Journal of Blood Medicine

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Joint Damage Prediction in Non-Severe Hemophilia A with Artificial Intelligence

Ana Marco-Rico[1], José Manuel Calvo-Villas[2], Francisco-Jose Lopez-Jaime[3], Mariana Canaro Hirnyk[4], María del Mar Nieto-Hernández[5], Sonia Herrero Martín[6], Laura Entrena-Ureña[7], Shally Marcellini-Antonio[8], Bolivar L Diaz-Jordan[9], Sergio Jurado-Herrera[10], Noelia F Pérez-González[10], Covadonga García-Díaz[11], Faustino García-Candel[12], Ihosvany Fernández-Bello[1], Jorge Mateo-Sotos[13,14], Pascual Marco-Vera[1,15]

[1]Department of Hematology and Hemotherapy, Dr. Balmis General University Hospital-ISABIAL, Alicante, Spain; [2]Department of Hematology and Hemotherapy, Miguel Servet University Hospital, Zaragoza, Spain; [3]Department of Hematology and Hemotherapy, Málaga Regional University Hospital-IBIMA, Málaga, Spain; [4]Department of Hematology and Hemotherapy, Son Espases University Hospital, Palma de Mallorca, Spain; [5]Department of Hematology and Hemotherapy, Jaén University Hospital, Jaén, Spain; [6]Department of Hematology and Hemotherapy, Guadalajara University Hospital, Guadalajara, Spain; [7]Department of Hematology and Hemotherapy, Virgen de Las Nieves University Hospital, Granada, Spain; [8]Department of Hematology and Hemotherapy, Segovia General Hospital, Segovia, Spain; [9]Department of Hematology and Hemotherapy, Valdepeñas General Hospital, Ciudad Real, Spain; [10]Department of Hematology and Hemotherapy, Torrecárdenas University Hospital, Almería, Spain; [11]Department of Hematology and Hemotherapy, Burgos University Hospital, Burgos, Spain; [12]Department of Hematology and Hemotherapy, Virgen de la Arrixaca University Hospital, Murcia, Spain; [13]Medical Analysis Expert Group, Castilla-La Mancha University, Cuenca, Spain; [14]Health Research Institute in Castilla-La Mancha (IDISCAM), Toledo, Spain; [15]Department of Clinical Medicine, Miguel Hernández University, Alicante, Spain

Correspondence: Ihosvany Fernández-Bello, Department of Hematology and Hemotherapy, Dr. Balmis General University Hospital-ISABIAL, Avenida Pintor Baeza 12, Alicante, 03010, Spain, Email ihosvanyf@yahoo.es

Purpose: Patients with non-severe hemophilia A (PwnSHA) can develop joint damage (JD). The objective was to identify a machine learning model based on routinely collected variables to predict the presence of JD in PwnSHA.

Patients and Methods: A nationwide, multicenter, cross-sectional study was conducted. Clinical and laboratory variables to assess joint health were included. Predictors were age, target joint history, thrombin generation capacity, baseline factor VIII (FVIII) measured by one-stage clotting (FVIII-CLOT) and chromogenic (FVIII-CHR) assays, and the FVIII-CLOT/FVIII-CHR ratio. The joint condition was described using the HEAD-US score. JD was defined as HEAD-US >0. A Random Forest (RF) ensemble was trained with regression-based multiple imputation, z-scaling, and Synthetic Minority Oversampling within a stratified five-fold stratified cross-validation repeated 100 times. Support Vector Machine, Decision Tree, Gaussian Naïve Bayes and k-Nearest Neighbors were used as comparators. Model performance was assessed on held-out test folds, and 95% confidence intervals (CIs) were obtained by bootstrap resampling with 10,000 repetitions.

Results: Eighty-four Spanish males ≥12 years old were enrolled. Forty-two percent (35/84) had JD. JD was present in 30% (3/10) of patients with moderate hemophilia and 43% (32/74) with mild hemophilia. The RF achieved an accuracy of 92.0% (95% CI: 90.72–93.31), a recall of 92.1% (95% CI: 90.87–93.41), a specificity of 91.9% (95% CI: 90.58–93.27), and an AUC-ROC of 0.92 (95% CI: 0.907–0.938), outperforming all alternative classifiers. Permutation-based feature importance identified age, target joint history, thrombin generation and the FVIII-CLOT/FVIII-CHR ratio as the most influential variables.

Conclusion: The RF model identifies PwnSHA more likely to have prevalent, occult JD in a cross-sectional setting, enabling rapid triage for targeted HEAD-US evaluation. External and prospective validation in larger cohorts is now warranted to confirm generalizability and to facilitate integration into electronic health-record decision-support systems aimed at preserving long-term joint health in PwnSHA.

Keywords: non-severe hemophilia a, joint damage, machine learning, random forest, thrombin generation

Introduction

Hemophilia A (HA), an X-linked inherited bleeding disorder caused by factor VIII activity (FVIII) deficiency, is classified according to baseline FVIII levels as severe (<1 IU/dL), moderate (1–5 IU/dL; MoH), or mild hemophilia (>5–40 IU/dL; MiH).[1] Although joint damage (JD) has long been associated with severe HA,[1–4] growing evidence

https://doi.org/10.2147/JBM.S569311
Received: 22 September 2025
Accepted: 10 March 2026
Published: 20 March 2026

indicates that patients with non-severe hemophilia A (PwnSHA) are also susceptible to JD.[5–8] In these patients, asymptomatic or minimally symptomatic microbleeds often go undetected, reducing the likelihood of early joint assessment and allowing irreversible JD to progress unnoticed in the absence of personalized treatment.[8,9]

To support early detection, the HEAD-US (Hemophilia Early Arthropathy Detection with Ultrasound) protocol was developed as a standardized musculoskeletal ultrasound scoring system designed to assist hematologists in the early detection and monitoring JD in patients with hemophilia.[10] Although the HEAD-US enables practical and reproducible detection of early structural changes in hemophilic arthropathy and correlates well with the Hemophilia Joint Health Score,[10,11] its implementation requires training and expertise. This may limit widespread use and delay the timely diagnosis of hemophilic arthropathy in PwnSHA, who are not commonly expected to experience significant joint deterioration.[6]

