EBM Experimental Biology and Medicine

TYPE Original Research
PUBLISHED 12 November 2026
DOI 10.3389/ebm.2026.11153

Check for updates

OPEN ACCESS

*CORRESPONDENCE
Ted Cory,
tcory1@uthsc.edu
Santosh Kumar,
ksantosh@uthsc.edu
Robert Davis,
rdavis88@uthsc.edu
Yan Cui,
ycui2@uthsc.edu

†These authors share first authorship

RECEIVED 18 April 2026
REVISED 16 September 2026
ACCEPTED 20 October 2026
PUBLISHED 12 November 2026

CITATION
Kumar S, Chinthala L, Cory T, Kumar S, Davis R and Cui Y (2026) Predicting stroke in HIV patients using machine learning and electronic health records: a national and regional analysis.
Exp. Biol. Med. 251:11153.
doi: 10.3389/ebm.2026.11153

COPYRIGHT
© 2026 Kumar, Chinthala, Cory, Kumar, Davis and Cui. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Predicting stroke in HIV patients using machine learning and electronic health records: a national and regional analysis

Sandeep Kumar[1,2†], Lokesh Chinthala[3†], Ted Cory[4*], Santosh Kumar[5*], Robert Davis[3*] and Yan Cui[1,2*]

[1]Department of Genetics, Genomics and Informatics, College of Medicine, University of Tennessee Health Science Center, Memphis, TN, United States, [2]Center for Integrative and Translational Genomics, College of Medicine, University of Tennessee Health Science Center, Memphis, TN, United States, [3]Center for Biomedical Informatics, Department of Pediatrics, College of Medicine, University of Tennessee Health Science Center, Memphis, TN, United States, [4]Department of Clinical Pharmacy and Translational Sciences, College of Pharmacy, University of Tennessee Health Science Center, Memphis, TN, United States, [5]Department of Pharmaceutical Sciences, College of Pharmacy, University of Tennessee Health Science Center, Memphis, TN, United States

Abstract

Stroke is a significant contributor to morbidity and mortality, particularly among people living with HIV (PLWH), who experience elevated incidence rates due to a higher prevalence of risk factors, such as hypertension, dyslipidemia, and chronic inflammation. The effective prediction and early identification of stroke in this population are critical but remain underexplored. This study employs machine learning (ML) algorithms, random forest and XGBoost, to predict stroke occurrence in PLWH using large-scale Electronic Health Record (EHR) datasets, including a national (U.S.) cohort and a regional (Tennessee) cohort. Both models demonstrated strong predictive performance, with random forest achieving an area under the receiver operating characteristic curve (AUROC) of 0.94 and XGBoost achieving 0.93 in the national dataset. The Tennessee cohort models exhibited lower yet notable predictive capability (AUROC ~0.82), likely reflecting smaller sample size and cohort differences. In both cohorts, the two ensemble models outperformed a logistic regression, a feedforward deep neural network, and an adapted Framingham Stroke Risk Profile. Permutation-based feature-importance analyses identified clinical features associated with stroke risk. These findings suggest that ML models can effectively predict stroke among PLWH and highlight potential modifiable risk factors that could guide targeted preventive interventions.

KEYWORDS
EHR, HIV, machine learning models, prediction models, stroke

Impact statement

This study is important because people living with HIV face a higher risk of stroke, yet tools to identify who is most at risk remain limited. Using large national and Tennessee Electronic Health Record (EHR) datasets, we develop machine learning models to predict stroke occurrence in people living with HIV. The models achieved high predictive ability in the national cohort and identified the clinical features that contributed most to stroke

prediction in this population. Several of these factors, including blood pressure, cholesterol, blood glucose, and smoking, are modifiable through existing treatments and interventions. By identifying these predictors from EHR data, our work provides a practical foundation for developing risk assessment models that could identify people living with HIV at elevated risk who may benefit from closer evaluation and management of established risk factors, ultimately reducing the burden of stroke in this vulnerable population.

Introduction

Stroke is one of the leading causes of morbidity and mortality worldwide, disproportionately impacting vulnerable populations, such as people living with HIV (PLWH), who face an elevated risk of stroke due to higher rates of hypertension, diabetes mellitus, smoking, hyperlipidemia, chronic inflammation, and immune activation [1, 2]. Thanks to the availability of antiretroviral therapy (ART), PLWH now experience extended lifespans

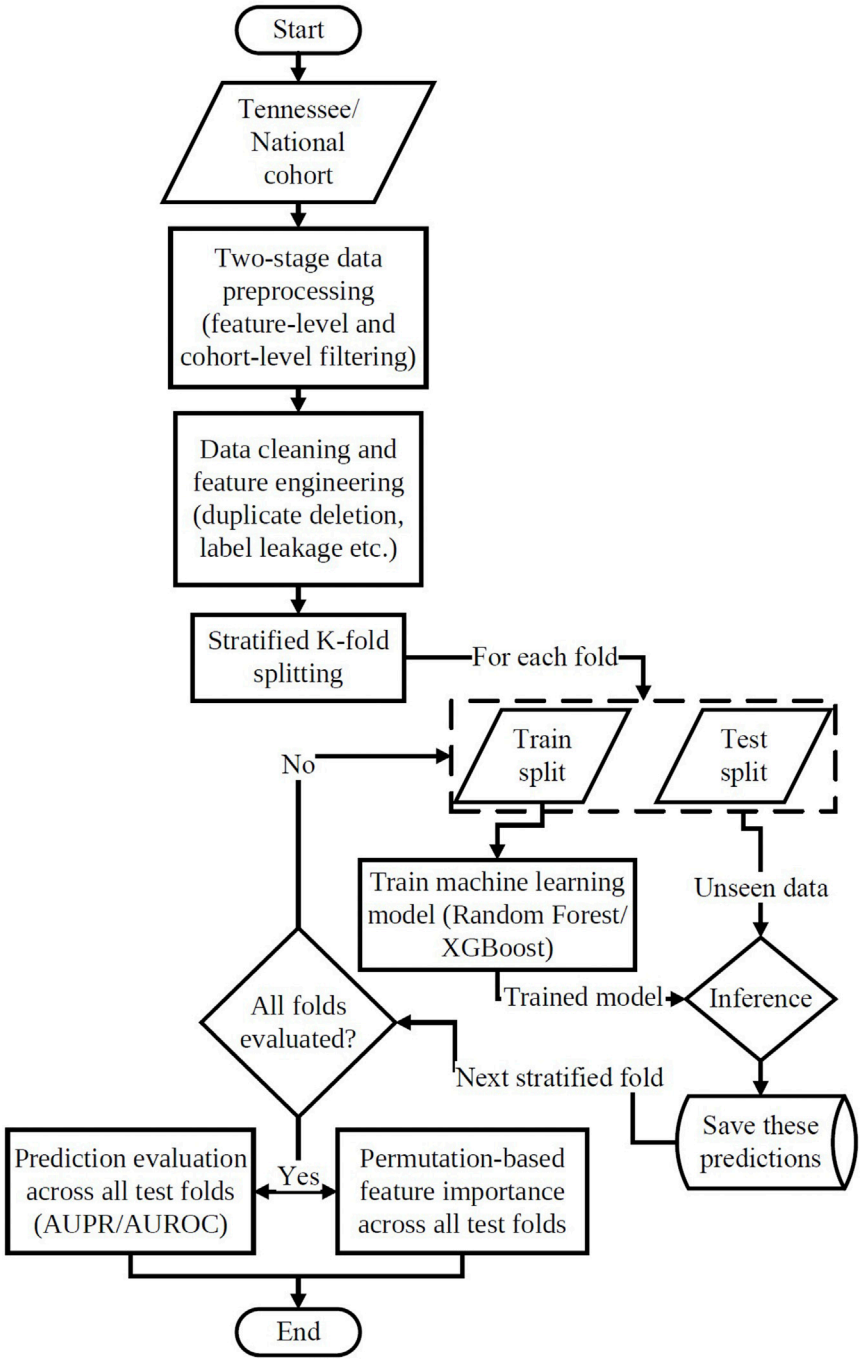


FIGURE 1
Flowchart of the proposed methodology for predicting stroke in PLWH.

TABLE 1 PLWH patient datasets.

