مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

Effective Data Mining Techniques Performance Analysis to Predict Anemia Disease Using Orange Tools

*Yahyia BenYahmed and Kamal Mohamed Elsanoussi[1]
Department of Computer Science, - Faculty of Education Taraghen - Fezzan University- LIBYA

A B S T R A C T

The most well-known type of dietary inadequacy is anemia disease. It is more common in undernourished people and affects humans equally. There has been a worldwide awareness of the use of anemia supplements for people due to this anaemia deficiency. Diagnosing the condition at an earlier stage of life is preferable to prevent further harm and create an appropriate treatment. In this study, anemia is taken into account for early disease prediction and diagnosis by analyzing the data and evaluating the training performance of classifiers in terms of accuracy, precision, sensitivity or recall, specificity, F-measure, Matthews correlation coefficient, and training duration. The volume of data in healthcare organizations is greater. An efficient way to extract knowledge from this kind of data is needed. Data mining is used to discover knowledge from large amounts of data in databases. A classification technique, which is a data mining technique, is used to classify the stages of anemia. The data was collected from 397 households of visitors to the Medical Alabideen Lab in Wadi Etabah, Libya. The research is carried out using Orange software. An experimental study will be conducted with the Anemia data set to determine the best prediction utilizing multiple data mining techniques. As a result, the performance of eight classification techniques is examined, and their training performance is compared using a confusion matrix. It has been determined that the Ensemble classifier outperforms the other techniques in terms of training performance precision.

Keywords: Anemia, Medical Prediction, Data Mining, Classification Technique, Orange Tools.

تحليل أداء تقنيات تنقيب البيانات الفعالة للتنبؤ بمرض فقر الدم باستخدام أدوات أورانج

يحي بن يحمد – كمال محمد السنوسي [*] [1]
[1] قسم الحاسوب – كلية التربية تراغن – جامعة فزان.

[*]Corresponding author:
E-mail yah.benyehmed@fezzanu.edu.ly.Received 7 mares 2023 - Received in revised form 20 june 2023 – Accepted 24 june 2023

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

الملخص

مرض فقر الدم هو النوع الأكثر شهرة من عدم كفاية النظام الغذائي. وهو أكثر شيوعًا عند الأشخاص الذين يعانون من نقص التغذية ويؤثر على البشر على حد سواء. حيث كان هناك وعي عالمي باستخدام مكملات فقر الدم للأشخاص نتيجة لنقص فقر الدم. ومن الأفضل تشخيص الحالة في مرحلة مبكرة من الحياة من أجل منع المزيد من الضرر وخلق العلاج المناسب. في هذه الدراسة، تم أخذ بيانات فقر الدم في الاعتبار للتنبؤ المبكر بالأمراض وتشخيصها من خلال تحليل البيانات وتقييم أداء تدريب المصنفات من حيث الدقة والدقة والحساسية أو الاسترجاع والنوعية وقياسF ومعامل ارتباط ماثيوز ومدة التدريب. حجم البيانات في مؤسسات الرعاية الصحية أكبر. هناك حاجة إلى طريقة فعالة لاستخراج المعرفة من هذا النوع من البيانات. يستخدم التنقيب عن البيانات لاكتشاف المعرفة من كميات كبيرة من البيانات في قواعد البيانات. تُستخدم تقنية التصنيف، وهي تقنية التنقيب عن البيانات، لتصنيف مراحل فقر الدم. تم جمع البيانات من 397 أسرة من زوار مختبر العابدين الطبي في وادي عتبة، ليبيا. تم إجراء البحث باستخدام برنامج Orange، حيث تم إجراء دراسة تجريبية باستخدام مجموعة بيانات فقر الدم لتحديد أفضل تنبؤ باستخدام تقنيات استخراج البيانات المتعددة. نتيجة لذلك، يتم فحص أداء ثماني تقنيات تصنيف، ومقارنة أداء التدريب باستخدام مصفوفة الارتباك Confusion Matrix. لقد تم تحديد أن تقنية مصنف المجموعة Ensemble يتفوق على التقنيات الأخرى من حيث دقة أداء التدريب.

الكلمات المفتاحية: فقر الدم، التنبؤ الطبي، التنقيب في البيانات، تقنية التصنيف، أدوات Orange.

Introduction

The incredible advancements in the health industry have resulted in a huge amount of data being produced in everyday life. This data must be processed to yield valuable information for analysis, prediction, suggestions, and decision-making. Data mining and machine learning techniques are used to convert accessible data into useful information [1, 2]. The key concern for specialists in medical research is disease prediction at the proper moment for prevention and effective treatment plans. In the lack of accuracy, this can sometimes result in death [3].

Anemia is defined as a decrease in red blood cells or hemoglobin levels in the blood. When the onset is gradual, symptoms such as fatigue, weakness, shortness of breath, or diminished capacity to exercise are common. Anemia that occurs more frequently has more severe consequences, such as weariness or increased thirst. Anemia is frequently severe before a person becomes visibly pale. Children may experience growth and developmental issues as a result of iron shortage or anemia. In medical research, the

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

primary concern for clinicians is the prediction of sickness at the appropriate time for prevention and therapy. [4, 5].

In many cases, the lack of clarity can lead to serious consequences. Data mining has enormous promise in the healthcare business, allowing health systems to actively use data and analytics to identify gaps and offer practice standards that maximize treatment and decrease costs. Machine learning techniques are successfully utilized for prediction in a variety of sectors such as healthcare, weather forecasting, stock price prediction, and product recommendation. The fundamental focus of medical science research is the prediction of illness causes and determinants. Healthcare data are utilized in the medical sphere to anticipate epidemics, detect disease, improve quality of life, and avert premature deaths. [6, 7].

There are various types of anemia caused by various factors. The most prevalent type of anemia is caused by a lack of iron and vitamin deficiency anemias in our bodies. This research is carried out by investigating three alternative classification techniques for anemia prediction, such as hemoglobin and low, normal, and high red blood cells. According to the World Health Organization (WHO) Report 2020, anemia sickness affected around 614 million women and 280 million children. Prediction of anemia disease is critical in recognizing other linked disorders. Anemia disorders are classified according to their morphology or root cause. Anemia is categorized into three kinds based on the cause: blood loss, insufficient normal blood production, and excessive blood cell death. [8].

