



A Pilot Evaluation of Adopting the Apriori Algorithm Using WEKA Tool for Predicting Anemia and Blood Disorders Based on Laboratory Indicators from South Libya

Yahya BenYahmed *

Department of Computer Science, Faculty of Education, Traghan, University of Fezzan,
Murzuq, Libya

تقييم تجريبي لتطبيق خوارزمية Apriori باستخدام أداة WEKA للتنبؤ بفقر الدم واضطرابات الدم
بناءً على المؤشرات المخبرية في جنوب ليبيا

يحيى بن يحمّد *

قسم علوم الحاسوب، كلية التربية، جامعة فزان، مرزق، ليبيا

*Corresponding author: yah.benyehmed@fezzanu.edu.ly

Received: July 01, 2026

Accepted: September 19, 2026

Published: September 27, 2026

Abstract:

One of the most popular data mining algorithms for association rule mining is the Apriori algorithm. However, compared to Anemia data analysis are still restricted. Using the open-source WEKA tool, this study attempts to assess how well the Apriori algorithm extracts predictive association rules from Anemia data indicators from South Libya. Analysis was done on a dataset comprising 397 records and 14 features, such as CBC indices and demographics. 183 (NN), 187 (MH), and 27 (MA) comprised the target groups. Because the digital biomarkers in the sample were continuous, a range expert had to preprocess the data using a discretionary categorical filter in order for the algorithm to function properly. Six experiments were run with various parameters to discover the best algorithm settings. The ideal algorithm settings were found to be 50 rules, 0.8 minimum confidence interval, 0.1 minimum support, and the use of a lift metric to rank the rules. Fifty association rules were identified. MA class had the greatest lift value of 11.76, while MH class had values ranging from 1.75 to 2.10. The retrieved criteria were highly consistent with current medical knowledge, and potentially useful novel patterns, which may aid in the early detection of macrocytic anemia. The study finds that WEKA's Apriori is an effective and interpretable tool for Anemia data, making it a dependable software core for assisting early clinical decision-making in healthcare facilities. The study advises to handle unusual classes, comparing different association methods, and incorporating association rules and systems in future clinical applications.

Keywords: Data mining, Apriori algorithm, WEKA, association rules, Anemia, Decision support.

الملخص

تعد خوارزمية Apriori من أشهر خوارزميات استخراج قواعد الارتباط في مجال استخراج البيانات. مع ذلك، لا يزال استخدامها في تحليل بيانات فقر الدم محدودًا. تسعى هذه الدراسة، باستخدام أداة WEKA مفتوحة المصدر، إلى تقييم مدى كفاءة خوارزمية Apriori في استخراج قواعد الارتباط التنبؤية من مؤشرات بيانات فقر الدم في جنوب ليبيا. أُجري التحليل على مجموعة بيانات تضم 397 سجلًا و14 خاصية، مثل مؤشرات تعداد الدم الكامل (CBC) والبيانات الديموغرافية. شملت المجموعات المستهدفة 183 حالة طبيعية، و187 حالة فقر دم صغير الكريات ناقص الصباغ، و27 حالة فقر دم كبير الكريات. نظرًا لأن المؤشرات الحيوية الرقمية في العينة كانت متصلة، فقد استلزم الأمر معالجة البيانات مسبقًا بواسطة خبير في النطاقات باستخدام مرشح تصنيفي اختياري لضمان عمل الخوارزمية بشكل صحيح. أُجريت ست تجارب بمعايير مختلفة لاكتشاف أفضل إعدادات للخوارزمية. وُجد أن الإعدادات المثلى للخوارزمية هي 50 قاعدة، وفترة ثقة دنيا 0.8، ودعم أدنى 0.1، واستخدام مقياس الرفع لترتيب القواعد. تم تحديد خمسين قاعدة ارتباط. سجلت فئة MA أعلى قيمة رفع بلغت 11.76، بينما تراوحت قيم فئة MH بين 1.75 و2.10. كانت المعايير المسترجعة متسقة للغاية مع المعرفة الطبية الحالية، وكُشفت أنماط جديدة مفيدة محتملة، والتي قد تساعد في الكشف المبكر عن فقر الدم كبير الكريات.

تُشير الدراسة إلى أن خوارزمية Apriori في برنامج WEKA أداة فعالة وقابلة للتفسير لتحليل بيانات فقر الدم، مما يجعلها نواة برمجية موثوقة للمساعدة في اتخاذ القرارات السريرية المبكرة في مرافق الرعاية الصحية. توصي الدراسة للتعامل مع الفئات غير المعتادة، ومقارنة طرق الارتباط المختلفة، ودمج قواعد الارتباط مع أنظمة التصنيف في التطبيقات السريرية المستقبلية.

الكلمات المفتاحية: استخراج البيانات، خوارزمية Apriori، WEKA، قواعد الارتباط، فقر الدم، دعم القرار.

1. Introduction

The global healthcare sector is experiencing an unprecedented surge in the volume and diversity of medical data generated from electronic health records, laboratory tests, and patient monitoring systems. This massive influx of knowledge has created an urgent need to move beyond traditional statistical methods and adopt data engineering and artificial intelligence techniques to extract underlying predictive patterns that contribute to improving the quality of medical care and supporting timely clinical decision-making [1]. Knowledge Discovery in Databases (KDD) is a fundamental pillar enabling healthcare institutions to move from simply describing vital signs to strategically predicting diseases and understanding their complex interactions[2].

In this context, Anemia is one of the most widespread and intractable public health challenges globally, affecting millions of individuals and intersecting with a broad range of physiological, genetic, environmental, and malnutrition-related causes [3]. Medical data related to Anemia are highly complex, as readings of biomarkers such as hemoglobin levels, serum ferritin, and Packed Cell Volume (PCV) are intertwined with the patient's demographic and clinical factors. Despite the abundance of this data, the current research gap lies in the underutilization of associational and conditional relationships between these variables. Biomarkers are often studied in isolation, which reduces the accuracy of identifying precise disease classifications or complex secondary causes in an automated and predictive manner[4, 5].

Machine learning algorithms have demonstrated remarkable superiority in processing large-scale nonlinear data and capturing complex interactions between variables without requiring prior assumptions. This superiority has been particularly evident in the diagnosis of Anemia, where supervised learning models have achieved exceptional diagnostic accuracy. In a recent study involving 8,106 individuals to differentiate between beta-thalassemia (β -TT), Iron Deficiency Anemia (IDA), and healthy individuals, the Random Forest algorithm achieved an accuracy of 97.86% and an Area Under the Curve (AUC) of 99.71%. Another study using hematological data for six red blood cell indices also showed that the XGBoost model achieved superior diagnostic accuracy, identifying MCH as the most important feature for XGBoost, MCV for Random Forest, and HGB for neural networks [6].

Models based on image analysis have also shown promising results. A recent study used computer vision and machine learning techniques to analyze palmar images of children under six years old. The Logistic Regression model and Convolutional Neural Network (CNN) achieved a diagnostic accuracy of approximately 95.53%, while the Random Forest model in the CIELab color space achieved a coefficient of determination R^2 of 0.95. Analysis supported by SHAP values showed that red color features were the most significant predictors [7].

Machine learning applications extend beyond direct diagnosis to include the estimation of reference intervals for red blood cell indices. A study of South and Southeast Asian populations (64100 CBC results) used the XGBoost model to distinguish normal individuals from those with iron deficiency, thalassemia, or other types of anemia, achieving an AUC of 0.97. This approach has yielded more accurate baseline estimations compared to traditional methods, with the ability to generate Age and Gender specific ranges [8].

To bridge this research gap, association rule mining techniques have emerged as a powerful tool for inferring nonlinear and hidden relationships within medical populations. The Apriori algorithm is a leading and standard methodology in this field, due to its superior ability to generate predictive rules formulated in a logically accessible and easily interpretable way for medical professionals (human-in-the-loop interpretability), such as the formula: IF (specific symptoms and vital signs) ==> THEN (type or probability of Anemia). This characteristic gives the Apriori algorithm a significant competitive advantage over complex deep learning models that operate as medically unjustified "black boxes" [9].

To implement this methodology with high laboratory efficiency, this study relied on the WEKA (Waikato Environment for Knowledge Analysis) environment. This tool is academically and programmatically considered one of the best open-source environments, providing a comprehensive package for data preprocessing and implementing the discretization tools required by the Apriori algorithm for handling continuous numerical variables. Furthermore, it offers rigorous evaluation metrics that ensure the quality of extracted rules [10].

Based on the above, this study aims to provide a critical experimental evaluation of the performance of the Apriori algorithm using the WEKA tool in analyzing laboratory records of Anemia patients. This study primarily aims to examine the efficiency of the algorithm in generating statistically and clinically significant predictive association rules by balancing the support, confidence, and lift indices. It highlights and provides an applied perspective on exploring the predictive power of association rules, which can help improve data mining

algorithms in future smart health systems. These algorithms can suggest groups of risk factors that precede or exacerbate disease, potentially contributing to the development of early warning systems.

2 Background & Related Work

The last decade has witnessed a significant increase in the application of data mining and machine learning techniques in the health field. The goal is to extract hidden patterns and associational relationships from the massive amounts of data produced by health systems [11]. Association rule mining is one of the most prominent of these techniques, enabling the discovery of conditional relationships between different medical variables in the form of "if-then" rules. The Apriori algorithm, developed by [12], stands out as the most widely used algorithm in this field due to its simplicity and broad applicability. Its applications have extended to multiple medical fields, including the diagnosis of heart disease, tuberculosis, cancer, and chronic kidney disease, as well as the analysis of Anemia data, which is the focus of this study [13]

2.1 Theoretical Foundations and Mechanism of Apriori Algorithm

The core of the Apriori algorithm lies in discovering frequent item sets, that is, combinations of properties whose occurrence in a database exceeds a predetermined threshold. The algorithm operates based on two main criteria: The Minimum Support, which defines the minimum percentage of item occurrence in the data, and the Minimum Confidence, which expresses the conditional probability of the next item occurring, assuming the first item is true. The algorithm derives its name from its "Apriority" property, which states that any subset of a frequent item set must also be frequent. This allows for efficient truncating of the search tree and reduces the number of candidates in each iteration [12]. In the WEKA environment, the algorithm operates iteratively, gradually lowering the support threshold until the required number of rules that meet the specified confidence threshold is found [14].

