Published by Nigerian Society of Physical Sciences. Hosted by FLAYOO Publishing House LTD

Proceedings of the Nigerian Society of Physical Sciences

Proceedings Homepage: flayoophl.com/journals/index.php/pnspsc

Machine learning-based prediction of climate-driven sickle cell anemia crises across Nigerian regions using synthetic multimodal data

Mary Ofuru Kama[a,b,e,*], Ifeyinwa Angela Ajah[b], Sylvester Agbo Igwe[b], Anayo Chukwu Ikegwu[a,c], Chinatu Michael Anyanwu[d,e]

[a]*Department of Software Engineering, Veritas University, Abuja, Nigeria*
[b]*Department of Computer Science, Ebonyi State University, Abakaliki, Nigeria*
[c]*Department of Computer Science, Federal University of Petroleum Resources, Effurun, Nigeria*
[d]*Department of Computer Science, Maduka University, Ekwegbe, Enugu State, Nigeria*
[e]*High Performance and Intelligence Computing Laboratory, University of Nigeria, Nsukka, Nigeria*

ABSTRACT

Sickle cell anemia (SCA) is a genetic hemoglobinopathy that imposes a major health burden in sub-Saharan Africa, particularly in Nigeria. Vaso-occlusive crises (VOCs), the principal acute manifestations of SCA, may be influenced by environmental factors such as temperature, low humidity, harmattan dust, and rainfall. Although associations between VOCs and environmental conditions have been reported, prediction models remain difficult to develop for data-constrained clinical settings. We developed a proof-of-concept VOC prediction framework for six Nigerian ecological regions using a synthetic multimodal health–environment dataset comprising 12,000 patient-month observations. Random Forest, support vector machine (SVM), XGBoost, neural network, and logistic regression models were evaluated. Logistic regression achieved the best overall performance (accuracy: 0.78; F1-score: 0.74; area under the receiver operating characteristic curve [ROC-AUC]: 0.84), followed by Random Forest (0.77, 0.72, and 0.83), SVM (0.78, 0.73, and 0.82), neural network (0.77, 0.72, and 0.82), and XGBoost (0.76, 0.70, and 0.82). Regional analysis showed a north–south gradient in simulated crisis occurrence: the North-East had the highest probability (80.1%), associated with temperature and particulate matter with aerodynamic diameter below 2.5 μm ($PM_{2.5}$), whereas the South-South had the lowest probability (10.4%). Temperature, $PM_{2.5}$, humidity, and previous hospitalization were the most influential predictors. These findings establish the methodological feasibility of a climate-sensitive prediction pipeline that should be externally validated with clinical and meteorological data before use in health-surveillance applications.

Keywords: Sickle cell anemia, Machine learning, Vaso-occlusive crisis, Synthetic multimodal data, Climate science.
DOI: 10.61298/pnspsc.2026.3.337

© 2026 The Author(s). Production and Hosting by FLAYOO Publishing House LTD on Behalf of the Nigerian Society of Physical Sciences (NSPS). Peer review under the responsibility of NSPS. This is an open access article under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

1. INTRODUCTION

Individuals living with sickle cell anemia (SCA) experience recurrent painful episodes known as vaso-occlusive crises (VOCs), which occur when sickled erythrocytes obstruct blood vessels and cause tissue ischemia, severe pain, organ injury, and increased mortality [1]. VOCs are sensitive to environmental and climatic conditions. Temperature, humidity, rainfall, and atmospheric dust have all been associated with crisis occurrence [2]. High temperatures may promote dehydration and increase blood viscosity, whereas cold exposure may induce vasocon-

*Corresponding author Tel. No.: +234-906-935-9954.
e-mail: kamam@veritas.edu.ng (Mary Ofuru Kama)

striction. During the West African harmattan season, dry and dusty winds may aggravate respiratory and physiological stress. A nationwide analysis also reported higher admissions for VOC and acute chest syndrome during colder periods [3].

SCA remains a substantial global and African health burden [4]. Global prevalence increased by an estimated 41.0% between 2000 and 2021 [5]. Nigeria contributes a large proportion of affected births, and a recent national meta-analysis estimated SCA and sickle cell trait prevalences of 4.0% and 21.0%, respectively, among Nigerian children and adolescents [6]. A continent-wide meta-analysis reported an overall HbSS prevalence of 1.43% in Africa [7], while burden estimates indicate that sub-Saharan Africa and South Asia account for most of the global disease burden [8]. The clinical burden is accompanied by substantial household expenditure in Nigeria [9], Uganda, and Malawi [10]; evidence from northwestern Nigeria likewise highlights the socioeconomic burden on affected families [11].

Recent work has applied artificial intelligence (AI) to point-of-care diagnosis [12], hematological severity classification [13], digitally monitored VOC prediction [14], explainable disease detection [15], and automated erythrocyte morphology analysis [16]. These approaches demonstrate the value of AI in sickle cell research, but they generally emphasize diagnostic, physiological, hematological, or imaging data rather than region-specific environmental exposure. Systematic reviews have reported promising results for multilayer perceptrons, support vector machines, Random Forest, logistic regression, and related algorithms [17]; deep-learning systems have also been used for sickle cell classification [18]. Machine-learning models can provide early warning information by identifying nonlinear relationships among heterogeneous variables, although clinically useful implementation depends on suitable data and external validation [19].

Data scarcity remains a major obstacle to climate-sensitive predictive modeling in Nigeria. Many health facilities lack comprehensive electronic records, and available datasets may be fragmented, geographically restricted, or inaccessible because of privacy constraints. Synthetic-data approaches can partly address these limitations. A corrected bibliographic source demonstrates the use of synthetic electronic health records for predictive modeling in low- and middle-income countries [20]; broader reviews describe open-source synthetic-data methods [21] and privacy-preserving clinical risk prediction [22]. Limited access to health services further complicates disease management in Nigeria [23].

Nigeria's ecological diversity adds an important spatial dimension. The country extends from Sahel and Sudan savannah zones in the north to humid rainforest and coastal zones in the south. Temperature, rainfall, humidity, and dust exposure vary accordingly, suggesting that climate-associated crisis risk may also vary by region. This study therefore develops a regional machine-learning framework for predicting climate-driven SCA crises from synthetic health–climate data representing six Nigerian geopolitical regions. The study is explicitly framed as a proof-of-concept simulation and modeling exercise rather than a clinically validated prediction tool. Its contributions are: (i) a synthetic multimodal dataset integrating regional climate and clinical variables; (ii) a comparative evaluation of five machine-learning algorithms; and (iii) a reproducible pipeline for stress-testing climate-informed crisis prediction under data scarcity.

1.1. RESEARCH GAP

Existing studies show substantial progress in automated diagnosis, biomarker-based severity grading, explainable AI, and digital health monitoring. Nevertheless, environmental determinants of VOCs remain poorly represented in current predictive models. To our knowledge, no prior study has integrated synthetic climate, demographic, and clinical variables to model regional VOC risk across Nigeria's distinct ecological zones. The present work addresses this methodological gap.

2. BACKGROUND AND RELATED WORK

2.1. TRIGGERS OF SICKLE CELL CRISES

SCA is caused by a single amino-acid substitution in the β-globin gene, resulting in abnormal hemoglobin S (HbS). Under hypoxic, dehydrating, or physiological stress, HbS polymerizes and deforms erythrocytes. The resulting rigid cells obstruct the microvasculature, initiating vaso-occlusion and tissue ischemia. Established VOC triggers include hypoxia, dehydration, infection, fever, temperature extremes, acidosis, and physical stress [24]. The pathogenesis and clinical expression of VOCs remain complex, particularly in resource-limited settings [25].

