Leukemia Research 109 (2021) 106639

Contents lists available at ScienceDirect

Leukemia Research

journal homepage: www.elsevier.com/locate/leukres

A machine learning approach to predicting risk of myelodysplastic syndrome

Ashwath Radhachandran [1], Anurag Garikipati, Zohora Iqbal [1,*], Anna Siefkas, Gina Barnes, Jana Hoffman, Qingqing Mao, Ritankar Das

Dascena, Inc., Houston, TX, United States

ARTICLE INFO

Keywords:
Myelodysplastic syndrome (MDS)
Early prediction
Risk assessment
Machine learning
Electronic health records (EHR)

ABSTRACT

Background: Early myelodysplastic syndrome (MDS) diagnosis can allow physicians to provide early treatment, which may delay advancement of MDS and improve quality of life. However, MDS often goes unrecognized and is difficult to distinguish from other disorders. We developed a machine learning algorithm for the prediction of MDS one year prior to clinical diagnosis of the disease.

Methods: Retrospective analysis was performed on 790,470 patients over the age of 45 seen in the United States between 2007 and 2020. A gradient boosted decision tree model (XGB) was built to predict MDS diagnosis using vital signs, lab results, and demographics from the prior two years of patient data. The XGB model was compared to logistic regression (LR) and artificial neural network (ANN) models. The models did not use blast percentage and cytogenetics information as inputs. Predictions were made one year prior to MDS diagnosis as determined by International Classification of Diseases (ICD) codes, 9th and 10th revisions. Performance was assessed with regard to area under the receiver operating characteristic curve (AUROC).

Results: On a hold-out test set, the XGB model achieved an AUROC value of 0.87 for prediction of MDS one year prior to diagnosis, with a sensitivity of 0.79 and specificity of 0.80. The XGB model was compared against LR and ANN models, which achieved an AUROC of 0.838 and 0.832, respectively.

Conclusions: Machine learning may allow for early MDS diagnosis MDS and more appropriate treatment administration.

1. Introduction

Myelodysplastic syndrome (MDS), a heterogeneous disease that occurs due to mutations in hematopoietic stem cells, manifests itself as diverse forms of cytopenia [1]. It is considered to be a "preleukemic" condition that evolves into acute myeloid leukemia (AML), which is highly aggressive and fatal, in approximately one-third of the patients [2]. MDS patients are also vulnerable to various other complications, such as increased risk of infections, bleeding, and cardiovascular disease [3,4]. Survival time for patients diagnosed with MDS ranges from less than 1 year to approximately 9 years, with mortality rates tied to factors such as age, gender, and bone marrow blast percentages [5]. MDS has an incidence rate of 20–50 cases per 100,000 persons-year for populations over 60 years of age, and mainly occurs in older male adults [6]. However, incidence statistics are much contested due to evolving diagnostic criteria and reporting and it is believed that MDS is underdiagnosed and underreported in databases [7].

Underdiagnosis of MDS may be attributable to multiple factors. Typically, clinical suspicion of MDS is raised when an older adult presents with symptoms of cytopenia, such as bleeding or recurrent infections. MDS diagnosis is confirmed by conducting a blood smear, a bone marrow biopsy, and a cytogenetic study [1]. A blood smear visually analyzes the degree of dysplasia in blood cells, while a bone marrow biopsy is used to measure blast percentages, the latter of which is critical for determining severity of the disease [8]. Cytogenetic data collected via karyotyping of bone marrow aspirate is used to identify chromosomal abnormalities and helps to subcategorize MDS patients, which is useful in determining treatment course [1,8]. However, MDS is often unrecognized by primary care physicians and is difficult to distinguish from other causes of bone marrow failures [9], cytopenias, and other clonal stem cell disorders [8,10]. These complexities often lead to delayed MDS diagnosis, at which point the condition is more advanced.

* Corresponding author at: 12333 Sowden Rd Ste B PMB 65148, Houston, TX 77080-2059, United States.
E-mail address: ziqbal@dascena.com (Z. Iqbal).
[1] Co-first author

https://doi.org/10.1016/j.leukres.2021.106639
Received 30 March 2021; Received in revised form 18 May 2021; Accepted 5 June 2021
Available online 8 June 2021

As a result, patients diagnosed with MDS suffer from poor quality of life as symptoms advance and the opportunity to benefit from earlier treatment is missed.

Treatments for MDS vary based on a patient's risk level and can include supportive care for MDS complications, therapeutics, chemotherapy, and stem cell transplants [11]. When MDS diagnosis occurs early in the disease onset, it affords clinicians the ability to monitor patients and provide early treatment, which yields better results for delaying advancement of MDS and improving quality of life in many patient populations [12–14]. For example, Cogle et al. have shown that early initiation of treatment for those with low risk MDS is more likely to result in higher successful blood transfusion treatments as well as earlier independence from these treatments, in turn reducing morbidity and mortality [15]. Thus, earlier recognition of MDS is imperative for improving patient outcomes.

