Engineered Science
DOI: https://dx.doi.org/10.30919/es1635

Application and Evaluation of Ensemble Machine Learning Models for the Differential Diagnosis of Clinical and Hematological Syndromes in the Health Passport Software Package

Indira Uvaliyeva,[1,*] Zhenisgul Rakhmetullina,[1,*] David Borozenets,[1] Farida Amenova[2,*] and Shynar Tezekpayeva[1]

Abstract

Anemia remains a widespread clinical and public health challenge, particularly in women and children, often leading to delayed or incorrect treatment due to diagnostic complexity. While traditional diagnosis relies on manual interpretation of laboratory data, recent advances in machine learning offer promising automated alternatives. This study proposes a novel diagnostic framework integrating ensemble and neural network models trained on a comprehensive clinical dataset comprising over 30 hematological and biochemical features-including hemoglobin, red blood cell count, hematocrit, red cell distribution width, serum B12, and iron levels. The approach systematically compares baseline classifiers (SVM, Logistic Regression, Decision Trees) with advanced ensemble techniques (Random Forest, Bagging, Voting, XGBoost, LightGBM, Stacking) and integrates them into the Health Passport software package. Experimental validation shows that ensemble models significantly improve diagnostic precision, with XGBoost achieving over 98% accuracy and neural networks providing probabilistic outputs for anemia subtypes. These results confirm the efficacy of intelligent diagnostic systems in enhancing clinical decision-making and supporting personalized medicine in hematology.

Keywords: Clinical and hematological syndromes; Ensemble algorithms; Machine learning algorithms; Anemia differential diagnosis; Machine learning methods.
Received: 21 April 2025; Revised: 02 July 2025; Accepted: 04 July 2025
Article type: Original research.

1. Introduction

Anemia is a common hematological disorder characterized by a reduction in hemoglobin concentration or red blood cell (RBC) count, leading to impaired oxygen transport and systemic fatigue.[1] It is not a single disease but rather a manifestation of various underlying conditions, including nutritional deficiencies, chronic inflammation, and renal impairment.[2] The process of erythropoiesis and RBC development is dynamic and regulated throughout life, beginning in embryonic stages and persisting into adulthood, where disruptions at any stage can contribute to anemia.[3] In pediatric populations, anemia often arises from diminished RBC production due to nutritional, hormonal, or genetic causes, and is frequently associated with low reticulocyte counts.[4]

In addition to being a diagnostic concern, RBCs are increasingly being explored in translational research domains such as drug delivery, highlighting the clinical significance of their structural and functional integrity.[5] In chronic kidney disease, anemia develops early due to disrupted erythropoietin production and iron metabolism, often manifesting in normochromic and hypochromic microcytic forms.[6] Among older adults, inflammation plays a key role in anemia development, with many cases being multifactorial and associated with erythropoietin resistance and oxidative stress.[7] Globally, anemia affects nearly one-third of the

[1]*D. Serikbaev East Kazakhstan Technical University, Oskemen, 070004, Kazakhstan*
[2]*S. Amanzholov East Kazakhstan University, Oskemen, 070004, Kazakhstan*
[*]*Email:* indirauvalieva@gmail.com (I. Uvaliyeva); zhrakhmetullina@edu.ektu.kz (Z. Rakhmetullina); faramen.74@mail.ru (F. Amenova)

Table 1: Comparative overview of traditional, alternative, and AI-based methods for anemia detection.

Category	Method	Description	Strengths	Limitations
Traditional Methods	CBC with threshold rules[15,16]	Use of static cut-off values for HGB, MCV, MCH, etc.	Widely available; low cost; easy to interpret	Cannot classify anemia type accurately; ignores inter-feature relationships
	hematology analyzer[17] and peripheral smear examination[18,19]	Microscopic analysis of blood morphology	Useful for identifying abnormal shapes and sizes of RBCs	Subjective; requires expert; low reproducibility
	Ferritin[20] and B12 assays[21,22]	Biochemical assays to detect iron or B12 deficiency	Etiological specificity	High cost; limited access in resource-constrained settings
Alternative Methods	Physiological monitoring[23]	Heart rate, SpO_2, palmar/nail color	Non-invasive; can be integrated with mobile tools	Low specificity; confounding variables; not standardized
	Imaging or computer vision[24]	Analyzing sclera, conjunctiva, or nail bed images	Emerging tool; potential for remote monitoring	Requires image standardization; prone to lighting/artifact issues
Machine Learning (ML)	Logistic Regression, SVM, Decision Tree, Random Forest[25,26]	Supervised learning based on CBC data	High interpretability; rapid classification	Performance drops with class imbalance; limited to known patterns
	Ensemble methods (Bagging, Boosting, Stacking, Voting) [27,28]	Aggregated decisions from multiple weak classifiers	Higher accuracy; handles nonlinearity and heterogeneity	Requires more computational resources; complex tuning
AI/Deep Learning	Neural Networks (DNN, CNN, RNN)[29,30]	Data-driven hierarchical models for pattern recognition	Can model complex nonlinear relationships; useful for image or time-series data	Low interpretability; prone to overfitting; requires large annotated datasets
	Explainable AI (SHAP, LIME, Anchor, QLattice)[31]	Interpretable ML models; reveals feature influence	Promotes trust and clinical usability	Requires model-specific adaptation; may oversimplify deep interactions

population, with a higher burden among women and children, requiring targeted approaches to identify its various subtypes and etiologies for effective intervention.[8]

Anemia affects individuals across all age groups but demonstrates particularly high prevalence among women of reproductive age and young children. In pediatric populations, anemia is the most common hematologic abnormality and is associated with nutritional deficiencies, comorbidities, and delayed cognitive development if left untreated.[9] Studies conducted in Mozambique and Portugal confirm that children under five and pregnant women remain particularly vulnerable groups, especially in low-resource settings.[10] Large-scale epidemiological analyses across 47 low- and middle-income countries indicate that over 56% of children under five and 40% of women of reproductive age are anemic, with prevalence increasing significantly during pregnancy, especially in the second and third trimesters.[11] A global meta-analysis confirms that nearly 37% of pregnant women worldwide suffer from anemia, with particularly high rates in the third trimester and in developing regions.[12] Furthermore, pooled estimates from 197 countries show insufficient global progress toward reducing anemia prevalence, with one-third of women and over 40% of children remaining affected as of 2019.[13] Diagnosing anemia is often complex and resource-intensive due to its diverse clinical presentations and a wide array of visually identifiable symptoms. Diagnosis of anemia has historically been based on traditional methods such as a general blood test, peripheral smear examination, and biochemical analyses related to iron content.[14] These approaches, while necessary for initial screening, are limited by static thresholds, operator dependence, and price barriers, especially in resource-constrained environments. To solve these problems, alternative diagnostic methods have emerged, such as non-invasive physiological monitoring and imaging-based methods that offer more accessible but less specific solutions. To contextualize current diagnostic practices, Table 1 summarizes traditional, alternative, machine learning, and AI-based methods used for anemia detection.

An analysis of the diagnostic approaches presented in the table shows that traditional methods, despite their accessibility and standardization, have limited sensitivity in classifying subtypes of anemia and require expert interpretation. Alternative non-invasive methods, such as physiological measurements and visual analysis, demonstrate potential for primary screening, but are inferior in specificity and reproducibility. At the same time, machine learning models and ensemble algorithms have proven effective in the

Table 2: Mathematical models of the first stage of morphological classification.

Pathology indicator	Calculation rules
Hemoglobin	$m_{\mathrm{HGB}} = \begin{cases} 0, & \mathrm{HGB} \geq 125 \\ \frac{125-\mathrm{HGB}}{125-115}, & 115 < \mathrm{HGB} < 125 \\ 1, & \mathrm{HGB} \leq 115 \end{cases}$ (1) where m_{HGB} represents the level of pathology in relation to hemoglobin, and HGB denotes the hemoglobin measurement from the patient's sample
Red Blood Cell Quantity	$m_{\mathrm{RBC}} = \begin{cases} 0, & \mathrm{RBC} \geq 4.0 \times 10^{12}/l \\ \frac{4.0-\mathrm{RBC}}{4.0-3.5}, & 3.5 \times 10^{12}/l < \mathrm{RBC} < 4.0 \times 10^{12}/l \\ 1, & \mathrm{RBC} \leq 3.5 \times 10^{12}/l \end{cases}$ (2) where m_{RBC} represents the level of pathology determined by the quantity of red blood cells, RBC indicates the result obtained from the patient's sample
Hematocrit	$m_{\mathrm{HCT}} = \begin{cases} 0, & \mathrm{HCT} \geq 0.38 \\ \frac{0.38-\mathrm{HCT}}{0.38-0.30}, & 0{,}3 < \mathrm{HCT} < 0.38 \\ 1, & \mathrm{HCT} \leq 0.3 \end{cases}$ (3) where m_{HCT} represents the level of pathology determined by hematocrit, and HCT indicates the result obtained from the patient's sample.
Mean corpuscular hemoglobin	$m_{\mathrm{MCH}} = \begin{cases} 1, & \mathrm{MCH} \leq 18.5\ \mathrm{pg} \\ \frac{27-\mathrm{MCH}}{27-18.5}, & 18{,}5\ \mathrm{pg} < \mathrm{MCH} < 27\ \mathrm{pg} \\ 0, & 27\ \mathrm{pg} < \mathrm{MCH} < 34\ \mathrm{pg} \\ \frac{MCH-34}{36.4-34}, & 34\ \mathrm{pg} < \mathrm{MCH} < 36.4\ \mathrm{pg} \\ 1, & \mathrm{MCH} \geq 36.4\ \mathrm{pg} \end{cases}$ (4) where m_{MCH} represents the level of pathology based on the mean corpuscular hemoglobin, MCH indicates the result of the patient's sample.
Mean corpuscular volume	$m_{\mathrm{MCV}} = \begin{cases} 1, & \mathrm{MCV} \leq 64\ fl \\ \frac{80-\mathrm{MCV}}{80-64}, & 64\ fl < \mathrm{MCV} < 80\ fl \\ 0, & 80\ fl < \mathrm{MCV} < 95\ fl \\ \frac{\mathrm{MCV}-95}{129-95}, & 95\ fl < \mathrm{MCV} < 129\ fl \\ 1, & \mathrm{MCV} \geq 129\ fl \end{cases}$ (5) where m_{MCV} represents the level of pathology determined by the mean corpuscular volume, and MCV indicates the result obtained from the patient's sample.
Mean corpuscular hemoglobin concentration	$m_{\mathrm{MCHC}} = \begin{cases} 0, & \mathrm{MCHC} \geq 32\ g/\mathrm{dl} \\ \frac{32-\mathrm{MCHC}}{32-28}, & 28\ g/\mathrm{dl} < \mathrm{MCHC} < 32\ g/\mathrm{dl} \\ 1, & \mathrm{MCHC} \leq 28\ g/\mathrm{dl} \end{cases}$ (6) where m_{MCHC} represents the level of pathology determined by the mean corpuscular hemoglobin concentration, $MCHC$ indicates the result obtained from the patient's sample.

The rule for determining the likeness of anemia M is given by Eq. (7):

$$M = \begin{cases} 0, & \mathrm{MCHC} \leq 0.2 \\ \sum M, & 0{,}2 < \mathrm{MCHC} < 0.5 \\ 1, & \mathrm{MCHC} \geq 0.5 \end{cases} \quad (7)$$

where M represents the overall degree of association with the anemic syndrome for all six indicators, as defined in Eq. (8):

$$\sum M = m_{HGB} * 0.5 + m_{HCT} * 0.1 + m_{MCHC} * 0.1 + m_{MCH} * 0.1 + m_{MCV} * 0.1 + m_{RBC} * 0.1 \quad (8)$$

automated interpretation of hematological data, providing high accuracy and stability in the diagnosis of anemia. Deep neural network architectures have the ability to identify complex nonlinear dependencies, but their practical application is hampered by insufficient interpretability. The use of explicable AI tools such as SHAP and LIME partially compensates for this shortcoming and increases the trust of clinicians. Thus, combined approaches combining ensemble methods, neural networks, and interpretation tools represent the most promising direction for the development of intelligent anemia diagnostic systems capable of providing both accuracy and clinical validity of conclusions. The purpose of this study is to develop, programmatically implement and experimentally evaluate an intelligent software package as part of the digital Health Passport platform for the automated diagnosis of clinical and hematological syndromes, primarily anemia of various morphological and etiological nature. As part of the work, the task was set to build a multiclass classification model

Table 3: Mathematical models of the second stage of morphological classification.

Cytometric type	Calculation rules
Microcytic	$M_{micro} = \begin{cases} 0, & m_{micro} < 0.5 \\ 1, & m_{micro} \geq 0.5 \end{cases}$ (9)
	where m_{micro} is the overall criterion for anemia's classification as microcytic (10):
	$m_{micro} = m_{MCH} * 0.5 + m_{MCV} * 0.5$ (10)
	$m_{\mathrm{MCH}} = \begin{cases} 0, & \mathrm{MCH} \geq 27\ pg \\ \frac{27-\mathrm{MCH}}{27-18.5}, & 18{,}5\ pg < \mathrm{MCH} < 27\ pg \\ 1, & \mathrm{MCH} \leq 18.5\ pg \end{cases}$ (11)
	$m_{\mathrm{MCV}} = \begin{cases} 0, & \mathrm{MCV} \geq 80\ fl \\ \frac{80-\mathrm{MCV}}{80-64}, & 64\ fl < \mathrm{MCV} < 80\ fl \\ 1, & \mathrm{MCV} \leq 64\ fl \end{cases}$ (12)
	$M_{macro} = \begin{cases} 0, & m_{macro} < 0.5 \\ 1, & m_{macro} \geq 0.5 \end{cases}$ (13)
	where m_{macro} is the overall criterion for anemia's classification as macrocytic (14):
	$m_{macro} = m_{MCH} * 0.5 + m_{MCV} * 0.5$ (14)
Macrocytic	$m_{\mathrm{MCH}} = \begin{cases} 0, & \mathrm{MCH} \leq 34\ pg \\ \frac{MCH-34}{36.4-34}, & 34\ pg < \mathrm{MCH} < 36.4\ pg \\ 1, & \mathrm{MCH} \geq 36.4\ pg \end{cases}$ (15)
	$m_{\mathrm{MCV}} = \begin{cases} 0, & \mathrm{MCV} \leq 95\ fl \\ \frac{\mathrm{MCV}-95}{129-95}, & 95\ fl < \mathrm{MCV} < 129\ fl \\ 1, & \mathrm{MCV} \geq 129\ fl \end{cases}$ (16)
	$M_{norm} = \begin{cases} 0, & m_{norm} < 0.5 \\ 1, & m_{norm} \geq 0.5 \end{cases}$ (17)
	where m_{norm} is the overall criterion for anemia's classification as normocytic (18):
	$m_{norm} = m_{MCH} * 0.5 + m_{MCV} * 0.5$ (18)
Normocytic	$m_{\mathrm{MCH}} = \begin{cases} 0, & \mathrm{MCH} \leq 18.5\ \mathrm{p}g \\ \frac{\mathrm{MCH}-18.5}{27-18.5}, & 18{,}5\ \mathrm{p}g < \mathrm{MCH} < 27\ \mathrm{p}g \\ 1, & 27\ \mathrm{p}g < \mathrm{MCH} < 34\ \mathrm{p}g \\ \frac{36.4-\mathrm{MCH}}{36{,}4-34}, & 34\ \mathrm{p}g < \mathrm{MCH} < 38\ \mathrm{p}g \\ 0, & \mathrm{MCH} \geq 38\ \mathrm{p}g \end{cases}$ (19)
	$m_{\mathrm{MCV}} = \begin{cases} 0, & \mathrm{MCV} \leq 80\ fl \\ \frac{\mathrm{MCV}-80}{90-80}, & 80\ fl < \mathrm{MCV} < 90\ fl \\ 1, & 90\ fl < \mathrm{MCV} < 95\ fl \\ \frac{100-\mathrm{MCV}}{100-95}, & 95\ fl < \mathrm{MCV} < 100\ fl \\ 0, & \mathrm{MCV} \geq 100\ fl \end{cases}$ (20)

implemented using ensemble machine learning algorithms and interpreted neural networks to improve the accuracy, reproducibility and clinical significance of diagnostic solutions.

Although the complete blood count (CBC) remains a widely used and cost-effective tool for initial screening for anemia, it does not provide sufficient information to accurately identify the specific type or cause of anemia. For example, iron deficiency anemia requires not only hemoglobin assessment but also assessment of iron stores, red blood cell morphology, and plasma ferritin levels to confirm the diagnosis.[32] Moreover, hemoglobin-based screening is prone to both false-positive and false-negative results, especially in the presence of inflammation, infection, or genetic hemoglobinopathies that may mask or mimic anemia.[33] The multifactorial nature of anemia, ranging from nutritional deficiencies and chronic diseases to genetic and environmental influences, requires a broader diagnostic approach that goes beyond standard hematological indices.[34] Consequently, there is growing interest in developing early detection strategies, including digital solutions and clinical decision support systems, particularly for vulnerable populations such as adolescents and women of reproductive age.[35]

Traditional diagnostic approaches for anemia typically rely on fixed threshold values of hemoglobin or other basic hematological parameters. However, such methods often lack sensitivity to subtle clinical nuances and may lead to diagnostic inaccuracies, particularly when differentiating

Table 4: Groups of signs of clinical and hematological syndromes.

Feature group	Signs
General demographic	Age (number)
	Sex (boolean)
Standard parameters of a complete blood count, which are directly related to the diagnosis of anemia	HGB, HCT, MCV, MCH, MCHC, RBC, RDW, PLT/mm3, TLC (numbers)
General anomaly indicators	Mhgb, Mhct, Mmchc, Mmch, Mmcv, Mrbc (numbers in [0;1] range)
Anomaly indicators for cytometric type classification	MCV_micro, MCV_macro, MCH_micro, MCH_macro (numbers in [0;1] range)
Affiliation values	siM (number in [0;1] range) - anemia presence
	MMicro, MMacro, MNorm (numbers in [0;1] range) - cytometric types
	Mmahz, Mmzhda, Mmb12 (numbers in [0;1] range) - subtypes of anemia (*e.g.*, iron deficiency, B12 deficiency)
Biochemical indicators	Fer (number) - ferritin
	B12 (number)
Target variables	Result - presence of anemia (boolean)
	Microcyt, Macrocyt, Normal - anemia cytometric types (booleans)
	Mahz, Mzhda, Mb12 - anemia characters (Booleans)

between morphologically or etiologically similar conditions such as iron deficiency anemia (IDA) and anemia of chronic disease (ACD).[33] For example, in inflammatory contexts, serum ferritin levels can be elevated independently of iron stores, thereby masking iron deficiency and complicating diagnosis.[38] Recent studies have emphasized that transferrin saturation (TSAT) may offer a more reliable diagnostic criterion under such conditions, especially in patients with heart failure or chronic kidney disease.[38] Additionally, novel biomarkers such as hepcidin and reticulocyte hemoglobin equivalent (RetHe) have demonstrated high diagnostic utility in distinguishing IDA from ACD, even in cases of mixed etiology.[37] These findings highlight the limitations of traditional static thresholds and support the need for more dynamic, context-sensitive diagnostic strategies.[36]

Accurate classification of anemia subtypes often requires comprehensive laboratory assessments, including biochemical and morphological parameters, which may be unavailable in low-resource settings due to the cost and infrastructure demands of advanced diagnostic equipment. To address these challenges, recent studies have explored the integration of machine learning (ML) and data mining techniques into diagnostic workflows as a scalable and cost-efficient alternative. ML has proven effective in analyzing large clinical datasets and uncovering complex patterns that may not be apparent through traditional analysis.[39] By leveraging classification algorithms such as decision trees, support vector machines, and random forests, predictive systems can be trained to detect multiple disease types-including hematological disorders-with high accuracy.[40]

In clinical settings, ML has been successfully applied to predict outcomes for conditions such as chronic kidney disease, diabetes, and cardiovascular diseases, often outperforming conventional diagnostic methods.[41] These approaches utilize a combination of structured clinical data and advanced algorithms to identify disease-specific signatures. Furthermore, the integration of ML in healthcare analytics has enabled early detection, reduced diagnostic delays, and optimized treatment planning, especially when dealing with overlapping or ambiguous clinical presentations.[42]

Among the diverse machine learning approaches, classification algorithms have demonstrated particularly strong capabilities in analyzing structured clinical datasets such as CBC parameters, aiding in early and differential diagnosis. In perinatology and neonatal care, the integration of machine learning tools with multiomics and CBC data has enabled the construction of predictive models capable of identifying at-risk patients before the onset of overt symptoms, thus supporting precision medicine initiatives.[43] In laboratory hematology, machine learning has been used to automate and improve the accuracy of tasks such as cell classification, morphology-based diagnostics, and parameter standardization, which traditionally relied on manual interpretation.[44]

Deep learning models, such as enhanced convolutional neural networks (CNNs), have been effectively applied to image-based blood cell analysis, achieving classification accuracies exceeding 95% in distinguishing normal and abnormal hematological images, thus offering viable support for clinical hematology laboratories. Furthermore, ML algorithms such as Lasso and Ridge regression have been used

to predict anemia using structured CBC parameters such as hemoglobin, hematocrit, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH), demonstrating the feasibility of automated diagnostic pathways even in resource-limited settings or physician variability.[46]

Advances in machine learning technologies have facilitated the development of intelligent decision support systems that can accurately interpret complex blood test data and support automated classification of anemia and its subtypes. For example, Vohra et al. demonstrated the effectiveness of multi-class classification models such as multilayer perceptrons to distinguish between mild, moderate, and severe anemia using CBC data, particularly emphasizing early prediction for preventive intervention.[47] Expanding the scope of traditional blood testing, Zhang et al. proposed a deep learning system that uses facial image recognition to predict anemia in an emergency setting, highlighting the clinical feasibility of real-time non-invasive diagnosis.[48] Furthermore, Kumar et al. presented a computer vision approach combining machine learning with conjunctival image analysis, providing a low-cost non-invasive alternative for hemoglobin level assessment that can be used in resource-limited or rural settings.[49]

Numerous studies have demonstrated the high efficacy of ML models in diagnosing anemia. For example, research based on data from the Department of Clinical Pathology and Laboratory Medicine at Gadjah Mada University (Indonesia) developed a model using the Extreme Learning Machine (ELM) algorithm to classify four anemia types. Validation on 190 cases yielded impressive results: accuracy (99.21%), sensitivity (98.44%), precision (99.30%), and F1-score (0.98), underscoring the potential of ML in differential diagnosis.[25] Another recent study conducted at Kasturba Medical College, Manipal, India, applied various Explainable AI tools-such as SHAP, LIME, Eli5, Qlattice, and Anchor-to interpret model predictions based on hematological traits. The most informative features identified included platelet count (PLT), plateletcrit (PCT), MCV, platelet distribution width (PDW), hemoglobin (HGB), absolute lymphocyte count (ABS LYMP), white blood cell count (WBC), MCH, and mean corpuscular hemoglobin concentration (MCHC).[33] Deotale T. and Saha S. (2025)[29] conducted a comprehensive comparison of ML models, including support vector machines (SVM), Random Forest, logistic regression, and several deep learning architectures. Model performance was evaluated based on metrics such as precision, recall, latency, cost-effectiveness, and scalability. The authors also introduced a composite metric-the Aggregated HB and Anemia Rating (HBADR)-to assess model performance holistically. Their findings demonstrate the potential of ML for optimizing diagnostic workflows and integrating these tools into electronic medical records for remote hemoglobin monitoring and real-time clinical decision support. Dhakal *et al.*[26] focused on anemia prediction in children under five using CBC data collected from 700 clinical records at Kanti Children's Hospital, Nepal. After preprocessing and balancing the dataset, various ML algorithms were evaluated. The Random Forest algorithm achieved the highest accuracy (98.4%). The authors further improved classification by employing feature selection techniques and ensemble approaches such as voting, stacking, bagging, and boosting.