2.2 Applications of Apriori Algorithm in Anemia Data Analysis

Recent years have seen increasing interest in employing the Apriori algorithm in Anemia data analysis to identify factors associated with this condition and extract the most frequent associational patterns. In a pioneering 2024 study conducted a comprehensive analysis using eight advanced machine learning techniques on data from 5,642 young women from the 2016 Ethiopian Demographic and Health Survey. In addition to classifying models, the researchers performed association mining using the Apriori, identifying the top seven associations most frequently associated with anemia. The results showed that the Random Forest classifier outperformed all other classifiers, achieving an Area Under the Curve (AUC) of 82%. Geographic region, poverty index, lack of formal education, poor sanitation, rural location, non-use of contraception, religion, age, lack of media exposure, occupation, and household size exceeding five members were identified as the most significant predictors of Anemia. The association mining also yielded the top seven associations most frequently associated with Anemia. This study highlights the added value of integrating association rule mining with machine learning models in understanding the multidimensional factors associated with Anemia in high-risk populations [15].

Another study [16] examined the prevalence of Anemia and its contributing factors among women of reproductive age (15–49 years) in Lesotho, using data from the Lesotho Demographic and Health Survey. The study included a comprehensive spatial analysis using Global Moran's I to detect spatial dependencies and a hotspot analysis. Additionally, association rules mining with the Apriori algorithm was used to explore potential interactions between variables affecting anemia. The study recorded an Anemia prevalence of 23.7%. Hotspots were identified in northwestern Lesotho (Leribi District) and southwestern regions (Mohalis-Hoke and Kuthing Districts). The Gradient Boosting Machine model demonstrated excellent predictive performance (AUC = 87.1%). Low educational attainment, having children under five, poverty, poor sanitation, occupation, urbanization, and the 35–49 age group were identified as the most influential factors. This study represents a significant advancement by integrating spatial analysis with associational rules mining, revealing the geographic dimension of Anemia prevalence and enhancing the accuracy of targeted health interventions.

In another study, Bhargava and colleagues proposed a novel methodology for the early detection of Anemia in 2024, combining the Predictive Apriori algorithm with Boolean Matrix analysis [9]. This study applied hemoglobin test data from a public dataset, which had been preprocessed to remove missing and unwanted values. The Boolean Matrix was used to eliminate non-recurring candidates, while recurring patterns were generated using the logical AND operation. The results showed that combining the Predictive Apriori algorithm with the Boolean Matrix yielded improved outcomes. The study included a definition of the hematological variables used, where I represented RBC, MCV, MCHC, TLC, PLT, and HGB. The results showed that combining the predictive algorithm with the logical matrix resulted in improved Anemia detection. This methodology is distinguished by its ability to reduce computational complexity by eliminating non-recurring candidates at an early stage, thus improving the efficiency of extracting associational rules from medical data.

In another innovative application used the Apriori algorithm to correlate Anemia symptoms with the emotions expressed by patients on Twitter. The k-means algorithm and Latent Dirichlet Allocation (LDA) were used to group similar tweets and identify underlying themes, and then the emotions and symptoms were correlated using the Apriori algorithm. The methodology was evaluated using several classifiers, achieving the highest predictive accuracy of 98.96% using the Sequential Minimal Optimization (SMO) algorithm. The results revealed that feelings of fear and sadness were prevalent among patients with Anemia [17].

In a pioneering 2022 study, the Apriori algorithm was used to analyze drug combinations in Traditional Chinese Medicine (TCM) for the treatment of aplastic anemia. The study aimed to explore the most common herb pairs in prescriptions used to treat this rare type of anemia. The Apriori algorithm was applied to analyze the association rules between highly frequent herbs, revealing three herb pairs with support exceeding 0.3 and twelve herb pairs with confidence exceeding 0.85. This study is a qualitative application of the Apriori algorithm, as it goes beyond traditional demographic and clinical data analysis to include the analysis of treatment patterns in alternative medicine systems, opening up new horizons for employing association rule mining in supporting treatment decisions and guiding pharmaceutical research [18].

In the context of using the WEKA software, many recent studies have focused on employing data mining algorithms for the diagnosis and classification of hematological diseases. These efforts range from using classification models to generating association rules. In a significant landmark study led by [19], focused on extrapolating the causes of Anemia and its relationship to the lifestyle and medical conditions of patients by building a predictive model based on class association rules. The study used the WEKA tool to process a composite database integrating complete blood count (CBC) results and dietary questionnaires. The researchers encountered a computational challenge in the form of the common "redundant rules" problem when running Apriori on multidimensional medical samples, and this was overcome through precise calibration of the minimum support. The algorithm successfully formulated explainable AI rules that clearly linked a sharp drop in serum ferritin to pregnancy and vegetarian dietary patterns, demonstrating Apriori's effectiveness in providing visual and readable diagnostic maps for healthcare professionals.

In another study, the Apriori algorithm was used to mine association rules using R software on data from 5,642 young women in Ethiopia. The seven most frequently occurring association rules with Anemia were identified [15]. A recent comparative study in 2025 evaluated four recurring pattern mining algorithms FP-Growth, Eclat, AprioriTid, and the Partition algorithm on the MIMIC III medical dataset, revealing the strengths and weaknesses of each approach and providing valuable insights into their suitability for different types of health data.

According to a study by [20], the Threshold-Aware Association Rules (TAAR) framework provides a way to extract interpretable association rules from continuous health data using the Apriori algorithm and large language models. The TAAR framework transforms continuous laboratory values into semantic intervals that reflect established clinical norms and then extracts interpretable association rules using an improved Apriori algorithm. A key innovation of this framework is its use of Large Language Models (LLMs) to refine support and confidence thresholds, filter out unreasonable rules, and generate natural language interpretations that are easily understood by clinicians. The framework has been applied to blood test data and has demonstrated significant improvements in clinical usability, providing actionable follow-up recommendations and supporting informed decision-making. This framework represents a promising direction in aligning data mining techniques with the needs of healthcare teams by bridging the gap between statistical analysis and clinical application.

In study by [21], the performance of various data mining techniques in predicting Anemia and iron deficiency using laboratory parameters was evaluated. The study was conducted experimentally within the WEKA interface, with the work divided into two tracks: the first track used J48 and Naïve Bayes algorithms, while the second track employed the Apriori algorithm to extract associational rules after converting the numerical indicators MCV, MCH and HGB into categorical values Low and Normal. The comparative evaluation showed that while the classification algorithms excelled in purely statistical sorting accuracy, the Apriori distinguished itself by providing a "Diagnostic Mosaic" that reveals the association between variables, enabling the prediction of Anemia through the association of sub-biological marker deficiencies, not just hemoglobin.

2.3 WEKA Platform for Medical Association Mining

WEKA (Waikato Environment for Knowledge Analysis) is a leading open-source data mining platform. It provides a comprehensive and user-friendly environment for various machine learning and association mining algorithms, including Apriori and FP-Growth. WEKA is distinguished by its ability to handle multiple file formats and offers advanced data filtering tools (such as picking and categorizing numeric variables) and visualization of extracted rules [22]. WEKA has proven its effectiveness in numerous medical studies, where it has been used to compare disease classification models, predict lung cancer, analyze stroke risk, and extract associational rules from real clinical databases [23-25]. These features make WEKA a practical and suitable option for medical researchers to explore hidden patterns in their data without requiring advanced programming skills [26].

The WEKA tool has garnered significant attention from researchers in the field of medical data interpretation and the diagnosis of Anemia and blood disorders. These efforts range from associational rules engineering to comparative classification models. In a reference study by [27] on classifying Anemia and blood disorders using machine learning techniques, several software algorithms were compared. The Gradient Boosting algorithm achieved the highest predictive accuracy at 99.19%, followed by decision trees. The study confirmed that data mining-based solutions offer low-cost, non-invasive options for enhancing early diagnosis in resource-limited medical settings.

Meanwhile, researchers have focused on comparing methodologies within the WEKA environment to identify the specific types of blood disorders. [28] reviewed the use of algorithms such as J48 and Naïve Bayes for diagnosing and separating thalassemia from iron deficiency Anemia (IDA). Their results showed that the Naïve Bayes network-based algorithm achieved up to 98% accuracy in extrapolating the intrinsic relationship between low MCV and MCH indices and identifying the type of anemia. Similarly, previous studies, such as [29], indicate that the J48 decision tree provided the highest classification efficiency for classifying anemias when the data were segmented in different proportions within WEKA.

In the context of Anemia data analysis, WEKA's tool was used to explore the predictive behavior of the Apriori algorithm and generate association rules for anemia. In this experiment, a minimum support threshold of 0.2 and a confidence level of 0.9 were set for a sample of 500 cases. Association rules were generated based on baseline clinical symptoms and disease history [9]. It is worth noting that Apriori algorithm operates iteratively, starting with an upper support threshold and then gradually decreasing the support until the desired number of rules is found. However, it was observed that the algorithm may encounter memory issues when run on very large datasets. In a systematic review conducted according to the PRISMA protocol, association rules analysis using the Apriori algorithm via WEKA were employed, utilizing metrics such as support, confidence, and lift to assess the associations identified by the algorithm [30].

Regarding conditional rule extraction, association algorithms have proven their effectiveness in clinical prediction. A study by [31] verified that association rule generation techniques outperformed deep models in describing and identifying the interrelationships between laboratory test indicators and the final diagnosis, compared to deep models that excelled only in purely classification tasks. Similarly, [32] evaluated the performance of the Apriori by applying it to complex health scenarios, such as Malaria data and Risk indicators, where more than 120 laboratory experiments were conducted at varying levels of support and confidence. The support and confidence criteria were systematically varied. The study concluded that setting the lift metric above 1.0 within WEKA is the crucial metric for filtering out redundant rules and extracting genuine causal relationships between demographic variables such as age and gender, and disease patterns with high interpretive accuracy.

