PUBLIC HEALTH
HACKATHON
ALMATY 2025

Early Detection of Blood Disorders: Predictive Modeling from Lab and Outpatient Data

AltynCode
Dmitry Shishkanov

GitHub Repository

Repository link

Hematology. ICD-10 classification

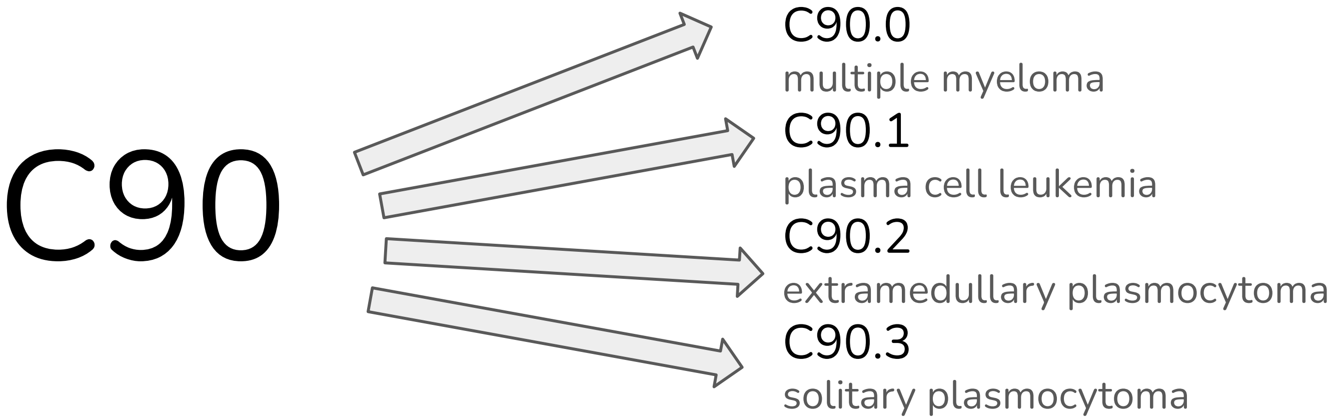
Non-malignant

1) Anemias (D50-D53)
2) Hemolytic anemias (D55-D59)
3) Apalastic anemias and other anemias (D60-D64)
4) Disorders of hemostasis (D65-D68)
5) Purpura and other conditions (D69)
6) Neutropenia (D70)
7) Other diseases of WBC (D72)
8) Other specified diseases of blood organs (D75)
9) Immune mechanism disorders (D80-D89)

Malignant

1) Hodgkin lymphoma
2) Non-Hodgkin lymphoma (C82-86, C85.9)
3) Multiple myeloma (C90)
4) Leukemias (C91-C95)

