Early Prediction of Sepsis from Clinical Data

Executive summary

החיזוי המוקדם של אלח דם מנתונים קליניים הוא מחקר שחוקר את השימוש באלגוריתמים של למידת מכונה לפיתוח מודל חיזוי לגילוי מוקדם של אלח דם. אלח דם הוא מצב חמור שעלול להוביל לאי ספיקת איברים ומוות, והתערבות מוקדמת היא קריטית לשיפור תוצאות המטופל.

המחקר השתמש במערך נתונים גדול של רישומי בריאות אלקטרוניים מ3 יחידות טיפול נמרץ מ-3 בתי חולים שונים בארה"ב. מערך הנתונים כולל המון פרמטרים ונתונים על כל חולה וחולה לאורך שהותו בטיפול הנמרץ בין אם אובחן כחולה באלח דם ובין אם לא אובחן לבסוף.

המשימה אותה עלינו היה לבצע היא להצליח ולזהות חולים באלח דם המגיעים לטיפול נמרץ בבית החולים וזאת לפחות תוך 6 שעות מרגע הגעתו לטיפול הנמרץ ותחילת תיעוד מצבו.

ראשית, ביצענו מחקר עומק על הנתונים שלו, תוך הבנת הפיצ'רים בנתונים, הבנת סטטיסטית של הנתונים הכוללת ניסיון לזיהוי התפלגות הנתונים וניסיון הבנת הקשר בין פרטי המידע השונים. לאחר ביצוע הצעדים המקדימים הללו, גילינו מה הם פרטי המידע הרלוונטים ביותר ממסד הנתונים בהם נרצה לעשות שימוש בשלב הבא, שלב החיזוי. בשלב זה, ניסינו לבחור בין מספר אלגוריתמים שונים של למידת מכונה מי מביא את התוצאות הטובות ביותר, אלגוריתמים הכוללים deep neural network, random forest ו-gradient boosting וזאת כדי לפתח את מודל הניבוי המיטבי. על מנת להבחין מי מהמודלים הוא המיטבי, הערכנו את ביצועי המודל באמצעות מדד f1.

לאחר סיכום מתומצת זה, נראה מכאן והלאה את כל אחד מהצעדים אותם ביצענו בצורה מפורטת ומנומקת.

Exploratory Data Analysis

Describing the features that are available in the dataset:

Variable	Description
HR	Heart rate (beats per minute)
O2Sat	Pulse oximetry (%)
Temp	Temperature (Deg C)
SBP	Systolic BP (mm Hg)
MAP	Mean arterial pressure (mm Hg)
DBP	Diastolic BP (mm Hg)