This study focuses on the performance analysis of multiple methods for data mining utilized to a data set to predict anemia disease and compares the performance of classifiers to identify which classifier predicts the disease correctly with high accuracy in terms of the confusion matrix and helps physicians predict the disease and make appropriate decisions using Orange open source Data Mining tools [9, 10]. Several strategies for problem classification have been developed [11]. To create a prediction model, the researchers used Nave Bayes (NB), Random Forest Tree (RFT), Classification Tree (ID3), Multi-layer Perceptron classifier (MLP), CN2 rules, k- Nearest Neighbors (kNN), Support Vector Machine (SVM), and Ensemble (AdaBoost.M1) approaches. These strategies' training performance can be quantified using Accuracy, Precision, Sensitivity or Recall, Specificity, F-measure, Matthews correlation coefficient, and training duration. Nevertheless, nothing is known about how to select categorization techniques in terms of training performance, time, and learning components. As a result, such performance of these strategies must be evaluated.

In this study, we investigate supervised data mining techniques for anemia illness prediction utilizing CBC (complete blood count) data acquired from pathology facilities in this study. The results suggest that AdaBoost.M1 and RFT approaches outperform other strategies in terms of training performance.

However, data mining can produce a knowledge-rich environment that can considerably enhance the caliber of healthcare judgments. Analysis of the performance of several data mining algorithms to identify the most efficient technique for end-user action on hematological data, which is used to forecast the set of relationships between the data and to choose the classifier that accurately and precisely predicts anemia illness. It is

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

crucial to invest in the advancement of computer technology since it will aid doctors in disease prediction and decision-making [12-14]. Data analysis is also helpful for creating automated tools that can support forecast outcomes, recommend additional diagnoses, and suggest treatments [3].

Review of Related work

Data Analysis is a knowledge field that combines domains from statistics and computer science, seeking to find relevant information from databases to facilitate the process of making decisions. Data mining is a logical technique for identifying interesting patterns in enormous amounts of data. It can handle big volumes of data with many properties. It aims to reduce complexity, find patterns in data, and shorten processing times [15]. Classification is a Data Mining task that learns from a set of instances to predict the target class for future cases [2, 16]. To accomplish classification, a variety of machine-learning approaches can be used. There are free and open source Data Mining software packages accessible via the web that can do classification using various techniques, such as KNIME, Orange, RapidMiner, and Weka [17].

In [18], four free and open source Data Mining tools were compared: Weka, RapidMiner, Orange, and KNIME. The study's goal is to identify the most accurate classification tools. Analysts can use the results to get a solid outcome quickly. The experimental results reveal that no single instrument or technique always produces the greatest outcomes, although some produce better results more frequently than others.

Free data mining tools such as Orange can be beneficial for exploratory data analyses and visualization. It gives a platform for varied experiment selection. When it comes to the concepts of innovation, dependability, or quality, orange is extremely effective [19, 20]. Orange Canvas is primarily used for visual programming interfaces. It offers a well-structured overview of several characteristics. These characteristics are Some of Orange's well-known capabilities include data visualization, classification, analysis, unsupervised learning, interaction, and visual analytics via the platform application. CSS, C++, and JavaScript-like languages can be used to accomplish this [10].

There are many Data Mining techniques such as Decision Trees, Naïve Bayes, k-means, and Neural Networks; that are used for analyzing a huge amount of data. Some used data mining techniques are discussed in the following:

Multilayer perceptron (MLP)

An uncomplicated, two-layer neural network called a multilayer perceptron has no hidden layers. In actuality, a two-layer perceptron is sufficient (input layer removed). The power of a multilayer perceptron is comparable to that of a decision tree. The fundamental categorization issue can be accurately predicted by the multilayer perceptron. Logistic regression and learning a multilayer perceptron are closely related. The advantage of multilayer perceptrons is that they may be trained to ignore irrelevant qualities. Multilayer perceptron may compete with more advanced learning techniques on a variety of real-world datasets, according to recent studies [21, 22].

k-Nearest Neighbors (kNN)

It is a straightforward approach to supervised learning in object recognition. Due to its ease of use and effectiveness in the field of machine learning, it is one of the most widely

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

used neighbourhood classifiers [23]. KNN technique explores the pattern space for the k training tuples that are almost equivalent to the unknown tuples. It then saves all cases and categorizes new instances according to similarity measurements. Performance is dependent on the appropriate number of k, which varies depending on the data sample [24].

Support Vector Machine (SVM)

It is a supervised learning technique for categorizing both linear and nonlinear data that is inspired by the concept of statistical learning. By optimizing the margin of hyperplane splitting, SVM divides the data into two categories over a hyperplane simultaneously to prevent over-fitting the data [21, 25].

Naïve Bayes (NB)

It is a probabilistic classifier and one of the most effective techniques for classifying data [3]. It applies the Bayes theorem with robust (naive) independent assumptions. About the class variable, it is presumptive that the feature's value is unrelated to the values of any other features. based on the greatest likelihood. It determines whether the provided tuple belongs to a specific class [4].

Decision tree technique (ID3)

Successful machine learning classification techniques include decision tree techniques. These are supervised learning techniques that make use of gathered and edited information to enhance outcomes. Moreover, decision tree techniques are frequently utilized in research for classification, including in the field of medical and health difficulties. Decision tree techniques come in numerous varieties, including ID3 and C4.5. The most widely used decision tree technique nevertheless, is J48. J48 is an extension of ID3 and the implementation of a better version of C4.5 [21].

Ensemble classifier (AdaBoost*)*

AdaBoost classifier is a boosting-based ensemble learning technique. AdaBoost [26] classifiers are meta-estimators that begin by adapting a classifier on the dataset and then adapt multiple copies of the classifier on the same dataset but with the weights of misclassified instances adjusted so that subsequent classifiers focus more on complicated cases. [27].

CN2 Rule Induction

The CN2 method is a classification system developed for the effective induction of basic, understandable rules of the kind "if *cond* then predict *class*" even in noisy environments. CN2 is a rule induction learning technique.\ It is intended to function even when the training data is flawed. It is inspired by the AQ technique and the ID3 algorithm. As a result, it generates a rule set similar to AQ's but is capable of handling noisy data such as ID3 [28].

Random Forest Tree

Leo Breiman created random forest ensembles of tree predictors in [29], drawing inspiration from prior work by [30]. To compete with boosting, Breiman (1996) developed Random Forests, which are an extension of his bagging theory. A categorical response variable, referred to as "classification" in [29], or a continuous response can both be utilized with random forests.