Information	Stratification	Tennessee cohort	National dataset (2009-2017)
Size of cohorts	-	5,320	212,663
Stroke (%)	-	199 (3.7%)	2,085 (1.0%)
Without stroke (%)	-	5,121 (96.3%)	210,578 (99%)
Age (Mean ± SD)	-	49 ± 11	48 ± 10
Age group (%)	21–30	17 (0.3%)	3,924 (1.8%)
	31–40	1,267 (23.8%)	45,640 (21.5%)
	41–50	1,543(29.0%)	75,709 (35.6%)
	51–60	1,623 (30.5%)	63,406 (29.8%)
	61–70	738 (13.9%)	19,424 (9.1%)
	>70	132 (2.5%)	4,560 (2.1%)
Gender (%)	Male	3,597 (67.6%)	146,546 (68.9%)
	Female	1,721 (32.3%)	65,588 (30.8%)
	Unknown	341 (6.4%)	529 (0.2%)
Race (%)	Black or African American	3,323 (62.5%)	108,466 (51.0%)
	White	1,656 (31.1%)	76,720 (36.1%)
	Other minority	55 (1.0%)	6,528 (3.1%)

relative to those affected by this epidemic earlier on. However, PLWH may experience accelerated aging and are particularly vulnerable to age-related neurological diseases, including stroke [3, 4]. HIV-associated stroke results in a substantial disease burden, decreased quality of life, a significant economic impact, and increased mortality [5].

Stroke prediction and early detection in PLWH are crucial because they enable healthcare providers to implement timely interventions, such as aggressive management of blood pressure, dyslipidemia, and other modifiable risk factors, in addition to optimized ART, which can reduce the occurrence and severity of cerebrovascular events in this vulnerable group [6]. Despite its importance, stroke prediction in PLWH remains largely unexplored.

Electronic Health Record (EHR) data are an important resource for predicting the occurrence of complex diseases because they provide comprehensive patient information, such as demographics, clinical histories, laboratory results, and medication records [7]. In recent years, machine learning has emerged as a critical tool for leveraging EHR data in disease diagnosis and prognosis, due to its ability to process large, complex datasets and to identify patterns that might not be detectable through traditional statistical methods [8, 9]. Machine learning algorithms capable of effectively processing high-dimensional data and capturing non-linear relationships and complex interactions have significantly improved stroke diagnosis and prognosis [10, 11, 12].

In this study, we developed machine learning models for stroke prediction in two cohorts of PLWH and identified potential stroke-predictive biomarkers in PLWH. We addressed the challenge of class imbalance and found that machine learning models, specifically random forest and XGBoost, achieved high predictive performance for the larger national (U.S.) cohort and reasonably good performance for the smaller regional (Tennessee) cohort.

Materials and methods

The flowchart in Figure 1 outlines the proposed methodology for predicting strokes in PLWH across national (U.S.) and Tennessee cohorts and the associated analysis on the importance of clinical features. The process began with two-stage data preprocessing, including feature-level and cohort-level filtering, followed by data cleaning and feature engineering steps, such as duplicate row deletion and addressing label leakage. The pre-processed data were then split into five stratified folds. For each fold, a model (Random Forest/XGBoost) was trained on the training split and evaluated on the held-out test split. Predictions from each fold's test set were recorded, and once all folds were processed, the overall performance was evaluated using the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPR). Finally, a permutation-based feature importance analysis was conducted to estimate the contribution of different features to the model's predictions.

Data processing and cleaning

This study investigated stroke occurrence among PLWH using two datasets for the national (U.S.) and the Tennessee cohorts (Table 1). The national dataset was extracted from the Cerner Health Facts database, and the Tennessee dataset was sourced from the University of Tennessee Health Science Center (UTHSC) Research Enterprise Data Warehouse (rEDW), a

10.3389/ebm.2026.11153

TABLE 2 Clinical features.

Tennessee cohort features	National cohort features
HIGH_CHOLESTROL	HIGH_CHOLESTROL
AGE_AT_HIV_DX	AGE_AT_HIV_DX
HYPERTENSION	HYPERTENSION
Protein serum/plasma	Protein total, serum
Alkaline phosphatase serum/plasma	Alkaline phosphatase, serum
SMOKING_STATUS	SMOKING_STATUS
CONGESTIVE_HEART_FAILURE	CONGESTIVE_HEART_FAILURE
Potassium serum/plasma	Potassium, serum
Chloride serum/plasma	Chloride, serum
DIABETES	DIABETES
Total bilirubin serum/plasma	Bilirubin total serum or plasma Mass/Volume
RACE	RACE
ATRIAL_FIBRILLATION	ATRIAL_FIBRILLATION
EXCESSIVE_ALCOHOL_USE	EXCESSIVE_ALCOHOL_USE
TRANSIENT_ISCHEMIC_ATTACK	TRANSIENT_ISCHEMIC_ATTACK
GENDER	GENDER
Leukocytes in blood by Automated count	White blood cell count
Basophils in blood by Automated count	Basophils percent
Platelets in blood by Automated count	Mean platelet volume
Hematocrit of blood by Automated count	Hematocrit
Glucose lab	Glucose, Serum/Plasma Quantitative
Creatinine serum/plasma	Creatinine, serum Quantitative
Carbon dioxide serum/plasma	Carbon dioxide CO2
Monocytes in blood by Automated count	Monocyte percent
Urea nitrogen serum/plasma	Blood urea nitrogen
Hemoglobin	Hemoglobin
Eosinophils in blood by Automated count	Eosinophil count
Sodium serum/plasma	Sodium, serum
Body temperature	Albumin, serum
Calcium serum/plasma	Triglyceride, serum
BP diastolic	HDL cholesterol

standardized, in-house-developed healthcare data repository. The rEDW aggregates data from multiple healthcare organizations across Tennessee, encompassing records for approximately 4.5 million patients. Both datasets comprised individuals with a history of stroke and a control group of those without.

PLWH were identified using ICD-9 code 042 and ICD-10 codes B20-B24. Stroke cases were identified using relevant stroke-related ICD codes across all healthcare encounter types. For cases, laboratory data from before the index stroke visit were extracted. For controls without a history of stroke, a pseudo-index date was assigned based on the most recent stroke visit date of a matched case. Controls were matched on gender, race, and age at the time of HIV diagnosis, and their laboratory data recorded prior to the pseudo-index date were extracted for the analysis. No fixed prediction horizon was imposed; the models discriminate PLWH who experienced a stroke from those who did not, using data recorded before the index or pseudo-index date.

Missing data are common in EHR datasets. To address this issue, we used a cohort-specific missingness threshold to select features for machine learning. For the Tennessee cohort, features with >32%

TABLE 3 PLWH dataset after preprocessing.

Cohorts	Tennessee cohort	National dataset
Non-stroke	2357	44901
Stroke	199	2085

missing values were excluded, yielding 44 features. Given the higher frequency of missing data in the national dataset, a more lenient threshold of ~77% was used, resulting in ~41 features. Across cohorts, 28 features were shared; each cohort included 3 additional features selected empirically for performance (Table 2), for a total of 31 features used per cohort. In addition, we performed patient-level filtering within the non-stroke groups: any record with >10% missing values across the retained features was removed. This step reduced the size of the non-stroke cohorts (to 46% of the original size in Tennessee and ~21% nationally) while retaining all stroke cases to preserve minority-class signal (Table 3).

Machine learning model development

Two machine learning models were developed and evaluated: a Random Forest model [13, 14] and an XGBoost model [15].

Random forest model configuration

A Random Forest classifier was developed for its robustness in handling high-dimensional data and its ability to mitigate overfitting. The model's hyperparameters for each dataset are stated below:

- *n_estimators:* 1,500 for the Tennessee Dataset and 2,500 for the National Dataset. The parameter is the number of trees in the forest.
- *max_depth:* None (no explicit limit on tree depth)
- *min_samples_split:* 2 (the minimum number of samples required to split an internal node)
- *min_samples_leaf:* 1 (the minimum number of samples required to be at a leaf node)

The *n_estimators* parameter was increased for the larger National Dataset to better capture complex patterns. Setting *max_depth* = None allows the trees to grow fully. The hyperparameters min_*samples_split* and *min_samples_leaf*, control when a tree branch can further split.

XGBoost model configuration

XGBoost (gradient-boosted decision trees) was trained without explicit imputation, leveraging its ability to route missing values during splits. The hyperparameters for the XGBoost model are selected as stated below:

- *n_estimators:* 10,000 (maximum number of boosting rounds)
- *max_depth:* 6 (maximum depth of each tree)
- *learning_rate:* 0.01 (scales the contribution of each tree to prevent overfitting)
- *subsample:* 0.8 (fraction of samples used for the training of each tree)
- *colsample_bytree:* 0.8 (fraction of features used for the training of each tree)

Here, *max_depth* = 6 ensures that each tree grows up to this depth without becoming too complex. A *learning rate* of 0.01 prevents overfitting by making incremental updates, while *subsample* = 0.8 and *colsample_bytree* = 0.8 introduce randomness, improving generalization. To prevent overfitting, early stopping was implemented during XGBoost training. The training process was halted if the model's performance did not improve for 50 consecutive boosting rounds, as measured by the Area Under the Precision-Recall Curve (AUPR) on a held-out test set.