Given the obstacles in diagnosing MDS, there is a clinical need for new technologies that may detect the onset of MDS. Utilization of automated hematology analyzers [16], next generation sequencing technologies [17], and novel methods in flow cytometry [18] may help detect MDS early in the disease state. However, these tools are not commonly available for standard tests. Additionally, because such tests are only ordered after clinical suspicion of MDS exists, they are unlikely to address issues related to underdiagnosis and detection of MDS. To address this need, we developed a machine learning algorithm (MLA) using two years of patient data available in the electronic health record (EHR) to predict MDS one year in advance of MDS diagnosis. These predictions are made without a reliance upon bone marrow biopsy and cytogenetic tests, and provide information that a patient is at high risk of receiving an MDS diagnosis in one year. This provides clinicians the opportunity to better serve patients who might be at risk of MDS and increase surveillance for MDS onset.

2. Methods

2.1. Data processing

Retrospective analysis was performed on 790,470 patients using EHR data drawn from over 700 healthcare sites across the United States between 2007 and 2020. We extracted information on patient demographics, vital signs, lab results, and diagnoses. Patient data was de-identified in compliance with the Health Insurance Portability and Accountability Act (HIPAA).

2.2. Cohort definition

Patients were included in the study sample if they had not previously been diagnosed with MDS or AML, were over the age of 45 at their retrospective algorithm time, and had at least one documented vital sign or laboratory measurement feature value in the 2 year window prior to their retrospective algorithm runtime. At least one vital sign or laboratory measure was required to generate algorithm predictions; patients for whom predictions could not be generated were therefore excluded. All model features used as input are presented in Table 1. Positive class patients are defined as those who were diagnosed with MDS. As in prior studies, MDS diagnoses was determined by the presence of International Classification of Diseases (ICD-9 and ICD-10) codes (Supplementary Table 1) [19–21]. Patients who were not diagnosed with MDS or AML during a two year follow-up period were included in the negative class. Patients who died during the study period were excluded. A patient inclusion flowchart is presented in Supplementary Fig. 1.

Patient characteristics were compared between the positive and negative class for both the training and test sets. Two-proportion z-test were used to compare binary data, and Mann-Whitney U tests were used to compare median lab values across classes.

Table 1
Input data used in all three models (XGBoost, Logistic regression, and artificial neural network) for prediction of MDS is presented. The average (mean) of 2-year data measurements across each feature for each patient was used as the model's inputs.

Demographics
Age
Sex
Vital Signs
Diastolic blood pressure
Heart rate
Respiratory rate
Systolic blood pressure
Temperature
Peripheral oxygen saturation (SpO_2)
Laboratory Measurements
Albumin
Blood urea nitrogen (BUN)
Calcium
Chloride
Creatinine
Glucose
Hematocrit
Hemoglobin
Absolute neutrophil count
Neutrophil percentage
Platelet count
Potassium
Red blood cell (RBC) count
Sodium
White blood cell (WBC) count

2.3. Algorithm runtime

The algorithm was designed to utilize 2 years of a patient's medical data as input and make predictions one year prior to the MDS diagnosis. For patients in the positive class, the retrospective algorithm runtime was one year prior to their MDS diagnosis, defined by the date and time of MDS ICD code, which was used as a proxy for MDS diagnosis date and time. For patients in the negative class, the retrospective algorithm runtime was defined differently. First, a two-year follow-up window after the algorithm runtime was defined. This follow-up window was then divided into two portions: a one year lookahead period (similar to the positive class) followed-by a one year outcome washout window. The purpose of the outcome washout window was to ensure that patients did not develop MDS soon after the lookahead period to avoid misclassification of patients with early stage MDS. The time windows with respect to algorithm runtime for positive and negative class patients are shown in Fig. 1.

2.4. Machine learning model development

The machine learning model, XGB, was built using gradient boosted decision trees using XGBoost [22] in Python. Gradient boosting allows the results of multiple decision trees to be iteratively combined to generate risk prediction scores. Each decision tree divides the patient population based on the values of their model features. At each branch, patients are divided into two groups based on whether their value of a given feature is above or below a certain value; for example, one branch might split patients based on whether their respiratory rate is above or below 20 breaths per minute. The branching process results in a set of "leaves," where all patients grouped within the same leaf are given the same risk score by the model. The features and values by which patients are divided during the branching process are those that minimize loss during the model training process. The XGB training process is designed to be able to handle missing data in the feature set through the identification of a default branch down which to proceed. The default branch

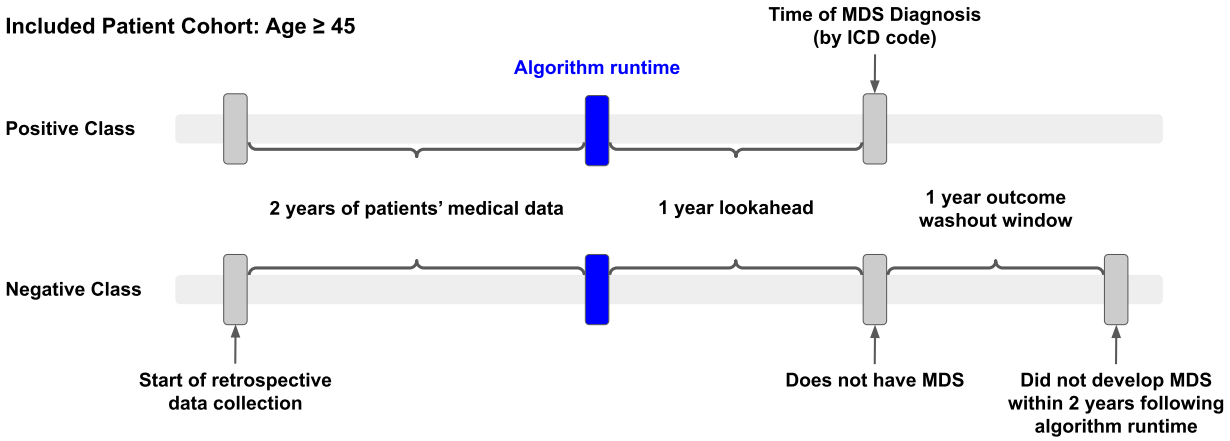

Fig. 1. Algorithm design describing patient data analysis time periods for positive and negative classes.

