International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Applications of Machine Learning to Haematological Diagnosis

***Srishti Adkar*[1], *Shreya Patki*[1], *Ameya Bhagwat*[1], *Chaitrali Ghanekar*[1], *Prof. Swati Shekapure*[2]**

[1]UG student, Department of Computer Engineering, Marathwada Mitra Mandal College of Engineering,Pune. Pune, India
[2]Professor, Department of Computer Engineering, Marathwada Mitra Mandal College of Engineering,Pune. Pune, India

ABSTRACT

Medical diagnosis must be precise and timely, in order to treat diseases effectively. This project was developed for predicting the possibility of a disease based on the parameters from the blood reports with the help of machine learning algorithms. The proposed model is built by applying the random forest algorithm. The model was trained by building the dataset from blood reports. The prediction model includes all available blood test parameters and the model performs well, achieving a prediction accuracy of 99.3% for the list of six most likely diseases. The model indicates that a smaller number of characteristics can reflect a relevant type of a disease. This information increases the model's usefulness for general practitioners and suggests that blood test results contain more information than most doctors realize. Our predictive algorithms were found to be as accurate as haematology specialists in a clinical test. This finding has the potential to open up new avenues for medical diagnostics.

Keywords—Machine learning, Random forest, flask, Haematological Diagnosis, healthspan.

1.Introduction

Over the last 100 years, the average lifespan has grown because of current technological breakthroughs. However, given the fact that chronic diseases currently affect the majority of the population, growing older does not always imply a better quality of life. Our health span is increasing as well.As we become older, maintaining our physical and mental health becomes more vital. How is it possible to quantify and increase our health span, as well as predict chronic diseases? A blood test is performed. It contains numerous factors that are an indication for particular diseases; it is an important indicator for many diseases. Disorders of the blood Patterns that lead to disease are used to forecast disease based on blood analysis. It is important to appreciate the importance of correctly defining the condition. The field of machine learning is in charge of this. To create models for projecting output based on historical data The machine's precision, the quality of the data acquired for the learning process determines the learning algorithms. The information was gathered from a variety of laboratories and double-checked by experts from reputable sources.

Medical diagnosis must be precise and timely in order to treat diseases effectively. For that,we developed a model based on machine learning techniques and laboratory blood test findings to detect health span and haematologic disorders.

Blood contains numerous mysteries that have an impact on human life. It is the postman who travels across the body and visits all of the organs. Blood should show the progression of age. The values of parameters in blood analysis tests could detect this shift. The physician might choose specific blood tests for diagnosing the disease based on numerous factors such as age, gender, symptoms, and any health issues. Many blood tests are routine and required for everyone. As a result, blood tests are widely used; most physicians may offer blood testing to estimate a patient's overall health. Using machine learning and artificial intelligence algorithms, current technological tools are being used to assist clinicians in improving the accuracy of disease diagnosis.

For general practitioners and other clinicians who are unfamiliar with specific hematological diagnosis, machine learning to interpret blood laboratory tests could be very valuable.

Using only the patient's common blood laboratory findings, our model predicts the most likely hematological illnesses. Patients and their relatives could use a diagnostic tool like ours to know and understand, and more importantly, predict the disease they may face in the future in advance, and also in the process of requesting a second opinion, at a time when patients are becoming increasingly aware and want to be actively involved in their treatment.

The premise of machine learning is that a computer programme can learn and adapt to new data without the need for human intervention. The premise of machine learning is that a computer programme can learn and adapt to new data without the need for human intervention. Machine learning is an AI branch that keeps a computer's built-in algorithms active despite global economic shifts.

1. Supervised Learning: Supervised learning algorithms create a mathematical model of a set of data that includes both the inputs and the outputs that are sought.The information is referred to as training data, and it consists of a collection of training instances. Each training example has one or more inputs and a supervisory signal as the desired output. Each training sample is represented by an array or vector, sometimes referred to as a feature vector, and the training data is represented by a matrix in the mathematical model. Supervised learning techniques develop a function that may be used to predict the output associated with fresh inputs by iteratively optimizing an objective function.

