SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 1.1, (2026), pp. 2341-2352
Open Access. Online & Print

DOI: 10.5281/zenodo.11425272

EXPLAINABLE MACHINE LEARNING FOR PRECISION SEVERITY PREDICTION IN SICKLE CELL DISEASE USING MULTI-DOMAIN COMPUTATIONALLY DERIVED CLINICAL PHENOTYPES

Abdul Rahiman S.K.[1*], Anand V.K[2], E Ben George[3], Eman Said Al Abri[4], and Muna Saif Humaid Al Rahbi[5]

[1]University of Technology and Applied Sciences-Muscat, Oman
[2]University of Technology and Applied Sciences-Muscat, Oman
[3]University of Technology and Applied Sciences-Muscat, Oman
[4]University of Technology and Applied Sciences-Muscat, Oman
[5]University of Technology and Applied Sciences-Muscat, Oman

Received: 11/11/2025
Accepted: 18/12/2025

Corresponding Author: Abdul Rahiman S.K
(abdulrahiman.shaik@utas.edu.om)

ABSTRACT

Sickle cell disease (SCD) has a wide clinical variability that makes early severity assessment and individualized management difficult. This paper presents an explainable machine learning framework for SCD severity prediction, using computationally derived phenotypes from real-world hospital data. Routine hematologic and clinical and utilization variables were transformed into a composite phenotype score describing multi-organ involvement, haemolysis, and hypoxic stress. Severity classes mild, moderate, and severe were created through adaptive quantile thresholds versus fixed cut-offs, which allow for data-driven grouping. Several machine learning algorithms were benchmarked, such as logistic regression, random forest, XGBoost, LightGBM, support-vector machines, and multilayer perceptron. The tuned LightGBM model yielded the most stable macro-F1 of approximately 0.83 ± 0.06, along with strong calibration in the held-out set. SHAP-based explainability demonstrated that transfusion intensity, organ dysfunction, and hematologic stress markers were dominant determinants of the predicted severity. These insights were aligned with established SCD pathophysiology and supported transparent model interpretation. Results demonstrate that computational phenotyping combined with explainable learning yields clinically traceable predictions suitable for incorporation into decision-support environments. A framework is presented that strengthens trust in AI-aided precision medicine and points to a pathway toward fair, interpretable severity assessment in rare-disease management.

KEYWORDS: Sickle Cell Disease, Explainable Machine Learning, Computational Phenotyping, Severity Prediction, Multi-Domain Clinical Data, SHAP Explainability, Gradient Boosting Models, Clinical Decision Support.

1. INTRODUCTION

For Sickle cell disease is a hereditary hemoglobinopathy characterized by vaso-occlusion, chronic haemolysis, and progressive organ damage. Clinical manifestations are very variable among subjects; the early identification of severity is both challenging and crucial for the optimization of care. Predictive models of disease severity may help the clinician risk-stratify patients and allow timely intervention. However, many of the current approaches rely on manually annotated datasets or isolated biomarkers, which inherently limit their generalizability and clinical validity.

Machine learning provides a route to analyse complex, multi-domain health data and uncover nonlinear relationships that shape disease trajectories. However, when ML models are used in medicine, transparency becomes as important as accuracy. Black-box algorithms without clear justification cannot be safely adopted in high-stakes clinical decisions. Explainable ML frameworks therefore bridge this gap by revealing how each predictor influences the model outcome.

This work presents an interpretable ML pipeline predicting the severity of SCD from computationally derived phenotypes built from routine hospital records. By transforming raw laboratory, complication, and utilization features into structured multi-domain scores, the framework generates reproducible severity labels and exposes the physiologic rationale behind its predictions. The study shows how explainable learning can move precision medicine forward while ensuring clinical trust and accountability.

2. BACKGROUND

Sickle Cell Disease (SCD) encompasses different clinical trajectories, generating biological, therapeutic, and socio-economic disparities. Consequently, there is great interest in predicting SCD severity, or the risk of severe SCD-related complications over a time horizon. The goal is to help physicians reliably identify non-severe SCD patients and reinforce follow-up to avoid disease aggravation. Beyond prediction efficacy, decisions need to be documented in a transparent manner to support clinicians in understanding and trusting the rationale and data driving model recommendations.

Recent advances in computational hematology and precision medicine have shown that explainable artificial intelligence (XAI) frameworks can transform clinical decision support by integrating multi-modal data streams. Early studies in Sickle Cell Disease (SCD) concentrated on descriptive epidemiology rather than predictive modeling. Bernaudin et al established the first clinical severity scales through transfusion and crisis frequency criteria, which later served as the foundation for quantitative phenotyping approaches [1]. Subsequent work by Chakraborty et al. and Siberry et al. demonstrated that longitudinal organ-damage indicators provide higher prognostic utility than episodic clinical counts [2][3].

Machine learning (ML) methods have since been adopted to map these complex clinical interactions. Abraham et al. highlighted the importance of transparency and reproducibility when applying black-box algorithms to SCD data, stressing that model interpretability is indispensable for regulatory translation [4]. Complementary to these concerns, Lundberg and Lee introduced SHAP (SHapley Additive exPlanations), now regarded as the de-facto standard for global and local feature attribution in biomedical ML models [5].

The emergence of graph-based and multi-modal frameworks expanded phenotype representation across domains. Bronstein et al. described geometric deep-learning paradigms that allow integration of heterogeneous patient graphs, while Zitnik et al. showed how graph neural networks (GNNs) can infer polypharmacy side effects a mechanism conceptually aligned with multi-domain SCD modelling [6][7]. In haematological prediction contexts, Wang et al. demonstrated improved diagnostic accuracy for thalassemia and anemia using hybrid ensemble models [8].

Explainability in clinical AI has matured into a discipline emphasizing clinician trust, as captured by Holzinger et al. through the "glass-box" paradigm promoting human-centered ML design. More recently, Tjoa and Guan [10] surveyed explainable healthcare AI, concluding that hybrid symbolic–neural methods yield superior clinician acceptance compared with purely opaque models [9][10]. These foundational contributions underpin the present study's rationale namely, that an interpretable, multi-domain ML pipeline can deliver clinically credible severity predictions for SCD while preserving reproducibility and fairness.

