Blood Diseases Detection using Classical Machine Learning Algorithms

Fahad Kamal Alsheref[1]
Information System Department
Faculty of Computers and Artificial Intelligence
Beni-Suef University
Beni-Suef, Egypt

Wael Hassan Gomaa[2]
Computer Science Department
Faculty of Computers and
Artificial Intelligence, Beni-Suef University
Beni-Suef, Egypt

***Abstract*—Blood analysis is an essential indicator for many diseases; it contains several parameters which are a sign for specific blood diseases. For predicting the disease according to the blood analysis, patterns that lead to identifying the disease precisely should be recognized. Machine learning is the field responsible for building models for predicting the output based on previous data. The accuracy of machine learning algorithms is based on the quality of collected data for the learning process; this research presents a novel benchmark data set that contains 668 records. The data set is collected and verified by expert physicians from highly trusted sources. Several classical machine learning algorithms are tested and achieved promising results.**

***Keywords*—Machine learning; classification algorithms; decision trees; KNN; k-means; blood disease**

I. Introduction

Blood has many secrets that affect human life. It is the postman that circulates through body and visits all organs [1]. The growth in age should be reflected in blood. This change could be detected by the values of parameters inside blood analysis tests [2]. Depending on several attributes like age, gender, symptoms, and any health conditions, the physician can choose the specific blood tests for diagnosing the disease. Many blood tests are standard and essential for everyone to get. Blood tests are widespread because of that; most physicians may recommend blood tests to predict the health level of the patient's body [3] [4].

Most of the blood tests do not need special conditions like fasting for 8 to 12 hours before the test or preventing some kinds of medicine [5]. By testing the fluid, different parameters in the blood van be measured. The results help to identify health problems in the early stages or nay predictable diseases [6]. Physicians cannot diagnose diseases and health problems with blood tests alone. However, they can use them as a factor to confirm a diagnosis. These factors may include some signs and symptoms, which could be integrated with other vital signs for diagnosing the diseases [7]. The disease is diagnosing, and prediction process is a necessary process which is based on the quality of data and physician's experience. Applying modern technological tools for helping physicians to improve the accuracy of disease diagnosing, become one of the hot topics of research, especially machine learning and artificial intelligence algorithms [8].

Machine learning is a data analysis technology that teaches computers to act like humans. It uses computational methods to extract information directly from data [8]. The performance of the machine learning algorithm is improved according to the quality of data, as well as enhancing the disease prediction process [9].

The main objective of this research is using machine learning techniques for detecting blood diseases according to the blood tests values; several techniques are performed for finding the most suitable algorithm that maximizes the prediction accuracy [9]. The rest of this paper is organized as follows. Section II introduces background information about the used techniques. Section III presents the different related methods on blood disease prediction using ML classifiers. Section IV describes the data set and the blood test attributes. Section V shows the experiments results. Finally, section VI presents the conclusion and future work of the research.

II. Background

Machine learning is a computer science branch that is responsible for the development of computer systems that can learn and change their reactions according to the situation [9]. The Machine Learning methodology is depending on learning from data inputs and evaluating the model results and trying to optimize the output [10]. It is also used in data analytics for making predictions on data. Figure 1 shows a brief of machine learning activity. Machine learning consists of 3 main models [11]:

- Supervised Learning: Computer is trained with presented inputs and their desired outputs, for predicting the output of future inputs.
- Unsupervised Learning: Computer is presented with inputs without desired outputs.
- Reinforcement learning: Computer interacts with the environment, and it must perform a specific goal without training.

Machine Learning techniques become an essential tool for prediction and decision-making in many disciplines [12]. The availability of clinical data leads machine learning to play a critical role in medical decision making. It serves as a valuable aid in identifying a disease for improving clinical decisions and choosing suitable medical procedures.

Fig. 1. MACHINE LEARNING ACTIVITIES [11].

We used the following classifiers for classifying the patients based on learning datasets; these classifiers are:

- Naive Bayes: it is based on the Bayes theorem. It considers that each attribute in unclassified tuple X is conditionally independent [13].

$$P(C_1|\boldsymbol{X}) = \prod_1^n P(X_i|C_1)P(C_1) \quad (1)$$

P(C1|X) is the probability of tuple X belongs to Class 1, P(C1) the probability of Class 1 that exists in the training set, and $\prod_1^n P(X_i|C_1)$ the production of each attribute in Tuple X the belongs to Class 1.The classification is done by calculating the probablility of tuple X for each labeled class , and the tuple will be classified to the class with the maximum probablility [13]. This algorithm needs a small amount of training data for estimating the vital parameters which made the algorithm extremely fast compared to more sophisticated methods.