When the outputs can have any numerical value within a range, regression procedures are applied. An incoming email, for example, would be the input to a classification algorithm that filters emails, and the output would be the name of the folder to file the email in.

The purpose of similarity learning, which is closely connected to regression and classification, is to learn from examples using a similarity function that quantifies how similar or related two items are. It can be used for ranking, recommendation systems, visual identification tracking, face verification etc.
2. Unsupervised Learning: Unsupervised learning algorithms take a set of data with only inputs and try to detect structure in it, such as grouping or clustering data points. As a result, the algorithms learn from unlabeled, uncategorized, and uncategorized test data.Unsupervised learning algorithms look for commonalities in data instead than responding to criticism.The data and react based on whether or not there are any commonalities in collection of information The topic of density estimation is one of the most common uses of unsupervised learning. Find the probability density function, for example, using statistics. Despite the fact that learning is unsupervised entails summarizing and interpreting data aspects in various disciplines.

Literature review

Recent advances in deep learning and large datasets have enabled algorithms to outperform medical professionals in a variety of medical imaging tasks, such as diabetic retinopathy detection (Gulshan et al., 2016), skin cancer classification (Esteva et al., 2017), arrhythmia detection (Rajpurkar et al., 2017), and hemorrhage detection (Rajpurkar et al., 2017). (Grewal et al., 2017).

There are many studies in the field of disease identification using machine learning techniques, but only a few of them are focused on blood diseases. One of the most recent studies on blood illness identification using machine learning approaches was written by Gregor Gunar and other co-authors. They used blood test results to train machine learning systems. They've created two blood illness prediction models. The first is a predictive model that uses the majority of blood test parameters, while the second uses only a smaller collection of parameters that are most prevalent inpatient admissions. Both models performed well, with the first achieving 88 percent accuracy and the second 59 percent. The study's main finding is that a machine learning predictive model based on blood tests may accurately predict haematologic outcomes. Some characteristics, such as f-measures and recall, were not calculated in this study, which could have resulted in better results.

Proposed methodology

Data acquisition: Acquiring raw data from the database.

Machine learning models learn extensively from the datasets . So it is recommended that large datasets should be used to train the model for better accuracy. For this project, we have collected around 400000 blood test samples from various pathology laboratories.These blood test samples primarily contained blood test parameters such as PCV, MCV, MCH, MCHC, Eosinophils, WBC count, RBC count, Lymphocytes, Monocytes, Neutrophils, Hemoglobin, RDW, Platelet count.

Data filtering: Selecting only those parameters from the blood reports, which are essential to predict anomalies in health, or deficiencies.

Out of the total blood test samples collected, around 800 samples were used for data preprocessing and training.Some of the parameters which we filtered out in the process were Eosinophils, RDW, Lymphocytes, Monocytes, Neutrophils. PCV, MCV, MCH, MCHC, WBC count, Hemoglobin, RDW, Platelet count are the key parameters in predicting diseases such as Anemia, Leukocytosis, Thrombocytosis, Thalassemia and Leukemia.

Data preprocessing: Canonizing blood parameters (comparing them to our reference parameter database, removing outliers, and dealing with missing values) (imputation).

Data cleaning process involved removing redundant, ambiguous, null and missing values from the dataset. To ensure homogeneity in the data, we converted the age of every patient from days to its equivalent number in years.

Furthermore, the parameters such as gender and diseases contained categorical values. Label encoding enabled us to convert these values into the numeric ones.

Data modeling: Building the predictive model.

Decision Tree: It represents properties and their values in the form of a tree, with nodes containing attributes and leaves containing decisions. The method takes into account all characteristics and performs a binary split on them. It arranges the characteristics on the tree in descending order based on the information gain value. Following the construction of the tree, fresh tuples will be categorized according to their values by traversing the tree till reaching the class leaf.