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

Many studies have been conducted to evaluate the efficacy of data mining classification systems utilizing the Orange by Healthcare forecasting model. The researchers examined the performance of a data mining classification method in Orange in the investigation[20, 31]. Another study [32] compared various classification approaches utilizing various datasets. The research in [33] examined the numerous clustering techniques of "Orange" tools. Furthermore, the performance and assessment of various data mining techniques utilized for the classification of breast cancer have been completed [34]. This is also utilized in artificial intelligence and anticipating abnormalities in the classification of heart illness [35-37]. In the study, data mining classifier techniques were utilized to compare multiple data analyses to establish an automated diagnostic of thalassemia [38] and another for the prediction of heart disease [4]. In addition, numerous data mining clustering techniques for health informatics were examined [39] [18]. Bioinformatics has also made use of data mining tools and numerous classification techniques, which have been contrasted [40, 41]. Data mining techniques were also employed to distinguish patients with typical blood diseases from those with blood cancer [42]. Another study [43] focused on contrasting classification methods for Sickle Cell Disease (SCD) and Chronic Kidney Disease (CKD) utilizing kNN and SVM techniques. These methods included Nave Bayes, Bagging, J48 techniques, Random tree, and CART using Orange [44, 45].

In more recent studies, anemia has been foreseen using different data mining classification algorithms [14, 46, 47] anemia has been predicted using several data mining classification algorithms, with the J48 method confirming its superior performance in categorizing kinds of anemia [48]. Besides, Orange was utilized in this experiment since it can be used to extract hidden predictive information from massive databases [34]. Also, since Orange is frequently used for data mining, the experiment was carried out for CBC (Complete Blood Count), which makes it rather reasonable to extract data using the planned technique [49, 50].

Methodology

The methodology of this study is adopted from [51] which provides a very useful methodology for experimental study. It consists of five steps. These steps are shown in Figure 1

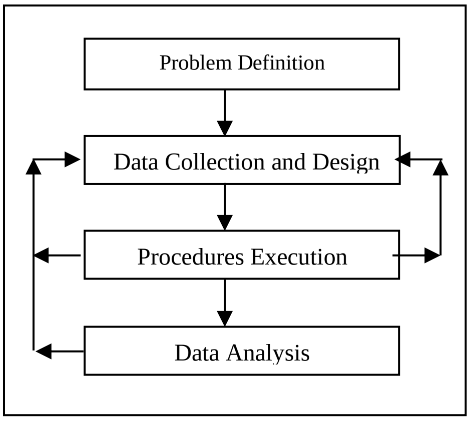

Figure 1. Four Steps to Perform Classification Techniques Experiments

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

1. Problem Definition

Orange machine learning tools techniques and comparison of prediction techniques based on its training performance have been utilized in this study as a tool to analyze and formulate the problem and study objectives during this step. comparing prediction techniques employing machine learning tools from Orange software in terms of accuracy and training time.

2. Data Collection and Design

Orange machine learning software techniques and comparisons of predictive models based on its training performance have been utilized in this study as a tool to analyze and formulate the problem and study objectives during this step. comparing prediction techniques employing machine learning tools from Orange software in terms of accuracy and training time, precision, sensitivity or recall, specificity, F-measure, and Matthews correlation coefficient and training time.

Data was collected from 397 households of visitors to the Medical Alabideen Lab in Wadi Etabah, Libya, using a straightforward pre-coded questionnaire that was created. The information also covers their demographic background and clinical information. The collection contains 397 cases overall. The study selected 14 attributes that would be utilized to categorize the data. Table 1 provides a summary of the Anima dataset's description.

Table 1. Dataset of Anemia Disease

No.	Attribute Name	Attribute Description	Type of Values
1	Age	Age patient	Numeric
2	Sex	Gender patient	Text (Male, Female)
3	HGB	Hemoglobin	Numeric
4	RBC	Red Blood Cell count	Numeric
5	HCT	Hematocrit	Numeric
6	PLT	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	Class label	Text: 1. Normocytic Normochromic (NN) 2. Microcytic Hypochromic (MH) 3. Macrocytic Normochromic (MN)

Figure 2. illustrates the suggested approach for data collecting and design. Pre-

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

processing involves transforming data into an understandable format, and it is a fundamental step in all data mining techniques. In general, real-world data collection is frequently uneven, lacking in some details regarding certain behaviours or patterns, and as a result, prone to several inaccuracies. This preprocessing may be a tried and true method of deconstructing such issues. This preparation method involves several steps to get data ready for any process. These tasks represent the integration, transformation, improvement, and reduction of information. The information is gathered from many guests' households at the Medical Alabideen Lab. The data set contains 3 classes Normocytic Normochromic (NN), Microcytic Hypochromic (MH) and Macrocytic Normochromic (MN) and 397 instances that are described by 14 attributes including the class attribute. 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.

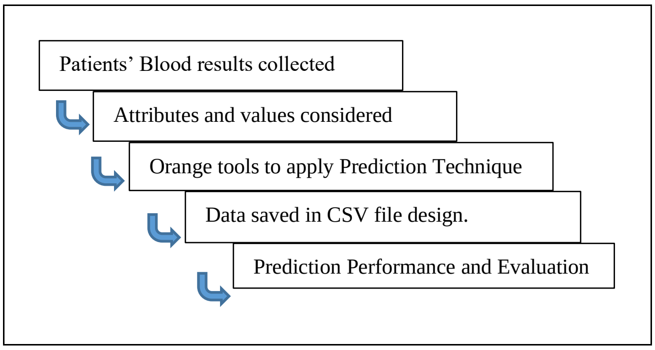

Figure 2: Proposed System Architecture

In the fields of data classification analysis and machine learning, prediction is a fundamental problem. The development of precise classifiers from data has been the focus of ongoing research to address medical issues. To learn a classifier from sets of pre-classified data, various strategies have been suggested. Lazy, Bayes, Statistical Neural Networks, Decision trees, and Support Vector Machines are a few of them. A prediction problem's goal is to create a classifier from a set of examples using a few qualities to characterize the individual objects and one attribute to describe the group of objects. The classifier is then used to either provide us with a better comprehension of the available data or to predict the group attribute of new cases from the domain based on the values of other characteristics.

The data analysis prediction techniques The information was categorized using a method that supported similarity between instances. There are two types of learning: supervised learning and unsupervised learning. The trained and specified data is provided during the supervised learning process. Several categorization strategies have been employed in the study to compare them and determine which is most appropriate for solving medical issues. This study introduced several techniques such as NB, ID3, MLP,

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

CN2, kNN, SVM, and Ensemble (adaBoost.M1) techniques for performance analysis to predict Anemia disease.