Environmental factors constitute an important but sometimes overlooked group of triggers. Low temperature and high wind exposure may increase VOC occurrence, while humidity has been associated with monthly pain scores [26]. High temperatures can cause dehydration and hyperosmolality, thereby increasing blood viscosity and sickling. Harmattan conditions combine cold, dry, and dusty winds and may create additional risk in West Africa. Genetic modifiers also influence the frequency and severity of VOCs [27].

Inflammation further complicates VOC pathogenesis. Activated endothelial cells, sickled erythrocytes, neutrophils, and platelets create an adhesive and inflammatory vascular environment that promotes further occlusion [28]. Environmental stressors may amplify these interactions, reinforcing the need to characterize the timing and distribution of climatic triggers.

2.2. CLIMATE–HEALTH RELATIONSHIPS IN SUB-SAHARAN AFRICA

The health effects of climate change have been documented across Africa, where vulnerable health systems may have limited adaptive capacity [29]. Nigeria contains six broad ecological zones: Sahel savannah, Sudan savannah, Guinea savannah, derived savannah, rainforest, and mangrove/coastal zones. Seasonal temperature, rainfall, humidity, and dust exposure vary markedly among these zones. Northern states such as Kano, Borno, and Sokoto experience extreme dry-season heat and harmattan dust, whereas Lagos, Rivers, and other southern states experience high humidity and rainfall. Observed rainfall and temperature trends in northeastern Nigeria reinforce the importance of regional climate analysis [30]. Nigeria is also experiencing climate-change-induced heatwave risk [31], which may be relevant to conditions that are sensitive to dehydration and air quality.

2.3. EXISTING PREDICTION MODELS

Earlier predictive studies relied on clinical variables such as hemoglobin concentration, white-cell count, and admission history. More recent machine-learning models have predicted acute kidney injury, organ failure, and transplantation outcomes in sickle cell disease. Acute organ failure has been modeled from intensive-care data [32], and registry-based machine learning has been used to predict hematopoietic cell transplantation outcomes [33]. Deep neural networks and explainable AI have also been applied to sickle cell image classification [34]. However, few models explicitly combine regional climatic exposure with patient-level crisis risk in an African setting.

2.4. MACHINE LEARNING AND SYNTHETIC DATA IN PUBLIC HEALTH

Machine learning supports public-health prediction and surveillance by modeling nonlinear relationships in heterogeneous data. Applications include infectious-disease forecasting, maternal-risk stratification, chronic-disease monitoring, and environmental-health surveillance. Synthetic clinical data can facilitate collaboration while reducing re-identification risk [35], but models trained in high-income settings may not transfer directly to low- and middle-income settings because disease presentation, healthcare use, and data infrastructure differ.

3. STUDY AREA AND REGIONAL CONTEXT

3.1. OVERVIEW OF NIGERIA

Nigeria occupies approximately $923{,}768\,\mathrm{km}^2$ in West Africa and comprises 36 states and the Federal Capital Territory. It is bordered by the Gulf of Guinea, Benin, Niger, Chad, and Cameroon. Nigeria reports a high burden of SCA [36], and stroke detection and prevention remain important components of disease management [37]. The combination of a large affected population, climatic diversity, rapid population growth, and uneven healthcare capacity makes the country a relevant and challenging setting for climate-sensitive early-warning research.

3.2. CLIMATE ZONES AND SELECTED REGIONS

The Sahel and Sudan savannah zones in northern Nigeria receive low rainfall, experience high daytime temperatures, and are exposed to harmattan dust. Guinea savannah occupies much of the Middle Belt and receives moderate rainfall. Derived savannah and rainforest zones receive greater rainfall and humidity, while mangrove and freshwater zones in the Niger Delta receive the highest annual rainfall.

Six representative geopolitical regions were selected: North-West (Kano State), North-East (Borno State), North-Central (Plateau State), South-West (Lagos State), South-East (Enugu State), and South-South (Rivers State). These regions span the principal climatic gradients represented in the simulation.

3.3. POPULATION DISTRIBUTION OF SICKLE CELL DISEASE

Sickle cell trait and SCA occur across Nigeria's major ethnic groups and ecological zones. The environmental stressors affecting individuals with SCA may nevertheless differ by location. Northern patients may face greater harmattan and heat-related risk, whereas southern patients may experience prolonged humidity, rainfall, and infection-related stress. The present simulation examines these regional contrasts without claiming that its synthetic rates represent real-world prevalence.

Table 1. Features of the synthetic climate–health dataset.

Variable	Type	Description
Age	Numerical	Simulated patient age (years)
Gender	Categorical	Male/female
Crisis frequency	Numerical	Number of crises per year
Hospital visits	Numerical	Annual SCA-related hospital visits
Temperature	Numerical	Average regional temperature (°C)
Humidity	Numerical	Relative humidity (%)
Rainfall	Numerical	Monthly rainfall (mm)
Dust index	Numerical	Harmattan dust/$PM_{2.5}$ indicator
Season	Categorical	Dry, harmattan, or wet

4. METHODOLOGY

4.1. RATIONALE FOR SYNTHETIC DATA

Publicly accessible Nigerian datasets rarely combine patient-level clinical records with temporally matched climate variables. Institutional datasets may be difficult to access because of ethical requirements, fragmented health-information systems, privacy constraints, and limited sample sizes. Transfer learning was considered but not adopted because no suitable pretrained model was identified for integrated climate–clinical crisis prediction. The study therefore generated a synthetic multimodal dataset of 12,000 patient-month observations for proof-of-concept model development, not clinical deployment.

4.2. DATASET CONSTRUCTION AND VARIABLES

The simulation was informed by published epidemiological evidence, reported crisis triggers, regional climate patterns, and global health information. Variables were divided into two broad groups:

1. *Health variables*: age, sex, genotype, hydroxyurea use, baseline hemoglobin, previous hospitalizations, crisis frequency, pain severity, hospital visits, and binary crisis occurrence.
2. *Climate and temporal variables*: temperature, relative humidity, rainfall, $PM_{2.5}$, season, calendar month, region, and cyclical month encodings.

Continuous variables were assigned distributions consistent with their intended ranges. Age was sampled uniformly from 5–60 years. Temperature and hemoglobin concentration were modeled with Gaussian distributions subject to physical or clinical bounds. Crisis counts and hospital visits were generated using Poisson processes. Categorical distributions were used for genotype, sex, season, and treatment status. Monte Carlo simulation captured variation across patient–environment combinations.

Descriptive statistics were used to assess the internal distribution of the synthetic data. Table 2 summarizes the numerical variables, Tables 3 and 4 summarize categorical and regional distributions, and Figure 1 shows the correlation structure.

