Artificial Intelligence with CBC based Morphometric Parameters aimed toward effective diagnostic practices for Dysplasia associated Hematological Malignancies

Zeeshan Haider (zeeshan3335@yahoo.com)
National Institute of Blood Diseases and Bone Marrow Transplantation https://orcid.org/0000-0002-2287-3895
Ikram Uddin Ujjan
Liaquat University of Medical and Health Sciences
Najeed Ahmed Khan
NED University of Engineering & Technology
Muhammad Umer Farooq
NED University of Engineering & Technology
Waseemullah Nazir
NED University of Engineering & Technology
Eloisa Urrechaga
Hospital Galdakao-Usansolo
Muhammad Rafiq Khanani
Baqai Medical University

Research Article

Keywords: Complete Blood Count (CBC), Morphometric parameters, Cell Population Data (CPD), Machine Learning, Artificial Intelligence, Myelodysplastic syndrome (MDS), Dysplasia

Posted Date: April 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1349008/v1

Abstract

Background: The diagnosis and classification of dysplasia associated hematological neoplasms is dominated by morphology. Current round of study made use of potential ‘fingerprints’ among routinely generated diagnostic data particularly morphological and immature fraction-related (morphometric) parameters produced during routine complete blood count (CBC) testing in hemat-oncology department through artificial intelligence predictive modeling.

Methods: Along conventional statistics, neural network models were trained on anonymized demographical, clinical and diagnostic data of total 1624 individuals with common hematological neoplasms. In addition, validation conducted on independent dataset. The frameworks were trained to differentiate hematological malignancies cases with various sub-entities of Dysplasia against Non-dysplastic group as a control cohort.

Results: Our predictive model attained greater precisions; percent incorrect prediction were remained 10.3 and 4.6 for training and testing phases, respectively along with a 95.4% negative predictive value (NPV). Moreover, higher accuracy (93.1%) was obtained during prospective validation in challenge of independent dataset. Model’s performance related metrics: the gain and lift chart, predictive-pseudo probability chart, and receiver operative curve (ROC) curve were noted as persuasive. Considering the sensitivity and specificity, area under the curve (AUC) values were also noted quite convincing; 0.954 for Non-dysplasia group while 0.994 for acute myeloid leukemia with dysplasia (AML-Dys) followed by 0.992, 0.988, 0.986, 0.984, 0.973, and 0.962 for chronic myelo-monocytic leukemia (CMML), refractory anemia with excess blast-I (RAEB-I), refractory anemia with excess blast-II (RAEB-II), refractory anemia with multi-lineage dysplasia (RCMD), Hypoplastic myelodysplastic syndrome (H-MDS), and refractory anemia with uni-lineage dysplasia (RCUD) respectively.

Conclusion: The negative predictive efficiency of our framework advocates its utility as a screening tool for the rapid expulsion of Dysplasia associated hematological neoplasms in hemat-oncology sections, aiding prompt care decisions.

Introduction:

The dysplasia associated hematological neoplasms are a group of clonal hematopoietic stem cell diseases featured by ineffective blood cell’s production, dysplasia in one or more blood cell’s lineages, cytopenias and higher transformational potential for more aggressive neoplasms particularly acute myeloid leukemia (AML)[1]. The peripheral and or bone marrow dysplasia is a diagnostic hallmark of this group that even symbolizes the substructure of World Health Organization (WHO) classification of this group[2]. The dysplasia-associated group has variable prognosis, usually aggressive requiring a start of the treatment even without any delay for definitive diagnosis[3]. The prolonged turnaround time of confirmatory tests and requirements of certain special diagnostic expertise and setup are among few key limitations of the current diagnostic practice. Accordingly, the rapid point-of-care diagnosis to prioritize appropriate healthcare delivery and directing case’s flow-from is still lacking. In this direction, the large medical datasets can be one of the promising sources to prospectively study pathological cases particularly in or for hemat-oncology emergencies. Hither, the current hospital’s acceptance towards the diagnostic utility of electronic healthcare record (HER) systems will inspirit the availability of quality medical datasets. Next, EHRs can assist us for rapid-digital diagnostic data incorporation at any clinical presentation, launching of high-output digital data extraction and handling tools to conceive large-rich datasets[4]. Over and above, it is capable to interface with modern artificial intelligence (AI) / machine learning (ML) tools that may potentially directs the development of predictive (screening) models[5, 6]. The diagnostic data particularly that readily available before the speck of screening is a promising ‘solo’ source for any predictive model to achieve the diagnostic applicability at earlier course of hospital presentations[7, 8]. Readily available data or data generated in initial one hour of hospital presentation includes clinical and diagnostic information came from patient clinical history, vital signs and baseline diagnostic tests. Among baseline testing, for complete blood cells count (CBC) analysis the development of innovative technologies and analytical principles especially in flow cytometry has allowed the commercialization of new-generation hematological analyzers from last two decades. These modern analyzers are cable to generate not only routine CBC parameters but also data (information) about the morphological characteristics and immature fractions of blood cells termed as "cell population data" (CPD) or “morphometric parameters”.

These morphometric parameters have a high degree of analytical efficiency for identifying many disease specifics cellular abnormalities (disease fingerprints) that have strong potential to be introduced as screening makers for several hematological and non-hematological disorders[[4], [5], [6], [7], [8]].

In present study, towards development of rapidly deployable model for early prediction and exclusion of dysplasia associated hematological neoplasms, we trained, tested and validated ML predictive methods on rich diagnostic dataset of extended CBC analysis in our EHR, by center focus to morphometric parameters. Apart from the microscopic classification systems and its limitations, eventually aim was through enhancing the predictive ability of concern morphological items through AI/ML modeling to accept the challenge of critical needs of early prediction in these pathological emergencies, which can be incorporated in real-time diagnostic practices.

Material And Methods:

This study includes extraction of anonymized clinical and diagnostic data of patients came under consultation-to-diagnosis flow with suspicion of hematological neoplasms in tertiary care hospital (National Institute of blood disease (NIBD), Karachi, Pakistan from February 2014 to December 2020. NIBD keeps three sub-centers and serves as teaching and tertiary referral hospitals system by serving not only national but also accepts patients from surrounding international regions (Afghanistan, Iran, and UAE). Extracted data for each presentation include details about vital signs, baseline and diagnostic hematological tests. Exclusion from analysis was performed in consideration of non-responders for EHR research consent, cases with age < 18yrs, and patients with missing/not tested for hematology analysis on hospital presentation. The whole extended CBC, however importantly white blood cells (WBC) related morphometric parameters (Table 1) at the time of diagnostic workup were remained main focus of our study. We clinically evaluated morphometric parameters usually labeled as ‘CBC research’ or ‘cell population data (CPD)’ items generated by flow cytometry–based modern hematology analyzers (Sysmex XN-1000, Kobe, Japan) in peripheral blood for early pre-microscopic exclusion of dysplasia associated hematological neoplasms from other common hematological neoplasms. A total of 1624 patients: 336 with Dysplasia associated hematological neoplasms while Non-Dysplasia associated hematological malignancies group were of 1288 cases included in study. 336 Dysplasia associated hematological neoplasms group include 84 refractory cytopenia with multi-lineage dysplasia (RCMD), 78 acute myeloid leukemia with dysplasia (AML with Dys), 54 refractory anemia with excess blast-I (RAEB-I), 36 hypoplastic myelodysplastic syndrome (Hypo-MDS), 30 refractory cytopenia with uni-lineage dysplasia (RCUD), 30 chronic myelomonocytic leukemia, and 24 refractory anemia with excess blast-II (RAEB-II).