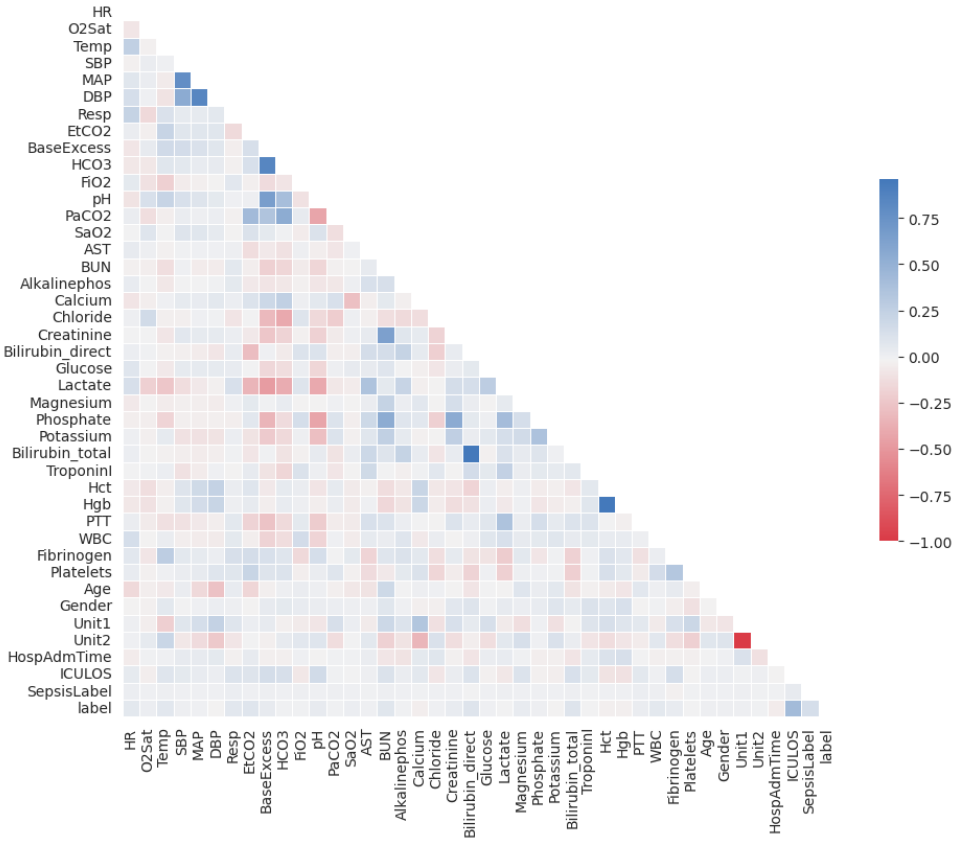
Resp	Respiration rate (breaths per minute)
EtCO2	End tidal carbon dioxide (mm Hg)
BaseExcess	Measure of excess bicarbonate (mmol/L)
HCO3	Bicarbonate (mmol/L)
FiO2	Fraction of inspired oxygen (%)
pH	N/A
PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
SaO2	Oxygen saturation from arterial blood (%)
AST	Aspartate transaminase (IU/L)
BUN	Blood urea nitrogen (mg/dL)
Alkalinephos	Alkaline phosphatase (IU/L)
Calcium	(mg/dL)
Chloride	(mmol/L)
Creatinine	(mg/dL)
Bilirubin_direct	Bilirubin direct (mg/dL)
Glucose	Serum glucose (mg/dL)
Lactate	Lactic acid (mg/dL)
Magnesium	(mmol/dL)
Phosphate	(mg/dL)
Potassium	(mmol/L)
Bilirubin_total	Total bilirubin (mg/dL)
TroponinI	Troponin I (ng/mL)
Hct	Hematocrit (%)
Hgb	Hemoglobin (g/dL)
PTT	Partial thromboplastin time (seconds)
WBC	Leukocyte count (count*10^3/µL)
Fibrinogen	(mg/dL)
Platelets	(count*10^3/µL)
Age	Years (100 for patients 90 or above)
Gender	Female (0) or Male (1)
Unit1	Administrative identifier for ICU unit (MICU)
Unit2	Administrative identifier for ICU unit (SICU)
HospAdmTime	Hours between hospital admit and ICU admit
ICULOS	ICU length-of-stay (hours since ICU admit)
SepsisLabel	For sepsis patients, SepsisLabel is 1 if $t \geq t_{sepsis} - 6$ and 0 if $t < t_{sepsis} - 6$. For non-sepsis patients, SepsisLabel is 0.

Inspecting the features distribution

על סמך ההיסטוגרמות שנמצאות מעלה, ניתן לראות כי יש מספר פיצ'רים להם יש צורת התפלגות שנראות ממרוכזת סביב התוחלת מה שיכול להעיד על דימיון להתפלגויות נורמליות (SBP, HR, MAP, TEMP, DBP, RESP, EtCO2, BaseExcess, HcO3, HCT, Hgb, PaCO2,Chloride, Magnesium, Age, PH, Phosphate). בנוסף, ניתן לראות שינם פיצ'רים עם זנב כבד מה שיכול להעיד על התפלגות זנב כבד (O2SAT, Bilirubin_total, troponinal, PTT, WBC, Platelets, SaO2, AST, BUN, ALKALINEPHOS, Creatinine, Bilirubin_direct, Lactate, ICULOS, HospAdmTime). כמובן שמדובר בקירובים ואין התפלגות מוחלטות של הפיצ'רים, אך באמצעות זיהוי המגמה הזאת, נוכל להסיק הרבה דברים בהקשר לפיצ'רים שלנו בצעדים הבאים.

בנוסף, ניתן לראות כי יש פער גדול בין כמות הנתונים לגבי אנשים חולים לבין אנשים בריאים במערך הנתונים שברשותינו מה שעשוי לגרום לdata imbalance. דבר זה יכול לגרום לבעיה בדיוק המודלים, דבר שנצטרך לתת לו מענה והתחשבות בשלבים הבאים של המחקר.