Logistic regression, deep neural network, and adapted Framingham Stroke Risk Profile

A penalized logistic regression was implemented in scikit-learn as a linear statistical baseline and fitted to the same restricted set of 31 features and the same stratified 5-fold partitions as the ensemble models. Because a linear model admits neither missing values nor unscaled inputs, missing values were mean-imputed, and features standardized, with the imputer and scaler fitted on the training fold only and applied to the held-out fold. Features whose absolute Pearson correlation with an earlier feature exceeded 0.5 were removed to limit collinearity, a filter computed on the training fold alone, leaving 26 of the 31 features in the Tennessee folds. The model used an L2 (ridge) penalty at the default inverse regularization strength of 1.0, optimized by the lbfgs solver for up to 5,000 iterations, raised from the default of 100 to ensure convergence. No class weighting or resampling was applied; thus, the fitted probabilities remained calibrated to the prevalence of the analyzed cohort rather than being rescaled toward the minority class. All random seeds were fixed at 1. Fitting every preprocessing step inside the training fold matters more for this model than for the tree ensembles, which require no imputation: imputing on the full cohort would allow held-out patients to influence the feature means used to fill their own missing values.

A feedforward neural network was implemented in PyTorch and trained on the same set of features as the ensemble models. Missing values were mean-imputed, and features standardized, with the imputer and scaler fitted on the training fold only and applied to the held-out fold. The network comprised three hidden layers of 256, 128, and 64 units, each followed by batch normalization, rectified linear activation, and dropout (rate 0.3), and a linear output layer producing a single logit, resulting in 50,305 trainable parameters. Within each training fold, a random 15% of patients formed an internal validation set, and the remaining 85% were used for weight updates. The loss function was binary cross-entropy, with stroke cases weighted by the ratio of non-stroke to stroke patients in the training fold to counteract class imbalance without resampling. Weights were optimized with AdamW (learning rate 0.001, weight decay 0.0001) in minibatches of 512 patients for the national cohort and 128 patients for the Tennessee cohort, for up to 80 epochs. Average precision (AUPR) on the internal validation set was computed after each epoch; the weights with the highest value were retained, and training stopped if no improvement occurred over 12 consecutive epochs.

As a clinical reference model, the Framingham Stroke Risk Profile [16] was applied to both cohorts after adapting it to the available variables. Several components of the original score were unavailable and were handled as follows: left ventricular hypertrophy on electrocardiogram was excluded; antihypertensive treatment status was not recorded, so the untreated systolic blood pressure (SBP) coefficients were used for all patients; SBP was entirely absent from the national dataset and was assigned a fixed value of 94 mmHg. Congestive heart failure served as a proxy for cardiovascular disease, age at HIV diagnosis was substituted for current age, and remaining missing values were mean-imputed. Diabetes status, smoking status, atrial fibrillation, and gender were used as recorded. The adapted score was computed for every patient in each analyzed cohort and evaluated with the same metrics as the machine learning models. Because these substitutions differ from the original specifications, the resulting score is an approximation of the Framingham profile rather than the validated instrument.

Machine learning model performance evaluation

Model performance was assessed using two key metrics:

- Area Under the Receiver Operating Characteristic Curve (AUROC): It measures a model's ability to distinguish between PLWH with stroke and without stroke. It is calculated as the area under a plot with the Y-axis as true positive rate (sensitivity) and the X-axis as the false positive rate across different classification thresholds.
- Area Under the Precision-Recall Curve (AUPR): It is particularly useful for evaluating models on imbalanced datasets. It is calculated as the area under the Precision-Recall curve, which plots the proportion of predicted stroke cases that are actual strokes (precision) on the Y-axis and the proportion of actual stroke cases correctly identified (recall) on the X-axis.

Both the above metrics range from 0 to 1; a value closer to 1 is considered an excellent prediction, while a value closer to 0 is considered the worst in both cases.

To ensure that the training and test sets adequately represented the minority class, the following strategy was used:

- Stratified Train-Test Split and Cross-Validation: Both the training and testing sets accurately reflected the same class distribution proportion as in the original dataset. Model performance was evaluated using 5-fold stratified cross-validation, providing a more robust, unbiased assessment of the model's generalization.

Feature importance analysis

Feature importance was determined using the permutation importance methodology. This technique assesses the importance of a feature by measuring the decrease in model performance when the feature's values are randomly shuffled (in the experiments in this study each feature was shuffled 20 times on the test fold). A larger decrease in performance indicates a more important feature. Only those features were included in the feature importance plots that exhibited positive values following summation of the permutation importance score and respective standard deviation across the five test folds. This criterion effectively excluded features with inconsistent or potentially spurious importance.

Results

Stroke statistics in Tennessee and national cohorts of PLWH

The age and sex distributions were broadly similar in the national and Tennessee cohorts. The mean age was 48–49 years, with comparable age-group profiles, and men comprised ~68% of PLWH in both datasets (~31% women; remainder unknown). Minor differences were present (e.g., a slightly higher proportion of individuals aged 51–60 in Tennessee), but the overall patterns were consistent. In the Tennessee cohort, 3.7% of PLWH had a stroke, while 96.3% did not. The national dataset, while larger, exhibits a more pronounced imbalance, as only 1.0% of PLWH experienced a stroke. The relatively larger percentage of stroke incidence in Tennessee compared with the national cohort may reflect the smaller dataset size in the Tennessee cohort or regional demographic composition (Table 1). These statistics are consistent with the literature, with previous research indicating that between 1% and 5% of patients with HIV experience stroke in clinical case series [3].

Racial composition differed between the cohorts (Table 1): Tennessee had a higher proportion of Black or African American PLWH (62.5% vs. 51.0% nationally) and a lower proportion of White PLWH (31.1% vs. 36.1%); "Other" was a small fraction in both. In both cohorts, stroke cases were disproportionately high among Black or African Americans (national: 59.2% of strokes vs. 50.9% of non-strokes; Tennessee: 76.4% vs. 61.9%), while the White and Other groups accounted for smaller shares of stroke cases (national: 30.9% and 9.9%; Tennessee: 19.6% and 4.0%) (Figures 2A,B), with greater disparities in the Tennessee cohort.

Machine learning prediction of stroke in PLWH

We evaluated XGBoost and random forest using two global performance metrics: AUROC and AUPR (Table 4; Figure 3), estimated from stratified 5-fold cross-validation.

National cohort

In the national cohort, both XGBoost and random forest achieved high predictive performance despite the severe class imbalance. XGBoost reached an AUPR of 0.81 (≈18.4 times the random baseline, with stroke cases representing 4.4% of the processed dataset) and an AUROC of 0.93, while random forest achieved an AUPR of 0.79 (≈18.0 times the random baseline) and an AUROC of 0.94.

Tennessee cohort

In the Tennessee dataset, performance was lower than in the national cohort but remained informative. Both XGBoost and random forest obtained an AUPR of 0.50 (≈6.4 times the random baseline, with stroke cases representing 7.8% of the processed dataset) and an AUROC of 0.82.

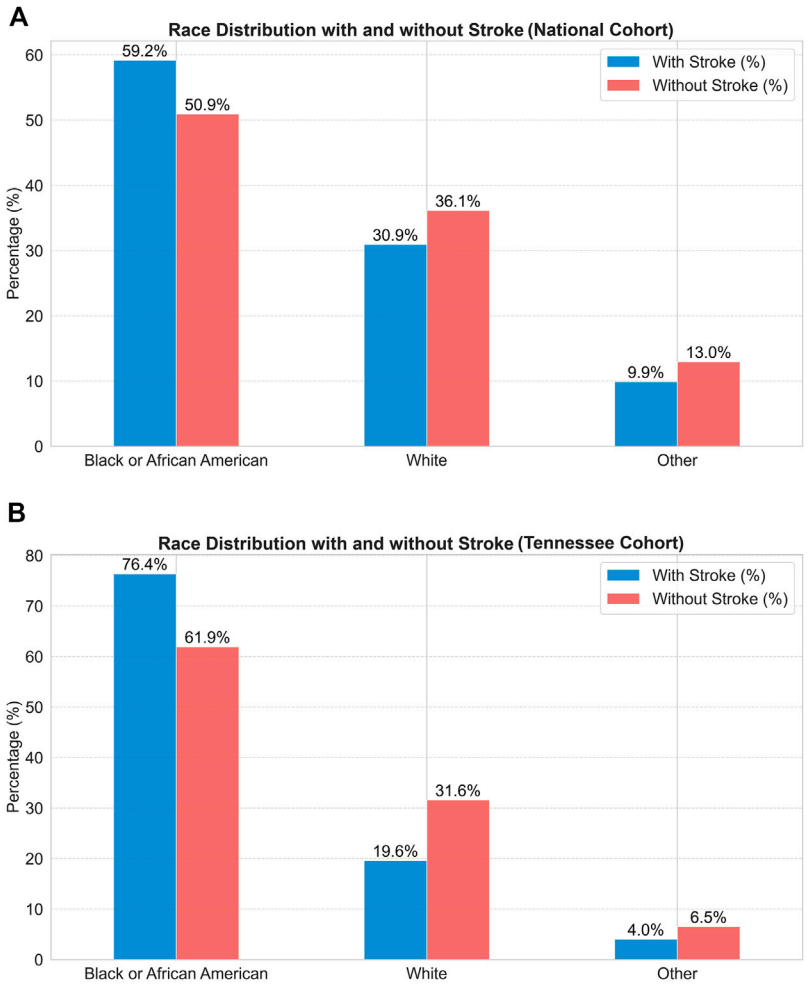


FIGURE 2
PLWH race-specific distribution [**(A)** National data, **(B)** Tennessee data].