Pre-process software was utilized in design selection to allow the explorer to load the data into Orange. The study's anemia data was saved in a CSV file, which was exported from Excel. Following the selection of classifiers used in this study, click the Classifier box, which displays the name of the presently selected classifier and its choices. Following the selection of a classifier, some of them have made adjustments to the choice, for example, the Pick classifier button allows the explorer to select one of the classifiers accessible in Orange and test the classifier on the selected dataset.

The creation of prediction algorithms in an Orange platform application is shown in Figure 3. The application of the selected classifier has been tested in accordance with the settings made by selecting the Test options box. In this study, the classifiers' Cross-validation performance was assessed using the test mode and the number of folds entered in the Folds text field

3. Procedures Execution

The Anemia data was checked on a sequence basis in this stage. Accuracy, Precision, Sensitivity or Recall, Specificity, F-measure, Matthews correlation coefficient, and training duration were used to test the data. The dataset was examined using eight methodologies to determine its accuracy and dependability. With the selection of data and classifier, the selected design will begin to run Orange, with Cross-validation using the number of folds entered in the Folds text field ranging from 3 to 15, as well as to evaluate the effectiveness of classifiers. The classifier output area has scroll bars that allow the explorer to browse through the findings. Also, the confusion matrix displays the number of instances assigned to each class. Elements display the number of test instances whose actual class is represented by the row and whose predicted class is represented by the column. Lastly, once the output has summarized the results, the result can be stored.

4. Data Analysis

The eight techniques are compared in a table to determine the best techniques based on training performance. Based on the data collected with Orange. Orange classifiers are designed to be trained to predict a single 'class' attribute represented in the data's last attribute, which is the objective for prediction. But, to enable the classifier to learn, the attribute of the data may be adjusted and modified under the filter option. The best techniques are determined by the outcomes of this analysis.

Experiments with Orange Tools

The WebLab97 at scenic Bled, Slovenia environs for Data Analysis (Orange) [9] is used in this study. It was created at the University of Ljubljana in Switzerland. The name Workshop represented Michie's vision that the tool should be a control of powerful C++ category libraries that implement various methods for data presentation, machine learning, data analysis, and data science. Orange is frequently obtained from the website depicted in Figure 3.

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

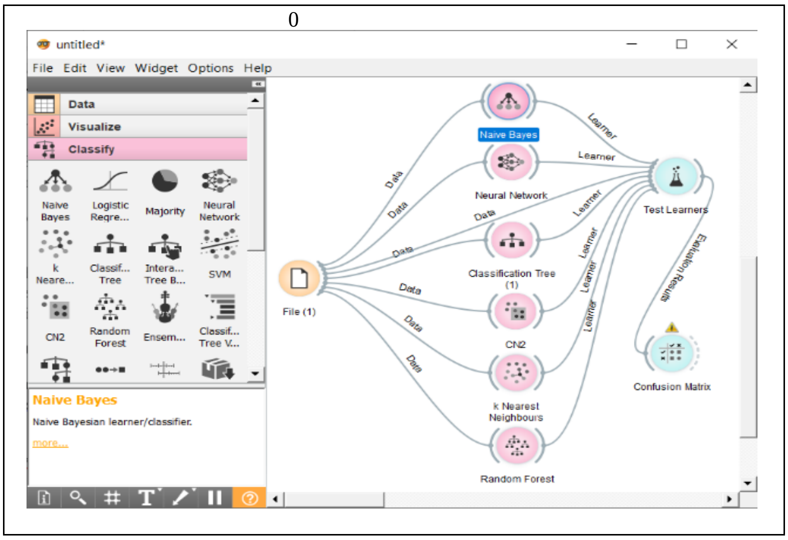

Figure 3: Orange Tools

Performance measures of classifier

In this experiment, anemia data is fed into a classifier that uses NB, RFT, ID3, MLP, CN2, kNN, SVM, and adaBoost.M1 algorithms to classify the data. The Confusion Matrix is used to assess the performance of the classifiers. Researchers utilize a range of performance criteria to evaluate the effectiveness of machine learning algorithms. The Confusion matrix is used to assess classifier performance. In the confusion matrix, the total of the diagonal elements is referred to as correctly categorized cases, while the remainder is referred to as wrongly classified instances. We employed the Accuracy (Acc), Precision (Prec), Sensitivity or Recall (Sens), Specificity (Spec), F-measure, Matthews correlation coefficient (MCC), and training duration performance indicators to assess and compare the performance of presented prediction models.

Accuracy is calculated by dividing the number of test records by the number of successfully classified records. The percentage of True Positive (TP) records to the total number of True Positive (TP) records in a certain class is called precision. There are two types of recall: true positives and false negatives. The total number of records properly categorized to the total number of records in a class is known as the recall ratio (FN). The most important and commonly used factor to measure the performance of the classifier is accuracy. Accuracy (Acc) is calculated by the ratio of correct prediction samples to the total samples in the dataset. Equation (1) defined Acc as:

By dividing the total number of test records by the total number of successfully categorized records, accuracy is calculated. Precision is defined as the ratio of True Positive (TP) records to all True Positive (TP) records in a given class. The recall comes in two flavours: true positives and false negatives. The recall ratio is the proportion of records that were correctly classified to all the records in a class (FN). Accuracy is the most significant and frequently used factor to assess the performance of the classifier. The ratio of valid prediction samples to all samples in the dataset is used to compute accuracy (Acc). According to Equation (1), ACC is:

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

However, error rate (ERR), which is determined as follows, represents the number of samples that were incorrectly classified into the positive and negative classes.

$$Error\ rate(ERR)\ =\ (1\ -\ Acc)\ \times\ 100\% \quad (2)$$

The following formulas were used to compute the precision:

$$Precision\ (Prec) = \frac{TP}{TP + FP} \quad (3)$$

The ratio of real projected positive samples to all positive samples is known as sensitivity (Sens) or recall. However, the ratio of genuine projected negative samples to all negative samples is known as specificity (Spec) or selectivity. Equations (4) and (5), respectively, stand for Sens and Spec.

$$Sensitivity\ or\ Recall\ (Sens) = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity\ (Spec) = \frac{TN}{TN + FP} \quad (5)$$

The cyclical mean between recall and precision was defined by the F-measure. A model is deemed successful if its value is one, while a value of 0 indicates ineffective performance. The following is the F-measure equation.

$$F - measure = \frac{2TP}{2TP + FP + FN} \quad (6)$$

The link between observed and expected categorization is depicted by the Matthews correlation coefficient (Mcc). The confusion matrix is used to calculate Mcc, and a value of +1 indicates perfect prediction while a value of 1 indicates the correlation between predicted and actual values.