Clinical appearance and differential diagnosis



C90

C90.0
multiple myeloma

C90.1
plasma cell leukemia

C90.2
extramedullary plasmocytoma

C90.3
solitary plasmocytoma

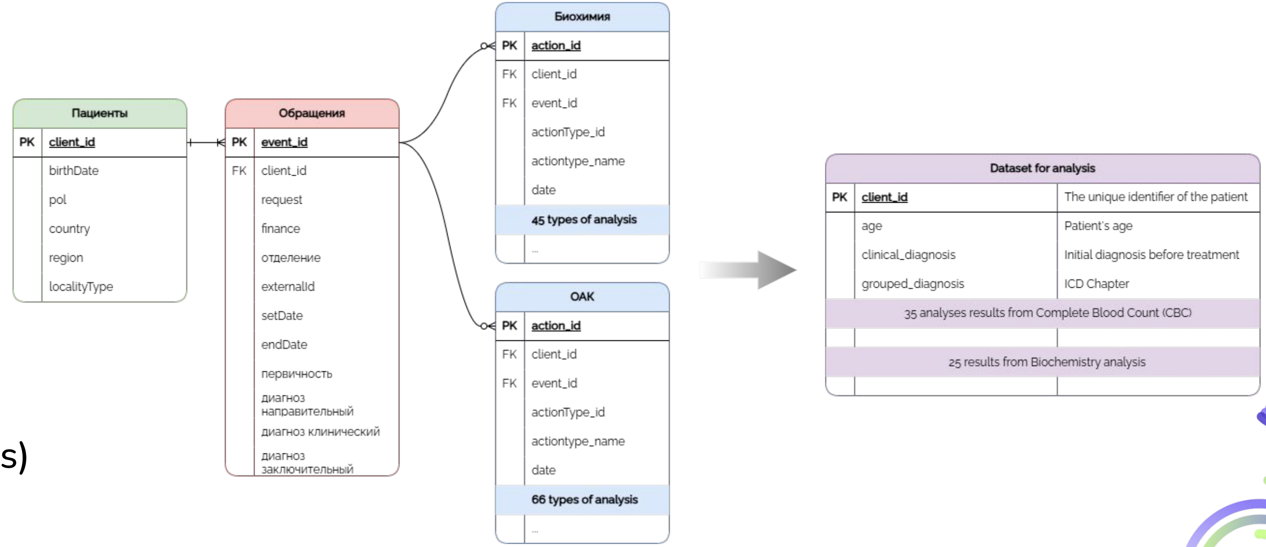
Dataset analysis and preparation

40000 patients diagnosed and treated in NMRC for hematology (Moscow, Russia) with 2014 - 2024 yy.

ICD-10:

- 21 groups of diseases
- 50 diseases
- Healthy group (bone marrow and blood donors)

- **Laboratory tests:**
  CBC: 52 features
- Biochemistry: 17 features



Пациенты	
PK	client_id
	birthDate
	pol
	country
	region
	localityType

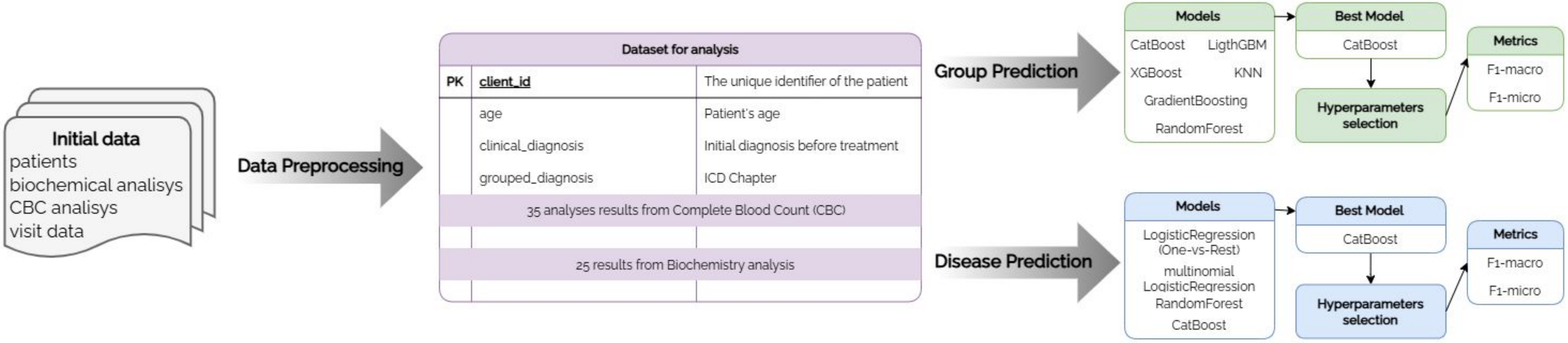
Обращения	
PK	event_id
FK	client_id
	request
	finance
	отделение
	externalId
	setDate
	endDate
	первичность
	диагноз направительный
	диагноз клинический
	диагноз заключительный

Биохимия	
PK	action_id
FK	client_id
FK	event_id
	actionType_id
	actiontype_name
	date
	45 types of analysis
	...

ОАК	
PK	action_id
FK	client_id
FK	event_id
	actionType_id
	actiontype_name
	date
	66 types of analysis
	...

Dataset for analysis		
PK	client_id	The unique identifier of the patient
	age	Patient's age
	clinical_diagnosis	Initial diagnosis before treatment
	grouped_diagnosis	ICD Chapter
	35 analyses results from Complete Blood Count (CBC)	
	25 results from Biochemistry analysis	

Workflow



Dataset for analysis

PK	client_id	The unique identifier of the patient
	age	Patient's age
	clinical_diagnosis	Initial diagnosis before treatment
	grouped_diagnosis	ICD Chapter
35 analyses results from Complete Blood Count (CBC)		
25 results from Biochemistry analysis		

