ISSN: 2755-0168

Journal of Biotechnology & Bioinformatics Research

SCIENTIFIC Research and Community

Research Article

Open Access

Analysis of Medical Laboratory Data and Biomarker Prediction Using Machine Learning

Weam Fakir*, Youssef Fakir and Rachid EL Ayachi

Laborotory of information processing and dicision support, Faculty of Sciences and Technics, Sultan Moulay Slimane University, Morocco

ABSTRACT

Medical laboratory data offer critical insights into patient health, enabling early detection of diseases and monitoring of treatment efficacy. This study investigates the application of machine learning (ML) algorithms specifically Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) neural networks in analyzing a dataset comprising routine biological parameters. The dataset includes complete blood count (CBC), platelets, erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), electrolytes, and renal function indicators. The primary objective is to classify patients based on anomaly risk and predict potential biomarkers indicative of underlying health conditions. The findings demonstrate that ML models can effectively process complex medical data, offering valuable tools for enhancing diagnostic accuracy and patient care.

***Corresponding author**

Weam Fakir, Laborotory of information processing and dicision support, Faculty of Sciences and Technics, Sultan Moulay Slimane University, Morocco.

Received: November 29, 2025; **Accepted:** December 04, 2025; **Published:** December 13, 2025

Keywords: Medical Laboratory Data, Machine Learning, Biomarker Prediction, Random Forest, Support Vector Machine, Multi Layer Perceptron, Classification, CRP, Creatinine, Urea

Introduction

Medical laboratory tests are fundamental in diagnosing diseases, monitoring treatment responses, and assessing overall health status [1, 2]. Traditionally, the interpretation of these tests relies heavily on clinical expertise and manual analysis. However, the increasing volume and complexity of laboratory data necessitate the adoption of advanced analytical techniques. Machine learning (ML) has emerged as a powerful tool in this domain, capable of identifying intricate patterns within large datasets. By leveraging ML algorithms, healthcare providers can enhance diagnostic accuracy, predict disease progression, and personalize treatment plans. This study explores the potential of ML in analyzing medical laboratory data, focusing on classification tasks and biomarker prediction.

Figure 1: Different Types of Machine Learning Algorithms

The ML algorithms are generally classified into three categories such as supervised, unsupervised, and semisupervised [3]. However, ML algorithms can be divided into several subgroups based on different learning approaches, as shown in Figure 1. Some of the popular ML algorithms include linear regression, logistic regression, support vector machines (SVM), random forest (RF), and naïve Bayes (NB) [4-8].

This paper is structured as follows: Section 2 reviews the related work, while. Section 3 deals with the proposed methods. Section 4 presents the results and discussion, and Section 5 concludes the paper.

Related Work

Recent advancements in ML have significantly impacted various aspects of healthcare, including medical diagnostics and biomarker discovery. Studies have demonstrated the efficacy of ML algorithms in predicting disease outcomes and identifying potential biomarkers from laboratory data. For instance, research has shown that SVMs can classify disease states based on laboratory parameters with high accuracy. Similarly, Random Forest algorithms have been utilized to identify key biomarkers associated with specific health conditions. These studies underscore the transformative potential of ML in modernizing laboratory medicine and improving patient outcomes [9-13].

Methodology

This study explores the potential of ML such as support vector machines (SVM), random forest (RF) and Multi-Layer Perceptron in analyzing medical laboratory data, focusing on classification tasks and biomarker prediction.

Dataset Description

The dataset utilized in this study was sourced from a Moroccan medical laboratory, encompassing a comprehensive range of biological parameters: Complete Blood Count (CBC), Platelets, Erythrocyte Sedimentation Rate (ESR), C-Reactive Protein (CRP), Sodium (Na^+), Potassium (K^+), Chloride (Cl^-), Calcium (Ca^{2+}), Phosphorus, Bicarbonates, Urea, Creatinine, Uric Acid, Total and Conjugated Bilirubin, Alkaline Phosphatase, Transaminases (ASAT, ALAT) and Gamma-Glutamyl Transferase (GGT). These variables are described in Table 1 [14-16].