The collection of patient data and samples (blood and bone marrow) has been carried out in accordance with the Declaration of Helsinki, under the terms of all relevant local legislation. The responsible ethical committee of the National Institute of Blood Disease (NIBD) reviewed and approved the study in accordance with the ‘medical research involving human subjects act’ on permit number: NIBD/RD-167/14-2014 dated 16th December 2013. Each study subject gave informed consent.

Data was analyzed using SPSS version 23.0 (New York, NY, USA) and visualized through Clustvis (Institute of Computer Science, University of Tartu, J. Liivi-Tartu, Estonia), which is a web tool for visualizing the clustering of multivariate data (inspired by the PREDECT project and mostly based on BoxPlotR codes). The calculation of mean, standard deviation (SD), and significance (*P*-) values among study groups were also carried out through SPSS.

To delve into and obtain visualization of the subtle patterns of the CBC morphometric parameters among study groups, heat map (a supervised data visualization tool) and principal component analyses (PCA) were conducted. The heat map and PCA plots were generated through “Clustvis” https://biit.cs.ut.ee/clustvis/ (accessed on 1st December 2021), which is a web tool for visualizing the clustering of multivariate data. Aimed clustering (nodding) of study parameters, the function ‘correlation’ for the “clustering distance”, ‘average’ for the “clustering methods” and ‘tightest cluster first’ considering the “tree ordering of columns” were used. For the color grading scheme of the heat map, the command of function ‘diverging: RdBu (Red to Blue)’ at ‘minimum-2 to maximum-2' was used. The diverging: RdBu contains diverging palette options, more suitable for data with both negative and positive values, which is the same as in our data.

For AI application, artificial neural network (ANN) from ML tools was selected for predictive modeling. Major reason in selecting ANN is its superiority over other AI tools for findings patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize. To conduct ANN predictive modeling aimed at prediction / differentiation, linear model (Radial Basis Function Network [RBFN]) and non-linear assemblage (Multilayer Perceptron [MLP]) were trained to differentiate hematological neoplasms cases with Dysplasia against Non-Dysplasia. Additionally, prediction for various types among dysplasia associated hematological neoplasms was also conducted.

By principle, these are computational data modeling networks/frameworks, have three types of layers: input, hidden and output. It worked on feed-forward and supervised learning methodology. In general, an input layer/factor (provide information from outside world to the network), a hidden layer (also called the radial basis function layer in RFBN) have no direct connection with the outside world and perform computations and transfer information from input nodes to output nodes, and an output layer (dependent variables) responsible for computations and transferring information from the network to the outside world. The hidden layer transforms the input vectors into radial basis functions/ linear assemblage for RBFN and MLP, respectively. The ANN models smartly classified the cases through the input layer similarity measurement to examples from a training (data) set. The input layer is basically our features (study parameters) that we fed for training, testing and validation cessions. Each hidden layer stores a 'prototype' that is an individual example of many more present in the training set. For classification of new case, each variable computes the 'Euclidean Distance' among a new input and its prototype.

Results:

Classical CBC parameters along with morphometric items were statisticized against study groups besides being in general format of Dysplasia vs. Non-Dysplasia groups (Table 1), but also from stem to stern manner in term of Non-Dysplasia vs. subgroups for Dysplasia associated hematological disorders including AML with Dysplasia, Hypoplastic MDS, RAEB-I, RAEB-II, RCUD and RCMD (Table 4). For initial statistical analysis (Table 1 and 4), we selectively displayed the 'only found significant' parameters. At a glance, there are multiple CBC items among classical as well as research category were found significant in comparison of Non-Dysplasia vs. Dysplasia group. Furthermore, aimed comparability of Non-Dysplasia vs. six common subgroups of Dysplasia associated hematological malignances, various study parameters (from both categories) showed decidedly predictive potential. Supplementary, peripheral differential leucocytes accompany with abnormal or dysplastic cells counts were also challenged statistically for their significance in consideration of aforementioned format (Table 2 and 4). The counts for neutrophil, monocyte, basophil, dysplastic neutrophil, myelocyte, promonocyte, blast, monoblast, atypical mononuclear cell (AMNC), abnormal promyelocyte, and abnormal lymphoid cell were noted significant for Dysplasia vs. Non-Dysplasia study group's comparison. Whilst, during detailed contrasting among Non-Dysplasia vs. six common subgroups of Dysplasia associated hematological neoplasms only limited counts including lymphocyte, monocyte, dysplastic neutrophil, metamyelocyte, blast and AMNC were found significantly deviated. If we conclude, in comparative extended subgroup's analysis over just major two Dysplasia vs. Non-dysplasia grouping, the neutrophil, basophil, myelocyte, promonocyte, monoblast, abnormal promyelocyte, and abnormal lymphoid cells loose their significance and two new parameters (counts) for lymphocyte and metamyelocyte were added in a list of significantly deviated study items.

Table 1
Mean (along with standard deviation) values for selected classical and morphometric CBC items for Non-Dysplasia vs. Dysplasia associated hematological neoplasms study groups.

Study Parameters	Dysplasia Group	Non-Dysplasia Group	Sig.
	Mean ± SD	Mean ± SD	
WBC(10^9/L)	10.38 ± 15.33	68.07 ± 107.32	< 0.005
PLT(10^3/uL)	91.77 ± 95.51	193.38 ± 287.67	< 0.005
NEUT#(10^3/uL)	4.26 ± 6.71	29.05 ± 72.95	< 0.005
LYMPH#(10^3/uL)	2.4 ± 2.23	24.16 ± 59.71	< 0.005
MONO#(10^3/uL)	3.58 ± 8.54	13.1 ± 32.22	0.002
EO#(10^3/uL)	0.1 ± 0.3	0.92 ± 2.73	< 0.005
BASO#(10^3/uL)	0.05 ± 0.14	0.83 ± 2.66	< 0.005
NEUT%	39 ± 18.52	36.19 ± 31.27	< 0.005
LYMPH%	41.88 ± 21.15	41.46 ± 29.73	< 0.005
MONO%	17.54 ± 17.6	20.25 ± 21.47	0.011
IG#(10^3/uL)	0.92 ± 2.85	10.05 ± 30.8	< 0.005
IG%	4.37 ± 7.35	6.88 ± 11.25	0.002
[TNC(10^9/L)]	11.66 ± 20.74	68.44 ± 107.24	< 0.005
[WBC-N(10^9/L)]	10.39 ± 15.32	67.63 ± 106.63	< 0.005
[BA-N#(10^3/uL)]	0.05 ± 0.14	0.97 ± 2.89	< 0.005
[WBC-D(10^9/L)]	10.24 ± 15.22	68.01 ± 107.67	< 0.005
[TNC-D(10^9/L)]	11.5 ± 20.6	68.55 ± 108.38	< 0.005
[NEUT#&(10^3/uL)]	3.33 ± 4.85	18.99 ± 43.7	< 0.005
[NEUT%&]	34.64 ± 17.73	29.31 ± 25.56	< 0.005
[LYMP#&(10^3/uL)]	2.34 ± 2.22	24.08 ± 59.9	< 0.005
[LYMP%&]	41.05 ± 21.18	40.95 ± 29.7	< 0.005
[BA-D#(10^3/uL)]	0.24 ± 0.84	0.82 ± 2.16	< 0.005
[BA-D%]	1.76 ± 2.55	0.87 ± 1.64	< 0.005

