the ratio is 70% (4213 columns), while for testing data it is 30% (1806 columns).

3.4 Data Modeling Phase

In the comprehensive analysis of the predictive modeling techniques used in this study, a multistage approach was adopted to assess the effectiveness of the Naive Bayes, J48 Decision Tree, and Random Forest classifiers. Each of these models played a distinct role in the modeling process, thus contributing to a more nuanced understanding of their performance characteristics.

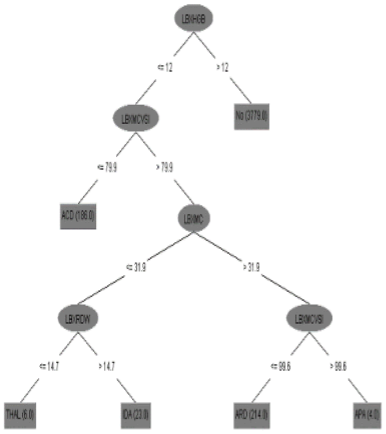

Figure 3. Tree Result of the J48 Decision Tree

In the initial modeling stage, Naive Bayes was employed in conjunction with a cross-validation operator, and its performance was assessed. This stage of the analysis revealed that the Naive Bayes model achieved a commendable accuracy rate of 92.69%, highlighting its competence in making predictions based on the given dataset. Subsequently, the second modeling stage saw the application of the J48 decision tree algorithm, again with the utilization of a cross-validation operator.

The results of this stage demonstrated a significantly high accuracy rate of 99.89%, underscoring the robust predictive capabilities of the J48 decision tree model. It should be noted that the J48 decision tree model was instrumental in the construction of the decision tree illustrated in Figure 3, characterized by a total of 6 leaves and a height of 11. This intricate tree structure reflects the complexity and depth of the decision-making process involved in the model.

Figure 4. Tree Result in Random Forest

In the third modeling stage, Random Forest was harnessed with the support of a cross-validation operator and its performance was meticulously evaluated. Impressively, the Random Forest model showed a remarkable accuracy rate of 99.61%, further attesting to its aptitude for predictive modeling. Figure 4 was used to visualize the results of the trees generated by the Random Forest method, providing insight into the multifaceted ensemble of decision trees that collectively contribute to its high predictive accuracy.

In summary, this multistage modeling approach used Naive Bayes, J48 Decision Tree, and Random Forest classifiers, each showing exceptional predictive accuracy rates. These results demonstrate the proficiency of these models and their different roles in the generation of accurate predictions. The subsequent sections of this article will provide a detailed discussion of the implications and insights derived from these modeling outcomes, ultimately contributing to a greater 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, the ROC (Receiver Operating Characteristic) data also corroborate the conclusion of the confusion matrix. 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 findings from the confusion matrix, one for each algorithm [16].

Table 2. Confusion Matrix Naïve 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 the confusion matrix of Table 2 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
Naïve 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 highest 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 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 (%) NB	J48	RF	Data set	Attribute	Tools
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
Manish [23]	96	95.5	95	200	18/5	WEKA
Nurul [24]	70.03	94.27	-	898	13/-	WEKA

This discrepancy may be caused by variations in the amount of data sets and technologies used, as seen by the comparison of research findings with earlier studies in Table 6. Unlike [12] that used only one variable of anemia parameters, namely hemoglobin, this study uses five variables of anemia parameters: hemoglobin, hematocrit, mean cell volume, mean cell hemoglobin, and red cell distribution. However, the study [25] did not include factors of the anemia parameter since it was more concerned with hematologic observations in the form of different blood diseases than with the classification of anemia.

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

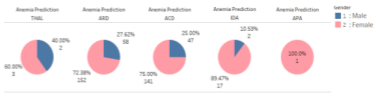

Figure 5. Visualization of Anemia Patients by Gender

According to Figure 5, women experience predominately all types of anemia. The prevalence of THAL-type anemia is 60% in women and 40% in men. 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.

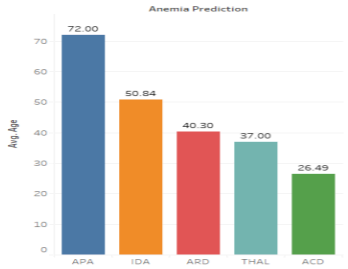

Figure 6. Visualization of Anemia Patients Based on Average Age

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

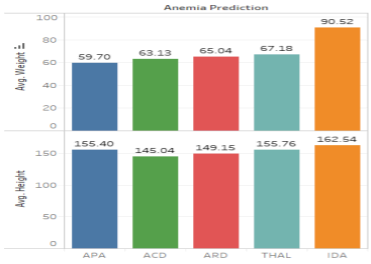

Figure 7. Visualization of Patients with Anemia According to Height and Weight

The average height and weight of each categorization are different, as shown in Figure 7. The range of average weight is 59.70-90.52, while the range of average height is 155.40-162.54. The type of APA anemia had the lowest average weight, whereas the type of ACD had the lowest average height.

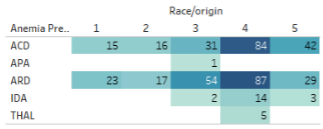

Figure 8. Visualization of Anemia Patients Based on Race/origin

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 of all racial / ethnic groups, with 190 anemia patients.

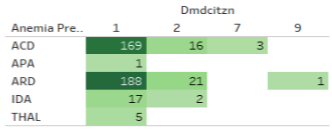

Figure 9. Visualization of Anemia Patients Based on Citizen

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 patients with anemia, category 1 (citizen by birth or naturalization) has the highest number of citizens.

4. Conclusions

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

Models of Naive Bayes, J48 Decision Tree, and Random Forest data mining algorithms are used to perform 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, modeling, evaluation, and deployment. A 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 highest 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 precision (99.89%), recall (71.67%), and precision (71.33%).

The data mining process used reveals new information on the classification of the many types of anemia, specifically that anemia patients are overwhelmingly women. Based on averages for age, weight, and height, it generates forecasts for patients with anemia that differ depending on the type of anemia. Patients of 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 to reduce anemia sufferers through health programs aimed at the non-Hispanic black female sex population and citizens by birth or naturalization.

Limitation:

Although this study provides valuable information on the prediction and classification of types of anemia

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

Future research:

Based on 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 improve 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 group methods, can lead to more robust prediction models for the classification of anemia. 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 models developed 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 better anemia diagnosis and personalized treatment strategies, ultimately benefiting patient care and results.

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

Acknowledgments

The achievement of this research has been achievable due to the steady support and invaluable assistance provided 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. Santhuram, 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/jfmpc.jfmpc_33_17.

[4] R. 2018, "Kemenkes RI. Hasil Riset Kesehatan Dasar Tahun 2018. Kementrian Kesehat RI.," J. Kesehat. Saintika 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. Hossin, 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/matecconf/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

Data Mining Techniques for Predictive Classification of Anemia Disease Subtypes

ORIGINALITY REPORT

8%	6%	3%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

< 1%

★ "Proceedings of International Conference on Recent Innovations in Computing", Springer Science and Business Media LLC, 2024

Publication

Exclude quotes	On	Exclude matches	< 8 words
Exclude bibliography	On		