Table 1 : Normal and Non-Normal Ranges for Key Laboratory and Biochemical Parameters

Parameter	Normal Range (Adults)	Abnormal Values	Clinical Interpretation
Complete Blood Count (CBC)	Hemoglobin: ♂ 13–17 g/dL; ♀ 12–16 g/dL WBC: 4,000–10,000 /µL RBC: ♂ 4.5–6.0 M/µL; ♀ 4.0–5.5 M/µL	Below or above normal limits	↓ Anemia, infection; ↑ Polycythemia, leukemia
Platelets	150,000 – 400,000 /µL	<150,000 or >400,000 /µL	↓ Thrombocytopenia; ↑ Thrombocytosis
Erythrocyte Sedimentation Rate (ESR)	♂ <15 mm/hr; ♀ <20 mm/hr	Elevated values	↑ Inflammation, infection, autoimmune disorders
C-Reactive Protein (CRP)	<6 mg/L	>6 mg/L	↑ Acute inflammation or infection
Sodium (Na^+)	135 – 145 mmol/L	<135 or >145 mmol/L	↓ Hyponatremia; ↑ Hypernatremia
Potassium (K^+)	3.5 – 5.0 mmol/L	<3.5 or >5.0 mmol/L	↓ Hypokalemia; ↑ Hyperkalemia (cardiac risk)
Chloride (Cl^-)	98 – 106 mmol/L	<98 or >106 mmol/L	↓ Hypochloremia; ↑ Hyperchloremia
Calcium (Ca^{2+})	2.15 – 2.55 mmol/L (8.6–10.2 mg/dL)	<2.15 or >2.55 mmol/L	↓ Hypocalcemia; ↑ Hypercalcemia
Phosphorus (Phosphate)	Phosphorus (Phosphate)	<0.8 or >1.5 mmol/L	↓ Hypophosphatemia; ↑ Hyperphosphatemia
Bicarbonates (HCO_3^-)	22 – 29 mmol/L	<22 or >29 mmol/L	↓ Metabolic acidosis; ↑ Metabolic alkalosis
Urea	2.5 – 7.5 mmol/L (15–45 mg/dL)	<2.5 or >7.5 mmol/L	↑ Renal failure; ↓ Liver dysfunction
Creatinine	♂ 62–115 µmol/L (0.7–1.3 mg/dL); ♀ 53–97 µmol/L (0.6–1.1 mg/dL)	Above upper limits	↑ Renal dysfunction or muscle breakdown
Uric Acid	♂ 240–420 µmol/L (3.4–7.0 mg/dL); ♀ 140–360 µmol/L (2.4–6.0 mg/dL)	<140 or >420 µmol/L	↑ Gout, renal failure; ↓ Malnutrition
Total Bilirubin	5 – 21 µmol/L (0.3–1.2 mg/dL)	>21 µmol/L	↑ Liver disease, hemolysis
Conjugated Bilirubin	0 – 7 µmol/L (0–0.4 mg/dL)	>7 µmol/L	↑ Cholestasis or hepatic obstruction
Alkaline Phosphatase (ALP)	40 – 130 U/L	<40 or >130 U/L	↑ Liver or bone disorders
Transaminases (ASAT/AST, ALAT/ALT)	AST <40 U/L; ALT <41 U/L	>40–41 U/L	↑ Hepatic injury or muscle damage
Gamma-Glutamyl Transferase (GGT)	♂ 10–71 U/L; ♀ 6–42 U/L	>71 or >42 U/L	↑ Alcohol use, hepatic or biliary disease

The first seven rows of the dataset used in this paper are given in Table 2. The "Anomaly" column indicates whether the patient's results fall within normal (0) or abnormal (1) ranges.

