Kafrelsheikh Journal of Information Sciences ISSN (Online): 2535-1478, ISSN (Print): 2537-0677 Volume 4, Issue 2, 2023, PP. 1–7
Date of publication 25 November 2023

The classification of mushroom using ML

Rahma Nabil Said Ahmed Sallam[1], Eman A.Shehab [2] Sara A.Shehab [3]
[1] Student, Bioinformatics Dep, Faculty of Computer &Artificial Intelligence, University of Sadat City.
[2] Faculty of Computer and Artificial Intelligence, University of Sadat City
[3] Faculty of Computer and Artificial Intelligence, University of Sadat City

Article Info

Keywords:
Mushroom.
fungi
Classification
Machine Learning.
Random Forest.

Abstract

The Mushroom is kind of fungi. Major health benefits of mushrooms include their ability to kill cancer cells. The goal of this research is to determine the most effective method for mushroom classification, with the categories of deadly and nonpoisonous mushrooms being used. Separate from plants and animals, they belong in their own realm. In terms of how they get nutrients, fungi are different from plants and mammals. Mushrooms are classified as edible and poisoned. To distinguish between two varieties of mushrooms, we can use machine learning, which is used in classification. There are numerous machine learning algorithms that perform classification, but in our model, I utilize random forest, MLP, Linear Regression and decision tree on the features of the mushroom to categorize it into edible and poisonous. Random Forest achieves high accuracy 98.70%. from these results, we can use Ml to differentiate between two varieties of mushrooms because it is used in classification efficiently.

1. Introduction

member of the fungus kingdom, mushrooms don't need sunlight to produce energy like plants do. Humans use some varieties of mushrooms some of them are edible, while others are poisonous and can result in serious disease or even death if consumed. To determine if these mushrooms are poisonous or edible, we must classify them.

Mushrooms can be poisonous due to the presence of toxins such as amatoxins, gyromitrin, muscarine, ibotenic acid, and psilocybin. These toxins can cause a range of symptoms from mild gastrointestinal distress to organ failure and death. Some mushrooms may also accumulate heavy metals or other harmful substances from their environment, making them unsafe for consumption. It is important to properly identify mushrooms before consuming them and to only eat those that are known to be safe.

1.1. About machine learning

"Machine learning" enables computer systems to automatically learn from their past performance and advancement. It entails analyzing data using algorithms and statistical models to spot patterns that can be used to make predictions or judgements to decrease anemia's prevalence,

Machine learning can be used to classify data about mushrooms by examining their shape, color, texture, and odour, among other attributes. A machine learning model can learn to recognize patterns in the data that distinguish between the two types by being trained on a dataset of labelled mushrooms (i.e., mushrooms that have been classified as either edible or toxic).

Once trained, the model may be used to categorize new mushrooms according to their traits. To make sure that only safe mushrooms are sold in markets or to identify potentially toxic mushrooms in the wild, this can be especially helpful.

1.2. applying ML for mushroom classification

Overall, machine learning is a potent tool for deciphering intricate datasets and producing precise predictions based on the data pattern. In our model I used algorithms from Ml and NN like logistic regression, decision tree, random forest. and, MlP comparing them and get the highest accuracy of these algorithm.by comparing all these algorithms I get the high accuracy by random forest algorithm with accuracy 0.9870994590095714 this is highest accuracy compared to others in the same data about the dataset.

2. Related Work

Although it contains deeper neural nets, deep learning (DL) is a descendant of artificial neural networks.