A recent study[36] applied ML models to identify iron deficiency anemia using the NHANES dataset (USA), comprising over 19,000 records. The model achieved a PR AUC of 0.87 and sensitivity scores of 0.98 (core dataset) and 0.89 (Kenya validation set), indicating strong generalizability across populations. In another investigation.,[28] researchers applied ML algorithms-including logistic regression, Cat Boost, XGBoost, decision trees, and k-nearest neighbors (kNN)-to predict anemia-related conditions using a heterogeneous hematological dataset. XGBoost and stacked classifier models delivered the best results, achieving 99% accuracy, sensitivity, and recall. Thakur and Mishra[30] explored the use of ML for detecting signs of iron deficiency anemia using alternative biomarkers such as physiological metrics and visual characteristics. The study employed exploratory data analysis (EDA) to uncover potential correlations between these indicators and iron levels. Rane *et al.*[50] provided a comprehensive review of current research in ML-based anemia diagnostics. The review discussed methods for predicting anemia prevalence, assessing severity, and classifying anemia types among diverse populations, particularly children aged 9 to 18 months and women of reproductive age. The authors emphasized the value of automated diagnostic systems in improving healthcare delivery for vulnerable groups. Ensemble algorithms such as Gradient Boosting, AdaBoost, XGBoost, and LightGBM have demonstrated high accuracy and robustness, particularly in imbalanced clinical datasets. For instance, one stuy evaluated nine ML algorithms for predicting chronic kidney disease using a publicly available Kaggle dataset (400 records, 14 features). Among these, the LightGBM model achieved the highest accuracy at 99.00%.[51]

Despite these advancements, much of the existing research focuses on binary anemia classification (anemic vs. non-anemic). The more nuanced task of classifying anemia by morphological subtype-such as microcytic, normocytic, and macrocytic-or by etiology (*e.g.*, iron deficiency, B12 deficiency) remains underexplored. This highlights the

Table 5: Fragment of the initial dataset, showcasing laboratory data and clinical indicators for anemia.

Age	HGB	HGB	MCH	MCHC	MCV	HCT	RBC	Mhgb	Mhct	Mahz	Mzhda	Mb12	Sex	RDW	PLT/mm3	TLC
28	96	9.6	17.0	28.2	60.1	0.340	5.66	1.0	0.5000	0.10	0.0	0.000	f	20.0	128.3	11.10
41	138	13.8	28.9	31.0	93.1	0.445	4.78	0.0	0.0000	1.00	0.0	0.033	f	13.0	419.0	7.02
40	134	13.4	28.8	32.2	89.5	0.416	4.65	0.0	0.0000	0.00	1.0	0.444	f	13.0	325.0	8.09
76	113	11.3	26.7	30.8	86.6	0.367	4.24	1.0	0.1625	0.75	0.0	0.613	f	14.9	264.0	13.41
20	115	11.5	27.8	31.2	89.1	0.369	4.14	1.0	0.1375	0.80	0.0	0.000	f	13.2	196.0	4.75

relevance of the current study, which aims to address this gap through multi-label classification using a diverse set of models and ensemble techniques. The central hypothesis of the study is that ML methods-particularly ensemble boosting techniques and neural networks-can automate the accurate and sensitive classification of anemia subtypes based on CBC parameters. Among the approaches tested, gradient boosting algorithms (*e.g.*, LightGBM) are hypothesized to yield superior results due to their ability to model complex, nonlinear relationships and handle class imbalance effectively. Methodologically, the study involves the application of modern ML algorithms to laboratory blood data. An initial EDA was conducted, including data cleaning, visualization, and correlation analysis. The classification models tested include basic models (Logistic Regression, SVM, Decision Tree), ensemble methods (Random Forest, Voting, Bagging, Boosting, Stacking), and neural networks with hyperparameter tuning via Optuna. Model performance was evaluated using metrics such as accuracy, precision, recall, standard deviation (STD), and F1-score, with K-fold cross-validation ensuring robustness.

Despite recent advances in ML-based anemia diagnostics, the literature reveals several key knowledge gaps. Most existing studies focus on binary classification, while clinically relevant multi-label classification across morphological and etiological subtypes remains largely underexplored. Additionally, few models offer explainable outputs or demonstrate seamless integration into clinical software environments. The lack of external validation and reliance on narrow feature sets further limit the generalizability and robustness of current approaches. This study addresses these gaps by proposing an interpretable, modular, multi-label classification framework designed to operate on real-world hematological data within a unified diagnostic platform.

2. Materials and methods

2.1 Description of the existing medical algorithm for morphological classification of clinical and hematological syndromes

This structured, algorithmic framework forms the basis of the proposed Health Passport platform and ensures consistency in diagnostic reasoning while maintaining alignment with clinical practice guidelines.

A diagnostic system for identifying clinical hematological syndromes is structured around a multi-step morphological classification algorithm that serves as the basis for formulating a diagnosis, storing information, and selecting appropriate treatment strategies. This approach builds on established diagnostic frameworks that integrate rule-based logic expert systems and evidence-based laboratory parameters. For example, the study[52] emphasizes automated analysis of blood biochemistry parameters to mimic expert reasoning to improve diagnostic accuracy and track abnormalities. In research by Z. Hevessy *et al.*[53] extend this concept by formalizing a comprehensive diagnostic algorithm that helps clinical pathologists evaluate common and rare causes of anemia through a systematic, test-based workflow using standard hematological indices. Furthermore, research by Tvedten[54] highlight the fundamental role of the complete blood count and blood smear evaluation in the diagnosis of anemia, illustrating how morphological and biochemical criteria are used to differentiate regenerative from non-regenerative anemia and to identify underlying systemic conditions. This structured algorithmic framework forms the basis of the proposed health passport platform and ensures consistency in diagnostic reasoning while maintaining compliance with clinical practice recommendations.

The first stage of diagnosis is to identify the presence and severity of a clinical hematological syndrome. Valuation is carried out according to the following six rules:

- assessment of the pathology degree based on the level of hemoglobin;
- assessment of the pathology by number of red blood cells;
- determination of the pathology degree based on the hematocrit value;
- pathology analysis by the average hemoglobin content in the erythrocyte;
- detection of pathology by the average volume of red blood cells;

- assessment of pathology by the average concentration of hemoglobin in the erythrocyte;
- The overall probability degree of anemic syndrome is calculated based on a combination of six indicators.

Before presenting the mathematical formulations, we briefly define the core hematological parameters used in the classification models:

- HGB (Hemoglobin, g/L) - a protein in red blood cells responsible for oxygen transport, low levels may indicate anemia (Eq. (1));
- RBC (Red Blood Cell Count, $\times 10^{12}$/L) -the total number of red blood cells in a volume of blood; essential for evaluating blood's oxygen-carrying capacity (Eq. (2));
- HCT (Hematocrit, %) - the proportion of blood volume occupied by red blood cells; reduced in most types of anemia (Eq. (3));
- MCV (Mean Corpuscular Volume, fL) - the average volume of individual red blood cells; used to classify anemia as microcytic, normocytic, or macrocytic (Eq. (4));
- MCH (Mean Corpuscular Hemoglobin, pg) - the average amount of hemoglobin per red blood cell (Eq. (5));
- MCHC (Mean Corpuscular Hemoglobin Concentration, g/dL) - the concentration of hemoglobin in a given volume of packed red blood cells (Eq. (6)).

These indicators form the input parameters for the morphological classification rules (1)-(8), as outlined in Table 2.

At the second stage, the type of anemia is identified according to morphological features, using the affiliation function for each type:

- Microcytic anemia: equations Eq. (9) - Eq. (12), where MCH and MCV are analyzed, disregarding cases with values above normal.
- Macrocytic anemia: equations Eq. (13) - Eq. (16), where MCH and MCV are analyzed, disregarding cases with values below normal.
- Normocytic anemia: equations Eq. (17) - Eq. (20), where MCH and MCV are analyzed in the same manner as defined in Eq. (4) - Eq. (5), considering all cases.

Mathematical models of the classification's second stage are presented in Table 3.

The third stage serves to confirm and clarify the type of anemia based on biochemical data, in particular, ferritin and vitamin B12 levels:

- Diagnosis of anemia in chronic diseases (normocytic form) - by ferritin level;
- Detection of iron deficiency anemia (micro- and normocytic forms) -by ferritin level;
- Diagnosis of anemia associated with vitamin B12 deficiency (macrocytic form) - by B12 vitamin content.

Thus, the presented calculations form the basis of the rules of the system designed to detect clinical and hematological syndromes, especially anemias.

The developed rule bases will be the methodological and algorithmic basis for the computational and analytical modules of the diagnostic software of the Health Passport web portal. The results of the morphological algorithm for the classification of clinical hematological syndromes will also be used as a complement to the machine learning algorithms.

2.2 Description of the initial data

2.2.1 Data source and preprocessing

In this study, we utilized publicly available clinical data from the Medical Information Mart for Intensive Care (MIMIC-IV)[55] database, a widely used resource that includes de-identified health-related data associated with over 40,000 patients admitted to critical care units at the Beth Israel Deaconess Medical Center between 2008 and 2019. Access to the dataset was granted through PhysioNet,[56] following successful completion of required training in data privacy and ethics.

From the MIMIC-IV database, we extracted structured laboratory data from adult patients (≥18 years) with CBC and anemia-related biochemical markers, including: Hematological parameters; Biochemical indicators; Demographic data; Diagnostic labels. Only records with complete CBC panels and at least one biochemical marker (Ferritin or B12) were included. This filtering yielded a dataset of 3,516 patients, balanced across different morphological anemia subtypes.

Since MIMIC contains some incomplete records, missing laboratory values were imputed using median values stratified by age group and sex. No patient was included if more than 30% of key features were missing.

2.2.2 Feature selection rationale and clinical justification

The selection of features such as HGB, RBC, and RDW was grounded in their well-established clinical relevance to anemia diagnostics:

- HGB is the primary diagnostic marker for anemia, as defined by WHO. Low HGB directly reflects the oxygen-carrying capacity of blood and is used universally for anemia screening and grading.
- RBC provides insight into erythropoiesis. In conditions like, RBC can be low despite relatively preserved hemoglobin levels, helping differentiate early-stage microcytic anemia from normocytic patterns.
- RDW measures anisocytosis - the variation in red blood.

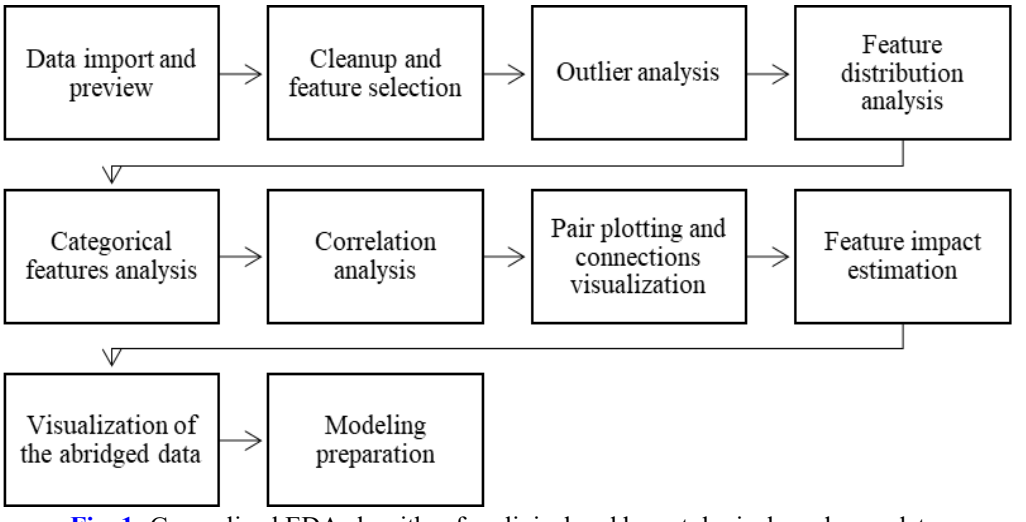

Fig. 1: Generalized EDA algorithm for clinical and hematological syndrome data.

cell size, which is critical in differentiating iron deficiency anemia (elevated RDW) from thalassemia trait (typically normal RDW), and in distinguishing mixed anemias Additional hematological indices such as MCV, MCH, and MCHC were included for their role in classifying anemia by morphological subtype (microcytic, normocytic, macrocytic), which has direct therapeutic implications.

Biochemical markers such as ferritin and vitamin B12 were used to confirm etiology. These features collectively support a clinically interpretable, mechanistically grounded, and diagnostically actionable framework.

Thus, the chosen features were not only statistically relevant (as shown in the correlation matrix and feature importance graphs), but also clinically interpretable, which enhances both the diagnostic validity and potential for integration into real-world decision-support systems.

2.2.3 The structure of the initial data

A medical sample with laboratory blood values of patients, including more than 30 traits, such as HGB, RBC, HCT, RDW, vitamin B12 and iron levels, was used as initial data. The groups of signs of clinical and hematological syndromes formed on the basis of the medical algorithm are listed in Table 4.

Note that AHZ and ZhDA acronyms mean “anemia in chronic diseases” and iron deficiency anemia” in this study’s native language.

Table 5 shows a fragment of the original dataset containing laboratory and clinical blood parameters of patients used in the study. The table shows the numerical values of such traits as age (Age), hemoglobin (HGB and HGB(adj.), since the methodic has different unit of measurement from the source dataset and needed to be adjusted), RBC, HCT, as well as derived signs associated with the diagnosis of anemia and its subtypes, for example, Mhgb, Mahz, Mb12, Mmzhda.

Some of the features are presented as binary indicators (0/1) reflecting the presence of corresponding states (*e.g.*, Sex, Mmb12, Mmahz), while others are continuous values obtained as a result of calculations or scaling (*e.g.*, Mhgb, Mb12).

This dataset formed the basis for the next stage of the analysis, which included training machine learning models to classify anemia and its clinical forms.

While the dataset includes fundamental hematological and biochemical parameters (*e.g.*, HGB, HCT, RBC, B12, Ferritin), the current feature set can be further enriched by incorporating additional clinical variables, such as:

- Patient comorbidities (*e.g.*, chronic kidney disease, autoimmune conditions);
- Nutritional status;
- Medication history (*e.g.*, iron or B12 supplementation);
- Menstrual history or pregnancy status (for women);
- Socioeconomic and geographic data.

These attributes are highly relevant in clinical practice and can contribute to more accurate context-aware classification. Future work will aim to integrate such variables either via electronic health records or structured medical history forms to enhance model specificity and real-world applicability.

2.3 Exploratory data analysis

Since MIMIC contains some incomplete records, missing laboratory values were imputed using median values stratified by age group and sex. No patient was included if more than 30% of key features were missing. Outliers were identified using the interquartile range method and verified against established clinical reference ranges. Implausible values were excluded.

To investigate the distribution of classes (*e.g.*, how many cases of anemia for each normal case), understand the class

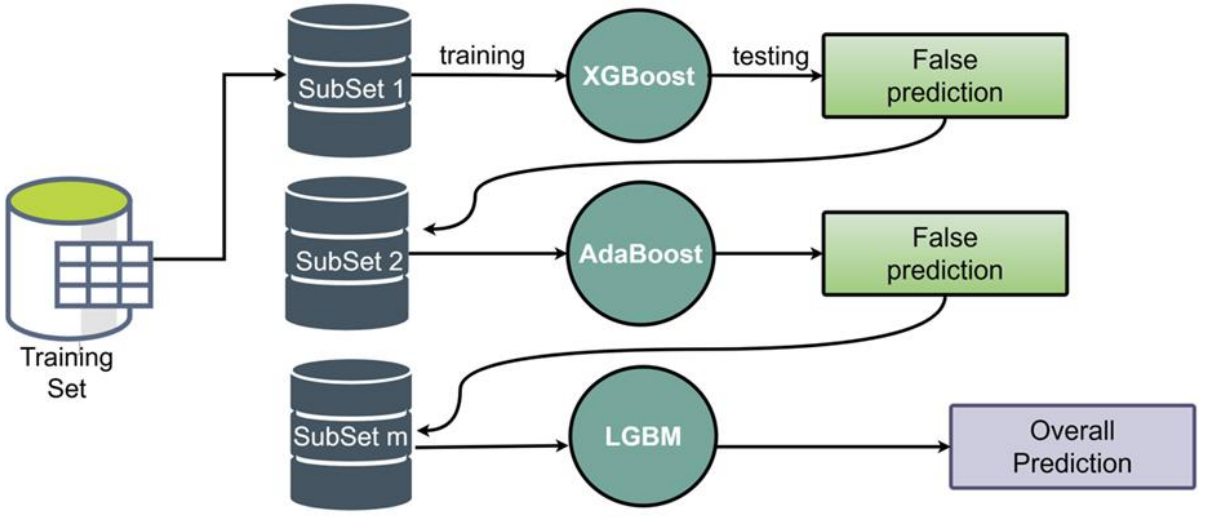

Fig. 2: Graphical model of the boosting ensemble algorithm.

imbalance that needs to be corrected and to identify informative features and correlations between them, an EDA was conducted.

EDA made it possible to detect correlated features (*e.g.*, MCH, MCHC, MCV) and uninformative or derived variables that can distort the model. This stage was also an important step before machine learning, which identified features and target variables, as well as assessed the need for scaling, transformation, and balancing of classes.[57] EDA allows to understand in advance what factors affect the outcome and which variables are potentially important in explaining the behavior of the model. This is especially important in medicine and biology, where interpretability is often more important than accuracy.[58]

A generalized EDA algorithm for these clinical and hematological syndromes is presented in Fig. 1

Fig. 1 illustrates the step-by-step EDA pipeline used in this study. The process begins with data import and screening, followed by cleaning and feature selection to remove incomplete, irrelevant, or noisy records. Outlier analysis is performed using both statistical methods (*e.g.*, IQR) and clinical validation to ensure medical validity. Feature distribution and categorical analysis of the data help identify asymmetries and dominant categories. Correlation analysis and pairwise plotting allow the detection of multicollinearity and feature dependencies, which is necessary for selecting meaningful inputs for machine learning. Feature influence assessment guides the identification of key clinical markers. Finally, preparation for modeling includes dimensionality reduction, normalization when necessary, and handling class imbalance using methods such as SMOTE. This structured EDA approach ensures data quality, supports interpretability, and improves the reliability of the subsequent modeling process in hematology diagnostics.

2.4 Base ML

To objectively compare the effectiveness of various algorithms for classifying anemia and its subtypes, a function was implemented that automatically conducts 10-fold cross-validation, training, and calculation of key quality metrics. The mathematical formalization of this function is given in Eq. (21) - Eq. (24):

$$M = \{M_1, M_2, \ldots M_n\} \tag{21}$$

$$T = \{T_1, T_2, \ldots T_k\} \tag{22}$$

$$S = \{s_1, s_2, \ldots s_r\} \tag{23}$$

$$R^{(i)} \in \mathbb{R}^{n*k} \tag{24}$$

where: M is a set of models (*e.g.* Random Forest, XGBoost, SVM, etc.)); T is a set of classification targets (*e.g.*, anemia, microcytic type, etc.); S is a set of metrics (*e.g.* accuracy, precision, etc.); $R^{(i)}$ - evaluation results matrix by metric s_i; $R_{mj}^{(i)}$ is the value of the metric s_i for the model M_m on the task T_j.

