# A Role of Data Mining Techniques to Predict Anemia Disease

Sasikala, PhD
Assistant Professor
Department of HI, FPHTM
Jazan University, KSA

Rasitha Banu, PhD
Assistant Professor
Department of HI, FPHTM
Jazan University, KSA

Thgani Babiker, PhD
Assistant Professor
Department of HE, FPHTM
Jazan University, KSA

Pushpalatha, PhD
Assistant Professor
Department of HE, FPHTM
Jazan University, KSA

## ABSTRACT
Iron deficiency is the most known form of nutritional deficiency. It is most common in undernourishment and is most prevalent in young children, women of childbearing age, and pregnant women. Iron deficiency in children causes developmental delays and behavioral disruptions, and in pregnant women it raises the risk of premature labor and delivery of a baby with low birth weight. There was an awareness in the past three decades, about the intake of iron supplement for infants that was resulted due to this childhood iron-deficiency anemia all over the world. Even though, it is always better to detect the disease at an earlier stage of life to prevent further harmful effects and to devise proper treatment. In this study, the anemia is taken into consideration for early prediction and diagnosis of the disease by using the data mining technique to analyze the data. In healthcare organizations the volume of data is more. To get knowledge from those data we need an efficient technique. Data mining is used for the purpose of discovering knowledge from vast amount of database. To classify the stages of anemia, classification technique which is one of the data mining technique is used. The data is collected from 200 household of students from Public Health College in Jazan University. The research work is done with WEKA open-source software under Windows7 environment. An experimental study is carried out using data mining techniques such as J48, Random Forest tree and hoeffding tree. As a result, the performance is evaluated for three classification techniques and their accuracy compared through confusion matrix. It has been concluded that Random Forest tree gives better accuracy than the J48 and Hoeffding tree technique.

## Keywords
Anemia, Data Mining, Classification Technique, J48, Random Forest tree and Hoeffding Tree

## 1. INTRODUCTION
Iron-deficiency is anemia, caused by the lack of iron. Anemia is characterized as a decrease in red blood cells, or a decrease in blood hemoglobin. When onset is slow, symptoms are often vague like feeling tired, weak, in need of breath, or having decreased ability to exercise. Anemia that develops increasingly often has more extreme effects, including fatigue or rising thirst.

Anemia is usually significant before an individual becomes noticeably pale. The growth and developmental problems may affect Children due to this iron deficiency or anemia. In medical research, prediction of illness at the right time is the core concern for the prevention and efficient treatment plan for physicians. Often, in the absence of this precision may lead for severity. Data mining gives the healthcare industry immense potential for allowing health systems to actively use data and analytics to recognize shortfalls and suggest practice guidelines that optimize treatment and reduce costs. The machine learning algorithms are used successfully in various domains like healthcare, weather forecasting, stock price prediction, product recommendation for making prediction. The main aspect of medical science research is the prediction of causes and factors for various diseases. The healthcare data are being used in medical domain to predict epidemics, to detect disease, to improve quality of life and avoid early deaths [1]. This work is processed by exploring three different classification algorithms for anemia prediction.

Since, Iron deficiency is normal, but the actual deficiency is that in iron storages. The most serious case is iron deficiency that leading to a decrease in the production of iron-containing compounds such as hemoglobin and low red blood cells. According to WHO report 2020, Iron-deficiency/ anemia affected several numbers of people about 614 million of Females and 280 million of children [2].

The rest of the paper is organized as follows: Section 2 gives the brief reviews of existing related work. In section 3, we discuss various types of anemia diagnosis tests. Section 4 presents proposed methodology. Section 5 presents experimental details and discussion. Finally, we conclude in section 6.

### 1.1 Impacts of Anemia and its stages
Hemoglobin is an iron-rich protein that helps the red blood cells bring oxygen to the rest of the body from the lungs. When a person has anemia, his body isn't getting enough blood that is rich in oxygen. This can make him to feel exhausted or sick. He may also experience shortness of breath, dizziness, headaches, or an erratic heartbeat. There are different types of anemia for different reasons. The most common type of anemia is caused by the shortage of iron in our body. Generally, our bone marrow needs hemoglobin to make iron. Our body cannot produce enough hemoglobin for the red blood cells without adequate iron.