Beyond imaging, global hemostasis assays, particularly thrombin generation assay (TGA), offer valuable insights into bleeding phenotype by evaluating thrombin generation dynamics, which reflect the patient's overall hemostatic capacity. In particular, TGA parameters correlate more closely with clinical bleeding risk than baseline FVIII levels alone and may enhance prognostic accuracy in PwnSHA.[12–15] At the same time, discrepancies between one-stage clotting and chromogenic FVIII assays are common in this population and can lead to misclassification of disease severity, thereby influencing treatment decisions and potentially contributing to JD.[16–19]

Other routinely collected variables, such as a history of target joints, might also serve as strong predictors of JD in this population.[1,20] By contrast, current evidence suggests that F8 gene mutations have limited clinical relevance to JD development in MiH,[6] and genetic data are often unavailable. Similarly, bleeding risk related to physical activity is seldom documented and may have minimal influence on JD onset in PwnSHA.[6]

These challenges highlight the need for integrative, data-driven tools that can capture the multifactorial nature of JD prediction in PwnSHA. Machine learning (ML) methods are well-suited for this task, as they can incorporate diverse clinical and laboratory variables to generate personalized scores and guide targeted intervention strategies.[21,22] TRIPOD (transparent reporting of studies on prediction models for individual prognosis or diagnosis)-compliant flow diagram is required to enhance transparency and reproducibility. External validation is needed to confirm its applicability to broader populations.[21,22]

Taken together, these insights suggest that artificial intelligence (AI) can integrate diverse clinical, imaging, and hemostatic variables to generate individualized scores aimed at the early identification of JD development in PwnSHA, enabling timely referral to the appropriate specialist for HEAD-US assessment and improving clinical management.

In this study, the objective was to identify an ML model based on routinely collected variables to predict JD development in PwnSHA.

Materials and Methods

Study Design

A national, multicenter, cross-sectional study was conducted in accordance with the Declaration of Helsinki. The study was approved by the Ethics Committee of Hospital Universitario Dr. Balmis (Alicante, Spain; reference PI2022-037) and ratified by the ethics committees of all participating hospitals. Written informed consent was obtained from all patients prior to inclusion. Data collection took place from October 2022 to December 2023.

Patients

Male PwnSHA aged ≥12 years with no history of prophylactic treatment were enrolled. All patients have signed the informed consent prior to inclusion in the study.

Exclusion criteria included other hemostatic disorders, conditions affecting joint health unrelated to hemophilia, history of inhibitors, participation in any experimental study at the time of enrollment or current/prior use of anticoagulant/antiplatelet therapy.

Sample Preparation

Blood samples were collected in 3.8% trisodium-citrate-anticoagulated Vacutainer® tubes (BD Diagnostics, Spain). Plasma was obtained by double centrifugation at 1,500×g for 15 minutes at room temperature and stored in aliquots at −80°C until analysis. For genetic studies, blood was collected in tubes containing EDTA as anticoagulant.

Basal-FVIII Activity

Baseline factor VIII levels measured by one-stage clotting (FVIII-CLOT, International Laboratory, Bedford, MA, USA) and chromogenic (FVIII-CHR, International Laboratory, Bedford, MA, USA) assays were obtained from routine clinical laboratories at each participating center, using standardized commercial platforms in accordance with local clinical practice. To reduce the impact of inter-laboratory variability, FVIII values were treated as continuous variables and further integrated using the FVIII-CLOT/FVIII-CHR ratio, which provides a harmonized measure of assay discrepancies and is less sensitive to systematic calibration differences across laboratories. Baseline FVIII-CHR levels were used to define hemophilia severity, in agreement with the International Society on Thrombosis and Hemostasis (ISTH) recommendations.[23]

Thrombin Generation Assay (TGA)

TGA was determined centrally using the ST®-BleedScreen assessment on the STA-Genesia® analyzer (Diagnostica Stago, Paris, France) designed by Hemker et al.[24] Lag-time (LT, start time of thrombin generation), peak of thrombin generation (Peak; maximum thrombin levels reached), the velocity index (VI; rate of reaching the peak), the endogenous thrombin potential (ETP; total amount of thrombin generated), time to peak (TTP, the time to reach the maximum peak of thrombin) and start tail (ST, the time to thrombin neutralization) were analyzed. A commercially certified reference normal plasma was provided with the ST®-BleedScreen kit. Accordingly, Peak, VI, and ETP were expressed as percentages relative to normal plasma, whereas Lag-time, Time-to-Peak, and Start-Tail were reported as ratios.

HEAD-US Scoring and Target Joints

The HEAD-US index was determined by an expert staff from each hospital. This index evaluates the health condition of the six major joints (elbows, knees and ankles) individually for signs of synovitis, as well as cartilage and osteochondral damage, utilizing a simplified scoring system.[7] The total value for every patient was calculated by summing the scores obtained for the six joints. JD was defined as a HEAD-US >0. A target joint was defined, in accordance with international consensus guidelines, as a joint that experienced three or more spontaneous bleeding episodes within a 6-month period.[1] Although formal inter-rater reliability metrics (eg., intraclass correlation coefficients) were not assessed in this study, the use of a validated scoring system and trained operators was intended to minimize inter-observer variability.

F8 Genotyping

To confirm the diagnosis of hemophilia, F8 mutations were obtained from clinical records or, if unavailable, determined at the Congenital Coagulopathies Laboratory, Blood and Tissue Bank (Barcelona, Spain) using next-generation sequencing with a congenital coagulopathy gene panel. This methodology included amplification of exonic, flanking intronic, and promoter regions using the CleanPlex kit and MiSeq platform, followed by variant analysis and Sanger confirmation.

Data Description and Group Comparisons

Numerical variables were described using the median (IQR: interquartile range) and categorical ones by percentages or counts. Statistical tests were chosen based on data distribution and performed with SPSS Statistics software (IBM, version 26). A p-value <0.05 was considered statistically significant.