This function, for each metric $s_i \in$ S builds the $R^{(i)}$ matrix as defined in Eq. (5), in which the rows correspond to models and the columns correspond to classification tasks, as shown in Eq. (25):

$$R^{(i)} = \begin{bmatrix} R_{11}^{(i)} & \ldots & R_{1k}^{(i)} \\ \ldots & \ldots & \ldots \\ R_{n1}^{(i)} & \ldots & R_{nk}^{(i)} \end{bmatrix} \tag{25}$$

This function is an automated method for comparing machine learning models by several quality metrics (accuracy, precision, recall, F1-score) for each of the target labels (anemia

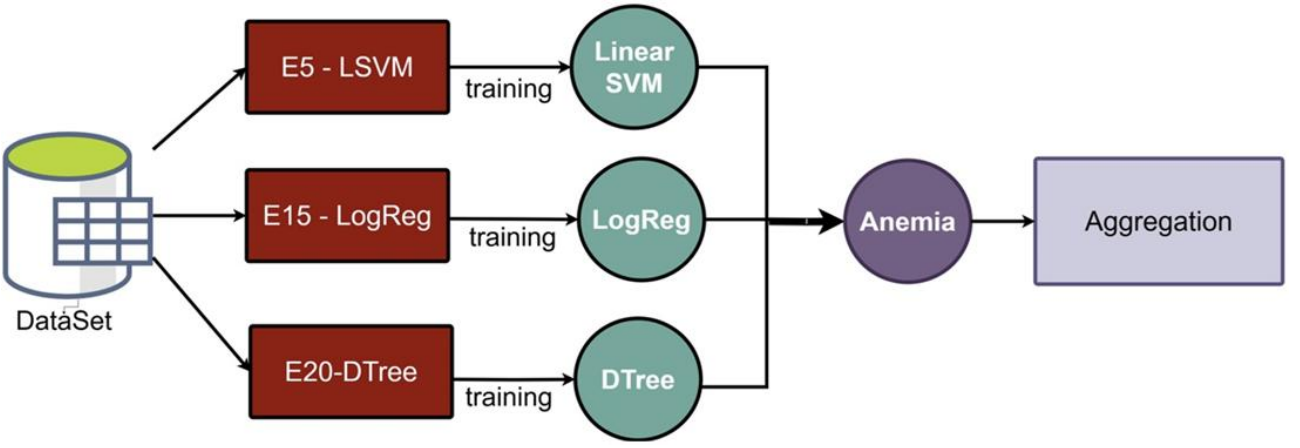

Fig. 3: Graphical model of the bagging ensemble algorithm.

and its 6 subtypes: microcytic anemia, macrocytic anemia, normocytic anemia, iron deficiency anemia, chronic anemia, B12-deficiency anemia).

2.5 Boosting

Boosting algorithms were used for automated diagnosis of anemia and its subtypes based on laboratory blood parameters. Boosting algorithms made it possible to build an ensemble of weak classifiers, where each subsequent model focused on the errors of the previous ones, thereby improving the final accuracy.[59]

In this study, the following main task classes (anemia and its forms) were identified for the task of multilabel classification of anemia and its subtypes using the boosting algorithm: Result, Microcyt, Macrocyt, Normal, Mmahz, Mmzhda. The metrics are cross-validation averages (CV Mean, STD) and key quality indicators (Accuracy, Precision, Recall, F1 Score). Three popular approaches were used as models: AdaBoost, XGBoost, and LightGBM. These models were trained on traits that included hemoglobin levels, red blood cells, hematocrit, RDW, vitamin B12 and iron levels.

The AdaBoost algorithm was chosen because it works well with a small amount of data and is resistant to overtraining on noisy data, and is also suitable for binary classification (*e.g.*, anemia: yes/no)[60] This algorithm for N objects and M iterations at the t-th step trains the weak classifier ht(x) and assigns it a weight, as defined in Eq. (26):

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \quad (26)$$

where ε_t is the error on weighted sampling.

Next, the algorithm selects a final solution based on Eq. (27)

$$H(x) = \text{sign}\left(\sum_{t=1}^{M} \alpha_t h_t(x)\right) \quad (27)$$

where M is the total number of weak models in the ensemble, is the weight (significance) of the t-th model, is the prediction of the weak classifier $\alpha_t h_t(x)$t on object x.

The XGBoost algorithm maintains the importance of features, which is useful in medical research. The LightGBM algorithm shows a high speed, which is useful when iterating over hyperparameters. Fig.2 shows a graphical model of the Boosting ensemble algorithm.

Fig. 2 illustrates the general architecture of the Boosting ensemble algorithm applied to anemia classification. The original training dataset is divided into several subsets, each of which is used to train a separate Boosting model, such as XGBoost, AdaBoost, and LightGBM. Each model is trained iteratively, learning from the errors of previous models to minimize the overall prediction error. Misclassified cases from earlier stages (*e.g.*, XGBoost and AdaBoost) are re-weighted or emphasized in subsequent rounds. The LightGBM model, known for its computational efficiency and ability to handle large feature spaces, contributes to the final solution by aggregating and adjusting earlier predictions. XGBoost, in turn, preserves feature importance metrics, promoting clinical interpretability. Together, this ensemble approach leverages the strengths of the individual algorithms to improve diagnostic performance and robustness in clinical prediction tasks.

Due to the ability to identify complex nonlinear dependencies in data, resistance to overfitting, and high interpretability (in particular, through the importance of features), boosting has demonstrated high efficiency in the task of multi-label classification of anemia, providing similar or better-quality metrics compared to neural network and classical machine learning methods.

To minimize the risk of overfitting, particularly in models prone to it such as AdaBoost and deep neural networks, several strategies were applied. All base and ensemble classifiers underwent stratified 10-fold cross-validation, ensuring that each fold preserved the proportion of anemia subtypes and

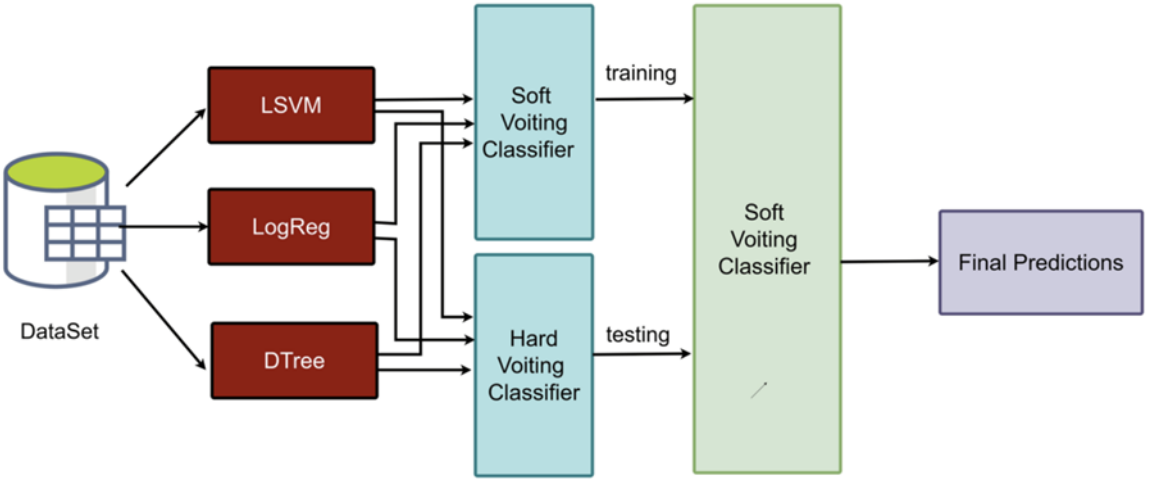

Fig. 4: Visualization of the voting ensemble algorithm.

class balance. This allowed for robust estimation of generalization performance.

For ensemble methods, including AdaBoost, we limited the number of base estimators (n_estimators) to 50-100 depending on model convergence and validation accuracy, and controlled model depth where applicable. Regularization techniques such as early stopping and minimum split gain (in XGBoost and LightGBM) were also used.

Hyperparameter tuning was performed via grid search for classical models and ensembles, and via Optuna for neural networks. The search was constrained to avoid overfitting by favoring simpler architectures and penalizing highly complex ones that demonstrated poor cross-fold stability.

Model reliability was evaluated by tracking not only accuracy but also standard deviation across folds, as well as standard error of the mean (SE), to assess the consistency of results. Performance was considered stable if the STD of key metrics remained below 0.05.

These precautions helped ensure that models were not only accurate on training data but also generalizable and robust under cross-validation.

2.6 Bagging

To study the quality of the classification of anemia and its subtypes, based on the use of various basic learning algorithms and varying the number of ensemble models, the Bagging ensemble technology was used.[61] The bagging method is an ensemble technique in which several weak models (basic classifiers) are trained on various bootstrap samples from training data.[60] The results are combined by voting (for classification) or averaging (for regression). This reduces model variance, increases resilience to overfitting, and improves generalizability.

To classify anemia and its subtypes, the following list of basic classifiers was chosen, based on their self-standing performance on this dataset:

- Support Vector Machine (Linear);
- Logistic Regression;
- Decision Tree.

The graphical structure of the Bagging ensemble is shown in Fig. 3. Each classifier is trained on separate resampled datasets, and their outputs are aggregated to produce a final anemia classification.

As shown in Fig. 3, the Bagging ensemble combines predictions from multiple base classifiers (Linear SVM, Logistic Regression, and Decision Tree), each trained on different bootstrapped subsets of the data. This approach improves classification robustness by reducing overfitting and variance, which is especially important when dealing with imbalanced classes such as anemia subtypes. By combining different decision boundaries, the ensemble captures finer patterns in the hematology data, leading to improved recall and F1 scores in most configurations.

The ensemble has the following number of basic models: 5, 10, 15, 20. All models perform 10-fold cross-validation and calculation of the following metrics: accuracy, precision, recall, F1-score. As a result, 12 models are obtained, which can be compared with each other in terms of the degree of influence of bagging on each of the basic classifiers and in terms of the dependence of quality on the number of ensemble models.

2.7 Voting

The next method of model ensemble was the Voting Classifier algorithm. This implementation uses three basic models: SVM (linear core), logistic regression and decision tree.

Two types of voting were considered: soft and hard. In the first case, predictions are averaged by probabilities, in the second - by the majority of votes. The assessment was carried

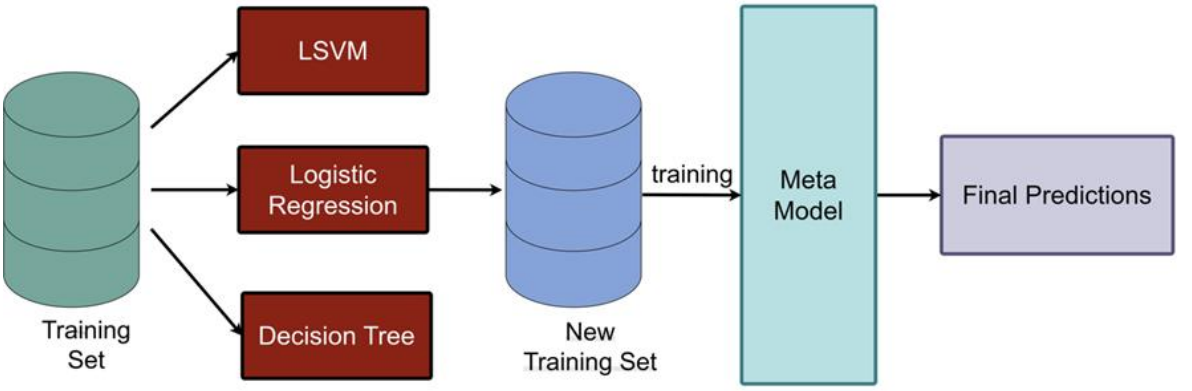

Fig. 5: Graphical model of the Stacking ensemble algorithm.

out using cross-validation by the metrics accuracy, precision, recall, F1-score.

A visualization of the Voting ensemble algorithm is shown in Fig.4.

Following the architecture illustrated in Fig. 4, the Bagging approach enabled a systematic comparison of base classifiers under varying ensemble sizes. By training Support Vector Machine, Logistic Regression, and Decision Tree models on resampled subsets of the data and aggregating their predictions, the method demonstrated improved robustness to overfitting and better generalizability. The results showed that increasing the number of estimators generally led to higher classification stability and reduced variance, particularly for less complex models like Decision Trees. This ensemble strategy proved especially beneficial in the presence of class imbalance and noisy features, reinforcing the utility of Bagging for medical classification tasks such as anemia subtype prediction.

This algorithm has two modes, both of which were evaluated in this study:

- Hard voting, when the winner is determined by most predictions from individual models, as defined in Eq. (28):

$$\hat{y} = mode(y_1, y_2, y_3) \tag{28}$$

where: y_1 - LSVM, y_2 - logistic regression, y_3 - decision tree estimations

- Soft voting, which calculates the value of the averaging of probabilities, and the selection by the highest mean value, as shown in Eq. (29):

$$\hat{y} = \underset{c \in \{1,..,C\}}{\arg max}\left(\frac{1}{K}\sum_{k=1}^{K} p_k^{(c)}(x)\right) \tag{29}$$

where: C is the number of classes, K is the number of basic classifiers, $p_k^{(c)}(x)$ is the probability of class c according to the model Mk.

The comparative evaluation of both voting strategies revealed that soft voting consistently outperformed hard voting across most anemia classification tasks. This superiority can be attributed to the probabilistic nature of soft voting, which allows integration of confidence levels from individual classifiers, making it particularly effective in borderline and imbalanced cases. As such, the Voting ensemble approach, especially in its soft implementation, offers a compelling balance between interpretability and performance, making it well-suited for integration into diagnostic support systems in clinical hematology. The results of this method are discussed in more detail in the Results section.

2.8 Stacking

The next ensemble algorithm was a Stacking algorithm implementation that combines the predictions of several basic models, passing them as input data to a meta-classifier (final estimator), which is trained to make a final decision.[59] The experiment used combinations of three basic models (SVM, Logistic Regression, Decision Tree), as well as three different final classifiers. All stacking models were evaluated against a variety of anemia classification tasks using cross-validation.

A visualization of the Stacking ensemble algorithm is shown in Fig. 5. The Stacking ensemble approach, illustrated in Fig. 5, efficiently integrates multiple base classifiers (SVM, logistic regression, and decision tree) by training them on the original dataset and then using their predictions to form a new training set for the meta-model.

This architecture allows the meta-classifier to learn from the strengths and weaknesses of individual base learners, thereby improving the overall predictive performance. In the context of anemia classification, Stacking has demonstrated strong generalizability and adaptability, especially when using robust meta-models such as gradient boosting or logistic regression. The multi-layered structure facilitates better discrimination of overlapping anemia subtypes and increased sensitivity in detecting rare classes.

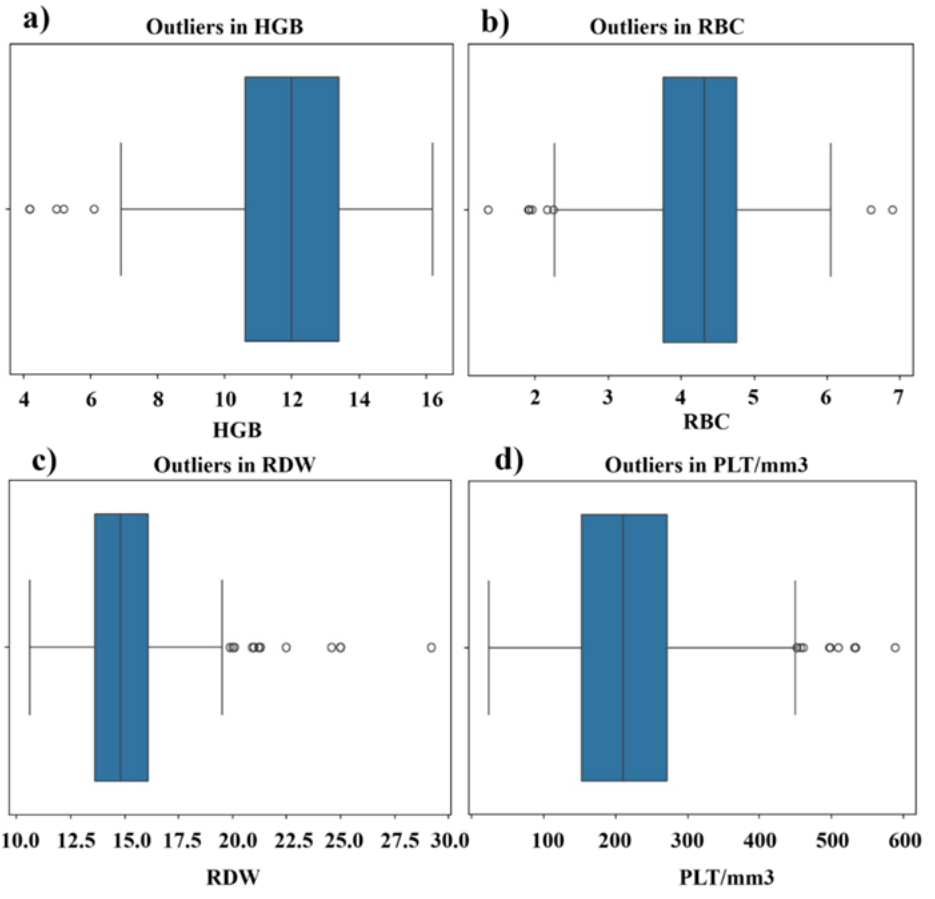

Fig. 6: Analysis of outliers in key hematological parameters: *(a)* HGB; *(b)* RBC; *(c)* RDW; *(d)* Platelet count (PLT/mm³).

2.9 Neural networks

Medical data often contain non-linear relationships between signs and diagnoses. Neural networks cope well with such structures, identifying complex patterns. In this study, neural networks are used as one of the methods for classifying anemia and its subtypes based on medical signs (RBC, HGB, RDW, etc.). Signs of anemia (laboratory test indicators) have nonlinear connections that are well captured by NNs. Also, they are easily scalable and adaptable to other medical tasks. Basic models (*e.g.*, logistic regression) are inferior in flexibility and expressiveness, especially in the case of complex, sparse, partially correlated information.

An artificial neural network is a multi-layered model that imitates the work of biological neurons.[62] Input variables (medical indicators) pass through several hidden layers, where nonlinear transformations are used. In the output layer, 7 neurons with the function of sigmoid activation are used, which makes it possible to solve the problem of multilabel classification - determining the presence of one or more subtypes of anemia at the same time.

To automatically configure the architecture of neural networks, the Optuna library was used, which made it possible to automatically select the number of layers and neurons, as well as select the activation function. This made it possible to obtain the optimal architecture without manual selection.

The model was trained with validation on a test sample and visualization of loss and accuracy graphs. Cross-validation was not used directly in the neural network (since it is resource-intensive) and instead validated on the deferred part (20%) and compared the results with other models based on the same data.

Thus, both basic algorithms (SVM, logistic regression, decision trees) and ensemble methods were used to build models: Bagging, Voting (soft and hard), Boosting (XGBoost, AdaBoost, LightGBM), and Stacking. In addition, a neural network with hyperparameter selection using Optuna was implemented. The models were evaluated according to the following metrics: Accuracy, Precision, Recall, F1-score, and CV Mean (accuracy during the fitting process).

An artificial neural network optimized using Optuna demonstrated high accuracy in the task of multi-label classification of anemia. In all key metrics, it surpassed or was comparable to ensemble methods. Due to its ability to automatically adjust and work with multiple labels, the proposed model has a high potential for practical application in clinical diagnostics. Neural network models optimized using

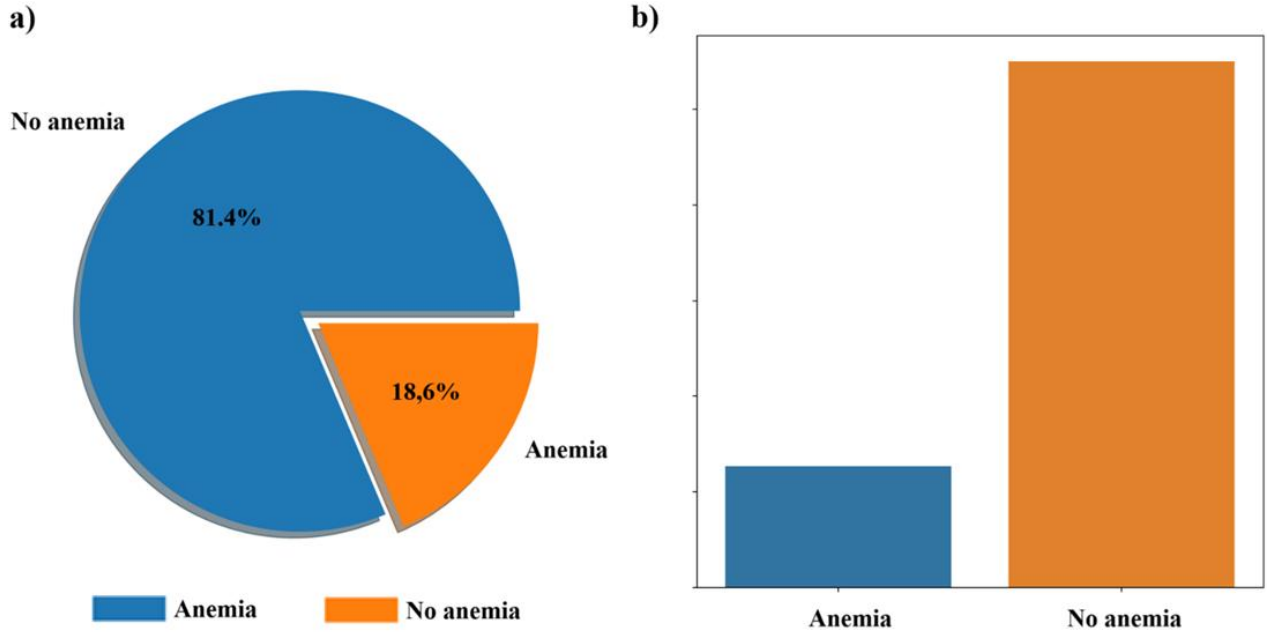

Fig. 7: Graphs of distribution of the results of anemia diagnosis: (a) Pie chart showing the proportion of patients diagnosed with anemia (red/orange) versus those without anemia (blue); (b) Bar chart displaying the absolute counts of anemia vs. non-anemia cases in the dataset.

the Optuna library showed decent results (F1-score > 0.9), but in some cases they turned out to be less stable with an increase in the number of epochs or the depth of the network. This may be due to the small size of the training sample and the presence of an imbalance between classes, especially at the Mmb12 label.