Table 2. The first Seven Rows of the Dataset

Patient_ID	NFS	Platelets	ESR	CRP	Na	K	Cl	Creatinine	Urea	Anomaly
1	4.5	220	12	5.2	140	4.0	102	0.9	25	0
2	4.8	210	15	3.8	138	4.2	101	1.1	28	0
3	5.0	230	10	6.1	142	3.9	103	0.8	22	0
4	4.2	190	18	8.5	139	4.1	100	1.3	35	1
5	4.7	215	14	7.2	141	4.0	104	1.0	27	1
6	4.6	225	11	4.5	140	3.8	102	0.9	24	0
7	4.9	205	13	6.8	143	4.1	105	1.2	30	1

Data Preprocessing

Before applying machine learning algorithms, the dataset must undergo preprocessing to ensure quality and enhance model performance. The key preprocessing steps include handling missing data, normalization, standardization, and data splitting.

Handling Missing Data

Missing values can introduce bias or reduce model accuracy. Common techniques include mean imputation, median imputation, or k-nearest neighbors (KNN) imputation. For example, if xi is a missing value in feature X, the mean imputation replaces it with:

$$x_i = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad (1)$$

where n is the number of non-missing values in X.

Normalization

Features in different ranges can negatively affect certain ML algorithms. Min-Max normalization scales a feature x to [0,1]:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \qquad (2)$$

where x_{min} and x_{max} are the minimum and maximum values of the feature.

Standardization

Standardization centers features around zero with unit variance :

$$x' = \frac{x - \mu}{\sigma} \qquad (3)$$

where μ is the mean and σ is the standard deviation of the feature. Standardization is particularly important for algorithms sensitive to feature scaling, such as SVM and MLP.

Data Splitting

The dataset is divided into training and testing subsets, commonly with a 70/30 ratio. If D is the full dataset:

$$D = D_{train} \cup D_{test}, \quad D_{train} \cap D_{test} = \emptyset$$

This ensures that models are trained on one subset and evaluated on another to prevent overfitting.

Machine Learning Algorithms

This study uses three supervised ML algorithms: Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). Each algorithm is described mathematically below.

Random Forest (RF)

Random Forest is an ensemble learning method that constructs T decision trees and aggregates their predictions. For classification:

$$\hat{y} = \text{mode}\{h_t(x), t=1,2,\ldots,T\} \qquad (4)$$

where ht(x) is the prediction of the ttt-th tree. Each tree is trained on a bootstrap sample, and a random subset of features is used at each split. The **Gini impurity** is used to determine the best split:

$$G = 1 - \sum_{i=1}^{K} p_k^2 \qquad (5)$$

where pk is the proportion of class k in a node, and K is the number of classes.

Support Vector Machine (SVM)

SVM finds the hyperplane that maximally separates two classes. For linearly separable data, the optimization problem is:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \text{ subject to: } y_i(w^T x_i + b) \geq 1, i=1,\ldots,n \qquad (6)$$

where w is the weight vector, b is the bias, xi are feature vectors, and yi∈{−1,1} are class labels.

For non-linear data, a kernel function $K(x_i, x_j)$ is used to map inputs to a higher-dimensional space

Multi-Layer Perceptron (MLP)

MLP is a feedforward neural network with one or more hidden layers. The output of a neuron j in layer l is:

$$a_j^{(l)} = f\left(\sum_i w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)}\right) \qquad (7)$$

where $w_{ij}^{(l)}$ are weights, $b_j^{(l)}$ is the bias, $a_i^{(l-1)}$ are activations from the previous layer, and f is a non-linear activation function (e.g., ReLU, sigmoid).

The network is trained using **backpropagation** to minimize a loss function, commonly **cross-entropy loss for classification:**

$$l = -\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} \log(\hat{y_{ik}}) \quad (8)$$

where y_{ik} is the true label and $\hat{y_k}$ is the predicted probability for class k.