is determined by calculating the branch down the tree which maximizes performance for the instances of data where the feature value is present. Due to this ability to handle this missing data, no feature imputation was performed during the training or testing of this algorithm. The algorithm required at least one vital sign or laboratory measure to be present in the patient chart during the two year data collection window.

To train the model, data were randomly divided in a 70:30 split, where 70% of the data was used to train the model and 30% was retained as a hold-out test set. Model hyperparameters were tuned using 5-fold cross validation on the training portion of the data. Final hyperparameters were a maximum tree depth of 8, L1 regularization (lambda) of 300, L2 regularization (alpha) of 0, scale positive weight of 1 and 100 estimators (boosting rounds).

The XGB model was compared to two other models generated using different machine learning techniques, logistic regression (LR) and artificial neural networks (ANN). These two models were provided with the same inputs as the XGB model. All three models used the mean value across each feature's measurements over the 2-year data window for each patient as the model's inputs. A 70:30 train-test split was also applied. Before training the LR and the ANN models, two data processing steps were applied. First, feature standardization was carried out by converting the raw feature value into its respective z-score (found by subtracting the feature's mean from the patient's feature value and then dividing by the feature's standard deviation). Feature standardization and related feature scaling methods are typically implemented to speed up backpropagation in ANN training and help LR models converge faster. Second, missing values in the dataset were handled using imputation. If a patient was missing a measurement for a feature, that value was imputed using the median measurement of that feature amongst the entire training portion of the data. The logistic regression model hyperparameters were also trained using 5-fold cross validation and L2 regularization was applied to help prevent the model from overfitting to the training portion of the data. There was only one optimized hyperparameter, which was a regularization strength of 1×10^{4}. The ANN was trained for 200 epochs using the Nesterov-accelerated Adam (NAdam) optimizer [23] with a learning rate of 5×10^{-4}.

2.5. Statistical analysis

Model performance was evaluated on a hold-out test set not seen during the model training process. The model was assessed in terms of area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive and negative likelihood ratios (+/-LR), and diagnostic odds ratio (DOR). In addition to measuring predictive performance, we assessed feature importance for the XGB model using SHapley Additive exPlanations (SHAP) values to identify those features that made the biggest contribution to accurate model predictions [24, 25]. This statistical analysis assesses the value of each feature for each patient in the training dataset, and generates a plot that evaluates how differing values for the features affect the model's prediction making capability. The SHAP plot ranks the features by importance and lists the most important features first.

3. Results

This study included a total of 790,470 patients, of which 1428 were eventually diagnosed with MDS and 789,042 had no recorded MDS diagnosis. Among patients in the test set, the patients who developed MDS had a median age of 77, which was higher than the median age of 61 for patients without MDS. Other notable observations were that ~33% of the test set positive class had a previous cancer diagnosis while only ~21% of the test set negative class had a previous cancer diagnosis. Additionally, 5% of the positive class had a history of receiving chemotherapy treatment, while only 1% of the negative class had such a treatment history (Table 2). Demographic and medical characteristics were similar between the test set and the training set (Supplementary Table 2).

3.1. Model results and feature importance

Three distinct models were trained on 23 features and then evaluated on the hold-out test set. Of the models evaluated, the XGB model performed the best with an AUROC of 0.872, with a sensitivity of 0.785 and a specificity of 0.804. The LR model had an AUROC of 0.838, with sensitivity and specificity values of 0.743, and 0.786, respectively. The ANN model had an AUROC of 0.832, with a sensitivity and specificity value of 0.731 and 0.794, respectively. The AUROC plots for all three models are depicted in Fig. 2 and the model performance is summarized in Table 3. Confusion matrices and precision recall curves for all models are presented as Supplementary Table 3 and Supplementary Fig. 2, respectively.

A feature importance plot was generated for the XGB model using a SHAP analysis. The patient's age at the algorithm runtime, hematocrit, red blood cell count and platelet count were the most important features for the XGB model. Other lab measurements including white blood cell count and neutrophil percentage were also important for the XGB model (Fig. 3).

4. Discussion

Diagnosing MDS continues to be a challenge in healthcare, as is evident by underdiagnosis of the disease [26,27]. Despite this, the majority of active research in this field focuses on progression of the disease and predicting outcomes following MDS diagnosis [28–33]. Limited

Table 2
Demographics and clinical characteristics of the hold-out test set.