The neural network model implemented in the study can be used as an auxiliary system for doctors in clinics and laboratories to assess the likelihood of anemia based on a blood test. The model is able to recognize not only the presence of anemia, but also its morphological type and possible cause. It is possible to integrate into medical information systems, learn from new data, expand to other diseases (for example, thrombocytosis, leukemia, etc.).

3. Results and discussion

3.1 Findings from exploratory data analysis

EDA served as a foundational step in preparing the clinical dataset for subsequent modeling. Its primary objectives were to evaluate data quality, identify missing or inconsistent values, detect outliers, assess class balance, and explore relationships between variables. In the context of anemia diagnostics, EDA not only ensured technical data integrity but also revealed biologically meaningful patterns and potential sources of diagnostic differentiation. This process helped refine feature selection, guide preprocessing strategies, and uncover clinically relevant structures within the dataset.

An analysis of outliers by several main characteristics is presented in Fig. 6.

Outlier analysis is an essential component of exploratory data analysis, particularly in clinical datasets where extreme values can reflect either pathological states or data entry errors. Fig. 7 illustrates the distribution and outlier detection results for four key hematological features. The boxplots reveal several data points beyond the interquartile range, which were further validated against clinical reference intervals to distinguish biologically plausible extremes from erroneous entries. For instance, outliers in RDW and PLT may reflect actual hematologic anomalies associated with anemia subtypes, whereas improbable values (*e.g.*, extreme PLT beyond 600/mm³ or HGB below 5 g/dL) were flagged and excluded during preprocessing. These insights informed subsequent data cleaning and model training steps, ensuring robustness and clinical reliability of the final predictions.

Hemoglobin levels show several low outliers, which may indicate potential cases of severe anemia. The red blood cell count displays an asymmetric distribution, suggesting possible deviations in the process of erythropoiesis. The red cell distribution width includes significant outliers skewed to the right, reflecting the presence of anisocytosis. The platelet count per cubic millimeter of blood contains extreme values, which may correspond to thrombocytosis or result from data artifacts.

To better understand the composition of the sample, a graph was built of the distribution of the target variable Result, which reflects the presence or absence of anemia in patients (Fig. 7).

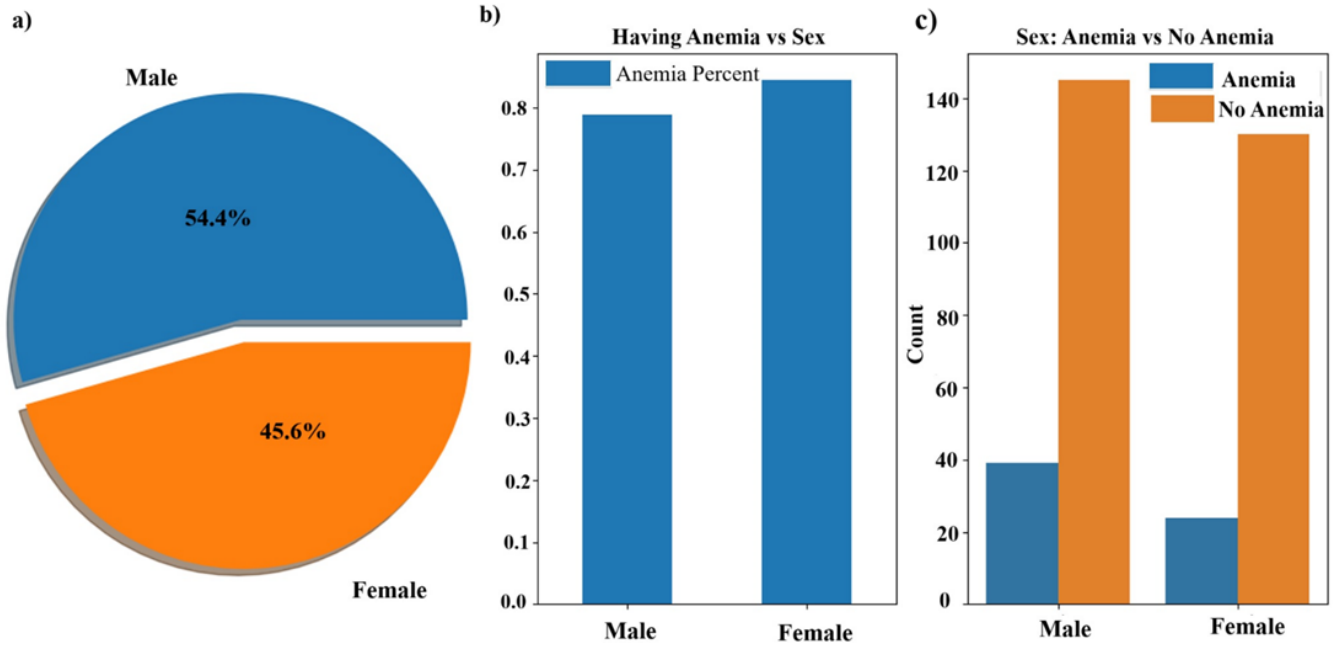

Fig. 8: Distribution of anemia status by gender: *(a)* Proportion of male and female participants in the dataset; *(b)* Anemia prevalence (%) among male and female groups; *(c)* Absolute counts of anemia vs. non-anemia cases stratified by sex.

Fig. 8 shows that the class "Anemia" occurs in ~81% of cases, and "No Anemia" occurs only in ~19%. Graphs of the distribution of results by gender is presented in Fig. 8.

Fig. 8 illustrates the gender-based distribution of anemia in the dataset. Subplot (a) indicates a slightly higher representation of male participants (54.4%) compared to females (45.6%). Interestingly, subplot (b) shows that the proportion of individuals with anemia is high for both sexes, with female participants demonstrating a slightly higher anemia prevalence rate. Subplot (c) provides absolute counts, confirming that anemia occurs across both genders, though the raw number of non-anemic individuals is markedly higher in both groups. These findings highlight the need to account for sex-specific factors in anemia prediction models and reinforce the clinical relevance of gender as a stratification variable in diagnostic algorithms.

The next stage of the EDA procedure is the correlation analysis[63] of clinical and hematological syndrome variables. Fig. 9 shows the correlation matrix of all the variables visualized in a heatmap. It shows how strong and in which direction the features are related to each other according to the Pearson correlation coefficient.

Fig. 9 illustrates the Pearson correlation matrix among key clinical and hematological features. As expected, HGB, HCT, and RBC exhibit high positive correlations, which aligns with physiological understanding of blood composition. These strong interdependencies confirm the need to manage multicollinearity when selecting input variables for machine learning models. Additionally, derived features such as “Microcyt,” “Macrocyt,” and “Normal” anemia flags show meaningful correlations with their biochemical counterparts (*e.g.*, MCV, MCH), reflecting the clinical logic used in morphological subtype classification. This analysis supports the reliability of the dataset structure and helps identify redundant or strongly correlated features that could be simplified or excluded to improve model generalization.

To better understand how traits affect each subtype of anemia and to what extent biomarkers (ferritin, B12) actually reflect the corresponding pathologies, a correlation between features and types of anemia was constructed in Fig. 10.

Fig. 10 presents a focused correlation matrix between clinical features and specific anemia subtypes. It reveals that RDW shows the strongest positive correlation with the general anemia label (Result) and subtypes such as Microcyt and Normal, indicating its critical role in detecting morphological abnormalities in red blood cells. Conversely, traditional indicators like HGB, HCT, and RBC show moderate to strong negative correlations with anemia outcomes, consistent with clinical expectations. Interestingly, the correlation between ferritin and B12 levels with their respective labels (Mmb12, Mmahz, Mmnzhda) is relatively weak, suggesting that although these biochemical markers are important diagnostically, they may not always align directly with morphological or statistical groupings in small or imbalanced

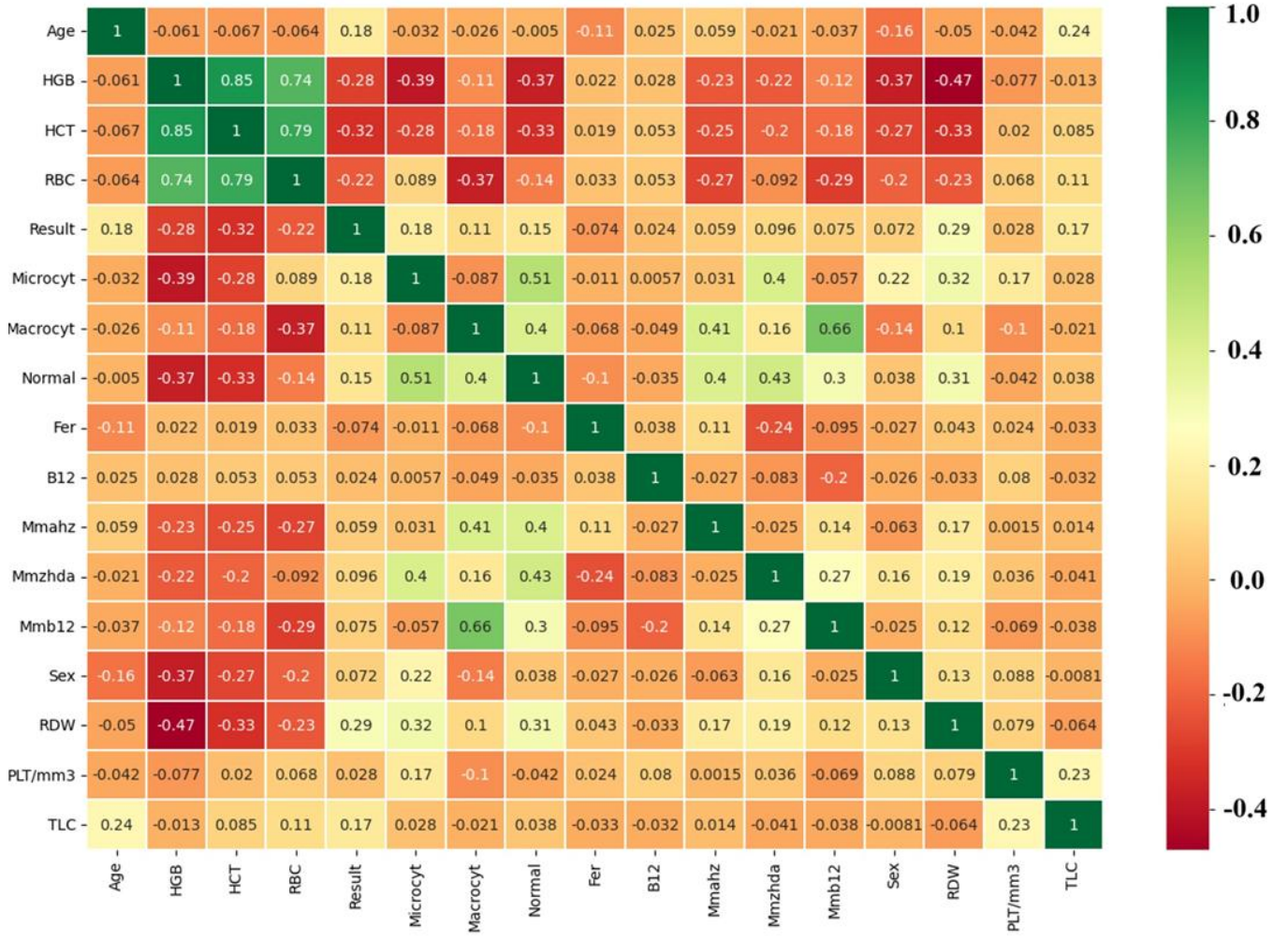

Fig. 9: Correlation matrix of all anemia variables.

datasets. These insights can guide both clinical interpretation and feature selection for machine learning models, emphasizing the importance of composite indicators over isolated measurements.

To identify differences between patients with and without anemia in terms of the main blood features, a pair plot was constructed to visually assess the data structures and the potential separability of classes (Fig. 11).

As demonstrated in Fig.11, the pair plot highlights the structural distinctions between anemia and non-anemia groups across key hematological features. Patients with anemia generally exhibit lower values of HGB and RBC, consistent with clinical diagnostic criteria. Additionally, the RDW is significantly elevated in the anemia group, which is indicative of increased variation in red blood cell size (anisocytosis), commonly observed in iron-deficiency and mixed-type anemias. These visual separations suggest that the selected features carry meaningful discriminatory power, supporting their relevance for subsequent machine learning classification and justifying their inclusion in model training pipelines.

To assess the influence of individual traits on the prediction of anemia by models, a feature importance graph was built, shown in Fig.12. The graph shows that HCT, red RDW, and RBC contribute the most to the prediction. This confirms the clinical significance of these parameters in the diagnosis of anemia.

As shown in Fig. 12, the importance ranking of features derived from the machine learning model highlights HCT, RDW, and RBC as the most influential variables in predicting anemia. These results align closely with established clinical knowledge, reinforcing the biological validity of the model. HCT reflects the proportion of red blood cells in blood, RDW captures size variability (anisocytosis), and RBC quantifies the total red cells - all central to the differential diagnosis of anemia. Lesser contributions from features such as B12, ferritin, and sex indicate a more nuanced or indirect influence, which is expected given the multicausal nature of anemia. This analysis demonstrates the model's capacity to reflect underlying medical mechanisms and supports its interpretability in clinical decision support.

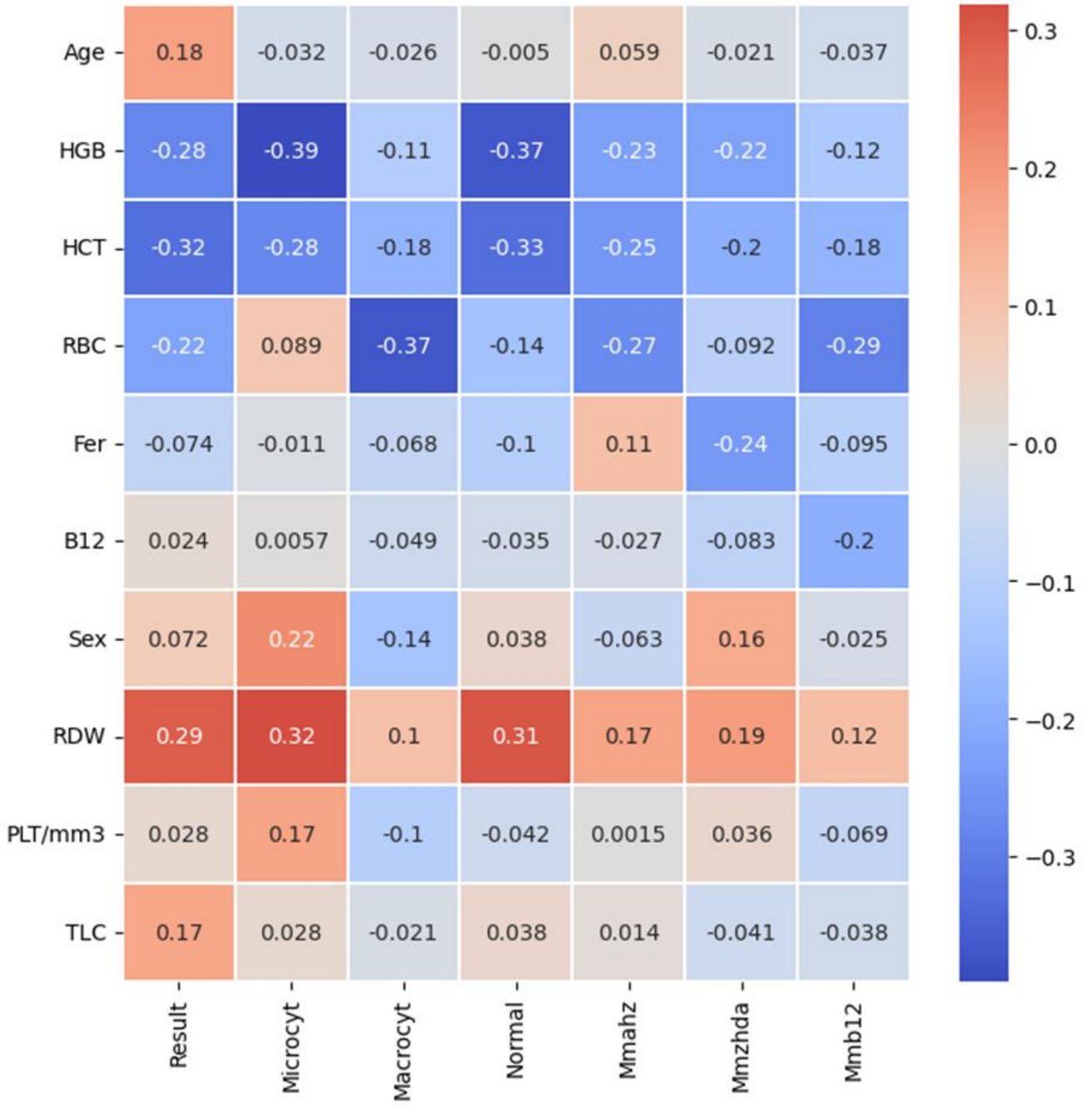

Fig. 10: Correlation matrix of variables and subtypes of anemia.

To assess the internal structure of the data and the possibility of linear division of classes, the projection of features into a two-dimensional space using the principal component method (PCA) was performed.[64] Fig. 13 shows a graph showing the partial separability of patients with and without anemia, which confirms the presence of significant patterns and the validity of the construction of a predictive model.

The PCA visualization in Fig. 13 provides a two-dimensional projection of the multidimensional feature space, revealing a partial clustering of patients with and without anemia. Although there is visible overlap between the two groups-reflecting the inherent complexity and heterogeneity of hematological data - distinct regions with predominantly one class indicate that the model can detect discriminative patterns. This structure justifies the application of classification algorithms, particularly those capable of capturing non-linear boundaries, such as XGBoost, SVM, or ensemble methods. The plot supports the premise that the dataset contains learnable signals that can facilitate effective anemia detection.

Thus, the use of EDA for the differential diagnosis of clinical and hematological syndromes is justified, because it creates an informed basis for subsequent modeling steps and helps to avoid errors associated with the "black box" of ML. It

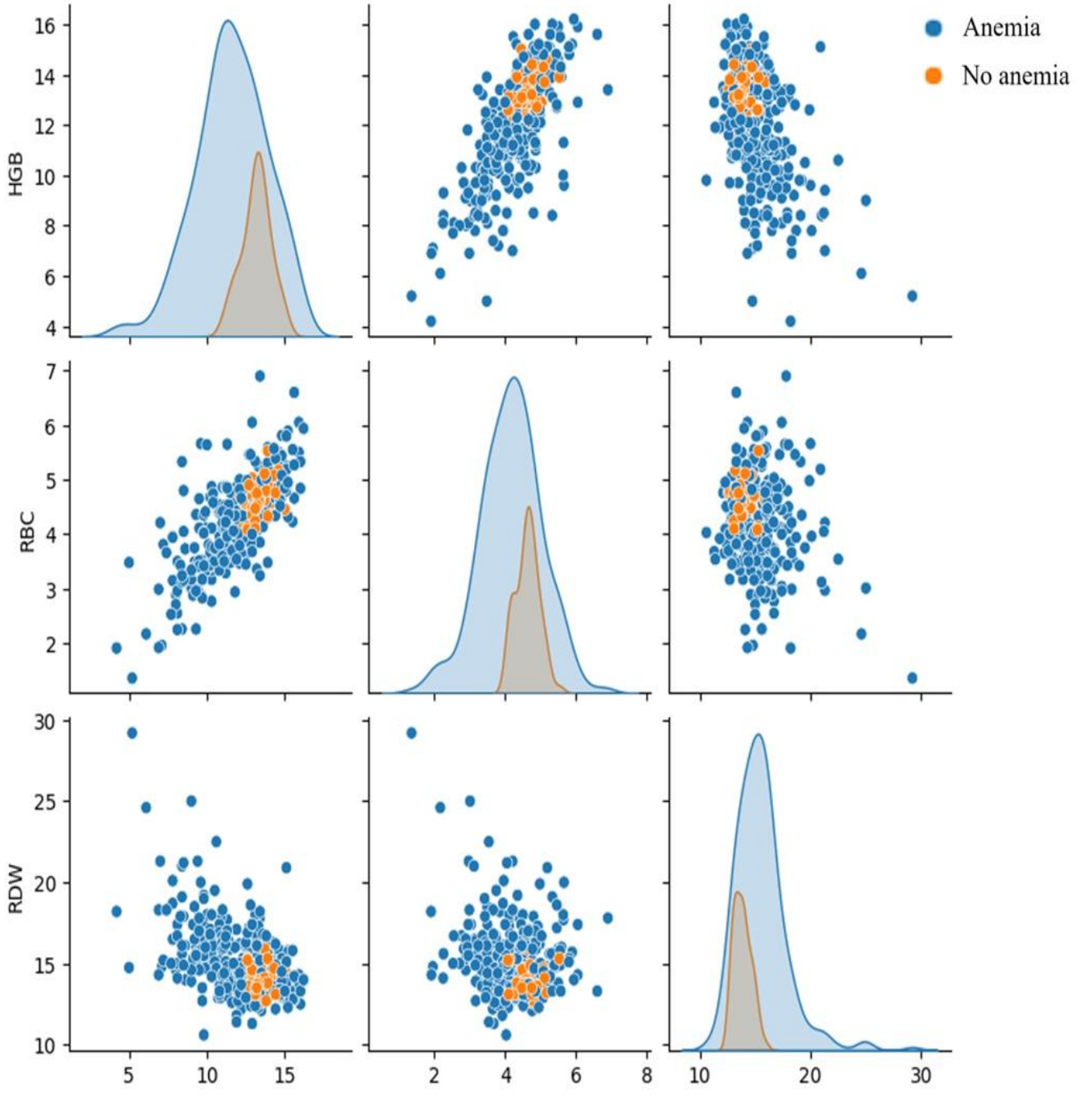

Fig. 11: Relationship between the main features and anemia.

also makes the analysis representative, transparent and reproducible.

