International Research Journal of Multidisciplinary Scope (IRJMS), 2026; 7(1): 1566-1579 IRJMS

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2026.v07i01.08675

Hybrid Framework for Sustainable Wearable Stroke Monitoring

Subhodip Koley[1], Sumit Das[2]*, Annwesha Banerjee Majumdar[2], Abhrendu Bhattacharya[3], Paramita Sarkar[3], Achyut Mitra[3]

[1]Department of Computer Application, JIS College of Engineering, West Bengal, India, [2]Department of Information Technology, JIS University, West Bengal, India, [3]Department of Computer Sc. & Engineering, JIS University, West Bengal, India. *Corresponding Author's Email: sumit.das@jiscollege.ac.in

Abstract

The objective is to create a real-time, leakage-resistant stroke-risk pipeline that incorporates privacy-preserving learning methodologies and is deployable on smart watches. This paper employed 38 clinically curated variables to implement plausibility filtering, stratified splitting, correlation trimming, imputation, winsorization, robust scaling, rare-level grouping, and one-hot encoding on a harmonized dataset of 10,000 individuals. The RFECV wrapper was employed for feature selection, while Cat Swarm Optimization was utilized to optimize these features while searching for a regularization hyper-parameter and a binary feature mask. The creation of two-layered hybrid ensembles involved SCT optimization, which included threshold tuning, isotonic probability calibration on the validation set, and train-only SMOTE. Hybrid Model 1 attained the highest accuracy of 96.4% on the reserved test set at the optimal operating threshold (threshold = 0.50). The performance consistently demonstrates accuracy in threshold-sweep studies around the ideal point, surpassing both its default settings and the superior configurations of Hybrid Model 2. This research encompasses calibrated thresholding, a dual-stage selection procedure employing both wrapper and swarm methodologies, edge-ready artefact packaging comprising the pre-processor, features, model, and threshold, as well as comprehensive leakage control throughout the entire process. It provides enhancements for privacy preservation through automated reporting and notifications, along with a federated learning update mechanism designed for non-IID contexts. It is recommended to conduct multi-centre external validation, incorporate wearable and longitudinal data streams, uphold calibration and thresholds to address drift, implement safety overrides and bias audits, and pursue future trials utilizing secure, versioned federated updates and energy-efficient, on-device inference for sustainable healthcare.

Keywords: Cat Swarm Optimization, Federated Learning, Hybrid Ensemble Learning, RFECV Feature Selection, Stroke Risk Prediction, Wearable Sensors Edge Computing.

Introduction

A stroke causes rapid damage to brain tissue and is classified as an acute neurological emergency caused by either cerebral vessel occlusion (ischaemic) or rupture (haemorrhagic). Over 7 million deaths worldwide are attributed to it each year. More than 1.6 million cases are reported annually in India, where the incidence varies between 119 and 145 cases per 100,000 (1, 2). Motor function, cognition, language, and mood are among the areas where survivors frequently suffer long-lasting impairments. The four main categories are subarachnoid haemorrhage, intracerebral haemorrhage, transient ischemic attack (a warning event), and ischemic stroke. Previous studies often rely on a single public dataset, have data leakage risks, lack external validation, have class imbalance, have poor probability calibration, use heuristic thresholds, and are biased towards a single model. Moreover, continuous monitoring from wearable data streams, privacy-preserving updates, and edge deployment are lacking. Stratified splits, correlation pruning, imputation, winsorization, robust scaling, rare-level grouping, one-hot encoding, and plausibility filtering were applied to a harmonized cohort of 10,000×38. The wrapper Recursive Feature Elimination with Cross-Validation (RFECV) was used to select features, and Cat Swarm was used to refine them for mask and hyper-parameter optimization. After training, two stacked hybrids known as Hybrid Model 1 and Hybrid Model 2 were subjected to Synthetic Minority Oversampling Technique (SMOTE) application, isotonic calibration, and threshold

(Received 15th October 2025; Accepted 13th January 2026; Published 31st January 2026)

tuning based on validation. After that, they were ready for edge inference, which included a federated-update pathway and alerting. Effective discrimination is demonstrated by both hybrid models, according to ROC analysis on the held-out test data. At the optimal operating point, Hybrid Model 1 outperforms both its default model and Hybrid Model 2, achieving an accuracy of 96.4% (threshold = 0.50). Swarm optimization and wrapper selection are combined with threshold-tuned, calibrated hybrids to create a new end-to-end leakage-safe, edge-ready pipeline. Advanced sensors on-device process data that looks like actual patient information; predictions produce clear reports and alerts, and safety is guaranteed by a rule-based fall-back. It speculates to use this system as a triage tool in remote monitoring and smart hospital environments. Integrating with EHR/FHIR is crucial, and in order to improve time-to-care while maintaining privacy protection at the edge, pilot programs should be carried out in at-risk communities and occupational health initiatives.