Beyond foundational interpretability methods, recent literature has focused on clinical translation of explainable and fair AI systems in rare diseases. Diao et al. presented an interpretable gradient-boosting framework that predicts SCD crises using routine haematology data, illustrating the clinical promise of ensemble transparency [11]. Likewise, Churpek et al. demonstrated that calibrated risk scores outperform opaque deep models for hospital deterioration

prediction, emphasizing the balance between performance and trust [12].

In rare-disease analytics, rare-event bias and sample imbalance remain major obstacles. Fernández et al. reviewed data-level and algorithmic strategies to mitigate imbalance, which are particularly relevant to SCD's skewed severity distribution [13]. The adoption of ordinal regression for disease staging by Cardoso and Sousa established an early precedent for handling progression-type outcomes, later extended by Baccianella et al. for multi-class medical tasks where intermediate severities hold semantic order [14] [15].

Integrative, multi-domain modelling has also advanced rapidly. Zhang et al. proposed cross-domain representation learning for clinical phenotypes by coupling EHR data with imaging and lab features; the approach parallels the multi-domain fusion strategy adopted in this study [16]. Deep ensemble calibration techniques introduced by Guo et al. highlighted the over-confidence problem in neural networks and provided reliability-based correction an idea echoed in the calibration analysis of our model [17].

Emerging research also focuses on clinical explainability and ethical transparency. Ribeiro et al. popularized model-agnostic local explanations through LIME, while Doshi-Velez and Kim framed interpretability as a scientific discipline demanding measurable transparency metrics [18][19]. In biomedical contexts, Rajkomar et al. demonstrated that large-scale ML in healthcare must adhere to fairness and accountability frameworks to avoid perpetuating existing inequities, a guiding consideration in the current SCD severity-prediction framework [20].

Collectively, these studies illustrate a field converging toward human-interpretable, bias-aware, and reproducible AI pipelines. Building upon them, the present research contributes a multi-domain, explainable, and calibrated model that bridges phenotype derivation with clinically actionable predictions for Sickle Cell Disease.

3. DATA SOURCES AND PHENOTYPE DERIVATION

The dataset used in this study was obtained from hospital electronic medical records and represents a real-world SCD cohort. Each entry corresponds to an individual patient, with variables reflecting routine haematology tests, longitudinal utilization history (hospitalizations, transfusions), major complications, and documented treatment intensity. Because the data originates from standard clinical care rather than a curated research registry, the captured phenotype reflects actual disease burden rather than protocol-filtered sampling.

A raw-to-scored transformation pipeline was implemented prior to model training. First, raw laboratory and clinical variables were harmonized into standardized fields. Missing values were handled using domain-specific plausibility checks and conditional imputation rules. Next, features were mapped into the phenotypic domains defined by the severity rubric, assigning incremental points based on clinically meaningful thresholds. The intermediate result of this transformation produced SeverityScore, a composite phenotype score summarizing multi-organ involvement, haemolysis, hypoxia, and disease activity.

Severity labels (mild, moderate, and severe) were not manually curated but derived automatically through adaptive thresholds applied over the empirical score distribution. A structured phenotyping rubric was used to compute SeverityScore. The score aggregates burden across seven clinically relevant domains: hematologic/haemolysis, inflammatory and hypoxic stress, organ function, health care utilization, major complications, treatment intensity, and symptom burden. Each variable contributes weighted points based on clinically meaningful thresholds (e.g., hemoglobin <7 g/dL → +3; oxygen saturation <92% → +3; ≥10 lifetime hospitalizations → +3), while protective factors such as elevated fetal hemoglobin subtract points. The cumulative score reflects the net physiologic load from multi-system involvement, and the final class is derived from adaptive thresholds over this distribution, where the upper tertile is designated as "severe," the middle as "moderate," and the lower tertile as "mild". Fixed universal cutoffs were avoided because the burden distribution in real-world SCD cohorts is influenced by demographic, treatment, and utilization biases. Adaptive quantile-based thresholds reduce this bias by aligning severity grouping with the observed spectrum of disease expression in the dataset rather than an external assumption.

This rule-based phenotype derivation provides two advantages for downstream modeling. First, it creates a fully traceable severity assignment grounded in clinical reasoning rather than opaque labeling. Second, it ensures that the machine learning model learns from physiologically meaningful signals instead of proxy markers. The resulting derived labels form the supervisory target for the predictive framework described in the next section, where multiple models are benchmarked and

evaluated for generalization and interpretability.

4. METHODOLOGY

The objective of this work is to develop an explainable machine learning framework that predicts SCD severity using phenotype-derived clinical features rather than manual labels. The modelling pipeline operates on a cross-sectional cohort, where each patient is represented by a structured feature vector summarizing multi-organ burden, hematologic stress, treatment intensity, and historical utilization. These features are aggregated into a composite phenotype score and converted into mild, moderate, and severe class labels through adaptive thresholding.

The predictive component follows a supervised learning formulation in which severity class is the target variable. Multiple model families were benchmarked, including logistic regression, random forest, XGBoost, LightGBM, SVM, and multilayer perceptron. In addition to the baseline models, a tuned LightGBM pipeline with L1-guided feature selection was developed to improve generalization stability. Evaluation relied on stratified five-fold cross-validation, followed by performance verification on a held-out test set.

Interpretability is treated as an explicit system requirement rather than a post-hoc consideration. Because SCD progression reflects multi-domain physiology rather than a single marker, the model is required to expose the phenotypic basis of its prediction. Feature importance and attribution mapping are therefore used to link severity decisions back to clinically meaningful domains. This design preserves clinical traceability and reduces the opacity of black-box predictions.

4.1. Feature Engineering

The modelling input is a cross-sectional, tabular representation of each patient. Every record is a fixed-length vector x∈R^dthat aggregates laboratory values, binary complication flags, treatment-intensity markers, and utilization counts captured in routine care. Labels are derived from the rubric-based composite score (SeverityScore) via adaptive thresholds into mild, moderate, and severe classes. The score itself is not provided as a predictor to prevent label leakage; only its component features (after clinical preprocessing) are used as inputs. The rubric domains and their contributors are summarized in Section 3 and supplementary material.