TABLE 4 **Performance analysis.**

Cohorts	Tennessee cohort	National dataset
Models	AUPR, AUROC	
XGBoost	0.50, 0.82	0.81, 0.93
Random forest	0.50, 0.82	0.79, 0.94
Logistic regression	0.18, 0.70	0.23, 0.79
Deep neural network	0.40, 0.76	0.71, 0.91
Framingham stroke risk profile (adapted)	0.13, 0.64	0.30, 0.44

Top features predictive of stroke in PLWH

National cohort

Permutation importance (Figure 4A) identified blood pressure, lipid measures, smoking status, age at HIV diagnosis, and hematologic indices (e.g., leukocyte/basophil counts) as influential for stroke prediction. Albumin, alkaline phosphatase, bilirubin, creatinine, glucose, and total protein also ranked highly, suggesting contributions from metabolic, hepatic, and renal function.

The XGBoost model achieved a high discriminative performance, with a mean AUPR of 0.81 and a Mean AUROC of 0.93 (as shown in Figures 3A,B, respectively), which identified key predictors within the National cohort. Feature importance analysis using mean permutation importance (across five test folds), as

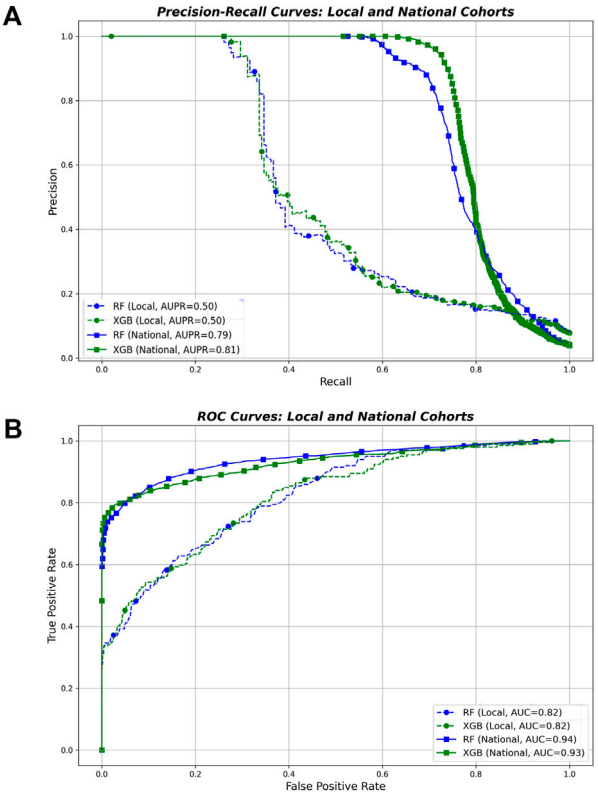


FIGURE 3
Performance of XGBoost on the National cohort, XGBoost on the Tennessee cohort, Random Forest on the National cohort, and Random Forest on the Tennessee cohort [**(A)** Precision-Recall Curves, **(B)** ROC Curves]. The curves were computed from the predictions on the held-out test folds of the stratified 5-fold cross-validation (Tennessee cohort: 199 stroke and 2,357 non-stroke PLWH; national cohort: 2,085 stroke and 44,901 non-stroke PLWH; Table 3). 'Local' in the figure key denotes the Tennessee cohort. The random-baseline AUPR equals the stroke prevalence in the analyzed cohort (0.078 in Tennessee, 0.044 in the national cohort), and the random-baseline AUROC is 0.5. AUPR and AUROC values are listed in Table 4.

depicted in Figure 4A, revealed that serum albumin levels, hypertension, high-density lipoprotein (HDL), monocyte percentage, serum Triglycerides, and a history of transient ischemic attack were the strongest predictors within the model. Other contributing factors include hematocrit, white blood cell count, age at HIV diagnosis, high cholesterol, carbon dioxide levels, mean platelet volume, smoking status, alkaline phosphatase, creatinine serum, bilirubin level, glucose, basophil percentage, congestive heart failure, diabetes, eosinophil count, potassium serum, race, sodium serum, atrial fibrillation [9, 17], gender [17], and excessive alcohol use, highlighting the non-linear relationship of the clinical variables captured by the model.

The performance of the XGBoost model in identifying strokes within the PLWH cohort in Tennessee, as depicted in Figure 3A, was notable, with a mean AUPR of 0.50 and a mean AUROC of 0.82 (in Figure 3B). However, it demonstrated a significant reduction compared to its performance on the larger national dataset. This decrease is likely due to the smaller size of the Tennessee dataset being approximately 2.5% the size of the national dataset (see Table 1). Therefore, the model had fewer examples to learn from, which impacted its performance. The AUPR of 0.50, while appearing low, was 6.4 times the random baseline score of 0.078 (see Table 3, as the imbalance is 7.8% in the training data), indicating a significant improvement over chance. An AUPR of 0.50 and an AUROC of 0.82 (compared to a random baseline AUROC score of 0.5) suggested that the model possesses a reasonably good ability to distinguish between PLWH with stroke and without stroke.

Figure 4B displays the mean permutation importance, representing the decrease in AUROC when a feature's values are randomly shuffled. High cholesterol, "AGE_AT_HIV_DX" (i.e., age at HIV diagnosis), hypertension, glucose lab, monocytes, protein serum/plasma, potassium levels, hemoglobin, smoking status, leukocytes, urea nitrogen levels, platelets, alkaline phosphatase levels, race, diastolic BP (blood pressure), calcium levels in plasma, congestive heart failure, and diabetes appeared to be the most influential features, as indicated by longer bars and lighter colors, which signify their significant impact on the model's performance. Conversely, features such as gender, excessive alcohol use, and bilirubin and chloride levels had minimal influence due to their near-zero or negative importance values.

XGBoost analysis on the national vs. Tennessee cohorts

Comparing the feature importance of XGBoost models in predicting an outcome in the National cohort and the Tennessee cohort revealed interesting observations. Hypertension emerged as a consistently important predictor across both groups. Monocyte/white blood cells/leukocytes were important in both models, supporting the role of infection and inflammation in stroke risk [18]. Other potential similarities included age at HIV diagnosis and smoking status, suggesting that these factors play a role in both populations. In general, key factors in both the National and Tennessee cohorts were related to hypertension, smoking, infection, and inflammation, all of which have been associated with increased stroke risk in the past [19]. While there are differences in the relative importance of these factors between the models, both models show that these factors may be important markers of stroke risk in PLWH.

Figure 3A shows the Precision-Recall curve for a Random Forest model evaluated on a national cohort, achieving an AUPR of 0.79. This metric suggests strong performance in identifying and ranking positive cases (i.e., PLWH with stroke) within the dataset. Furthermore, the Random Forest model achieved an AUROC of 0.94 on the National cohort (as depicted in Figure 3B), indicating excellent performance. Interestingly, the performance of the Random Forest model was similar to that of an XGBoost model that achieved an AUPR of 0.81 (Figure 3A) and an AUROC of 0.93 (Figure 3B) on the same National dataset. Despite their different underlying mechanisms, both models leverage the characteristics of the national dataset to achieve comparable levels of performance. The XGBoost model was optimized using early stopping on AUPR and hence had slightly superior performance in that particular metric. Both the Random Forest and XGBoost models