As the primary objective was to develop an ML algorithm for predicting JD, traditional sample size calculations were not used due to the nature of the ML methodologies (see Machine Learning for predicting JD section below). Instead, data availability guided the dataset size, focusing on maximizing sample diversity and representativeness for robust model training and validation.

ML for Predicting JD

A Random Forest (RF) model was developed and benchmarked in MATLAB R2024a (The MathWorks, Natick, MA, USA) to classify PwnSHA with/without JD according to the HEAD-US score. The selection of RF was motivated by three clinically relevant considerations: 1) The bootstrap-aggregation design reduces the risk of overfitting in small, heterogeneous patient cohorts; 2) The tree-based structure can capture complex, non-linear interactions frequently observed in biological data; 3) The variable-importance measures provide intuitive rankings of predictors that enhance clinical interpretability. All computations were performed on a workstation equipped with an Intel® Core™ i9-13900K processor and 64 GB of RAM running Windows 11 Pro.

Preprocessing began with a systematic review of the raw variables. Continuous predictors were adjusted by limiting extreme values at the 1st and 99th percentiles to reduce the influence of transcription outliers, whereas categorical variables were converted into binary indicator variables (1 = yes, 0 = no). Missing values, which never exceeded 4% for any single variable, were imputed using regression-based multiple imputation (five replicates with predictive mean matching). Rubin's rules were applied to ensure that the uncertainty introduced by imputation propagated through all subsequent analytical steps. Normalization (z-scoring) of continuous features was performed within each resampling fold to prevent information leakage.

Hyperparameter optimization followed a nested grid-search protocol embedded within an internal five-fold stratified cross-validation loop. The final RF configuration consisted of 500 trees, bootstrap aggregation, unlimited maximum tree depth (MaxNumSplits = Inf) and the square root of the total number of predictors sampled at each split. The minimum leaf size was set to 1, and a minimum of two observations were required to attempt a split. Surrogate splits were disabled, class priors were estimated empirically from the training data, and a uniform misclassification cost matrix was used. All analyses were conducted in MATLAB R2024a (The MathWorks, Natick, MA, USA) using the Statistics and Machine Learning Toolbox. The configuration yielding the highest mean area under the ROC curve (AUC-ROC) across inner folds was retained for outer-loop evaluation.

Class imbalance-damaged joints accounted for roughly 40% of the sample, and overall data scarcity was addressed using the Synthetic Minority Oversampling Technique (SMOTE). Within each training fold, SMOTE interpolated each minority observation with its 5 nearest neighbors, thereby generating synthetic joints that preserved the multivariate structure of the data without contaminating the validation set. The corresponding test folds remained completely untouched. This design ensured an unbiased evaluation of model performance and prevented any form of information leakage.

Calibration performance was evaluated using the Brier score, calibration slope, and calibration-in-the-large (CITL) derived from out-of-fold predictions.

To assess performance stability, variability across the 100 repeated stratified five-fold cross-validation runs was quantified.

Comparator models (support vector machine (SVM) with radial basis function kernel, decision tree (DT), Gaussian naïve Bayes (GNB), and k-nearest neighbors (KNN)) were subjected to identical resampling procedures, with hyperparameters tuned independently to ensure fair comparison.

The optimal cut-off was determined by maximizing Youden's statistic (sensitivity + specificity − 1) on the outer-fold predictions, thereby tailoring the operating point to the clinical costs of false positives and false negatives. Model discrimination and calibration were assessed primarily using the AUC-ROC, and supplemented with accuracy, sensitivity, specificity, precision, F1-score, Matthews correlation coefficient, Cohen's κ, and the discrimination-Youden index. Ninety-five percent confidence intervals (CI) for all metrics were obtained by bootstrapping with 10,000 repetitions, and visual diagnostics included ROC curves, precision–recall curves, and calibration curves (see Supplementary Material 1).

To mitigate the risk of overfitting in the context of a limited sample size, all preprocessing steps, including multiple imputation, feature scaling, and SMOTE, were performed strictly within the training partitions of each cross-validation fold. Hyperparameter tuning was embedded within a nested cross-validation framework. Model performance metrics

were derived exclusively from held-out test folds and aggregated across 100 repeated stratified five-fold cross-validation runs, providing a robust estimate of generalization performance and minimizing optimistic bias.

Results

Patient Demographic and Clinical Characteristics

A total of 84 PwnSHA were included: 10 MoH and 74 MiH patients. Demographic and clinical characteristics of patients were summarized in Table 1. Detailed information on individual age, target joint history, FVIII-CLOT, FVIII-CHR and ratios are available upon request.

Thrombin Generation Assay (TGA)

TGA results were summarized in Table 1. Not statistically significant differences were found in TGA between MoH and MiH patients. FVIII-CHR exhibited mild correlations with VI ($r = 0.251$, $p = 0.043$), LT ($r = -0.258$, $p = 0.038$), and TTP ($r = -0.412$, $p = 0.001$), while FVIII-CLOT only showed mild correlations with LT ($r = -0.255$, $p = 0.041$) and TTP ($r = -0.462$, $p < 0.001$).

Joint Damage (JD)

A total of 502 joints were examined using the HEAD-US scoring system. The assessment of joint condition revealed varying degrees of severity across the patient cohort. Detailed information on individual joint assessments, including specific clinical findings are provided upon request.