Random Forest: Random Forest is a well-known supervised learning machine learning algorithm. It is based on ensemble learning, which is the practice of combining multiple classifiers to solve a difficult problem and increase the model's performance.

"Random Forest is a classifier that enhances the projected accuracy of a given dataset while aggregating a number of decision trees on diverse subsets of that dataset," according to the name.

The random forest collects forecasts from each tree and predicts the final output based on the majority votes of projections, rather than depending on a single decision tree.

Using the train test split class from the scikit learn library, we split the data in the ratio of 70:30.

Model training:

In scikit learn library, there is a class called RandomizedSearchCV containing features like:

n_estimators: number of trees in random forest

max_features: number of features to be considered at every split.

max_depth: maximum number of levels in the tree.

min_samples_split: minimum number of samples required to split a node.

min_samples_leaf: minimum number of samples required at each node.

We stored the results of these features in the form of a grid.

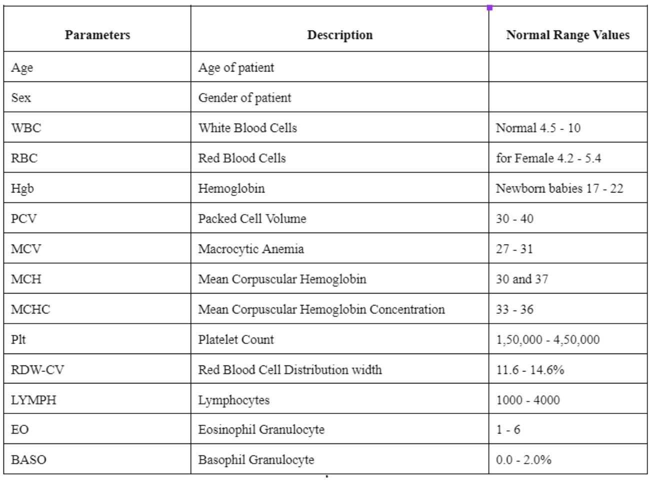
Model Testing:

For calculating the accuracy of the model, we used the accuracy_score function. The model performed greatly achieving an accuracy of 99% overall.

Confusion matrix: A confusion matrix is a tabular representation of your prediction model's performance. The number of predictions produced by the model where it categorized the classes correctly or erroneously is represented by each entry in a confusion matrix.

We have used a confusion matrix for checking the performance of the prediction model.

Actual parameters used for blood disease detection are as follows:



Parameters	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 - 5.4
Hgb	Hemoglobin	Newborn babies 17 - 22
PCV	Packed Cell Volume	30 - 40
MCV	Macrocytic Anemia	27 - 31
MCH	Mean Corpuscular Hemoglobin	30 and 37
MCHC	Mean Corpuscular Hemoglobin Concentration	33 - 36
Plt	Platelet Count	1,50,000 - 4,50,000
RDW-CV	Red Blood Cell Distribution width	11.6 - 14.6%
LYMPH	Lymphocytes	1000 - 4000
EO	Eosinophil Granulocyte	1 - 6
BASO	Basophil Granulocyte	0.0 - 2.0%

On the basis of these parameters, we predict

Thrombocytopenia: Thrombocytopenia is a condition in which platelets are insufficient. It is not hazardous, although it can cause excessive bleeding.

Thrombocytosis: Thrombocytosis is a condition in which your body creates an excessive number of platelets. When the cause is an underlying ailment, such as an infection, it's called reactive thrombocytosis or secondary thrombocytosis.

Leukemia: Increased quantities of immature or aberrant leukocytes are produced by the bone marrow and other blood-forming organs in this malignant, progressive disease. These inhibit the creation of normal blood cells, resulting in anemia and other side effects.

Leukocytosis: Leukocytosis is a condition in which the number of white blood cells in the blood rises over the usual range. It has the potential to produce parasite infections, bone tumors, and leukemia.

Anemia: Anemia is a condition in which the number of hemoglobin or red blood cells in the blood is reduced. It can cause a variety of symptoms, including fatigue, shortness of breath, and weakness.