	Characteristic	MDS Patients (%) n = 428	Non-MDS Patients (%) n = 236,713	p-value
Age Median (IQR) MDS: 77 (68–82) non-MDS: 61 (53–71)	45–49	5 (1.17 %)	32,915 (13.91 %)	<.001
	50–54	16 (3.74 %)	35,069 (14.81 %)	<.001
	55–59	18 (4.21 %)	37,913 (16.02 %)	<.001
	60–64	42 (9.81 %)	35,532 (15.01 %)	.003
	65–69	44 (10.28 %)	29,016 (12.26 %)	.21
	70–74	48 (11.21 %)	23,444 (9.9 %)	.36
	75–79	100 (23.36 %)	16,899 (7.14 %)	<.001
	80–84	113 (26.4 %)	11,860 (5.01 %)	<.001
	85+	42 (9.81 %)	14,065 (5.94 %)	<.001
Sex	Male	239 (55.84 %)	105,089 (44.4 %)	<.001
	Female	189 (44.16 %)	131,624 (55.6 %)	<.001
Ethnicity	Hispanic	7 (1.64 %)	11,078 (4.68 %)	.003
	Not Hispanic	388 (90.65 %)	200,881 (84.86 %)	<.001
	Unknown	33 (7.71 %)	24,754 (10.46 %)	.06
Race	Caucasian	381 (89.02 %)	186,407 (78.75 %)	<.001
	African American	26 (6.07 %)	22,680 (9.58 %)	.014
	Asian	4 (0.93 %)	4919 (2.08 %)	.1
	Other/Unknown	17 (3.97 %)	22,707 (9.59 %)	<.001
Comorbidities	Arrhythmia	138 (32.24 %)	40,881 (17.27 %)	<.001
	Heart Valve Disease	68 (15.89 %)	21,880 (9.24 %)	<.001
	Prior Myocardial Infarction	46 (10.75 %)	10,792 (4.56 %)	<.001
	Chronic Heart Failure	65 (15.19 %)	12,800 (5.41 %)	<.001
	Hypertension	257 (60.05 %)	109,207 (46.13 %)	<.001
	Vascular Disease	99 (23.13 %)	26,879 (11.36 %)	<.001
	Cerebrovascular Disease	77 (17.99 %)	17,988 (7.6 %)	<.001
	Hepatic Cirrhosis	7 (1.64 %)	1835 (0.78 %)	.04
	Chronic Kidney Disease	90 (21.03 %)	17,168 (7.25 %)	<.001
	Renal Failure	55 (12.85 %)	10,714 (4.53 %)	<.001
	COPD	82 (19.16 %)	30,141 (12.73 %)	<.001
	Cancer	143 (33.41 %)	49,600 (20.95 %)	<.001
	Diabetes	104 (24.3 %)	42,204 (17.83 %)	<.001
	Obesity	67 (15.65 %)	43,897 (18.54 %)	.12
Clinical Measures	Hematocrit (%)	35.50	41.05	<.001
	Red blood cell count (x10^6/μL)	3.76	4.54	<.001
	Platelet count (x10^3/μL)	194.61	229.50	<.001
	Peripheral oxygen saturation (SpO_2) (%)	96.65	97.04	<.001
	Diastolic Blood Pressure (mm Hg)	69.58	76.00	<.001
	Temperature (°C)	36.64	36.68	.02
	White blood cell count (x10^3/μL)	6.35	7.10	<.001
	Systolic Blood Pressure (mm Hg)	128.56	128.06	.02
	Albumin (g/dL)	3.9	4.1	<.001
	Neutrophil percentage (%)	62.66	62.40	<.001
	Heart Rate (bpm)	73.73	75.33	.63
	Respiratory Rate (breaths/min)	17.44	17.00	.002
	Blood urea nitrogen (mg/dL)	19.78	15.60	.004
	Absolute neutrophil count (x10^3/μL)	4.06	4.40	<.001
	Glucose (mg/dL)	107.34	102.00	<.001
	Chloride (mEq/L)	103.95	103.00	<.001
Procedures	Radiation Therapy	11 (2.57 %)	2043 (0.86 %)	<.001
	Chemotherapy	21 (4.91 %)	2508 (1.06 %)	<.001

research related to MDS diagnosis revolves around confirmation and subcategorization of MDS using novel tools and technologies, such as next generation sequencing to look for genetic mutations [33–36] and automated hematology [16], that are not readily available, and would require MDS diagnosis to be already under consideration. However, diagnostic barriers to MDS are far more dependent upon lack of recognition by clinicians [26,37] and difficulties in distinguishing MDS from other disorders [9]. To our knowledge, little research has been done to develop clinical decision support (CDS) tools to improve early MDS diagnosis.

In this longitudinal retrospective study, we analyzed data for patients over the age of 45, from over 700 healthcare sites across the United States between the years 2007 and 2020. The aim of this study was to investigate the use of machine learning for early prediction of MDS diagnosis using only EHR data. To this end, we examined three separate models: XGB, LR, and ANN. All three models yielded AUROC values above 0.83 (Table 3), demonstrating high accuracy in MDS prediction one year prior to patient MDS diagnosis. Best performance was achieved by the XGB model, with an AUROC of 0.872, and sensitivity and specificity of 0.785, and 0.804, respectively. An assessment of feature importance for the XGB model revealed that the model utilized several known risk factors for MDS when generating MDS risk predictions. For example, the algorithm identified increased age as predictive of MDS, in agreement with epidemiologic research showing that MDS prevalence increases with age [5,6]. Similarly, the model found sex to be predictive of MDS, again in accord with epidemiologic research finding differential prevalence of MDS across sex [5]. Additionally, several important features were related to hematopoietic cell lineages (e.g. platelet count, red blood cell count, hemoglobin, hematocrit, and white blood cell count). As MDS is characterized by abnormalities in hematopoietic cells, the importance of these features may indicate that the algorithm is identifying trends within and values these features to predict MDS. Finally, it