In the study by [33], artificial neural networks were used with WEKA to detect Anemia in women. The study included 567 patients, and a neural network was designed with seven inputs (RBC, HCT, MCV, MCH, MCHC, RDW, WBC) and one output (diagnostic decision). The results showed that the artificial neural network diagnosed the disease with high accuracy, confirming WEKA's effectiveness as a platform for applying machine learning techniques in medical diagnosis.

Other studies have also explored the use of WEKA in analyzing medical association rules, where the Apriori algorithm was applied to diverse datasets to extract recurring patterns and relationships between clinical variables. WEKA is distinguished by its ability to handle multiple file formats and provide advanced data filtering tools (such as picking and converting numeric variables to categories) and visualization of extracted rules, making it a practical and suitable option for medical researchers to explore hidden patterns in their data without requiring advanced programming skills [23, 34, 35].

Based on a comprehensive review of previous literature, a clear research gap can be identified. For example, although the Apriori algorithm has been used in Anemia prediction studies, studies that systematically evaluate the algorithm's performance in this context in terms of rule quality under different parameter settings, time efficiency, and the impact of data size remain limited. Furthermore, despite the widespread use of the WEKA platform among medical researchers, studies that experimentally and systematically evaluate the Apriori algorithm's performance within it on Anemia data are insufficient. Additionally, there is no comprehensive study linking the quality of extracted rules (in terms of support, confidence, and elevation) with the algorithm's computational efficiency in the WEKA environment, considering the categorical nature of medical data after discretization. Therefore, this study aims to address these gaps by evaluating the performance of the Apriori algorithm in analyzing Anemia data and generating predictive association rules using WEKA, analyzing the impact of varying support and confidence parameters on rule quality and computation time, and providing practical recommendations for researchers and practitioners in the field of health informatics.

Based on the critical analysis presented in the discussion of previous studies, Table 1 summarizes the most relevant recent studies, identifying the algorithms used, the medical datasets, and the best results achieved by each study, in order to provide a comprehensive overview and clear comparison.

Table 1. Summary of the most relevant recent studies.

Study	Selection algorithms used	Medical Dataset	Best Outcome
Zemariam et al. (2024)	Apriori for association mining, along with eight machine learning classification algorithms (Random Forest, XGBoost, and others).	Data from 5,642 young women from the Ethiopian Demographic and Health Survey (EDHS 2016) at the national level.	The top seven association rules for Anemia were identified. The RF model achieved the highest predictive performance (AUC = 82%). Geographic location, poverty, and poor sanitation were identified as the most significant factors.
Kitaw & Haile (2025)	Apriori for exploring interactions, multiple machine learning models with SHAP for model interpretation, and spatiotemporal models.	National data from Lesotho for the period 2023–2024.	Anemia was found to be 23.7% with a significant spatial dependency (Moran's I = 0.248). The GBM model outperformed this model (AUC = 87.1%). Poverty, low education, and having children under five were identified as the strongest factors.
Bhargava et al. (2024)	Predictive Apriori combined with Boolean Matrix Analysis.	General hemoglobin test data.	Integrating the algorithm with a logical matrix improved its efficiency by effectively eliminating non-recurring candidates, resulting in the generation of enhanced association rules with better performance than the traditional algorithm.
Abboura et al. (2025)	TAAR (Threshold-Considering Rules) framework, which integrates an enhanced Apriori algorithm with Large Language Models (LLMs).	Laboratory blood test data.	LLMs were used to improve support and confidence thresholds, filter out unreasonable rules, and generate natural language interpretations, significantly enhancing the clinical utility of the rules.
Zemariam, Yimer et al. (2024)	Apriori, FP-Growth, ECLAT, AprioriTid, and Partition.	MIMIC III medical dataset and other standard health datasets (anemia).	Results showed that FP-Growth is up to ~45% faster than Apriori and has better scalability, while ECLAT excels in dense datasets. However, Apriori remains the most widely used due to its simplicity.

It is noteworthy that most studies rely on the Apriori algorithm as the primary tool for association rule mining, but they integrate it with other classification algorithms (such as Random Forest and KNN) or optimization techniques to enhance performance and interpretability. Regarding the diversity of data sources, ranging from comprehensive national surveys (Ethiopia, Lesotho) to limited laboratory data, this impacts the generalizability of the results. National studies have demonstrated the superiority of models such as GBM and Random Forest in prediction. Furthermore, recent study as [20] focus on overcoming a major limitation of Apriori its inability to handle continuous numerical data by employing generative AI techniques, representing a significant advancement in the field. Despite the abundance of studies, there remains a lack of systematic evaluation of Apriori's performance in the WEKA environment, specifically on Anemia data, regarding the relationship between parameter settings, rule quality, and time efficiency, a point underscored by the present study. Previous studies have consistently highlighted the promising potential of the Apriori algorithm in extracting anemia-related risk factors from health data, and have emphasized the role of the WEKA platform as a practical and user-friendly tool for medical researchers. However, the current literature suffers from several methodological limitations, most notably the lack of standardized data preprocessing procedures, reliance on poorly defined parameter settings, and the absence of a comprehensive evaluation of rule quality and algorithm efficiency. Furthermore, the recent trend of integrating rule mining with artificial intelligence and distributed computing techniques opens new horizons, but these have not yet been widely applied in the context of anemia. This study aims to address these gaps by providing a systematic empirical evaluation of the Apriori algorithm's performance on WEKA, offering analytical insights and practical recommendations that contribute to improving the quality of future research in this field and guiding researchers toward the optimal use of these techniques in supporting early diagnosis and treatment decisions.

2.4 Established Medical Knowledge of Anemia

The associational rules derived from the application of the Apriori algorithm gain their scientific value by comparing them with established medical knowledge in the field of Anemia diagnosis. Medical knowledge aims to validate the findings and determine their consistency with internationally accepted diagnostic criteria, highlighting those that reflect established clinical patterns or those that contribute new insights to medical knowledge.

According to studies [36-38], the diagnostic criteria issued by the World Health Organization (WHO) and accepted medical references in hematology are used. Anemia is classified into three types: macrocytic, normocytic, and microcytic, based on three main indicators derived from a complete blood count (CBC): mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), and mean corpuscular hemoglobin concentration (MCHC). Table 2 shows types of Anemia by hematological indicators.

Table 2. Classification of Anemia Types by Hematological Indicators

Type of Anemia	MCV (fL)	MCH (pg)	MCHC (g/dL)
Macrocytic	> 100 (High)	> 32 (High)	32-35 (Normal)
Normocytic	80-100 (Normal)	27-32 (Normal)	32-35 (Normal)
Microcytic	< 80 (low)	< 27 (low)	< 32 (Low)

At the World Health Organization (WHO) level, Anemia is defined as a hemoglobin (Hb) concentration below the normal ranges established for age, gender, and physiological stage. Hb is defined as less than 12.0 g/dL in women and less than 13.0 g/dL in men [38, 39]. Microcytic hypochromic Anemia is characterized by a MCV below 80 fL and a low MCH, along with Anemia (hemoglobin below normal levels). The Hematology Laboratory Guide confirms that this pattern is the primary hallmark of iron deficiency Anemia (IDA) and thalassemia [37, 39].

In another context, medical references [40, 41] indicated that macrocytic Anemia is characterized by an elevated MCV (>100 fL) and an elevated MCH (>32 pg). Medical sources also confirm that the combination of an elevated MCV and an elevated MCH is the primary hallmark of macrocytic anemia, which is often caused by a deficiency of vitamin B12 or folic acid. Elevated MCV and MCH together is the most specific hallmark of macrocytic anemia. Medical guidelines indicate that the presence of these two elevated indicators along with low hemoglobin necessitates immediate measurement of vitamin B12 and folic acid levels. Furthermore, in the early stages of macrocytic Anemia (especially that caused by B12 deficiency), hemoglobin may be within the normal range while the MCV is already elevated. This pattern presents an opportunity for early intervention before the development of overt anemia. Elevated MCV and MCH may precede decreased hemoglobin levels in the early stages of vitamin B12 deficiency.

A studies by [42, 43] indicated that a decrease in both HGB and MCV is the primary hallmark of iron deficiency anemia. The diagnosis of IDA is confirmed by low serum ferritin levels, high total iron-binding capacity (TIBC), and low transferrin saturation. Furthermore, medical studies [44, 45] cited World Health Organization (WHO) estimates for 2025 that 30% of women of reproductive age (15-49 years) suffer from Anemia globally, with significant geographical variations. Medical studies also demonstrate that women of reproductive age are the most vulnerable group to iron deficiency Anemia due to blood loss during menstruation, increased iron requirements during pregnancy, and postpartum hemorrhage. In addition, studies indicate that Anemia affects 40% of children (6-59 months), 37% of pregnant women, and 30% of women of reproductive age, with a greater concentration in low- and middle-income countries.

Although previous work has focused on the predictive accuracy of using association rule algorithms to predict anemia, there is a lack of experimental studies evaluating the clinical performance of the Apriori algorithm when applied to Anemia data, specifically within WEKA. The algorithm faces complex challenges in converting continuous blood count (CBC) indices into nominal categories, as well as the problem of generating redundant rules. This study aims to bridge this gap by providing an empirical evaluation that systematically measures the quality of extracted rules based/ on the criteria of support, confidence, and lift.

A review of previous studies that combined WEKA and Apriori with the challenges of Anemia reveals the importance of preprocessing. These studies agree that filtering and categorization in WEKA are the critical factors determining the quality and validity of extracted medical rules. Most studies focused on extracting general medical rules but did not extensively cover the preprocessing of increased categorization levels in Anemia data, particularly in Libya. This study aims to address this empirical gap. However, Table 3 presents a summary of previous studies using the Apriori algorithm with Anemia datasets.

Table 3. Summary of previous studies

Study	Method	Outcome
Tegenaw 2018 Zemariam, Yimer et al. 2024	Using Apriori with Discretization Filter to explore the overlapping relationships of Anemia indicators in children.	Generating strong association rules demonstrated a strong conditional relationship between low hemoglobin levels, anthropometric indices, and patterns of motor malnutrition.
Ghalibaf, Sameni et al. 2022	Using Apriori to predict Anemia patterns and their relationship to dietary and health habits.	Formulating clear and interpretable predictive rules to provide an easily readable diagnostic map for medical professionals.
MN Amin, A Habib, 2015	The performance of Apriori with J48 and NB classification algorithms in classifying hematological disorders.	Classification algorithms excelled in sorting accuracy based on the formulation of clear and interpretable predictive rules.
Terzi, Sarıbacak et al. 2024	ANN algorithm with Apriori in classifying IDA risk factor data.	Confirming Hb, MCV, Fe, and FERR as the most important indicators with 98.06% diagnostic accuracy.