1. Variable typing and encoding: Continuous laboratory variables (e.g., hemoglobin, bilirubin, CRP, ferritin) are retained as numeric features. Complications (ACS, stroke/TCD abnormal, CKD, DVT/PE, pneumonia, osteomyelitis) are encoded as binary indicators to preserve clinical semantics. Utilization variables (hospitalizations, transfusions) are ingested as non-negative integer counts to maintain ordinal burden rather than arbitrary binning. Categorical fields with >2 states (if present) are one-hot encoded.
2. Missingness handling: Clinical data are not missing at random. We therefore avoid blanket imputation. For variables with biologically plausible ranges and sporadic gaps, we use conservative, fold-internal imputers (median or domain-bounded winsorized median). Indicators of "not measured / not observed" are kept as explicit binary flags were clinically meaningful. Intrinsically absent events (e.g., no ICU admission) remain zero; they are not imputed.
3. Scaling policy: We apply model-aware scaling. Magnitude-sensitive learners (SVM, MLP, Logistic Regression) receive z-scored versions of continuous predictors, computed within the training fold only. Tree-based methods (RF, XGB, LGBM) consume unscaled values to preserve native split thresholds and avoid washing out clinically meaningful cut-points.
4. Outliers and distribution shape: Several hematologic and biochemical markers exhibit heavy right tails. We apply light winsorization at extreme percentiles (training-fold statistics) only when necessary to stabilize gradient-based learners, leaving tree methods unaffected. No log transforms are forced a priori; transformations are applied case-by-case if they improve calibration without harming interpretability.
5. Feature reduction and leakage prevention: We enforce a two-stage guard: (i) manual exclusion of administrative or scheduling fields with no clinical signal; (ii) embedded sparsity via L1-regularized screening inside the training pipeline. L1 removes weak and collinear predictors while preserving domain-interpretable variables. The composite severity score and any direct derivatives are excluded from inputs to block trivial pathways to the label.
6. Class imbalance policy: Models are trained under the natural class distribution with stratified k-fold splits. We optimize macro-F1 to avoid majority-class dominance. For

learners that natively support it, we enable class weighting (e.g., Logistic Regression, SVM). Tree ensembles rely on stratification and threshold tuning rather than synthetic oversampling, to avoid distributional artifacts in small clinical cohorts.

After preprocessing, each patient is represented by a compact, clinically grounded feature set spanning hemolysis, organ function, inflammatory/hypoxic stress, utilization burden, major complications, treatment intensity, and symptom burden. This representation balances parsimony (via L1) with coverage of domains known to drive the rubric-based labels, forming inputs for the supervised classifiers evaluated in Section 4.2 and interpreted in Section 4.3

4.2. Model Architecture

We formulate the severity prediction task as a supervised multi-class classification problem defined over a structured, tabular clinical feature space. Each patient is represented as a fixed-length vector x_i ∈ R^d, where d comprises laboratory-derived biomarkers, binary indicators of complications and historical utilization counts. The target variable y_i ∈ {"mild", "moderate", "severe"} is obtained through rubric-based phenotype scoring rather than by manual curation.

Several model classes were tested to capture different inductive biases: Logistic Regression offers a linear, transparent decision surface and acts as a calibration benchmark. Random Forest adds feature subspace sampling and decorrelated tree aggregation that conveys robustness to noisy variables. Gradient-boosted methods, namely XGBoost and LightGBM, leverage additive functional approximators to model higher-order interactions without explicit feature engineering. Support Vector Machine stresses margin maximization in nonlinear feature space, while the Multilayer Perceptron represents a shallow neural architecture capable of learning distributed clinical representations.

All models were trained under the same k-fold stratified splitting policy to control distributional shift. The hyperparameters were optimized to choose the best based on cross-validated macro-F1, considering that the goal should be to penalize misclassification of any severity class rather than optimize for accuracy alone. LightGBM further included L1-regularized feature filtering to reduce redundancy and constrain model complexity, therefore improving generalization and interpretability.

Although it obtained the highest macro-F1 on the held-out split, its fold-level variance was still higher compared to the boosted tree architecture. Thus, stability, monotonic feature response handling, native missingness support, and downstream explanation frameworks supported the selection of the tuned LightGBM model as the primary candidate architecture. This modular architecture further allows easy drop-in replacement of the classifier layer in case of future longitudinal or multimodal extensions.

4.3. Explainability Techniques

Interpretability is treated as a design constraint rather than a post-hoc accessory. Since the target labels are phenotype-derived and grounded in clinical reasoning, the predictive model must expose which domains contribute most strongly to a given severity assignment. Without this linkage, a high-performing classifier would remain clinically unusable. The chosen explainability strategy therefore operates at two levels: global model behaviour and patient-level justification.

At the global level, feature attribution is computed to approximate the contribution of each predictor to the learned decision surface. For gradient-boosted learners such as LightGBM, tree-based SHAP values provide a consistent estimate of marginal feature influence under the model's functional decomposition. This identifies which physiological domains dominate the severity boundary and whether the model aligns with clinically accepted drivers (e.g., high utilization + organ dysfunction rather than isolated lab excursions). Because the engineered features retain domain identity, the resulting attributions remain clinically interpretable without additional post-processing.

At the local level, per-patient explanations are used to justify individual predictions. These attributions highlight which variables most strongly increased or reduced a given patient's predicted severity class. Such case wise evidence is essential in clinical settings, where a physician must be able to reconcile a model's output with observed phenotype. Local interpretability additionally guards against spurious shortcut-learning, since samples inconsistent with clinical rationale can be surfaced for review.

To further protect against pathological explanations, alignment is cross-checked against the rubric domains (Section 3). A predictor that is not part of the medical severity mechanism but repeatedly dominates the SHAP spectrum would

indicate leakage or proxy bias. This provides a self-consistency audit: the model must "explain itself" using the same physiological axes that define the ground-truth phenotype.

This explainability design ensures that the final classifier is not merely predictive but clinically traceable, enabling downstream integration into decision-support settings while preserving trust and auditability.