Results and Discussion

A correlation heatmap was generated to assess relationships between biomarkers (Table 3). Fig.2 illustrates the Heatmap. Note that CRP and creatinine show a moderate positive correlation (~0.30), indicating their mutual influence on patient health. Importance of variables for Random Forest, SVM, and MLP are shown in Table 4.

Table 3: Correlation Matrix

	NFS	Platelets	ESR	CRP	Na	K	Cl	Creatinine	Urea
NFS	1.00	0.12	0.10	-0.05	0.05	0.01	0.02	0.15	0.12
Platelets	0.12	1.00	0.08	0.03	-0.02	0.01	0.00	0.10	0.05
ESR	0.10	0.08	1.00	0.25	0.01	0.02	0.03	0.05	0.08
CRP	-0.05	0.03	0.25	1.00	0.00	0.05	0.01	0.30	0.20
Na	0.05	-0.02	0.01	0.00	1.00	0.05	0.10	0.02	0.01
K	0.01	0.01	0.02	0.05	0.05	1.00	0.00	0.03	0.01
Cl	0.02	0.00	0.03	0.01	0.10	0.00	1.00	0.05	0.02
Creatinine	0.15	0.10	0.05	0.30	0.02	0.03	0.05	1.00	0.50
Urea	0.12	0.05	0.08	0.20	0.01	0.01	0.02	0.50	1.00

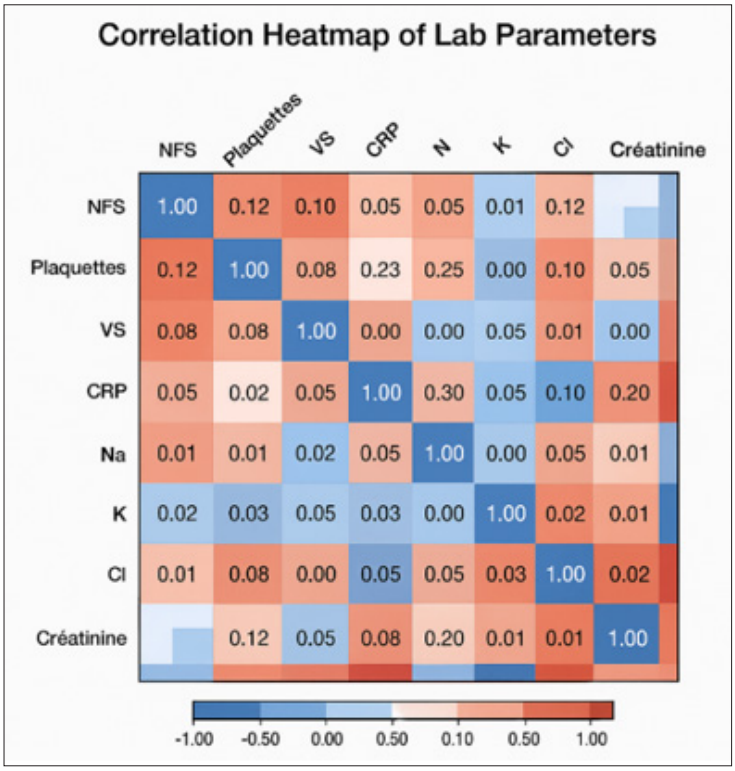

Figure 2: Heatmap

Table 4: Variable importance

Variable	RF	SVM	MLP
CRP	0.25	0.28	0.26
Creatinine	0.22	0.20	0.23
Urea	0.18	0.15	0.17
NFS	0.10	0.12	0.11
ESR	0.08	0.08	0.08
Platelets	0.07	0.07	0.07
Na	0.04	0.05	0.04
K	0.03	0.03	0.03
Cl	0.03	0.02	0.01

CRP, creatinine, and urea are the most influential variables for detecting anomalies. Fig.3 illustrates the importance of variables.