- A Bayesian network: it is a probabilistic directed acyclic graphical model; (DAG) it represents a set of variables and their conditional relies on a directed acyclic graph. It is ideal for dealing with an event that occurred and predicting the likelihood that any one of several possible known causes [14].
- A multilayer perceptron: it is a feedforward neural network. It consists of three layers of nodes or more: an input layer, a hidden layer, and an output layer. Each node is a neuron that uses an activation function. It uses a backpropagation supervised learning technique for training; it can distinguish data that is not learned before [15].
- Logit Boost: it is one of the boosting algorithms; its primary purpose is predicting basic protein classes. It performs classification using regression as the base learner, which can deal with multi-class problems [16].
- Random forests classifier: it is a band learning method for classification that operates by constructing a multitude of decision trees by training records with their labeled classes. After building the tree, the unknown records could be classified [17].
- Support vector machine: it represents the training data as points in a flat separated space by an apparent gap. New examples are mapped into space with the forecast category based on which side of the gap they fall [18].
- K-Nearest Neighbor (KNN): it classifies the object based on the distance between the new object and the defined objects. The object is assigned to the class k that has the shortest distance to class k that defined as the nearest neighbor [19].
- Regression analysis: it is a process for rating the relationships among variables. It includes many techniques for modeling and analyzing several variables for finding the relationship between a dependent variable and one or more independent variables. After finding the relation, the missing values of the variable could be predicted with high accuracy [20].
- Decision Tree: it models the attributes and its values with decisions in the tree; where the nodes contain attributes with its values and leaves contain decisions. The algorithm considers all features and makes a binary split on them. It orders the attributes on the tree according to the information gain value in descending order. After building the tree, new tuples will be classified according to its values by traversing the tree until reaching the leaf that contains the class [21].

All these classifiers are used in the diseases prediction process for improving the clinical decision making and minimize the medical errors, in the next section, we listed the recent researches that using the machine learning in blood disease analysis.

III. RELATED WORK

There are many studies in the field of machine learning techniques in disease detection, but a few numbers of them interested in blood diseases detection. Gregor Gunčar [22] and other co-authors write one of the most recent researches that worked on blood disease detection by using machine learning techniques. They used machine learning algorithms based on blood test results. They have built two models to predict blood disease. The first one is a predictive model used most of blood test parameters, and the second one used only a reduced set that is most common inpatient admittance [22]. The two models achieved good results; they get 88% accuracy in the first model, 59% in the second. The key point of this study shows that a machine learning predictive model based on blood tests can predict haematologic accurately. This research contains some limitations; some parameters were not calculated like f-measures and recall that may lead to better results [22].

David Martinez [23] and other co-authors are also interested in blood disease detection, but they concentrate on the textual content of the clinical reports other than the values of blood analysis parameters.

They collected free-text computed tomography (CT) over a specific hospitalization period (2003–2011); this collection contains 264 Invasive fungal diseases (IFDs) and 289 control patients. They worked with text mining methods and on the

sentence level [23]. They tested a variety of Machine Learning, rule-based, and hybrid systems. Also, it extracts the bags of words, bags of phrases, and bags of concepts. The proposed model used Support Vector Machines and achieved a high recall and precision at 95% at 71% respectively. The core of this model is the high quality of the collected documents and the extraction of information from textual reports and uses them in the disease prediction [24].

IV. Blood Diseases Analysis Data Set

This research presents a new benchmark dataset; it contains 668 patient's blood analysis. Each blood analysis contains 28 parameters; these parameters are presented on table I.

The dataset contains four main classes related to four different blood diseases:

- **Thrombocytopenia**: it is about the lack of platelets. It is not so dangerous but sometimes leads to bleed too much [25].
- **Leukocytosis**: it causes an increase in white cells above the normal range in the blood. It may cause certain parasitic infections or bone tumors, as well as leukemia [26].
- **Anemia**: it **is** a decrease in the amount of hemoglobin or red blood cells in the blood. It may cause vague and may include feeling tired, shortness of breath, or weakness [27].
- **Normal**: in this class, which all parameters values are normal, and there are no essential notifications in the blood analysis.