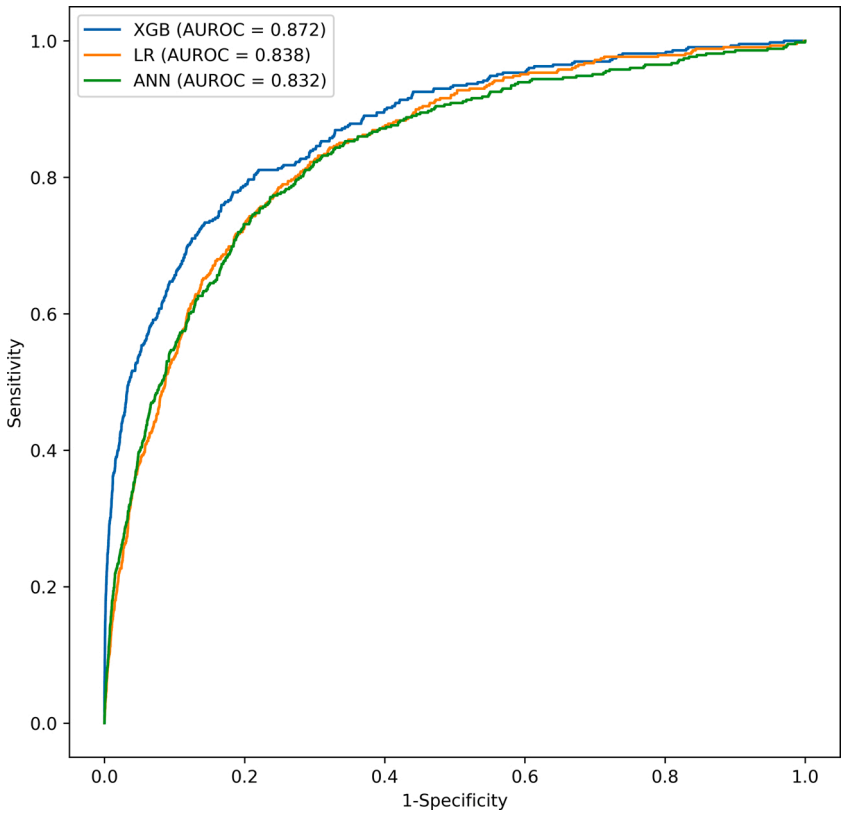

Fig. 2. The area under the receiver operating characteristic curve (AUROC) plots for XGB, LR and ANN models.

Table 3
Summary of performance for XGB, LR and ANN models on hold-out test set. Area under the receiver operating characteristic curve (AUROC) value, sensitivity, specificity, positive and negative likelihood ratios (+/-LR), and diagnostic odds ratio (DOR) demonstrate better performance from XGB model compared to LR and ANN models.

	XGB	LR	ANN
AUROC (95 % CI)	0.872 (0.853–0.89)	0.838 (0.819–0.855)	0.832 (0.811–0.851)
Sensitivity (95 % CI)	0.785 (0.743–0.822)	0.743 (0.702–0.784)	0.731 (0.69–0.772)
Specificity (95 % CI)	0.804 (0.802–0.805)	0.786 (0.785–0.788)	0.794 (0.793–0.796)
LR+	6.456	3.479	3.558
LR-	0.377	0.327	0.338
DOR	17.104	10.645	10.522

has been noted that MDS patients are at risk of infection due to neutropenia [3,38]. Neutrophil count and percentage may be important model features due to the correlation between MDS and neutropenia.

Several factors may have contributed to the XGB model outperforming LR and ANN models [39]. Firstly, compared to XGB, LR is a structurally simpler model, where decisions are based on a set of coefficients that determines a complex boundary to separate the dataset by class. Second, neural networks are typically geared towards classification tasks on unstructured data, such as images or texts. In these cases, neural networks may outperform XGB, since the former evaluates how different features may be related (i.e. relationship between pixels in an image, words in a sentence, etc) and the latter evaluates features independent of one another. Since this study is based on structured data extracted from the EHR, which in this case is not a continuous set of data, XGB may perform better due to its ability to examine features independently as opposed to looking at the feature set as a whole. Third, XGB resembles a rule-based decision process since it follows a deterministic method that attempts to maximize information gain. In a clinical setting, a hematopathologist follows a decision tree-like process when diagnosing MDS, i.e. values of WBC, platelet, etc. above or below a certain count. Since the XGB model makes decisions in a similar pattern, it may contribute to improved performance of XGB over the ANN model. Finally, XGB is able to handle missing or null values, while ANN and LR require imputation techniques to be implemented. Depending on how much data is missing, imputation of missing values could decrease ANN or LR performance, or reduce precision.