WBC; white blood cell, PLT; platelet, NEUT; neutrophil, LYMPH; lymphocyte, MONO; monocyte, EO; eosinophil, BASO; basophil, IG; immature granulocyte, TNC; total nucleated cells, WBC-N; WBC from WBC and Nucleated Red Cell (WNR) channel, BA-N; basophile from WNR channel, WBC-D; WBC from differential (D) channel, TNC-D; TNC from D channel, NEUT#&; neutrophil count after deduction of immature granulocyte, LYMPH#&; lymphocyte count after deduction of high fluorescence lymphocyte count (HFLC), BA-D; basophile from D channel, NE-SSC; mean neutrophil side scatter light, LY-X; mean lymphocyte side scatter light, LY-Y; mean lymphocyte side fluorescence light, LY-Z; mean lymphocyte forward scatter light, MO-X; mean monocyte side scatter light, MO-Z; mean monocyte forward scatter light, NE-WY; distribution width neutrophil side fluorescence light, LY-WX; distribution width lymphocyte side scatter light, LY-WY; distribution width lymphocyte side fluorescence light, LY-WZ; distribution width lymphocyte forward scatter light, NRBC; nucleated red blood cell, MicroR; RBC with small than normal size, MacroR; RBC with large than normal size, PLT-I; platelet count from impedance channel, MPV; mean platelet volume, P-LCR; platelet-large cell ratio, PCT; platocrit, IPF; immature platelet fraction, Q-Flag(Blasts?); alert for the presence of blasts, Q-Flag(Abn Lympho?); alert for the presence of abnormal lymphocyte, Q-Flag(Left shift?); alert for the presence of neutrophil-precursor cells, Q-Flag(PLT Clumps?); alert for the presence of platelet clumps.

Study Parameters	Dysplasia Group	Non-Dysplasia Group	Sig.
	Mean ± SD	Mean ± SD	
[NE-SSC(ch)]	137.76 ± 13.81	147.07 ± 10.66	< 0.005
[LY-X(ch)]	83 ± 6.11	83.73 ± 8.47	0.004
[LY-Y(ch)]	68.53 ± 7.95	62.99 ± 15.55	0.011
[LY-Z(ch)]	56.63 ± 2.53	56.57 ± 3.78	0.017
[MO-X(ch)]	119.93 ± 5.94	117.1 ± 9.06	< 0.005
[MO-Z(ch)]	63.09 ± 4.69	63.3 ± 5.5	0.004
[NE-WY]	1050.11 ± 584.17	1307.74 ± 810.41	< 0.005
[LY-WX]	528.23 ± 99.67	557.27 ± 137.21	0.031
[LY-WY]	986.54 ± 208.64	1157.2 ± 548.36	0.012
[LY-WZ]	510.14 ± 113.39	585.11 ± 160.67	0.003
NRBC#(10^3/uL)	1.27 ± 8.46	0.54 ± 1.71	< 0.005
NRBC%	3.41 ± 15.47	1.14 ± 2.85	< 0.005
[MicroR(%)]	6.46 ± 5.08	10.44 ± 12.64	0.001
[MacroR(%)]	8.43 ± 6.62	5.57 ± 4.16	< 0.005
[PLT-I(10^3/uL)]	89.62 ± 93.71	191.77 ± 283.81	< 0.005
MPV(fL)	6.02 ± 5.91	7.34 ± 5.18	< 0.005
P-LCR(%)	19.15 ± 19.43	21.46 ± 16.58	< 0.005
PCT(%)	0.08 ± 0.12	0.18 ± 0.3	0.001
IPF(%)	13.21 ± 8.97	6.06 ± 5.89	0.003
Q-Flag(Blasts?)	119.09 ± 90.82	190.64 ± 126.21	0.003
Q-Flag(Abn Lympho?)	30.91 ± 39.1	77.61 ± 114.53	0.003
Q-Flag(Left Shift?)	51.07 ± 83.55	90.34 ± 116.84	< 0.005
Q-Flag(PLT Clumps?)	28.21 ± 68.15	47.75 ± 82.03	0.012

WBC; white blood cell, PLT; platelet, NEUT; neutrophil, LYMPH; lymphocyte, MONO; monocyte, EO; eosinophil, BASO; basophil, IG; immature granulocyte, TNC; total nucleated cells, WBC-N; WBC from WBC and Nucleated Red Cell (WNR) channel, BA-N; basophile from WNR channel, WBC-D; WBC from differential (D) channel, TNC-D; TNC from D channel, NEUT#&; neutrophil count after deduction of immature granulocyte, LYMPH#&; lymphocyte count after deduction of high fluorescence lymphocyte count (HFLC), BA-D; basophile from D channel, NE-SSC; mean neutrophil side scatter light, LY-X; mean lymphocyte side scatter light, LY-Y; mean lymphocyte side fluorescence light, LY-Z; mean lymphocyte forward scatter light, MO-X; mean monocyte side scatter light, MO-Z; mean monocyte forward scatter light, NE-WY; distribution width neutrophil side fluorescence light, LY-WX; distribution width lymphocyte side scatter light, LY-WY; distribution width lymphocyte side fluorescence light, LY-WZ; distribution width lymphocyte forward scatter light, NRBC; nucleated red blood cell, MicroR; RBC with small than normal size, MacroR; RBC with large than normal size, PLT-I; platelet count from impedance channel, MPV; mean platelet volume, P-LCR; platelet-large cell ratio, PCT; platocrit, IPF; immature platelet fraction, Q-Flag(Blasts?); alert for the presence of blasts, Q-Flag(Abn Lympho?); alert for the presence of abnormal lymphocyte, Q-Flag(Left shift?); alert for the presence of neutrophil-precursor cells, Q-Flag(PLT Clumps?); alert for the presence of platelet clumps.