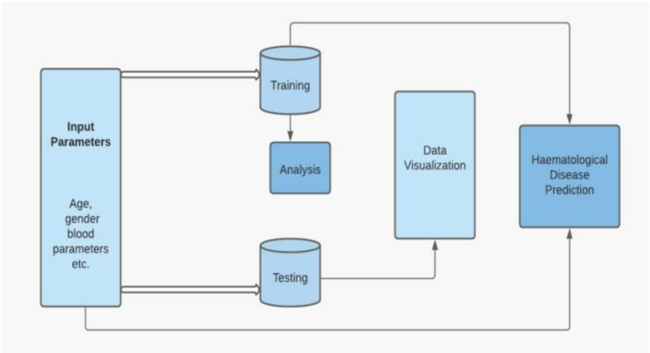
Microcytic Anemia: When the MCV range is below 80fl.

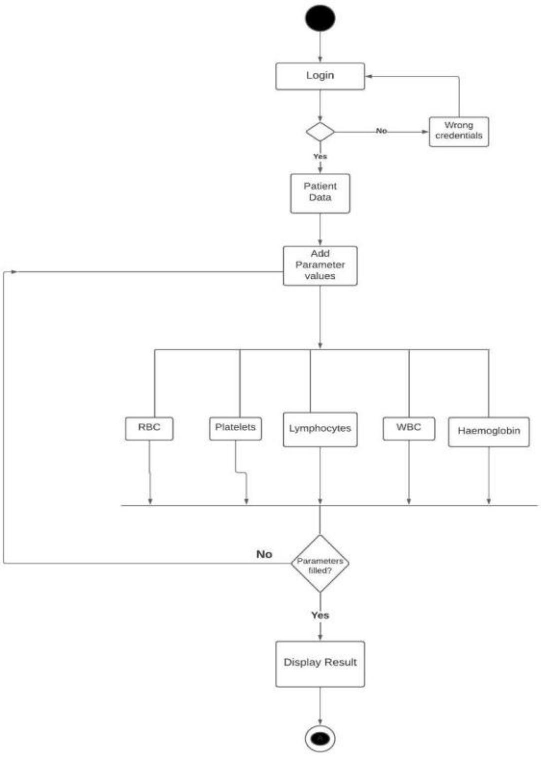
Macrocytic Anemia: When the MCV range is above 100fl.

Thalassemia: Thalassemia is a hereditary (i.e., passed down through the generations) blood illness characterized by the body's inability to produce enough hemoglobin, a key component of red blood cells. When the body's red blood cells don't have enough hemoglobin, they don't function effectively and survive for shorter periods of time, resulting in fewer healthy red blood cells flowing through the bloodstream.

In the Normal class, all parameter values are normal, and the blood analysis contains no essential notifications.

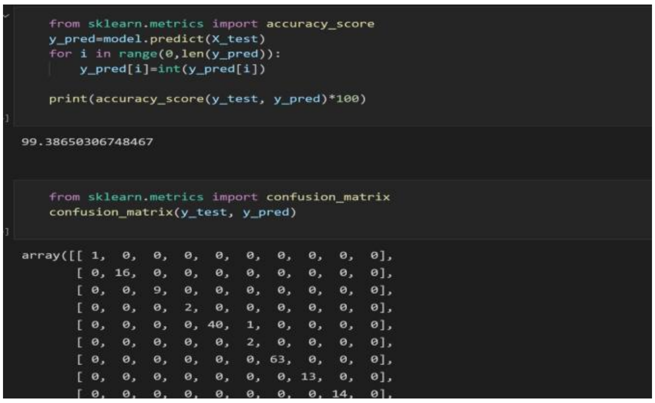
Evaluation: evaluating the predictive model


Input Parameters
Age, gender blood parameters etc.
Training
Analysis
Testing
Data Visualization
Haematological Disease Prediction



Output and results

The model works perfectly and is able to classify the diseases on the basis of the given parameters.. The model gave a great accuracy of 99.38% on the training set as well. The quality of acquired data for the learning process determines the accuracy of machine learning algorithms; this study introduces a new benchmark data set with 600 records. Expert physicians gathered and confirmed the data collection from highly reputable sources.