Through its possible applications, DL is a growing new current method that is succeeding in all disciplines. The DL model incorporates more real-world system capabilities, which aids in creating higher trained models for classification accuracy. According to the survey, many studies use convolution neural networks using ANNs with various convolution layer levels for image-related applications to handle classification, prediction, or identification problems.
CNN is used in to predict the detection of thirteen different plant species' leaf diseases. CNN is frequently used in the agriculture sector for a variety of applications.
GoogleNet is employed to detect plant diseases in 14 types of crops in. Although it contains deeper neural nets, deep learning (DL) is a descendant of artificial neural networks. Through its possible applications, DL is a growing new current method that is succeeding in all disciplines. The DL model incorporates more real-world system capabilities, which aids in creating higher trained models for classification accuracy. According to the survey, many studies use convolution neural networks using ANNs with various convolution layer levels for image-related applications to handle classification, prediction, or identification problems. CNN is used to predict the detection of thirteen different plant species' leaf diseases. CNN is frequently used in the agriculture sector for a variety of applications. GoogleNet employed to detect plant diseases in 14 types of crops in Because of its commercial worth, mushrooms. benefits to commerce and nutrition. Among these 7000, 0.14 million are mushrooms worldwide.
are edible, while the remaining 14,000 are difficult to determine if they are harmful or edible. Numerous research papers use supervised, semi-supervised, or unsupervised machine learning techniques to solve classification challenges. Multi-layer preceptor (MLP) and base radical network (BRF) are used in the classification of mushrooms, and MLP produces better results than BRF. A forecasting tool built on MLP is shown in. Mushrooms are harvested while a robot vision system is used to check for damage using a support vector machine (SVM). Different combinations of variables are analysed using the UCI repository for clustering with K-modes to classify mushrooms as poisonous or edible. The Kaggle dataset is used, but the attributes considered are unclear and few, hence the accuracy is few. As segmentation scales up and minimises the negative error impacts dramatically at large scales, classification accuracy will rise with the use of images.
A web application, unified database, and mobile phone application are all included in the mushroom diagnosis assistance system (MDAS) described in. For the classification of mushrooms, a comparison between the decision tree (DT) algorithm and naive Bayes algorithm is conducted. When SVM and Navie Bayes algorithms are compared for the classification of mushrooms in, SVM produces the best results out of the simulated strategies. When the categorization of mushrooms using three algorithms—decision tree method, SVM, and ANN with clustering—is examined in, ANN outperforms the others.

3. About the dataset

This section's introduction provides details about the dataset's characteristics and uses them to describe it. This study's dataset was obtained via Kaggle. [21]. Sources: (a) The Audubon Society Field Guide to North American Mushrooms (1981), which served as the basis for the mushroom data. Jeff Schlimmer, Donor (Jeffrey.Schlimmer@a.gp.cs.cmu.edu), G. H. Lincoff (President), New York: Alfred A. Knopf (c) Time: April 27, 1987 .
On pages 500–525 of this data set, 23 species of gilled mushrooms belonging to the Agaricus and Lepiota Family are described as hypothetical samples. Each species is classified as either unquestionably edible, unquestionably poisonous, or maybe edible but not advised. The toxic class was joined with the latter. The guide makes it obvious that there is no easy rule—like "leaflets three, let it be" for poisonous oak and ivy—to determine if a fungus is edible.

4. Proposed Methodology

The proposed methodology is divided into three phases: I) Data pre-processing. II)Classifying using different ML algorithms. III) Measure performance. In this section is a summary of these phases. See figure 1.

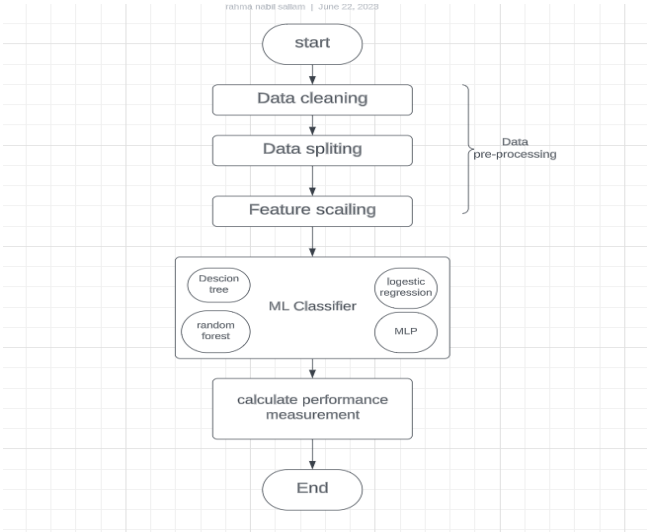


Figure 1. Proposed Methodology

4.1 Data Preprocessing

There are correlations between data because some attributes like veil- type, veil-color, ring-number, and gill-attachment. veil-type was constant this make data correlation. correlation make overfitting, then check data duplicated after this I check data by data profiling to make sure there are no correlation between data. also applying data scaling. I didn't apply feature selection.

4.2 Data Cleaning

The duplicate data and the missing data are examined at this stage of the pre-processing procedures. There are no messing data as shown in figure 2.