MDS diagnosis requires evidence of cytopenias, morphologic or immunophenotypic evidence of dysplasia, and cytogenetic abnormalities [1,8,40]. According to current standard of care, these characteristics are established by conducting 1) physical examination, 2) complete blood count (CBC), 3) blood smear to visualize dysplasia, 4) bone marrow biopsy to determine blast percentages, and 5) karyotyping or molecular cytogenetic tests to look for chromosomal abnormalities, such as fluorescent in situ hybridization (FISH), or comparative genomic hybridization (CGH) [1]. Although some of these tests are performed routinely, such as CBC and physical examinations, a bone marrow biopsy for blast percentage measurements and cytogenetic examinations may not be conducted until MDS diagnosis is suspected. Therefore, we designed an MLA to predict MDS accurately without using blast percentages, cytogenetic tests or specialized lab results, to ensure that it is capable of providing early MDS warning in patients not yet clinically suspected of having the condition. AUROC values > 0.83 across all our models demonstrated that without this data, we are able to predict MDS prior to diagnosis, using only information from the patient EHR. Our feature importance analysis, as shown by the SHAP plot in Fig. 3, determined the patient's age, hematocrit, red blood cell and platelet count as the most important features for our XGB model.

Machine learning has emerged as a supplementary tool to aid clinical care in many areas of medicine. Radakovich et al. details the various uses of machine learning for cancers of the blood and hematological systems, where such tools typically assess risk, disease progression, and

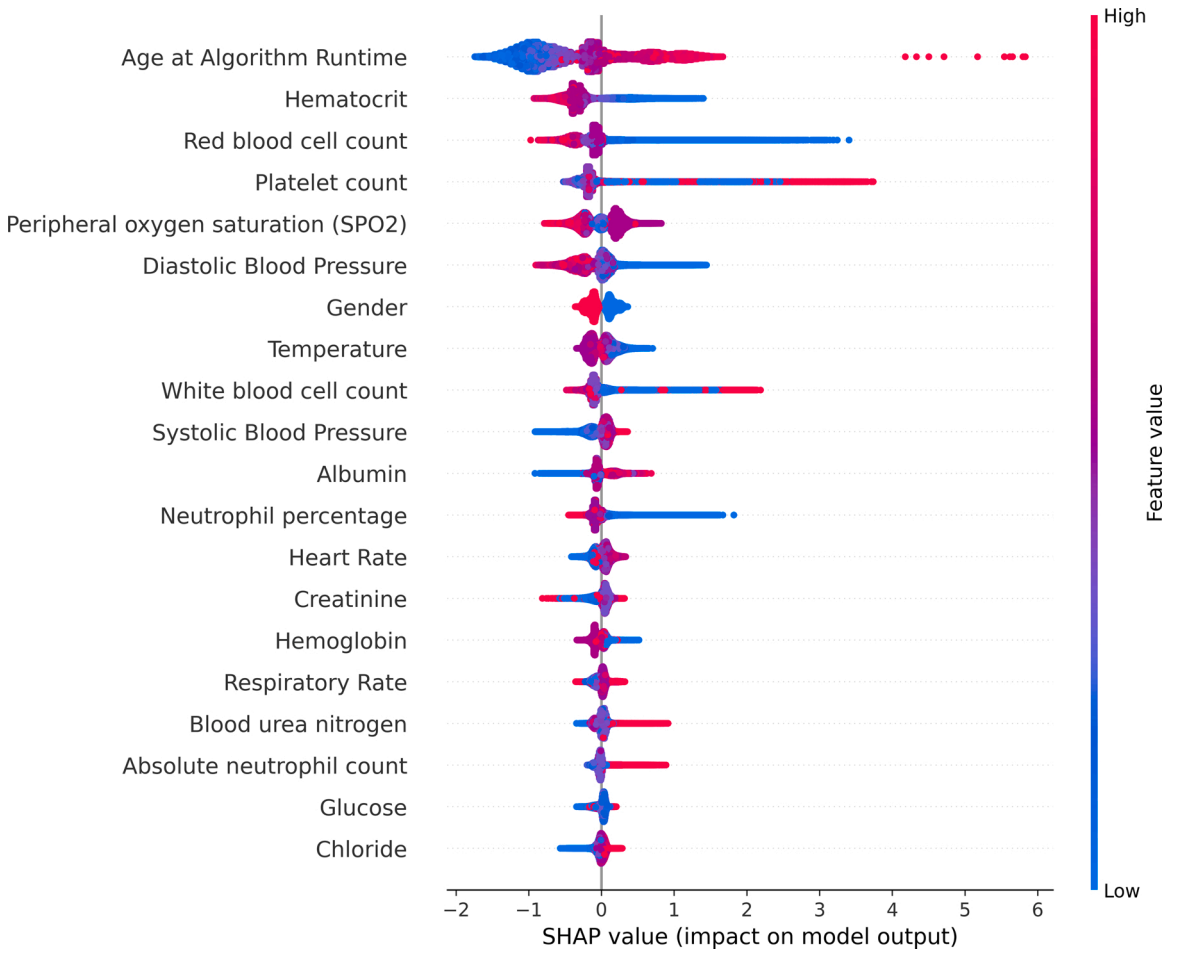

Fig. 3. Feature correlations and distribution of feature importance for the XGB model. Model input variables are ranked in descending order of feature importance. Red is indicative of a high feature value and blue is indicative of a low feature value. Points to the right of the line of neutral contribution resulted in a higher score; points to the left of this line resulted in a lower score.

response to therapeutics and treatment [41]. However, to our knowledge, no approach, including machine learning or other predictive methods, exist to ascertain an individual's likelihood of developing MDS using only EHR data.