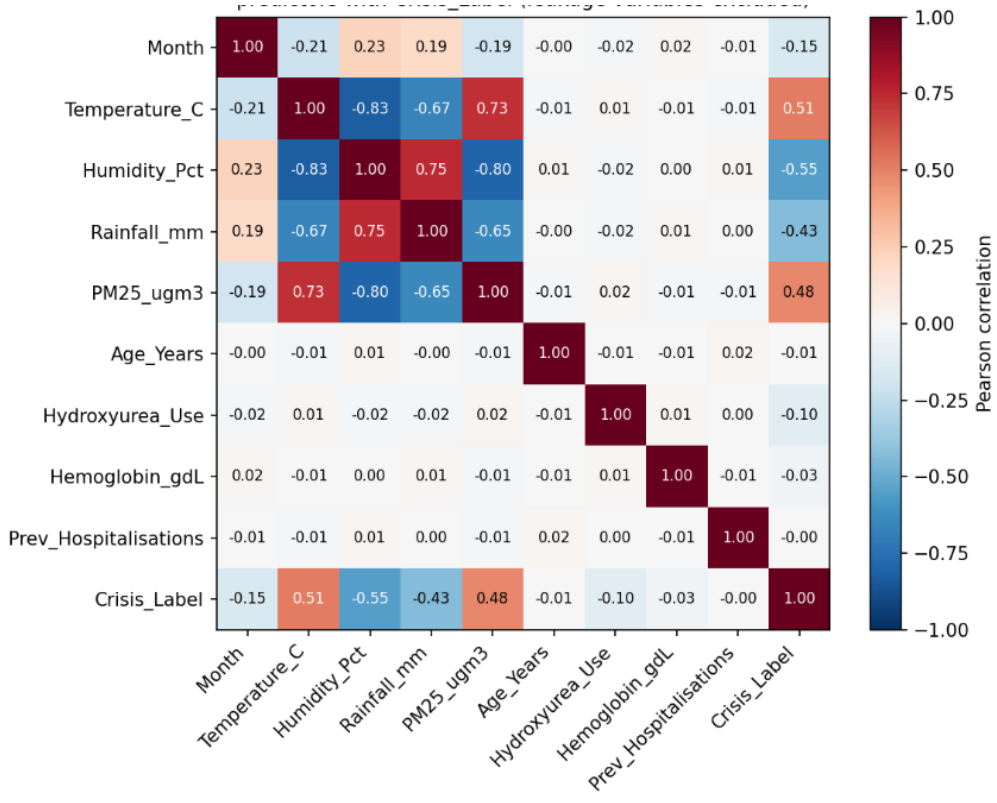


Figure 1. Correlation heatmap for the simulated environmental, temporal, and clinical variables.

Table 2. Numerical description of the synthetic dataset.

Variable	Mean	SD	Minimum	Maximum	Interpretation
Temperature (°C)	32.7	5.8	20	50	Higher in northern regions
Humidity (%)	54.8	29.1	5	100	Higher in southern regions
Rainfall (mm)	149.1	106.6	0.4	600	Peaks in the South-South
$PM_{2.5}$ ($\mu g\,m^{-3}$)	53.6	30.0	5	174.6	Higher in dry northern regions
Age (years)	32.7	16.2	5	60	Uniform distribution
Hemoglobin ($g\,dL^{-1}$)	7.5	1.2	4	12	Reflects anemia in SCA
Previous hospitalizations	2.0	1.4	0	10	Indicates prior health burden
Crisis frequency	1.6	1.8	0	15	Key count outcome
Pain-severity score	1.5	0.6	1	5	Linked to crisis severity
Hospital visits per quarter	1.0	1.5	0	12	Healthcare utilization
Crisis label	0.41	0.49	0	1	Binary classification target

Table 3. Categorical variables in the synthetic dataset.

Variable	Categories	Distribution
Region	Six regions	2,000 records each
Season	Dry, harmattan, wet	Region-dependent
Gender	Male, female	50% each
Genotype	HbSS, HbSC, HbS-β	70%, 25%, 5%
Hydroxyurea use	No, yes	65%, 35%

Table 4. Simulated regional statistical pattern.

Region	Average crisis frequency	Interpretation
North-East	Highest (4–5)	Extreme heat and dust
North-West	High (3–4)	Dry climate
North-Central	Moderate (3)	Transitional zone
South-West	Lower (2–3)	Humid climate
South-East	Low (2)	High rainfall
South-South	Lowest (1–2)	Very humid climate

4.3. REGIONAL CLIMATE PARAMETERIZATION

Each geopolitical region was assigned a climate profile based on published ecological and meteorological patterns. The North-East profile represented extreme heat, low humidity, and high dust exposure, with an illustrative mean temperature near $40\,°C$, relative humidity near 18%, and $PM_{2.5}$ near $95\,\mu g\,m^{-3}$. The South-South profile represented milder temperature, high humidity, and high rainfall. These parameters were used to construct a plausible simulation and should not be interpreted as validated regional clinical-risk estimates.

Table 5. Sample records from the synthetic multimodal health–climate dataset.

Region	Month	Season	Temp.	Humidity	Rainfall	$PM_{2.5}$	Age	Gender	Genotype	Hydroxyurea	Hb	Prev. hosp.	Crisis freq.	Pain	Visits/qtr	Label
North-East	8	Wet	44.0	11.5	90.0	104.9	23	Female	HbSS	No	6.0	2	3	2	1	1
North-West	2	Harmattan	44.0	13.6	9.0	74.6	53	Female	HbSC	No	7.8	0	5	3	5	1
South-South	10	Wet	26.4	89.6	373.9	42.0	9	Female	HbSC	No	6.9	3	0	1	0	0
South-South	12	Harmattan	27.7	89.7	247.2	40.3	52	Male	HbSS	No	5.7	3	1	1	1	0
North-East	5	Dry	36.0	19.4	51.4	86.5	10	Female	HbSS	No	8.1	5	2	1	2	1

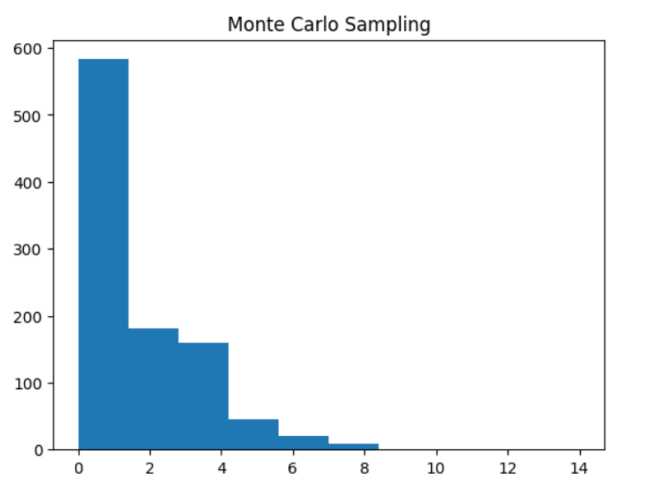


Figure 2. Monte Carlo sampling distribution of crisis frequency.

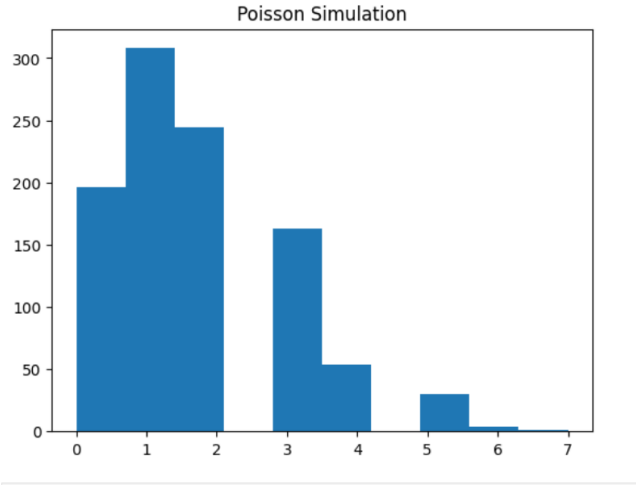


Figure 3. Simulated Poisson distribution of monthly crisis frequency.

Table 6. Performance metrics of the machine-learning models.