Table 2
Peripheral film based manual (selected) differential leucocyte counts between Non-Dysplasia and Dysplasia associated hematological neoplasms study groups

Study Parameters	Dysplasia Group	Non-Dysplasia Group	Sig.
	Mean ± SD	Mean ± SD	
%Neutrophil	38.89 ± 20.49	31.36 ± 26.37	0.001
%Monocyte	5.5 ± 7.52	2.16 ± 4.55	< 0.005
%Basophil	0.34 ± 1.64	1.18 ± 4.46	0.028
%Dysplastic Neutrophil	2 ± 9.41	0.22 ± 2.47	< 0.005
%Myelocyte	2.25 ± 4.46	4.05 ± 8.79	< 0.005
%Promonocyte	0 ± 0	0.51 ± 3.5	0.030
%Blast	5.09 ± 11.73	21.3 ± 30.94	< 0.005
%Monoblast	0 ± 0	0.29 ± 2.23	0.048
%AMNC	0.84 ± 3.95	0.19 ± 1.66	< 0.005
%Abnormal Promyelocyte	0 ± 0	3.23 ± 15.36	0.001
%Abnormal Lymphoid cell	0 ± 0	7.14 ± 21.78	< 0.005
%: percent, AMNC; atypical mononuclear cell			

Table 3

Mean (along with standard deviation) values for selected classical and morphometric CBC items between Non-Dysplasia and subgroups of Dysplasia associated hematological neoplasms extended study groups.

Study Parameters	Non-Dysplasia Group	AML with Dysplasia	Hypoplastic MDS	RAEB-I	RAEB-II	RCMD	RCUD	CMML	Sig.
	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	
RBC(10^12/L)	3.34 ± 1.26	3.1 ± 0.84	3.23 ± 0.69	2.53 ± 0.95	2.75 ± 1.08	2.44 ± 1	2.45 ± 0.98	3.77 ± 0.74	0.038
MCV	86.18 ± 10.41	86.37 ± 7.11	88.5 ± 5.86	96 ± 9.15	92.5 ± 10.41	94.07 ± 10.15	101.4 ± 4.1	85.4 ± 5.18	< 0.005
MCH	27.72 ± 3.7	27.72 ± 3.38	29.17 ± 2.71	31.29 ± 3.96	31.25 ± 4.65	30.07 ± 3.87	33.6 ± 2.61	28.2 ± 2.86	< 0.005
WBC(10^9/L)	68.07 ± 107.32	11.3 ± 15.45	7.75 ± 12.88	11.21 ± 19.78	6.75 ± 9.11	3.72 ± 2.68	6.58 ± 4.5	35.03 ± 21.39	0.023
[TNC(10^9/L)]	68.44 ± 107.24	11.36 ± 15.45	7.99 ± 13.41	18.76 ± 40.66	6.82 ± 9.2	3.75 ± 2.67	6.7 ± 4.66	35.11 ± 21.52	0.027
[WBC-N(10^9/L)]	67.63 ± 106.63	11.31 ± 15.45	7.78 ± 12.9	11.21 ± 19.78	6.77 ± 9.1	3.74 ± 2.67	6.59 ± 4.5	34.98 ± 21.43	0.023
[WBC-D(10^9/L)]	68.01 ± 107.67	11.17 ± 15.31	7.66 ± 12.6	11.05 ± 19.66	6.67 ± 8.99	3.57 ± 2.49	6.45 ± 4.4	34.74 ± 21.45	0.024
[TNC-D(10^9/L)]	68.55 ± 108.38	11.22 ± 15.32	7.87 ± 13.1	18.53 ± 40.39	6.71 ± 9.08	3.58 ± 2.49	6.56 ± 4.56	34.87 ± 21.54	0.029
[BA-D%]	0.87 ± 1.64	1.61 ± 2.17	0.33 ± 0.72	1.93 ± 1.42	2.52 ± 4.13	2.09 ± 2.92	1.22 ± 2.08	2.5 ± 4.54	0.005
[NE-SSC(ch)]	147.07 ± 10.66	131.91 ± 11.42	148.17 ± 7.03	133.91 ± 15.11	134.4 ± 10.96	140.06 ± 15.01	143.28 ± 13.02	138.16 ± 18.62	< 0.005
[NE-FSC(ch)]	78.68 ± 9.43	66.93 ± 7.33	76.57 ± 6.88	72.83 ± 11.56	74.55 ± 4.93	74.37 ± 7.34	78.18 ± 9.73	70.42 ± 13.81	< 0.005
[NE-WX]	401.8 ± 114.68	462.92 ± 65.11	345.17 ± 55.78	511.78 ± 101.11	406.5 ± 207.33	402.79 ± 87.75	376.2 ± 65.49	433.2 ± 47.15	0.049
[MO-WX]	319.64 ± 81.22	296.46 ± 63.51	286 ± 149.58	375.22 ± 113.75	300.5 ± 44.79	272.57 ± 38.86	285.4 ± 35.68	257.6 ± 16.1	0.040
[MO-WY]	869.27 ± 324.3	776.85 ± 238.28	648.83 ± 428.38	887.56 ± 350.04	614.75 ± 554.31	639.57 ± 246.67	658.2 ± 156.36	762.6 ± 83.07	0.034
RDW-SD(fL)	54.93 ± 13.49	59.56 ± 9.61	54.72 ± 10.33	54.91 ± 21.2	56.4 ± 4.73	64.88 ± 23.49	74.66 ± 14.62	53.9 ± 9.94	0.011
NRBC# (10^3/uL)	0.54 ± 1.71	0.05 ± 0.09	0.22 ± 0.51	7.55 ± 20.96	0.05 ± 0.1	0.01 ± 0.01	0.11 ± 0.19	0.13 ± 0.19	< 0.005
NRBC%	1.14 ± 2.85	0.76 ± 1.33	0.78 ± 1.45	18.14 ± 36.72	0.22 ± 0.45	0.34 ± 0.7	1.08 ± 1.35	0.36 ± 0.31	< 0.005
[MacroR(%)]	5.57 ± 4.16	5.79 ± 3.8	5.47 ± 2.8	9.99 ± 8.57	6 ± 2.23	11.23 ± 8.79	14 ± 3.87	4.56 ± 0.87	< 0.005

Study Parameters	Non-Dysplasia Group	AML with Dysplasia	Hypoplastic MDS	RAEB-I	RAEB-II	RCMD	RCUD	CMML	Sig.
	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	
Q-Flag(Blasts/Abn Lympho?)	223.71 ± 94.37	198.18 ± 106.66	148.33 ± 99.88	208.33 ± 112.86	252.5 ± 76.32	125.56 ± 60.44	78 ± 39.62	182.5 ± 83.42	< 0.005
Q-Flag(Left Shift?)	90.34 ± 116.84	60.77 ± 91.6	58.33 ± 80.85	55.56 ± 101.87	17.5 ± 22.17	12.86 ± 16.84	10 ± 17.32	184 ± 97.11	0.031
Q-Flag(RBC Agglutination?)	64.36 ± 9.43	65.38 ± 8.77	68.33 ± 7.53	72.22 ± 6.67	72.5 ± 9.57	69.29 ± 9.17	76 ± 5.48	66 ± 5.48	0.004
Q-Flag(Iron Deficiency?)	81.23 ± 8.51	81.54 ± 8.01	78.33 ± 4.08	72.5 ± 7.07	77.5 ± 9.57	78.46 ± 6.89	70 ± 0	82 ± 8.37	0.007
Q-Flag(HGB Defect?)	72.89 ± 10.4	71.54 ± 6.89	76.67 ± 8.16	65 ± 11.95	72.5 ± 5	64.62 ± 9.67	64 ± 8.94	72 ± 13.04	0.017

Table 4
Peripheral film based manual (selected) differential leucocyte counts between Non-Dysplasia and subgroups of Dysplasia associated hematological neoplasms extended study groups.