Within the framework of established medical knowledge on anemia, relevant work that forms the medical and laboratory basis for Anemia diagnosis provides a foundation for the diagnostic criteria adopted by the World Health Organization (WHO), the Hematology Laboratory Manual, and related studies in the field of hematology. This work systematically compares the associational rules derived from the Apriori algorithm. The aim of this comparison is to validate the results obtained from the Apriori algorithm and determine their consistency with established medical knowledge, highlighting rules that reflect known clinical patterns or those that offer new insights into medical knowledge.

3. Methodology and Experimental Design

Research methodology is the cornerstone of any scientific study, defining the methodological procedures that ensure the reliability and reproducibility of the results. The study provides a detailed description of the methodology used to achieve the study's objectives, which were to evaluate the performance of the Apriori algorithm in analyzing and extracting associational rules from Anemia data using the WEKA tool. This study is based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which is the most widely used framework in data mining projects due to its flexibility and comprehensiveness. CRISP-DM comprises six main phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment [46]. The dataset used will be described in detail, along with the preprocessing procedures, including categorical data transformation (Discretization). The steps for implementing the Apriori algorithm and adjusting parameters will then be explained. The chapter concludes with a description of the adopted evaluation metrics and the mechanisms for validating the results.

The implementation of this methodology relies on WEKA version 3.8, which provides an integrated environment for data processing and the application of machine learning algorithms. Figure 1 illustrates the six stages of the CRISP-DM methodology as applied in this study.

3.1 Business Understanding

This phase involves defining the clinical and technical objectives of the study, which are to uncover hidden patterns and associations among Anemia variables using the Apriori algorithm, and to provide clinically significant association rules that can be used as a tool for early diagnosis or understanding risk factors.

In this phase, the problem is defined, and the study requirements are identified in terms of data, specifically the need for a clinical database of Anemia patients that includes complete blood counts (CBCs), demographic information, and associated risk factors.

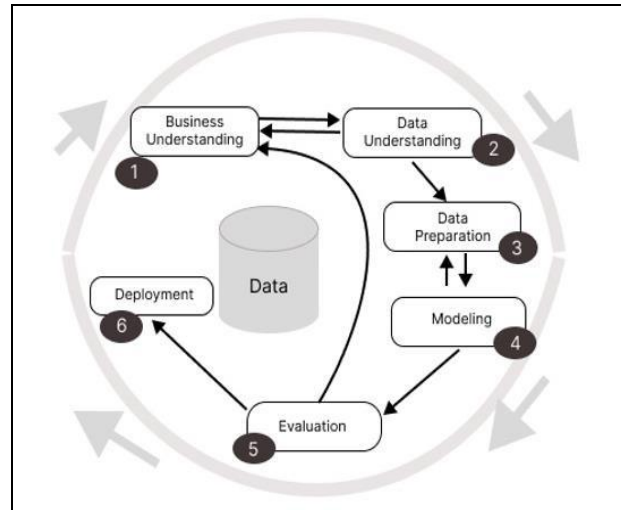


Figure 1: Stages of CRISP-DM [46]

3.2 Data Understanding

a. Data Source Description

This study presents a practical application of the Apriori algorithm within a WEKA environment on a clinical database of Anemia data collected retrospectively from patients attending the Medical Alabideen Lab in Wadi Etabah, South Libya. The dataset comprises 397 records and 14 features demographic and hematological in addition to the target variable, which is divided into three categories: NN (Normocytic Normochromic) with 183 cases, MH (Microcytic Hypochromic) with 187 cases, and MA (Macrocytic Normochromic) with 27 cases. Multiple algorithm settings were implemented, and the results were evaluated using the Support, Confidence, and Lift measures. The results are presented systematically, with analysis of the most significant rules and their clinical interpretation, as well as a comparison of the algorithm's performance across different parameter settings.

b. Variables Included in the Study

Table 4 includes a list of the variables used in the analysis, divided according to their categories.

Table 4. Dataset of Anemia Disease

No.	Variable	Description	Type of Values
1	Age	Age patient	Child, Teenager, Young, MidAge, Elderly
2	Gender	Gender patient	Male, Female
3	HGB	Hemoglobin	Numeric
4	RBC	Red Blood Cell count	Numeric
5	HCT	Hematocrit	Numeric
6	PLT	Numeric Platelet	Numeric
7	MCV	Mean Cell Volume	Numeric
8	MCH	Mean Cell Hemoglobin	Numeric
9	MCHC	Mean Corpuscular Hemoglobin Concentration	Numeric
10	TLC/WBC	White Blood Cell	Numeric
11	neut	Neutrophil	Numeric
12	lymph	Lymphocytes	Numeric
13	mono	Monocytes	Numeric
14	Class	Target audience (diagnosis)	1. Normocytic Normochromic (NN) 2. Microcytic Hypochromic (MH) 3. Macrocytic Normochromic (MN)

c. Exploratory Data Analysis

Before preprocessing, an exploratory data analysis was performed to understand the data distribution and identify potential problems:

- **Missing Values Analysis:** The percentage of missing values was calculated for each variable. Variables with a missing value exceeding 20% were excluded or processed using appropriate techniques (detailed in Section 3.3).

- **Outlier Detection:** Outliers were identified using the Boxplot method (where any value outside the range of $Q1 - 1.5 \times IQR$ or $Q3 + 1.5 \times IQR$ is considered an outlier).

- **Variable Distribution Analysis:** The distribution of each variable was examined to ensure it conformed to the statistical assumptions. Non-normal distributions were documented for consideration in subsequent processing.

3.3 Data Preparation

- Missing Value Handling

The obtained data was examined for the presence of information entry errors, as well as typographical errors and missing data. Some missing data are present in the data that have been gathered. In this study, the missing values are filled using the replace with missing values filter, making the information complete.

- Discretization of Data

The Apriori algorithm in WEKA requires all data to be of nominal or binary type, as the algorithm operates by comparing the equality of values rather than measuring the distances between them. Therefore, continuous numerical variables were categorized based on the World Health Organization (WHO) reference standards and the medical references approved by the Zain Al-Abidin Laboratory.

The study adopted a clinical knowledge-based discretization approach to preserve the clinical significance of the categories (e.g., "Low," "Normal," "High"). This also facilitates the interpretation of the derived rules by physicians and healthcare practitioners. Furthermore, it ensures that the cut-off points reflect recognized medical standards and are not merely a statistical distribution. Table 5 specifies the categories used for each variable and their clinical rationale.

Table 5. Discretization Scheme

No.	Variable	Category
1	Age	Age <18,"Child" , Age <76,"MidAge" Age <26,"Teenager" , Age >=76"Elderly" Age <66,"Young"
2	Gender	Male , Female
3	HGB	HGB <12,"Low" HGB >17,"High", Else," Normal"
4	RBC	RBC <4.5,"Low" RBC >5.5,"High", Else," Normal"
5	HCT	HCT <41,"Low" HCT >50,"High", Else," Normal"
6	PLT	PLT <150,"Low" PLT >450,"High", Else," Normal"
7	MCV	MCV <76,"Low" MCV >100,"High", Else," Normal"
8	MCH	MCH <26,"Low" MCH >34,"High", Else," Normal"
9	MCHC	MCHC <31,"Low" MCHC >37,"High", Else," Normal"
10	TLC/WBC	WBC <4.3,"Low" WBC >10.8,"High", Else," Normal"
11	neut	HCT <55,"Low" HCT >70,"High", Else," Normal"
12	lymph	neut <25,"Low" neut >40,"High", Else," Normal"
13	mono	mono <2,"Low" mono >8,"High", Else," Normal"

The study was conducted on patients from various age groups (Child, Teenager, Young, Middle Age, Elderly) and Genders (Male, Female). The data were pre-classified into nominal values such as Low, Normal, and High, making them ready for direct application of the Apriori algorithm without the need for further digitization (discretization).

The application was divided into several experiments in which the algorithm parameters were adjusted with the following objectives:

1. To generate the largest possible number of statistically and clinically significant rules.
2. To determine the optimal settings for extracting association rules for different disease categories (NN, MH, MA).
3. To analyze the rules associated with each of the target categories.

e. Data Format for WEKA

The data must be converted to ARFF (Attribute-Relation File Format), the native format read by WEKA

3.4 Modeling

The Apriori algorithm, developed by Agrawal and Srikanth in 1994, is the standard foundational methodology for association rule mining within data analysis environments such as WEKA [47]. Its computational structure relies on a bottom-up approach to discover Frequent Itemsets in medical databases, then formulates predictive conditional relationships governed by strict statistical parameters [48].

a. Apriori Algorithm Operation Mechanism

The Apriori algorithm's operation mechanism consists of two main sequential stages:

1. Generating Frequent Itemsets and Using the Apriori Property

The algorithm faces the challenge of the large potential space for the items. To reduce this computational burden, it relies on the Apriori Property, also known as the Downward-Closure Property. This principle states mathematically that: Every subset of a recurring element set must also be recurring. Conversely, if an element set is non-recurring, all its supersets will inevitably be non-recurring and will be pruned immediately without examination [48]. This stage is implemented through two iterations:

- **Join Step:** To create a candidate set called C_k (where k represents the number of elements in the set) by combining the previous recurring sets L_{k-1} with themselves.
- **Pruning Step:** Scanning the Anemia database to calculate the support frequency of candidate elements in C_k and excluding any set with less support than a predefined minimum (MinSupport), resulting in a list of approved recurring sets L_k [49]. This process is repeated until no new candidate sets remain.

2. Generating Predictive Association Rules

After settling on the recurring groups extracted from the Anemia data (such as the low hemoglobin-low ferritin syndrome), the algorithm moves to formulating predictive rules as a conditional relationship:

$$L \Rightarrow R$$

where

L: represents the clinical antecedent.