Dataset statistics	
Number of variables	23
Number of observations	8124
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.4 MiB
Average record size in memory	184.0 B

Variable types	
Categorical	22
Boolean	1

Figure 2: Dataset statistics

4.3 Data Splitting

A machine learning model is overfit when it reliably fits its training data too well but cannot fit new data. Data splitting is commonly used in machine learning to prevent overfitting. The initial data is frequently divided into two or three groups by a machine learning model. The training set, the validation set, and the testing set are the three sets that are typically used. The training set refers to the set of data that was utilized to train the model. The model must observe and learn from the training set in order to improve any one of its parameters. The validation set is a data set of examples used to change the learning process's settings. This data set's goal is to rate the model's accuracy to help choose the best one. The testing set is the data set that is compared to the prior data sets and tested in the final model. The testing set is used to evaluate the algorithm and mode that has been selected. (See figure 3)

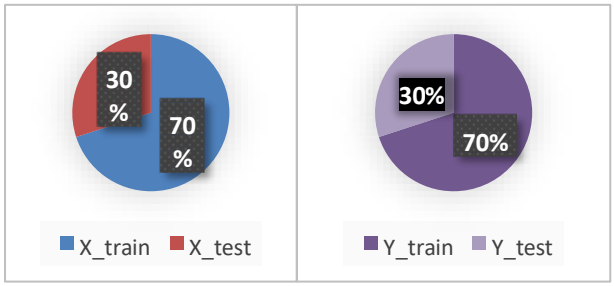


Figure 3 Data Splitting

4.5 Feature Scaling

A data preparation method called "feature scaling" is used to scale down the range of features or variables in a dataset. This ensures that all features contribute evenly to the model and prevents features with higher values from dominating it. Feature scaling is essential when working with datasets where the features have a variety of ranges, units of measurement, or orders of magnitude. Common techniques for feature scaling include standardization, normalization, and min-max scaling. Feature scaling can be used to scale the data more uniformly, making the development of accurate and effective machine learning models easier.

A Standard Scaler (Standardization) is used in this study which uses the following equation:

$$z = \frac{x - \mu}{\sigma} \qquad (1)$$

Where **z** is the new data, **x** the old data, $\boldsymbol{\mu}$ is the mean of the feature values and $\boldsymbol{\sigma}$ is the standard deviation of the feature values. After the data has been cleansed, we can now use ML algorithms to build models and make predictions.

4.6 Machine Learning (ML) classification algorithms

The goal of machine learning (ML) is to develop algorithms that enable computers to learn on their own without requiring programming for each rule to plan or recognize a certain pattern. With the aim of making accurate predictions for future data inputs, supervised learning algorithms create a model using inputs (features) and outputs (targets). To predict results from new data, algorithms understand correlations and patterns [25]. One of the most important and well-liked supervised ML techniques is classification. The two classification types that depend on the total number of outputs are binary and multiple classification [26]. Since we can identify and forecast whether a person would be anemic or not, binary classification was used in this study. There were four classifiers used in this study:

4.7 Random Forest (RF)

A classification technique called random forest (RF) relies on "growing" a set of classifiers using a tree-structure. A new individual is categorized using their traits utilizing each categorization tree in the forest. The trees are built randomly, and each growing tree offers a categorization (or "voting") for a class label. The decision is made using the majority of the trees in the forest's votes [27].

4.8 Decision Tree

One of the supervised learning techniques used for classification and regression is the decision tree. With a root node, branches, internal nodes, and leaf nodes, it is arranged hierarchically. Each leaf node represents a decision among numerous choices, whereas each branch node represents a decision. Since the objective of data mining is to find as much hidden information as possible, this strategy has been thoroughly studied. There is still a lot of uncharted ground in the field of medicine [28].

4.9 Logistic Regression (LR)

LR is a supervised classification method that estimates the likelihood of an event by fitting data to a logistic curve. A binary variable is used to evaluate the outcome. Several expected variables—some of which may be categorical or numerical—are used in logistic regression. The social sciences and the medical field both frequently use LR. In addition,utilized frequently in marketing to ascertain whether consumers are most likely to buy a product [30].

4.10 Multilayer Perceptron MLP

A multilayer representation perceptron classifier, which refers to a neural network itself, is a common name for the MLP Classifier. In contrast to other classification algorithms like SVM and NB, MLP completes the classification task based on the underlying neural network. The multilayer sensory structure of MLP is composed of three layers of nodes: an input layer, a hidden layer, and an output layer. The layer on top is the input layer. The initial layer takes the training input and multiplies it by weights that were chosen at random during initialization. The final product is then given an activation function once biases are added later. again the exception of the first layer, each layer's input data comes from the layer before it, and the procedure is repeated again the output is being sent to the layer below [31][32].