Model	Accuracy	F1-score	ROC-AUC
Logistic regression	0.78	0.74	0.84
Random Forest	0.77	0.72	0.83
SVM	0.78	0.73	0.82
XGBoost	0.76	0.70	0.82
Neural network	0.77	0.72	0.82

4.4. SIMULATION OF CLIMATE AND HEALTH VARIABLES

Temperature and humidity were simulated from region-specific Gaussian distributions and constrained to physically plausible ranges of 15–50 °C and 5–100%, respectively. Age was sampled uniformly, and sex was generated from a Bernoulli distribution with equal probabilities. Genotype probabilities were set to 70% HbSS, 25% HbSC, and 5% HbS-β thalassemia. Hydroxyurea use was assigned a probability of 35%. Hemoglobin concentration was normally distributed with mean $7.5\,\mathrm{g\,dL^{-1}}$ and standard deviation $1.2\,\mathrm{g\,dL^{-1}}$. Previous hospitalizations were simulated from a Poisson distribution with $\lambda = 2$.

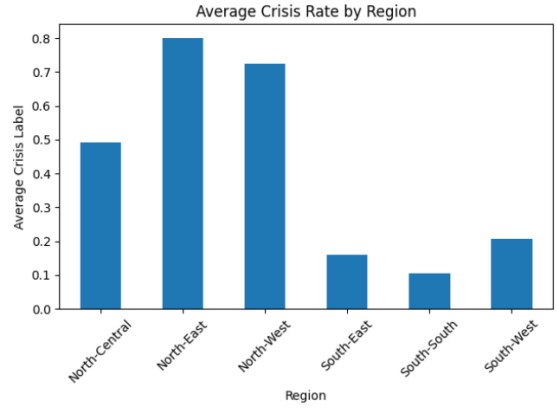


Figure 4. Average simulated crisis rate by Nigerian region.

4.5. CRISIS-OUTCOME MODELING

Crisis outcomes were generated from a domain-informed risk score rather than assigned at random. The baseline score combined normalized temperature, inverse humidity, $PM_{2.5}$, and stochastic noise:

$$R = 0.35T_n + 0.25(1 - H_n) + 0.20PM_n + 0.20\varepsilon, \qquad (1)$$

where T_n, H_n, and PM_n are normalized temperature, humidity, and $PM_{2.5}$, respectively, and ε is a stochastic noise term.

The baseline score was then modified by season, genotype, and hydroxyurea use:

$$R' = R\,M_{\text{season}}M_{\text{genotype}}M_{\text{hydroxyurea}}, \qquad (2)$$

where

$$M_{\text{season}} \in \{1.4\ (\text{harmattan}),\ 1.1\ (\text{dry}),\ 0.8\ (\text{wet})\},$$

$$M_{\text{genotype}} \in \{1.3\ (\text{HbSS}),\ 1.0\ (\text{HbSC}),\ 0.8\ (\text{HbS-}\beta)\},$$

and $M_{\text{hydroxyurea}} = 0.7$ for hydroxyurea use and 1.0 otherwise. Additional adjustments were applied for low hemoglobin and age extremes. Monthly crisis frequency was generated using a Poisson process parameterized by R'. Pain severity, hospital visits, and binary crisis status were derived through clinically informed transformations.

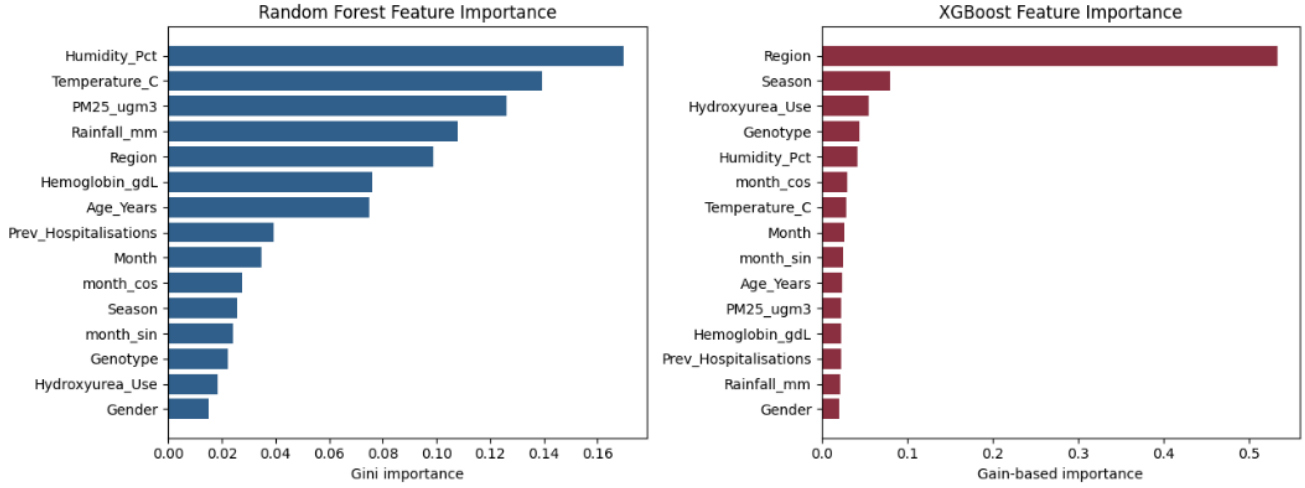
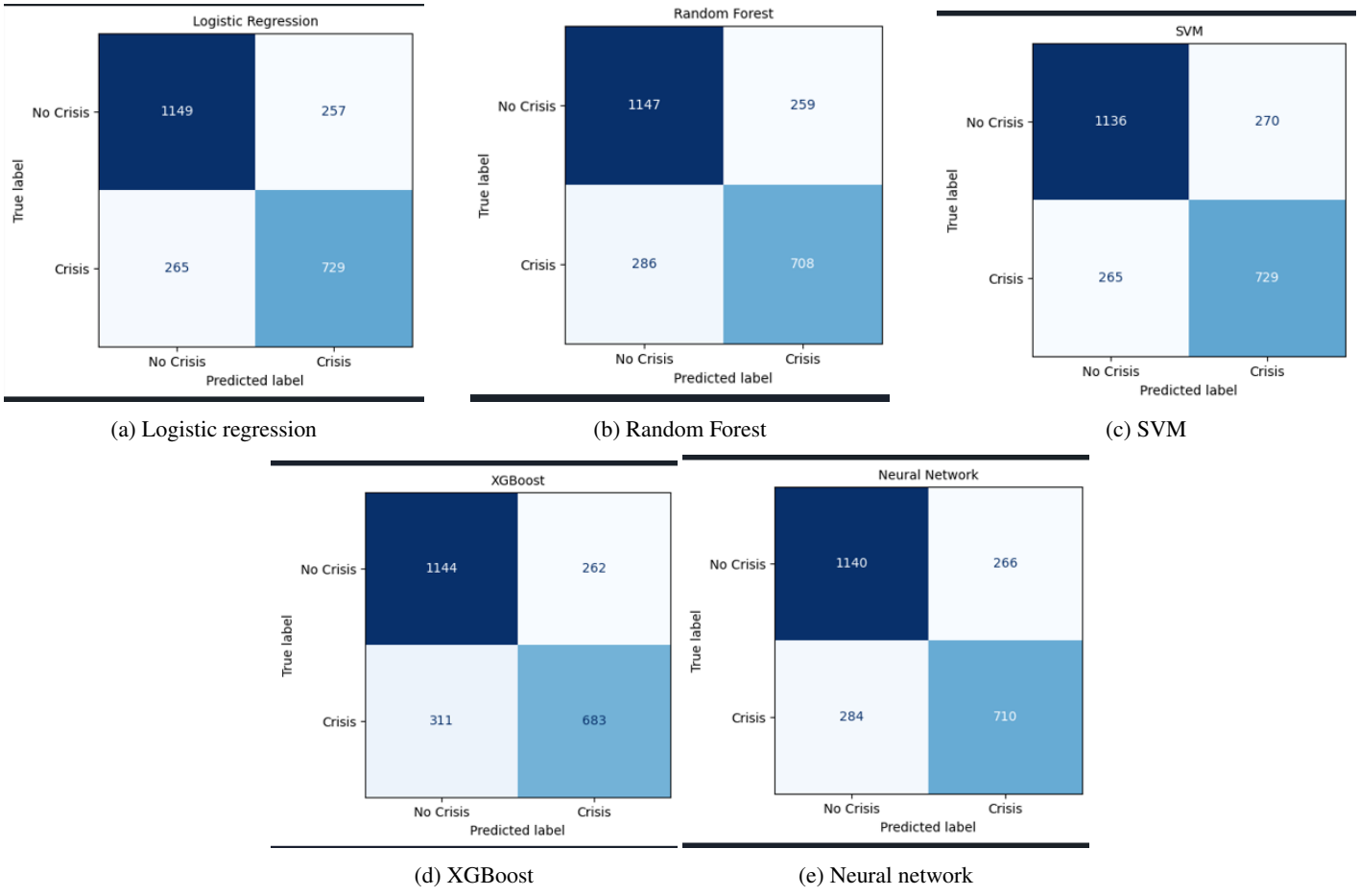