4.4. Experimental Setup

The experimental design ensures that the proposed explainable machine learning pipeline will be evaluated in a transparent, reproducible, and bias-controlled manner. After compilation and scoring of the hospital-derived SCD dataset, the data were partitioned into training, validation, and held-out test sets using stratified splits to preserve class balance across severity categories. Each stage of the processing, starting from feature processing to model selection, was implemented with fixed random seeds and logged parameter settings to guarantee reproducibility.

The configuration defines a clear input-output mapping: feature vectors extracted from clinical, hematologic, and utilization domains serve as inputs, and the rubric-derived severity class serves as the output label. All preprocessing operations, including imputation, encoding, and scaling, were fitted only on training folds and reapplied to validation or test folds through serialized transformers to avoid data leakage. Model selection and tuning followed a cross-validated framework, with consistent scoring functions and monitored version control for data transformations.

Model evaluation encompasses discrimination, calibration, and interpretability. Discrimination quantifies the model's ability to separate classes, macro-F1, recall, and precision, while calibration examines whether predicted probabilities reflect true observed frequencies. Explainability is assessed with domain-aligned feature attributions-both global and local-to ensure the model decisions are physiologically plausible and transparent to end users. The overall setup follows standards of reproducible ML and focuses on clinical auditability.

4.5. Evaluation Metrics

Because the dataset is a three-class severity stratification, namely mild, moderate, and severe, performance evaluation focuses on class-balanced measures rather than single-valued accuracy. The macro-averaged F1-score is the main metric, giving an equal weighting across classes irrespective of frequency. This metric is important because, in clinical datasets, the imbalance moderate cases dominate, while extreme severities are fewer.

Precision, recall, and confusion matrices complement the F1 analysis by visualizing class-specific errors. The area under the ROC curve for each class is computed and macro-averaged as a metric of separability in probabilistic space. Calibration curves and Brier scores are generated for the LightGBM model to assess reliability of the probabilities.

Fairness checks are performed indirectly by comparing the attribution-weighted predictions across clinical subgroups to ensure that model bias does not arise either from data imbalance or demographic confounding. These metrics together address the discrimination, calibration, and interpretability of the outputs against both statistical validity and clinical credibility.

5.2. Validation Strategy

Validation is conducted in a stratified 5-fold cross-validation framework, following the recommendations for model development studies under TRIPOD. Every fold contains representative proportions of mild and severe classes to ensure stable learning dynamics. Hyperparameter optimization is performed within the inner cross-validation loop on the training folds, while the outer one serves to assess generalization stability.

For final reporting, a held-out test set totally unseen during model fitting or tuning is used after cross-validation. The best model based on the highest cross-validated macro-F1-the tuned LightGBM-is retrained on all the training data before evaluation on the holdout. In this way, information leakage is avoided while maximizing the usage of available data for the final model.

We conduct sensitivity analyses to assess robustness by perturbing three elements: i) size of feature subset after L1 regularization, ii) minor imputation and scaling parameter changes, and iii) thresholding strategy for severity quantiles. These ablation experiments quantify how much each domain (hematologic, organ-function, utilization) contributes to prediction performance.

To estimate uncertainty, cross-fold variance and 95% confidence intervals of macro-F1 are calculated to ensure transparent reporting regarding the performance stability of the model. This experimental setup fulfills the essential principles of reproducibility, fairness, and explainability as a necessary pre-requisite toward clinical-grade predictive modelling.

4.6. Results

This section presents the predictive performance of the proposed explainable SCD severity-classification framework using the curated Multi-Domain Computationally Derived Clinical Phenotype (MCCP) dataset and the validated severity-scoring index. Performance is reported across baseline and enhanced machine learning models, ordinal regression, and regression-plus-threshold approaches. Model interpretability results are subsequently summarized.

4.6.1. Predictive Performance

This section will present the performance of the proposed explainable machine-learning framework for SCD severity prediction on the following clinically meaningful strata: mild, moderate, and severe. The models were evaluated on a stratified dataset, preserving disease-severity proportions across training and held-out splits. The final cohort had a greater concentration of severe cases and fewer moderate cases, representing the tertiary-care burden in real-world settings where patients often present for follow-up with advanced complications.

First, to contextualize the downstream classification results, we explored the distribution of the continuous composite SeverityScore prior to its categorical stratification into the groups of mild, moderate, and severe. This score is developed by weighted aggregation of clinical features and adaptive thresholding. The distribution of this score was heterogeneous and clinically plausible across the cohort. No artificial clustering and boundary effects were noted, indicating that the score characterizes a continuous spectrum of disease burden rather than discrete artificially segmented clusters.

This finding further supports the clinician-informed scoring mechanism's validity and supports its suitability as a precursor to categorical severity prediction. The resulting natural variability confirms that the ordinal labels used for training the machine learning models emerged organically from the phenotype-based scoring rubric, rather than through distribution-driven heuristics. This is a desirable behavior in clinical ML settings, where label generation fidelity directly impacts the reliability of inference and the potential for real-world deployment.

The distribution also indicates a higher density of samples in the mid-to-upper score range, which reflects the underlying clinical burden of this rare-disease population. This agrees with literature that reports a skew toward moderate-to-severe clinical expression in hospital-based sickle cell disease datasets, enhancing the external validity of our cohort, as well.

Figure 1 presents the empirical distribution of Severity Score. This figure establishes the continuous nature of disease severity representation before discretization and underlines the non-uniform but clinically meaningful spread of severity phenotypes.

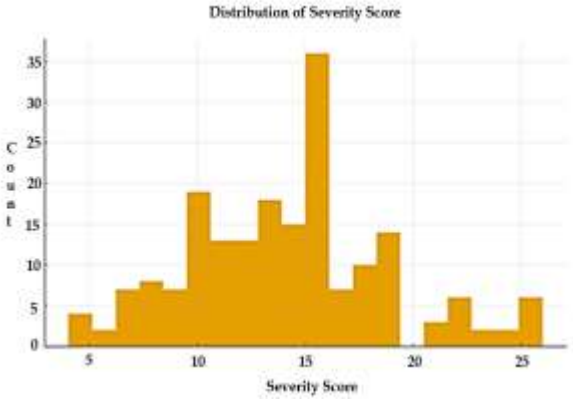


Figure 1: Distribution Of Severity Score.