Briefly, 42% (35/84) of PwnSHA had JD. Within this subset, JD was present in 30% (3/10) of patients with MoH and 43% (32/74) with MiH. No significant differences in the HEAD-US score were observed between MoH and MiH patients (see Table 1). Similarly, FVIII-CHR and FVIII-CLOT levels did not differ between patients with and without JD. Patients

Table 1 Summary of Demographic and Clinical Data for Study Subjects

Parameters	MoH	MiH	p-value
	n= 10	n= 74	
Age (years)	37.6 (21.7–45.0)	38.4 (17.7–48.9)	0.961
FVIII-CHR (IU/dL)	4.0 (1.0–5.0)	15.0 (10.0–22.0)	< 0.001
FVIII-CLOT (IU/dL)	12.7 (7.1–18.1)	16.6 (11.0–23.8)	0.419
Target joint history (n)	0	4	NA
HEAD-US	0.0 (0.0–6.0)	0.0 (0.0–3.0)	0.627
Peak (%)	34.3 (27.8–51.5)	46.2 (31.0–58.7)	0.405
ETP (%)	64.2 (49.2–78.3)	69.0 (49.3–83.3)	0.566
Velocity Index (%)	14.6 (12.5–21.5)	25.4 (14.2–38.7)	0.225
Lag-time (ratio)	1.2 (1.0–1.5)	1.2 (1.0–1.4)	0.971
Time-To-Peak (ratio)	1.6 (1.3–1.8)	1.5 (1.3–1.7)	0.758
Start-Tail (ratio)	1.6 (1.4–1.9)	1.4 (1.3–1.6)	0.105

Abbreviations: MoH, moderate hemophilia; MiH, mild hemophilia; FVIII-Chr, basal levels of factor VIII obtained using a chromogenic assay; FVIII-Clot, basal levels of factor VIII obtained by a one-step clotting assay; NA, not applicable; HEAD-US, Hemophilia Early Arthropathy Detection with UltraSound; Peak, Maximum thrombin concentration reached; ETP, endogenous thrombin potential.

https://doi.org/10.2147/JBM.S569311

with JD exhibited significantly lower thrombin generation, as evidenced by a longer TTP (1.7 [IQR: 1.5–1.8] vs. 1.5 [IQR: 1.3–1.6]; p = 0.022) and a lower VI (19.1% [IQR: 12.3–29.3] vs. 26.8% [IQR: 18.4–41.2]; p = 0.03).

Also, to complement the main ML analysis, a separate subgroup analysis focused on patients with a history of target joints. Four individuals with MiH reported at least one target joint, although none opted for prophylaxis treatment. As with JD status, no significant differences in FVIII-CHR or FVIII-CLOT levels were observed between those with and without target joints. Regarding TGA parameters, a significant prolonged ST ratio (1.7 [IQR: 1.7–2.0] vs. 1.4 [IQR: 1.3–1.6], p = 0.004) and a trend toward reduced Peak and VI values were described in patients with target joints compared to those without target joints (Peak: 30.9% [IQR: 21.8–37.8] vs. 46.9% [IQR: 30.9–59.2], p = 0.068; VI: 12.5% [IQR: 11.7–16.6] vs. 25.2% [IQR: 14.5–39.4], p = 0.057), suggesting that TGA may serve as a potential test for predicting joint bleeding risk in this population.

Complementarily, in the overall patient cohort, the HEAD-US score showed significant correlations with key TGA parameters, including VI (r = –0.253, p = 0.042) and TTP (r = 0.356, p = 0.004), reinforcing the potential role of TGA as a biomarker of joint health in this population.

F8 Mutations

A total of 77 patients had a known mutation: 92.2% presented missense mutations, 3.1% had splice site mutations, 1.6% were non-sense mutations, and 3.1% were structural mutations. The large number of mutation types and the low frequency of cases per mutation precluded the inclusion of mutation type as a predictive variable for the presence or absence of JD. This database is available upon request.

ML for Predicting JD

Five ML models were applied for this study: SVM, DT, GNB, KNN and the proposed RF.

As shown in Tables 2 and 3, RF clearly topped every column, achieving an accuracy of 92.0% (95% CI: 90.72–93.31), and an AUC-ROC of 0.92 (95% CI: 0.907–0.938), outperforming all alternative classifiers by 6.9–14.6 percentage points. The superiority of the RF was equally obvious with DYI and Kappa index (92.02% (95% IC 90.76–93.41) and 81.92% (95% IC 80.72–83.17) respectively) and confirmed that the ensemble maintained a balanced trade-off between sensitivity and specificity.

RF turned out to be the most efficient model, with a recall of 92.13% (95% IC 90.87–93.41) and a precision of 91.37% (95% IC 90.13–92.71).

The RF model demonstrated excellent calibration, achieving the lowest Brier score (0.12) among all evaluated models, a calibration slope of 1.03, and a near-zero CITL (0.03), indicating minimal systematic bias and reliable probabilistic estimates. Calibration metrics for all models are reported in Table S1.

Figures 1 and 2 provided a compact, yet comprehensive, visual synthesis of the different key performance indices obtained in the cross-validated training rounds (see Figure 1) and reproduced on the completely test fold (see Figure 2). In the training phase (see Figure 1), the proposed RF trace approached that ideal geometry, signaling that the ensemble

Table 2 Discrimination Performance of the ML Models on the Testing Phase

Model	Accuracy (%)	Specificity (%)	AUC	Cohen's Kappa (%)
SVM	82.06 ± 0.84 (95% CI: 80.43–83.73)	81.96 ± 0.82 (95% CI: 80.41–83.55)	0.82 ± 0.008 (95% CI: 0.808–0.838)	73.05 ± 0.79 (95% CI: 71.62–74.64)
DT	80.19 ± 0.81 (95% CI: 78.63–81.72)	80.09 ± 0.72 (95% CI: 78.65–81.53)	0.80 ± 0.007 (95% CI: 0.782–0.815)	71.39 ± 0.74 (95% CI: 69.93–72.85)
GNB	77.39 ± 1.12 (95% CI: 75.22–79.57)	77.30 ± 1.08 (95% CI: 75.19–79.44)	0.77 ± 0.012 (95% CI: 0.749–0.797)	68.89 ± 0.81 (95% CI: 67.38–70.51)
KNN	85.10 ± 0.72 (95% CI: 83.75–86.51)	85.00 ± 0.71 (95% CI: 83.67–86.34)	0.85 ± 0.007 (95% CI: 0.839–0.864)	75.76 ± 0.69 (95% CI: 74.45–77.12)
RF	92.02 ± 0.63 (95% CI: 90.72–93.31)	91.91 ± 0.63 (95% CI: 90.58–93.27)	0.92 ± 0.006 (95% CI: 0.907–0.938)	81.92 ± 0.63 (95% CI: 80.72–83.17)

Notes: Values are reported as mean ± standard deviation (SD) across 100 repeated five-fold cross-validation runs. Ninety-five percent confidence intervals (95% CI) were estimated by bootstrap resampling (10,000 iterations) of aggregated out-of-fold predictions.