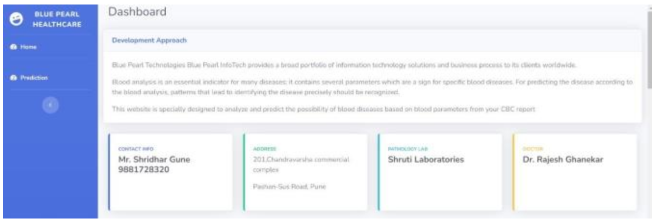
```
from sklearn.metrics import accuracy_score
y_pred=model.predict(X_test)
for i in range(0,len(y_pred)):
    y_pred[i]=int(y_pred[i])

print(accuracy_score(y_test, y_pred)*100)
```

```
99.38650306748467
```

```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred)
```

```
array([[ 1,  0,  0,  0,  0,  0,  0,  0,  0,  0],
       [ 0, 16,  0,  0,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  9,  0,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  2,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0, 40,  1,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  2,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0, 63,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0, 13,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  0,  0, 14,  0],
```



Development Approach

Blue Pearl Technologies Blue Pearl InfoTech provides a broad portfolio of information technology solutions and business process to its clients worldwide.

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.

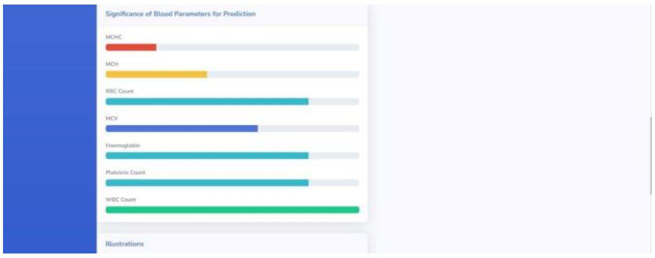
This website is specially designed to analyze and predict the possibility of blood diseases based on blood parameters from your CBC report

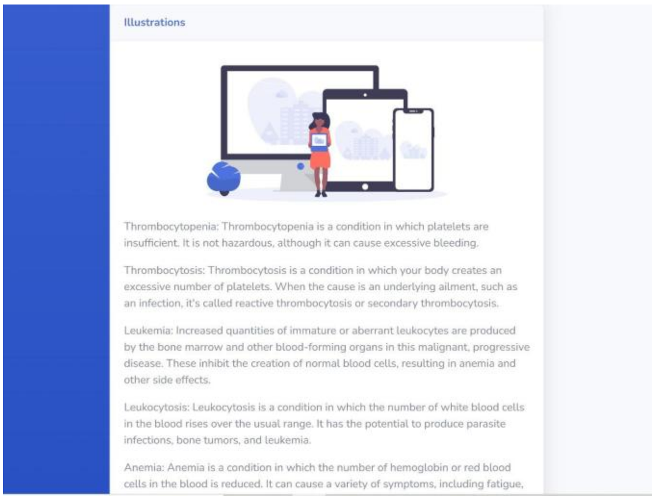
CONTACT INFO
Mr. Shridhar Gune
9881728320

ADDRESS
201,Chandravarsha commercial complex
Pashan-Sus Road, Pune

PATHOLOGY LAB
Shruti Laboratories

DOCTOR
Dr. Rajesh Ghanekar





Illustrations

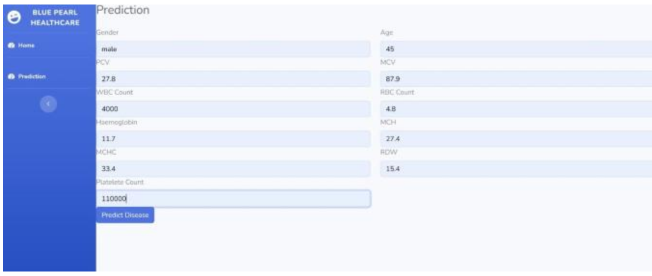
Thrombocytopenia: Thrombocytopenia is a condition in which platelets are insufficient. It is not hazardous, although it can cause excessive bleeding.