Figure 5. Feature importance for the random forest and XGBoost models.



(a) Logistic regression (b) Random Forest (c) SVM

(d) XGBoost (e) Neural network

Figure 6. Confusion matrices for the five classifiers.

The study also described a logistic probability for the artificial target variable:

$$P(\text{crisis}) = \left[1 + \exp\left(- \left[\beta_0 + \beta_1 T + \beta_2 H + \beta_3 D + \beta_4 C_{\text{prior}}\right]\right)\right]^{-1}, \tag{3}$$

where T is ambient temperature, H is relative humidity, D is the dust index, and C_{prior} is prior crisis history. Positive coefficients increase the modeled probability, whereas negative coefficients reduce it. Monte Carlo sampling converted the probability into a binary event.

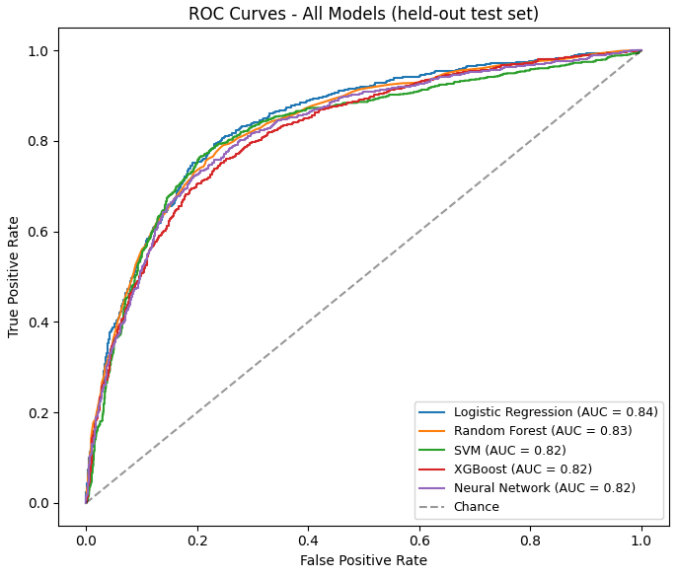


Figure 7. Receiver operating characteristic curves for all models on the held-out test set.

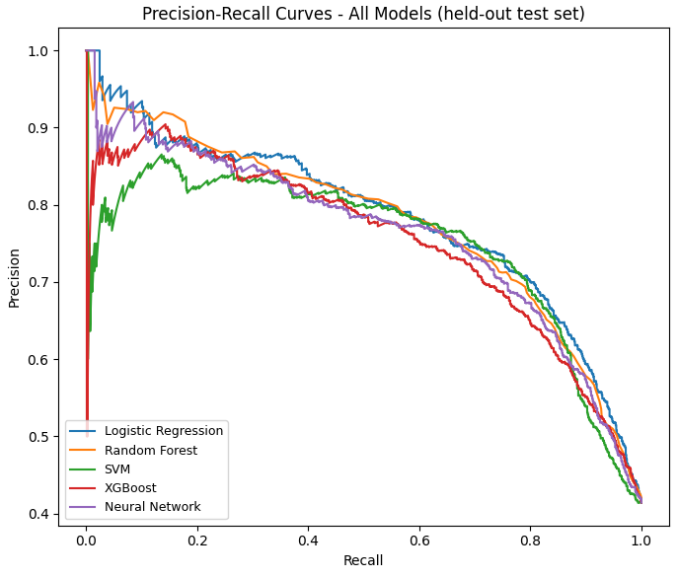


Figure 8. Precision–recall curves for all models on the held-out test set.

4.6. DATA ASSEMBLY AND INTERNAL PLAUSIBILITY CHECKS

The six regional datasets contained 2,000 observations each and were merged and randomly shuffled. Each record contained 18 climate, demographic, clinical, outcome, and temporal features. Because real patient data were unavailable, formal external validation was not possible. Internal checks indicated positive correlations of crisis occurrence with temperature and $PM_{2.5}$ and a negative correlation with humidity. The manuscript reports correlations of $r = 0.55$ for temperature, $r = -0.58$ for humidity, and $r = 0.53$ for $PM_{2.5}$. The final class distribution was 41.4% crisis and 58.6% non-crisis. These checks establish directional plausibility only.

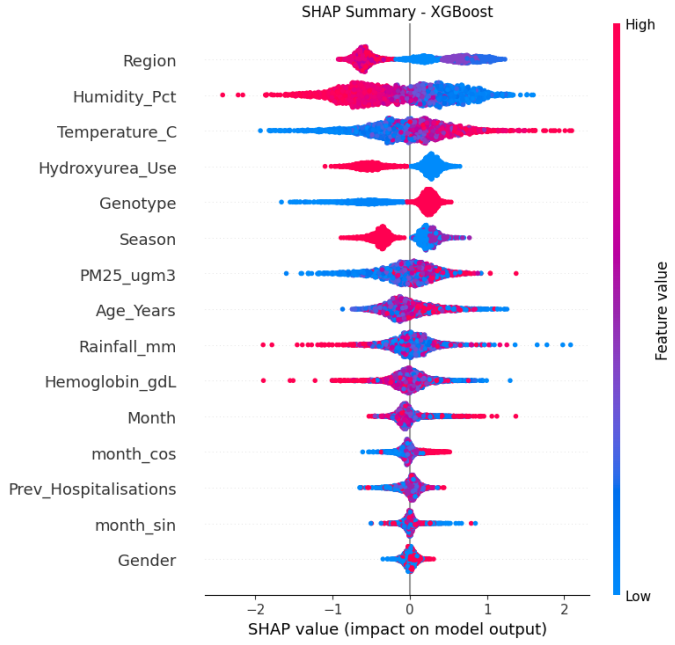


Figure 9. SHAP summary plot for the XGBoost model.

5. MACHINE-LEARNING MODELS

Five classifiers were evaluated.