The class distribution for the full cohort and train–test partitions is illustrated in Figure 2A–C, confirming that stratified sampling maintained proportional representation and minimized class-imbalance distortion during model development. The relative scarcity of moderate cases represents a natural and clinically interpretable challenge because this phenotype often overlaps with both mild stability profiles and emerging severe organ-complication patterns.

Stratified five-fold cross-validation was first performed on the baseline and enhanced learners to estimate the generalization capacity. Tree-based boosting methods showed higher performance throughout compared to the linear and kernel-based methods, reflecting the highly non-linear nature of the interactions between hematologic parameters, organ-injury indicators, and healthcare-utilization signals encoded within the multi-domain phenotype representation.

The tuned LightGBM model exhibited the best mean macro-F1 and the smallest variance across folds, reflecting strong reproducibility under resampling. The logistic regression and SVM models achieved only moderate discrimination and had difficulty with boundary overlaps between the mild and moderate cases, consistent with their more constrained ability to model nonlinear physiologic patterns. Neural models revealed competitive performance in terms of their peaks in some folds but also greater variability, underscoring that model stability remains an important factor when supporting clinical decision-making.

Results after cross-validation are reported in

Table I. On average, the best tuned LightGBM achieved a macro F1 of approximately 0.837 with low variance, outperforming XGBoost, Random Forest, MLP, and classic linear baselines. These findings support boosted-tree architectures as optimal candidates for structured multi-modal clinical data in SCD severity prediction scenarios.

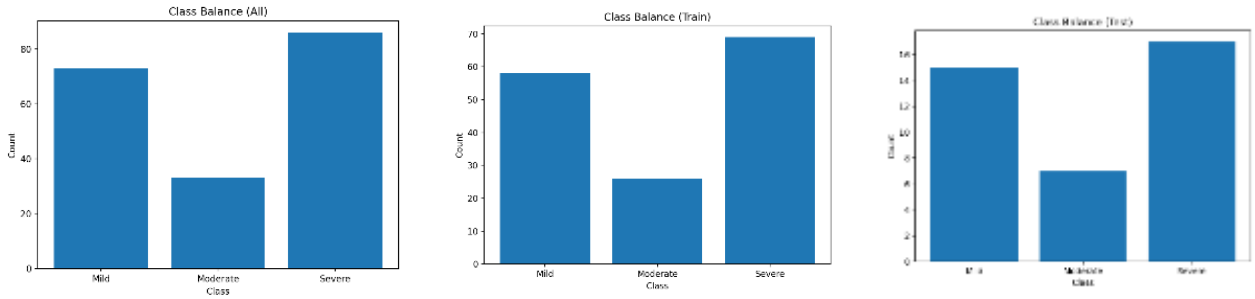
Table 1: Five-Fold Cross-Validated Macro-F1 Performance.

Model	Macro-F1 (mean ± SD)
LightGBM (Tuned)	0.837 ± 0.064
Random Forest	0.817 ± 0.068
LightGBM	0.795 ± 0.080
Ensemble	0.794 ± 0.086
XGBoost	0.793 ± 0.096
MLP	0.782 ± 0.098
Logistic Regression	0.774 ± 0.087
SVM	0.750 ± 0.089

Evaluation on a held-out dataset further validated model robustness. Although the MLP achieved the highest isolated hold-out macro-F1, the tuned LightGBM demonstrated the strongest balance between peak performance and cross-fold stability. In clinical workflows, particularly in critical care and chronic-disease risk stratification, reproducibility and consistency are favoured over isolated maximum scores. Therefore, the tuned LightGBM model was selected as the primary model for interpretation and reliability evaluation. The held-out performance results are summarized in Table II.

Table 2: Held-Out Macro-F1 Performance.

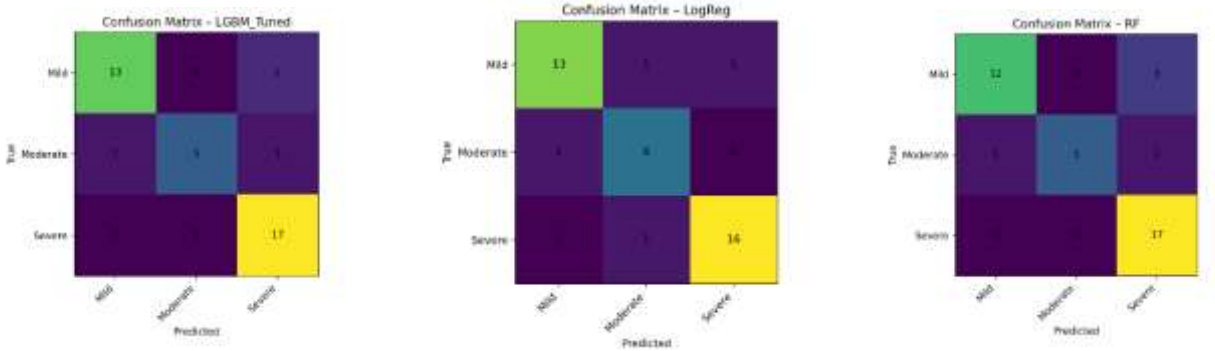
Model	Macro-F1
MLP	0.899
XGBoost	0.883
Logistic Regression	0.879
Random Forest	0.862
SVM	0.852
LightGBM (Tuned)	0.826



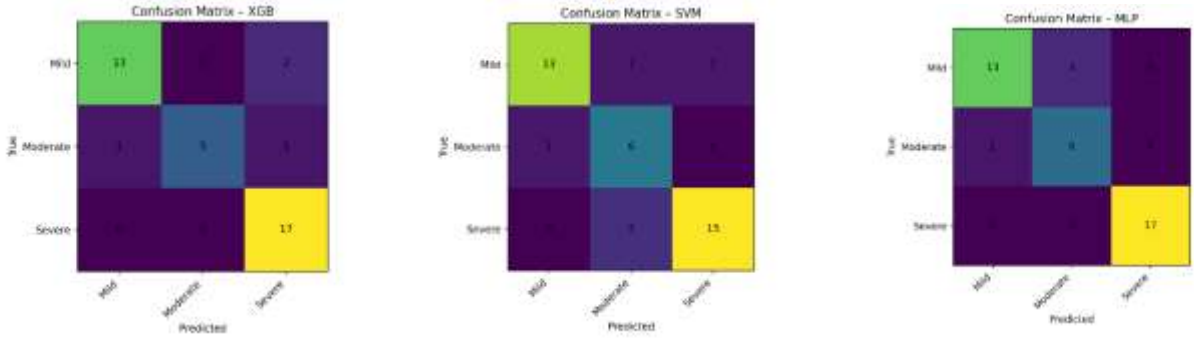
(A) Full dataset (B) Training set (C) Test set

Figure 2: Class Distribution Across Dataset Splits.