:comparative analysis between features



על פי טבלת ה-heatmap עבור הקורלציות בין הפיצ'רים השונים ב-data, ניתן לראות כמה קשרים מעניינים. ראשית, אנו הוספנו לטבלה המקורית עמודת label שמטרתה לסווג את המטופל באופן סופי כחולה או לא חולה במחלת אלח דם. לאחר הסתכלות על שורת הקורלציות עם עמודת label, ניתן לראות כי באופן כללי רוב הקורלציות מעט חלשות, אם כי ניתן לראות קורלציה חיובית חזקה עם עמודת iculos (זמן שהייה בטיפול נמרץ, ערך הקורלציה = 0.4206) בנוסף, ניתן לראות עוד כמה קורלציות חיוביות עם עמודת label אך לא מספיק חזקות ולכן לא נוכל להסיק קשר משמעותי ביניהם שיעזור לנו בהמשך.

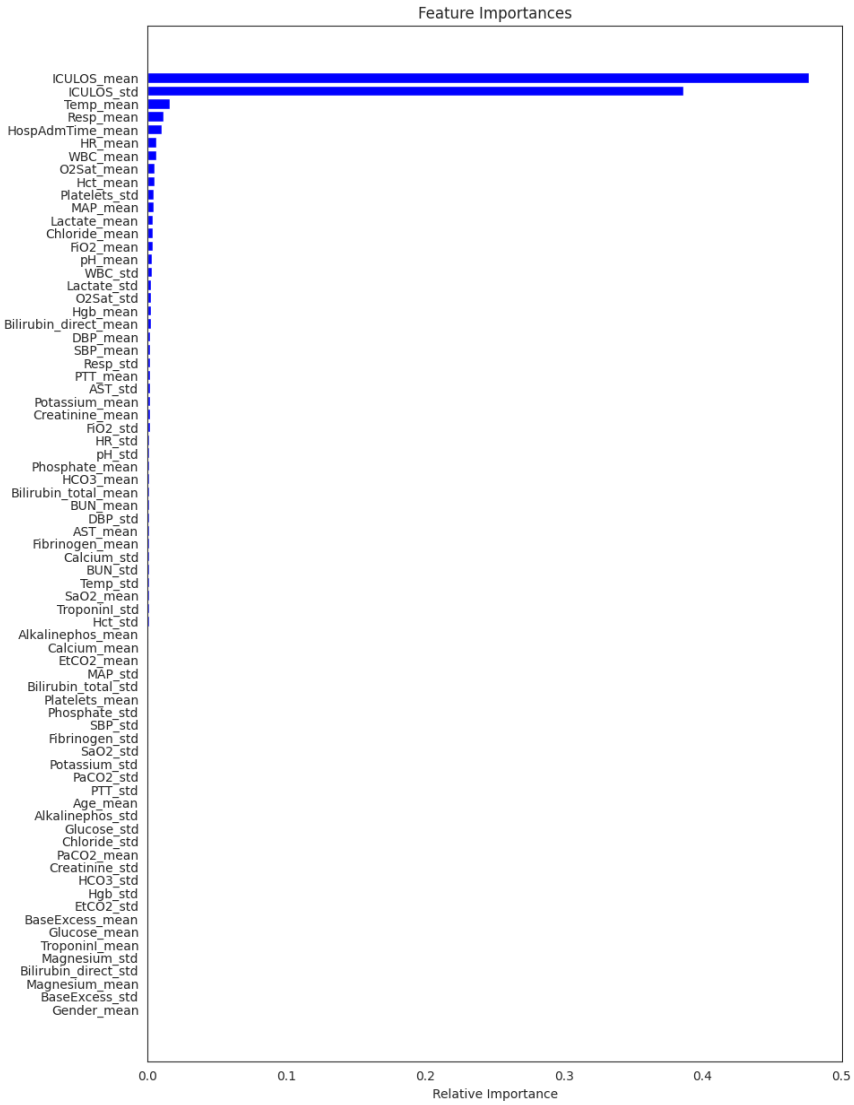
Handling missing data:

ראשית, החלטנו להניח כי מודל החסרות ב-data הוא MAR (missing at random).