The Mcc Equation (7) is shown below.

$$\mathrm{Mcc} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP}\ \times \mathrm{FN}}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (7)$$

Results and Analysis

The anemia dataset consists of 397 items overall and 14 characteristics. We created a classification model of the applicants using these characteristics to determine whether or not they had anemia. The model created from that classification can then determine based on a new patient's characteristics if they are affected by anemia. The records are divided into three categories for this purpose: MN, with 27 instances, MH, having 187 cases, and NN, with 183 instances.

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

1. Prediction Models

To choose the best classifier, the experiment of training and learning duration was based on a cross-validation fold that began with 3, 4, 5, 6, 7, 8, and 15. With 11, the cross-validation fold produced the greatest results. The D data set was used to test the measurements of Acc, Prec, Sens, Spec, F-measure, and Mcc as well as training time. Fig. 4 and Table 5 show the experiment's outcomes in detail.

To categorize the anemia data, the classifier uses the NB, RFT, ID3, MLP, CN2, kNN, SVM, and Ensemble (adaBoost.M1) algorithms. The Confusion Matrix is used to assess the effectiveness of the classifiers.

2. Confusion Matrix

Using the Confusion matrix, the classifier is assessed for effectiveness. The confusion matrix refers to instances that are correctly classified as the sum of the diagonal elements and occurrences that are wrongly classified as the other elements. the subsequent Figure 4. The output of the NB, RFT, ID3, CN2, MLP, kNN, SVM, and adaBoost.M1 algorithms was depicted in Figure 4.

The adaBoost.M1 classifier produced the greatest results when compared to other methods in the experiment with the Confusion Matrix, correctly identifying 393 occurrences and mistakenly identifying 4. The second-best result was scored using the RFT and CN2 rules techniques, with 390 cases properly detected and 7 incorrectly identified. The fourth-best result in terms of classification accuracy for the NB classifier is

384 correctly identified cases and 13 wrongly identified instances. Confusion Matrix performance can be compared to other training performance measures as Sens, Spec, or F-measure, as shown in Table 2.

Contents
Learner: Naive Bayes
Data: Number of examples
Matrix

	MN	MH	NN	
MN	27	0	0	27
MH	0	184	3	187
NN	4	6	173	183
	31	190	176	397

Contents
Learner: Random Forest
Data: Number of examples
Matrix

	MN	MH	NN	
MN	26	0	1	27
MH	0	186	1	187
NN	0	5	178	183
	26	191	180	397

Contents
Learner: CN2 rules
Data: Number of examples
Matrix

	MN	MH	NN	
MN	26	0	1	27
MH	0	187	0	187
NN	0	6	177	183
	26	193	178	397

Contents
Learner: SVM
Data: Number of examples
Matrix

	MN	MH	NN	
MN	25	0	2	27
MH	1	185	1	187
NN	2	6	175	183
	28	191	178	397

Contents
Learner: Classification Tree
Data: Number of examples
Matrix

	MN	MH	NN	
MN	0	0	27	27
MH	0	187	0	187
NN	0	9	174	183
	0	196	201	397

Contents
Learner: kNN
Data: Number of examples
Matrix

	MN	MH	NN	
MN	22	1	4	27
MH	0	182	5	187
NN	3	10	170	183
	25	193	179	397

Contents
Learner: Neural Network
Data: Number of examples
Matrix

	MN	MH	NN	
MN	24	0	3	27
MH	2	184	1	187
NN	1	7	175	183
	27	191	179	397

Contents
Learner: AdaBoost.M1
Data: Number of examples
Matrix

	MN	MH	NN	
MN	26	0	1	27
MH	0	187	0	187
NN	0	3	180	183
	26	190	181	397

Notes: columns represent prediction, row represent true classes

Figure 4: Confusion matrix of classification technique based on Cross-validation method by folds (11)

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

3. Prediction Performance

Table 2. and Figure 5. shown the performance of the prediction models by considering all features. From the table and graph, we can see that the classification accuracies of the NB, RFT, ID3, MLP, CN2, kNN, SVM and Ensemble (adaBoost.M1) techniques-based prediction models with all features and were 98.99%, 98.23%, 98.23%, 96.99%, 96.72%, 96.47%, 94.21%, and 90.93% respectively. Table 2. indicated the results of the prediction models with all features.

Table 2: Test learners, prediction performances with training time taken to build mode bases on Cross-validation by 11

Validation method							
Method: Random sampling							
Method: Cross-validation							
Folds: 11							
Target class: Class							
Data							
Examples: 397 Cases							
Attributes: 13 (patientage, patientsex, hgb, rbc, hct, plt, mcv, mch, mchc, wbc, neut, lymph, mono)							
Class: Class							
Results							
Technique classifier	**Acc**	**Sens**	**Spec**	**Mcc**	**F1**	**Prec**	**Time (sc)**
Naive Bayes (NB)	0.967	1.000	0.989	0.928	0.931	0.871	0.056
Random Forest Tree (RFT)	**0.982**	**0.963**	**1.000**	**0.980**	**0.981**	**1.000**	**0.050**
Classification Tree (ID3)	0.909	0.000	1.000	0.000	0.000	0.000	0.161
AdaBoost.M1 (Ensemble)	**0.989**	**0.963**	**1.000**	**0.980**	**0.981**	**1.000**	**0.093**
Neural Network (MLP)	0.964	0.888	0.991	0.880	0.888	0.888	0.065
Support Vector Machine (SVM)	0.969	0.925	0.991	0.902	0.909	0.892	0.070
CN2 rules	**0.982**	**0.963**	**1.000**	**0.980**	**0.981**	**1.000**	**0.030**
k- Nearest Neighbors (kNN)	0.942	0.814	0.991	0.836	0.846	0.880	0.090

The performance of various classifiers for the anemic disease is broken down in Table 2 in detail. Based on the output confusion matrix, various performance evaluation measures including accuracy, recall, fall-out, and error rate were used for this comparison study. Table 2 makes it clear that AdaBoost.M1 (Ensemble) accuracy produces more accurate findings. Using the Euclidean Distance function, Random Forest also performed well and reached 98.2% accuracy. CN2 dominates the Random Forest Tree (RFT) by 98.2%. With an accuracy of 96.9% and 96.7%, respectively, SVM and Naive Bayes (NB) both achieved satisfactory performance. The decision tree had a five-level depth and the lowest accuracy of all the classifiers at 74.19%.