TABLE. I. Blood Analysis Parameters [25]

Parameter	Description	Normal Range Values
Age	Age of patient	
Sex	Gender of patient	
WBC	white blood cells	normal 4.5-10
RBC	RED blood cells	for female 4.2 to 5.4
Hgb	Hemoglobin	Newborn Babies 17- 22
HCT	Hematocrit	For women: 36.1% to 44.3%
MCV	Macrocytic Anemia:	27 to 31
MCH	Mean Corpuscular Hemoglobin	30 and 37
MCHC	Mean Corpuscular Hemoglobin Concentration	33-36
PLT	Platelet Count	150,000 to 400,000
RDW-SD	Red blood cell distribution width	29-46
RDW-CV	Red blood cell distribution width	11.6 – 14.6%
PDW	platelet distribution width	
MPV	Mean platelet volume	7.5-11.5
P-LCR	Platelet larger cell ratio	
PCT	Procalcitonin	0.1-0.5
NEUT	Neutropenia	1500-8000
LYMPH	Lymphocytes	1000 to 4000
MONO	Mononucleosis	
EO	Eosinophil granulocyte	1 to 6
BASO	Basophil granulocyte	0.0 - 2.0 %
IG	Intravenous immunoglobulin	1-2 grams
NRBC	nucleated red blood cells	0
RET	Reticulocytes	0.5% to 2.5%
IRF	Immature reticulocyte fraction	0.8–4.7% of reticulocytes
LFR	low fluorescence ratio of reticulocytes	87.9–98.4%
MFR	medium fluorescence ratio of reticulocytes	1.6–11.0%
HFR	high fluorescence ratio of reticulocytes	0.0–1.7%

Each record in the proposed dataset is labeled with his related class; this classification is performed manually by expert physicians.

V. Experiments Results and Discussion

Using the Weka tool, a classical machine learning algorithms are applied on 668 records that belong to four different classes as described in the data set section. 10-fold cross-validation is used for all of the experiments after performing the required preprocessing modules presented in fig.1. Cross-Validation is a statistical method of evaluating and comparing learning classifiers by dividing data into two segments: one used to learn or train a model and the other used to validate the model. The training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against.

For each classifier several metrics were measured for determining the accuracy. Furthermore, the parameters values of each classifier were changed according to the specifications of each classifier. Table II presents the evaluation metrics used in the experiments and their description. Table III shows the experiments results. The accuracy of all classifiers is ranged between 71.2% and 98.16%. The LogitBoost classifier has the highest accuracy, where Support Vector Machine classifier has the lowest value. Table IV shows the classifiers accuracy in descending order. The overall results prove the success of applying the classical machine learning algorithms in the process of blood diseases prediction.

TABLE. II. Evaluation Metrics [28]

Metric	Description
TP Rate	True Positive Rate
FP Rate	False Positive Rate
Precision	A measure of statistical variability
Recall	Classifier Sensitivity
F-Measure	A measure of a test's accuracy
MCC	A measure of the quality of binary (two-class) classifications
ROC Area	A graph showing the performance of a classification model at all classification thresholds
PRC Area	Precision/Recall
Accuracy	Accuracy of classifier
Mean absolute error	Assessing the quality of a machine learning model

TABLE. III. Experiments Results

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Accuracy	Mean absolute error
NaiveBayes	0.816	0.059	0.862	0.816	0.835	0.753	0.933	0.857	81.60%	0.09
Bayesian network	0.929	0.04	0.936	0.929	0.93	0.898	0.984	0.967	92.86%	0.0362
MultilayerPerceptron	0.918	0.04	0.918	0.918	0.918	0.879	0.974	0.95	91.80%	0.04
LogitBoost	0.982	0.01	0.982	0.982	0.98	0.972	0.995	0.987	98.16%	0.023
Random forests	0.971	0.022	0.971	0.971	0.969	0.956	0.996	0.99	97.12%	0.042
Support Vector Machine	0.712	0.329	0.799	0.712	0.64	0.494	0.691	0.584	71.20%	0.14
K-Nearest Neighbor	0.93	0.048	0.928	0.93	0.927	0.892	0.94	0.883	92.97%	0.04
Regression analysis	0.965	0.02	0.965	0.965	0.964	0.948	0.992	0.979	96.54%	0.0447
Decision Tree	0.97	0.018	0.969	0.97	0.969	0.955	0.979	0.955	97.00%	0.018

TABLE. IV. Accuracy Results in Descending Order

Classifier	Accuracy
LogitBoost	**98.16%**
Random forests	97.12%
Decision Tree	97.00%
Regression analysis	96.54%
K-Nearest Neighbor	92.97%
Bayesian network	92.86%
MultilayerPerceptron	91.80%
NaiveBayes	81.60%
Support Vector Machine	71.20%

VI. Conclusion and Future Work

Machine learning becomes an essential technique for modeling the human process in many disciplines, especially in the medical field, because of the high availability of data. One of the essential disease detectors is the blood analysis; as it contains many parameters with different values that indicates definite proof for the existence of the disease. The machine learning algorithm accuracy depends mainly on the quality of the dataset; for this reason, a high-quality dataset is collected and verified from expert physicians. This dataset is used for training the classifiers for obtaining high accuracy. We tested several classifiers and achieved accuracy up to 98.16% which realize the research objective, which is helping the physicians to predict the blood diseases according to general blood test.