Different models for grouped diagnosis

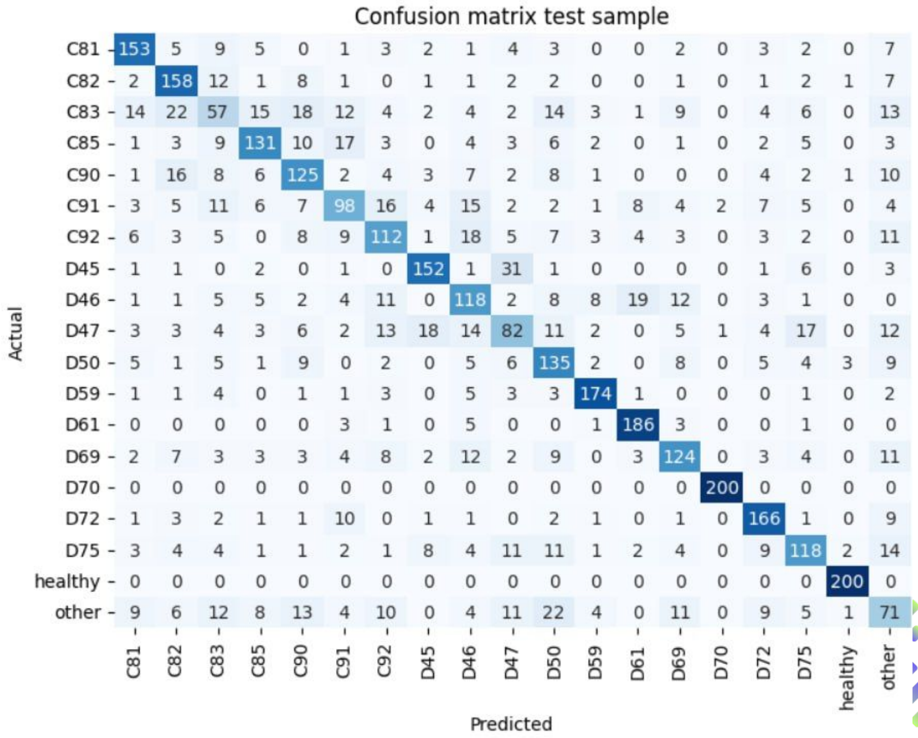
C81, C82, C83, C90, C91, C92, D45, D46, D47, D50, D59, D61, D69, D70, D72, D75, healthy (Z52.3), others

Table with **validation** metrics of different models on categories

Grouped Disease	Accuracy	F1-macro	F1-micro	Precision	Recall
models					
CatBoost	0.679934	**0.670501**	0.679934	0.667430	0.679934
XGboost	0.678684	**0.671063**	0.678684	0.668267	0.678684
LigthGBM	0.631842	**0.621459**	0.631842	0.620270	0.631842
KNN	0.44	0.42	0.44	0.44	0.44
LogistRegression (OvR)	0.42	0.4	0.42	0.43	0.43
LogistRegression (Softmax)	0.42	0.4	0.42	0.43	0.43
GradientBoosting	0.58	0.56	0.58	0.557	0.57
RandomForest	0.65	0.66	0.67	0.66	0.67

Best CatBoost model for grouped diseases

	Accuracy	F1-macro	F1-micro
CatBoost	0.673684	0.668039	0.673684



	C81	C82	C83	C85	C90	C91	C92	D45	D46	D47	D50	D59	D61	D69	D70	D72	D75	healthy	other
C81	153	5	9	5	0	1	3	2	1	4	3	0	0	2	0	3	2	0	7
C82	2	158	12	1	8	1	0	1	1	2	2	0	0	1	0	1	2	1	7
C83	14	22	57	15	18	12	4	2	4	2	14	3	1	9	0	4	6	0	13
C85	1	3	9	131	10	17	3	0	4	3	6	2	0	1	0	2	5	0	3
C90	1	16	8	6	125	2	4	3	7	2	8	1	0	0	0	4	2	1	10
C91	3	5	11	6	7	98	16	4	15	2	2	1	8	4	2	7	5	0	4
C92	6	3	5	0	8	9	112	1	18	5	7	3	4	3	0	3	2	0	11
D45	1	1	0	2	0	1	0	152	1	31	1	0	0	0	0	1	6	0	3
D46	1	1	5	5	2	4	11	0	118	2	8	8	19	12	0	3	1	0	0
D47	3	3	4	3	6	2	13	18	14	82	11	2	0	5	1	4	17	0	12
D50	5	1	5	1	9	0	2	0	5	6	135	2	0	8	0	5	4	3	9
D59	1	1	4	0	1	1	3	0	5	3	3	174	1	0	0	0	1	0	2
D61	0	0	0	0	0	3	1	0	5	0	0	1	186	3	0	0	1	0	0
D69	2	7	3	3	3	4	8	2	12	2	9	0	3	124	0	3	4	0	11
D70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	200	0	0	0	0
D72	1	3	2	1	1	10	0	1	1	0	2	1	0	1	0	166	1	0	9
D75	3	4	4	1	1	2	1	8	4	11	11	1	2	4	0	9	118	2	14
healthy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	200	0
other	9	6	12	8	13	4	10	0	4	11	22	4	0	11	0	9	5	1	71