Another main reason for anemia is the vitamin deficiency. Besides iron, folate and vitamin B-12 are needed in our body to produce enough healthy red blood cells. A diet that lacks these and other main nutrients may cause reduced development of red blood cells. The other reasons are rare like Anemia of inflammation, Aplastic anemia, Hemolytic anemias. Anemia is categorized as severe, moderate and

mild; This categorization level of Hemoglobin level varies for both Male and Female; If the level given for female is <7.0, 7.0 – 9.9 and 10.0 – 12.0 respectively; For Male it is given as <9.0,9.0 – 11.9 and 12.0 – 13 respectively.

Iron deficiency anemias and Vitamin deficiency anemias can be avoided by having a diet that includes a variety of vitamins and minerals. Iron is rich in foods like meats, beans, lentils, iron-fortified cereals, dark green leafy vegetables, and dried fruit. Folate is a nutrient, and its synthetic form folic acid, that can be found in fruits and fruit juices, dark green leafy vegetables, green peas, kidney beans, peanuts, and enriched grain products, such as bread, cereal, pasta, and rice. The vitamin B-12 is rich in meat, dairy products, and fortified cereal and soy products. The Vitamin C Foods also good to avoid anemia. The foods rich in vitamin C include citrus fruits and juices, peppers, broccoli, tomatoes, melons, and strawberries. While eating these foods we can increase iron absorption and so we it can prevent from anemia disease.

## 1.2 PROBLEM STATEMENT

This research focuses on the performance analysis of different data mining techniques applied on data set to predict anemia disease.

## 1.3 OBJECTIVES

### General Objective:

The aim of this study is to develop a predictive model using J48 and Random forest tree and hoffeding tree classifier.

### Specific objectives

a. To implement data mining techniques such as J48, hoffeding tree and Random forest tree to predict Anemia from the Anemia dataset.

b. To compare the performance of classifiers to identify which classifier predict the disease correctly with high accuracy in terms of confusion matrix and help the physicians to predict the disease and take decision at proper time.

## 2. REVIEW OF LITERATURE

[Arun,et.al], stated in his paper that Iron deficiency anemia (IDA), a type of microcytic and hypochromic anemia, occurs when an individual's iron supply is lower than the physiological amount required to produce hemoglobin (Hgb). He suggested that Pregnancy anemia can be aggravated by various conditions such as uterine or placental bleedings, gastrointestinal bleedings, and peripartum blood loss. [3]

[Gupta.P, et al.], described in his work that, Iron deficiency and anemia are associated with impaired neurocognitive development and immune function in young children. In his work he analyzed and described the prevalence of iron deficiency (ID), anemia, and iron deficiency anemia (IDA) among children 1–5 years using data from the 2007–2010 National Health and Nutrition Examination Survey (NHANES).[4]

[Jimenez, K., et,al], they pointed in his work that anemia affects one-fourth of the world's population, and iron deficiency is the predominant cause. He stated that Anemia is associated with chronic fatigue, impaired cognitive function, and diminished well-being. Patients with iron deficiency anemia of unknown etiology are frequently referred to a gastroenterologist because in most cases the condition has a gastrointestinal origin. He suggested that only proper management improves quality of life, alleviates the symptoms of iron deficiency, and reduces the need for blood transfusions.[5]

[Reid, S., et al], were recently identified that PALB2 as a nuclear binding partner of BRCA2. The Fanconi anemia subtype FA-D1 and predispose to childhood malignancies were caused by Biallelic BRCA2 mutations. The pathogenic mutations in PALB2 (also known as FANCN) was identified in seven families of USA and were affected with Fanconi anemia and cancer in early childhood which demonstrating that biallelic PALB2 mutations cause a new subtype of Fanconi anemia, FA-N, and, similar to biallelic BRCA2 mutations, leads a high risk of childhood cancer.[6]