Though machine learning has yet to be applied as an early diagnostic tool for MDS using EHR data, some studies have examined the ability of these methodologies to predict cancer types that are similar to MDS, such as AML [42]. Warnat-Herresthal et al. tested the ability of nine machine learning models to discriminate between AML and non-AML patients, as well as to predict various types of leukemic cancers [43]. Though accuracy was high, the study required data derived from DNA sequencing. Ohno-Machado et al. notes that this type of data is not consistently available within the EHR and is not typically included in a manner that makes the data readily interpretable [44]. Our present study builds upon this literature and demonstrates the potential of machine learning for early hematological cancer detection.

4.1. Limitations

The retrospective nature of the study makes it difficult to determine if this MLA would ultimately impact patient outcomes. Although early identification can help many MDS patients, there are a high number of patients who do not qualify for any curative treatment, such as bone marrow transplantation, due to advanced age or other comorbidities [45]. Another limitation of this study is that we maintained a 2-year observation period from the time of algorithm MDS diagnosis prediction to ensure that patients did not develop MDS during this time. However, that does not imply that they would not develop MDS past this 2-year period. Lastly, the data from which these scores were derived was comprised heavily of non-Hispanic whites. This lack of diverse representation in the dataset may mean that this MLA cannot be accurately used in broad populations.

5. Conclusions

While there is significant ongoing active research on MDS, there is a lack of MDS CDS tools being investigated or developed for early and accurate prediction of MDS. Machine learning can make a meaningful impact in this area. We have developed methods to accurately predict MDS prior to clinical diagnosis. Our algorithm relies on a limited number of inputs readily available in EHR data, without utilizing blast percentages or cytogenetics. In clinical practice, use of such a tool may enable earlier diagnosis of MDS and supportive treatment administration.

Declaration of Competing Interest

All authors who have affiliations listed with Dascena (Houston, Texas, U.S.A) are employees or contractors of Dascena.

Acknowledgements

We would like to thank Dr. Matt Schwede for providing valuable medical perspective into management of myelodysplastic syndromes.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.leukres.2021.106639.