לאחר הנחה מודל החסרות הנוכחי, ניסנו למדל את הקשרים בין הדגימות השונות על מנת להשלים את החסרות ב-data ולהשתמש באלגוריתם היעיל ביותר להשלמת הנתונים בצורה המדויקת ביותר. תחילה, ניסינו לבצע זאת בעזרת השלמת ממוצע העמודה ורגרסיה לינארית, אך ראינו ששיטות אלו לא מביאות תוצאות מיטביות. לבסוף, בחרנו להשתמש באלגוריתם KNN להשלמת הנתונים עם היפר פרמטר 3 של מספר השכנים זאת לאחר ביצוע cross validation על מספר השכנים המיטבי.

Feature Engineering

Which feature you will be using (and why)



אנו השתמשנו בכמה כלים על מנת להחליט מהם הפיצ'רים האיכותים ביותר על מנת להשיג תוצאות מיטביות במודלים השונים. ראשית, ניסנו להבין מהם הפיצ'רים המשמעותיים ביותר בתהליך הלמידה של המודלים השונים. ניתן לראות למטה את התוצאות שקיבלנו כך שהפיצ'ר הראשון בטבלה הוא הפיצ'רים בעל ההשפעה הגדולה ביותר על תהליך הלמידה של המודלים. זהו המדד הראשון שעזר לנו להבין אילו פיצ'רים משמעותיים יותר ופחות בדאטא.

בנוסף, ביצענו cross validation על הפיצ'רים והמודלים ונמצא כי איגוד הפיצ'רים הבאים הם אלו שמניבים את התוצאות המיטביות ביותר :

['HR_mean', 'O2Sat_mean', 'Temp_mean', 'MAP_mean', 'Resp_mean', 'Chloride_mean', 'Lactate_mean', 'Hct_mean', 'Hgb_mean', 'WBC_mean', 'Platelets_std', 'HospAdmTime_mean', 'ICULOS_mean', 'ICULOS_std','label']

ניתן לראות כי כל הפיצ'רים שנבחרו ב-cross validation ממוקמים במקומות הראשונים בטבלת "חשיבות פי'צרים" שפירטנו עליה מעלה מה שמחזק עוד יותר את זה שאכן אלו פיצ'רים משמעותיים שידייקו את המודל. לכן, אלו הם הפיצ'רים שבחרנו להרצת המודל.

Features transformations

כיוון שאנו עובדים עם data סדרתי, ראינו לנכון לבצע התאמות של הדאטא כך שהמודלים השונים יוכלו להפיק תועלת מיטבית מהמידע. בשל כך, אנו החלטנו לבצע אגריגציה למידע כך שעבור כל פיצ'ר מקורי מה-data ניצור 2 פרמטרים חדשים עימם נבצע את למידת המודל – חישוב ממוצע וסטיית תקן. הטרנספורמציה מחושבת כך שעבור כל אדם שיש לו תיעוד רפואי, מחושבים הממוצעים וסטיות התקן עבור כל מדד שנאסף על אותו המטופל בפרק זמן השהייה בבית החולים ואלו הם הפרמטרים הסופיים המייצגים כל חולה בלמידת המודלים. כלומר, כעת כל חולה מיוצג על ידי שורה אחת בדאטא בה יש מספר כפול של פיצ'רים כיוון שעבור כל פיצ'ר שבחרנו (לפי ה-cross validation) יהיה כעת פיצ'ר ממוצע ופיצ'ר סטיית תקן שלו.

החלטנו להשתמש בטרנספורמציה זו על סמך מחקרים ומאמרים שונים שראינו שנעשה שימוש בטכניקה זו אשר הניבה דיוק והצלחה גבוהה יותר בזיהוי מוקדם של אלח דם.

Prediction :