Models for concrete diseases

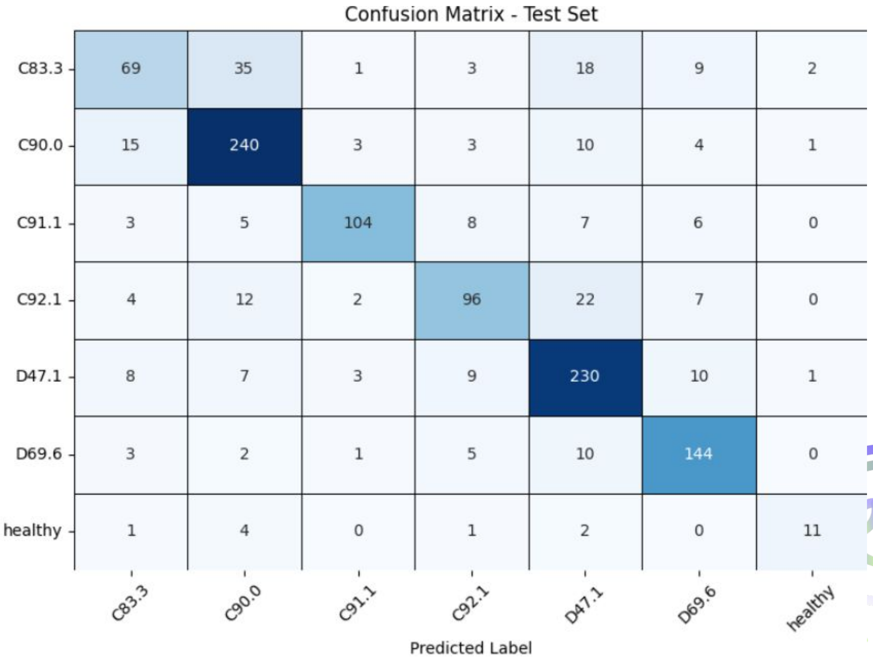
D68.8, C90.0, D47.1, D68.9, D66, D69.6, C92.1, C83.3, C91.1, C92.0, D45, D69.5, D69.3, C91.0, C81.1, C91.4, D46.9, D50, D47.4, D47.3, D61.3, I89.8, D50.9, C85.7, D68.0, D75.9, E75.2, D64.9, D72, D72.9, C82.0, D68.5, C81, D75, C92.4, C83.0, I70.8, D50.8, C82, D59.1, D75.1, C85.9, D47.2, C88.0, D75.2, D46.7, M16.1, D56.1, C90

Table with **validation** metrics of different models on concrete diseases

	Accuracy	F1-macro	F1-micro
LogistRegression (OvR)	0.733537	0.675154	0.723537
LogistRegression (Softmax)	0.731009	0.685489	0.731009
RandomForest	0.813200	0.747070	0.813200