3.2 Average accuracy of models on anemia classification tasks

To evaluate the performance of various machine learning models in the classification of anemia and its subtypes, cross-validation was conducted using a comprehensive set of performance metrics: accuracy, recall, precision, F1-score, cross-validation mean (CV mean), and standard deviation (STD). This approach ensures a robust comparison of model behavior across different data partitions and helps identify algorithms that demonstrate both high accuracy and stability. The results are presented as horizontal bar charts, enabling both visual and quantitative analysis (Fig. 14).

As shown in Fig. 14, a comparative evaluation across multiple metrics (precision, confidence, recall, F1-score, mean CV, and standard deviation) demonstrates that the logistic regression and decision tree models consistently deliver strong and balanced performance on the anemia classification task. While models such as linear SVM demonstrate high precision, they also exhibit higher variability (as seen in STD), suggesting potential overfitting or sensitivity to data splits.

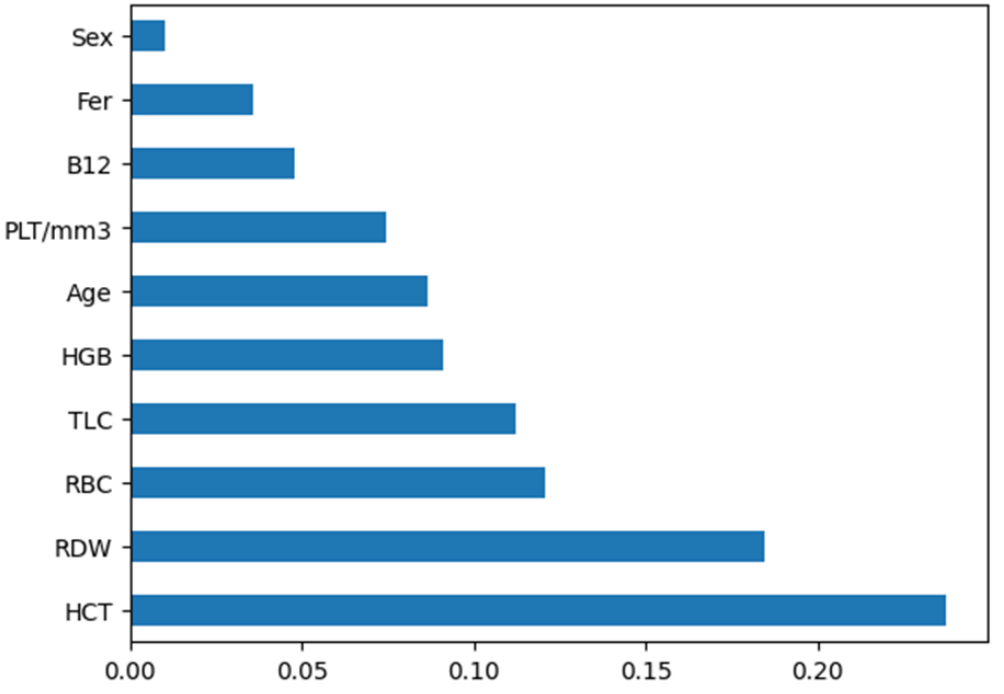

Fig. 12: Ranking of traits by the degree of their contribution to the anemia prediction model.

Random Forest and Naive Bayes also perform robustly across most metrics, with Random Forest demonstrating robust F1-scores and low standard deviation, indicating stability. These comparative results support the choice of ensemble or regularized models for robust anemia classification in real-world applications.

3.3 Results of ensemble-based classification

3.3.1 Results of boosting-based classification

Boosting algorithms were used for automated diagnosis of anemia and its subtypes based on laboratory blood parameters. In particular, three popular approaches were used: AdaBoost, XGBoost, and LightGBM. These models were trained on traits that included hemoglobin levels, red blood cells, hematocrit, RDW, vitamin B12 and iron levels.

Boosting algorithms made it possible to build an ensemble of weak classifiers, where each subsequent model focused on the errors of the previous ones, thereby improving the final accuracy. The results of the boosting algorithm are presented in Table 6.

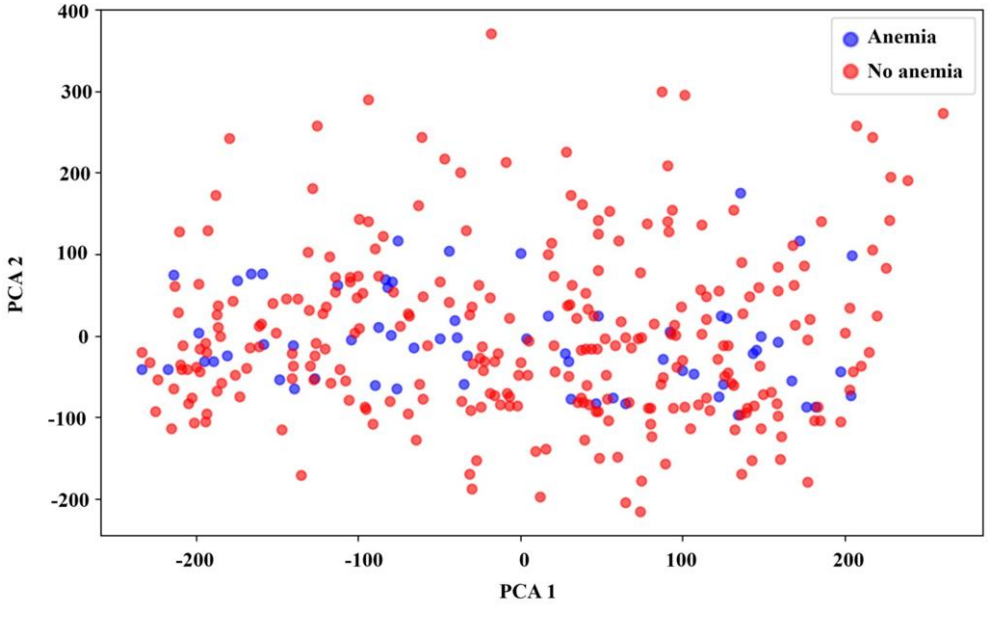

Fig. 13: PCA visualization of anemia classes based on blood features.

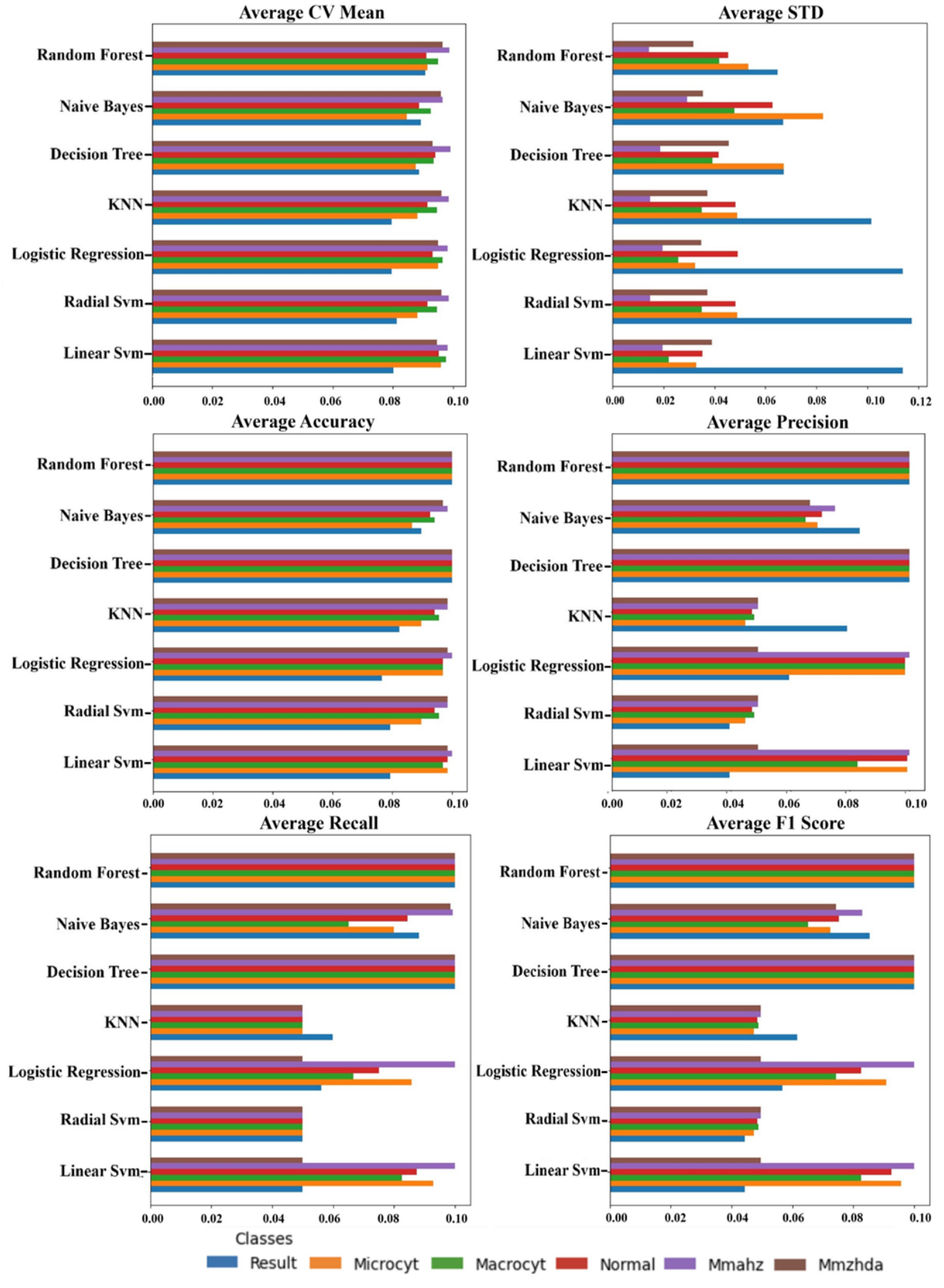

Fig. 14: Graph of the average accuracy of models cross-validation for anemia classification tasks.

Table 6: Results of the boosting algorithms.

Classification quality assessment metric	Classifier	Result	Microcyt	Macrocyt	Normal	Mmahz	Mmzhda
CV Mean	XGBoost	0.899	0.911	0.947	0.929	0.988	0.959
	AdaBoost	0.905	0.920	0.947	0.917	0.991	0.953
	LGBMClassifier	0.923	0.923	0.959	0.920	0.994	0.962
STD	XGBoost	0.069	0.056	0.037	0.048	0.015	0.044
	AdaBoost	0.079	0.042	0.047	0.039	0.013	0.042
	LGBMClassifier	0.047	0.038	0.027	0.046	0.012	0.042
Accuracy	XGBoost	1.000	1.000	1.000	1.000	1.000	1.000
	AdaBoost	0.956	1.000	1.000	0.956	1.000	1.000
	LGBMClassifier	1.000	1.000	1.000	1.000	1.000	1.000
Precision	XGBoost	1.000	1.000	1.000	1.000	1.000	1.000
	AdaBoost	0.891	1.000	1.000	0.976	1.000	1.000
	LGBMClassifier	1.000	1.000	1.000	1.000	1.000	1.000
Recall	XGBoost	1.000	1.000	1.000	1.000	1.000	1.000
	AdaBoost	0.927	1.000	1.000	0.812	1.000	1.000
	LGBMClassifier	1.000	1.000	1.000	1.000	1.000	1.000
F1 Score	XGBoost	1.000	1.000	1.000	1.000	1.000	1.000
	AdaBoost	0.908	1.000	1.000	0.872	1.000	1.000
	LGBMClassifier	1.000	1.000	1.000	1.000	1.000	1.000

As shown in Table 6, all three boosting models demonstrated strong classification performance across anemia subtypes, achieving perfect or near-perfect accuracy, precision, recall, and F1 scores in most categories. LightGBM slightly outperformed the others in terms of cross-validation (CV) mean and standard deviation (STD), indicating both higher average predictive power and greater consistency across folds. XGBoost also maintained excellent performance, particularly in precision and recall, while AdaBoost, although slightly less stable in some categories (*e.g.*, Mmahz), still provided competitive results. These findings confirm that boosting algorithms are well-suited for complex, multi-label classification tasks in hematological diagnostics and can serve as reliable tools for supporting clinical decision-making.

Thus, due to the ability to identify complex nonlinear dependencies in data, resistance to overfitting, and high interpretability (in particular, through the importance of features), boosting has demonstrated high efficiency in the task of multi-label classification of anemia, providing comparable quality metrics compared to classical machine learning methods.

3.3.2 Results of bagging-based classification

Further, for a visual comparison of the quality metrics of the classifiers for all subtypes of anemia, the function of each indicator was implemented. This allows you to effectively interpret the results and choose the best model for each task. The results of the Bagging algorithm are presented in Table 7.

As evident from Table 7, the Random Forest and Decision Tree ensembles achieved the highest scores across all metrics, including accuracy, precision, recall, and F1-score, consistently reaching values close to or equal to 1.0. These results suggest that tree-based methods benefit the most from the bagging approach due to their sensitivity to data variance and their ability to capture nonlinear feature interactions. In contrast, models such as Radial SVM and KNN performed less consistently, especially on minority anemia subtypes, possibly due to class imbalance and lower robustness to noisy features. Naive Bayes also demonstrated moderate performance but with larger standard deviations, indicating less stability. Overall, the findings confirm that bagging significantly enhances classification performance for decision trees and further justifies the use of Random Forests in clinical diagnostic tasks involving complex and imbalanced datasets.

3.3.3 Results of voting-based classification

The next algorithm for assembling machine learning methods was Voting, which integrates the Soft Voting and Hard Voting strategies. To compare the quality of ensemble classifiers on anemia diagnostic subtasks (Result, Microcyt, Macrocyt, Normal, Mmahz, Mmzhda), the same metrics were used as in the previous algorithms (Average, CV Mean, Average, STD, Accuracy, Precision, Recall, F1 Score). The results of the experimental study of the Voting ensemble algorithm are presented in Table 8.

As shown in Table 8, the Soft Voting ensemble (Vsoft-LSVM+LR+DT) consistently outperformed the Hard Voting variant across most anemia subtypes and classification metrics.

Table 7: Results of the bagging algorithm.

Classification quality assessment metric	Classifier	Result	Microcyt	Macrocyt	Normal	Mmahz	Mmzhda
CV Mean	Linear Svm	0.801	0.959	0.976	0.953	0.982	0.947
	Radial Svm	0.813	0.882	0.947	0.914	0.985	0.962
	Logistic Regression	0.795	0.950	0.965	0.932	0.982	0.950
	KNN	0.796	0.882	0.947	0.914	0.985	0.962
	Decision Tree	0.887	0.876	0.935	0.941	0.991	0.932
	Naive Bayes	0.893	0.846	0.926	0.888	0.964	0.959
	Random Forest	0.908	0.915	0.950	0.911	0.988	0.965
STD	Linear Svm	0.114	0.033	0.022	0.035	0.020	0.039
	Radial Svm	0.117	0.049	0.035	0.048	0.015	0.037
	Logistic Regression	0.114	0.032	0.026	0.049	0.020	0.035
	KNN	0.102	0.049	0.035	0.048	0.015	0.037
	Decision Tree	0.067	0.067	0.039	0.042	0.019	0.046
	Naive Bayes	0.067	0.082	0.048	0.063	0.029	0.035
	Random Forest	0.065	0.053	0.042	0.045	0.014	0.032
Accuracy	Linear Svm	0.794	0.985	0.971	0.985	1.0	0.985
	Radial Svm	0.794	0.897	0.956	0.941	0.985	0.985
	Logistic Regression	0.765	0.971	0.971	0.971	1.0	0.985
	KNN	0.824	0.897	0.956	0.941	0.985	0.985
	Decision Tree	1.0	1.0	1.0	1.0	1.0	1.0
	Naive Bayes	0.897	0.868	0.941	0.926	0.985	0.971
	Random Forest	1.0	1.0	1.0	1.0	1.0	1.0
Precision	Linear Svm	0.397	0.992	0.826	0.992	1.0	0.493
	Radial Svm	0.397	0.449	0.478	0.471	0.493	0.493
	Logistic Regression	0.596	0.984	0.985	0.985	1.0	0.493
	KNN	0.789	0.449	0.478	0.471	0.493	0.493
	Decision Tree	1.0	1.0	1.0	1.0	1.0	1.0
	Naive Bayes	0.833	0.69	0.651	0.706	0.75	0.667
	Random Forest	1.0	1.0	1.0	1.0	1.0	1.0
Recall	Linear Svm	0.5	0.929	0.826	0.875	1.0	0.5
	Radial Svm	0.5	0.5	0.5	0.5	0.5	0.5
	Logistic Regression	0.561	0.857	0.667	0.75	1.0	0.5
	KNN	0.598	0.5	0.5	0.5	0.5	0.5
	Decision Tree	1.0	1.0	1.0	1.0	1.0	1.0
	Naive Bayes	0.882	0.8	0.651	0.844	0.993	0.985
	Random Forest	1.0	1.0	1.0	1.0	1.0	1.0
F1 Score	Linear Svm	0.443	0.957	0.826	0.925	1.0	0.496
	Radial Svm	0.443	0.473	0.489	0.485	0.496	0.496
	Logistic Regression	0.566	0.909	0.742	0.826	1.0	0.496
	KNN	0.616	0.473	0.489	0.485	0.496	0.496
	Decision Tree	1.0	1.0	1.0	1.0	1.0	1.0
	Naive Bayes	0.854	0.725	0.651	0.753	0.83	0.742
	Random Forest	1.0	1.0	1.0	1.0	1.0	1.0

Soft Voting achieved higher average accuracy and F1-score, particularly in complex or low-prevalence categories such as Mmahz and Mmzhda. This advantage stems from its ability to aggregate probabilistic outputs from base classifiers, which provides a more nuanced decision boundary compared to the simple majority rule of Hard Voting. Additionally, Soft Voting exhibited lower standard deviation values, indicating more stable and reliable predictions across cross-validation folds. These results confirm the suitability of probabilistic ensemble strategies in medical diagnosis tasks where interpretability and sensitivity are critical, and further support the integration of Soft Voting into clinical decision-support systems.

Table 8: Results of the experimental study of the voting ensemble algorithm.

Classification quality assessment metric	Classifier	Result	Microcyt	Macrocyt	Normal	Mmahz	Mmzhda
CV Mean	Vsoft-LSVM+LR+DT	0.870	0.947	0.968	0.938	0.985	0.965
	Vhard-LSVM+LR+DT	0.819	0.953	0.968	0.950	0.982	0.953
Average STD	Vsoft-LSVM+LR+DT	0.079	0.026	0.024	0.038	0.015	0.039
STD	Vhard-LSVM+LR+DT	0.104	0.030	0.024	0.032	0.020	0.038
Accuracy	Vsoft-LSVM+LR+DT	0.985	0.971	1.000	0.956	1.000	0.956
	Vhard-LSVM+LR+DT	0.912	0.971	1.000	0.926	1.000	0.941
Recall	Vsoft-LSVM+LR+DT	0.944	0.917	1.000	0.867	1.000	0.700
Precision	Vhard-LSVM+LR+DT	0.667	0.917	1.000	0.742	1.000	0.600
F1 Score	Vsoft-LSVM+LR+DT	0.966	0.946	1.000	0.888	1.000	0.774
	Vhard-LSVM+LR+DT	0.726	0.946	1.000	0.787	1.000	0.651

Analyzing the results of the experimental study of the Voting ensemble algorithm, the following conclusions can be drawn:

- Soft Voting summarizes data better, especially for the Result label;
- Vsoft outperforms Vhard in almost every subtask;
- Both methods are high on the Precision metric, but Vsoft is higher on Result, Normal, and Mmzhda.

To improve the quality of anemia classification, an ensemble approach was implemented using the Voting Classifier. A comparison of soft and hard voting strategies showed that the model with soft voting (Vsoft-LSVM+LR+DT) demonstrates the best values for all metrics: accuracy, precision, and F1-score. The advantage is especially noticeable in the Result label (F1-score: 0.966 vs. 0.726), which emphasizes the critical importance of probabilistic averaging for medical tasks. In addition, soft voting provided more consistent results in cross-validation (low STD value), in contrast to hard voting, which showed high variability and reduced quality in less pronounced classes (Mmzhda).

3.3.4 Results of stacking-based classification

To improve the quality of classification, a stacking technique was used, where LSVM, Logistic Regression and Decision Tree were used as basic models. Each one of them was also selected in turn as the final classifier. The Stacking results presented in Table 9 show which model was used as the final estimator (FE-*). Models trained on the tasks of diagnosing anemia and its subtypes were also evaluated using the CV Mean, STD, Accuracy, Precision, Recall, and F1 Score metrics. Based on Table 9, the following conclusions can be drawn:

- The best metrics (accuracy, F1-score, stability) were shown by stacking with the final LSVM model;
- LogReg also showed equally high accuracy, especially on the main classes;
- Decision Tree was inferior in all key metrics, especially in sensitivity (Recall) and F1;

As such, the LSVM is the preferred option for the final estimator in an ensemble.

3.3.5 Comparative analysis of ML algorithms

As part of the study, a comparative analysis of anemia classification models using the Bagging ensemble method was carried out. Ensembles with a different number of sub models (5 to 20) were built for the basic classifiers (Support Vector Machine, Logistic Regression, Decision Tree). All models were evaluated using 10-fold cross-validation for accuracy, precision, recall, and F1-score metrics. analysis of the impact of ensemble on performance.