5.1. RANDOM FOREST

Random Forest was selected for its tolerance of noise, capacity to model nonlinear interactions, and provision of feature-importance rankings. Ensemble methods have demonstrated utility in predictive healthcare, including disease-prediction applications [38].

5.2. SUPPORT VECTOR MACHINE

Support vector machines (SVMs) can provide robust classification in high-dimensional datasets and have been widely applied to public-health data [39].

5.3. XGBOOST

XGBoost was included because gradient-boosted trees can capture nonlinear relationships and interactions. A three-variable XGBoost model has previously predicted outcomes after acute ischemic stroke [40].

5.4. NEURAL NETWORK

A multilayer perceptron was evaluated because deep neural networks can model complex multivariable relationships and have demonstrated strong performance in disease-severity prediction [41].

5.5. LOGISTIC REGRESSION

Logistic regression served as an interpretable and computationally efficient baseline. Comparative evidence shows that performance gaps between machine learning and logistic regression can narrow as predictor richness increases. Climate-linked disease prediction in Damaturu also provides a methodologically relevant Nigerian precedent [42].

5.6. EXPERIMENTAL SETUP

The 12,000 patient-month observations were divided using an 80/20 stratified train–test split with `random_state=42`. `Crisis_Frequency`, `Pain_Severity_Score`, and `Hospital_Visits_Qtr` were excluded from the predictor set because they are downstream consequences of crisis occurrence. Numerical features were standardized with z-score normalization, and the categorical variables region, season, genotype, and gender were label-encoded.

No hyperparameter tuning or class-imbalance correction was performed. Logistic regression, Random Forest, SVM, and the neural network were implemented using scikit-learn defaults; XGBoost used its library defaults. The neural network contained one hidden layer of 100 units and was trained for a maximum of 500 iterations. Performance was assessed on a single held-out test set using accuracy, F1-score, and ROC-AUC. K-fold cross-validation was not performed and is identified as a priority for future work.

6. EXPERIMENTAL RESULTS

All five models achieved similar performance on the held-out synthetic test set (Table 6). Logistic regression had the highest ROC-AUC (0.84) and F1-score (0.74). Random Forest followed with ROC-AUC 0.83, while SVM, XGBoost, and the neural network each achieved ROC-AUC 0.82. The close clustering suggests that the engineered risk relationship was largely recoverable by both linear and nonlinear classifiers.

Figure 4 shows a marked simulated north–south gradient. Crisis rates were approximately 0.80 in the North-East, 0.72 in the North-West, 0.49 in the North-Central, 0.21 in the South-West, 0.16 in the South-East, and 0.10 in the South-South. Because the target was synthetically generated from climate-linked rules, these differences demonstrate recovery of the simulation design rather than observed regional prevalence.

The confusion matrices in Figure 6 show the counts of correct and incorrect crisis classifications. The ROC curves in Figure 7 nearly overlap, consistent with the narrow AUC range of 0.82–0.84. Figure 8 presents the precision–recall curves, which provide an additional diagnostic view for the 58.6%/41.4% class distribution.

Feature importance differed by model (Figure 5). Random Forest emphasized humidity, temperature, $PM_{2.5}$, rainfall, and region. XGBoost assigned particularly high gain importance to region, which may act as a proxy for region-specific climate profiles. The SHAP summary in Figure 9 similarly shows region, humidity, temperature, hydroxyurea use, genotype, season, and $PM_{2.5}$ among the influential features.

7. DISCUSSION

This study demonstrates that a simulation-and-modeling pipeline can recover an engineered climate–crisis relationship across ecologically heterogeneous Nigerian regions. The results establish methodological feasibility, but external clinical validation is required before any claim of real-world predictive accuracy or regional disease prevalence can be made.

7.1. REGIONAL VULNERABILITY

The simulation produced a clear north–south gradient: the North-East had the highest crisis probability (80.1%), followed by the North-West (72.4%) and North-Central (49.2%). The South-West, South-East, and South-South had lower probabilities of 20.6%, 15.8%, and 10.4%, respectively. These patterns arise from the imposed regional climate profiles. Northern regions were parameterized with higher temperature, lower humidity, and greater dust exposure, whereas southern regions were parameterized with higher humidity and rainfall. The results should therefore be interpreted as scenario outputs rather than epidemiological estimates.

7.2. CLIMATIC AND CLINICAL DRIVERS

Temperature, $PM_{2.5}$, humidity, and previous hospitalization emerged as influential features. High temperature can contribute to dehydration, while dust exposure may impose respiratory and physiological stress. Humidity showed a protective direction in the simulated risk function because it reduced dehydration-related risk. Rainfall contributed less after humidity and region were considered, suggesting that its modeled effect was partly mediated by other climate variables.

Hydroxyurea use reduced simulated risk, consistent with the multiplier imposed during target generation. Previous hospitalizations represented prior health burden. These associations demonstrate internal consistency with the data-generating mechanism; they do not independently validate causal effects.

7.3. POLICY IMPLICATIONS AND EARLY-WARNING POTENTIAL

A clinically validated climate-sensitive model could support seasonal counseling, hydroxyurea programs, and hospital preparedness. For example, weather-informed risk alerts might encourage hydration and avoidance of excessive outdoor exposure during high-risk periods. Hospitals could use validated forecasts to plan beds, blood supplies, analgesics, intravenous fluids, and staffing. These applications remain prospective and require validation with real patient and meteorological data.

7.4. HEALTHCARE PLANNING BENEFITS

Regional and seasonal risk estimates could eventually support resource allocation in areas with high disease burden. A validated system might help northern hospitals prepare for harmattan-related increases in admissions and could be adapted to other African countries with substantial SCA burden, including Ghana, Cameroon, the Democratic Republic of the Congo, and Tanzania. Country-specific calibration would be essential.

7.5. LIMITATIONS

The principal limitation is the exclusive use of synthetic data. Even a domain-informed simulation cannot reproduce the full heterogeneity, stochasticity, comorbidity, treatment response, care-seeking behavior, and longitudinal dependence observed in real patients. The use of an explicit risk function to generate the target also creates circularity: models are evaluated on their ability to recover relationships that were built into the data.

Patient-month observations were treated as statistically independent, whereas actual crisis risk changes with treatment ef-

fectiveness, age, cumulative organ damage, infection, and prior events. The study did not model repeated measures within individuals, long-term treatment effects, or comorbidity. Regional climate profiles were coarse and did not capture microclimatic variation, urban heat-island effects, or interannual variability. Higher-resolution data from sources such as ERA5 or the Copernicus Climate Change Service could improve ecological realism.

Potentially important variables were not included, such as household water access, hydration behavior, healthcare proximity, nutrition, concurrent infection, and psychosocial stress. In addition, performance was estimated from a single train–test split without cross-validation, hyperparameter tuning, calibration analysis, or external validation.

8. CONCLUSION

This proof-of-concept study demonstrates the feasibility of using machine learning to recover an engineered relationship between climate variables and synthetic SCA crisis outcomes across Nigerian regions. All five models achieved closely clustered performance (ROC-AUC 0.82–0.84), with logistic regression performing best (accuracy 0.78, F1-score 0.74, and ROC-AUC 0.84). Random Forest, SVM, XGBoost, and the neural network performed comparably.

The study contributes a reproducible synthetic-data framework for multimodal climate–health modeling, a simulated regional comparison across six Nigerian geopolitical zones, and an analysis of influential climate and clinical features. The observed regional gradient and model performance are properties of the simulation and must not be interpreted as validated clinical predictions. Real-world multicenter validation is required before deployment.