[Nemeth, E. and T. Ganz],they stated in their work about Anemia of inflammation (AI, also called anemia of chronic disease) is a common, typically normocytic, normochromic anemia that is caused by an underlying inflammatory disease. they diagnosed when serum iron concentrations are low despite adequate iron stores, as evidenced by serum ferritin that is not low. In the setting of inflammation, it may be difficult to differentiate AI from iron deficiency anemia, and the 2 conditions may coexist. [7]

## 3. METHODOLOGY

## 3.1 DATA COLLECTION

A simple pre-coded questionnaire was developed, and data was collected from 206 household of students from Public Health College in Jazan University. The data includes clinical and their demographic background. There are totally 206 instances in the dataset. In our research work we have taken 9 attributes which will be used to classify the data. The data set is given below table 1.

**Table 1: Anemia data set**

| SN | Attribute Name | Values |
|---|---|---|
| 1 | Age | Numeric |
| 2 | Gender | Male, Female |
| 3 | Iron supplement | T or F |
| 4 | Reproductive Status | Numeric |
| 5 | Working hours | Numeric |
| 6 | Education | Numeric |
| 7 | How serious anemia | Numeric |
| 8 | Anemia awareness | Numeric |
| 9 | Anemia serious | Numeric |
| 10 | Quality of care | Numeric |

## 3.2 PROPOSED SYSTEM

The proposed system is given below:

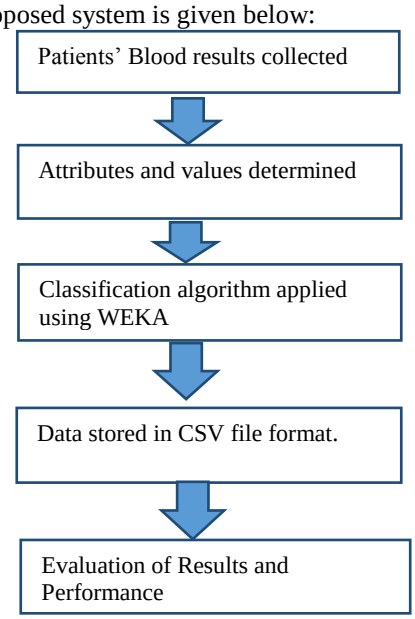


**Fig 1: Proposed System Architecture**

### a. Preprocessing

The pre-processing of data is basic process that resides in all the Data mining technique that consists of remodeling data into a noticeable format. In general, the gathering of Real-world information is typically incomplete, inconsistent, and some missing information in bound behaviors or trends that leads for several errors. This pre-processing might be a verified technique of breakdown such problems. There are several tasks in this pre-processing technique to prepare data for any process. These tasks embody information improvement, integration, transformation, and information reduction. The Data is collected from the 206-social unit of students from Public Health school in Jazan University. The collected Data/ information was checked for the presence of error in information entry together with misspellings and missing information. There is an availability of some missing information inside the collected data. In this research, replace with missing values filter is used to fill the missing values to form the information complete.

### b. Classification

Classification is also one of the techniques in Data mining Technique. it is used to classify the information supported similarity of instances. There are a unit of 2 varieties of learnings like supervised and unsupervised. During supervised learning, the trained and predefined data is offered. The familiar classification techniques area Decision trees and neural networks so on.

### c. Decision tree

It is the most popular classification techniques in data processing. it is tree-like graph. Testing every attribute delineated as internal node and every branch represents associate degree outcome of the check, and also the leaf node represents categories. it is a graphical illustration of attainable solutions, supported these solutions, optimum course of action is disbursed. During this analysis, 3 decision tree classifiers are used. They are Random Forest tree, Hoeffding tree and J48 to classify the anemia information set.

The algorithmic rule of J48 and Random Forest tree and hoeffding tree is given below.