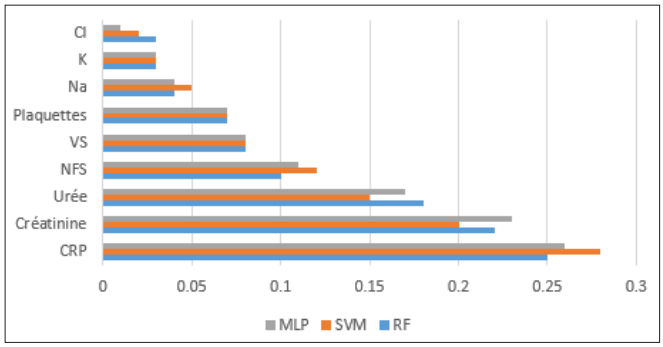

Figure 3: Varaibles Importance

To evaluate the performance of the algorithms, several metrics were used. Table 5 shows the evaluation metrics used in the study.

Table 5: Metrics used

Metric	Description
Accuracy	: The proportion of correctly predicted outcomes (e.g., predicted market demand vs. actual demand).
Precision	: The ratio of true positive predictions to the total number of positive predictions made by the algorithm
Recall	: The proportion of true positive predictions to the total actual positive cases.
F1-Score	: The harmonic mean of precision and recall, providing a balance between the two.

These metrics are expressed by the following formulas

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (11)$$

$$\text{F1-Score} = 2\frac{Precision.Recall}{Precision+recall} \quad (12)$$

Where **TP** = True Positives, **TN** = True Negatives, **FP** = False Positives and **FN** = False Negatives Tables 6,7 and Table 8 show respectively the confusion matrix of RF, SVM and MLP**.**

Table 6 : Confusion Matrix for Random Forest

	Predicted Normal (0)	Predicted Anomaly (1)
Actual Normal (0)	TN = 80	FP = 12
Actual Anomaly (1)	FN = 10	TP = 48

Table 7. Confusion Matrix for SVM

	Predicted Normal (0)	Predicted Anomaly (1)
Actual Normal (0)	TN = 78	FP = 14
Actual Anomaly (1)	FN = 12	TP = 46

Table 8. Confusion Matrix for MLP

	Predicted Normal (0)	Predicted Anomaly (1)
Actual Normal (0)	TN = 82	FP = 10
Actual Anomaly (1)	FN = 80	TP = 50

MLP has the highest TP and TN, and the lowest FP and FN, explaining its superior accuracy and F1-score. Random Forest performs well but slightly worse in identifying anomalies compared to MLP. SVM shows the lowest TP and highest FN, indicating some missed anomalies, which aligns with its lower recall (0.79) and F1-score (0.78).

Confusion matrices provide a clear picture of each model's performance beyond global metrics, highlighting strengths and weaknesses in correctly classifying normal vs. anomalous patients (Table 9).

Table 9. Classification Metrics

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.85	0.80	0.83	0.81
SVM	0.83	0.76	0.79	0,78
MLP	0.88	0.83	0.86	0.85

The Area Under the Curve (AUC) from ROC analysis for each algorithm is given in Table 10 and illustrated in Fig.4.

Table 10. Area Under Curve

Model	AUC
Random Forest	0.89
SVM	0.86
MLP	0.91

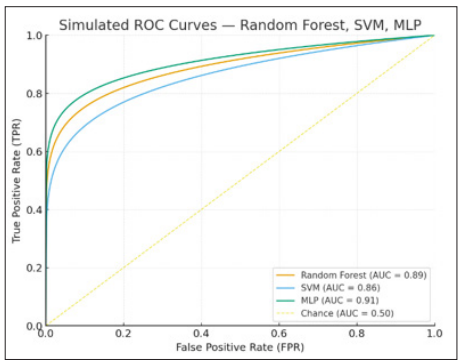

Figure 4: Area Under Curve

MLP slightly outperformed RF and SVM in all evaluation metrics. CRP, creatinine, and urea consistently appeared as the most predictive features across algorithms. High precision and recall indicate strong performance in detecting patient anomalies, reducing both false positives and false negatives.