Study Parameters	Non-Dysplasia Group	AML with Dysplasia	Hypoplastic MDS	RAEB-I	RAEB-II	RCMD	RCUD	CMML	Sig.
	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	
%Blast	21.3 ± 30.94	16.15 ± 20	0 ± 0	4.11 ± 6.03	5.5 ± 6.03	0 ± 0	0 ± 0	3.2 ± 4.87	0.015
%Lymphocyte	22.48 ± 23	31.08 ± 20.28	48.83 ± 26.19	40.89 ± 20.73	57 ± 29.22	54.36 ± 21.25	45.2 ± 33.12	12.4 ± 7.77	< 0.005
%Monocyte	2.16 ± 4.55	4.08 ± 3.35	4.17 ± 1.83	2.56 ± 3.13	12.75 ± 22.19	4.93 ± 5.59	5.4 ± 1.67	12 ± 10.56	< 0.005
%Metamyelocyte	1.05 ± 2.69	4.08 ± 11.15	0.33 ± 0.82	0.67 ± 1.66	1 ± 2	0.43 ± 1.09	0 ± 0	1.8 ± 2.68	0.036
%AMNC	0.19 ± 1.66	2.85 ± 8.04	0.33 ± 0.82	0.33 ± 0.71	0 ± 0	0.14 ± 0.36	0.2 ± 0.45	0.4 ± 0.89	0.001
%Dysplastic Neutrophil	0.22 ± 2.47	1 ± 3.61	0 ± 0	5.89 ± 17.67	0 ± 0	0 ± 0	0 ± 0	9.2 ± 20.57	< 0.005

Next, selected classical CBC attributes inclusive of Hb, RBC, WBC, Platelet, and morphometric parameters driven heat map through correlation based clustering of extended study groups was created, aimed at underpinning the subtle trends ‘disease signature’ (Fig. 1). The heat map illustration is not only just a color grading of study parameters (rows) assisting at a glance for hot and cold spots within dataset, but conversely clusters (rearranges) the study groups (columns) having identical patterns by nodding (branching) them. At a glance, a wide distribution of hot (higher) and cold spots (lower values) for classical and morphometric parameters were noted in comparison of Non-Dysplasia and Dysplasia associated hematological neoplasms subgroups. In the way that, hot spots against values of WBC, PLT and LY-WY for Non- Dysplasia

group, Hb and NE-WY for CMML, LY-X and LY-Y for AML with Dysplasia, LY-Z for RCMD, MO-Y and MO-Z for Hypoplastic MDS, and NE-WX and MO-WX for RAEB-I were noted. Exceptionally, only limited cold spots in conjunction with Hb, LY-WX and LY-WY for RCMD, NE-FSC for AML with Dysplasia, and MO-WZ for RAEB-II were noticed. Over and above, the nodding trends help us to find how closely patterned to each other our study parameters are. The step/level of any particular node where it groups to other node/s describes its degree of clustering (correlation). As a whole, the morphometric parameters driven heat map remained suggestive for the predictive potential of these selected CBC items for differentiation of Non-Dysplasia from Dysplasia groups and subgroups, too. Importantly, the clustering (predictive potential) limitations were also ascertained for differentiation betwixt AML with dysplasia from CMML, and RCMD from RCUD.

In Fig. 2, the PCA plot (from class of supervised machine learning tools) driven by aforementioned selected CBC based classical and morphometric parameters endorsed findings of heat map concerning the predictive potential of our study CBC attributes. Principal components (PC1 and PC2) were calculated by calling the 'singular value decomposition (SVD)' function and displayed on X and Y-axis. Which explained notable 33.3% and 24.2% of the total variance, respectively. Predictions of Dysplasia and or Non-Dysplasia are such that with probability of 0.95, a new observation from the same study group will fall inside the respective dysplasia study group and or subgroups.

Table 5
Assessment of performance 'Cross entropy error' and 'Percent in-correct prediction' of our MLP and RFB framework during training and testing by identifying extended dysplasia's study groups with 70 and 30 percent distribution for training and testing set respectively.

Metrics		MLP*	RBFN~
Training	Cross Entropy Error	16.084	30.827
	Percent In-correct Predictions	**10.3**	17.5
	Training Time	00:19.7	00:39.0
Testing	Cross Entropy Error	5.271	12.046
	Percent In-correct Predictions	**4.6**	7.3
*Stopping Rule Used: 1 consecutive step(s) with no decrease in error. Error computations are based on the testing sample.			
~The number of hidden units is determined by the testing data criterion: The "best" number of hidden units is the one that yields the smaller error in the testing data.			

The metrics in consideration of practical results of our CBC driven ML predictive frameworks presented promising elements; higher the values for area under the curve (AUC) and lower the number of cross entropy error for training and testing, as shown in Fig. 3 and Table 5. In comparison, MLP out performed RFB by generating notably superior AUC values for differentiation of our study groups. In addition, during training and testing phases the percent in-correct predictions rates along with cross entropy error values were remained noticeably lower for MLP as to RBF (Table 5). Smaller the value of square error in testing over training indicates the most-fitted number of hidden units (layers) to minimize error function. Furthermore, MLP model's performance related scales in term of predictive-pseudo probability, sensitivity and specificity (AUC), gain, and lift charts found promisingly convincing over RBF framework for the predictive ability of the MLP network (Fig. 3). Receiver operating characteristics (ROC) curve gives more powerful and much cleaner visual presentation of the specificity and sensitivity in a single plot than series of tables. The ROC chart presents all eight curves of Non-dysplasia group, AML with dysplasia, Hypoplastic MDS, RAEB-I, RAEB-II, RCMD, RCUD, and CMML significant with the area value of 0.954, 0.994, 0.973, 0.988, 0.986, 0.984, 0.962 and 0.992, respectively. In predicted-by-observed chart, the 'observed response' and 'predicted categories' were aligned with x-axis and y-axis respectively. The predicted-by-observed chart for the combined training and testing samples displayed predicted pseudo-probabilities as clustered boxplots. The prediction is considered as 'Correct' when boxplot found near the level of '0.5' for y-axis as for MLP framework, most of the boxplots are well separated

and noted near the 0.5 mark. Here, the lift chart is also displayed where the values along y-axis represent the ratio of the cumulative gain for each curve (category/ subgroups) against baseline curve. A 'baseline' curve: reference line is indicated in shape of the diagonal line, indicating greater gain if any particular line found above the baseline, and it will be lower gain provided that any individual curve observed below the reference line. In our case, the MLP predictive model present greater 'lift' (gain) values by touching '100%/10%=10.0' over RBF framework where value remained 60%/10%=6.0. Altogether, the lift at 10% for the all categories (study groups) found above 6.0 for MLP while it remained 2.0 for RFB. It showed that if we score a dataset with the MLP network and sort all the cases by predicted pseudo-probability for all study groups, we can expect top 10% to contain approximately 60% of all the cases considering their respective category (study group).