#### *3.2.1. J48 algorithmic rule*

J48 may be a tree-based learning approach. The divide-and-conquer algorithmic rule is employed in J48 to separate a root node into a set of 2 partitions until leaf node (target node) occur in tree. the subsequent steps area unit want to construct the tree structure for given set T of total instances.

Step 1: If all the instances in T belong to constant cluster category or T has fewer instances, than the tree is leaf tagged with the foremost frequent category in T.

Step 2: If step one doesn't occur then choose a check supported one attribute with a minimum of 2 or bigger attainable outcomes. This check is taken into account as a root node of the tree with one branch of every outcome of the check, partition T into corresponding T1, T2, T3, etc., in keeping with the result for every various case, and the same is also applied to every sub node in algorithmic manner.

Step 3: The ranking of knowledge gain and default gain magnitude relation is completed by heuristic criteria in algorithmic rule J48. [12]

#### *3.2.2. Random Forest Tree algorithmic rule*

In machine learning the Random forest Tree is one in all the economical learning strategies for classification, regression and alternative tasks that operates by constructing a mess of call trees at coaching time [10],[11]. The output of the class with the mode of the categories for classification or mean prediction is regression of the individual trees. {the decision|the choice} trees' habit of over fitting to their coaching set makes of Random decision forest.

#### *3.2.3. Hoeffding tree algorithmic rule*

The Hoeffding tree is associate degree progressive call tree learner for large information streams. It grows incrementally a choice tree supported the theoretical assurances of the Hoeffding certain.

In this algorithmic rule a node is expanded as before long as there's adequate applied math proof that associate degree best cacophonic feature exists, a choice supported the distribution freelance Hoeffding certain. The model learned by the Hoeffding tree is asymptotically nearly a twin of the one engineered by a non-incremental learner once the quantity of coaching instances is giant enough.[13]

In this algorithmic rule, the information is hold on within the main memory and tree organization with one root node is initialized in its start. In second step each coaching information is filter down incrementally to an acceptable leaf node. In third step every leaf node has enough information needed to form call regarding next step. The leaf node information estimates the knowledge gain once the other attribute is split. In fourth step we discover the simplest attribute at a node and perform a check on provided information to plan that attribute will manufacture the higher result than alternative attributes mistreatment Hoeffding certain. The quantity of tests is applied to the attribute to decide that attribute can give higher result than the other node, which node leads to cacophonic the node for growth of tree.

The attribute comparisons area unit higher during this algorithmic rule than alternative algorithms. the good thing about this algorithmic rule is its memory consumption. It will use terribly less memory and delivers increased utilization with sampling of information.

# 4. EXPERIMENTS WITH WEKA

In this study, the Waikato surroundings for data Analysis (WEKA)[9] is used. It is a comprehensive suite of Java category libraries that implement several algorithms for data processing clump, classification, regression, analysis of results. WEKA are often downloaded from the web site that is shown in figure2.

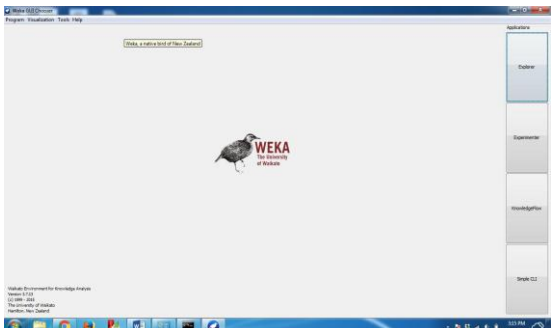


**Fig2: WEKA Tool**

## 4.1 PERFORMANCE MEASURES OF CLASSIFIER

In this experiment the data is supplied to classifier of J48 Algorithm, Random forest tree and Hoeffding Tree to classify the data. The Confusion Matrix is used to evaluate the classifiers performance.

### A. Confusion Matrix

The performance of classifiers is measured using Confusion matrix. In the confusion matrix, sum of the diagonal elements are called correctly classified instances and others are called incorrectly classified instances.