4.11 Performance Measurements

The output of each classification algorithm is visualized and compiled in a confusion matrix when all the suggested algorithms have finished classifying data. A confusion matrix is a table that evaluates the effectiveness of a classification system figure 4. The confusion matrix, which is utilised to supply the classifier's measurement parameters, has four basic qualities as illustrated in Figure 4. The four parameters are as follows:

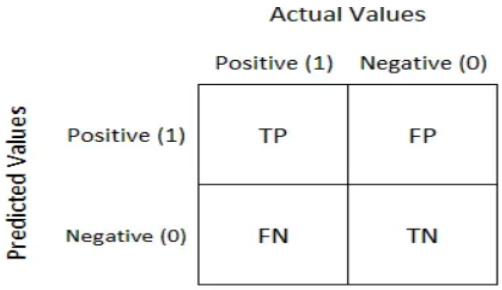


Figure 4. Confusion Matrix

TP (True positive): Total number of cases that were predicted yes and was correctly predicted (in our study means that they really have the disease and the model predict that correctly).

TN (True negative): Total number of cases that were predicted no and was correctly predicted (in our study means that they don't have the disease and the model predicted that correctly).

FP (False positive): Total number of cases that were predicted yes and was a wrong prediction (in our study means that they don't really have the disease, but the model predicted that they had).

FN (False negative): Total number of cases that were predicted no and was a wrong prediction (in our study means that they really have the disease, but the model predicted that they don't have).

Performance measurements include classification accuracy, precision, recall, specificity, and F1-score. It is calculated as given in equations (2), (3), (4), (5), and (6) below for the purposes of comparing and evaluating our suggested methods.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (2)$$

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

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

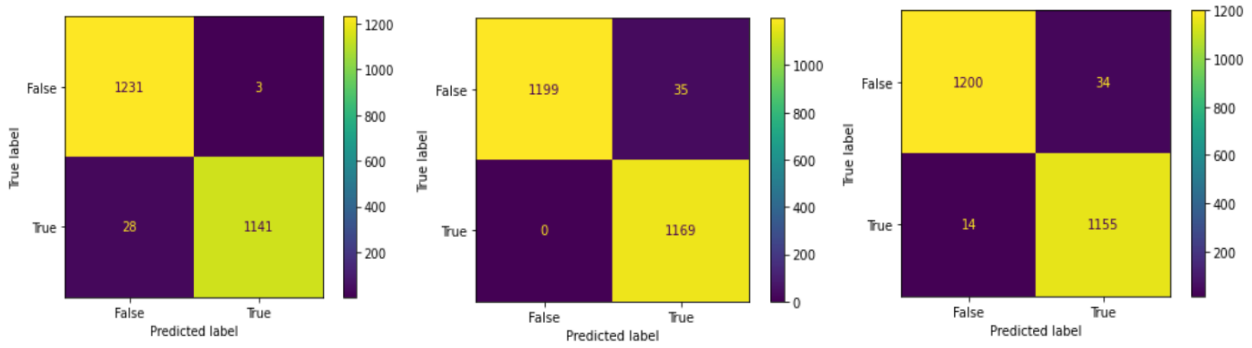
$$Specifity = \frac{TN}{(TN+FP)} \quad (5)$$

$$F1 = \frac{2\,.\,precsion\,.\,recall}{precision+recall} \quad (6)$$

In addition, the receiver operating characteristic curve (ROC) curve was also plotted and the area under the curve was calculated (AUC).

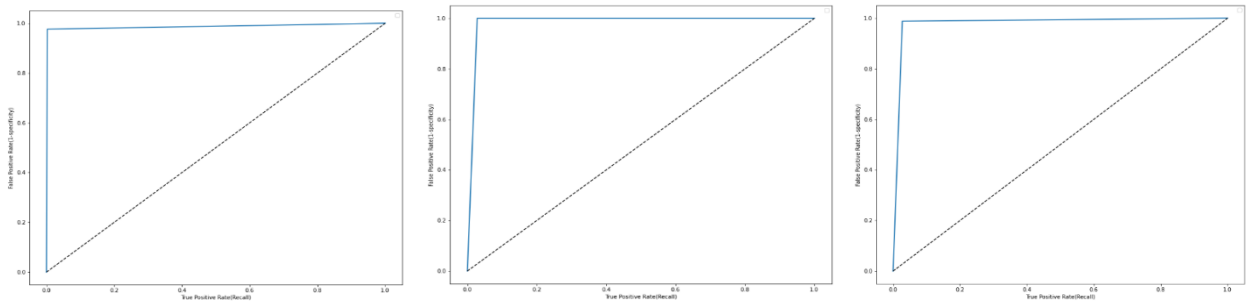
5. Results and Discussions

Following the completion of all the prior steps in our investigation, we are now presenting all the findings and performance metrics for each ML classifier that was employed. Figure 5 displays the confusion matrix for random forest, MLP and decision tree. The ROC curve for three ML models is presented in figure 6.