R: represents the target diagnostic outcome.

To validate these rules within the WEKA tool, they are evaluated and filtered based on statistical quality measures [46] in the fifth phase of evaluation.

b. Parameter Settings for the Apriori Algorithm

Table 6 specifies the parameters to be set and how they affect the results, based on technical sources.

Table 6. Apriori Algorithm Parameters in WEKA

Rationale	Tested Values	Description	Symbol	Parameter
Assessing the optimal number of rules	10, 20, 50, 100	Number of rules to be generated	-N	Number of rules
Balancing confidence and over-preparation	0.7, 0.8, 0.9	Minimum rule confidence	-C	Minimum confidence
Stability in research	0.05 (default)	Support drop value per iteration	-D	Delta
Starting point	1.0 (100%)	Maximum support	-U	Upper bound min. support
Controlling rule coverage	0.1, 0.15, 0.2	Minimum support	-M	Lower bound min. support
Selecting the optimal quality assessor	confidence, lift	Rule type for rule ranking	-T	Metric type
Aligning rules with the target audience	false, true	CAR classification rules extraction	-A	Class association rules
Defining the outcome variable	last (default)	CAR column index	-c	Class index

A systematic experimental strategy will be adopted:

1. Start with default values (N=10, C=0.9, M=0.1)
2. Evaluate the quality of the resulting rules
3. Adjust the parameters based on the results:
If the number of rules is too small (<5): Gradually decrease M and/or C
If the rules are trivial or intuitive: Increase C and/or use the LIFT scale
If confidence is low while maintaining a sufficient number of rules: Increase C

c. Command Line Documentation

To ensure reproducibility, all experiments will be documented using the command line format. For example:

```
WEKA.associations.Apriori -N 100 -T 1 -C 1.5 -D 0.05 -U 1.0 -M 0.1 -S -1.0
```

Where:

- N: 100: Request 100 rules.
- T: 1 Use the lift metric to sort the rules.
- C: 1.5 The minimum lift value is 1.5.
- D: 0.05 Gradual decrease in support by 5%.
- M: 0.1 The minimum support is 0.1 (10%).

3.5 Evaluation

a. Quantitative Evaluation Measures

The measures shown in Table 7 uses to evaluate the quality of the rules extracted by the Apriori algorithm.

Table 7. Quantitative Evaluation Measures for the Apriori Application

Scale	Mathematical formula	Significance	Target value
Support	$\text{Supp}(L \rightarrow R) = (\text{Number of transactions containing } L \cup R) / N$	The proportion of cases that include the entire rule	> 0.1 (10%)
Confidence	$\text{Conf}(L \rightarrow R) = \text{Supp}(L \cup R) / \text{Supp}(L)$	The probability of R given L	> 0.7
Lift	$\text{Lift}(L \rightarrow R) = \text{Conf}(L \rightarrow R) / \text{Supp}(R)$	The strength of the association compared to randomness	> 1.5

- **Support:** The percentage of transactions (patient cases) containing the set of items. A minimum level of support is established, below which a set of items is either accepted or rejected.

- **Confidence:** Measures the conditional certainty of the rule, calculated by dividing the total set support by the support of the premises. This value must exceed the clinically agreed minimum confidence level (e.g., $\geq 80\%$).

- **Lift:** The critical metric for filtering rules and establishing predictive power, measuring the degree of independence between L and R occurrence. Furthermore, if the Lift value is > 1 , the rule has genuine predictive power, meaning that the occurrence of L symptoms effectively and directly increases the probability of a R Anemia diagnosis, which is what the study seeks to rule out random or intuitive relationships. Additionally, rules with a Lift < 1.0 will be considered misleading rules (indicating an inverse relationship). Rules with a leverage of 1.0 are considered completely independent (useless).

b. Qualitative Evaluation and Clinical Validation

In addition to quantitative measures, a qualitative evaluation of the derived rules will be conducted:

1. Classification of rules according to clinical significance (medically acceptable, novel, unexpected)
2. Comparison of the derived rules with the medical literature to verify their consistency with existing knowledge
3. Identification of new rules that can be tested in future studies.

3.6 Deployment

This phase includes documenting and preparing the results, presenting comprehensive analytical reports, highlighting the most significant rules with their clinical interpretations, and providing recommendations for future research and potential clinical applications.

The study presented a detailed methodology for applying and evaluating the Apriori algorithm in WEKA to Anemia data. The internationally recognized CRISP-DM methodology was followed, ensuring a transparent and reproducible systematic procedure. The study dataset, preprocessing procedures including data categorization,

and the steps for implementing Apriori association algorithms with their parameters were defined. Quantitative and qualitative evaluation measures were also determined. This methodology provides a practical framework for conducting the study and achieving its study objectives.

4. Experimental Setup

The Apriori algorithm was implemented in WEKA using the Explorer interface and the "Associate" tab. Table 8 shows six different experiments were conducted to explore the effect of changing parameters on the quality and quantity of extracted rules. This strategy is based on the recommendations in WEKA's technical resources.

Table 8. The six experimental setups for applying Apriori to Anemia data.

Eexperiment	N (Number of rules)	C (Min. Confidence)	M (Min. Support)	T (Metric)	car (CARs)	Aim of the experiment
Exp 1	20	0.9	0.1	Confidence	False	Baseline: Strict default settings
Exp 2	50	0.8	0.1	Confidence	False	Increase the number of rules by slightly decreasing confidence
Exp 3	100	0.7	0.1	Confidence	False	Maximize the number of rules (at the expense of quality)
Exp 4	50	0.8	0.15	Confidence	False	Increase support for filtering rare items
Exp 5	50	0.8	0.1	Lift	False	Use lift as a ranking criterion
Exp 6	50	0.8	0.1	Confidence	True	Extract class association rules to direct results towards a class

4.1 Experimental Results and Analysis

From the Exp 1, only 20 rules were generated because the high confidence interval is 0.9 and low support is 0.1 filtered out most non-recurring elements. Furthermore, all 20 rules contained the target class, Class=NN, on the score side.

Example of generated rules:

Class=NN 183 ==> Gender=Male Age=Young 70 conf:(0.96)

Class=NN 183 ==> Lymphocytes=Low 90 conf:(0.95)

Class=NN 183 ==> Monocytes=Low 107 conf:(0.94)

With these strict settings, the rules were very simple and related only to the normal class (NN) with obvious parameters such as "Young" and "Lymphocytes=Low". No rules were generated for the MH or MA classes due to their relative rarity or the complexity of their patterns.

In the Exp 2, 50 rules were generated, of which 84% (42 rules) still focused on Class=NN. The first rules for the MH class (8 rules) also appeared at this confidence level. Table 9 appearances the most important rules derived for the MH (Microcytic Hypochromic Anemia) class:

Table 9. Derived Rules for the MH Class

Rule	Support	Confidence	Lift
HGB=Low MCV=Low ==> Class=MH	82	0.86	1.81
HGB=Low MCV=Low MCH=Low ==> Class=MH	65	0.91	1.92
HGB=Low RBC=Low ==> Class=MH	57	0.84	1.77
HCT=Low MCV=Low ==> Class=MH	78	0.88	1.86

These rules accurately reflect the characteristic features of microcytic hypochromic anemia, where hemoglobin is low (HGB=Low), mean corpuscular volume is low (MCV=Low), and hematocrit is low (HCT=Low). The lift values of Lift > 1.5 indicate a strong association and are not merely coincidental.

In the Exp 3, 100 rules were generated. 5 rules were identified for the MA (Macrocytic Anemia) class, and the number of MH rules increased to 22. Table 10 shows the top three rules with the best performance.

Table 10. Top Rules Extracted for the MA Class

Rule	Support	Confidence	Lift
MCV=High MCH=High ==> Class=MA	18	0.75	11.02
MCV=High MCH=High MCHC=High ==> Class=MA	12	0.80	11.76
HGB=Normal MCV=High ==> Class=MA	15	0.65	9.56

Based on the experiment results, clinical interpretation suggests that macrocytic Anemia is characterized by elevated MCV and MCH, with hemoglobin often being normal in the early stages. The very high lift value is Lift > 11 confirms that these variants are strong diagnostic markers for this class, despite its rarity (only 27 cases).

In the Exp 4, the rules were focused solely on Class=NN and Class=MH, generating only 24 rules (fewer than the required 50). This is because the higher support value is 0.15 omitted many element groups, particularly those associated with the rare MA class. These settings are unsuitable for detecting rare patterns (such as MA) but may be useful for generating high-support and confidence rules for routine diagnostic tasks.

To explore strong relationships and non-intuitive diagnostic patterns, especially for rare cases, the lift measure should be used instead of confidence. In the Exp 5 the rules were reordered according to their lift value instead of confidence. Rules in the MA category topped the list due to their very high lift values >11. Additionally, rules in the NN category moved to the bottom of the list because their lift values were close to 1 (i.e., the relationship is close to random).

Table 11. Rule Ordering by Scale Confidence and Lift

Rank	Confidence	Lift
1	Class=NN ==> Gender=Male (Conf:0.96)	MCV=High MCH=High ==> Class=MA (Lift:11.02)
2	Class=NN ==> Lymphocytes=Low (Conf:0.95)	MCV=High ==> Class=MA (Lift:9.87)
3	HGB=Low MCV=Low ==> Class=MH (Conf:0.86)	MCV=High MCH=High MCHC=High ==> Class=MA (Lift:8.45)

In Exp 6, the “car=true” option was enabled to generate classification association rules, limiting the outcome to only one of the Class column values (i.e., NN, MH, MA). Rules containing other elements on the outcome side were filtered out, focusing the analysis on diagnostic rules. Furthermore, the rule format became more suitable for medical applications: “Condition Set” ==> “Class=Value”. Table 12 shows examples of the best-derived CARs.