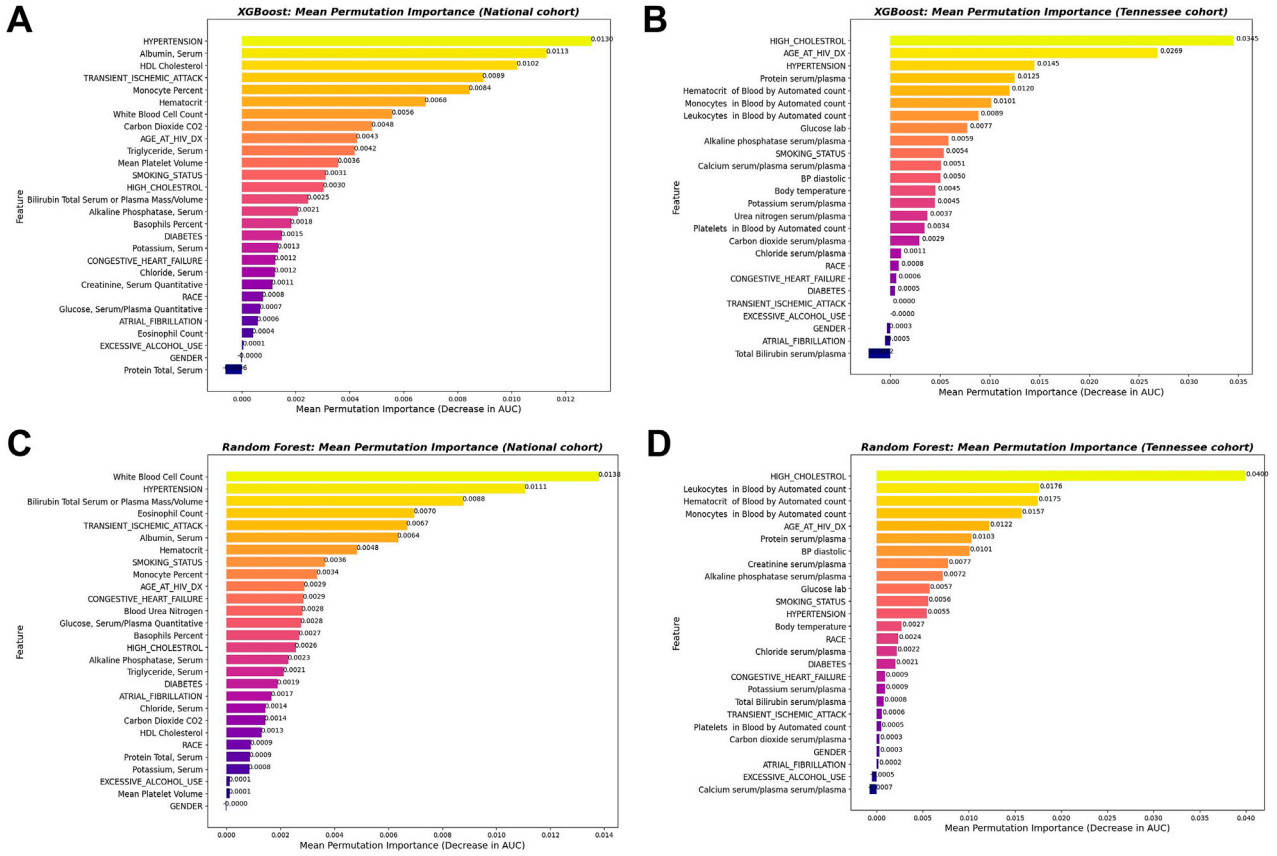


FIGURE 4
Permutation Feature importance scores [**(A)** National cohort using XGBoost, **(B)** Tennessee cohort using XGBoost, **(C)** National cohort using Random Forest, **(D)** Tennessee cohort using Random Forest]. The bars show the mean decrease in AUROC across the five test folds when the feature is permuted (20 permutations per feature per fold). Only features for which the mean importance plus its standard deviation across folds exceeded zero are shown. Feature names follow the source-system labels listed in Table 2.

demonstrated impressive capabilities for stroke prediction in PLWH using national-level data.

XGBoost and random forest models for the national cohort

The analysis of feature importance for stroke prediction in PLWH revealed key points between the Random Forest (Figure 4C) and XGBoost (Figure 4A) machine learning models. Both models achieved high predictive accuracy (AUCs of 0.94 and 0.93, respectively) and converged on the importance of certain clinical factors. Both algorithms identified hypertension, albumin levels, white blood cell count, and a history of transient ischemic attack as top predictors of stroke risk in PLWH. Furthermore, several other overlapping features, including hematocrit, age at HIV diagnosis, cholesterol levels, bilirubin levels, smoking status, and various blood chemistry markers, were consistently associated with stroke risk in this specific PLWH population.

The two models also exhibited differences in the specific features and their relative ranking. While both models acknowledged the importance of liver function, the XGBoost model placed greater importance on albumin levels, whereas the Random Forest model assigned more weight to bilirubin (another liver function marker). Also, HDL cholesterol (a lipid marker) was identified as a more influential predictor in the XGBoost model than in the Random Forest model, indicating variations in capturing the relationship between lipid metabolism and stroke risk in PLWH between these algorithms. Variations were also observed for blood chemistry markers, such as carbon dioxide levels and blood urea nitrogen levels. These differences are likely due to the algorithms' distinct approaches to handling feature interactions.

Figure 5A depicts the weighted average importance score for both models in the National dataset (for calculating, first the importance was scaled by each model's respective maximum score, then averaged over, and finally rescaled by the mean of the maximum scores of both models). Both models commonly focus on the following clinical features, in decreasing order of their importance: hypertension, white blood cell count, albumin levels, history of transient ischemic attack, monocyte percentage, HDL cholesterol [17, 20], hematocrit, liver function, eosinophil count, age at HIV diagnosis, smoking status, carbon dioxide levels, triglycerides, high cholesterol, basophil percentage, alkaline phosphatase serum, glucose, congestive heart failure, diabetes, and atrial fibrillation in PLWH for stroke prevention, in addition to other markers. Random Forest assigned negligible importance to platelet volume, serum creatinine, serum sodium, hemoglobin, and

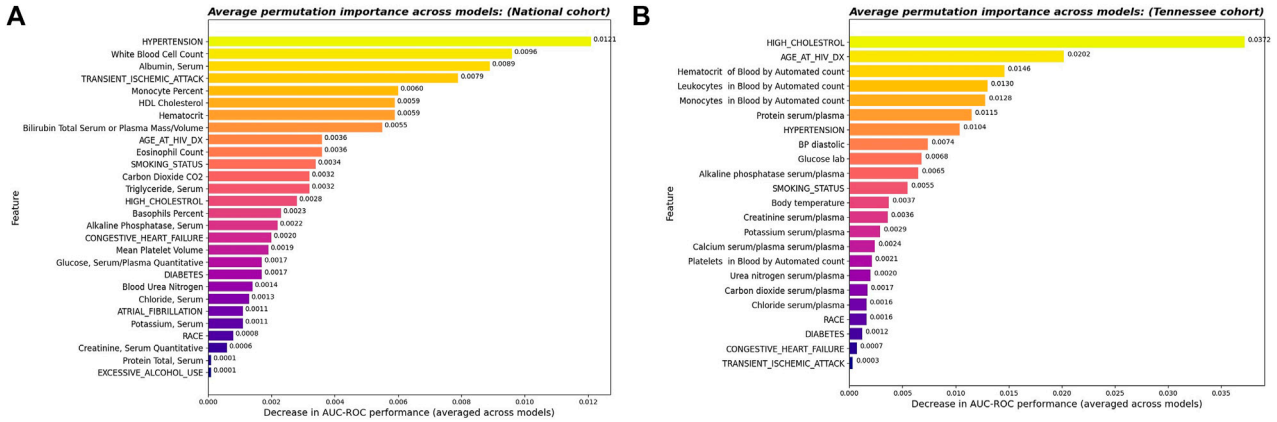


FIGURE 5
Average permutation importance of common features across different models [**(A)** National cohort, **(B)** Tennessee cohort]. For each feature, the permutation importance of each model was scaled by that model's maximum importance, averaged across the two models, and rescaled by the mean of the two maxima, as described in the Results.

gender, while XGBoost, apart from hemoglobin, serum sodium, and gender, also designated the following features as unimportant: excessive alcohol usage, blood urea nitrogen, and serum total protein. On average, the models placed less importance on race, creatinine, serum sodium and serum total protein, excessive alcohol usage, gender, and hemoglobin.

The performance of the Random Forest model in identifying strokes within PLWH in Tennessee was similar to that of the XGBoost model. The Random Forest model achieved a mean AUPR of 0.50 (Figure 3A), similar to the performance of XGBoost. This indicates that both models face similar challenges when predicting stroke cases within this specific dataset. Furthermore, the Random Forest model also attained a similar AUROC of 0.82 (Figure 3B) as XGBoost. The consistent AUROC and comparative AUPR values across both models highlight that the characteristics of the Tennessee dataset (including its low availability) are likely the primary drivers of the observed performance, rather than specific model limitations.

XGBoost and random forest models for the Tennessee cohort

When comparing the feature importance for stroke prediction in HIV patients between the XGBoost (Figure 4B) and Random Forest (Figure 4D) models within the Tennessee cohort, the significance of certain key factors was identified. The average importance scores of these models (in the same way as for the national dataset) are plotted in Figure 5B. High cholesterol emerged as a prominent predictor in both models, suggesting its consistent relevance in this specific population. Age at HIV diagnosis, hematocrit, leukocyte count, monocytes in blood, hypertension, glucose lab [20], serum protein, diastolic blood pressure, smoking status, alkaline phosphatase, serum urea nitrogen, serum potassium levels, race [17], diabetes [9, 17, 20], platelet counts, serum calcium [20], body temperature, and congestive heart failure, were other influential shared factors. These overlapping features consistently contributed to stroke prediction across the different algorithmic models within the Tennessee cohort. Features including serum sodium, past transient ischemic attacks, hemoglobin, eosinophil count, atrial fibrillation, excessive alcohol use, basophil counts in blood, gender, and total bilirubin serum levels were at the bottom of the two models, indicating their diminished/negligible contribution.