Thrombocytosis: Thrombocytosis is a condition in which your body creates an excessive number of platelets. When the cause is an underlying ailment, such as an infection, it's called reactive thrombocytosis or secondary thrombocytosis.

Leukemia: Increased quantities of immature or aberrant leukocytes are produced by the bone marrow and other blood-forming organs in this malignant, progressive disease. These inhibit the creation of normal blood cells, resulting in anemia and other side effects.

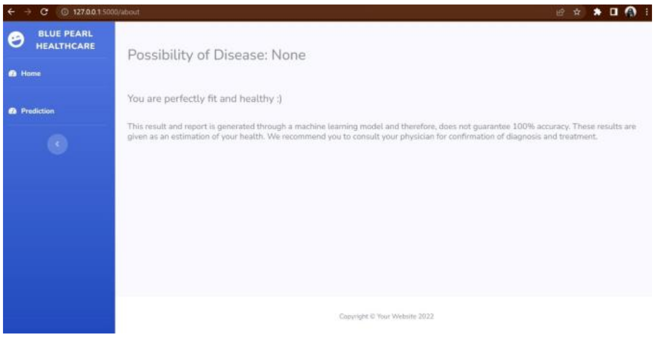
Leukocytosis: Leukocytosis is a condition in which the number of white blood cells in the blood rises over the usual range. It has the potential to produce parasite infections, bone tumors, and leukemia.

Anemia: Anemia is a condition in which the number of hemoglobin or red blood cells in the blood is reduced. It can cause a variety of symptoms, including fatigue,



Prediction

Gender: male
Age: 45
PCV: 27.8
MCV: 87.9
WBC Count: 4000
RBC Count: 4.8
Haemoglobin: 11.7
MCH: 27.4
MCHC: 33.4
RDW: 15.4
Platelete Count: 110000

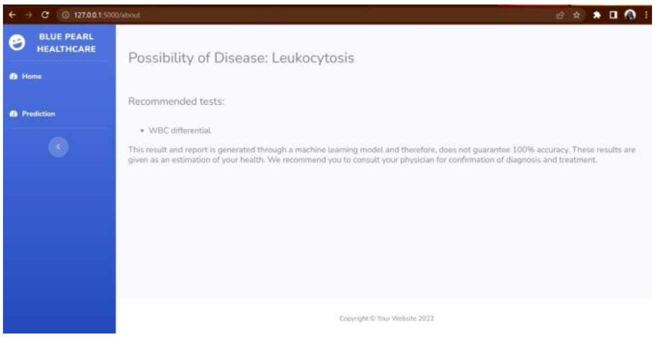


Possibility of Disease: None

You are perfectly fit and healthy :)

This result and report is generated through a machine learning model and therefore, does not guarantee 100% accuracy. These results are given as an estimation of your health. We recommend you to consult your physician for confirmation of diagnosis and treatment.

Copyright © Your Website 2022



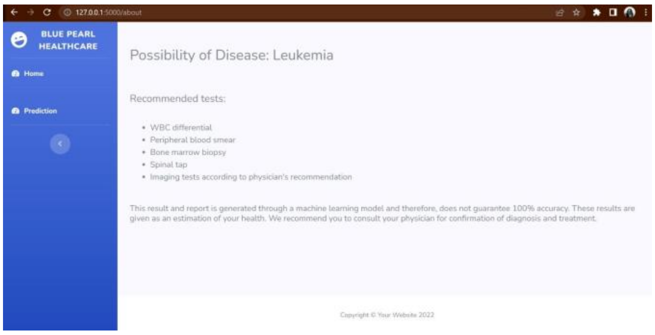
Possibility of Disease: Leukocytosis

Recommended tests:

- WBC differential

This result and report is generated through a machine learning model and therefore, does not guarantee 100% accuracy. These results are given as an estimation of your health. We recommend you to consult your physician for confirmation of diagnosis and treatment.