Defolt CatBoost model for concrete diseases

	Accuracy	F1-macro	F1-micro
CatBoost	0.796276	0.765584	0.796276

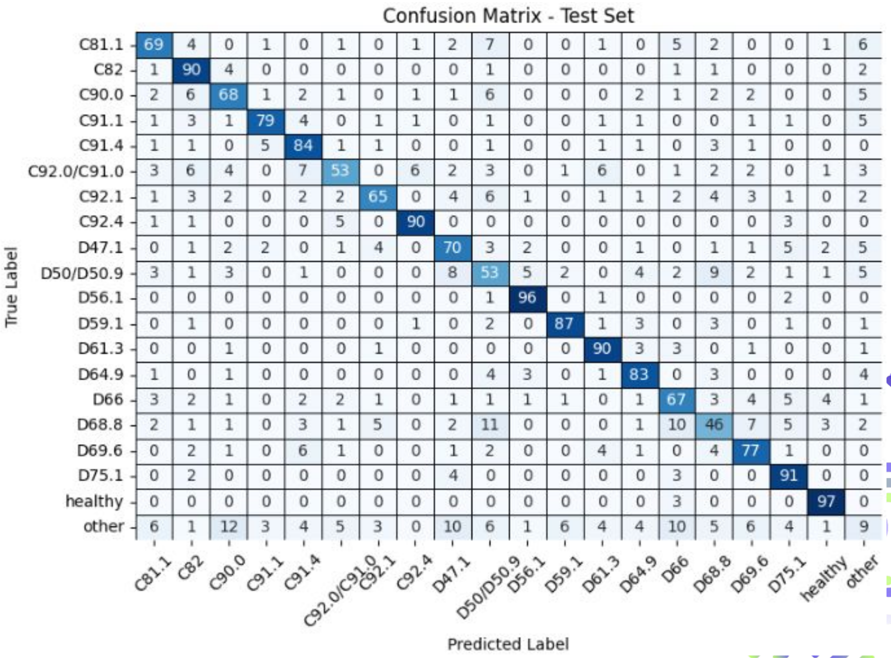


Confusion Matrix - Test Set

	C83.3	C90.0	C91.1	C92.1	D47.1	D69.6	healthy
C83.3	69	35	1	3	18	9	2
C90.0	15	240	3	3	10	4	1
C91.1	3	5	104	8	7	6	0
C92.1	4	12	2	96	22	7	0
D47.1	8	7	3	9	230	10	1
D69.6	3	2	1	5	10	144	0
healthy	1	4	0	1	2	0	11

Best Model with wide range of diseases

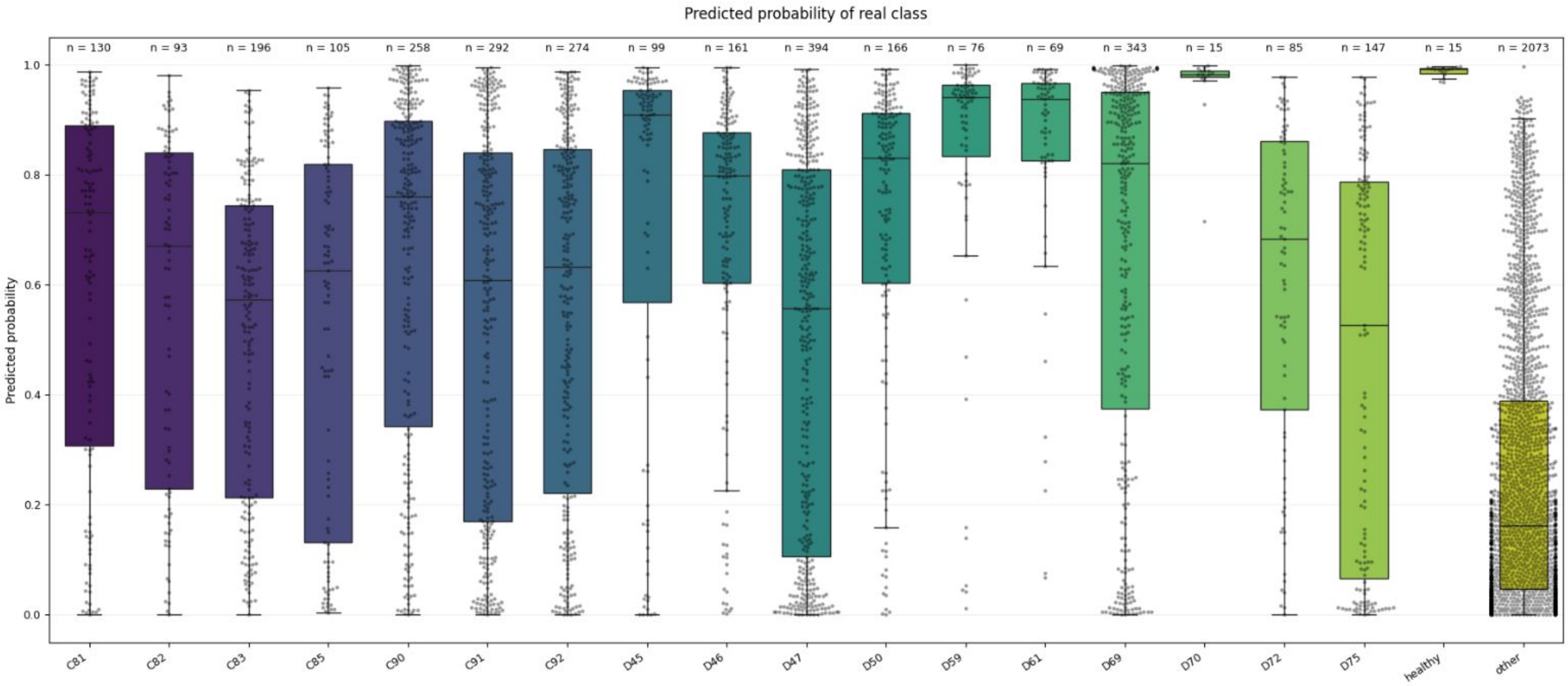
	Accuracy	F1-macro	F1-micro
CatBoost	0.732000	0.721436	0.732000