Abbreviations: ML, machine learning; CI, confidence interval; SVM, Support Vector Machine; DT, Decision Tree; GNB, Gaussian Naive Bayes; KNN, K-nearest neighbors; RF, Random Forest; AUC, area under the ROC curve.

Table 3 Classification Metrics Describing Predictive Agreement of the ML Models on the Testing Phase

Model	Recall (%)	Precision (%)	F1 score (%)	MCC (%)	DYI (%)
SVM	82.15 ± 0.87 (95% CI: 80.42–83.88)	81.47 ± 0.91 (95% CI: 79.73–83.25)	81.81 ± 0.88 (95% CI: 80.22–83.46)	72.81 ± 0.76 (95% CI: 71.36–74.31)	82.06 ± 0.87 (95% CI: 80.42–83.75)
DT	80.28 ± 0.79 (95% CI: 78.76–81.83)	79.61 ± 0.81 (95% CI: 78.03–81.24)	79.95 ± 0.79 (95% CI: 78.42–81.51)	71.15 ± 0.72 (95% CI: 69.79–72.54)	80.19 ± 0.80 (95% CI: 78.63–81.78)
GNB	77.48 ± 1.14 (95% CI: 75.26–79.69)	76.84 ± 1.15 (95% CI: 74.61–79.11)	77.16 ± 0.87 (95% CI: 75.43–78.89)	68.67 ± 0.84 (95% CI: 67.03–70.36)	77.39 ± 0.96 (95% CI: 75.54–79.27)
KNN	85.20 ± 0.64 (95% CI: 83.83–86.56)	84.50 ± 0.71 (95% CI: 83.11–85.96)	84.85 ± 0.67 (95% CI: 83.46–86.23)	75.51 ± 0.66 (95% CI: 74.26–76.79)	85.10 ± 0.68 (95% CI: 83.74–86.49)
RF	92.13 ± 0.63 (95% CI: 90.87–93.41)	91.37 ± 0.67 (95% CI: 90.13–92.71)	91.75 ± 0.64 (95% CI: 90.51–92.97)	81.65 ± 0.61 (95% CI: 80.49–82.87)	92.02 ± 0.65 (95% CI: 90.76–93.31)

Notes: Values are reported as mean ± standard deviation (SD) across 100 repeated five-fold cross-validation runs. Ninety-five percent confidence intervals (95% CI) were estimated by bootstrap resampling (10,000 iterations) of aggregated out-of-fold predictions.
Abbreviations: ML, machine learning; CI, confidence interval; SVM, Support Vector Machine; DT, Decision Tree; GNB, Gaussian Naive Bayes; KNN, K-nearest neighbors; RF, Random Forest; MCC, Matthews Correlation Coefficient; DYI, degenerated Youden index.

captured almost every damaged joint. Crucially, the polygon drawn for the test set (see Figure 2) mirrored the training shape with only a minimal inward deflection.

In Figure 3, RF model achieved an area under the curve (AUC) of 0.92 (95% CI: 0.907–0.938).

Figure 4 summarized how much the RF ensemble relies on each candidate variable when it decided whether a joint is already damaged or still intact.

Age dominated the ranking, followed by the history of target joints, and this is entirely consistent with the natural history of hemophilic arthropathy.

TTP and VI were the most influenced TGA parameters, while the FVIII-CLOT/FVIII-CHR assay ratio emerged as the most informative biochemical variable.

Absolute FVIII-CLOT still contributed, but to a much lesser extent than the kinetic parameters of TGA or the assay ratio. LT and ETP added incremental value by refining the early and late phases of thrombin dynamics, respectively.

A TRIPOD-compliant flow diagram summarizing patient inclusion, predictor selection, and model development is provided in the Supplementary Material 2.

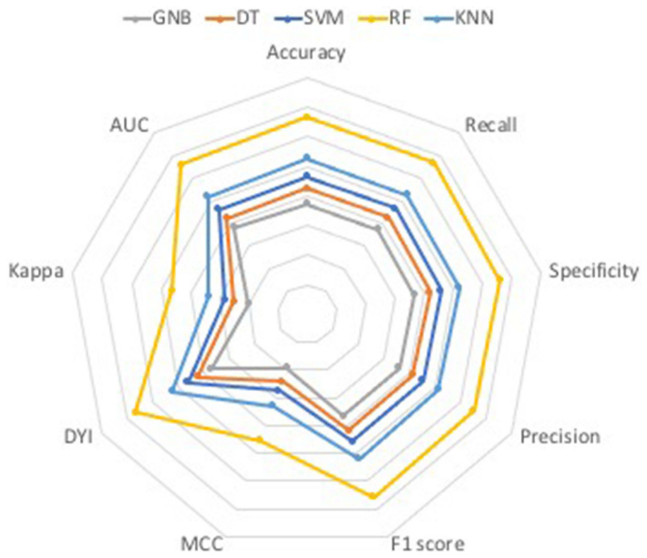

Figure 1 Radar Chart of machine learning algorithm performance during the training phase. All metrics are expressed in percentages (range 60–100%).
Abbreviations: RF, Random Forest; KNN, K-nearest neighbors; SVM, Support Vector Machine; DT, Decision Tree; GNB, Gaussian Naive Bayes; AUC, area under the curve; MCC, Matthews Correlation Coefficient; DYI, degenerated Youden index.