a) RF confusion b) MLP confusion c) DS confusion matrix

Figure 5. Confusion matrix:(a)Random Forest (b) MLP classifier, (c) Decision tree classifier



a) RF ROC Curve b) MLP ROC Curve c) DS ROC Curve

Figure 6. ROC :(a)Random Forest (b) MLP classifier, (c) Decision tree classifier

By using the performance metrics equations (2), (3), (4),(5) and (6) on the previous confusion matrix we get this outcomes in table1 for proposed 4 models.

Table 1 Performance Measurement

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 score (%)
Random Forest	98.70	99.73	1.0	99.75	98.65
Decision Tree	98.00	97.14	1.0	97.24	98.65
MLP	98.54	97.09	1.0	97.16	98.52
Logistic regression	94.38	94.26	1.0	94.57	94.22

The measures in Table 2 above allow us to see that the outcomes of our prediction models are outstanding and even beyond expectations. The Random Forest (RF) 98.70% is clearly visible. Decision tree, logistic regression, and MLP [20] are contrasted with one other.

6. Conclusion

In this paper, we evaluate the performance of four alternative classifiers in identifying poisonous or edible mushrooms. The experimental results on a test dataset show that Decision and Logistic Regression Technique classification approach outperforms Random Forest, and MLP , in terms of accuracy. The amount of manual labour necessary for prediction can be reduced using automatic prediction.

References

[1] Rial Adity and Setia Hadi Purwono, Jamur – Info Lengkap dan Kiat Sukses Agribisnis. Depok, Indonesia/West Java: Agriflo, 2012.

[2] Kristianus Sunarjon Dasa, "Pemanfaatan bagas sebagai campuran media pertumbuhan jamur tiram putih," vol. 11, pp. 195-201, 2011.

[3] N. Zahan, M. Z. Hasan, M. A. Malek and S. S. Reya, "A Deep Learning-Based Approach for Edible, Inedible and Poisonous Mushroom Classification," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021, pp. 440-444, doi: 10.1109/ICICT4SD50815.2021.9396845.

[4] Al-Mejibli and D. Hamed Abd, "Mushroom Diagnosis Assistance System Based on Machine Learning by Using Mobile Devices Intisar Shadeed Al-Mejibli University of Information Technology and Communications Dhafar Hamed Abd Al-Maaref University College," vol. 9, no. 2, pp. 103–113, 2017.https://doi.org/10.29304/jqcm.2017.9.2.319

[5] M. Alameady, "Classifying Poisonous and Edible Mushrooms in the Agaricus," International Journal of Engineering Sciences & Research Technology, vol. 6, no. 1, pp. 154–164, 2017.

[6] Duong, L.T.; Nguyen, P.; Di Sipio, C.; Di Ruscio, D. Automated fruit recognition using EfficientNet and MixNet. Comput. Electron. Agric. 2020, 171, 105326.

[7] J. M. J. Ward, E. L. Stromberg, D. C. Nowell, F. W. Jr. Nutter, "Gray leaf spot: a disease of global importance in maize production," Plant Disease, vol. 83, no. 10, pp. 884-895, Oct. 1999.

[8] G. Wang, Y. Sun, J. X. Wang, "Automatic ImageBased Plant Disease Severity Estimation Using Deep Learning," Computational Intelligence and Neuroscience, vol. 2017, no. 2, pp. 1-8, Jul. 2017.

[9] Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Networks 61, 85–117.

[10] Izza SabillaS, Sarno R, Siswantoto J, Estimating concentration using Artificial neural network for electronic .Proceds compu. Sci. 2017, 124, 181-188

[11] Sven Behnke, N B Karayiannis, CNet competitive neuraltrees for pattern classification, IEEE transaction son neural network, 9, 6, 1996

[12] Phongsakhon Tongcham, Pichaya Supa, Peerapong Pornwongthong, Pitcha Prasitmeeboon, Mushroom spawn quality classification with machine learning, Computers and Electronics in Agriculture, Vol. 179, 2020, 105865, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2020.105865.

[13] Peng P, Xiaofang P, Classification using deep convolution neural networks, Sensors, 2018,18, 157.