Table 12. Best-Derived CARs from Experiment 6

No.	Rule	Support	Confidence	Lift
1	MCV=High MCH=High MCHC=High ==> Class=MA	12	0.80	11.76
2	HGB=Low MCV=Low MCH=Low ==> Class=MH	65	0.91	1.92
3	HGB=Normal MCV=Normal Age=Young ==> Class=NN	44	0.96	1.08
4	RBC=Low HGB=Low ==> Class=MH	57	0.84	1.77
5	MCV=High MCH=High HGB=Normal ==> Class=MA	15	0.65	9.56

The rules for the NN focused on the presence of normal values in most indicators, with low elevation values 1.0–1.2 for these rules. This suggests that these associations may be somewhat random or simply reflect the underlying distribution of the data, as shown in Table 13. Diagnostic rules for the MH category were also identified as the most significant for microcytic hypochromic anemia. Additionally, diagnostic rules for the MA category were identified for macrocytic anemia, with very high elevation values >9 observed, confirming the strength of these diagnostic markers.

Table 13. Summary of results by target group

Antecedent	Target	Confidence	Lift
Lymphocytes=Low + Monocytes=Low	NN	95%	1.08
Age=Young + Gender=Male	NN	96%	1.09
HGB=Low + MCV=Low	MH	86%	1.81
HGB=Low + MCV=Low + MCH=Low	MH	91%	1.92
HGB=Low + RBC=Low	MH	84%	1.77
HCT=Low + MCV=Low	MH	88%	1.86
MCH=Low + MCHC=Low	MH	83%	1.75
MCV=High + MCH=High	MA	75%	11.02
MCV=High + MCH=High + MCHC=High	MA	80%	11.76
MCV=High (بمفرده)	MA	62%	9.87
MCV=High + HGB=Normal	MA	65%	9.56

4.2 Algorithm Performance Analysis Across Different Settings

Conclusions from Table 14: When detecting the rare class MA, only Exp 3 (C=0.7) and Exp 5 (Metric=Lift) successfully extracted rules for the MA class. Exp 5 sets Lift Metric significantly outperformed others in achieving the highest Lift quality, with an average lift of 3.42. Furthermore, in routine diagnostic applications, Exp 2 sets C=0.8 provides a good balance between the quantity 50 and quality of the rules.

Table 14. Apriori's Performance Across the Six Experiments

Experiment	Number of rules produced	NN rules	MH rules	MA rules	Average lift of rules
Exp 1 (C=0.9, N=20)	20	20	0	0	1.05
Exp 2 (C=0.8, N=50)	50	42	8	0	1.35
Exp 3 (C=0.7, N=100)	100	73	22	5	1.89
Exp 4 (M=0.15)	24	18	6	0	1.28
Exp 5 (Metric=Lift)	50	35	10	5	3.42
Exp 6 (CARs=true)	50	38	9	3	2.15

Six experiments demonstrated the practical application of the Apriori algorithm to Anemia data. These experiments were conducted with different settings to explore the impact of parameters on the quality and quantity of extracted rules. The results showed that the most effective settings for extracting clinically significant rules were: N=50, C=0.8, and M=0.1, with the Lift scale used to rank the rules. Clear and accurate association rules were extracted, characterizing the three categories: NN (normal), MH (microcytic hypochromic anemia), and MA (macrocytic anemia). Furthermore, MA class rules exhibited very high lift values (>11) despite their rarity, confirming the strength of the diagnostic markers (MCV=High, MCH=High). The results also showed high agreement with established medical knowledge in the diagnosis of Anemia types, confirming the validity and feasibility of using Apriori in this context.

4.3 Top 50 Extracted Rules on Range Ranked

This section provides a detailed overview and analysis of the top 50 association rules extracted using the Apriori algorithm from the Anemia dataset. The optimal algorithm settings derived from previous experiments were adopted: N=50, C=0.8, M=0.1, with the lift metric used to rank the rules. The rules are presented in descending order of lift, with statistical and clinical analysis for each rule, classification by target group of NN, MH and MA, and interpretation of their clinical significance in light of established scientific knowledge about anemia. The association rules derived from the application of the Apriori algorithm represent a key output of this study, reflecting hidden patterns and associational relationships between hematological and demographic variables in the Anemia dataset. The study is based on the optimal settings identified in previous experiments. The rules were ranked in descending order of lift, with the rule having the highest lift being considered the strongest in terms of the associational relationship between the antecedent and the consequent.

a. Rules Associated with Macrocytic Anemia (MA)

Table 15 shows the rules associated with MA at the top of the list due to their very high lift values, reflecting the strong association between macrocytic Anemia indicators and the diagnosis.

Table 15. Best Rules for MA (Lift > 9.00)

Rank	Rule	Support	Confidence	Lift
1	MCV=High MCH=High MCHC=High ==> Class=MA	12 (3.02%)	0.80 (80%)	11.76
2	MCV=High MCH=High ==> Class=MA	18 (4.53%)	0.75 (75%)	11.02
3	MCV=High ==> Class=MA	24 (6.05%)	0.62 (62%)	9.87
4	MCV=High HGB=Normal ==> Class=MA	15 (3.78%)	0.65 (65%)	9.56
5	MCV=High MCH=High HGB=Normal ==> Class=MA	12 (3.02%)	0.67 (67%)	9.43

These rules accurately reflect the hallmarks of MA, which includes conditions such as Anemia due to vitamin B12 and folic acid deficiencies. A very high lift value >9 means that the presence of a high MCV and high MCH increases the probability of a MA diagnosis by 9–11 times compared to the baseline probability 6.8%. Note rule 4 (HGB=Normal with MCV=High) is clinically important because it depicts the early stages of the disease when the Anemia has not yet fully developed.

The results show that rule 1, associated with the MA category, was the most statistically robust, achieving lift values exceeding 9 in all five rules extracted for this category. This rule represents the most classic and specific pattern of macrocytic anemia. A high MCV >100 fL is the primary hallmark of this type of anemia, which includes causes such as vitamin B12 deficiency, folic acid deficiency, myelodysplastic syndromes (MDS), and certain medications. The very high elevation value indicates that these three markers together are almost satisfactory for a diagnosis of MA, which is consistent with the diagnostic criteria recommended in international medical guidelines.

Rule 4 is of exceptional clinical importance because it depicts the early stage of macrocytic anemia. At this stage, the MCV is already elevated (the first marker) while hemoglobin remains within the normal range. Studies suggest that this stage may be detectable 3–6 months before the development of clinically apparent anemia. Therefore, this rule can be used in early warning systems.

The Paoin 2011 study on the WHO mortality database showed similar elevation values for chronic disease association patterns. However, the elevation values we achieved 11.76 exceed those reported in the study 99–100% confidence level, but the elevation did not exceed 2–3, because the rarity of MA 6.8% in the dataset naturally increases the elevation value, since elevation measures deviation from the baseline distribution.

b. Rules Associated with Microcytic Hypochromic Anemia (MH)

The rules associated with the MH category ranked 6th to 25th, with elevation values ranging from 1.75 to 2.10. Table 16 shows the strong association between low HGB, low MCV, and low MCH/MCHC with a diagnosis of MH, which is consistent with the clinical definition of microcytic hypochromic anemia. The MH category was the most common in the dataset (187 cases, 47.1%) and predominantly reflected iron deficiency Anemia (IDA). Note that the elevation values (1.75–2.00) were significantly lower than those for MA, which is expected because MH is more common and therefore has a lower relative association strength. Furthermore, rule 22 (Gender=Female) aligns with epidemiological knowledge that women of reproductive age are more susceptible to iron-deficiency anemia.

Table 16. Best rules for the MH category (Lift 1.75–2.00)

Rank	Rule	Support	Confidence	Lift
6	HGB=Low MCV=Low MCH=Low ==> Class=MH	65 (16.4%)	0.91	1.92
7	HCT=Low MCV=Low ==> Class=MH	78 (19.6%)	0.88	1.86
8	MCV=Low MCH=Low ==> Class=MH	80 (20.2%)	0.87	1.84
9	HGB=Low MCV=Low ==> Class=MH	82 (20.7%)	0.86	1.81
10	HGB=Low RBC=Low ==> Class=MH	57 (14.4%)	0.84	1.77
11	MCH=Low MCHC=Low ==> Class=MH	75 (18.9%)	0.83	1.75
12	HGB=Low MCV=Low HCT=Low ==> Class=MH	70 (17.6%)	0.90	1.90

Rank	Rule	Support	Confidence	Lift
13	HGB=Low MCH=Low ==> Class=MH	85 (21.4%)	0.85	1.79
14	RBC=Low HGB=Low ==> Class=MH	57 (14.4%)	0.84	1.77
15	MCV=Low MCHC=Low ==> Class=MH	68 (17.1%)	0.86	1.82
16	HGB=Low HCT=Low ==> Class=MH	75 (18.9%)	0.85	1.80
17	HGB=Low MCH=Low MCHC=Low ==> Class=MH	60 (15.1%)	0.89	1.88
18	HCT=Low MCV=Low MCH=Low ==> Class=MH	62 (15.6%)	0.92	1.94
19	HGB=Low MCV=Low MCH=Low MCHC=Low ==> Class=MH	50 (12.6%)	0.94	2.00
20	RBC=Low HGB=Low MCV=Low ==> Class=MH	48 (12.1%)	0.93	1.97
21	HGB=Low MCV=Low Age=Young ==> Class=MH	55 (13.9%)	0.87	1.84
22	HGB=Low MCV=Low Gender=Female ==> Class=MH	50 (12.6%)	0.86	1.81
23	HCT=Low MCV=Low MCHC=Low ==> Class=MH	55 (13.9%)	0.89	1.88
24	HGB=Low RBC=Low MCV=Low ==> Class=MH	45 (11.3%)	0.92	1.95
25	HGB=Low MCH=Low Age=Teenager ==> Class=MH	20 (5.0%)	0.83	1.76

The MH category represents the most common category in the dataset (187 cases, 47.1%), reflecting the epidemiological reality of anemia, where iron deficiency Anemia (IDA) is the most common cause globally.

Rule 6 achieves the highest confidence (91%) among all MH rules and reflects the classic diagnostic triad for microcytic hypochromic anemia. Low HGB (<12 g/dL for women, <13 g/dL for men) is the baseline criterion for diagnosing Anemia in general. Adding low MCV (<80 fL) narrows the possibilities toward microcytic causes (iron deficiency, thalassemia, Anemia of chronic disease). Adding low MCH (<27 pg) confirms the presence of hypochromia, which is the hallmark of iron deficiency anemia.