Discussion:

The peripheral and or bone marrow film examination even after the addition of advance confirmatory investigations in diagnostic workup of MDS is still considered as key opponent[9]. May be because we know phenotypic genotypic associations in hematology, inclusive of the classic hyper-granular promyelocytes of acute promyelocytic leukemia and the PML-RARA, myelokathexis and germline GATA2 mutations, and etc. The said examples are also reported for SF3B1 mutations and MDS with ring sideroblasts or else isolated del (5q) and megakaryocyte dys-morphology[10]. However, morphologic interpretation itself has limitations, including its qualitative nature (not a quantitative), differing accuracy in pathologic evaluation mostly due to intra and inter-observer reproducibility, and need of unification of cell's indices generated in complete blood cell count (CBC) testing[11]. Morphologic features shape CBC based cell's indices, but the complexity of cell's indices and morphologic alterations enact clear correlations challenging. The present research work evaluate dysplasia associated common hematological neoplasms adopting machine learning tools hatched to underpin patterns "fingerprints" among CBC morphometric parameters. Interestingly, publications on promising predictive ability of CBC morphometric items for MDS in general and it various types, too is date back < 10 years[12] [13] [14] [15]. Additionally, morphometric parameters driven-various basic classifiers are also reported [16] [17]. Exceptionally, in present study we have been abled to not only decipher the subtle dysplasia associated hematological neoplasms' signatures (fingerprints) among morphometric parameters, but also CBC morphometric data driven ML predictive modeling is suggested.

On the whole, ML application aimed assessment of MDS; the present work is not on first. For instance, to predict mutational profiles of MDS in distinction to cellular morphologic characters from pathology reports was published by Nagata and colleagues[18]. Similarly, another study reported the accessibility of dysplasia through bone marrow smears by inducting a deep neural network models[19]. Further, work from Bruck and colleagues favored the utility of ML models for prediction of MDS on bone marrow images by least vagueness[20]. The goal of such proposed ML aided practices is to go beyond the limitations caused by demand of expert eyes and morphologist to morphologist variations in conjunction with granting early prediction through avoiding delay due to shared symptoms and prolonged turnaround time for confirmatory tests. Proposed to accept the challenge of critical diagnostic needs in third-world regions and resource constraint setups where extended confirmatory (like molecular, cytogenetic, flowcytometery, and related) testing is not under flag of routine testing panel. CBC based morphometric parameters driven ML tools can offer efficient dysplasia associated hematological neoplasms screening to backing early expulsion and direct intact case flow-from.

The aim of this particular a hybrid practice (morphometric items and ML tools) is to go beyond the classic old "art" of morphology in provisional MDS predication and classification. So far, these morphometric parameters are discussed under flag of 'CBC research parameters', beside their strong potential to be introduced as automated quantitative morphological items. The approach suggested in this study presumably will edge the current diagnostic practices of MDS and other common hematological neoplasms, too. Among dysplasia associated hematological neoplasms, various morphometric parameters become deviated but neutrophils (NE) associated items are more considerable, and with the exception of NE-SFL, may be as DNA/RNA content variation of neutrophil (NE-SFL) has not been associated to any specific dysplasia associated hematological neoplasms phenotype. Here, it is more logical to use ML predictive models to identify and learned disease' patterns / signatures subtle by multidimensionality of these morphometric parameters. Actually, the routine in-practice

approach established on the gold standard of morphology (e.g., panel testing) is barley a provisional stage. The expert eyes being as morphologist on peripheral or bone marrow smear can see beyond what an in / less experienced observes. Hither, for translational linkage between the art of observation and natural human intelligence, a lot of practice, experience, and the talent make it possible. Contrarily, for experience (data sets) based learning, the machines (computers) are much faster over human. This is a reason, ML-backend tools / devices come up with different names such as Siri or Alexa, language translation programs, and driver assistance technology into our private individual environments. Assuredly, it will be a paradigms shift in routine diagnostics and clinic's follow-ups. The unending exploding diagnostic / clinical knowledge and the limited mutable innate comprehension ability of human brains can partially justify the aforesaid trend shift.

If aimed hassle free timely and correct diagnosis in the future, the AI / ML predictive tools must be introduced in our routine practices. The workflow will be like; train, test, validate, improve, and approve them[21]. Without any doubt, predictive modeling is a part of tomorrow's diagnostics, thus ideally we can't miss our today (present) to challenge and establish it, and to establish ourselves, too[22, 23]. Antonio Di leva well said hereof, "AI will not replace physicians. However, physicians who use AI will replace those who don't"[24]. Same as, "the CBC based morphometric parameters driven AI modeling can't replace the peripheral film morphologist, but promisingly peripheral film morphologist who use morphometric parameters and AI will replace who don't.

Abbreviations:

CBC Complete blood count

NPV Negative predictive value

ROC Receiver operative curve

AUC Area under the curve

AML-Dys Acute myeloid leukemia with dysplasia

CMML Chronic myelo-monocytic leukemia

RAEB-I Refractory anemia with excess blast-I

RAEB-II Refractory anemia with excess blast-II

RCMD Refractory anemia with multi-lineage dysplasia

H-MDS Hypoplastic myelodysplastic syndrome

RCUD Refractory anemia with uni-lineage dysplasia

WHO World Health Organization

EHR Electronic healthcare record

AI Artificial Intelligence

ML Machine learning

CPD Cell population data

WBC White blood cells

SD Standard deviation

PCA Principal component analyses

ANN Artificial neural network

RBFN Radial basis function network

MLP Multilayer perceptron

AMNC Atypical mononuclear cell

NPV Negative predictive values

PPV Positive predictive value

Declarations:

Ethical Approval and Consent to Participate:

The National Institute of Blood Disease Research Ethics Committee approved our study (Permit number: NIBD/RD-167/14-2014 dated 16th December 2013). All patients provided written informed consent prior to enrollment in the study.

Consent for Publication:

All authors read and approved the submitted version.

Availability of Data and Material:

The codes and data studied are accessible from National Institute of Blood Disease Research Database, condition to a request matching NIBD Research Ethics.

Competing Interests:

All authors have no competent interest that need to be disclosed.

Funding:

The authors declare that they did not receive funding.

Author's Contribution:

Conceptualization, R.Z.H.; Data curation, R.Z.H.; Methodology, R.Z.H., N.A.K. and E.U.; Resources, I.U.U. and N.A.K.; Software, N.A.K. M.U.F., W.N.; Supervision; I.U.U.; Visualization, R.Z.H., M.U.F., W.N.; Writing—original draft, R.Z.H.; Writing—review & editing, I.U.U., N.A.K., E.U., and M.R.K. All authors have read and agreed to the published version of the manuscript.

Acknowledgments:

Authors are grateful to NIBD diagnostic laboratory staff for their assistance on identification of COVID-19 cases. Also we are grateful for the research mentorship from the late Prof. Dr. Tahir Shamsi.

References:

1. Ogawa, S., Genetics of MDS. Blood, The Journal of the American Society of Hematology, 2019. **133**(10): p. 1049-1059.
2. Della Porta, M.G., et al., Minimal morphological criteria for defining bone marrow dysplasia: a basis for clinical implementation of WHO classification of myelodysplastic syndromes. Leukemia, 2015. **29**(1): p. 66-75.