### B. Accuracy

**Accuracy** is defined as the ratio between correctly classified instances and total number of instances in the dataset.

Where correctly classified instances are termed as True positive and True Negative others are called False Positive and False Negative.

$$\text{Accuracy} = \text{TP+TN/TP+TN+FP+FN}$$

### C. Error Rate

The Misclassification error rate is calculated by the following formula:

**Misclassification error rate = 1-Accuracy.**

# 5. RESULTS AND ANALYSIS

There are totally 206 records and 12 attributes in the anemia dataset. With these attributes we made a classification model of the applicants who were affected or not affected for anemia. The model derived from that classification can then predict if a new patient affected by anemia according to his/her attributes. For this the records are classified into 2 classes such as Yes with 108 instances and No with 98 instances. The following figure 3 and 4 represents the output of J48 and Random Forest tree Algorithm.

The following Table 2 represents confusion matrix of J48 Algorithm.

**Table 2: confusion matrix of J48 Algorithm.**

| Target Class | YES | NO |
|---|---|---|
| YES | **105** | 3 |
| NO | 5 | **93** |

In J48 classifier, the correctly identified instances are 198 and incorrectly identified instances are 8**.**

The following Table 3 represents confusion matrix of Random Forest Tree Algorithm.

**Table 3: confusion matrix of RFT Algorithm**

| Target Class | YES | NO |
|---|---|---|
| YES | **108** | 0 |
| NO | 5 | **93** |

In RFT classifier, the correctly identified instances are 201 and incorrectly identified instances are 5**.**

The following Table 4 represents confusion matrix of and hoeffding tree Algorithm.

**Table 4: confusion matrix of Hoeffding tree Algorithm**

| Target Class | YES | NO |
|---|---|---|
| YES | **105** | 3 |
| NO | 12 | **86** |

In hoeffding tree classifier, the correctly identified instances are 191 and incorrectly identified instances are 15**.**

The following table 5 shows the accuracy, time taken to build the model and error rate of J48 and RFT and hoeffding tree Algorithm.

**Table 5: Accuracy, error rate and time taken to build model**

| Classifier | Accuracy | Time Taken (Secs) | Error rate |
|---|---|---|---|
| J48 | 96.11% | 0.02 | 3.88% |
| Random Forest tree | 97.57% | 0.07 | 2.42% |
| Hoeffding tree | 92.71% | 0.07 | 7.28% |

Table 5 shows that the accuracy of J48 (96.11%), accuracy of Random forest (97.57%) and accuracy of hoeffding tree (92.71%). J48 takes 0.02 seconds to build the model, Random forest takes 0.07 seconds and hoeffding tree takes 0.07 seconds to build the model. The error rate of J48 is 3.88% and the error rate of Random Forest tree is 2.42% and the error rate of hoeffding tree is 7.28%. While comparing with other classifiers, Random Forest tree is

giving highest accuracy (97.57%), minimum error rate (2.42%) with time (0.07 seconds) than J48 and Hoeffding tree algorithms.

The following chart1 shows the accuracy, error rate and time taken to build model of classifiers.

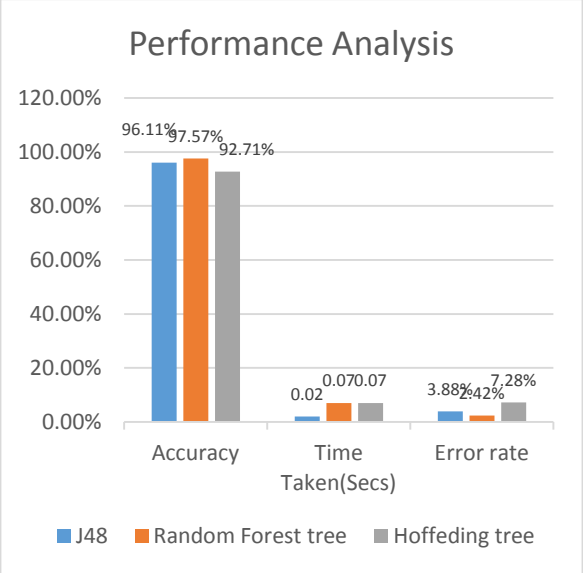


**Chart1: Performance Analysis of classifiers**

In this chart 1, X axis represent the classifiers and Y axis represent the accuracy, error rate and time. It shows that the accuracy of Random Forest Tree is giving highest accuracy (97.57%), minimum error rate (2.42%) and with time (0.07 seconds) than J48 and Hoeffding algorithms.