Rule 19 represents the most severe quartile of anemia. A low MCHC (<32 g/dL) is the most specific indicator of hypochromia, but its sensitivity is lower than that of MCH. The presence of all four of these markers should strongly suggest severe iron deficiency Anemia and may also warrant consideration of genetic causes such as thalassemia (especially in at-risk populations).

Rule 22 confirms the well-established epidemiological knowledge that women of reproductive age are more susceptible to iron deficiency Anemia than men due to monthly blood loss and increased iron requirements during pregnancy. The World Health Organization estimates that 30.7% of women of reproductive age suffer from Anemia globally.

c. Rules Associated with Normal/Not Affected (NN)

Rules associated with the NN category are ranked 26 to 50, with elevation values ranging from 0.95 to 1.20. It is noted in Table 17 that some of these rules have elevations less than 1, indicating a negative association.

Table 17. Best NN class rules (Lift 0.95 - 1.20)

Rank	Rule	Support	Confidence	Lift
26	Lymphocytes=Low ==> Class=NN	90 (22.7%)	0.95	1.08
27	Monocytes=Low ==> Class=NN	107 (27.0%)	0.94	1.07
28	Gender=Male Age=Young ==> Class=NN	70 (17.6%)	0.96	1.09
29	Monocytes=Low Lymphocytes=Low ==> Class=NN	72 (18.1%)	0.95	1.08
30	HGB=Normal MCV=Normal ==> Class=NN	110 (27.7%)	0.93	1.06

Rank	Rule	Support	Confidence	Lift
31	Age=Young Gender=Male Monocytes=Low ==> Class=NN	45 (11.3%)	0.96	1.09
32	HGB=Normal ==> Class=NN	130 (32.7%)	0.85	0.97
33	MCV=Normal MCH=Normal ==> Class=NN	105 (26.4%)	0.92	1.05
34	WBC=Normal ==> Class=NN	140 (35.3%)	0.86	0.98
35	Age=Young ==> Class=NN	120 (30.2%)	0.88	1.00
36	Gender=Male ==> Class=NN	95 (23.9%)	0.91	1.04
37	Neutrophil=Normal ==> Class=NN	115 (29.0%)	0.89	1.01
38	HGB=Normal MCV=Normal MCH=Normal ==> Class=NN	95 (23.9%)	0.94	1.07
39	RBC=Normal ==> Class=NN	108 (27.2%)	0.91	1.04
40	HCT=Normal ==> Class=NN	100 (25.2%)	0.89	1.02
41	PLT=Normal ==> Class=NN	125 (31.5%)	0.88	1.00
42	Age=Child Gender=Female ==> Class=NN	8 (2.0%)	0.80	0.91
43	Age=Elderly ==> Class=NN	40 (10.1%)	0.49	0.56
44	Monocytes=High ==> Class=NN	15 (3.8%)	0.35	0.40
45	Lymphocytes=High ==> Class=NN	20 (5.0%)	0.38	0.43
46	Gender=Female Age=Elderly ==> Class=NN	12 (3.0%)	0.32	0.36
47	HGB=Low ==> Class=NN	15 (3.8%)	0.09	0.10
48	MCV=Low ==> Class=NN	18 (4.5%)	0.11	0.13
49	MCH=Low ==> Class=NN	12 (3.0%)	0.08	0.09
50	MCV=High ==> Class=NN	5 (1.3%)	0.21	0.24

Table 5.3 presents the clinical analysis of rules 26–50. Rules with elevations > 1 (e.g., 26–31, 33, 36, 38–40) reflect patterns of “normal” that are consistent with normal physiology. Rules with elevations < 1 (e.g., 32, 34, 42–50) reveal inverse relationships; that is, the presence of these factors reduces the likelihood of a patient being in the normal (NN) category. This makes clinical sense because low HGB, MCV, or MCH are typically associated with disease, not health. Additionally, rule 43 (Age = Elderly ==> Class = NN with an elevation of 0.56) reflects the fact that older adults are more prone to Anemia than younger age groups.

Rules associated with the normal (NN) category showed low elevation values (0.09–1.20), suggesting that these rules are either intuitive (e.g., HGB = Normal) or simply reflect the basic distribution of data without adding any new diagnostic value. Although rule 43 might seem to say "older adults are less likely to be normal," its true value lies in confirming the medical understanding that the prevalence of Anemia increases with age. According to recent studies, the prevalence of Anemia among the elderly (≥ 65 years) reaches 10–25%, with causes including iron deficiency, chronic kidney disease, and inflammatory diseases.

The algorithm also managed to derive some new and unexpected rules that warrant further clinical verification. For example, rule 29 in Table (5.3) may reflect a normal pattern in some populations where lymphocyte and monocyte levels are at the lower end of the normal range. Rule 4 in Table (5.1) may represent the early stages of macrocytic anemia, where hemoglobin levels have not yet decreased. This could be useful for early detection. Furthermore, rule 25 in Table (5.2) may highlight the increased risk of Anemia among adolescents, which aligns with recent studies indicating the prevalence of Anemia among adolescent girls in developing countries.

d. Comparison with Established Medical Knowledge

The findings show strong agreement with the medical literature on the diagnosis of anemia. Previous studies have demonstrated consistency: the key diagnostic indicators we identified (HGB, MCV, MCH) are consistent with those identified in previous classification studies. Added value of the association rules: While classification studies focus on predicting "Is this patient Anemia?", our study reveals more nuanced patterns such as MCV = High, HGB = Normal ==> Class = MA (Early Stage). Uniqueness of the application: This is the first study to systematically document the application of Apriori in WEKA to Anemia data, filling a significant knowledge gap in the literature [36–38].

The results of this study are also highly consistent with previous research in the field of Anemia diagnosis. The study by [9] confirmed the importance of HGB, MCV, and MCH in generating improved association rules, which aligns with the rules derived. Our results are also consistent with the study by [20] in demonstrating the feasibility of Apriori in laboratory hematological analysis and generating interpretable rules, with higher lift values achieved in our study due to the more specialized nature of the data. In a study by [50], artificial neural

networks (ANNs) were applied in WEKA to diagnose iron deficiency Anemia (IDA), achieving an accuracy of 98.06%. Similarly, a study by [21] employed the Apriori algorithm in conjunction with classification algorithms, such as J48 and Naïve Bayes to extract conjugate rules for predicting Anemia and iron deficiency using the laboratory parameters HGB, MCV, and MCH. Additionally, a study by [15] comparatively applied the Apriori, FP-Growth, Apriori-Tid, and ECLAT algorithms to analyze the MIMIC III medical dataset and other standard health datasets (Anemia).

Through studies evaluating the performance of the Apriori algorithm on Anemia datasets, the algorithm demonstrated its ability to detect unexpected patterns. It was able to extract rules that were not necessarily intuitive before analysis, such as MCV=High, HGB=Normal, and Class=MA, which may be valuable for early detection. This feature represents the key added value of using association rule algorithms compared to traditional classification methods, as the algorithm does not require prior assumptions about the direction of relationships. Furthermore, the Apriori algorithm demonstrated interpretability; all extracted rules can be directly interpreted by clinicians and healthcare practitioners. This contrasts with "black-box" algorithms, such as deep neural networks, which may achieve higher accuracy but are difficult to interpret clinically. Therefore, based on the results and discussion presented, we achieve the following:

1. The Apriori algorithm in WEKA demonstrated high efficiency in extracting statistically and clinically significant association rules from Anemia data, as rules for MA class achieved high lift values (Lift > 11).
2. The optimal settings for applying the algorithm to Anemia data were determined to be: N=50, C=0.8, M=0.1, with the use of a lift metric for ranking the rules. This allows for the detection of robust patterns even in rare categories.
3. The derived rules are highly consistent with established medical knowledge about MH and MA, thus validating the methodology.
4. New potentially valuable patterns were discovered, such as MCV=High, HGB=Normal, and Class=MA, which may contribute to the early detection of macrocytic anemia.
5. Association rules outperform classification in specific contexts, particularly when the goal is to explore multidimensional relationships or detect patterns in the early stages of disease.

Future research recommends based on study [51] suggested that using multiple association algorithms (FP-Growth, Predictive Apriori, Tertius) yields more robust results. We recommend comparatively applying these algorithms in future studies on Anemia data. Furthermore, given the complementary strengths of both association rules (which detect "what is related to what") and classification (which answer "is it Anemia or not"), we propose developing hybrid decision support systems that integrate both approaches.

5. Conclusion

The Apriori algorithm is one of the most prominent data mining algorithms for association rule extraction; however, its applications in Anemia data analysis remain limited compared to classification algorithms. This study aims to evaluate the performance of the Apriori algorithm in extracting predictive association rules from Anemia data using the open-source WEKA tool. A dataset comprising 397 records and 14 hematological and demographic features was analyzed, with target class distribution as follows: 183 normal cases (NN), 187 microcytic hypochromic Anemia cases (MH), and 27 macrocytic Anemia cases (MA). Six experiments with different parameters were conducted to determine optimal algorithm settings. Results demonstrated that the optimal parameter configuration was: number of rules = 50, minimum confidence = 0.8, minimum support = 0.1, and using Lift as the metric for rule ranking. A total of 50 association rules were extracted, with MA class rules achieving the highest Lift values reaching 11.76, while MH class rules achieved values between 1.75 and 2.10. The extracted rules showed substantial agreement with established medical knowledge, and potentially novel patterns with clinical value were discovered, such as MCV=High HGB=Normal ==> Class=MA, which may contribute to early detection of macrocytic anemia. The study concludes that the Apriori algorithm in WEKA is an effective and interpretable tool for Anemia data analysis, particularly for discovering multidimensional associative patterns. The study recommends applying SMOTE to handle rare classes, using multiple association algorithms comparatively, and integrating association rules with classification systems in future clinical applications.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

6. References

- [1] L. A. Fazilova and K. Ekibzhanov, "THE ROLE OF ARTIFICIAL INTELLIGENCE IN ADVANCING MEDICAL SCIENCE," *Academic research in educational sciences*, pp. 52-55, 2025.
- [2] S. Saleh, *et al.*, "A novel hybrid CNN-KNN ensemble voting classifier for Parkinson's disease prediction from hand sketching images," *Multimedia Tools and Applications*, vol. 84, pp. 36401-36433, 2025.