https://doi.org/10.2147/JBM.S569311

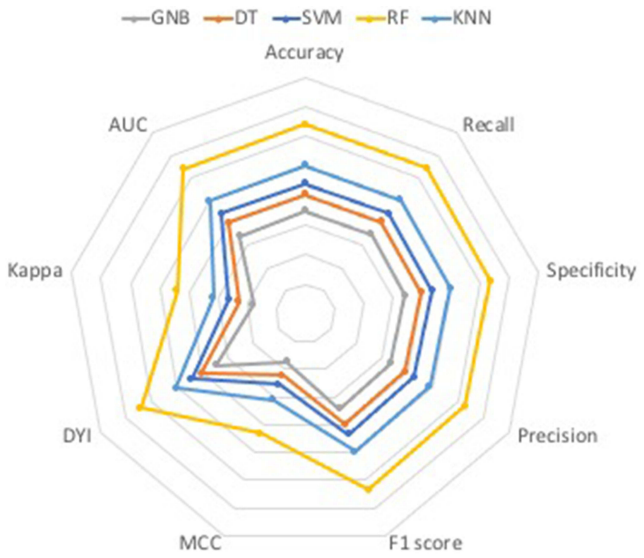

Figure 2 Radar Chart of machine learning algorithm performance during the testing phase. All metrics are expressed in percentages (range 60–100%).
Abbreviations: RF, Random Forest; KNN, K-nearest neighbors; SVM, Support Vector Machine; DT, Decision Tree; GNB, Gaussian Naive Bayes; AUC, area under the curve; MCC, Matthews Correlation Coefficient; DYI, degenerated Youden index.

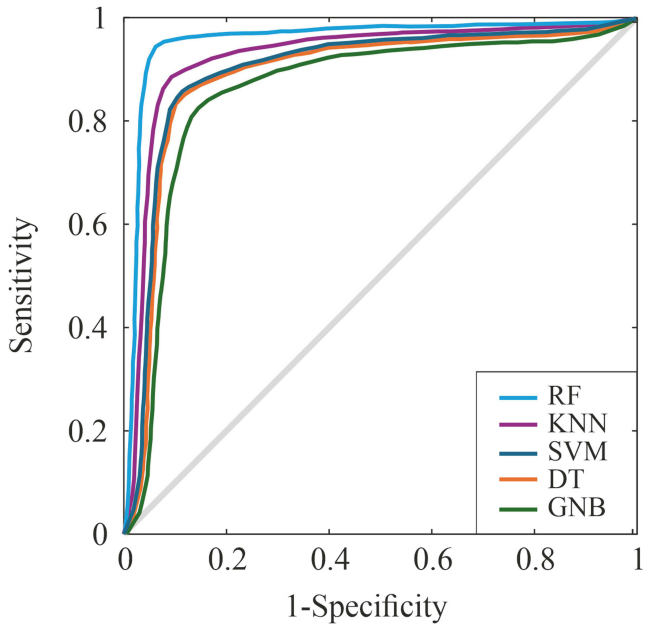

Figure 3 Representation of Receiver Operating Characteristic (ROC) curves of the analyzed algorithms in the testing phase.
Abbreviations: RF, Random Forest; KNN, K-nearest neighbors; SVM, Support Vector Machine; DT, Decision Tree; GNB, Gaussian Naive Bayes.

Discussion

Our study leverages ML models to predict JD in PwnSHA, highlighting key predictive factors such as age, history of target joints and TGA. These findings are pivotal, given the substantial heterogeneity in clinical presentations and treatment responses in hemophilia, which are often not fully reflected by traditional metrics such as baseline factor VIII levels.[25]

Recent data have shown that 36% of patients with MiH may present JD.[4] Our results corroborate this, revealing JD presence in 30% of patients with MoH and 43% with MiH. Additionally, apart from those with a history of target joint, none of the patients with JD reported a previous joint bleeding. This implies that asymptomatic joint microbleeding

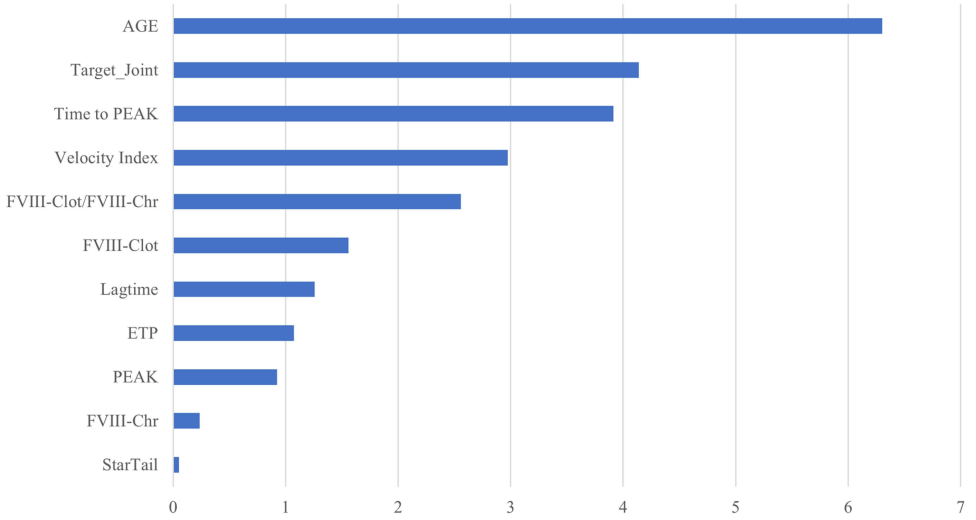

Figure 4 Variable importance rankings in hemophilia prognostic modelling using the RF algorithm. The x-axis quantifies the relative importance of each variable.
Abbreviations: ETP, endogenous thrombin potential; FVIII-CHR, basal levels of factor VIII obtained by a chromogenic assay; FVIII-CLOT, basal levels of factor VIII obtained by a one-step clotting assay; Ratio, FVIII-CLOT/FVIII-CHR.

might progressively deteriorate joint condition, a phenomenon already described in severe hemophilia.[2–4] This underscores the importance of assessing joint condition in PwnSHA, representing a paradigm shift in the management of these patients who have traditionally been considered free from JD.