The use of the Bagging algorithm with decision trees and Random Forest made it possible to achieve the maximum values of all metrics (Accuracy, Recall, F1-score), which confirms the high resistance of ensembles to outliers and class imbalances. Limited capabilities in multi-class and multi-label layouts. of particular interest is the comparison of voting strategies in ensemble models.

The Soft Voting model outperformed Hard Voting in almost all metrics, which confirms the advantage of the probabilistic approach in medical tasks, where sensitivity to small changes is important. Stacking using SVM and logistic regression as final classifiers also showed high generalizability and balanced metrics, especially under multiclass classification conditions.

As part of the experiment, ensemble models were built based on the Voting Classifier algorithm, which combines the predictions of three basic models: SVM (linear core), logistic regression and decision tree. Two types of voting were considered: soft and hard. In the first case, predictions are averaged by probabilities, in the second - by a majority of votes. The assessment was carried out using cross-validation for the accuracy, precision, recall, and F1-score metrics. The results

Table 9: Results of the experimental study of the stacking ensemble algorithm.

Classification quality assessment metric	Classifier	Result	Microcyt	Macrocyt	Normal	Mmahz	Mmzhda
CV Mean	FE-LSVM	0.899	0.959	0.962	0.941	0.982	0.962
	FE-LogReg	0.884	0.959	0.965	0.935	0.985	0.962
	FE-DTree	0.819	0.911	0.950	0.920	0.973	0.932
STD	FE-LSVM	0.062	0.033	0.026	0.035	0.020	0.037
	FE-LogReg	0.076	0.033	0.026	0.039	0.015	0.037
	FE-DTree	0.077	0.046	0.035	0.037	0.021	0.054
SE	FE-LSVM	0.020	0.010	0.008	0.011	0.006	0.012
	FE-LogReg	0.024	0.010	0.008	0.012	0.005	0.012
	FE-DTree	0.024	0.015	0.011	0.012	0.007	0.017
Accuracy	FE-LSVM	1.000	0.971	1.000	0.941	0.985	0.926
	FE-LogReg	1.000	0.971	1.000	0.956	0.985	0.926
	FE-DTree	0.926	0.941	1.000	0.956	1.000	0.956
Precision	FE-LSVM	1.000	0.983	1.000	0.892	0.493	0.463
	FE-LogReg	1.000	0.983	1.000	0.976	0.493	0.463
	FE-DTree	0.850	0.924	1.000	0.880	1.000	0.812
	FE-LSVM	1.000	0.917	1.000	0.804	0.500	0.500
Recall	FE-LogReg	1.000	0.917	1.000	0.812	0.500	0.500
	FE-DTree	0.816	0.866	1.000	0.921	1.000	0.976
F1 Score	FE-LSVM	1.000	0.946	1.000	0.841	0.496	0.481
	FE-LogReg	1.000	0.946	1.000	0.872	0.496	0.481
	FE-DTree	0.832	0.892	1.000	0.899	1.000	0.872

obtained demonstrate the high applicability of machine learning methods to the task of diagnosing anemia. Ensemble approaches, especially gradient boosting, have shown high accuracy and stability, outperforming neural networks with less sensitivity to hyperparameters. Soft Voting and Stacking demonstrated efficiency by combining predictions from different models, and the use of logistic regression and SVM as final classifiers allowed for high interpretability. Particular attention is paid to the Recall metric, which is critical for medical tasks. LightGBM and Voting provided the best Recall when classifying key tags. It was found that the choice of ensemble strategy and architecture can significantly affect the resulting accuracy and F1 metric, especially in subtypes of anemia (*e.g.*, Mmzhda).

3.4 Results of anemia classification using artificial neural networks

To improve the quality of recognition of anemia and its subtypes, the method of artificial neural networks was used. The model was trained on features obtained from medical tests and returned the probabilities of the presence of each of the 7 conditions. Due to the ability of neural networks to model nonlinear dependencies and support multi-label output, they are an effective tool in medical diagnostics.

The experiment confirmed the feasibility of using artificial neural networks for multi-label classification of anemia subtypes. An artificial neural network optimized using Optuna demonstrated high accuracy in the task of multi-label classification of anemia. In all key metrics, it surpassed or was

Table 10: Results of the experimental study of the stacking ensemble algorithm.

Parameter	Possible values
n_units - number of neurons in the first hidden layer, and double the number of neurons in other hidden layers	2,4,8,16
layers - number of hidden layers	1,2,3,4,5
epochs	10, 20, 30, 40, 50, 60, 70, 80, 90, 100
batch_size	2,4,6,8,10,12,14,16,18,20,22,24,26,28,30,32
activation - activation function used in all layers except for output	relu, tanh
lr - learning rate	Float value between 1e-5, 1e-2

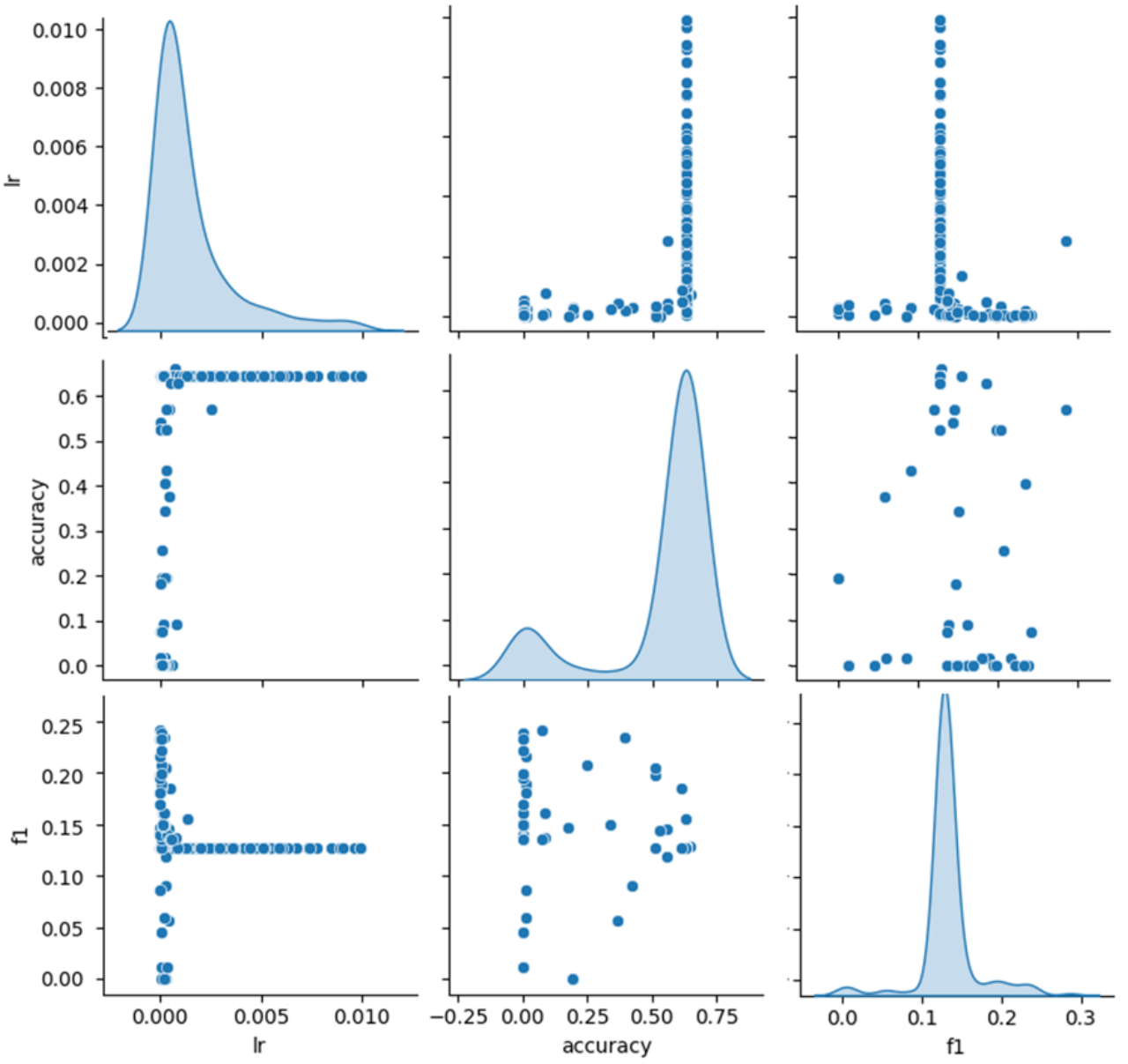

Fig. 15: Model parameters distribution and their pairing with metrics.

comparable to ensemble methods. Due to its ability to automatically adjust and work with multiple labels, the proposed methodic has a high potential for practical application in clinical diagnostics.

Next, the relationship between the hyperparameters of the neural network and quality metrics (accuracy, precision, recall, F1) was analyzed. Two hundred NN architectures with different parameters have been tested, with parameters listed in Table 10.

The hyperparameter space defined in Table 10 was systematically explored to identify optimal neural network configurations for anemia classification tasks. A total of 200 models were trained using combinations of these parameters. The diversity in architecture allowed for evaluation across a broad spectrum of complexity, from shallow networks with a small number of neurons to deeper architectures with increased representational capacity. Particular attention was given to the learning rate (lr), which strongly influenced model convergence and performance. The goal of this experimental setup was to understand which parameter combinations lead to better generalization, especially under conditions of data imbalance and clinical variability. The distribution and impact of these parameters on model quality metrics such as accuracy and F1-score are illustrated in Fig. 15.

As shown in Fig.15, most architectures tend to cluster around low learning rate values (lr < 0.002), which correlates with higher accuracy and F1-score, indicating more stable convergence. The distribution of accuracy values reveals a clear separation between underperforming and adequately performing networks, with few intermediate cases. Similarly,

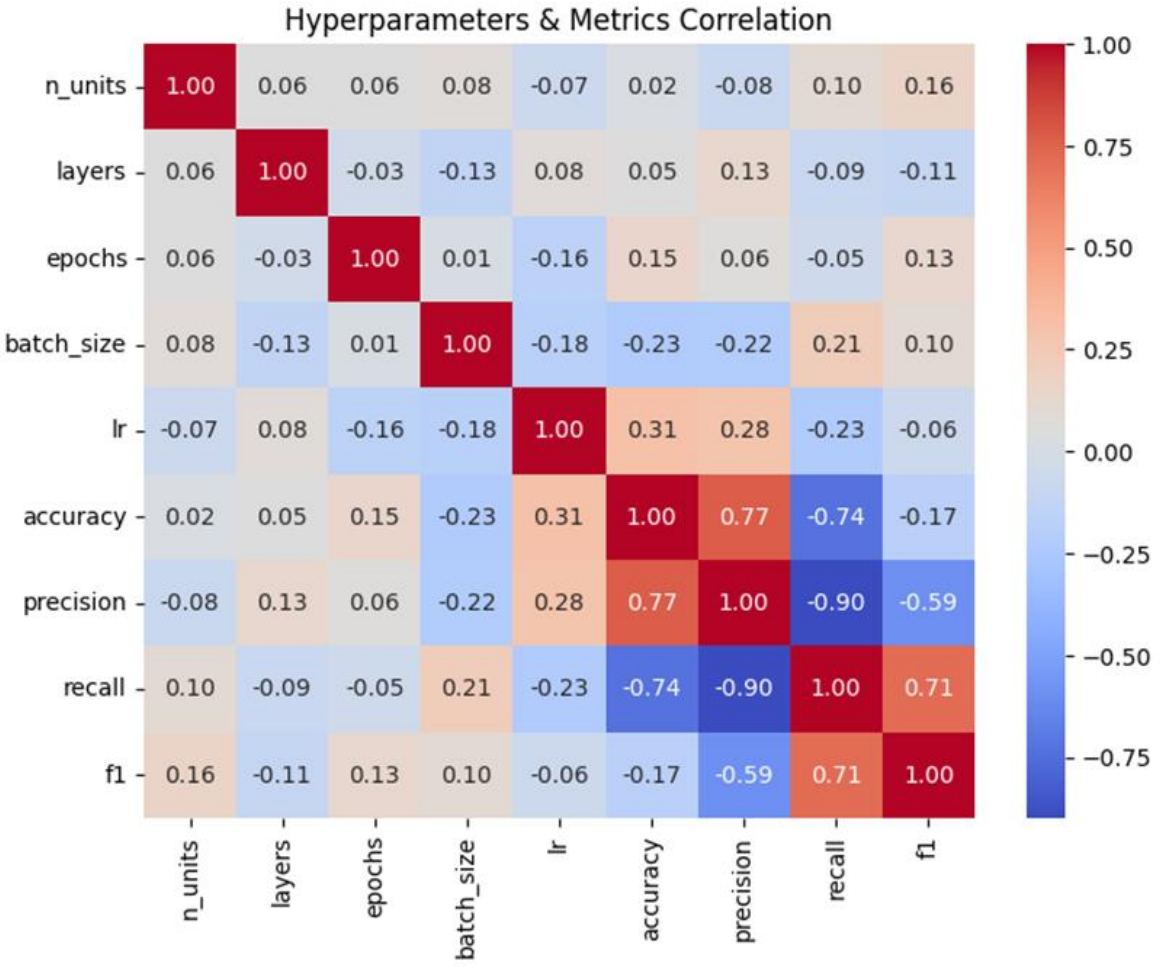

Fig. 16: Model parameters distribution and their pairing with metrics.

F1-scores are concentrated within a narrow range (0.05–0.20), which corresponds to the imbalanced nature of the dataset and the challenge of achieving both high precision and recall simultaneously. This exploratory visualization highlights the sensitivity of neural network performance to hyperparameter tuning and justifies the use of automated search techniques such as Optuna in this study.

Overall, as seen on Fig. 16, most metrics are primarily tied with model learning rate, and each other.

As shown in Fig.16, the learning rate (lr) shows the strongest positive correlation with key performance metrics such as accuracy (r = 0.31) and precision (r = 0.28), indicating that optimal tuning of this parameter is critical to the model performance. In contrast, batch size and number of epochs showed a weak or even slightly negative relationship with the prediction quality, suggesting that overfitting or unstable mini-batches gradients may have a detrimental effect on the generalization of the model. Interestingly, while accuracy and precision are highly correlated (r = 0.77), recall is inversely proportional to both (r = -0.90 with accuracy), highlighting the trade-off between network sensitivity and specificity. This interaction highlights the importance of carefully balancing model parameters to meet clinical priorities-whether minimizing false negatives (high recall) or false positives (high precision)-depending on the specific anemia subtype being diagnosed.

Graphs of learning dynamics (accuracy and loss) by epochs for all models are presented in Fig. 17. The model made it possible to demonstrate the paired dependencies between all hyperparameters and metrics, as well as to determine which parameters affect the metrics and how exactly (linear, abrupt, insignificant).

Model accuracy changes during epoch-based training for all configurations selected during hyperparametric optimization. It can be seen that many models achieve 100% accuracy in 5 to 20 epochs - especially those that converge quickly. Some configurations "hang" at the bottom for a long time, and then jump sharply, this is a sign of a low learning rate or excessive depth (layers). There are models that are not trained at all (accuracy remains around 0) - most often due to unsuccessful parameters. But fast and steady learning is observed in a small number of successful models.

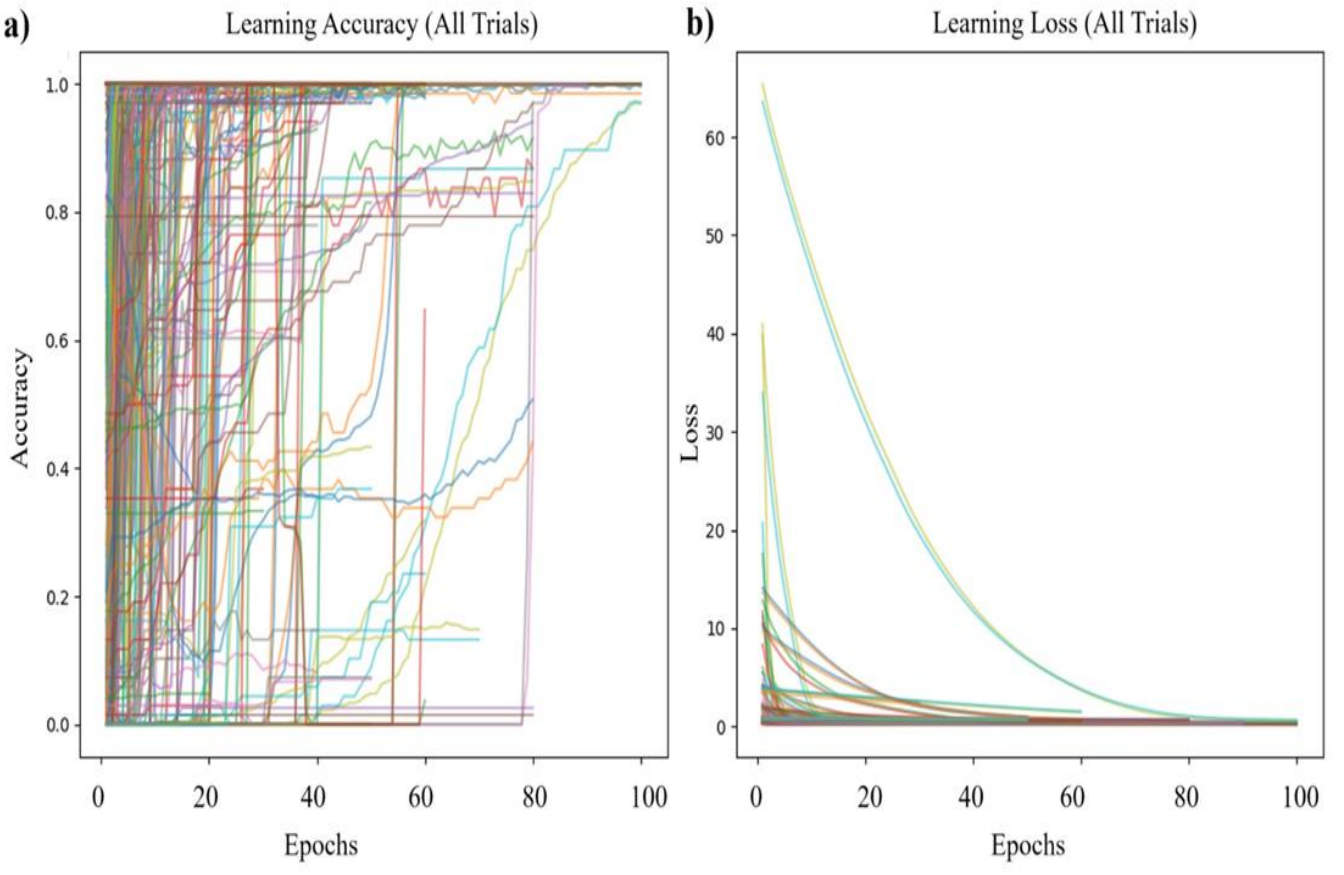

Fig. 17: Accuracy and loss over epochs.

In the Loss Over Epochs chart, some curves show a perfect rapid loss drop, with several patterns starting with a very high loss (~10-12), which means they are not approaching the target function at all. Thus, the graphs depict the training trajectories of models built during the automatic selection of hyperparameters. Despite a total of 200 trials, only a fraction of the configurations demonstrates sustained and rapid learning, achieving high learning accuracy values (accuracy ≈ 1.0) and minimizing the loss function in the first 10 to 20 epochs. The rest of the models either converge very slowly or do not learn at all. This indicates the high sensitivity of the neural network architecture to combinations of hyperparameters and justifies the need for automated optimization. All metrics have been calculated based on the 20% of the initial dataset, unknown to the models during the fitting process, and aggregated by multiplying their values label-wise (i.e. accuracy for anemia, accuracy for macrocytic type, etc.). At first glance, they seem to be performing much worse than other considered ML types, however, ensembling algorithms would show similar values if the same aggregation is applied (*e.g.* for accuracy in Bagging-Naïve Bayes, it would result in 0.897*0.868*0.941*0.926*0.985*0.971 = 0.6489 accuracy value), which makes it more sensitive to models misclassifying anemia and helps display the results of evaluation in a more concise manner. Top 5 best-performing architectures, sorted by accuracy, are shown in Table 11.

Table 11: Best-performing NN models.

#	n_units	layers	epochs	batch_size	activation	lr	accuracy	precision	recall	f1
1	10	1	100	20	relu	0.000708	0.647	0.691	0.137662	0.128
2	2	5	30	10	tanh	0.000478	0.632	0.973	0.142857	0.128
3	12	1	50	24	tanh	0.001187	0.632	0.973	0.142857	0.128
4	2	2	70	10	tanh	0.002682	0.632	0.973	0.142857	0.128
5	4	2	60	24	tanh	0.000458	0.632	0.973	0.142857	0.128

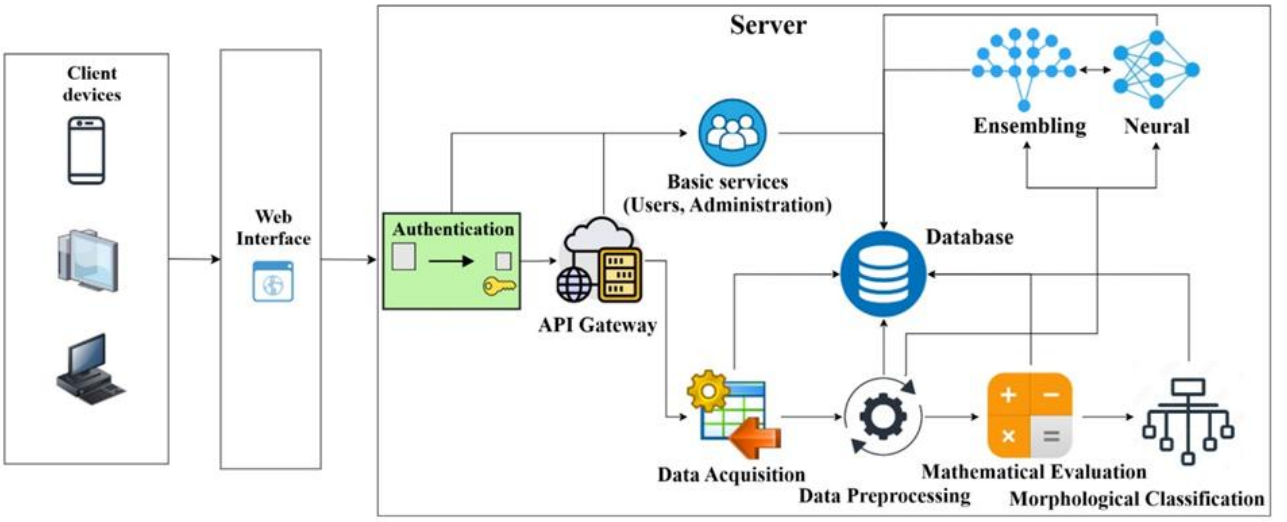

Fig. 18: The architectural model of a hardware and software system for diagnosing clinical and hematologic syndromes for the health passport.