3. Platzbecker, U., Treatment of MDS. Blood, The Journal of the American Society of Hematology, 2019. **133**(10): p. 1096-1107.
4. Levin, S., et al., Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. Annals of emergency medicine, 2018. **71**(5): p. 565-574. e2.
5. Menni, C., et al., Real-time tracking of self-reported symptoms to predict potential COVID-19. Nature medicine, 2020: p. 1-4.
6. Soltan, A.A., et al., Artificial intelligence driven assessment of routinely collected healthcare data is an effective screening test for COVID-19 in patients presenting to hospital. medRxiv, 2020.
7. Collins, G.S., et al., Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) The TRIPOD Statement. Circulation, 2015. **131**(2): p. 211-219.
8. Wynants, L., et al., Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. bmj, 2020. **369**.
9. Goyal, A., et al., Genotype-Resultant Morphology of Myelodysplastic Syndromes (MDS). Blood, 2018. **132**(Supplement 1): p. 1824-1824.
10. Swerdlow, S.H., et al., WHO classification of tumours of haematopoietic and lymphoid tissues. Vol. 2. 2008: International agency for research on cancer Lyon, France.
11. Zhang, L., et al., Diagnosis of myelodysplastic syndromes and related conditions: rates of discordance between local and central review in the NHLBI MDS Natural History Study. Blood, 2018. **132**: p. 4370.
12. Rocco, V., et al., Possibility of myelodysplastic syndromes screening using a complete blood automated cell count. Leukemia research, 2011. **35**(12): p. 1623-1627.
13. Kim, S.Y., et al., Discriminating myelodysplastic syndrome and other myeloid malignancies from non-clonal disorders by multiparametric analysis of automated cell data. Clinica Chimica Acta, 2018. **480**: p. 56-64.
14. Park, S.H., et al., Establishment of age-and gender-specific reference ranges for 36 routine and 57 cell population data items in a new automated blood cell analyzer, Sysmex XN-2000. Annals of laboratory medicine, 2016. **36**(3): p. 244-249.
15. Pozdnyakova, O., et al., Beyond the Routine CBC: Research CBC Parameters Associated with Myelodysplastic Syndromes and Underlying Mutations. 2019, American Society of Hematology Washington, DC.
16. Raess, P.W., et al., Automated screening for myelodysplastic syndromes through analysis of complete blood count and cell population data parameters. American journal of hematology, 2014. **89**(4): p. 369-374.
17. Ravalet, N., et al., Automated Early Detection of Myelodysplastic Syndrome within the General Population Using the Research Parameters of Beckman–Coulter DxH 800 Hematology Analyzer. Cancers, 2021. **13**(3): p. 389.
18. Nagata, Y., et al., Machine learning demonstrates that somatic mutations imprint invariant morphologic features in myelodysplastic syndromes. Blood, 2020. **136**(20): p. 2249-2262.
19. Mori, J., et al., Assessment of dysplasia in bone marrow smear with convolutional neural network. Scientific reports, 2020. **10**(1): p. 1-8.
20. Brück, O.E., et al., Machine learning of bone marrow histopathology identifies genetic and clinical determinants in patients with MDS. Blood cancer discovery, 2021. **2**(3): p. 238.
21. Topol, E.J., High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 2019. **25**(1): p. 44-56.
22. Rajkomar, A., J. Dean, and I. Kohane, Machine learning in medicine. New England Journal of Medicine, 2019. **380**(14): p. 1347-1358.
23. Rahwan, I., et al., Machine behaviour. Nature, 2019. **568**(7753): p. 477-486.
24. Di Ieva, A., AI-augmented multidisciplinary teams: hype or hope? The Lancet, 2019. **394**(10211): p. 1801.

Supplementary File:

Material and Methods

***RBFN/MLP* algorithm:** To create *these* predictive modeling, SPSS syntax-programming language was used that allowed operator for any possible modification, and in this way we tried various dataset partition of 50, 60, 70 and 50, 40, 30 for training and test, respectively for our networks. In this regard, cases were randomly assigned based on relative numbers of cases without using any portioning variable to assigned cases. In architecture of our predictive network: we call function "automatically compute range" in finding the best number of units in hidden layer. Next, normalized radial basis function and linear assemblage were selected for activation function of hidden layers. Alongside an "automatic computation" function was called for overlapping among hidden units. Methodologically, these models were educated in two steps; 1) in first step by using clustering methods 'radial basis functions (in RBFN)' and 'linear assemblage' (in MLP), the width and center of layers were calculated, 2) while during second step, the networks determined their synaptic weights. Both classification and prediction through output layer was laid by sum-of-squares error functions with identity activation function.

Model's Training-Calibration-Testing flow: Training and testing of our frameworks were completed on data collected from February 2014 to December 2020. Against Non-dysplasia associated hematological neoplasms to predict Dysplasia associated hematological malignancies CBC data on presentations was use to train models. The 'Cross entropy error' along with 'percent correct' attained during training through 10-fold cross validation were reported as model's performance metrics. Test set (30% hold-out data) was used in evaluation of ML models for their predictive efficiency. Model testing was initiated with equal numbers of Non-Dysplasia and Dysplasia associated hematological neoplasms cases. Hither, the percent correct, specificity and sensitivity against each calibrated threshold were stated. To visualize the network behavior, a predicted-by-observed chart, receiver operating characteristics (ROC) curve, and lift chart were generated.

Validation: Validation of models was conducted in direct analogy of framework prediction of Non-Dysplasia vs. Dysplasia associated hematological neoplasms by employing independent data set of patients pursue consultation-to-admission flow from Jan-2021 to July 2021. The values of cross entropy error and percent correct were stated all along validation step.

Figures

Figure 1

The heat map: color grading along with clustering inclinations of Hb, RBC, WBC, platelets and morphometric parameters among study groups. Whereas in heat map color grading for higher to lower values a 'diverging Red to Blue scheme' was used, respectively. The clustering of study groups (columns) is presented by calling a 'correlation' function.