Performance comparison of the two models on the two cohorts is shown in Table 4. It can be seen from the table that the performance of the two models is comparable despite their different underlying algorithmic architecture. Hence, to isolate the most robust and consistent predictors of stroke in PLWH, a comprehensive comparison of the four feature importance charts (XGBoost and Random Forest for both National and Tennessee cohorts, as depicted in Figure 4) and their weighted average (depicted in Figure 5) was analyzed with a focus on identifying similarities across all models. The following features across both the National and Tennessee cohorts were predictive of stroke in PLWH: hypertension, high cholesterol, white blood cell count/leukocytes, age at HIV diagnosis, serum albumin, monocyte percent, hematocrit, past transient ischemic attack, HDL cholesterol, smoking status, serum glucose, bilirubin, eosinophil count, serum triglycerides, serum alkaline phosphatase, carbon dioxide, serum total protein, blood urea nitrogen, mean platelet volume, congestive heart failure, serum potassium, diabetes, diastolic BP, serum chloride, and race.

Logistic regression, deep neural networks, and the adapted Framingham Stroke Risk Profile underperformed the ensemble models in both cohorts

The penalized logistic regression was the weakest of the machine learning models in both cohorts, and the size of the shortfall differed sharply between them (Table 4). In the national cohort, it reached an AUPR of 0.23, approximately 5.2 times the random baseline of 4.4%, and an AUROC of 0.79, against compared with AUPR values of 0.81 and 0.79 and AUROC values of 0.93 and 0.94 for XGBoost and random forest, respectively. In the Tennessee cohort, it reached an AUPR of 0.18, approximately 2.3 times the random baseline of 7.8%,

and an AUROC of 0.70, against an AUPR of 0.50 and an AUROC of 0.82 for both ensembles. The gap was therefore wider nationally in absolute terms, where the ensembles benefit more from the larger cohort than the linear model does.

The feedforward neural network performed below both ensemble models in each cohort (Table 4). In the national cohort, it reached an AUPR of 0.71 (approximately 16 times the random baseline of 4.4%) and an AUROC of 0.91, against AUPR values of 0.81 and 0.79 and AUROC values of 0.93 and 0.94 for XGBoost and random forest. The shortfall was larger in AUPR than in AUROC, indicating that the network ranked the small number of stroke cases less precisely at the top of the risk distribution while separating the two classes almost as well overall. In the Tennessee cohort, the network achieved an AUPR of 0.40 (approximately 5.1 times the random baseline of 7.8%) and an AUROC of 0.76, against an AUPR of 0.50 and an AUROC of 0.82 for both XGBoost and random forest. Because interval estimates were not computed, these comparisons are descriptive, and we cannot infer a difference between model classes from them. A plausible contributor to the gap is that the tree ensembles route missing values natively during splitting, whereas the network receives mean-imputed inputs in which the pattern of missing laboratory tests is no longer represented. A further contributor in the Tennessee cohort is that the number of training epochs was selected on an internal validation split rather than on the evaluation fold, unlike the ensembles.

The adapted Framingham score performed well below both ensemble models in both cohorts (Table 4). In the Tennessee cohort, it achieved an AUROC of 0.64 and an AUPR of 0.13, approximately 1.7 times the stroke prevalence of 7.8% in the analyzed cohort, indicating modest discrimination and low precision. In the national cohort, it achieved an AUPR of 0.30, approximately 6.8 times the stroke prevalence of 4.4%, but an AUROC of 0.44, below the value of 0.5 expected from random classification. The divergence between the AUROC and AUPR values across the cohorts likely reflects differences in outcome prevalence, population characteristics, and data completeness. The weak performance is attributable to the substitutions required to apply the score, in particular the absence of systolic blood pressure in the national dataset and the use of age at HIV diagnosis in place of current age, and should not be read as a property of the original model. These results indicate that a general population stroke score does not transfer to PLWH EHR cohorts without recalibration and the incorporation of HIV-specific and contemporary clinical risk factors.

Permutation feature importance (Figures 4, 5) was evaluated for the two primary models, random forest and XGBoost; the three comparator models were included for performance comparison only and are reported in Table 4.

Biological interpretation of the top factors predictive of stroke in PLWH

Prominent blood cell counts

- White Blood Cell Count/Leukocyte/Basophils/Monocyte percentages: These factors consistently appeared as top predictors across all four models. This suggests that abnormalities in white blood cell levels carry information about stroke risk in PLWH. This relationship may be due to the increased production of inflammatory cytokines and other inflammatory markers associated with elevated white blood cell counts, potentially due to impacts on coagulation. Patients with white blood cell counts exceeding $9.7 \times 10^3 \mu L^{-1}$ (third tertile) face an over eightfold increase in in-hospital mortality risk compared to others. The rapid increase in cytokines and chemokines, in the case of ischemic stroke, leads to the recruitment of white blood cells to the affected brain tissue. Elevated white blood cell counts are key mediators of the inflammatory response and can cause endothelial damage. This damage can reduce the arterial wall's anti-thrombogenic properties, thereby increasing the risk of thrombosis [18, 21, 22, 23]. High basophil and neutrophil blood counts have both been identified as potential predictors of mortality in patients with stable coronary artery disease. This association may be linked to increased factor II plasma coagulant activity, suggesting a prothrombotic mechanism [21].
- Hematocrit: While its ranking varied slightly, hematocrit was consistently identified as an important predictor in all models. This points to the relevance of red blood cell volume in assessing stroke risk in PLWH. This has been observed by other studies in the past, albeit with lower relative importance than other traditional risk factors, and is consistent with previously reported risk factors for stroke [24, 25].
- Eosinophil count: The National cohort showed that eosinophil count was important for stroke prediction, as shown in Figure 5A. Since eosinophils are a type of white blood cell, their altered values could be associated with infection and medication therapy, including antiretroviral therapy, especially in the case of PLWH. Eosinophil-related markers have been associated with carotid plaque and incident stroke in the general population [26].

Traditional stroke risk factors

- Hypertension (Blood Pressure): Hypertension was a consistently strong predictor in the National cohort. In the Tennessee models, the importance of blood pressure was reinforced by the presence of "diastolic BP" [20]. It is well known that elevated blood pressure in hypertension induces heightened stress on the endothelium, which can impact the blood–brain barrier's integrity, enhancing permeability and subsequent local or multifocal brain edema. Damage to endothelial cells and altered interactions between blood cells and the endothelium promote the formation of local thrombi and can lead to ischemic lesions. Lacunar infarcts can be caused by fibrinoid necrosis due to occlusions and focal stenosis. Degenerative alterations in the endothelium and smooth muscle cells increase the risk of intracerebral hemorrhages. Moreover, hypertension accelerates arteriosclerosis, thereby heightening the risk of cerebral lesions associated with stenosis and embolism [9, 19, 20, 27].
- High cholesterol: Hyperlipidemia is associated with clotting risk and can predispose patients to strokes. Elevated levels of lipids or lipoproteins in the blood pose a significant stroke risk due to the promotion of atherosclerosis in both extracranial and intracranial blood vessels [9, 17, 19, 20, 28]. Hypercholesterolemia consistently appeared as an important

factor and was the most significant factor in the Tennessee cohort models, highlighting the need for lipid management.
- Smoking Status: Smoking is a known risk factor for stroke, and its consistent presence across the models underscores the importance of smoking cessation interventions for people with HIV. Numerous studies have confirmed the risk of stroke due to exposure to carbon monoxide and nicotine. Nicotine, tar, and carbon monoxide cause vascular endothelial damage, reduce oxygen transport, and increase blood viscosity, thus leading to increased stroke risk [9, 17, 19, 20, 29].

Organ function markers

- Liver Function (Albumin, Bilirubin, and Alkaline Phosphatase): Albumin was consistently significant in the National cohort models. Alkaline phosphatase serum levels appear across all four models, suggesting a potential link to stroke in PLWH. Recent research has found an association between liver function and risk of ischemic stroke, most likely due to increased systemic inflammation. Additional research into the link between these factors is warranted [30].

Age at HIV diagnosis

- Age at HIV diagnosis appeared as a significant predictor across all four models. This suggests that the timing of HIV infection and its long-term effects influence stroke risk. Since stroke is an age-dependent disease, with an increased incidence at higher ages, HIV, which also accelerates aging [31, 32], likely plays a significant role in stroke prediction.