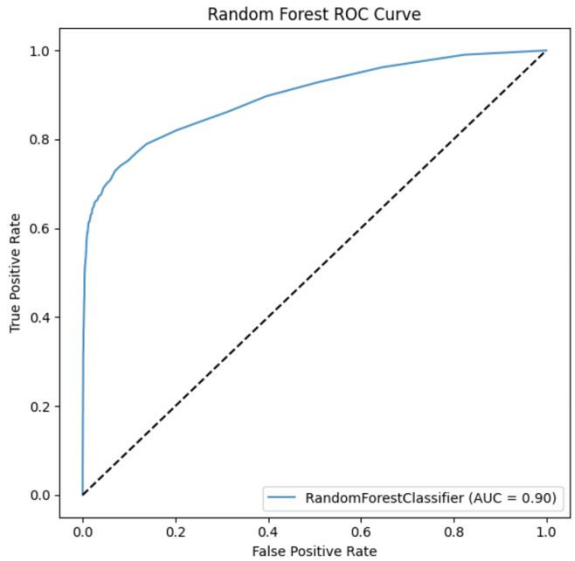
אלגוריתם	Random Forest	Neural Network	XGBoost
	שיטת לימוד אנסמבל הבונה מספר עצי החלטה ומשלבת את התחזיות שלהם למשל באמצעות מיצוע.	רשתות נוירונים מורכבות משכבת קלט, שכבה נסתרת אחת או יותר ושכבת פלט. כל שכבה מורכבת מנוירונים מחוברים, כאשר המידע זורם משכבת הקלט דרך השכבות הנסתרות אל שכבת הפלט	XGBoost (Extreme Gradient Boosting): שיטת לימוד אנסמבל אשר משתמשת בגישת gradient boosting, זוהי שיטה הבונה סדרת לומדים חלשים כדי לתקן את הטעויות של המודלים הקודמים.
Hyperparameter Selection and Regularization	ההיפר פרמטר היחיד שלבסוף בחרנו להשתמש באלגוריתם זה הוא n_estimators = 200 כלומר 200 הוא מספר העצים בהם נשתמש ב-random forest	מספר שכבות חבויות-1, אקטיבציית -'relu' ובשכבה האחרונה 'sigmoid' , פונקציית לוס- binarry cross entropy, גודל השכבות לפי הסדר מהראשונה לאחרונה 256-<32-<1. מאפטמם מסוג 'adam' , עם 100 אפוקים וגודל batch של 16	באלגוריתם זה, מספר העצים שבחרנו להשתמש בו הוא 1000, שיעור למידה של 0.001 שבחרנו לאחר CV ושיפרנו רגולריזציה על ידי בחירה רנדומית של 50% מכל עץ באמצעות 'subsample'
Training and Validation Results	F1 על סט האימון-0.999 F1 על סט המבחן-0.6721	F1 על סט האימון-0.6965 F1 על סט המבחן-0.6488	F1 על סט האימון- 0.7288 F1 על סט המבחן-0.695
Post Analysis	מדדנו את ביצועי המודל על שלוש קבוצות-1)כל הפיצ'רים ועוד שתי קבוצות עליהן פירטנו בסעיף	גם במקרה זה בדקנו על שלושת הקבוצות על התוצאות כפי שהוסבר	גם במקרה זה בדקנו על שלושת הקבוצות על התוצאות כפי שהוסבר

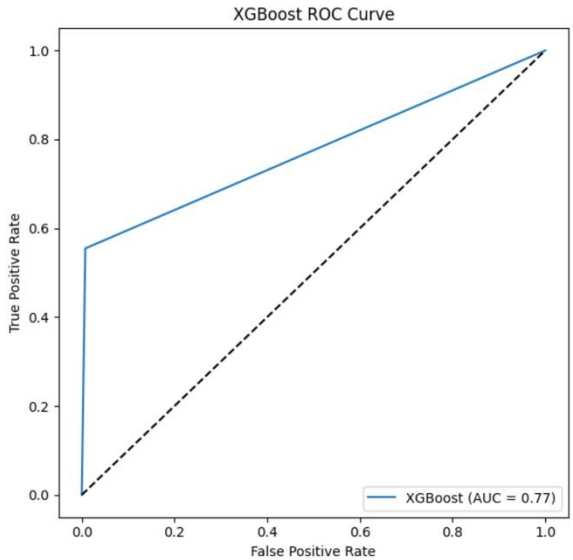
	הפיצ'רים, לבסוף בחרנו את הקבוצה שנתנה את התוצאה הטובה ביותר שזו הקבוצה אותה מצאנו באמצעות אלגוריתם ה cross validation כפי שתארנו. בנוסף ניסינו לשפר את הרגולריזציה כיוון שניתן לראות במקרה זה כי נוצר overfitting באמצעות הגבלת עומק העץ אך הדבר פגע בתוצאות, לכן קיבלנו כי במקרה זה הoverfitting לא הזיק	בrandom forest, ניסינו מספר שינויים גם בפונקציית האקטיבצייה וגם בגודל ומספר שכבות הרשת אך השינויים רק גרעו.	בrandom forest, בנוסף ניסינו באופן דומה עם מבחינת פרמטרים להריץ אלגוריתם Gradient boosting רגיל אשר נתן לנו בשני סוגי הסטים תוצאות פחות טובות.