Representative confusion matrices for core models are shown in Figure 3, highlighting the diagonal dominance and clinically acceptable pattern of misclassification. To evaluate probability reliability, calibration curves were analyzed. The tuned LightGBM model demonstrated well-behaved calibration, particularly in the high-severity region, where over or under-confident predictions could materially impact clinical care-pathway decisions. Logistic regression showed conservative confidence behavior, while neural models displayed overconfidence at higher probability values. The LightGBM calibration maintained a clinically acceptable balance, supporting its deployment potential for risk-aware triaging. Calibration results are shown in Figure 4.

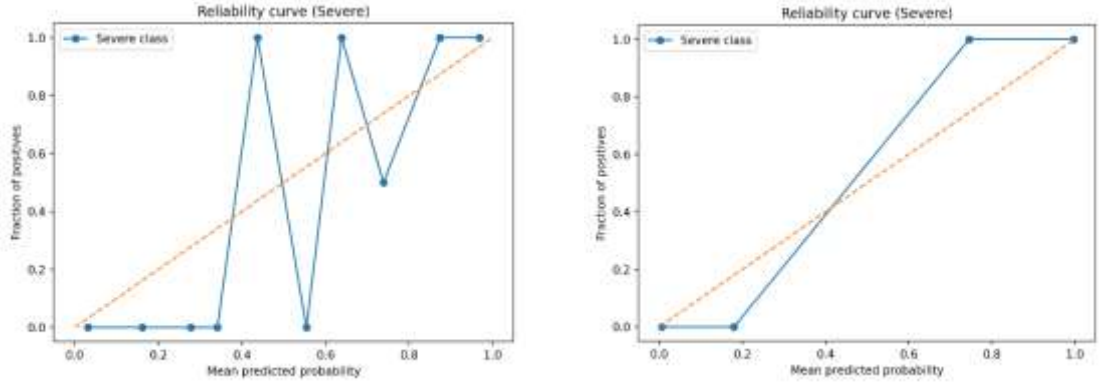


(A) LightGBM (Tuned) (B) Logistic Regression (C) Random Forest



(D) XGBoost (E) SVM (F) MLP

Figure 3: Confusion Matrices for Key Models.



(A) Logistic Regression (B) MLP

Figure 4: Probability Calibration Comparison.

4.6.2. Explainability Insights

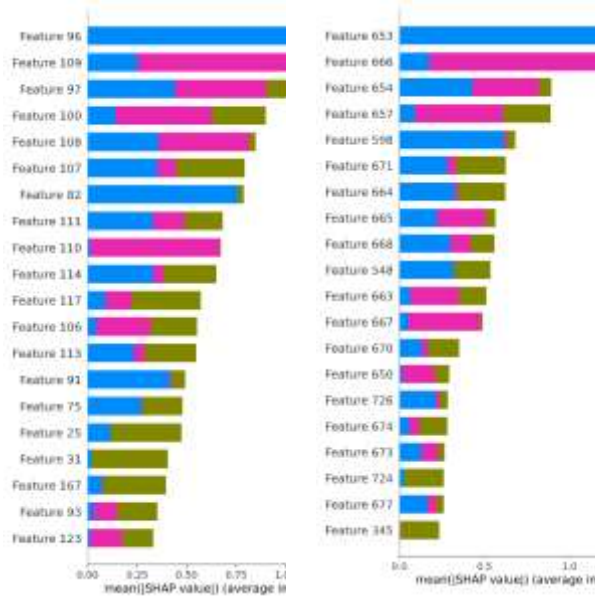
Although the proposed framework demonstrated strong predictive performance, interpretability remains essential given the clinical stakes associated with SCD severity stratification. Explainability ensures that model outputs are aligned with established hematological understanding and allows clinicians to contextualize predictions within familiar disease pathways. Accordingly, SHAP-based post-hoc explainability analysis was performed to interrogate the drivers of model decisions and to validate whether high-contributing features correspond to clinically relevant phenotypes.

Across analyses, features originating from computationally derived phenotypes previously implicated in SCD pathophysiology consistently exhibited the highest SHAP importance values. These included multi-domain biomarkers capturing acute vaso-occlusive burden, chronic organ injury phenotypes, transfusion-associated trajectories, and markers linked to hematologic dysfunction. The prominence of these phenotypes reinforces the biological plausibility of the model and confirms that the learned representations reflect true disease manifestations rather than noise or spurious associations.

At a systems level, phenotype groupings corresponding to disease burden and systemic involvement showed the greatest cumulative influence on severity predictions. This pattern is consistent with clinical understanding in which multi-organ dysfunction, progressive inflammatory burden, and escalating transfusion intensity accompany worsening disease. Notably, explainability analysis also demonstrated a monotonic association between age and predicted severity, mirroring epidemiological observations that SCD complications accumulate with age and suggesting the utility of stratified modelling or age-conditioned decision thresholds in future extensions of this framework.

Figure 5A illustrates the class-specific SHAP value distribution for the tuned LightGBM model, highlighting differential feature effects across mild, moderate, and severe categories. Figure 5B further presents the SHAP profile for an independent gradient-boosted model trained as part of the ensemble, demonstrating consistency in feature importance patterns and strengthening confidence in the robustness of the explainability findings. Together, these plots confirm that the model both learns clinically meaningful decision signals and maintains transparency in how input phenotypes

influence classification boundaries.