Confusion Matrix - Test Set

	C81.1	C82	C90.0	C91.1	C91.4	C92.0/C91.0	C92.1	C92.4	D47.1	D50/D50.9	D56.1	D59.1	D61.3	D64.9	D66	D68.8	D69.6	D75.1	healthy	other
C81.1	69	4	0	1	0	1	0	1	2	7	0	0	1	0	5	2	0	0	1	6
C82	1	90	4	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	2
C90.0	2	6	68	1	2	1	0	1	1	6	0	0	0	2	1	2	2	0	0	5
C91.1	1	3	1	79	4	0	1	1	0	1	0	0	1	1	0	0	1	1	0	5
C91.4	1	1	0	5	84	1	1	0	0	1	0	0	1	1	0	3	1	0	0	0
C92.0/C91.0	3	6	4	0	7	53	0	6	2	3	0	1	6	0	1	2	2	0	1	3
C92.1	1	3	2	0	2	2	65	0	4	6	1	0	1	1	2	4	3	1	0	2
C92.4	1	1	0	0	0	5	0	90	0	0	0	0	0	0	0	0	0	3	0	0
D47.1	0	1	2	2	0	1	4	0	70	3	2	0	0	1	0	1	1	5	2	5
D50/D50.9	3	1	3	0	1	0	0	0	8	53	5	2	0	4	2	9	2	1	1	5
D56.1	0	0	0	0	0	0	0	0	0	1	96	0	1	0	0	0	0	2	0	0
D59.1	0	1	0	0	0	0	0	1	0	2	0	87	1	3	0	3	0	1	0	1
D61.3	0	0	1	0	0	0	1	0	0	0	0	0	90	3	3	0	1	0	0	1
D64.9	1	0	1	0	0	0	0	0	0	4	3	0	1	83	0	3	0	0	0	4
D66	3	2	1	0	2	2	1	0	1	1	1	1	0	1	67	3	4	5	4	1
D68.8	2	1	1	0	3	1	5	0	2	11	0	0	0	1	10	46	7	5	3	2
D69.6	0	2	1	0	6	1	0	0	1	2	0	0	4	1	0	4	77	1	0	0
D75.1	0	2	0	0	0	0	0	0	4	0	0	0	0	0	3	0	0	91	0	0
healthy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	97	0
other	6	1	12	3	4	5	3	0	10	6	1	6	4	4	10	5	6	4	1	9

Interpretation 1



Interpretation 2

Disease	Hemoglobin	Leukocyte	Platelets	Erythrocyte	Erythrocyte Volume	Globulin	Total calcium
C90	110.1	7.0	231.0	3.59	93.3	53.2	2.5
D50	93.1	6.1	335.0	4.35	72.7	29.3	2.4
D69	135.4	8.1	40.0	4.75	85.3	29.7	2.3
Healthy	139.9	6.1	263.4	4.92	86.0	28.3	2.4

C90: low hemoglobin, high calcium and globulin

D50: low red blood cells and hemoglobin, decrease in the average volume and content of red blood cells

D69: low platelets

Conclusion

- two models
- ability to identify specific disorders
- 100% accuracy in identifying healthy clients

Thank you for attention!