Additionally, Table 2 demonstrates the classification accuracy of Naive Bayes (96.7%), Random Forest (98.2%), Classification Tree (90.93%), AdaBoost.M1 (Ensemble) (98.9%), Neural Network (96.4%), SVM (96.9%), CN2 rules (98.2%), and kNN (94.2%). The best prediction result is in Ensemble (AdaBoost.M1) prediction, with time training building requiring 0.05 seconds. The Sens, Spec, Prec, F-measure, and Mcc

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

classifiers produce identical results (96.3%, 1.0%, 1.0%, 98.1% and 98.00%) in the first tree result approaches, which are AdaBoost.M1 (Ensemble), Random Forest Tree (RFT), and Support Vector Machine (SVM), respectively

Furthermore, Table 2 shows that the classification accuracy of Naive Bayes (96.7%), accuracy of Random Forest (98.2%), accuracy of Classification Tree (90.9%), and accuracy of AdaBoost.M1 (Ensemble) (98.9%), accuracy of Neural Network (96.47%), accuracy SVM (96.9%), accuracy of CN2 rules (98.2%) and accuracy of kNN (94.2%). The best result shows in Ensemble (AdaBoost.M1) prediction with time training building taking 0.05 seconds. The Sens, Spec, Prec, F-measure, and Mcc record the similar result (96.3%, 1.0%, 1.0%, 98.1% and 98.0%) in the first tree result techniques which are AdaBoost.M1 (Ensemble), Random Forest Tree (RFT) and Support Vector Machine (SVM) classifiers respectively. However, the Ensemble technique by AdaBoost.M1 method provides the highest classification accuracy (98.9%), the lowest error rate of 0.01 with time is 0.09 seconds when compared to other classifiers, while Classification Tree (ID3) and k- Nearest Neighbors (kNN) provided the lowest accuracy by 94.2% and 94.2%, respectively. Figure 5. depicts the accuracy, error rate, and time required to develop a classifier model.

Table 2. shows that the RFT classifier, when employed on Anemia datasets, has the shortest training time. Yet, when compared to other classifiers, RFT performs the best in terms of Sensitivity.

RFT has the shortest training time among the four classifiers that can handle the Anemia dataset used in this work, as shown in Table 2. The ID3 and AdaBoost.M1 (Ensemble) methods take more learning time than the other algorithms by 0.161sc and 0.093sc, respectively. RBF, on the other hand, has the largest complexity of all classifiers.

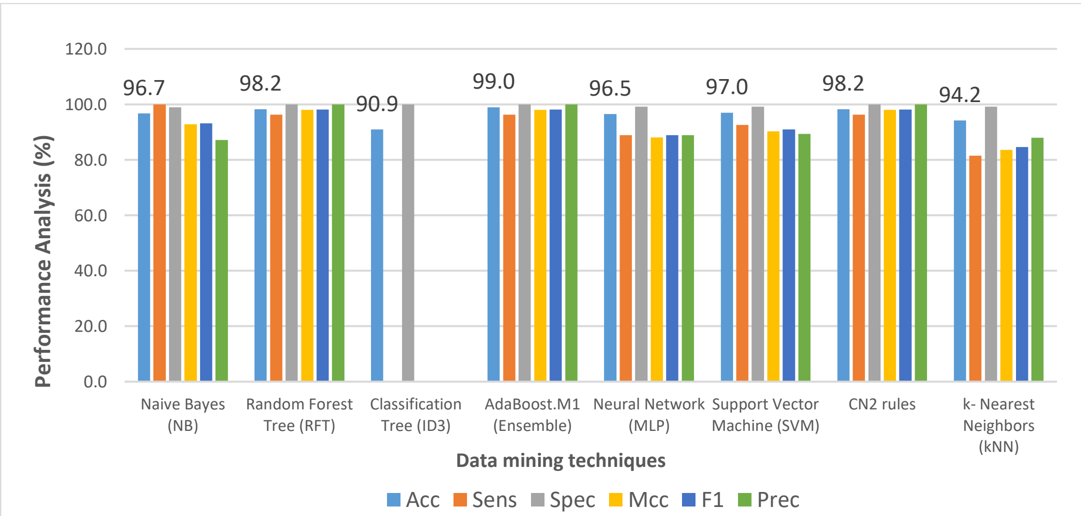

Figure 5: Comparison of classifier algorithms based on performance

The classifiers were represented by the X axis in Figure 5., and the training performances and training time were represented by the Y axis. It demonstrates that

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

AdaBoost.M1 (Ensemble) provides the highest accuracy (98.9%), the lowest error rate (0.010), and the shortest duration (0.05 seconds) when compared to other techniques.