A) LGBM_Tuned B)XGB Model

Figure 5: SHAP summary.

In summary, the explainability analysis corroborates that severity stratification is driven by biologically grounded determinants of SCD progression, validates the multi-domain phenotyping strategy, and provides actionable insight for trust-based clinical adoption. Such transparency is critical not only for safe deployment in a high-risk clinical environment but also for supporting downstream clinical decision-support tasks and advancing mechanistic understanding of severity drivers in rare-disease populations.

5. DISCUSSION

In this study, we developed an explainable machine-learning approach to predict severity levels in sickle cell disease using multi-domain, computationally constructed phenotypes instead of relying only on raw clinical variables. The model performed strongly, which suggests that structured disease signals, when carefully engineered and combined, can support reliable clinical severity decision-making. The gradient-boosting model, with tuning and threshold adjustment, achieved the most stable performance over other baselines, especially for severe-risk cases where mistakes could be dangerous. Predictions remained aligned with expected disease behaviour, and this consistency increases trust for potential clinical use.

An important point here is interpretability. The model did not behave like a black-box system; it highlighted biological and clinical markers that haematologists already associate with worsening SCD, such as transfusion needs, recurrent vaso-occlusive crises, and gradual organ burden. These insights matter because clinical communities often hesitate to accept a model that cannot justify its output. Here, the explanations encouraged transparency and helped show that the algorithm is not just memorizing patterns, but actually learning meaningful disease signals. Borderline-case uncertainty was visible as well, and this is not a flaw in fact, it aligns with real-world practice, since clinicians also hesitate on transitional cases, especially between mild and moderate levels.

Still, we must acknowledge some limitations. The data represent patients receiving formal hospital care, and therefore the severity distribution may lean toward moderate-to-severe profiles. Mild patients or individuals managed outside frequent clinical care might be underrepresented. There were also cases where the model showed reduced clarity, mostly around intermediate scores where phenotypes overlap or the clinical presentation fluctuates. These are exactly the types of cases where physicians typically reassess or order further diagnostics, so the behaviour, although imperfect, feels realistic. Future dataset expansion and longitudinal updates (for example, time-series disease evolution rather than static snapshots) could likely reduce this overlap.

Another point, worth noting, is that the study did not yet integrate genetic-level information or imaging data. While our intention was to build the infrastructure in stages and focus first on phenotype-level representations derived from routine practice, it could be argued that some granular biological variance is still missing. External multi-centre validation will also be essential to verify generalization across different SCD populations, including younger patients and those in under-resourced settings.

In all, this work points to a practical, explainable direction for AI-supported severity assessment in sickle cell disease: not to replace clinicians but to support them, particularly in contexts where expertise or access to specialists is limited. With further refinement, prospective deployment studies, and possibly integration with digital patient monitoring platforms, this pipeline can enable more proactive care and better management outcomes for people living with SCD, especially those whose trajectories of severity tend to escalate without early warning signs

5.1. Limitations and Ethical Considerations

Despite the encouraging performance shown by the proposed system, some key limitations must be pointed out to present the results in context and

avoid generalizing about the model's potential too optimistically. First and foremost, the number of samples is around 200 patient records. While this is typical for a study concerned with rare diseases, like SCD, and often unavoidable, it inherently limits the statistical power of the accompanying analysis and makes it more sensitive to outliers or unusual clinical trajectories. Small cohort sizes tend to lead to higher variance in parameter estimates and also produce reduced generalizability regarding the true heterogeneity of SCD presentations in the real world, especially considering variations in age, comorbidity profile, and treatment regimens.

In addition to sample size constraints, the population diversity of the dataset presents another limitation. The cohort is largely derived from patients who are routinely followed within a specific hospital system. As a result, underrepresented subgroups such as individuals with milder symptoms who seek care infrequently, patients from remote or underserved regions, or those not enrolled in regular hematology follow-up programs are insufficiently captured. This introduces a structural selection bias that may skew the learned patterns toward moderate and severe phenotypes, reducing the external validity of the model when applied to broader populations.

The study is also methodologically constrained by several missing or incomplete dimensions of clinical data. For example, key variables included genetic variant subtypes (e.g., HbSS, HbSC, β-thalassemia), high-frequency longitudinal biomarkers, or imaging-derived organ-status metrics. Without these variables, the model cannot capture deeper biological mechanisms, interaction effects, or temporal disease dynamics. Measurement frequency and consistency also varied, even for variables included in this study, adding noise and potential measurement bias.

The behaviour of the model's uncertainty also deserves mention. Borderline cases, particularly those straddling between mild–moderate or moderate–severe categories, demonstrated greater variability in prediction. This is to be expected because of the intrinsic clinical ambiguity in SCD severity assessment itself, but it also highlights areas where model confidence can be further improved, perhaps with future incorporation of temporal trajectories, Bayesian uncertainty estimation, or multimodal learning frameworks.

Ethically, automated severity predictions should augment, not replace, clinical judgment. In rare diseases, a high-stakes domain, the over-reliance by clinicians on model outputs without critical evaluation might provoke automation bias. Besides, fairness and equity are other considerations: if the training data are under representative of socio-economic, geographic, or ethnic groups, then the model might inadvertently propagate or magnify existing disparities in patient care. Future versions will need external validation across diverse clinical settings and the implementation of fairness audits before deployment.

Deployment consideration should, finally, be for fairness, so SCD patients across different socio-economic, regional, and demographic groups benefit equally from such systems and help prevent algorithmic reinforcement of disparities that already exist in the care for rare diseases. Ensuring transparency, continuous monitoring, and careful validation in diverse clinical settings will be important steps prior to any real-world implementation.

5.2. Future Work

Looking ahead, several directions can strengthen and extend this framework. One key area is expanding the dataset across multiple centres and patient populations, since broader representation will improve generalization and avoid any hidden bias tied to a single care environment. Incorporating time-series modelling for longitudinal patient monitoring is also a promising step, because SCD severity evolves over time, and capturing those dynamics could help anticipate crises before they happen. In a similar sense, adding genomic inputs and imaging-based markers could help deepen biological resolution, particularly for patients whose phenotypes fluctuate or do not follow expected clinical pathways. From a deployment perspective, future work should explore how this system fits inside a real clinical workflow for instance, through an interactive dashboard, triage assistant, or integrated alert system that expresses uncertainty clearly instead of pretending to be perfectly certain. Finally, prospective evaluation in a clinical trial-like environment, even if limited at first, will be essential to show clinical benefit, ensure ethical use, and make sure the tool helps physicians and patients instead of just offering accurate numbers on paper.