References

[1] S. Swerdlow, et al., WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues, IARC, 2017.
[2] Myelodysplastic Syndromes - MDS - Statistics, Cancer.Net, 2012. https://www.cancer.net/cancer-types/myelodysplastic-syndromes-mds/statistics.
[3] G. Leone, L. Pagano, Infections in myelodysplastic syndrome in relation to stage and therapy, Mediterr. J. Hematol. Infect. Dis. 10 (2018).
[4] A.M. Brunner, et al., Risk and timing of cardiovascular death among patients with myelodysplastic syndromes, Blood Adv. 1 (2017) 2032–2040.
[5] P.L. Greenberg, et al., Revised international prognostic scoring system for myelodysplastic syndromes, Blood 120 (2012) 2454–2465.
[6] M. Cazzola, L. Malcovati, Myelodysplastic syndromes — coping with ineffective hematopoiesis, N. Engl. J. Med. 352 (2005) 536–538.
[7] A.M. Zeidan, R.M. Shallis, R. Wang, A. Davidoff, X. Ma, Epidemiology of myelodysplastic syndromes: why characterizing the beast is a prerequisite to taming it, Blood Rev. 34 (2019) 1–15.
[8] G.A. Hamid, A.W. Al-Nehmi, S. Shukry, Diagnosis and classification of myelodysplastic syndrome. Recent Dev. Myelodysplastic Syndr., 2019, https://doi.org/10.5772/intechopen.82532.
[9] A.E. DeZern, M.A. Sekeres, The challenging world of cytopenias: distinguishing myelodysplastic syndromes from other disorders of marrow failure, Oncologist 19 (2014) 735–745.
[10] D.P. Steensma, Does early diagnosis and treatment of myelodysplastic syndromes make a difference? Best Pract. Res. Clin. Haematol. 32 (2019), 101099.
[11] Myelodysplastic Syndromes Treatment (PDQ®)–Health Professional Version -, National Cancer Institute, 2020. https://www.cancer.gov/types/myeloproliferative/hp/myelodysplastic-treatment-pdq.
[12] A.M. Zeidan, et al., Deferasirox therapy is associated with reduced mortality risk in a medicare population with myelodysplastic syndromes, J. Comp. Eff. Res. 4 (2015) 327–340.
[13] M. Delforge, et al., Adequate iron chelation therapy for at least six months improves survival in transfusion-dependent patients with lower risk myelodysplastic syndromes, Leuk. Res. 38 (2014) 557–563.
[14] V. Runde, et al., Bone marrow transplantation from HLA-identical siblings as first-line treatment in patients with myelodysplastic syndromes: early transplantation is associated with improved outcome, Bone Marrow Transplant. 21 (1998) 255–261.
[15] C.R. Cogle, et al., Early treatment initiation in lower-risk myelodysplastic syndromes produces an earlier and higher rate of transfusion independence, Leuk. Res. 60 (2017) 123–128.
[16] Y.L. Carattini, et al., Early detection of myelodysplastic syndromes: maximizing the utility of automated hematology, Blood 128 (2016) 5527.
[17] B.B. Ganguly, N.N. Kadam, Mutations of myelodysplastic syndromes (MDS): an update, Mutat. Res. Rev. Mutat. Res. 769 (2016) 47–62.
[18] C. Duetz, et al., Machine learning-based flow cytometry diagnostics in myelodysplastic syndromes: validation in the HOVON89 clinical trial (EudraCT 2008-002195-10), Blood 136 (2020) 10–12.
[19] A.M. Zeidan, et al., Lenalidomide performance in the real world, Cancer 119 (2013) 3870–3878.
[20] C.R. Cogle, B.M. Craig, D.E. Rollison, A.F. List, Incidence of the myelodysplastic syndromes using a novel claims-based algorithm: high number of uncaptured cases by cancer registries, Blood 117 (2011) 7121–7125.
[21] S.L. Goldberg, et al., Economic impact on US Medicare of a new diagnosis of myelodysplastic syndromes and the incremental costs associated with blood transfusion need, Transfusion (Paris) 52 (2012) 2131–2138.
[22] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (Association for Computing Machinery) (2016), https://doi.org/10.1145/2939672.2939785.
[23] T. Dozat, Incorporating Nesterov Momentum Into Adam, 2016, p. 4.
[24] S. Lundberg, S.-I.A. Lee, Unified Approach to Interpreting Model Predictions, ArXiv170507874 Cs Stat, 2017.
[25] R. Rodríguez-Pérez, J. Bajorath, Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values, J. Med. Chem. 63 (2020) 8761–8777.
[26] A.M. Khan, Why are myelodysplastic syndromes unrecognized and underdiagnosed? A primary care perspective, Am. J. Med. 125 (2012) S15–S17.
[27] C.R. Cogle, Incidence and burden of the myelodysplastic syndromes, Curr. Hematol. Malig. Rep. 10 (2015) 272–281.
[28] H. Tamaki, et al., The Wilms' tumor gene WT1 is a good marker for diagnosis of disease progression of myelodysplastic syndromes, Leukemia 13 (1999) 393–399.
[29] I. Triantafyllidis, A. Ciobanu, O. Stanca, A.R. Lupu, Prognostic factors in myelodysplastic syndromes, Mdica 7 (2012) 295–302.
[30] J. Shreve, A. Nazha, The evolving landscape of myelodysplastic syndrome prognostication, Clin. Hematol. Int. 2 (2020) 43–48.
[31] U. Oelschlaegel, et al., Machine learning approach identifies independent prognostic value of flow cytometry (FCM) in myelodysplastic syndromes (MDS), Blood 134 (2019) 4251.
[32] M.G. Della Porta, et al., Validation of WHO classification-based Prognostic Scoring System (WPSS) for myelodysplastic syndromes and comparison with the revised International Prognostic Scoring System (IPSS-R). A study of the International Working Group for Prognosis in Myelodysplasia (IWG-PM), Leukemia 29 (2015) 1502–1513.
[33] L.M. Drusbosky, C.R. Cogle, Computational modeling and treatment identification in the myelodysplastic syndromes, Curr. Hematol. Malig. Rep. 12 (2017) 478–483.
[34] M. Meggendorfer, W. Walter, C. Haferlach, W. Kern, T. Haferlach, Challenging blast counts by machine learning techniques and genome sequencing for discriminating AML and MDS, Blood 134 (2019) 4663.
[35] E.J. Duncavage, B. Tandon, The utility of next-generation sequencing in diagnosis and monitoring of acute myeloid leukemia and myelodysplastic syndromes, Int. J. Lab. Hematol. 37 (2015) 115–121.
[36] A. Tefferi, et al., Targeted next-generation sequencing in myelodysplastic syndromes and prognostic interaction between mutations and IPSS-R, Am. J. Hematol. 92 (2017) 1311–1317.
[37] T.A. Glauser, et al., Current pathology practices in and barriers to MDS diagnosis, Leuk. Res. 37 (2013) 1656–1661.
[38] C. Pomeroy, M.M. Oken, R.E. Rydell, G.A. Filice, Infection in the myelodysplastic syndromes, Am. J. Med. 90 (1991) 338–344.
[39] R. Strandberg, J. Låås, A Comparison Between Neural Networks, Lasso Regularized Logistic Regression, and Gradient Boosted Trees in Modeling Binary Sales, 2021, p. 56.
[40] U. Germing, G. Kobbe, R. Haas, N. Gattermann, Myelodysplastic syndromes: diagnosis, prognosis, and treatment, Dtsch. Ärztebl. Int. 110 (2013) 783–790.
[41] N. Radakovich, M. Nagy, A. Nazha, Machine learning in haematological malignancies, Lancet Haematol. 7 (2020) e541–e550.
[42] J.-N. Eckardt, M. Bornhäuser, K. Wendt, J.M. Middeke, Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects, Blood Adv. 4 (2020) 6077–6085.
[43] S. Warnat-Herresthal, et al., Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics, iScience 23 (2020) 100780.
[44] Genomics and electronic health record systems, Human Molecular Genetics, Academic, Oxford, 2021. https://academic.oup.com/hmg/article/27/R1/R48/4975618.
[45] L. Malcovati, et al., Diagnosis and treatment of primary myelodysplastic syndromes in adults: recommendations from the European LeukemiaNet, Blood 122 (2013) 2943–2964.