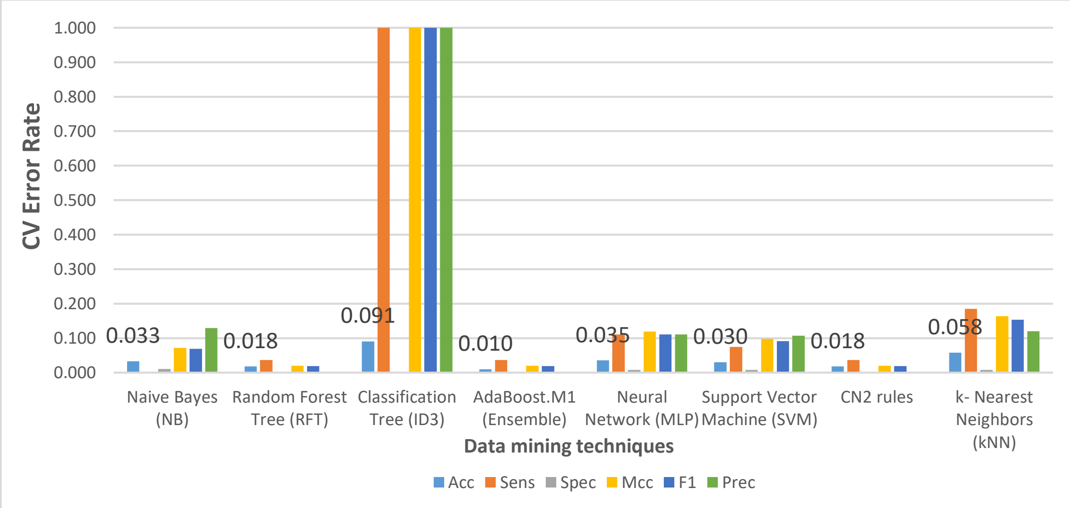

Figure 6: Performance of Cross Validation Error Rate

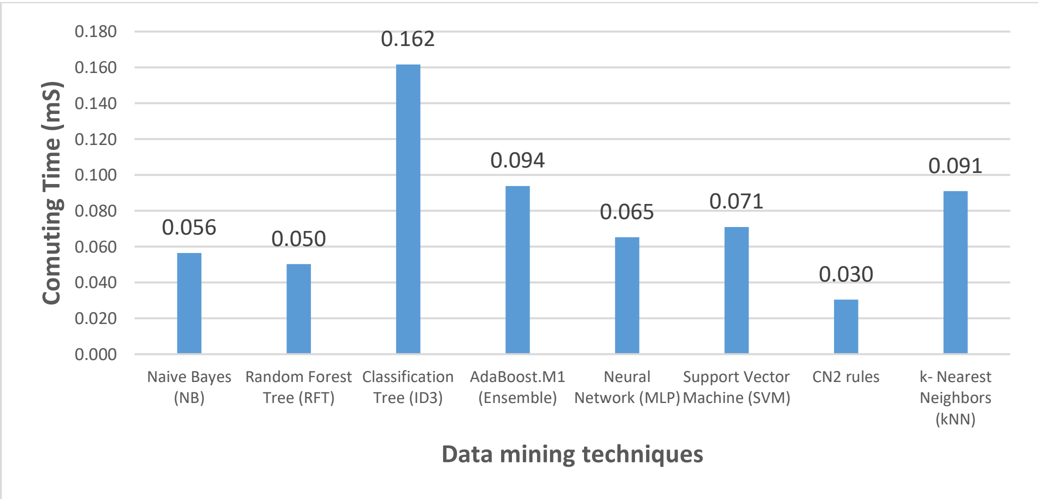

Figure 7: Performance of Computing Time

The figures are composed of various classification values. The accuracy was computed based on these numbers. Figures (5-7) illustrate the final values for the data set described above using classification techniques for data mining, with the best accuracy and least computability. The graph suggests comparing based on performance, compute time, accuracy value, error rate (11- fold Cross Validation), and lastly greatest accuracy and lowest computing time. When compared to other techniques, the AdaBoost.M1 technique (Ensemble) performs better.

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

Conclusion and Future Work

Disease diagnosis is an extremely difficult undertaking in the realm of health care. Different data mining strategies are incredibly beneficial in decision-making. This work analyzes a set of factors associated with the patient's CBC test result and helps to enhance the standard of prediction by recognizing anemic patients so that doctors can aid the patients by quickly improving their treatment level. In this study, we used data cleaning tasks to fill in the missing values and applied data mining prediction techniques such as Nave Bayes, Random Forest Tree, Classification Tree, Multi-layer Perceptron, CN2 rules, k-Nearest Neighbors, Support Vector Machine, and Ensemble to predict the Anemia disease By using the confusion matrix, the prediction's performances are assessed in terms of classification accuracy, precision, sensitivity or recall, specificity, f-measure, Matthews correlation coefficient, and training duration. The experimental outcome on a test dataset indicates that, when compared to the other seven ways, the Ensemble technique, which uses the AdaBoost.M1 method, performs best in terms of accuracy by 98.99%.

The same method will be utilized in future work to use feature selection techniques to choose the most appropriate characteristics that might affect classification accuracy, training duration, and other methods with other diseases including kidney disease, lung cancer, heart disease, and so forth.

References

[1]J. Wang, Y. Yang, T. Wang, R. S. Sherratt, and J. Zhang, "Big data service architecture: a survey," Journal of Internet Technology, vol. 21, pp. 393-405, 2020.
[2]A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," Expert Systems with Applications, vol. 166, p. 114060, 2021.
[3]M. Jaiswal, A. Srivastava, and T. J. Siddiqui, "Machine learning algorithms for anemia disease prediction," in Recent Trends in Communication, Computing, and Electronics: Select Proceedings of IC3E 2018, 2019, pp. 463-469.
[4]E. El-kenawy, "A Machine Learning Model for Hemoglobin Estimation and Anemia Classification," International Journal of Computer Science and Information Security (IJCSIS), vol. 17, pp. 100-108, 2019.
[5]P. Bhadra and A. Deb, "A review on nutritional anemia," Indian Journal of Natural Sciences, vol. 10, pp. 18466-18474, 2020.
[6]J. Rössler, J. Breckwoldt, and D. R. Spahn, "Management of Anemia," in Senior Trauma Patients: An Integrated Approach, ed: Springer, 2022, pp. 233-236.
[7]V. Arun, V. Shyam, and S. Padma, "Privacy of health information in telemedicine on private cloud," Internation journal on Family Medicine and Medical Science Research, vol. 4, p. 2, 2015.
[8]W. H. Organization, "WHO Guidance Helps Detect Iron Deficiency and Protect Brain Development," World Health Organization, 2020.
[9]J. Demšar, B. Zupan, G. Leban, and T. Curk, "Orange: From experimental machine learning to interactive data mining," in Knowledge Discovery in Databases: PKDD 2004: 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, September 20-24, 2004. Proceedings 8, 2004, pp. 537-539.