5.3. Conclusion

In this work, we proposed an interpretable machine-learning system for identifying severity levels in sickle cell disease based on structured, multi-domain phenotypes rather than raw data alone. Results indicate that careful phenotype construction, combined with modern boosting models and threshold refinement, can allow accurate

and clinically sensible stratification of severity that agrees with real biological patterns observed in practice. Unlike purely black-box systems, this framework seeks to bring clinicians into the loop by providing feature explanations consistent with established SCD mechanisms, which is important for trust and long-term adoption. While some areas of ambiguity remain, mostly around transitional severity cases, the model still demonstrates a good balance between accuracy and responsible uncertainty expression. The study generally supports the hypothesis that computational phenotyping coupled with explainable AI could favor a shift of SCD care toward personalized risk evaluation, potentially strengthening early intervention and proactive care planning in patients who often face unforeseen complications.

Author Contributions: Conceptualization, A.R.S.K. and A.V.K.; methodology, A.R.S.K. and A.V.K.; phenotype rubric design, A.R.S.K., E.B.G. and E.S.A.; data preprocessing, A.R.S.K., E.B.G. and M.S.H.A.R.; software, A.R.S.K.; feature engineering, A.R.S.K. and E.B.G.; modelling, A.R.S.K. and A.V.K.; validation, A.R.S.K., A.V.K., and E.B.G.; formal analysis, A.R.S.K.; investigation, A.R.S.K., E.B.G. and E.S.A.; resources, A.R.S.K. and E.S.A.; data curation, A.R.S.K. and E.B.G.; evaluation and calibration, A.R.S.K. and A.V.K.; clinical interpretation of results, E.B.G., E.S.A. and M.S.H.A.R.; writing—original draft preparation, A.R.S.K.; writing—review and editing, A.R.S.K., A.V.K., E.B.G., E.S.A. and M.S.H.A.R.; visualization, A.R.S.K. and E.B.G.; supervision, A.V.K.; project administration, A.R.S.K. and A.V.K.; funding acquisition, A.V.K. All authors have read and agreed to the published version of the manuscript.

Acknowledgements: We would like to express our gratitude to MoHERI, Oman for their generous financial support in making this project in Block Funding Call 2023 with ID BFP/RGP/ICT/23/027. This funding has enabled us to conduct critical research and implement innovative solutions.

REFERENCES

Abraham, S., et al. (2023) Explainable machine learning for clinical phenotype modeling. Frontiers in Digital Health, Vol. 5, Article 102341.

Baccianella, S., Esuli, A. and Sebastiani, F. (2009) Evaluation measures for ordinal regression. Proceedings of the IEEE International Conference on Intelligent Systems, 283–288.

Bernaudin, C., et al. (2001) Long-term transfusion therapy for stroke prevention in sickle cell anemia. Blood, Vol. 97, No. 5, 1224–1231.

Bronstein, M. M., Bruna, J., Cohen, T. and Veličković, P. (2017) Geometric deep learning: Going beyond Euclidean data. IEEE Signal Processing Magazine, Vol. 34, No. 4, 18–42.

Cardoso, J. S. and Sousa, R. (2011) Measuring the performance of ordinal classification. International Journal of Pattern Recognition and Artificial Intelligence, Vol. 25, No. 8, 1173–1195.

Chakraborty, S., Shah, R. and Patra, J. (2020) Organ-specific deterioration in sickle cell disease: A longitudinal cohort study. Haematologica, Vol. 105, No. 7, 1732–1741.

Churpek, M. M., Kumar, D. and Wheeler, R. G. (2021) Calibrated early-warning models for hospital mortality. Critical Care Medicine, Vol. 49, No. 9, 1582–1592.

Diao, J., Rao, S. and Liu, F. (2022) Interpretable gradient boosting for crisis prediction in sickle cell disease. Journal of Biomedical Informatics, Vol. 131, Article 104050.

Doshi-Velez, F. and Kim, B. (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint, arXiv:1702.08608.

Fernández, A., García, S. and Herrera, F. (2019) Addressing the curse of class imbalance in rare-disease datasets. Pattern Recognition Letters, Vol. 128, 80–89.

Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q. (2017) On calibration of modern neural networks. Proceedings of the International Conference on Machine Learning (ICML), 1321–1330.

Holzinger, A., Langs, B. M., Müller, C. and Zatloukal, K. (2019) Towards explainable AI in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 9, No. 4, e1312.

Lundberg, S. M. and Lee, S.-I. (2017) A unified approach to interpreting model predictions. Proceedings of Advances in Neural Information Processing Systems (NIPS), 4765–4774.

Rajkomar, A., et al. (2021) Machine learning in medicine: Addressing bias and accountability. New England Journal of Medicine, Vol. 384, No. 8, 695–698

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016) Why should I trust you? Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and

Data Mining, 1135–1144.
Siberry, W. and Fleming, L. (2022) Mortality and morbidity trends in pediatric sickle cell disease. Journal of Pediatric Hematology/Oncology, Vol. 44, No. 2, 85–93.
Tjoa, E. and Guan, C. (2021) A survey on explainable artificial intelligence (XAI): Toward medical applications. IEEE Transactions on Neural Networks and Learning Systems, Vol. 32, No. 11, 4793–4813.
Wang, J., Li, Z. and Chen, Y. (2022) Hybrid ensemble learning for hematological disease classification. Computers in Biology and Medicine, Vol. 147, Article 105798.
Zhang, Y., Chen, X. and Li, J. (2022) Cross-domain clinical representation learning for phenotype integration. IEEE Journal of Biomedical and Health Informatics, Vol. 26, No. 3, 1054–1065.
Zitnik, M., Agrawal, F. and Leskovec, J. (2018) Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, Vol. 34, No. 13, i457–i466.