Copyright © Your Website 2022



Possibility of Disease: Leukemia

Recommended tests:

- WBC differential
- Peripheral blood smear
- Bone marrow biopsy
- Spinal tap
- Imaging tests according to physician's recommendation

This result and report is generated through a machine learning model and therefore, does not guarantee 100% accuracy. These results are given as an estimation of your health. We recommend you to consult your physician for confirmation of diagnosis and treatment.

Copyright © Your Website 2022

Conclusion and Future work

Because of the abundance of data, machine learning has become a vital tool for modeling human processes in a variety of areas, particularly in medicine. The blood analysis is an important illness detector since it comprises several parameters with varied values that indicate clear confirmation of the disease's existence. The accuracy of the machine learning algorithm is largely determined by the dataset quality; as a result, a high-quality dataset is gathered and evaluated by expert physicians. This dataset is used to train classifiers so that they can achieve high accuracy. We examined many classifiers and achieved accuracy of up to 99.28 percent, achieving the research goal of assisting clinicians in the prediction of blood disorders based on routine blood tests.

Machine learning could be particularly effective for interpreting blood laboratory testing. Because our approach predicts the most likely hematological disorders using only a few variables, it is suitable for general practitioners and other clinicians who are unfamiliar with specialized hematological diagnostics. The results of a patient's routine blood tests This capacity could also aid physicians in accurately referring patients. The utility of such models would be considerably enhanced if they were integrated into medical information systems, i.e., as decision assistance for health care practitioners who included a list of the most likely diseases in laboratory results on their own.

Furthermore, such integration could encourage the creation of unified, large-scale digital databases of patient data, which could influence the path of medicine on their own. Patients and their relatives could use a diagnostic tool like ours to know and understand, and more importantly, predict the disease they may face in the future in advance, and also in the process of requesting a second opinion, at a time when patients are becoming increasingly aware and want to be actively involved in their treatment.

References

[1] Zhavoronkov, A., Mamoshina, P., Vanhaelen, Q., Scheibye-Knudsen, M., Moskalev, A., Aliper, A. (2019). Artificial intelligence for aging and longevity research: Recent advances and perspectives. Ageing Research Reviews, 49, 49-66. doi:10.1016/j.arr.2018.11.003

[2] Tsai, M., Tao, Y. (2019). Machine Learning Based Common Radiologist Level Pneumonia Detection on Chest X-rays. 2019 13th International Conference on Signal Processing and Communication Systems (ICSPCS) doi:10.1109/icspcs47537.2019.9008684

[3] Grewal, M., Srivastava, M. M., Kumar, P., Varadarajan, S. (2018). RADnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans.

[4] 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). doi:10.1109/isbi.2018.8363574

[5] Alsheref, F. K., Hassan, W. (2019). Blood Diseases Detection using Classical Machine Learning Algorithms. International Journal of Advanced Computer Science and Applications, 10(7). doi:10.14569/ijacsa.2019.0100712

[6] Gunˇcar, G., Kukar, M., Notar, M., Brvar, M., Cernelˇc, P., Notar, M., Notar, M.(2018). An application of machine learning to haematological diagnosis. Scientific Reports, 8(1). doi:10.1038/s41598-017-18564-8

[7] Luo, Y., Szolovits, P., Dighe, A. S., Baron, J. M. (2016). Using Machine Learning to Predict Laboratory Test Results. American Journal of Clinical Pathology, 145(6), 778-788. doi:10.1093/ajcp/aqw064

[8] Bruijne, M. D. (2016). Machine learning approaches in medical image analysis: From detection to diagnosis. Medical Image Analysis, 33, 94-97.doi:10.1016/j.media.2016.06.032

[9] Gunčar, Gregor, et al. "An application of machine learning to haematological diagnosis." Scientific reports 8.1.2018: 411.

[10] 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.

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

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

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

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

[15] 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.