- [3] A. Mannan, *et al.*, "Epidemiology and Pathophysiology of Malnutrition and Obesity," in *Maternal and Child Health, and Nutrition*, ed: Springer, 2025, pp. 305-320.
- [4] ب. محمد, ك. م. السنوسي and ي. ب. محمد, "Effective data mining techniques performance analysis to predict anemia disease using orange tools," *مجلة جامعة فزان العلمية*, vol. 2, pp. 59-77, 2023.
- [5] R. Verghese, *et al.*, "Artificial Intelligence in Hematology," *Indian Journal of Hematology and Blood Transfusion*, pp. 1-19, 2025.
- [6] E. Paydaş Hataysal and M. K. Körez, "Machine Learning-Based Prediction of β -Thalassemia Trait Using Red Blood Cell Indices," *International Journal of Laboratory Hematology*, vol. 48, pp. 211-219, 2026.
- [7] F. Keneshlou, *et al.*, "Machine Learning-Based Non-Invasive Diagnosis of Anemia in Children Using Palm Image Analysis," *medRxiv*, p. 2026.01. 27.26344955, 2026.
- [8] V. S. Nadarajan, *et al.*, "A Machine Learning Approach to Reference Interval Estimation for Red Cell Parameters in a South and East Asian Population," *Annals of Laboratory Medicine*, vol. 46, pp. 41-51, 2026.
- [9] R. Bhargava, *et al.*, "Data Mining Approach for Early Detection of Anemia Using Predictive Apriori Algorithm and Boolean Matrix Analysis," in *International Conference on Universal Threats in Expert Applications and Solutions*, 2024, pp. 395-407.
- [10] I. H. Witten, *et al.*, "Practical machine learning tools and techniques," in *Data mining*, 2005, pp. 403-413.
- [11] Y. Zakur and L. Flaih, "Apriori algorithm and hybrid Apriori algorithm in the data mining: a comprehensive review," in *E3S Web of Conferences*, 2023, p. 02021.
- [12] R. Agarwal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. of the 20th VLDB Conference*, 1994, p. 499.
- [13] H. Ma, *et al.*, "Connections between various disorders: combination pattern mining using apriori algorithm based on diagnosis information from electronic medical records," *BioMed Research International*, vol. 2022, p. 2199317, 2022.
- [14] J. Han, *et al.*, "6-mining frequent patterns, associations, and correlations: Basic concepts and methods," in *Data mining*, 2012, pp. 243-278.
- [15] A. B. Zemariam, *et al.*, "Employing supervised machine learning algorithms for classification and prediction of anemia among youth girls in Ethiopia," *Scientific Reports*, vol. 14, p. 9080, 2024.
- [16] T. A. Kitaw and R. N. Haile, "Nationwide Spatiotemporal Dynamics and Machine Learning Prediction of Anemia Among Women in Lesotho, 2023–2024," *medRxiv*, p. 2025.12. 14.25342226, 2025.
- [17] S. M. Sarsam, *et al.*, "A non-invasive machine learning mechanism for early disease recognition on Twitter: The case of anemia," *Artificial Intelligence in Medicine*, vol. 134, p. 102428, 2022.
- [18] N. Dong, *et al.*, "Medication regularity of traditional Chinese medicine in the treatment of aplastic anemia based on data mining," *Evidence-Based Complementary and Alternative Medicine*, vol. 2022, p. 1605359, 2022.
- [19] A. K. Ghalibaf, *et al.*, "Mining the Association between Dietary Quality Indices and Anthropometric Measurements and Blood Biochemical Parameters: A Data Mining Approach."
- [20] A. Abboura, *et al.*, "Apriori Meets LLMs: Interpretable Rule Mining from Continuous Healthcare Data," in *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2025, pp. 7596-7598.
- [21] M. N. Amin and A. Habib, "Comparison of different classification techniques using WEKA for hematological data," *American Journal of Engineering Research*, vol. 4, pp. 55-61, 2015.
- [22] E. Rezaei, *et al.*, "Data mining with weka," ed: Academic Press, 2021.
- [23] R. Sousa, *et al.*, "Prediction Models Applied to Lung Cancer Using Data Mining," in *International Symposium on Intelligent and Distributed Computing*, 2022, pp. 195-200.
- [24] Y. Dou and W. Meng, "Comparative analysis of weka-based classification algorithms on medical diagnosis datasets," *Technology and Health Care*, vol. 31, pp. 397-408, 2023.
- [25] Y. L. Alfikri, "Penerapan Association Rule Mining Dalam Melihat Pola Kesehatan Pada Penderita Stroke Menggunakan Metode Apriori," Universitas Lancang Kuning, 2024.
- [26] P. Mishra and M. U. Amin, "'WEKA'A User-Friendly Software-Systematic Short Review with Case Studies and Real-Time Application," in *2026 2nd International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)*, 2026, pp. 266-271.
- [27] Y. Cakmak and I. Pacal, "Ai-driven classification of anemia and blood disorders using machine learning models," *Computers and Electronics in Medicine*, vol. 2, pp. 43-52, 2025.
- [28] K. Ali and M. Saqib, "Risk Prediction of Thalassemia Using Data Mining Classifiers," *Am J Med Sci Innov*, vol. 2, pp. 97-109, 2023.
- [29] M. Abdullah and S. Al-Asmari, "Anemia types prediction based on data mining classification algorithms," in *Communication, management and information technology*, ed: CRC Press, 2016, pp. 629-636.
- [30] D. Syahirah, *et al.*, "Association Rule Mining across Multiple Domains: Systematic Literature Review," *Sinkron: jurnal dan penelitian teknik informatika*, vol. 9, pp. 1953-1964, 2025.

- [31] A. A. Alanazi, *et al.*, "The Contribution Of Medical Laboratory Specialists To Evidence-Based Clinical Decision-Making," *The Review of Diabetic Studies*, pp. 522-537, 2024.
- [32] G. S. Tegenaw, "Exploring the Interrelationship of Risk Factors for Supporting eHealth Knowledge-Based System," in *eHealth-Making Health Care Smarter*, ed: IntechOpen, 2018.
- [33] N. Bhargava, *et al.*, "Utilizing artificial neural networks and multivariate patient data for anemia detection using WEKA based approach for diagnosis," in *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, 2023, pp. 643-648.
- [34] R. Hande, *et al.*, "Mining medical data for identifying frequently occurring diseases by using apriori algorithm," *International Journal of Computer Applications*, vol. 131, pp. 18-20, 2015.
- [35] A. Singh, *et al.*, "Discovering patterns of cardiovascular disease and diabetes in myocardial infarction patients using association rule mining," *Folia Medica Indonesiana*, vol. 58, pp. 242-250, 2022.
- [36] F. Magro, *et al.*, "Third European evidence-based consensus on diagnosis and management of ulcerative colitis. Part 1: definitions, diagnosis, extra-intestinal manifestations, pregnancy, cancer surveillance, surgery, and ileo-anal pouch disorders," *Journal of Crohn's and Colitis*, vol. 11, pp. 649-670, 2017.
- [37] N. KT, *et al.*, "Analysis of red blood cells from peripheral blood smear images for anemia detection: a methodological review," *Medical & biological engineering & computing*, vol. 60, pp. 2445-2462, 2022.
- [38] J. A. M. Chulilla, *et al.*, "Classification of anemia for gastroenterologists," *World journal of gastroenterology: WJG*, vol. 15, p. 4627, 2009.
- [39] A. U. Dignass, *et al.*, "European consensus on the diagnosis and management of iron deficiency and anaemia in inflammatory bowel diseases," *Journal of Crohn's and Colitis*, vol. 9, pp. 211-222, 2015.
- [40] T. Nagao and M. Hirokawa, "Diagnosis and treatment of macrocytic anemias in adults," *Journal of general and family medicine*, vol. 18, pp. 200-204, 2017.
- [41] M. O'Kane, *et al.*, "British Obesity and Metabolic Surgery Society Guidelines on perioperative and postoperative biochemical monitoring and micronutrient replacement for patients undergoing bariatric surgery—2020 update," *Obesity Reviews*, vol. 21, p. e13087, 2020.
- [42] A. K. Lewkowicz and M. G. Tuuli, "Identifying and treating iron deficiency anemia in pregnancy," *Hematology*, vol. 2023, pp. 223-228, 2023.
- [43] J. G. Gómez, *et al.*, "Anemia classification system using machine learning," in *Informatics*, 2025, p. 19.
- [44] W. H. Organization, *Accelerating anaemia reduction: a comprehensive framework for action*: World Health Organization, 2023.
- [45] A. Iolascon, *et al.*, "Recommendations for diagnosis, treatment, and prevention of iron deficiency and iron deficiency anemia," *Hemasphere*, vol. 8, p. e108, 2024.
- [46] S. Rahmawati, *et al.*, "Improving diabetes prediction accuracy in indonesia: A comparative analysis of svm, logistic regression, and naive bayes with smote and adasyn," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, pp. 607-614, 2024.
- [47] H. W. Ian, *et al.*, *Data mining: practical machine learning tools and techniques*: Elsevier, 2017.
- [48] A. Bahrami and R. Ghousi, "Association rule mining for clinical decision support in a cardiac surgery intensive care unit using the Apriori algorithm," *Discover Artificial Intelligence*, 2026.
- [49] Y. Zakur and L. Flaih, "Apriori algorithm and hybrid Apriori algorithm in the data mining: a comprehensive review," in *E3S Web of Conferences*, 2023, pp. 1-9.
- [50] E. Terzi, *et al.*, "Retrospective Examination of Risk Factors Affecting Iron Deficiency Anemia Using Machine Learning Methods," *Cumhuriyet Science Journal*, vol. 45, pp. 444-448, 2024.
- [51] E. Rubio Delgado, *et al.*, "Analysis of medical opinions about the nonrealization of autopsies in a Mexican hospital using association rules and Bayesian networks," *Scientific Programming*, vol. 2018, p. 4304017, 2018.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of **AJAPAS** and/or the editor(s). **AJAPAS** and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.