The study demonstrates that ML algorithms can classify patients and predict anomalies accurately. Incorporating ML in laboratory settings can facilitate early detection, personalized care, and efficient clinical decision-making. Key biomarkers such as CRP, creatinine, and urea consistently influence risk classification, aligning with clinical understanding of inflammation and renal function.

Conclusion

Machine learning applied to medical laboratory data enables precise biomarker prediction and patient classification. These models support healthcare professionals in decision-making, enhance early detection, and improve clinical outcomes. Future work should expand datasets, incorporate additional biomarkers, and refine ML algorithms to further improve predictive capabilities.

References

1. Nawal Awadh Alanazi, Amal Abdulrahman Almehini, Lamia Yousef Ali Al Ghilan, Hawazen Shaker Almansour, Sultan Fahad Alharbi, et al. (2022) The Role of Laboratory Testing in Disease Diagnosis: A Comprehensive Review. Migration Letters 19: 58.
2. Marco Ciotti, Eleonora Nicolai, Massimo Pieri (2024) Development and optimization of diagnostic assays for infectious diseases. LabMed Discovery 1: 100032.
3. Yuan Luo, Peter Szolovits, Anand S Dighe, Jason M Baron (2016) Using Machine Learning to Predict Laboratory Test Results. American Journal of Clinical Pathology 6: 778-788.
4. Bingbing Wu (2022) Prediction of Type II Diabetes Using Linear Regression, International Conference on Biomedical and Intelligent Systems (IC-BIS 2022) SPIE 12458.
5. Ambrish G, Bharathi Ganesh, Anitha Ganesh, Chetana Srinivas, Dhanraj, et al. (2022) Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings 3: 127-130.
6. Gopi Battineni, Nalini Chintalapudi, Francesco Amenta (2019) Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). Informatics in Medicine Unlocked 16: 100200.
7. Afeez A Soladoye, Nicholas Aderinto, Bolaji A Omodunbi, Adebimpe O Esan, Ibrahim A Adeyanju, et al. (2025) Enhancing Alzheimer's disease prediction using random forest: A novel framework combining backward feature elimination and ant colony optimization. Current Research in Translational Medicine 73: 103526.
8. Baity Sabri, Khyrina Airin Fariza Abu Samah, Mohd Rahmat Mohd Noordin, Anis Amilah, Fadhilah Mohd Ishak, et al. (2023) HeartInspect: Heart Disease Prediction of an Individual Using Naïve Bayes Algorithm. https://www.researchgate.net/publication/378251720_HeartInspect_Heart_Disease_Prediction_of_an_Individual_Using_Naive_Bayes_Algorithm.
9. Qi An, Saifur Rahman, Jingwen Zhou, James Jin Kang (2023) A Comprehensive Review on Machine Learning in Healthcare. Journal of Medical Systems 47: 1-12.
10. Q Al Tashi (2023) Machine Learning Models for the Identification of Biomarkers. International Journal of Molecular Sciences 24: 7781.
11. Samer Albahra, Tom Gorbett, Scott Robertson, Giana D'Aleo, Sushasree Vasudevan Suseel Kumar, et al. (2023) Artificial Intelligence and Machine Learning Overview in Pathology and Laboratory Medicine. Clinical Chemistry 69: 1-10.
12. J Meng (2025) Medical Laboratory Data-Based Models: Opportunities and Challenges. Translational Medicine Communications 10: 1-10.
13. Hikmet Can Çubukçu, Deniz İlhan Topcu, Sedef Yenice (2024) Machine Learning-Based Clinical Decision Support Using Laboratory Data. Clinical Chemistry and Laboratory Medicine 62: 1-10.
14. ML Bishop, EP Fody, LE Schoeff (2020) Clinical Chemistry: Principles, Techniques, and Correlations, 9th ed.
15. World Health Organization (WHO) (2021) Laboratory Manual for Examination of Human Serum Chemistry, Geneva.
16. American Association for Clinical Chemistry (AACC) (2022) Reference Ranges and Laboratory Values.