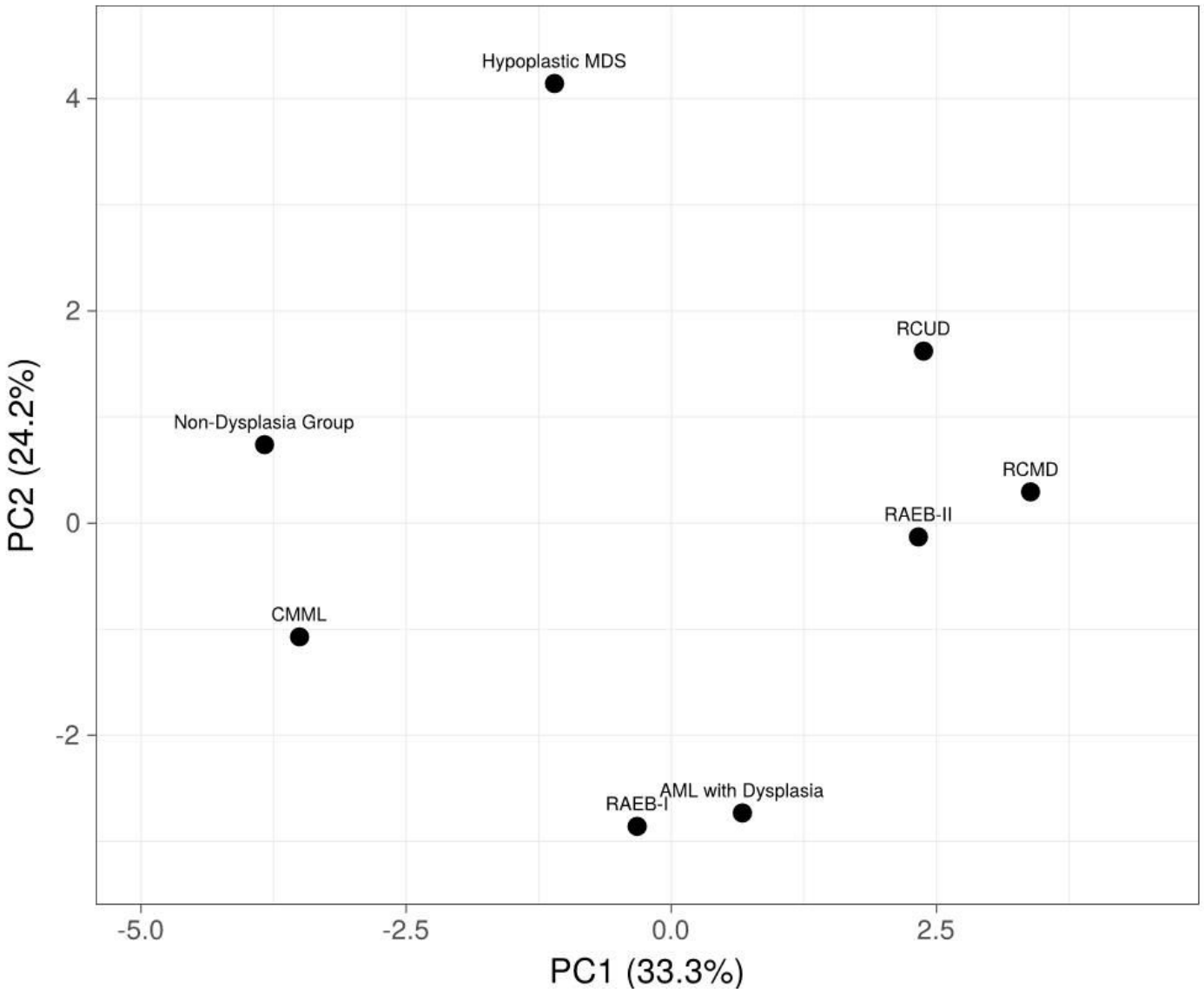

Figure 2

Visualization of latent pattern of Hb, RBC, WBC, platelets and morphometric parameters extended for study groups through Principle component analysis (PCA). By unsupervised machine learning tools (PCA) three-dimensional data reduced into two dimensions so that we can plot and understand our data in a better way. Together, both components (PC1 and PC2) covered 57.5% of the information (variance). In plot, groups are labeled with their names.

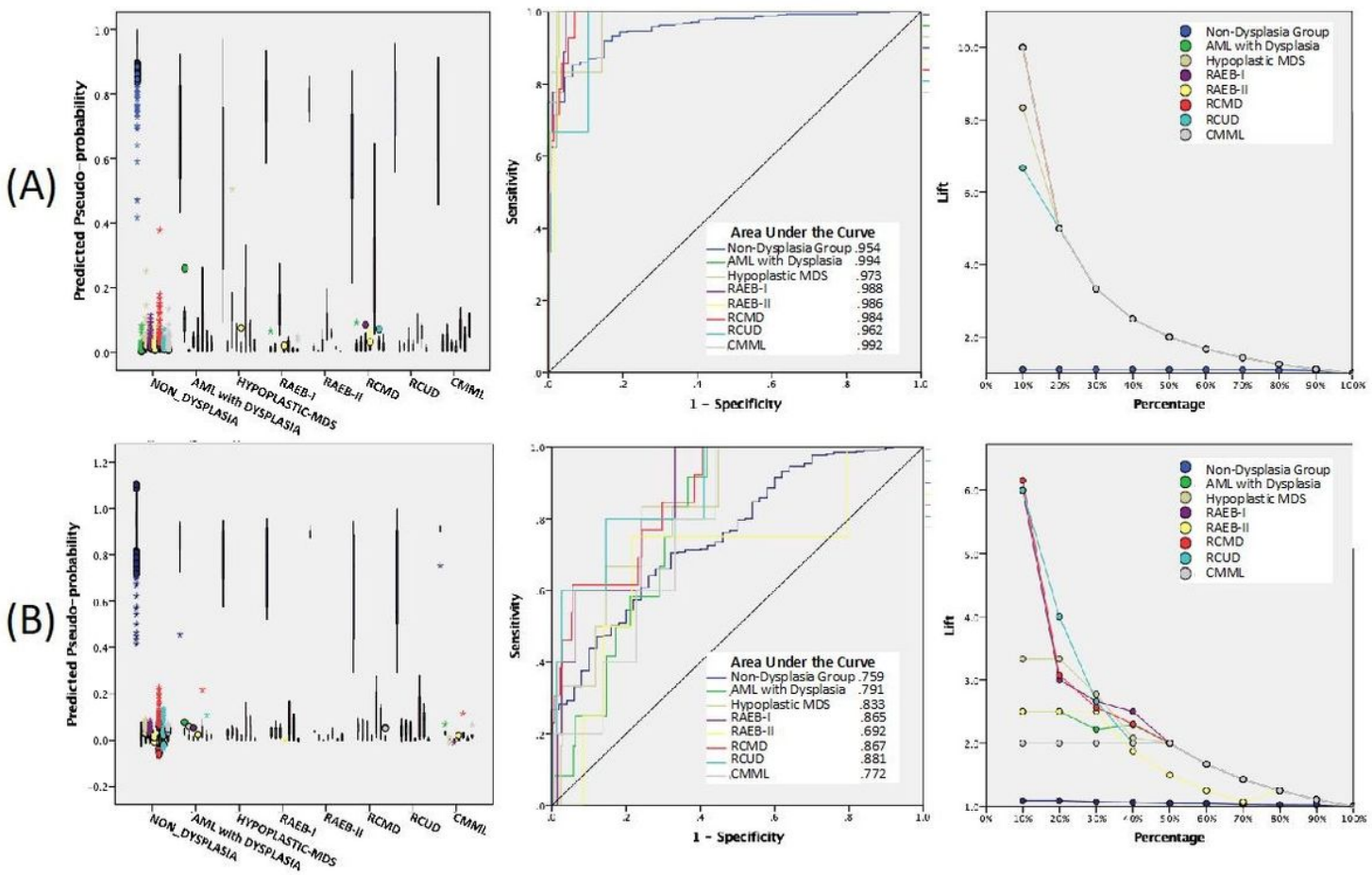

Figure 3

The predicted-by-observed chart, ROC curve and lift chart for selected CBC based classical and morphometric parameters driven (A) MLP and (B) RBF predictive model performance metrics.