Notably, the RF algorithm captured this silent progression more effectively than all ML comparator models, as reflected by its superior recall (92.13% (95% IC 90.87–93.41)) and F1-score (91.75% (95% IC 90.51–92.97)), metrics that emphasize its capacity to catch clinically unapparent damage.

Age emerged as the most significant predictor of JD, likely reflecting the cumulative effect of repeated asymptomatic microbleedings over time, which reinforces the importance of early intervention to halt joint deterioration.[7,8]

Discrepancies between FVIII-CLOT and FVIII-CHR significantly impact treatment management. These inconsistencies can lead to undertreatment, risking inadequate protection against bleeding and contributing to JD over time.[16,18] Our group has previously reported different thrombin generation profiles and baseline FVIII levels, even in patients with the same F8 mutation. On top of that, some patients with the same F8 mutation had discrepancies and others not.[26] This reinforces the clinical relevance of resolving assay discrepancies to ensure accurate classification and appropriate treatment decisions in PwnSHA.

Our results show that the target joint history in PwnSHA is a strong predictor of JD development, reflecting the local impact of recurrent bleeds. This strengthens the established notion that joints with prior bleeds are more susceptible to further injury and degeneration.[6,7] The identification of target joint history as an important variable in the predictive ML model highlights the potential value of aggressive treatment to prevent or minimize joint deterioration in patients with target joints.

Our findings also suggest that, alongside baseline FVIII levels, other factors are important in predicting JD in PwnSHA. This supports the idea that while baseline factor levels are essential for diagnosing hemophilia and guiding initial treatment, they do not strongly correlate with long term joint outcomes, highlighting the need for a more holistic approach. TGA demonstrated a mild correlation with FVIII-CLOT, FVIII-CHR, and the HEAD-US score. This mild correlation may be attributable to other factors that contribute to joint deterioration, such as differential local inflammation in response to similar bleeding events,[14,15] and the patient's comprehensive treatment history (time to start treatment, availability of home treatment, etc). All together, these findings underscore the multifactorial influences affecting joint health outcomes, emphasizing the significant challenge of identifying all factors contributing to the development of JD in PwnSHA.

To the best of our knowledge, no ML models have been previously reported in the literature to predict the presence of JD and the associated key factors in PwnSHA.

As shown in the training phase, the RF model outperformed all other algorithms across key performance metrics, demonstrating excellent classification ability in an imbalanced dataset. The high MCC and Kappa values for the RF

https://doi.org/10.2147/JBM.S569311

model further underscore its efficacy in high-quality classification, particularly in settings with imbalanced data. These challenges are frequent in hemophilia, where limited sample sizes and clinical variability often constrain the use of conventional statistical approaches.

The strong performance of the RF model in the testing phase confirmed its generalization capability and supported its reliability beyond the training data. Although the other models remained effective, they showed a slight decrease in performance compared to their training phase results, which is typical when transitioning from the training to the testing phase due to variability and the unique characteristics of the new data. While the RF model showed strong performance during internal validation, supporting its robustness on test data, future efforts will aim to externally validate the model to confirm its applicability to broader populations.

This study presents potential limitations, including the possibility of limited generalizability beyond the studied population. Although the RF model demonstrated robustness during the test phase, it has not yet been validated in an external cohort. This limitation is partly due to the rarity of hemophilia, which poses significant challenges for recruiting a sufficiently large number of patients. Furthermore, the availability of the HEAD-US assessment remains limited in many hospitals, further constraining the inclusion of larger and more diverse patient groups. Additionally, variability between intra- and inter-evaluator assessments of the HEAD-US score can influence diagnostic accuracy and, consequently, the performance of predictive models.

Despite these limitations, this study represents a significant step forward in the field of hemophilia care. It introduces the novel concept that AI, combined with routine available clinical and laboratory data, can be used to predict JD without the need for direct joint imaging, an approach that may be particularly valuable in settings where HEAD-US is not accessible. Our findings lay the groundwork for a more rapid and scalable method to identify patients at risk of joint deterioration, enabling earlier referral to specialized centers. This strategy could help optimize healthcare resources while promoting more personalized care and underscores the importance of sharing these results with the clinical and research communities interested in improving outcomes for PwnSHA.

Conclusion

The present work confirms that JD is not uncommon in people with MiH or MoH and suggests that AI-based data analysis may help identify JD development in PwnSHA who can benefit from regular prophylaxis. External and prospective validation is required.

Acknowledgments

The authors thank the patients and their families for their contribution to the collection of clinical data and samples. They also appreciate the work of the staff who conducted the joint health evaluations, as well as the nursing and laboratory personnel who supported the study procedures. Finally, the Mediterranean Group, composed of the participating study centers, for their significant involvement, scientific guidance, and effort in conducting the study. This work was supported by the Institute of Technology (University of Castilla-La Mancha), Chair of Artificial Intelligence (sponsored by Bayer) and Castilla-La Mancha Institute of Health Research (IDISCAM).

Disclosure

Francisco-Jose Lopez-Jaime reports grants from Bayer, during the conduct of the study; personal fees from Bayer, personal fees from Novo Nordisk, personal fees from Sobi, personal fees from CSL Behring, personal fees from Amgen, personal fees from Octapharma, personal fees from Pfizer, outside the submitted work. The authors declare no other conflicts of interest in this work.