8.1. FUTURE WORK

The immediate priority is to obtain patient-level data linking crisis events with contemporaneous climate exposure. Collaboration with the Federal Ministry of Health, tertiary hospitals, and national sickle cell programs could support an SCA–climate registry for external validation and calibration.

Future models should represent longitudinal patient trajectories, cumulative climate exposure, treatment adherence, and disease progression. Recurrent neural networks, transformer architectures, survival models, and repeated-event methods may be suitable. High-resolution satellite temperature, $PM_{2.5}$, and Nigerian Meteorological Agency feeds could replace coarse region-level parameters. A low-data mobile application could then be pilot-tested with patients, caregivers, community health workers, and referral facilities, subject to clinical, ethical, usability, and regulatory evaluation. Future work should also examine hydroxyurea pharmacokinetics and fetal hemoglobin to support more individualized risk estimation.

DATA AVAILABILITY

The synthetic dataset and supporting material are available from the corresponding author upon reasonable request.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no financial or non-financial competing interests.

FUNDING

The authors received no external funding for this study.

References

[1] P. Wasnik, P. Das, A. Kumar, P. K. Kannauje, R. R., V. Pandit, T. Sahu & J. Sahu, "Hospitalization events among adolescents and adults with sickle cell disease in a tertiary care center in central India", Cureus **16** (2024) e61185. https://doi.org/10.7759/cureus.61185.

[2] E. I. Obeagu & G. U. Obeagu, "Implications of climatic change on sickle cell anemia: a review", Medicine **103** (2024) e37127. https://doi.org/10.1097/MD.0000000000037127.

[3] N. J. Desai, A. L. Gonzalez-Herrera, S. Malay, T. Brown, A. Owusu-Ansah & S. Ahuja, "Influence of weather on sickle cell disease vaso-occlusive episodes and acute chest syndrome: a nationwide sample analysis", Pediatric Blood & Cancer **72** (2025) e31504. https://doi.org/10.1002/pbc.31504.

[4] O. P. Adigwe, G. Onavbavba & S. O. Onoja, "A critical review of sickle cell disease burden and challenges in sub-Saharan Africa", Journal of Blood Medicine **14** (2023) 367. https://pmc.ncbi.nlm.nih.gov/articles/PMC10239624/.

[5] A. M. Thomson *et al.*, "Global, regional, and national prevalence and mortality burden of sickle cell disease, 2000–2021: a systematic analysis from the Global Burden of Disease Study 2021", The Lancet Haematology **10** (2023) e585. https://doi.org/10.1016/S2352-3026(23)00118-7.

[6] A. Issa, O. R. Ibrahim, R. M. Ibraheem, A. F. Lawal, M. Abdulbaki, M. B. AbdulKadir, B. S. Alabi & K. S. Ernest, "Prevalence of sickle cell disease and sickle cell traits among children and adolescents in Nigeria: a systematic review and meta-analysis", Systematic Reviews **15** (2026) 32. https://doi.org/10.1186/s13643-025-03029-1.

[7] B. Jonani, E. C. Kasule, B. R. Herman, J. F. Arturo, M. C. Mugambwa, S. Stephen, J. B. Mundaka, R. Kwizera, G. Mboowa & F. Bongomin, "Burden of sickle cell anemia in Africa: a systematic review and meta-analysis", PLOS ONE **20** (2025) e0337090. https://doi.org/10.1371/journal.pone.0337090.

[8] S. Sahu, S. Bhavanam, S. Younas, S. Gold-Olufadi, D. Saparov, H. Becerra, K. Kriplani, U. Soni & P. G. Bhasin, "Sickle cell disease in the global south: mapping the unmet burden using global burden of disease DALY estimates and its relevance to United States hematologists", Blood **146** (2025) 2601. https://doi.org/10.1182/blood-2025-2601.

[9] C. N. Amarachukwu, I. L. Okoronkwo, M. C. Nweke & M. K. Ukwuoma, "Economic burden and catastrophic cost among people living with sickle cell disease, attending a tertiary health institution in south-east zone, Nigeria", PLOS ONE **17** (2022) e0272491. https://doi.org/10.1371/journal.pone.0272491.

[10] C. Kamya, R. Idro, D. Kalibbala, K. S. Phiri, T. Nkosi-Gondwe, P. Akun, J. Kirikumwino, O. Kaarboe & R. Bjarne, "The economic burden and catastrophic health expenditures among children with sickle cell anaemia on households in malaria-endemic areas: insights from Uganda and Malawi", BMC Public Health **25** (2025) 2070. https://doi.org/10.1186/s12889-025-23209-x.

[11] B. I. Beli, L. I. Ali, C. C. Onuoha, M. Z. Jasseh, M. Zentari & N. Belakoul, "Socio-economic burden of sickle cell disease on families attending sickle cell clinic in Kano State, northwestern Nigeria", IJID Regions **12** (2024) 100385. https://www.sciencedirect.com/science/article/pii/S2667009724000617.

[12] S. N. Karatas, V. De Pierrefeu, S. Moutereau, N. Hebert, T. Mikobi, D. Baigl & P. Bartolucci, "Antibody- and equipment-free, ultralow-cost point-of-care sickle cell disease diagnostic: an equitably accessible powerful assay", Blood **146** (2025) 1171. https://doi.org/10.1182/blood-2025-1171.

[13] F. A. Ussher, E. F. Laing, A. B. Atta-Owusu, E. K. Kontor, N. Jain, S. Y. Lokpo, E. A. Adu, S. Ametepe, R. Tetteh, Y. Dei-Adomakoh & R. A. Ngala, "Identification of hematological biomarkers and assessment of machine learning models for sickle cell anemia severity classification", Journal of Sickle Cell Disease **2** (2025) yoaf020. https://doi.org/10.1093/jscdis/yoaf020.

[14] K. Z. Summers, O. Agrippa, D. Ade-Odunlade, K. A. Anie, P. Telfer, S. Lugthart & S. Kim, "Enhanced artificial intelligence (AI)-driven prediction of vaso-occlusive crises in sickle cell disease: precision through advanced machine-learning frameworks and digital remote monitoring", Blood **144** (2024) 522. https://doi.org/10.1182/blood-2024-209961.

[15] J. Patel, H. Muralikrishna, K. Chadaga, A. Thalengala & N. Sampathila, "Sickle cell disease detection in low-resource conditions using transfer-

learning and contrastive-learning coupled with XAI", Scientific Reports **16** (2026) 6104. https://doi.org/10.1038/s41598-026-35831-9.

[16] L. A. De Ligt, E. Brandsma, A. E. Gaartman, G. Van Gameren, C. Blans, M. Veldthuis, H. Matlung, B. J. Biemond, E. Nur, E. Van Den Akker & R. van Bruggen, "Artificial-intelligence driven quantification of sickle red blood cell morphology as an objective disease marker", Blood **144** (2024) 4996. https://doi.org/10.1182/blood-2024-199994.

[17] T. F. Machado, F. das C. Barros Neto, M. dos S. Goncalves, C. G. Barbosa & M. E. Barreto, "Exploring machine learning algorithms in sickle cell disease patient data: a systematic review", PLOS ONE **19** (2024) e0313315. https://doi.org/10.1371/journal.pone.0313315.

[18] S. S. Jennifer, M. H. Shamim, A. W. Reza & N. Siddique, "Sickle cell disease classification using deep learning", Heliyon **9** (2023) e22203. https://doi.org/10.1016/j.heliyon.2023.e22203.