## 6. CONCLUSION AND FUTURE SCOPE

Diagnosis of disease is an incredibly challenging task in the field of health care. Various data mining techniques have proven to be extremely helpful in decision making. This work analysis set of attributes that is related to the patient CBC test result and helps to improve the standard of prediction by identifying the anemic patients, so that the doctors can help the patients by immediately improving their level of treatments. In our work, we have used data cleaning task to fill up the missing values and we have applied J48, RFT and hoeffding Tree data mining classification techniques which are used to classify the anemia disease. The performances of classifiers are evaluated through the confusion matrix in terms of accuracy and error rate. The experimental result on a sample dataset suggests that Random Forest Tree algorithm provides best performances in terms of accuracy as compared with other two. As a future work the same technique is used to apply for other disease datasets such as heart disease, Lung cancer and so on.

## 7. REFERENCES

[1] Manish Jaiswal, Anima Srivastava, and Tanveer J. Siddiqui, "Machine Learning Algorithms for Anemia diseasePrediction",https://www.researchgate.net/publication/329484705; Chapter · January 2019, DOI: 10.1007/978-981-13-2685-1_44

[2] https://www.who.int/news-room/detail/20-04-2020-who guidance-helps-detect-iron-deficiency-and-protect-brain-development.

[3] Arun, V, et al.: Privacy of Health Information in Telemedicine on Private Cloud, International Journal of Family Medicine & Medical Science Research. (2015)

[4] Gupta, P., Perrine, C., Mei, Z., & Scanlon, K. (2016). Iron, anemia, and iron deficiency anemia among young children in the United States. *Nutrients* ,(6)8, .330.

[5] Jimenez, K., Kulnigg-Dabsch, S., & Gasche, C. (2015). Management of iron deficiency anemia. *Gastroenterology & hepatology*, *11*(4), 241

[6] Reid, S., et al. (2007). "Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer." Nature genetics39(2): 162.

[7] Nemeth, E. and T. Ganz (2014). "Anemia of inflammation." Hematology/Oncology Clinics28(4): 671-681.

[8] Introduction to Data Science By Jeffrey Stanton, © 2012, 2013 and Portions © 2013, By Robert De Graaf

[9] http://www.cs.waikato.ac.nz/ml/WEKA/

[10] https://en.wikipedia.org/wiki/Random_forest

[11] https://builtin.com/data-science/random-forest-algorithm

[12] D. Parameswari, Dr.V. Khanaa, "INTRUSION DETECTION SYSTEM USING MODIFIED J48 DECISION TREE ALGORITHM", Journal of Critical Reviews, ISSN- 2394-5125 Vol 7, Issue 4, 2020 J48 Algorithms of machine learning for predicting user's the acceptance of an E-orientation Systems Rachida IHYA J48 Algorithms of machine learning for predicting user's the acceptance of an E-orientation Systems Rachida IHYA

[13] https://en.wikipedia.org/wiki/Massive_Online_Analysis-Hoeffding Tree

[14] 13-Achebe, M. M. and A. Gafter-Gvili (2017). "How I treat anemia in pregnancy: iron, cobalamin, and folate." Blood129(8): 940-949

[15] https://www.hematology.org/education/patients/anemia

[16] A. and M. Javidroozi (2016). "The patient with anemia." Current opinion in anaesthesiology29(3): 438-445.(Shander and Javidroozi 2016).

[17] https://www.hematology.org/education/patients/anemia