[14] Wang, P.; Liu, J.; Xu, L.; Huang, P.; Luo, X.; Hu, Y.; Kang, Z. Classification of Amanita Species Based on Bilinear Networks with Attention Mechanism. Agriculture 2021, 11, 393. https://doi.org/ 10.3390/agriculture11050393

[15] Schmidhuber J, Deep learning in neural networks: an overview. Neural Networks 61, 85–117, 2015.

[16] Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R and Muharemagic E (2015) Deep learning applications and challenges in big data analytics. Journal of Big Data 2, 1

[17] Sladojevic S, Arsenovic M, Anderla A, Culibrk D and Stefanovic D (2016 Deep neural networks based recognition of plant diseases by leaf image classification. Computational Intelligence and Neuroscience 2016, 3289801.

[18] Mohanty SP, Hughes DP and Salathé M (2016) Using deep learning for image-based plant disease detection. Frontiers in Plant Science 7, 1419.doi: 10.3389/fpls.2016.01419.

[19] Amara J, Bouaziz B and Algergawy A (2017) A deep learning-based approach for banana leaf diseases classification. In Mitschang B (ed.),Datenbanksysteme für Business, Technologie und Web (BTW 2017) – Workshopband. Lecture Notes in Informatics (LNI). Stuttgart, Germany:Gesellschaft für Informatik, pp. 79–88.

[20] Kussul N, Lavreniuk M, Skakun S and Shelestov A (2017) Deep learning classification of land cover and crop types using remote sensing data.IEEE Geoscience and Remote Sensing Letters 14, 778–782

[21] Grinblat GL, Uzal LC, Larese MG and Granitto PM (2016) Deep learning for plant identification using vein morphological patterns. Computers and Electronics in Agriculture 127, 418–424.

[22] Kuwata K and Shibasaki R (2015) Estimating crop yields with deep learning and remotely sensed data. In IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Milan, Italy: IEEE, pp. 858–861.

[23] Rahnemoonfar M and Sheppard C (2017) Deep count: fruit counting based on deep simulated learning. Sensors 17, 905. [24] Chen SW, Shivakumar SS, Dcunha S, Das J, Okon E, Qu C, Taylor CJ andKumar V (2017) Counting apples and oranges with deep learning: a datadriven approach. IEEE Robotics and Automation Letters 2, 781–788

[24]Mukherjee, K.L., Ghosh, S., 2012. Medical laboratory Technology. Procedure Manual for Routine Diagnostic Tests. Vol I (Second edition), 263-266.

[25]Lanier, J. Brian, James J. Park, and Robert C. Callahan. "Anemia in older adults." American family physician 98.7 (2018): 437-442.

[26]Jaiswal, Manish, Anima Srivastava, and Tanveer J. Siddiqui. "Machine learning algorithms for anemia disease prediction." Recent Trends in Communication, Computing, and Electronics: Select Proceedings of IC3E 2018. Springer Singapore, 2019.

[27]Verma, Parth, and Vinay Chopra. "A Review on Machine Learning Algorithms for Anemia disease Prediction." (2022).

[28]Shilpa, S. A., Nagori, M., & Kshirsaga, V. (2011). Classification of anemia using data mining techniques. In Swarm, evolutionary, and memetic computing (pp. 113–121). Springer.

[29]El-kenawy, E.M.T. A Machine Learning Model for Hemoglobin Estimation and Anemia Classification. Int. J. Comput. Sci. Inf. Secur. 2019, 17, 100–108.

[30]Sow, B., et al.: Assessing the relative importance of social determinants of health in malaria and anemia classification based on machine learning techniques. Inform. Health Soc. Care 45(3), 229–241 (2020)

[31]Amin, N., & Habib, A. (2015). Comparison of different classification techniques using WEKA for hematological data. American Journal of Engineering Research, 4(3), 55–61.

[32]Saidi, R., Bouaguel, W., Essoussi, N.: Hybrid feature selection method based on the genetic algorithm and pearson correlation coefficient. In: Hassanien, A.E. (ed.) Machine Learning Paradigms: Theory and Application. SCI, vol. 801, pp. 3–24.

Springer, Cham (2019). https://doi.org/10.1007/978-3-03002357-7_1

[33] Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18–22.

[34]Díaz-Martínez, M.A., Ahumada-Cervantes, M. de los A., Melo-Morín, J.P.: Decision trees as a methodology to determine academic performance in higher education. Rev. Lasallista Investig. 18, 94–104 (2021)