Metabolic function

- Serum Glucose was associated with stroke, especially in the Tennessee cohort, as shown in Figure 5B. Serum Glucose is associated with diabetes and metabolic health, which can influence the inflammatory response and immune activation, especially in PLWH. This increase in systemic inflammation is likely responsible for the increased stroke risk in our model.
- Serum Triglyceride was associated with stroke prediction, especially in the National cohort, as shown in Figure 5A. Serum Triglyceride is linked to metabolic syndrome and, consequently, cardiovascular risk. Cardiovascular risk increases as patients age, especially in PLWH [33, 34].
- Serum total Protein is associated with stroke risk, especially in the Tennessee cohort, as shown in Figure 5B. Serum total protein is an indicator of overall nutritional and metabolic status, both of which are also compromised in aging populations [35].

Cardiovascular and systemic conditions

- Transient Ischemic Attack (TIA) (a precursor to stroke) was associated with stroke, especially in the National cohort, as shown in Figure 5A. A TIA is a recognized clinical antecedent of an ischemic stroke and is included in routine stroke risk assessments; because all of the predictors were recorded before the index date, its contribution reflects prior patient history rather than the index event.

Factors with less predictive value for strokes

In these models, atrial fibrillation, serum sodium, excessive alcohol use, hemoglobin, and gender received low permutation-importance ranks in both cohorts (Figure 5). Apart from these, features such as serum potassium, race, total protein serum, and creatinine were found to be less significant, especially in the national dataset. These factors were found to be more important in the Tennessee cohort, especially total serum protein, as shown in Figure 5B. Platelet count was found to be consistently in the bottom half of the feature importance graphs in both datasets (Figures 5A,B). In the case of the Tennessee cohort, features such as transient ischemic attack, eosinophils, basophils, and bilirubin did not contribute to the model, while they were very important in stroke prediction in PLWH in the National cohort. Previous research has suggested that a history of TIAs is associated with an elevated risk of stroke [36].

Diabetes is an important marker for stroke risk due to its accelerating impact on vascular aging [9, 19, 37]. However, in our analysis, diabetes was found to be in the bottom half of the feature importance figures (Figures 5A,B in both the National and Tennessee cohorts). This may be because diabetes is often associated with abnormalities in laboratory parameters, such as glucose levels, lipid profiles, and inflammatory markers. In our analysis, the ML models attribute more predictive power to these directly measured laboratory values rather than to clinical diabetes diagnoses. For example, elevated serum glucose levels may be a stronger predictor in the model than a diagnosis of diabetes, as shown in Figure 5.

It is well documented that atrial fibrillation and diabetes are strongly and significantly associated with stroke. The discrepancy between the univariable association and the low permutation importance rank arises because correlated features may absorb the signal. Permutation importance measures a feature's marginal contribution given the other features in the model. Permuting the diabetes variable (DIABETES in Table 2) leaves the serum glucose value intact, and the model recovers the same information from the correlated laboratory value; the same applies to atrial fibrillation relative to the hematologic and cardiovascular features. A low permutation-importance rank under feature correlation is evidence of marginal contribution, not clinical importance.

Discussion

Previous efforts to predict stroke risk among PLWH have been limited. This study applies machine learning to predict stroke in PLWH, who are at an elevated stroke risk, up to threefold higher incidence compared to HIV-negative individuals [38], and demonstrates the effectiveness of ensemble machine learning models, specifically Random Forest and XGBoost, for stroke prediction in this population using large-scale EHR data.

Overall, the use of ML and model interpretation analysis allows for the integration of diverse clinical features and complex nonlinear relationships [39, 40, 41, 42], while providing insights into key predictors of stroke in PLWH. Our ML-driven findings, consistent

with the literature [6, 19, 33], show that PLWH share many of the same key stroke determinants as the general population. However, HIV infection and its sequelae likely act as accelerants or additional stressors in the pathogenesis of stroke. This convergence of evidence supports the clinical relevance of our identified predictors and suggests that interventions proven effective in general stroke prevention (e.g., blood pressure, lipid, and inflammation management) are generally applicable to PLWH [43]. At the same time, the results emphasize the value of incorporating HIV-specific health measures into risk assessment models to capture the full spectrum of stroke risk factors in PLWH.

From a clinical standpoint, the predictors fall into two groups. Hypertension, hypercholesterolemia, smoking status, and serum glucose are established, modifiable stroke risk factors for which preventive management is well-defined, and their prominence across models supports paying attention to these factors in PLWH. Hematologic, hepatic, and renal laboratory values, such as white blood cell indices, albumin, and alkaline phosphatase, are best regarded as indicators of inflammation, illness burden, or organ function that contribute to risk stratification rather than as targets of intervention in themselves. Many of these values are likely to reflect comorbid conditions and the clinical circumstances that prompted testing, and no clinical action is implied by an isolated abnormal value, such as an elevated basophil count. The intended use of models of this kind is to identify PLWH at elevated risk who may benefit from closer evaluation and management of established risk factors, not to guide decisions about individual laboratory results.

Comparing the two cohorts, the regional (Tennessee) cohort is smaller and demographically distinct, resulting in machine learning models with lower performance. This shows how limited sample size and region-specific differences can affect predictive accuracy.

In summary, this study demonstrates that machine learning models trained on EHR data can stratify stroke risk in PLWH and identify candidate predictors, providing a first step toward personalized risk assessment and intervention to mitigate stroke risk among PLWH.

Limitations

Apart from age at HIV diagnosis, neither dataset contained HIV-specific variables such as CD4 count, viral load, ART regimen, or duration of infection. The models are therefore, in effect, general cardiovascular risk models applied to a PLWH population, and the mechanisms by which HIV contributes to stroke cannot be examined with these data. Incorporating immunological and virological measures, including nadir CD4 count, cumulative viremia, and ART class and duration [44] is essential for a genuinely HIV-specific stroke model.

The national and Tennessee cohorts were modeled separately. Although the two datasets differ in source, geography, and coding practice (Table 2), comparing models trained separately on each cohort does not constitute external validation in the formal sense, and the Tennessee case count (n = 199) limits the precision of any estimate derived from it.

Demographic characteristics, including race, gender, and age at HIV diagnosis, were incorporated into model development, and controls were matched to cases based on these characteristics. We therefore focused on overall model discrimination rather than estimating performance separately within demographic strata. Nevertheless, subgroup-specific performance was not formally evaluated, and the relatively small number of stroke cases in the Tennessee cohort limited the precision of such comparisons. Future studies with larger, independent cohorts should assess whether discrimination and calibration are consistent across demographic groups.

Finally, permutation importance measures the marginal contribution of a feature given the other features in the model and does not establish causation. The identified predictors are candidates for further investigation rather than established HIV-specific risk factors.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Data availability

The data analyzed in this study is subject to the following licenses/restrictions: The national dataset was extracted from the Cerner Health Facts database and the Tennessee dataset was sourced from the UTHSC Research Enterprise Data Warehouse. Requests to access these datasets should be directed to RD, rdavis88@uthsc.edu.

Ethics statement

The studies involving humans were approved by UTHSC Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

References

1. Moawad MHED, Mohamed Shalaby MM, Hamouda E, Mahfouz A, Mouffokes A, Hamouda H, et al. Risk of stroke among HIV patients: a systematic review and meta-analysis of global studies and associated comorbidities. *JAIDS J Acquired Immune Deficiency Syndromes* (2024) 95:399–410. doi: 10.1097/QAI.0000000000003382

2. Abissegue G, Yakubu SI, Ajay AS, Niyi-Odumosu F. A systematic review of the epidemiology and the public health implications of stroke in Sub-Saharan Africa. *J Stroke Cerebrovasc Dis* (2024) 33:107733. doi: 10.1016/j.jstrokecerebrovasdis.2024.107733

3. Du M, Wang Y, Qin C, Mi D, Liu M, Liu J. Prevalence and incidence of stroke among people with HIV. *AIDS* (2023) 37:1747–56. doi: 10.1097/QAD.0000000000003632

4. Wing EJ. The aging population with HIV infection. *Trans Am Clin Climatological Assoc* (2017) 128:131–44.

5. Yang Y, Yao X, Liu Y, Zhao J, Sun P, Zhang Y, et al. Global and regional estimate of HIV-associated stroke burden: a meta-analysis and population attributable modeling study. *Stroke* (2023) 54:2390–400. doi: 10.1161/STROKEAHA.123.043410

6. Benjamin LA, Bryer A, Emsley HCA, Khoo S, Solomon T, Connor MD. HIV infection and stroke: current perspectives and future directions. *The Lancet Neurol* (2012) 11:878–90. doi: 10.1016/S1474-4422(12)70205-3

7. Grout R, Gupta R, Bryant R, Elmahgoub MA, Li Y, Irfanullah K, et al. Predicting disease onset from electronic health records for population health management: a scalable and explainable deep learning approach. *Front Artif Intelligence* (2024) 6:1287541. doi: 10.3389/frai.2023.1287541

8. Yang Z, Mitra A, Liu W, Berlowitz D, Yu H. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat Commun* (2023) 14:7857. doi: 10.1038/s41467-023-43715-z

9. Chakraborty P, Bandyopadhyay A, Sahu PP, Burman A, Mallik S, Alsubaie N, et al. Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing. *BMC Bioinformatics* (2024) 25:329. doi: 10.1186/s12859-024-05866-8

10. Daidone M, Ferrantelli S, Tuttolomondo A. Machine learning applications in stroke medicine: advancements, challenges, and future prospectives. *Neural Regen Res* (2024) 19:769–73. doi: 10.4103/1673-5374.382228

11. Sau A, Barker J, Pastika L, Sieliwonczyk E, Patlatzoglou K, McGurk KA, et al. Artificial intelligence–enhanced electrocardiography for prediction of incident hypertension. *JAMA Cardiol* (2025) 10:214–23. doi: 10.1001/jamacardio.2024.4796

12. Yu Y, Xie Y, Thamm T, Gong E, Ouyang J, Huang C, et al. Use of deep learning to predict final ischemic stroke lesions from initial magnetic resonance imaging. *JAMA Netw Open* (2020) 3:e200772. doi: 10.1001/jamanetworkopen.2020.0772

13. Breiman L. Random forests. *Machine Learn* (2001) 45:5–32. doi: 10.1023/A:1010933404324

14. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Machine Learn Res* (2011) 12:2825–30.

15. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery (2016). p. 785–94. doi: 10.1145/2939672.2939785

16. D'Agostino RB, Wolf PA, Belanger AJ, Kannel WB. Stroke risk profile: adjustment for antihypertensive medication. The Framingham study. *Stroke* (1994) 25:40–3. doi: 10.1161/01.STR.25.1.40

17. Hong C, Pencina MJ, Wojdyla DM, Hall JL, Judd SE, Cary M, et al. Predictive accuracy of stroke risk prediction models across Black and white race, sex, and age groups. *JAMA* (2023) 329:306–17. doi: 10.1001/jama.2022.24683

18. Kazmierski R, Guzik P, Ambrosius W, Ciesielska A, Moskal J, Kozubski W. Predictive value of white blood cell count on admission for in-hospital mortality in acute stroke patients. *Clin Neurol Neurosurg* (2004) 107:38–43. doi: 10.1016/j.clineuro.2004.03.003

19. Boehme AK, Esenwa C, Elkind MSV. Stroke risk factors, genetics, and prevention. *Circ Res* (2017) 120:472–95. doi: 10.1161/CIRCRESAHA.116.308398

20. Vu T, Kokubo Y, Inoue M, Yamamoto M, Mohsen A, Martin-Morales A, et al. Machine learning approaches for stroke risk prediction: findings from the suita study. *J Cardiovasc Development Dis* (2024) 11:207. doi: 10.3390/jcdd11070207

21. Pizzolo F, Castagna A, Olivieri O, Girelli D, Friso S, Stefanoni F, et al. Basophil blood cell count is associated with enhanced factor II plasma coagulant activity and increased risk of mortality in patients with stable coronary artery disease: not only neutrophils as prognostic marker in ischemic heart disease. *J Am Heart Assoc* (2021) 10:e018243. doi: 10.1161/JAHA.120.018243

22. Hu Z-B, Lu Z-X, Zhu F, Jiang C-Q, Zhang W-S, Pan J, et al. Higher total white blood cell and neutrophil counts are associated with an increased risk of fatal stroke occurrence: the Guangzhou biobank cohort study. *BMC Neurol* (2021) 21:470. doi: 10.1186/s12883-021-02495-z

23. Pantoni L, Sarti C, Inzitari D. Cytokines and cell adhesion molecules in cerebral ischemia: experimental bases and therapeutic perspectives. *Arterioscler Thromb Vasc Biol* (1998) 18:503–13. doi: 10.1161/01.ATV.18.4.503

24. Zheng W, Chen Y-H, Sawan M. Longitudinal data to enhance dynamic stroke risk prediction. *Healthcare* (2022) 10:2134. doi: 10.3390/healthcare10112134

25. Tohgi H, Yamanouchi H, Murakami M, Kameyama M. Importance of the hematocrit as a risk factor in cerebral infarction. *Stroke* (1978) 9:369–74. doi: 10.1161/01.STR.9.4.369

26. Sundström J, Söderholm M, Borné Y, Nilsson J, Persson M, Östling G, et al. Eosinophil cationic protein, carotid plaque, and incidence of stroke. *Stroke* (2017) 48:2686–92. doi: 10.1161/STROKEAHA.117.018450

27. Johansson BB. Hypertension mechanisms causing stroke. *Clin Exp Pharmacol Physiol* (1999) 26:563–5. doi: 10.1046/j.1440-1681.1999.03081.x

28. Tirschwell DL, Smith NL, Heckbert SR, Lemaitre RN, Longstreth WT, Psaty BM. Association of cholesterol with stroke risk varies in stroke subtypes and patient subgroups. *Neurology* (2004) 63:1868–75. doi: 10.1212/01.WNL.0000144282.42222

29. Wang Y, Ge Y, Yan W, Wang L, Zhuang Z, He D. From smoke to stroke: quantifying the impact of smoking on stroke prevalence. *BMC Public Health* (2024) 24:2301. doi: 10.1186/s12889-024-19754-6

30. Ogunniyi KE, Alagbo HO, Tolu-Akinnawo OZ, Ibrahim S, Adeleke OD, Mahtani A, et al. Cirrhosis is an independent risk factor for mortality in ischemic stroke—a nationwide analysis. *Int J Hepatol* (2025) 2025:9250819. doi: 10.1155/ijh/9250819

31. Aung HL, Aghvinian M, Gouse H, Robbins RN, Brew BJ, Mao L, et al. Is there any evidence of premature, accentuated and accelerated aging effects on neurocognition in people living with HIV? A systematic review. *AIDS Behav* (2021) 25:917–60. doi: 10.1007/s10461-020-03053-3

32. Wing EJ. HIV and aging. *Int J Infect Dis* (2016) 53:61–8. doi: 10.1016/j.ijid.2016.10.004

33. Paternò Raddusa MS, Marino A, Celesia BM, Spampinato S, Giarratana C, Venanzi Rullo E, et al. Atherosclerosis and cardiovascular complications in people living with HIV: a focused review. *Infect Dis Rep* (2024) 16:846–63. doi: 10.3390/idr16050066

34. Palios J, Kadoglou NPE, Lampropoulos S. The pathophysiology of HIV-/HAART-related metabolic syndrome leading to cardiovascular disorders: the emerging role of adipokines. *Exp Diabetes Res* (2012) 2012:103063. doi: 10.1155/2012/103063

35. Tian C-R, Qian L, Shen X-Z, Li J-J, Wen J-T. Distribution of serum total protein in elderly Chinese. *PLoS ONE* (2014) 9:e101242. doi: 10.1371/journal.pone.0101242

36. Amarenco P, Lavallée PC, Monteiro Tavares L, Labreuche J, Albers GW, Abboud H, et al. Five-year risk of stroke after TIA or minor ischemic stroke. *New Engl J Med* (2018) 378:2182–90. doi: 10.1056/NEJMoa1802712

37. Hill MD. Stroke and diabetes mellitus. *Handbook Clin Neurol* (2014) 126:167–74. doi: 10.1016/B978-0-444-53480-4.00012-6

38. Ismael S, Moshahid Khan M, Kumar P, Kodidela S, Mirzahosseini G, Kumar S, et al. HIV associated risk factors for ischemic stroke and future perspectives. *Int J Mol Sci* (2020) 21:5306. doi: 10.3390/ijms21155306

39. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *New Engl J Med* (2017) 376:2507–9. doi: 10.1056/NEJMp1702071

40. Rose S. Machine learning for prediction in electronic health data. *JAMA Netw Open* (2018) 1:e181404. doi: 10.1001/jamanetworkopen.2018.1404

41. Orfanoudaki A, Chesley E, Cadisch C, Stein B, Nouh A, Alberts MJ, et al. Machine learning provides evidence that stroke risk is not linear: the non-linear Framingham stroke risk score. *PLoS ONE* (2020) 15:e0232414. doi: 10.1371/journal.pone.0232414

42. Gao Y, Cui Y. Clinical time-to-event prediction enhanced by incorporating compatible related outcomes. *PLOS Digital Health* (2022) 1:e0000038. doi: 10.1371/journal.pdig.0000038

43. Lembas A, Załęski A, Peller M, Mikuła T, Wiercińska-Drapało A. Human immunodeficiency virus as a risk factor for cardiovascular disease. *Cardiovasc Toxicol* (2024) 24:1–14. doi: 10.1007/s12012-023-09815-4

44. Nweke M, Mshunqane N. Characterization and stratification of risk factors of stroke in people living with HIV: a theory-informed systematic review. *BMC Cardiovasc Disord* (2025) 25:405. doi: 10.1186/s12872-025-04833-2