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

[10] R. Ratra and P. Gulia, "Experimental evaluation of open source data mining tools (WEKA and Orange)," International Journal of Engineering Trends and Technology, vol. 68, pp. 30-35, 2020.
[11] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: classification and comparison," International Journal of Computer Trends and Technology (IJCTT), vol. 48, pp. 128-138, 2017.
[12] K. Lakshmi, M. V. Krishna, and S. P. Kumar, "Performance comparison of data mining techniques for predicting of heart disease survivability," International Journal of Scientific and Research Publications, vol. 3, pp. 1-10, 2013.
[13] F. Akter, M. A. Hossin, G. M. Daiyan, and M. M. Hossain, "Classification of hematological data using data mining technique to predict diseases," Journal of Computer and Communications, vol. 6, p. 76, 2018.
[14] P. Verma and V. Chopra, "A Review on Machine Learning Algorithms for Anemia disease Prediction," 2022.
[15] R. R. Asaad and R. M. Abdulhakim, "The Concept of Data Mining and Knowledge Extraction Techniques," Qubahan Academic Journal, vol. 1, pp. 17-20, 2021.
[16] R. J. Roiger, Data mining: a tutorial-based primer: Chapman and Hall/CRC, 2017.
[17] S. Slater, S. Joksimović, V. Kovanovic, R. S. Baker, and D. Gasevic, "Tools for educational data mining: A review," Journal of Educational and Behavioral Statistics, vol. 42, pp. 85-106, 2017.
[18] L. C. Borges, V. M. Marques, and J. Bernardino, "Comparison of data mining techniques and tools for data classification," in Proceedings of the International C* Conference on Computer Science and Software Engineering, 2013, pp. 113-116.
[19] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, et al., "Orange: data mining toolbox in Python," the Journal of machine Learning research, vol. 14, pp. 2349-2353, 2013.
[20] J. Demšar and B. Zupan, "Orange: Data mining fruitful and fun-a historical perspective," Informatica, vol. 37, 2013.
[21] S. Agarwal, "Data mining: Data mining concepts and techniques," in 2013 international conference on machine intelligence and research advancement, 2013, pp. 203-207.
[22] E. Frank and M. A. Hall, Data mining: practical machine learning tools and techniques: Morgan Kaufmann, 2011.
[23] B. Deekshatulu and P. Chandra, "Classification of heart disease using k-nearest neighbor and genetic algorithm," Procedia technology, vol. 10, pp. 85-94, 2013.
[24] P. Iswarya and V. Radha, "Ensemble learning approach in improved k nearest neighbor algorithm for text categorization," in 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015, pp. 1-5.
[25] R. Srinivas, "Managing Large Data Sets Using Support Vector Machines," 2010.
[26] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in Computational Learning Theory: Second

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

European Conference, EuroCOLT'95 Barcelona, Spain, March 13–15, 1995 Proceedings 2, 1995, pp. 23-37.
[27] T.-K. An and M.-H. Kim, "A new diverse AdaBoost classifier," in 2010 International conference on artificial intelligence and computational intelligence, 2010, pp. 359-363.
[28] P. Clark and T. Niblett, "The CN2 induction algorithm," Machine learning, vol. 3, pp. 261-283, 1989.
[29] L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5-32, 2001.
[30] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," Neural computation, vol. 9, pp. 1545-1588, 1997.
[31] D. Umer, "Analysis of Heart Patients Disease Using Data Mining Tool Orange," vol, vol. 9, pp. 1146-1150, 2020.
[32] A. Ishak, K. Siregar, R. Ginting, and M. Afif, "Orange software usage in data mining classification method on the dataset lenses," in IOP Conference Series: Materials Science and Engineering, 2020, p. 012113.
[33] S. Kodati, K. P. Reddy, G. Ravi, and N. Sreekanth, "Analysis of Heart Disease Data Using K-Means Clustering Algorithm in Orange Tool," in Intelligent Manufacturing and Energy Sustainability: Proceedings of ICIMES 2020, 2021, pp. 417-423.
[34] G. Amala, "Orange tool approach for comparative analysis of supervised learning algorithm in classification mining," Journal of Analysis and Computation, vol. 13, pp. 1-10, 2019.
[35] I. Tougui, A. Jilbab, and J. El Mhamdi, "Heart disease classification using data mining tools and machine learning techniques," Health and Technology, vol. 10, pp. 1137-1144, 2020.
[36] S. Kodati, "Analysis of heart disease using in data mining tools Orange and Weka," Global journal of computer science and technology, vol. 18, pp. 17-21, 2018.
[37] Z. R. Mohi, "Orange Data Mining as a tool to compare Classification Algorithms," Dijlah Journal of Sciences and Engineering, vol. 3, pp. 13-23, 2020.
[38] A. K. Ramotra, A. Mahajan, R. Kumar, and V. Mansotra, "Comparative analysis of data mining classification techniques for prediction of heart disease using the Weka and SPSS modeler tools," in Smart Trends in Computing and Communications: Proceedings of SmartCom 2019, 2020, pp. 89-97.
[39] P. Saxena and S. Lehri, "Analysis of various clustering algorithms of data mining on health informatics," International Journal of Computer & Communication Technology, vol. 4, pp. 108-112, 2013.
[40] S. K. David, A. Saeb, and K. Al Rubeaan, "Comparative analysis of data mining tools and classification techniques using weka in medical bioinformatics," Computer Engineering and Intelligent Systems, vol. 4, pp. 28-38, 2013.
[41] A. Naik and L. Samant, "Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer, Tanagra, Orange and Knime," Procedia Computer Science, vol. 85, pp. 662-668, 2016.
[42] N. B. Tayfor and S. J. Mohammed, "A comparison study of data mining algorithms for blood cancer prediction," Passer Journal of Basic and Applied Sciences, vol. 3, pp. 174-179, 2021.

مجلة جامعة فزان العلمية
Fezzan University Scientific Journal
Journal homepage: wwwhttps://fezzanu.edu.ly/

[43] A. Sharma and D. Kaladhar, "Phylogeny, Predicton and Analysis of Sickle Cell Anaemia Using Data Mining Tools from Chhattisgarh State, India," EasyChair 2516-2314, 2021.
[44] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," in 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), 2016, pp. 300-305.
[45] P. Sinha and P. Sinha, "Comparative study of chronic kidney disease prediction using KNN and SVM," International Journal of Engineering Research and Technology, vol. 4, pp. 608-12, 2015.
[46] 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.
[47] K. Meena, D. K. Tayal, V. Gupta, and A. Fatima, "Using classification techniques for statistical analysis of Anemia," Artificial Intelligence in Medicine, vol. 94, pp. 138-152, 2019.
[48] V. Geetha, C. Gomathy, K. Keerthi, and N. Pavithra, "Diagnostic Approach To Anemia In Adults Using Machine Learning," Journal of Pharmaceutical Negative Results, pp. 3713-3717, 2022.
[49] P. Appiahene, J. W. Asare, E. T. Donkoh, G. Dimauro, and R. Maglietta, "Detection of iron deficiency anemia by medical images: a comparative study of machine learning algorithms," BioData Mining, vol. 16, pp. 1-20, 2023.
[50] M. Almaliki and E. Atlam, "Complete Blood Analysis: An Android OCR-Based Interpretation," in Intelligent Computing: Proceedings of the 2022 Computing Conference, Volume 2, 2022, pp. 278-293.
[51] F.-Y. Lo, A. Rey-Martí, and D. Botella-Carrubi, "Research methods in business: Quantitative and qualitative comparative analysis," vol. 115, ed: Elsevier, 2020, pp. 221-224.