בהקשר להשוואת המודלים שביצענו, ניתן לראות מקרה מעניין שלמרות שבמרוצת השנים מודלי למידה עמוקה עקפו ביכולתם את מודלי ה-machine learning הקלאסיים, במחקר שלנו, דווקא מודלי העצים הם אלו שהנפיקו תוצאות טובות יותר. ניתן להסיק כי למרות שרוב הטכנולוגיות העכשוויות הולכות לשימוש במודלי למידה עמוקה, ישנם עדיין תחומים (עולמות הרפואה וכו') שבהם מודלים קלאסיים עדיין משיגים תוצאות מרשימות ואף טובות יותר ממודלי למידה עמוקה.

<u>roc curve :</u>

עקומת ROC מודדת את ה trade-off בין השיעור החיובי האמיתי (true positive rate) לשיעור החיובי השקרי (false positive rate). כמו כן, עקומת ROC טובה יותר פירושה שלמודל ה-random forest יש שיעור חיובי אמיתי גבוה יותר עבור שיעור חיובי שגוי נתון בהשוואה למודל xgboost. במילים אחרות, מודל random forest טוב יותר בלהבחין בין מקרים חיוביים לשליליים.





מצד שני, ציון F1 מתייחס גם ל- precision וגם ל-recall , המושפעים ממספר התוצאות החיוביות האמיתיות, החיוביות השגויות ושליליות השגויות. כלומר, מדד ה-f1 מספק איזון בין שני המדדים. לכן, ייתכן שמודל ה-xgboost משיג ציון F1 גבוה יותר כמו שראינו מכיוון שיש לו איזון טוב יותר בין recall ל-precision, למרות שהיכולת שלו להבחין בין מקרים חיוביים ושליליים (כפי שמוצג על ידי עקומת ROC) עשויה להיות מעט פחות טובה מזו של ה-random forest.

כמו כן, במערכי נתונים לא מאוזנים, שבהם ההתפלגות של מקרים חיוביים ושליליים מוטה מאוד, הערכת ביצועי המודל רק על סמך דיוק עשויה שלא לספק ייצוג מדויק של יעילותו. במקרים כאלה, מדדים כמו ציון F1 הופכים לבעלי ערך רב יותר מכיוון שהם לוקחים בחשבון גם דיוק וגם זכירה. במקרה שלנו, אכן יש בידנו מערך נתונים לא מאוזן מה שמעלה מאוד את ערכו של מדד f1 ביחס למדדים אחרים.

כתוצאה מכך, בשל מערך הנתונים הלא מאוזן שבידנו, אכן הגיוני שלמודל ה-random forest עשוי להיות עקומת ROC טובה יותר אך מספר גבוה יותר של שליליות שגויות. לעומת זאת, מודל ה-xgboost משיג איזון טוב יותר בין recall ו-precision, וכתוצאה מכך ציון F1 גבוה יותר. זה מציג את החשיבות של התחשבות במדדי הערכה כמו ציון F1 במערכים לא מאוזנים, מכיוון שהוא מספק הערכה מקיפה יותר של ביצועי המודל.

Summary and Discussion:

לסיכום, לאחר ביצוע כל הצעדים והסקת המסקנות הנדרשות אותן פירטנו מעלה, המחקר מצא כי מודל הניבוי xgboost השיג דיוק ורגישות הגבוהים ביותר בזיהוי חולים שעלולים לפתח אלח דם. המודל הצליח לזהות חולים עם אלח דם לפחות 6 שעות לפני זיהוי קליני, מה שיכול לאפשר לרופאים להתערב מוקדם יותר ולשפר את תוצאות המטופל.

לממצאי המחקר יש השלכות חשובות על ניהול אלח דם ויכולים לעזור להפחית את שיעורי התמותה הקשורים למצב. גילוי מוקדם של אלח דם הוא קריטי לטיפול יעיל, ואלגוריתמי למידת מכונה יכולים לספק כלי רב ערך לשיפור הטיפול באלח דם. עם זאת, דרוש מחקר נוסף כדי לאמת את ממצאי המחקר ולקבוע כיצד לשלב בצורה הטובה ביותר מודלים חזויים בפרקטיקה הקלינית.