All models, as also seen in Table 11, have shown high precision and low recall levels (their imbalance resulting in lower F1 scores), meaning that most of the detected anemic types are correct, but many cases remain undetected. The best-performing models have lower number of neurons overall (n_units*layers), which means that complicated models suffer from overfitting and while having a perfect 100% in learning accuracy, perform extremely poor on the testing dataset.

3.5 System architecture and clinical integration of the health passport platform

The architectural model of a hardware and software system for diagnosing clinical and hematologic syndromes for the Health Passport describes the structure and interaction between the various components of the system. It includes software and hardware components that work together to collect, process and analyze medical data. Fig. 18 shows a general schematic of the architectural model.

As illustrated in Fig. 18, the system architecture ensures seamless interaction between client devices (*e.g.*, smartphones, tablets, and computers) and the server-side analytical modules via a secure web interface. The API Gateway handles authentication and directs requests to core components such as data acquisition, preprocessing, and mathematical evaluation. A centralized database serves as the backbone for storing user data, processed inputs, and diagnostic outputs. Analytical modules include ensemble and neural network models, as well as a rule-based morphological classification system. This hybrid design supports flexible, scalable, and accurate diagnostics. The modularity of the system allows easy integration with hospital information systems and adaptability to new clinical guidelines or datasets, making it well-suited for deployment in both centralized medical facilities and remote e-health settings.

The machine learning algorithms and neural network models developed as part of the study were integrated into an interactive platform that performs the functions of a digital health passport. Each system module implements the key stages of automated anemia diagnostics - from data acquisition and preprocessing to classification and statistical analysis. The application, Health Passport, includes modules for data preprocessing, neural network construction, morphological classification, analysis of diagnosis statistics, and AI-based anemia prediction.

The practical feasibility of clinical deployment is supported by the existing software implementation of the Health Passport system, as shown in Fig. 19. The platform includes a fully developed graphical user interface (GUI) designed for healthcare professionals, featuring dedicated modules for data acquisition, preprocessing, diagnostic index calculation, morphological classification, and machine learning-based anemia prediction.

This modular design facilitates seamless integration with hospital information systems and supports real-time interaction with clinical workflows. Specifically, the interface allows users (*e.g.*, physicians, lab staff) to upload patient data, view prediction results, track diagnostic statistics, and interact with trained neural or ensemble models, all without the need for programming skills. The inclusion of patient history and diagnostic history modules also enables longitudinal tracking and decision support. These features provide the foundation for scalable clinical implementation, with potential extensions for EHR integration, model retraining, and improved explainability using visualizations.

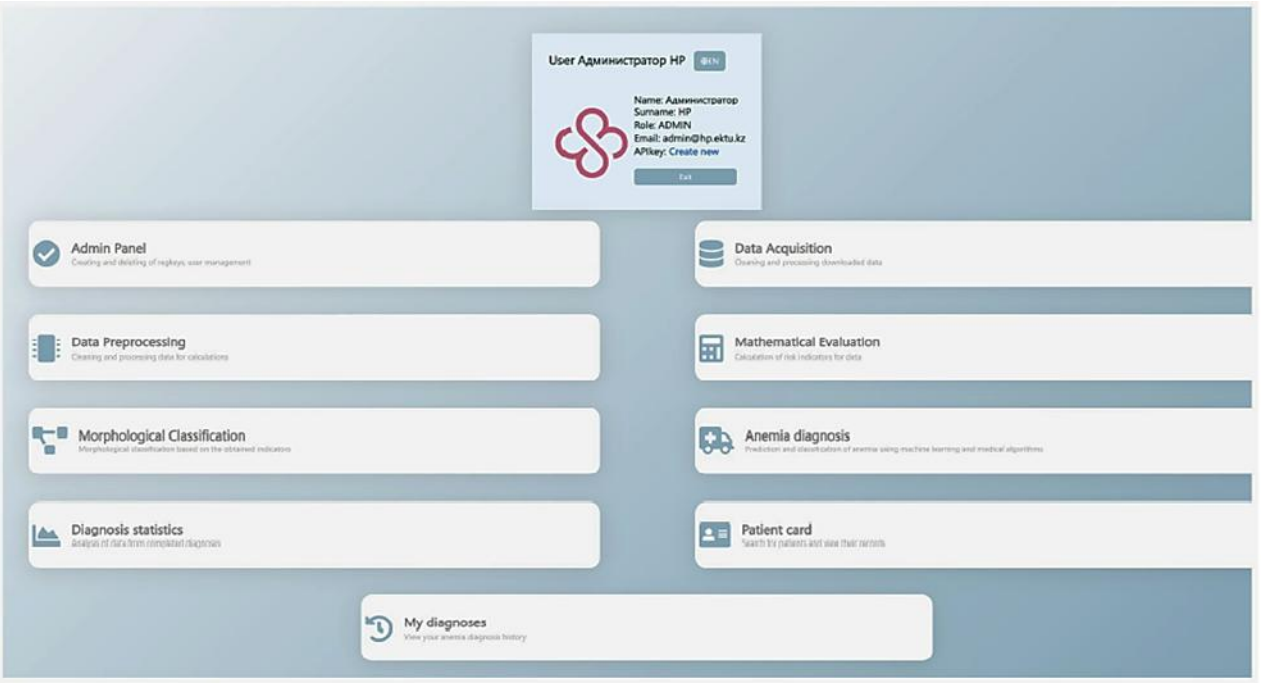

Fig. 19: Main window of the Health Passport software package.

This interface not only provides a user-friendly way to experiment with different ensembles and datasets, but also enables reproducibility and transparency in model selection. Users can visually compare performance metrics such as accuracy for each model, and easily switch between ensemble configurations without the need for deep programming expertise. Moreover, the ability to download prediction results supports seamless integration with clinical workflows or further research pipelines. Such modularity and interactivity enhance the system's practical applicability and make it accessible to a broader range of medical professionals and researchers.

3.6 Patterns and feature contribution

The results of the preliminary data analysis showed that women in the sample were more prone to anemia, which was consistent with clinical observations (*e.g.*, menstrual blood loss, pregnancy).[65] This fact can be used in split analysis by sex or building models based on Sex × Features interactions. The absolute number of anemia cases is lower in women (because there are fewer of them in the sample), but the relative incidence is higher. This once again confirmed the fact that gender is a significant predictor of the risk of anemia.

The heat maps built made it possible to understand which medical and laboratory indicators are important predictors for each type of anemia - and how to use this knowledge to build interpretable and effective models.

Based on the ranking of traits according to the degree of their contribution to the model of anemia prediction, the following conclusions can be drawn, presented in Table 12.

The results in Table 12 support clinical reasoning and demonstrate the model's alignment with medical expertise. HCT, RDW, and RBC emerged as the most influential predictors, highlighting the importance of core hematological parameters in anemia diagnostics. Notably, while HGB is traditionally used to define anemia, its slightly lower model weight compared to HCT suggests the added value of considering cell volume and distribution metrics. Biochemical markers like vitamin B12 and Fer had moderate importance, reinforcing their diagnostic role in specific anemia subtypes, such as macrocytic and iron-deficiency anemia. The relatively low direct contribution of sex as a feature indicates that its influence may be better captured through interaction effects or stratified modeling. These findings underscore the potential of combining clinical insight with data-driven feature evaluation to enhance the interpretability and robustness of machine learning–based anemia classification systems.

3.7 Performance interpretation and dataset limitations

3.7.1 Interpretation of model performance differences

The observed differences in model performance can be attributed to both algorithmic properties and data characteristics.[66] Among the ensemble models tested, boosting methods-particularly LightGBM and XGBoost-demonstrated the highest accuracy and stability across all anemia classification tasks. This can be explained by their ability to capture complex non-linear relationships and handle imbalanced data through instance weighting during training.

Table 12: Interpretation of the ranking graph of the importance of traits for predicting anemia.

Sign	Importance	Interpretation
HCT	~0.23	Hematocrit is the main indicator of anemia
RDW	0.18	The width of distribution of erythrocytes reflects anisocytosis
RBC	0.12	The number of red blood cells is directly related to anemia
TLC	0.11	Possibly reflects general inflammation or infection
HGB	0.09	Hemoglobin - high importance, but slightly below HCT
Age	0.09	Age affects predisposition
PLT/mm3, B12, Fer	0.05-0.07	Biochemistry has an effect, but less than hematology
Sex	0.01	Gender has little direct effect, but it can be an indirect factor

LightGBM consistently showed lower STD in cross-validation compared to XGBoost, likely due to its histogram-based gradient boosting and efficient handling of categorical features. In contrast, AdaBoost underperformed in sensitivity and F1-score, particularly on the "Normal" class. This may be due to its reliance on weak learners such as decision stumps, which are more susceptible to noise and less effective in capturing subtle class boundaries.[67]

Among ensemble methods, soft voting consistently outperformed hard voting across all classification tasks. This superiority is attributed to soft voting's probabilistic averaging of predictions, which allows the model to integrate the confidence levels of each base classifier. In contrast, hard voting may underperform in imbalanced or borderline cases, as it disregards probability scores and relies solely on majority class selection, which can misclassify low-prevalence anemia subtypes. Soft voting thus proves more suitable in medical diagnostics, where class overlap and sensitivity are critical. Stacking further improved performance by learning optimal combinations of base model predictions, leveraging their diverse decision boundaries. However, it requires more computational resources and is prone to overfitting if not regularized. These trade-offs justify our choice to prioritize soft voting and stacking as preferred ensemble approaches in this study.

Neural networks also achieved high training accuracy but exhibited reduced sensitivity (recall) for rare classes such as B12-deficiency anemia. This discrepancy suggests overfitting to dominant patterns in the dataset and reflects the networks' sensitivity to hyperparameter configurations. The use of Optuna for hyperparameter optimization revealed that simpler architectures (*e.g.*, fewer hidden layers and moderate learning rates) generalized better, while deeper models were more prone to overfitting, especially given the limited data available for minority classes.

Overall, ensemble algorithms outperformed individual classifiers and neural networks in terms of robustness and generalizability. Their ability to aggregate weak learners, reduce variance, and perform reliably in multi-label, imbalanced classification settings make them especially suitable for clinical diagnostic tasks. Neural networks, although flexible and capable of modeling complex feature interactions, require careful tuning and are more sensitive to sample size and class distribution.

3.7.2 Limitations and challenges in multi-label anemia classification

Despite the strong overall performance of the evaluated models, several challenges intrinsic to multi-label classification in clinical diagnostics must be acknowledged. Anemia subtypes often share overlapping clinical symptoms and laboratory markers, which introduces label ambiguity and inter-label correlation. For instance, normocytic anemia frequently co-occurs with chronic inflammation or B12 deficiency, making it difficult to establish clear class boundaries based solely on hematological indices.

A further complication arises from pronounced class imbalance and the limited number of samples in underrepresented subtypes, such as macrocytic and B12-deficiency anemia. This imbalance can lead to model bias toward more prevalent conditions and result in suboptimal sensitivity for rare but clinically significant classes. Although ensemble methods and neural network regularization can reduce overfitting,[68] they are not sufficient to fully address the degradation in performance for minority labels. Moreover, the use of macro-averaged performance metrics may mask these shortcomings by diluting the contribution of rare class predictions.[69]

To improve robustness and fairness, future work should include label dependence modeling approaches such as classifier chains or conditional random fields,[70] as well as advanced resampling methods including the synthetic minority method of extraction [71] and class-weighted loss functions.[72] Ensuring stratified training and testing splits is equally important to preserve the label distribution and prevent leaky or biased validation results.[73]

3.7.3 Generalizability and representativeness of the dataset

Given the limitations in the size and source of the dataset, we recognize the risk that our models may reflect sampling bias or not fully reflect the heterogeneity of the broader clinical population. The dataset used in this study, although rich in labeled features, was obtained from a single-source retrospective clinical database that may not reflect the full range of demographic, geographical features, or disease variability.

To solve this problem, we plan to evaluate the prepared models using external datasets from various medical institutions, including community hospitals, rural clinics, and multinational populations. This test will focus on the indicators of subgroups, especially depending on gender, age groups, and the etiology of anemia. Indicators such as recall for a specific subgroup and equity indicators will be used to identify and correct potential model errors.

In future work, we also intend to integrate domain adaptation and data augmentation strategies to better model underrepresented groups and rare clinical conditions. These improvements will allow models to adapt more reliably to real-world scenarios, reducing the risk of overfitting for a narrow group of patients and increasing the fairness and reliability of clinical trials.

3.8 Clinical applicability and deployment considerations

3.8.1 Scalability and clinical integration considerations

While the proposed machine learning models demonstrated high accuracy in classifying anemia and its subtypes on the given dataset, it is crucial to assess their scalability and applicability in real-world clinical environments. The ensemble algorithms, particularly LightGBM and XGBoost, are inherently scalable due to their support for parallel computation and efficient memory usage.[74] These properties make them suitable for deployment on larger and more heterogeneous datasets, such as those from multi-institutional electronic health records.

However, real-world clinical data are often more noisy, incomplete, and demographically diverse than the curated datasets used in this study. To ensure generalizability, future work should validate model performance on external datasets across different healthcare settings and populations. Additional steps such as domain adaptation, retraining with site-specific data, and incorporation of structured missingness handling will be essential for reliable deployment.

From a clinical workflow perspective, the integration of ML models into health information systems must be seamless, non-intrusive, and interpretable.[75] Models should provide clear, probabilistic outputs alongside human-readable explanations to support-rather than replace-clinical decision-making.

In summary, although the proposed models are technically scalable, their successful integration into clinical practice will require not only technical validation but also careful attention to human, regulatory, and infrastructural factors.

3.8.2 Clinical relevance and decision support impact

Although the performance indicators of the tested models, including accuracy, responsiveness, and F1 score, show excellent results (often exceeding 95%), it is extremely important to use these results in clinical practice. The practical value of the proposed ensemble and neural network models lies in their ability to automate the differential diagnosis of anemia subtypes, which are often impossible to distinguish from each other based solely on the threshold interpretation of CBC.

For example, it can be difficult to distinguish microcytic anemia from normocytic anemia when the MCV and MCH values approach the borderline values.[76] The models demonstrated that in microcytic and iron deficiency anemia, the sensitivity is more than 98%, which allows for earlier and targeted intervention.

By integrating these models into clinical workflows, such as decision support systems or electronic medical records, healthcare providers can receive real-time notifications of possible types of anemia, offering appropriate follow-up (such as ferritin or vitamin B12 testing), without relying solely on a doctor's interpretation.

3.9 Ethical considerations

The integration of ML algorithms into medical diagnostics introduces not only technical but also ethical challenges that must be addressed to ensure responsible and clinically acceptable deployment.

Firstly, data privacy and patient confidentiality remain critical concerns. Despite the use of anonymized datasets (*e.g.*, MIMIC), the potential for re-identification or misuse of sensitive medical information requires strict adherence to data governance standards and compliance with international data protection laws such as HIPAA or GDPR.

Secondly, ML models trained on imbalanced or demographically skewed datasets may inherit and perpetuate algorithmic bias. This may lead to disparities in diagnostic accuracy across patient subgroups, particularly affecting underrepresented populations. Ensuring diversity in training data and employing fairness-aware learning techniques are essential steps to mitigate these risks.

Thirdly, while ML-based systems such as the Health Passport offer valuable support in identifying and classifying

hematological conditions, they must not be regarded as replacements for clinical judgment. These tools are best positioned as DSS that augment clinicians' ability to make informed, timely, and individualized diagnoses. Final decision-making responsibility must always remain with trained healthcare professionals.

Future implementations of ML diagnostic tools should therefore prioritize transparency, explainability, auditability, and accountability, ensuring that model outputs are interpretable, verifiable, and ethically justifiable in real clinical environments.

4. Conclusion

The study demonstrated the high efficiency of machine learning in the diagnosis of anemia. The best results were achieved using efficiency improvement algorithms (in particular, LightGBM), as well as soft voting and grouping programs. These results support the development of an intelligent clinical decision support system for hematologists.

In general, the results of the experiment confirmed the central hypothesis: machine learning algorithms are able to accurately classify anemia and its subtypes based on laboratory data. Among all the evaluated models, efficiency enhancement algorithms, especially LightGBM, showed the highest accuracy and stability, with F1 scores and recall rates approaching 1.0 in all target classes, including microcytic, macrocytic, and B12-deficient anemia. This suggests that the models have successfully captured complex, potentially nonlinear relationships between hematological features such as HGB, RBC, RDW, and B12. A comparison with the existing literature once again confirms these conclusions. While most previous studies have focused on the binary classification of anemia (anemic and non-anemic), this work suggests a more clinically detailed approach using multiple labels, which makes it possible to distinguish between subtypes of morphological and deficiency anemia, which increases its practical value. To further enhance the reliability and clinical applicability of the proposed models, future research should focus on several key areas. First, expanding the dataset to include more diverse and representative clinical populations will improve generalizability and reduce demographic biases. Particular attention should be paid to balancing the underrepresented subtypes of anemia, such as vitamin B12 deficiency and macrocytic anemia, to eliminate class imbalance. Secondly, the introduction of domain adaptation and knowledge transfer methods can allow models to adapt to data from different hospitals or geographical regions. Third, validation in real-world clinical settings is necessary to assess the integration of models, their usability, and their impact on diagnostic work flows. Finally, collaboration with clinicians and hematologists is crucial to clarify diagnostic rules and bring machine-generated predictions in line with medical standards. These steps will help transform the existing system into a reliable, scalable, and ethically sound decision support tool for anemia diagnosis and clinical treatment.

Ethical statement

This study utilized the publicly available «MIMIC-IV Waveform Database Matched Subset» dataset (https://doi.org/10.13026/C2XW26) available through PhysioNet under the PhysioNet Credentialed Health Data Use Agreement version 1.5.0. Data are unidentified and were collected with Institutional Review Board (IRB) approval by the original researchers. This study does not involve direct interaction with humans or animals. This study also includes synthetic and retrospective analysis of data from the Passport to Health software package and meets all ethical standards for the use of secondary data. Thus, this study does not require ethical approval from an institutional review board or ethics committee.

Acknowledgments

This research was carried out with the support of grant funding for scientific and (or) scientific and technical projects for 2023-2025 of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant №AP19679525). The authors express their sincere gratitude to the staff of the medical organizations «Sau Med Group» LLP (Astana, Kazakhstan)[77] and «Aksel&A» LLP (Almaty, Kazakhstan)[78] for their expert participation in this study.

Conflict of Interest

The authors declare no conflicts of interest.

Supplementary Information

Not applicable.

Abbreviations

Abbreviation	Definition
ABS LYMP	Absolute Lymphocyte Count
ACD	Anemia of Chronic Disease
ANN	Artificial Neural Networks
B12	Vitamin B12
CBC	Complete Blood Count
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DSS	Decision-Support Systems
DT	Decision Tree
EDA	Exploratory Data Analysis
ELM	Extreme Learning Machine

Abbreviation	Definition
HCT	Hematocrit
HGB	Hemoglobin
IDA	Iron Deficiency Anemia
kNN	K-Nearest Neighbors
LightGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
LR	Logistic Regression
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
MCV	Mean Corpuscular Volume
MIMIC	Medical Information Mart for Intensive Care
ML	Machine Learning
PCA	Principal Component Method
PCT	Plateletcrit
PDW	Platelet Distribution Width
PLT	Platelet Count
RBC	Red Blood Cell Count
RDW	Red Cell Distribution Width
RNN	Recurrent Neural Network
SE	Error of the Mean
SHAP	SHapley Additive Explanations
STD	Standard Deviation
SVM	Support Vector Machine
TLC	Total Leukocyte Count
WBC	White Blood Cell Count
XGBoost	Extreme Gradient Boosting

CRediT Statement

Indira Uvaliyeva: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing, Visualization, Supervision. **Zhenisgul Rakhmetullina:** Conceptualization, Methodology, Formal analysis, Writing, Visualization. **David Borozenets:** Software, Validation, Investigation, Resources, Data Curation. **Farida Amenova:** Conceptualization, Methodology, Validation, Formal analysis, Writing. **Shynar Tezekpayeva:** Validation, Formal analysis, Investigation, Resources, Data Curation.

References

[1] D. T. Lee, M. L. Plesa, Anemia, *Family Medicine*, Cham: Springer International Publishing, 2022, 1815-1829, doi: 10.1007/978-3-030-54441-6_132.

[2] V. Sankar, A. Villa, Hematologic diseases, *Burket's Oral Medicine*, 2021, **17**, 627-664, doi: 10.1002/9781119597797.ch17.