[19] R. S. Zahr, A. Mohammed, S. Naik, D. Faradji, K. I. Ataga, J. Lebensburger & R. L. Davis, "Machine learning predicts acute kidney injury in hospitalized patients with sickle cell disease", American Journal of Nephrology **55** (2024) 18. https://doi.org/10.1159/000534864.

[20] G. O. Ghosheh, C. L. Thwaites & T. Zhu, "Synthesizing electronic health records for predictive models in low-middle-income countries (LMICs)", Biomedicines **11** (2023) 1749. https://doi.org/10.3390/biomedicines11061749.

[21] V. C. Pezoulas, D. I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. S. Tachos & D. I. Fotiadis, "Synthetic data generation methods in healthcare: a review on open-source tools and methods", Computational and Structural Biotechnology Journal **23** (2024) 2892. https://doi.org/10.1016/j.csbj.2024.07.005.

[22] Z. Qian, T. Callender, B. Cebere, S. M. Janes, N. Navani & M. van der Schaar, "Synthetic data for privacy-preserving clinical risk prediction", Scientific Reports **14** (2024) 25676. https://doi.org/10.1038/s41598-024-72894-y.

[23] O. P. Adigwe, G. Onavbavba & S. O. Onoja, "Improving access to healthcare services for sickle cell disease patients in Nigeria: perspectives and views of healthcare professionals", Frontiers in Health Services **5** (2025) 1466299. https://doi.org/10.3389/frhs.2025.1466299.

[24] Cleveland Clinic, "Sickle cell crisis: how to know you are having one (and what you can do about it)", 2024. Available online: https://my.clevelandclinic.org/health/diseases/sickle-cell-crisis.

[25] A. Kaponda, K. Muya, J. Panda, K. K. Koto & B. Bonnechere, "Unraveling the complexity of vaso-occlusive crises in sickle cell disease: insights from a resource-limited setting", Journal of Clinical Medicine **13** (2024) 2528. https://doi.org/10.3390/jcm13092528.

[26] D. S. Darbari, E. Sheehan & V. R. Bhatt, "Vaso-occlusive pain crisis in sickle cell disease: definition, pathophysiology, and management", European Journal of Haematology **105** (2020) 237. https://doi.org/10.1111/ejh.13430.

[27] F. Sophocleous, N. M. Archer, C. W. Lederer & International Hemoglobinopathy Research Network, "Genetic modifiers associated with vaso-occlusive crises and acute pain phenomena in sickle cell disease: a scoping review", International Journal of Molecular Sciences **26** (2025) 4456. https://doi.org/10.3390/ijms26094456.

[28] S. Ndour & B. Bonnechere, "The vaso-occlusive pain crisis in sickle cell patients: a focus on pathogenesis", Current Research in Translational Medicine **73** (2025) 103512. https://doi.org/10.1016/j.retram.2025.103512.

[29] C. Y. Wright, T. Kapwata, N. Naidoo, K. P. Asante, R. E. Arku, G. Cissé, B. Simane, L. Atuyambe & K. Berhane, "Climate change and human health in Africa in relation to opportunities to strengthen mitigating potential and adaptive capacity: strategies to inform an African brains trust", Annals of Global Health **90** (2024) 7. https://doi.org/10.5334/aogh.4260.

[30] D. Ishaku, E. T. Umaru, A. A. Adebayo, R. Löwner & A. A. Okhimamhe, "Analysis of the observed trends in rainfall and temperature patterns in North-Eastern Nigeria", Climate **12** (2024) 219. https://doi.org/10.3390/cli12120219.

[31] B. M. Akatah, H. Onyeaka, I. Onungwe, P. P. Akpan, K. T. Igulu & E. P. Elenwa, "Climate change-induced heatwaves in Nigeria: causes, challenges, and adaptive strategies", Journal of Environmental Management **394** (2025) 127433. https://doi.org/10.1016/j.jenvman.2025.127433.

[32] A. Mohammed, P. S. Podila, R. L. Davis, K. I. Ataga, J. S. Hankins & R. Kamaleswaran, "Using machine learning to predict early onset acute organ failure in critically ill intensive care unit patients with sickle cell disease: retrospective study", Journal of Medical Internet Research **22** (2020) e14693. https://doi.org/10.2196/14693.

[33] R. S. Chandrasekar, M. Kane & L. Krishnamurti, "Machine-learning predictive tool for the individualized prediction of outcomes of hematopoietic cell transplantation for sickle cell disease: registry-based study", JMIR AI **4** (2025) e64519. https://doi.org/10.2196/64519.

[34] N. G. Goswami, A. Goswami, N. Sampathila, M. G. Bairy, K. Chadaga & S. Belurkar, "Detection of sickle cell disease using deep neural networks and explainable artificial intelligence", Journal of Intelligent Systems **33** (2024) 20230179. https://doi.org/10.1515/jisys-2023-0179.

[35] S. Marino, R. Cassidy, J. Nanni, Y. Wang, Y. Liu, M. Tang, Y. Yuan, T. Chen, A. Sinha, B. Pandian, I. D. Dinov & M. L. Burns, "Medical data sharing and synthetic clinical data generation: maximizing biomedical resource utilization and minimizing participant re-identification risks", npj Digital Medicine **8** (2025) 526. https://doi.org/10.1038/s41746-025-01935-1.

[36] O. P. Adigwe, G. Onavbavba & S. O. Onoja, "Impact of sickle cell disease on affected individuals in Nigeria: a critical review", International Journal of General Medicine **16** (2023) 3503. https://doi.org/10.2147/IJGM.S410015.

[37] K. A. Ahmed, H. Bello-Manga & L. C. Jordan, "Sickle cell anemia and early stroke detection and prevention in Nigeria", Frontiers in Stroke **3** (2024) 1368576. https://doi.org/10.3389/fstro.2024.1368576.

[38] R. Mishra, A. K. Tiwari, C. Srivastava, P. Singh, P. Srivastava & R. K. Pandey, "Disease prediction and medicine recommendation using HistGradientBoosting and Random Forest algorithm", in Proceedings of Data Analytics and Management, ICDAM 2025, Lecture Notes in Networks and Systems, vol. 1608, Springer, Cham, Switzerland, 2026, pp. 444–451. https://doi.org/10.1007/978-3-032-03740-4_36.

[39] G. Khyathi, K. P. Indumathi, A. J. Hasin, M. L. F. Jency, S. S. Siluvai & G. Krishnaprakash, "Support vector machines: a literature review on their application in analyzing mass data for public health", Cureus **17** (2025) e77169. https://doi.org/10.7759/cureus.77169.

[40] C. C. Chung, E. C. Y. Su, J. H. Chen, Y. T. Chen & C. Y. Kuo, "XGBoost-based simple three-item model accurately predicts outcomes of acute ischemic stroke", Diagnostics **13** (2023) 842. https://doi.org/10.3390/diagnostics13050842.

[41] M. Rakovics, F. A. Meznerics, P. Fehérvári, T. Kói, D. Csupor, A. Bánvölgyi, G. A. Rapszky, M. A. Engh, P. Hegyi & A. Harnos, "Deep neural networks excel in COVID-19 disease severity prediction: a meta-regression analysis", Scientific Reports **15** (2025) 10350. https://doi.org/10.1038/s41598-025-95282-6.

[42] S. Jinad, M. O. Kama, M. Baba-Adamu, A. C. Ikegwu & O. Machete, "Analysis of malaria and climate in Damaturu City of Nigeria using predictive supervised learning", Discover Sustainability **6** (2025) 1239. https://doi.org/10.1007/s43621-025-02168-8.