References

1. Srivastava A, Santagostino E, Dougall A, et al. WFH Guidelines for the Management of Hemophilia, 3rd edition. *Haemophilia*. 2020;26(suppl 6):1–158. doi:10.1111/hae.14046
2. Chen CM, Huang KC, Chen CC, et al. The impact of joint range of motion limitations on health-related quality of life in patients with haemophilia A: a prospective study. *Haemophilia*. 2015;21:e176–e184. doi:10.1111/hae.12644
3. Simpson ML, Valentino LA. Management of joint bleeding in hemophilia. *Expert Rev Hematol*. 2012;5:459–468. doi:10.1586/ehm.12.27

4. Manco-Johnson MJ, Le B, Acharya S, et al. Risk factors for joint bleeding in severe hemophilia A and B: analysis of the community counts longitudinal surveillance cohort. *Blood Vessel Thromb Hemost*. 2025;2:100047. doi:10.1016/j.bvth.2025.100047
5. Wang M, Recht M, Iyer NN, Cooper DL, Soucie JM. Hemophilia without prophylaxis: assessment of joint range of motion and factor activity. *Res Pract Thromb Haemost*. 2020;4:1035–1045. doi:10.1002/rth2.12347
6. De la Corte-Rodriguez H, Rodriguez-Merchan EC, Alvarez-Roman MT, Martin-Salces M, Rivas-Pollmar I, Jiménez-Yuste V. Arthropathy in people with mild haemophilia: exploring risk factors. *Thromb Res*. 2022;211:19–26. doi:10.1016/j.thromres.2022.01.010
7. Chiari JB, Prozora S, Feinn R, Louizos E, Gallagher PG, Bona R. Joint bleeds in mild hemophilia: prevalence and clinical characteristics. *Haemophilia*. 2024;30:331–335. doi:10.1111/hae.14939
8. Zwagemaker AF, Kloosterman FR, Hemke R, et al. Joint status of patients with non-severe hemophilia A. *J Thromb Haemost*. 2022;20:1126–1137. doi:10.1111/jth.15676
9. Måseide RJ, Berntorp E, Astermark J, et al. Haemophilia early arthropathy detection with ultrasound and haemophilia joint health score in the moderate haemophilia (MoHem) study. *Haemophilia*. 2021;27:e253–e259. doi:10.1111/hae.14245
10. Martinoli C, Della Casa Alberighi O, Di Minno G, et al. Development and definition of a simplified scanning procedure and scoring method for Haemophilia Early Arthropathy Detection with Ultrasound (HEAD-US). *Thromb Haemost*. 2013;109:1170–1179. doi:10.1160/TH12-11-0874
11. Prasetyo M, Moniqa R, Tulaar A, Prihartono J, Setiawan SI. Correlation between hemophilia early arthropathy detection with ultrasound (HEAD-US) score and hemophilia joint health score (HJHS) in patients with hemophilic arthropathy. *PLoS One*. 2021;16:e0248952. doi:10.1371/journal.pone.0248952
12. Brummel-Ziedins KE, Whelihan MF, Gissel M, Mann KG, Rivard GE. Thrombin generation and bleeding in haemophilia A. *Haemophilia*. 2009;15:1118–1125. doi:10.1111/j.1365-2516.2009.01994.x
13. Sidonio JRF, Hoffman M, Kenet G, Dargaud Y. Thrombin generation and implications for hemophilia therapies: a narrative review. *Res Pract Thromb Haemost*. 2023;7:100018. doi:10.1016/j.rpth.2022.100018
14. Verhagen MJ, van Heerde WL, van der Bom G, et al. In patients with hemophilia, a decreased thrombin generation profile is associated with a severe bleeding phenotype. *Res Pract Thromb Haemost*. 2023;7:100062. doi:10.1016/j.rpth.2023.100062
15. Dargaud Y, Béguin S, Lienhart A, et al. Evaluation of thrombin generating capacity in plasma from patients with haemophilia A and B. *Thromb Haemost*. 2005;93:475–480. doi:10.1160/TH04-10-0706
16. Oldenburg J, Pavlova A. Discrepancy between one-stage and chromogenic factor VIII activity assay results can lead to misdiagnosis of haemophilia A phenotype. *Hamostaseologie*. 2010;30:207–211. doi:10.1055/s-0037-1619052
17. Kitchen S, Bowyer A, Makris M. Both one-stage and chromogenic factor VIII assays are required for the diagnosis of mild hemophilia A. *J Thromb Haemost*. 2023;21:773–775. doi:10.1016/j.jtha.2022.12.013
18. Srivaths L, Larson J, Saroukhani S, et al. Comparing one stage, chromogenic assay results and discrepancies with bleeding phenotype and genetic variants in females with hemophilia A. *J Thromb Haemost*. 2025;23:504–512. doi:10.1016/j.jtha.2024.10.030
19. Verhagen MJ, van Balen EC, Blijlevens NM, et al. Patients with moderate hemophilia A and B with a severe bleeding phenotype have an increased burden of disease. *J Thromb Haemost*. 2024;22:152–162. doi:10.1016/j.jtha.2023.09.029
20. Manco-Johnson MJ, Abshire TC, Shapiro AD, et al. Prophylaxis versus episodic treatment to prevent joint disease in boys with severe hemophilia. *N Eng J Med*. 2007;357:535–544. doi:10.1056/NEJMoa067659
21. Rodriguez-Merchan EC. The current role of artificial intelligence in hemophilia. *Exp Rev Hematol*. 2022;15:927–931. doi:10.1080/17474086.2022.2114895
22. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44–56. doi:10.1038/s41591-018-0300-7
23. Rezende SM, Neumann I, Angchaisuksiri P, et al. International Society on Thrombosis and Haemostasis clinical practice guideline for treatment of congenital hemophilia A and B based on the grading of recommendations assessment, development, and evaluation methodology. *J Thromb Haemost*. 2024;22:2629–2652. doi:10.1016/j.jtha.2024.05.026
24. Hemker HC, Al Dieri R, De Smedt E, Beguin S. Thrombin generation, a function test of the hemostatic-thrombotic system. *Thromb Haemost*. 2006;96:553–561. doi:10.1160/TH06-07-0408
25. Pavlova A, Oldenburg J. Defining severity of hemophilia: more than factor levels. *Semin Thromb Haemost*. 2013;39:702–710. doi:10.1055/s-0033-1354426
26. Marco-Rico A, Calvo-Villas JM, López-Jaime FJ, et al. Real-world evidence on joint condition in non-severe hemophilia A patients: a multicenter study. *J Blood Med*. 2025;16:251–258. doi:10.2147/JBM.S517596