[3] A. P. Singh, N. K. Maurya, R. Saxena, S. Saxena, An overview of red blood cell properties and functions, *Journal of International Research in Medical and Pharmaceutical Sciences*, 2024, **19**, 14-23, doi: 10.56557/jirmeps/2024/v19i28667.

[4] O. V. Chinelo., E. Chukwuka, A. C. Ifeoma, I. U. Stella N., B. Okoro Chinonso, C. Peterson Jesse, Causes of anemia due to diminished red blood cell production in pediatrics, *International Journal of Scientific Advances*, 2022, **3**(5) 711-8, doi: 10.51542/ijscia.v3i5.6.

[5] J. S. Brenner, S. Mitragotri, V. R. Muzykantov, Red blood cell hitchhiking: a novel approach for vascular delivery of nanocarriers, *Annual Review of Biomedical Engineering*, 2021, **23**, 225-248, doi: 10.1146/annurev-bioeng-121219-024239.

[6] M. R. Mohammed, B. Mahmood, Morphological types of anemia associated with chronic renal diseases, *Open Access Macedonian Journal of Medical Sciences*, 2022, **10**, 905-908, doi: 10.3889/oamjms.2022.9338.

[7] E. Wacka, J. Nicikowski, P. Jarmuzek, A. Zembron-Lacny, Anemia and its connections to inflammation in older adults: a review, *Journal of Clinical Medicine*, 2024, **13**, 2049, doi: 10.3390/jcm13072049.

[8] H. S. Mutlaq, K. A. T. Alotaibi, I. M. Husaykan, S. S. F. Alotaibi, F. H. H. Alotaibi, F. A. A. Alotibi, E. S. E. Alotaibi, A. A. Abualssayl, A. M. Mubarak, M. H. Almutairi, M. S. Binnwejim, M. A. A. Khabrani, Anemia: An Updated Review for Healthcare Professionals, *Journal of Ecohumanism*, 2024, **3**(8), 13225-13237, doi: 10.62754/joe.v3i8.6232

[9] V. Martinez-Torres, N. Torres, J. A. Davis, F. F. Corrales-Medina, Anemia and associated risk factors in pediatric patients, *Pediatric Health, Medicine and Therapeutics*, 2023, **14**, 267-280, doi: 10.2147/PHMT.S389105.

[10] R. Maulide Cane, J. B. Chidassicua, L. Varandas, I. Craveiro, Anemia in pregnant women and children aged 6 to 59 months living in Mozambique and Portugal: an overview of systematic reviews, *International Journal of Environmental Research and Public Health*, 2022, **19**, 4685, doi: 10.3390/ijerph19084685.

[11] J. Sun, H. Wu, M. Zhao, C. G. Magnussen, B. Xi, Prevalence and changes of anemia among young children and women in 47 low- and middle-income countries, 2000-2018, *eClinicalMedicine*, 2021, **41**, 101136, doi: 10.1016/j.eclinm.2021.101136.

[12] M. Karami, M. Chaleshgar, N. Salari, H. Akbari, M. Mohammadi, Global prevalence of anemia in pregnant women: a comprehensive systematic review and meta-analysis, *Maternal and Child Health Journal*, 2022, **26**, 1473-1487, doi: 10.1007/s10995-022-03450-1.

[13] G. A. Stevens, C. J. Paciorek, M. C. Flores-Urrutia, E. Borghi, S. Namaste, J. P. Wirth, P. S. Suchdev, M. Ezzati, F. Rohner, S. R. Flaxman, L. M. Rogers, National, regional, and global estimates of anaemia by severity in women and children for 2000–19: a pooled analysis of population-representative data, *The Lancet Global Health*, 2022, **10**, e627-e639, doi: 10.1016/S2214-109X(22)00084-5.

[14] A. K. B. Rivera, A. A. E. Latorre, K. Nakamura, K. Seino, Using complete blood count parameters in the diagnosis of iron deficiency and iron deficiency anemia in Filipino women, *Journal*

of Rural Medicine, 2023, **18**, 79-86, doi: 10.2185/jrm.2022-047.
[15] K. Doig, B. Zhang, A methodical approach to interpreting the red blood cell parameters of the complete blood count, *American Society for Clinical Laboratory Science*, 2017, **30**, 173-185, doi: 10.29074/ascls.30.3.173.
[16] B. F. Casanova, M. D. Sammel, G. A. Macones, Development of a clinical prediction rule for iron deficiency anemia in pregnancy, *American Journal of Obstetrics and Gynecology*, 2005, **193**, 460-466, doi: 10.1016/j.ajog.2004.12.008.
[17] N. K T, K. Prasad, B. M. K. Singh, Analysis of red blood cells from peripheral blood smear images for anemia detection: a methodological review, *Medical & Biological Engineering & Computing*, 2022, **60**, 2445-2462, doi: 10.1007/s11517-022-02614-z.
[18] S. Farrukh, Q. U. A. Ayyaz, F. Ali Khanzada, H. Sheikh, A. Anwar, S. Cheema, Comparison of classification of anemia based on mean corpuscular volume by hematology analyzer and peripheral smear examination, *Pakistan Journal of Pathology*, 2024, **35**, 81-86, doi: 10.55629/pakjpathol.v35i2.793.
[19] R. An, Y. Huang, Y. Man, R. W. Valentine, E. Kucukal, U. Goreke, Z. Sekyonda, C. Piccone, A. Owusu-Ansah, S. Ahuja, J. A. Little, U. A. Gurkan, Emerging point-of-care technologies for anemia detection, *Lab on a Chip*, 2021, **21**, 1843-1865, doi: 10.1039/d0lc01235a.
[20] A. E. Benson, M. C. Smid. Maternal Anemia. Maternal-Fetal Evidence Based Guidelines, *CRC Press*, 2022, 144-152. ISBN: 9781003099062.
[21] J. Prajapati, V. Uduthalapally, D. Das, R. Mahapatra, P. N. Wasnik, XAIA: an explainable AI approach for classification and analysis of blood anemia, *OITS International Conference on Information Technology (OCIT)*, December 13-15, 2023, Raipur, India, IEEE, 88-93, doi: 10.1109/OCIT59427.2023.10430938.
[22] E. S. Aslan, S. Gür, Evaluation and epigenetic impact of B12, vitamin D, folic acid and anemia in Hashimoto's thyroiditis: a clinical and molecular docking study, *Journal of Health Sciences and Medicine*, 2023, **6**, 705-712, doi: 10.32322/jhsm.1243597.
[23] M. Ramzan, J. Sheng, M. U. Saeed, B. Wang, F. Z. Duraihem, Revolutionizing anemia detection: integrative machine learning models and advanced attention mechanisms, *Visual Computing for Industry, Biomedicine and Art*, 2024, **7**, 18, doi: 10.1186/s42492-024-00169-4.
[24] J. W. Asare, P. Appiahene, E. T. Donkoh, G. Dimauro, Iron deficiency anemia detection using machine learning models: A comparative study of fingernails, palm and conjunctiva of the eye images, *Engineering Reports*, 2023, **5**, e12667, doi: 10.1002/eng2.12667.
[25] D. C. E. Saputra, K. Sunat, T. Ratnaningsih, A new artificial intelligence approach using extreme learning machine as the potentially effective model to predict and analyze the diagnosis of anemia, *Healthcare*, 2023, **11**, 697, doi: 10.3390/healthcare11050697.
[26] P. Dhakal, Prediction of anemia using machine learning algorithms, *International Journal of Computer Science and Information Technology*, 2023, **15**, 15-30, doi: 10.5121/ijcsit.2023.15102.
[27] S. Pullakhandam, S. McRoy, Classification and explanation of iron deficiency anemia from complete blood count data using machine learning, *BioMedInformatics*, 2024, **4**, 661-672, doi: 10.3390/biomedinformatics4010036.
[28] D. D. B S, P. Sharma, K. Chadaga, N. Sampathila, G. M. Bairy, S. Belurkar, S. Prabhu, S. K S, An ensemble machine learning framework with explainable artificial intelligence for predicting haemoglobin anaemia considering haematological markers, *Systems Science & Control Engineering*, 2024, **12**, 2420927, doi: 10.1080/21642583.2024.2420927.
[29] T. Deotale, S. Saha, Review of machine learning applications in HB detection and anaemia screening prognosis under clinical conditions. *Advances in Information and Communication*, Cham: Springer Nature Switzerland, 2025, **1285** 124-155, doi: 10.1007/978-3-031-84460-7_9.
[30] D. K. Thakur, D. Mishra, Exploratory data analysis for detecting of iron deficiency using machine learning techniques, *7th International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, Greater Noida, India, September 18-20, 2024, 1130-1135, doi: 10.1109/IC3I61595.2024.10828826.
[31] B. S. Dhruva Darshan, N. Sampathila, G. Muralidhar Bairy, S. Prabhu, S. Belurkar, K. Chadaga, S. Nandish, Differential diagnosis of iron deficiency anemia from aplastic anemia using machine learning and explainable Artificial Intelligence utilizing blood attributes, *Scientific Reports*, 2025, **15**, 505, doi: 10.1038/s41598-024-84120-w.
[32] A. Iolascon, I. Andolfo, R. Russo, M. Sanchez, F. Busti, D. Swinkels, P. Aguilar Martinez, R. Bou-Fakhredin, M. U. Muckenthaler, S. Unal, G. Porto, T. Ganz, A. Kattamis, L. De Franceschi, M. D. Cappellini, M. G. Munro, A. Taher, from EHA-SWG Red Cell and Iron, Recommendations for diagnosis, treatment, and prevention of iron deficiency and iron deficiency anemia, *HemaSphere*, 2024, **8**, e108, doi: 10.1002/hem3.108.
[33] M. N. Garcia-Casal, O. Dary, M. E. Jefferds, S. R. Pasricha, diagnosing anemia: Challenges selecting methods, addressing underlying causes, and implementing actions at the public health level, *Annals of the New York Academy of Sciences*, 2023, **1524**, 37-50, doi: 10.1111/nyas.14996.
[34] C. U. Loechl, A. Datta-Mitra, L. Fenlason, R. Green, L. Hackl, L. Itzkowitz, M. Koso-Thomas, D. Moorthy, V. O. Owino, H. Pachón, N. Stoffel, M. B. Zimmerman, D. J. Raiten, Approaches to address the anemia challenge, *The Journal of*

Nutrition, 2023, **153**, S42-S59, doi: 10.1016/j.tjnut.2023.07.017.
[35] Y. Deivita, S. Syafruddin, U. Andi Nilawati, A. Aminuddin, B. Burhanuddin, Z. Zahir, Overview of Anemia; risk factors and solution offering, *Gaceta Sanitaria*, 2021, **35**, S235-S241, doi: 10.1016/j.gaceta.2021.07.034.
[36] J. A. Rusch, D. J. van der Westhuizen, R. S. Gill, V. J. Louw, Diagnosing iron deficiency: Controversies and novel metrics, *Best Practice & Research Clinical Anaesthesiology*, 2023, **37**, 451-467, doi: 10.1016/j.bpa.2023.11.001.
[37] N. Svenson, J. Bailey, S. Durairaj, N. Dempsey-Hibbert, A simplified diagnostic pathway for the differential diagnosis of iron deficiency anaemia and anaemia of chronic disease, *International Journal of Laboratory Hematology*, 2021, **43**, 1644-1652, doi: 10.1111/ijlh.13666.
[38] M. Rohr, V. Brandenburg, H. P. Brunner-La Rocca, how to diagnose iron deficiency in chronic disease: A review of current methods and potential marker for the outcome, *European Journal of Medical Research*, 2023, **28**, 15, doi: 10.1186/s40001-022-00922-6.
[39] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, T. Gonzales-Yanac, Multiple disease prediction using Machine learning algorithms, *Materials Today: Proceedings*, 2023, **80**, 3682-3685, doi: 10.1016/j.matpr.2021.07.361.
[40] M. Nabeel, S. Majeed, M. Javed Awan, H. M. ud-Din, M. Wasique, R. Nasir, Review on effective disease prediction through data mining techniques, *International Journal on Electrical Engineering and Informatics*, 2021, **13**, 717-733, doi: 10.15676/ijeei.2021.13.3.13.
[41] M. Saberi-Karimian, Z. Khorasanchi, H. Ghazizadeh, M. Tayefi, S. Saffar, G. A. Ferns, M. Ghayour-Mobarhan, Potential value and impact of data mining and machine learning in clinical diagnostics, *Critical Reviews in Clinical Laboratory Sciences*, 2021, **58**, 275-296, doi: 10.1080/10408363.2020.1857681.
[42] S. G. Kanakaraddi, K. C. Gull, J. Bali, A. K. Chikaraddi, S. Giraddi, Disease prediction using data mining and machine learning techniques, *Advanced Prognostic Predictive Modelling in Healthcare Data Analytics*, Singapore: Springer Singapore, 2021, 71-92, doi: 10.1007/978-981-16-0538-3_4.
[43] M. Pammi, N. Aghaeepour, J. Neu, Multiomics, artificial intelligence, and precision medicine in perinatology, *Pediatric Research*, 2023, **93**, 308-315, doi: 10.1038/s41390-022-02181-x.
[44] A. E. Obstfeld, Hematology and machine learning, *The Journal of Applied Laboratory Medicine*, 2023, **8**, 129-144, doi: 10.1093/jalm/jfac108.
[45] B. Hemalatha, B. Karthik, C. V. Krishna Reddy, A. Latha, Deep learning approach for segmentation and classification of blood cells using enhanced CNN, *Measurement: Sensors*, 2022, **24**, 100582, doi: 10.1016/j.measen.2022.100582.
[46] V. Geetha, C. K. Gomathy, K. Keerthi, N. Pavithra, Diagnostic approach to anemia in adults using machine learning, *Journal of Pharmaceutical Negative Results*, 2022, **13**(9), 3713-3717, doi: 10.47750/pnr.2022.13. S09.456.
[47] R. Vohra, A. Hussain, A. K. Dudyala, J. Pahareeya, W. Khan, Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting, *PLoS One*, 2022, **17**, e0269685, doi: 10.1371/journal.pone.0269685.
[48] A. Zhang, J. Lou, Z. Pan, J. Luo, X. Zhang, H. Zhang, J. Li, L. Wang, X. Cui, B. Ji, L. Chen, Prediction of anemia using facial images and deep learning technology in the emergency department, *Frontiers in Public Health*, 2022, **10**, 964385, doi: 10.3389/fpubh.2022.964385.
[49] A. Kumar, V. Kumar, D. Debnath, D. Sorout, P. Chawla, S. Tyagi, Predicting blood hemoglobin level with machine learning and spectroscopy, *2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT),* IEEE, Faridabad, India, November 28-29, 2024, 710-716, doi: 10.1109/ICAICCIT64383.2024.10912215.
[50] S. Rane, A. Yadav, G. Patel, R. Gurjwar, A. Barve, K. Gagan Kumar, A survey on the use of machine learning approaches for analysis of anemia, *Machine Learning and Information Processing: Proceedings of Icmlip*, Ranchi, India, AIP Publishing, December 11, 2023, 2855, doi: 10.1063/5.0174115.
[51] A. Farjana, F. T. Liza, P. P. Pandit, M. C. Das, M. Hasan, F. Tabassum, M. H. Hossen, predicting chronic kidney disease using machine learning algorithms, *IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, March 8-11, 2023, 1267-1271, doi: 10.1109/CCWC57344.2023.10099221.
[52] U. Indira, S. Belginova, A. Ismukhamedova, Informational and Analytical System to Diagnose Anemia *Proceedings of the Fourth International Conference on Engineering & MIS*, Istanbul Turkey, ACM, 18, 2018, 1-8, doi: 10.1145/3234698.3234716.
[53] Z. Hevessy, G. Toth, P. Antal-Szalmas, M. Tokes-Fuzesi, J. Kappelmayer, B. Karai, E. Ajzner, W. G. O. Guidelines, Algorithms of the Hungarian Society of Laboratory Medicine, Algorithm of differential diagnosis of anemia involving laboratory medicine specialists to advance diagnostic excellence, *Clinical Chemistry and Laboratory Medicine*, 2023, **62**, 410-420, doi: 10.1515/cclm-2023-0807.
[54] H. Tvedten, Classification and laboratory evaluation of anemia, *Schalm's veterinary hematology*, 2022, 198-208, doi: 10.1002/9781119500537.ch25.
[55] A.E.W. Johnson, L. Bulgarelli, L. Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset, *Scientific data*, 2023, **10**(1), 1, doi: 10.1038/s41597-022-01899-x.
[56] G. B. Moody, PhysioNet, *Encyclopedia of Computational Neuroscience*, New York, NY: Springer New York, 2022, 2806-

2808, doi: 10.1007/978-1-0716-1006-0_496.
[57] A. S. Rao, B. V. Vardhan, H. Shaik, Role of Exploratory Data Analysis in Data Science, *6th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, Coimbatore, India, July 8-10, 2021, 1457-1461, doi: 10.1109/icces51350.2021.9488986.
[58] S. Dhummad, The imperative of exploratory data analysis in machine learning, *Scholars Journal of Engineering and Technology*, 2025, **13**, 30-44, doi: 10.36347/sjet.2025.v13i01.005.
[59] Y. Wu, L. Liu, Z. Xie, K. H. Chow, W. Wei, Boosting ensemble accuracy by revisiting ensemble diversity metrics, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 16469-16477, url: https://openaccess.thecvf.com/content/CVPR2021/html/Wu_Boosting_Ensemble_Accuracy_by_Revisiting_Ensemble_Diversity_Metrics_CVPR_2021_paper.html.
[60] Z. Li, K. Ren, Y. Yang, X. Jiang, Y. Yang, D. Li, Towards inference efficient deep ensemble learning, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, **37**(7), 8711-8719, doi: 10.1609/aaai.v37i7.26048.
[61] G. Ngo, R. Beard, R. Chandra, Evolutionary bagging for ensemble learning, *Neurocomputing*, 2022, **510**, 1-14, doi: 10.1016/j.neucom.2022.08.055.
[62] M. Krichen, Convolutional neural networks: a survey, *Computers*, 2023, **12**, 151, doi: 10.3390/computers12080151.
[63] M. Franzese, A. Iuliano, Correlation analysis, *Encyclopedia of Bioinformatics and Computational Biology*, Amsterdam: Elsevier, 2018, 706-721, doi: 10.1016/b978-0-12-809633-8.20358-0.
[64] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, E. Tuzhilina, Principal component analysis, *Nature Reviews Methods Primers*, 2022, **2**, 100, doi: 10.1038/s43586-022-00184-w.
[65] M. E. Bentley, P. L. Griffiths, The burden of anemia among women in India, *European Journal of Clinical Nutrition*, 2003, **57**, 52-60, doi: 10.1038/sj.ejcn.1601504.
[66] O. Kwon, J. M. Sim, Effects of data set features on the performances of classification algorithms, *Expert Systems with Applications*, 2013, **40**, 1847-1857, doi: 10.1016/j.eswa.2012.09.017.
[67] J. M. Johnson, T. M. Khoshgoftaar, Survey on deep learning with class imbalance, *Journal of Big Data*, 2019, **6**, 27, doi: 10.1186/s40537-019-0192-5.
[68] C. F. G. Dos Santos, J. P. Papa, Avoiding overfitting: a survey on regularization methods for convolutional neural networks, *ACM Computing Surveys*, 2022, **54**, 1-25, doi: 10.1145/3510413.
[69] M. I. H. Siddiqui, S. Khan, Z. H. Limon, H. Rahman, M. A. Khan, A. Al Sakib, S. M. M. R. Swapno, R. Haque, A. W. Reza, A. Appaji, Accelerated and accurate cervical cancer diagnosis using a novel stacking ensemble method with explainable AI, *Informatics in Medicine Unlocked*, 2025, **56**, 101657, doi: 10.1016/j.imu.2025.101657.
[70] X. Wei, H. Wang, Stochastic stratigraphic modeling using Bayesian machine learning, *Engineering Geology*, 2022, **307**, 106789, doi: 10.1016/j.enggeo.2022.106789.
[71] I. M. Alkhawaldeh, I. Albalkhi, A. J. Naswhan, Challenges and limitations of synthetic minority oversampling techniques in machine learning, *World Journal of Methodology*, **13**, 373-378, doi: 10.5662/wjm.v13.i5.373.
[72] J. Terven, D.-M. Cordova-Esparza, J.-A. Romero-González, A. Ramírez-Pedraza, E. A. Chávez-Urbiola, A comprehensive survey of loss functions and metrics in deep learning, *Artificial Intelligence Review*, 2025, **58**, 195, doi: 10.1007/s10462-025-11198-7.
[73] J. Sadaiyandi, P. Arumugam, A. K. Sangaiah, C. Zhang, Stratified sampling-based deep learning approach to increase prediction accuracy of unbalanced dataset, *Electronics*, 2023, **12**, 4423, doi: 10.3390/electronics12214423.
[74] N. Rane, S. P. Choudhary, J. Rane, Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions, *Studies in Medical and Health Sciences*, 2024, **1**, 18-41, doi: 10.48185/smhs.v1i2.1225.
[75] A. Sheikh, M. Anderson, S. Albala, B. Casadei, B. D. Franklin, M. Richards, D. Taylor, H. Tibble, E. Mossialos, Health information technology and digital innovation for national learning health and care systems, *The Lancet Digital Health*, 2021, **3**, e383-e396, doi: 10.1016/S2589-7500(21)00005-4.
[76] T. W. Htut, K. Z. Thein, T. H. Oo, Pernicious anemia: pathophysiology and diagnostic difficulties, *Journal of Evidence-Based Medicine*, 2021, **14**, 161-169, doi: 10.1111/jebm.12435.
[77] "Sau Med Group" LLP. (n.d.). Official website. Retrieved June 02, 2025, from https://saumedgroup.kz/
[78] "Aksel&A" LLP. (n.d.). Official website. Retrieved June 02, 2025, from https://aksel.kz/