The future work will focus on testing the proposed data set using different deep learning algorithms to compare between classical and deep learning approaches in this research area. Furthermore, an online Internet of Things (IOT) application will be implemented to collect and test more blood data.

References

[1] Lewontin, Richard C. It ain't necessarily so: The dream of the human genome and other illusions. New York Review of Books, 2001.

[2] Feldman, Eric A., Eric Feldman, and Ronald Bayer, eds. Blood feuds: AIDS, blood, and the politics of medical disaster. Oxford University Press, USA, 1999.

[3] Fekkes, Minne, et al. "Do bullied children get ill, or do ill children get bullied? A prospective cohort study on the relationship between bullying and health-related symptoms." Pediatrics 117.5 ;2006: 1568-1574.

[4] ESHRE, The Rotterdam, and ASRM-Sponsored PCOS Consensus Workshop Group. "Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome." Fertility and sterility 81.1 ;2004: 19-25.

[5] Schalm, Oscar William, Nemi Chand Jain, and Edward James Carroll. Veterinary hematology. No. 3rd edition. Lea & Febiger., 1975.

[6] Allison, James E., et al. "A comparison of fecal occult-blood tests for colorectal-cancer screening." New England Journal of Medicine 334.3 ;1996: 155-160.

[7] Park, Sang Hyuk, et al. "Establishment of age-and gender-specific reference ranges for 36 routine and 57 cell population data items in a new automated blood cell analyzer, Sysmex XN-2000." Annals of laboratory medicine 36.3 ;2016: 244-249.

[8] Cabitza, Federico, Raffaele Rasoini, and Gian Franco Gensini. "Unintended consequences of machine learning in medicine." Jama 318.6 ;2017: 517-518.

[9] Darcy, Alison M., Alan K. Louie, and Laura Weiss Roberts. "Machine learning and the profession of medicine." Jama 315.6 ;2016: 551-552.

[10] Jiang, Min, et al. "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries." Journal of the American Medical Informatics Association 18.5 ;2011: 601-606.

[11] Lison, Pierre. "An introduction to machine learning." ;2015.

[12] Michalski, Ryszard S., and Yves Kodratoff. "Research in machine learning: Recent progress, classification of methods, and future directions." Machine learning. Morgan Kaufmann, 1990. 3-30.

[13] Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.

[14] Friedman, Nir, Dan Geiger, and Moises Goldszmidt. "Bayesian network classifiers." Machine learning 29.2-3 ;1997: 131-163.

[15] Ruck, Dennis W., et al. "The multilayer perceptron as an approximation to a Bayes optimal discriminant function." IEEE Transactions on Neural Networks 1.4 ;1990: 296-298.

[16] Otero, José, and Luciano Sánchez. "Induction of descriptive fuzzy classifiers with the Logitboost algorithm." Soft Computing 10.9 ;2006: 825-835.

[17] Breiman, Leo. "Random forests." Machine learning 45.1 ;2001: 5-32.

[18] Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers" Neural processing letters 9.3 ;1999: 293-300.

[19] Keller, James M., Michael R. Gray, and James A. Givens. "A fuzzy k-nearest neighbor algorithm." IEEE transactions on systems, man, and cybernetics 4 ;1985: 580-585.

[20] Seber, George AF, and Alan J. Lee. Linear regression analysis. Vol. 329. John Wiley & Sons, 2012.

[21] Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics 21.3 ;1991: 660-674.

[22] Gunčar, Gregor, et al. "An application of machine learning to haematological diagnosis." Scientific reports 8.1 ;2018: 411.

[23] Martinez, David, et al. "Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans." Journal of biomedical informatics 53 ;2015: 251-260.

[24] Pekelharing, J. M., et al. "Haematology reference intervals for established and novel parameters in healthy adults." Sysmex Journal International 20.1 ;2010: 1-9.

[25] Warkentin, Theodore E., and John G. Kelton. "A 14-year study of heparin-induced thrombocytopenia." The American journal of medicine 101.5 ;1996: 502-507.

[26] Shopsin, Baron, Richard Friedmann, and Samuel Gershon. "Lithium and leukocytosis" Clinical Pharmacology & Therapeutics 12.6;1971:923-928.

[27] Weiss, Guenter, and Lawrence T. Goodnough. "Anemia of chronic disease." New England Journal of Medicine 352.10 ;2005: 1011-1023.

[28] Ragab, Abdul Hamid M., et al. "A comparative analysis of classification algorithms for students college enrollment approval using data mining." Proceedings of the 2014 Workshop on Interaction Design in Educational Environments. ACM, 2014.