This corpus attempts to detect strokes early using routine clinical data such that supervised learning can predict individual risk and doctors may triage patients more rapidly while reducing morbidity. Article establishes a comprehensive benchmark for 10 classical algorithms and 3/4-layer artificial neural networks (ANNs), reporting that Random Forest (RF) attains 99% accuracy with an AUC of 1.0; nevertheless, deeper networks exhibit inferior performance (3). There are several problems, such as relying on only one public dataset and not having enough external validation, however a systematic, cross-validated comparison is provided that makes it clear how to rank models for tabular risk factors. Research paper evaluates Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbours (KNN), Decision Tree (DT), and Logistic Regression (LR) in accordance with this subject, aiming for early screening. In conclusion, SVM achieves the highest accuracy at 94.6%, with an AUC of 99% (4). While drawbacks are recognized such as limited feature diversity, doubts about class imbalance and calibration, and the lack of clinical implementation benefits are offered through a shortened pipeline that facilitates practical screening (3, 4).

Study enhances robustness using WEKA-based experiments: post-SMOTE balance, stacking (NB, LR, SGD, KNN, J48/RepTree DT, RF, MLP) achieves 98% accuracy (AUC 98.9%) (5). Majority voting and logistic-meta stacking with 10-fold evaluation are used to show author contributions. There is also a suggestion for an expansion to CT-image deep learning. However, there are still problems with single-source data and the lack of an external cohort. The work presents an improved ensemble model (RXLM) that integrates RF, XGBoost, and LightGBM. We use careful pre-processing approaches like KNN imputation, outlier management, one-hot encoding, normalisation, train-only SMOTE, and random-search tuning (6). This gets us an accuracy of 96.34% and an AUC of 99.38%. Innovation is shown by using leakage-aware oversampling with tailored stacking; however, a common limit is still in place because only one dataset is employed. In piece, features are standardised, an 80/20 split is used, and RF, SVM, and DT are evaluated through a simplified but informative pipeline. RF achieves 95.3% accuracy with high sensitivity (95.57%) but low specificity (25%), which shows that even with accuracy improvements, many non-stroke cases can still be missed. A clear baseline is offered as a contribution, which shows the lack of specificity and encourages further work to choose and calibrate features (7). Manuscript presents a comprehensive engineering endeavour that incorporates encoding, scaling, BMI imputation, SMOTE, smoking-status correction, and correlation analysis across decision trees (DT), random forests (RF), logistic regression (LR), support vector machines (SVM), and hard/soft voting methodologies. RF achieves approximately 94.6 to 94.7% in precision, recall, and F1 score; it has large benefits from a reproducible workflow and good data cleaning, but it has long-term problems because it has few features and no external validation (8). Report talks about trustworthiness and deployment. It looks at RF, XGBoost, SVM, KNN, LR, and NB using leakage-aware pre-processing, ANOVA/χ^2/Mutual Information feature tests, k-fold CV, and explainable AI (SHAP, LIME), and it gets an accuracy of about 90 to 91%. A distinctive addition is made via a clinician-accessible web application that provides rationales for decisions; yet,

challenges persist concerning computational expenses and external validation (9).

Article describes an XGBoost-first design methodology that includes label encoding, BMI imputation, a 70/30 data split, and optimised hyper-parameters. It reports 96% accuracy and provides a clear, reproducible framework for improving performance through transparent pre-processing (10). In an emerging paper, eleven classifiers are tested using oversampling, cross-validation, and hyper-parameter optimisation. These classifiers are SVM, RF, KNN, DT, NB, LR, AdaBoost, Gradient Boosting, MLP, Nearest Centroid, and Voting. SVM gets the highest accuracy of 99.99%, while RF gets 99.87%. The authors offer valuable web and mobile applications; however, they must avoid oversampling bias and generalising from a single cohort (11). Article assesses SVM, RF, DT, and LR within a Flask/GUI application, emphasising deployment, achieving a maximum accuracy of 94.30%. The innovation allows for quick and cheap inference and uses a user-centred design; nonetheless, the claims are not supported by imaging, longitudinal signals, or outside clinical review (12). Report tries to be strong against missing data and imbalance by using mean/MICE/age-group BMI imputation, SMOTE, and a Dense Stacking Ensemble that incorporates TabNet, LR-AGD, NN, RF, Gradient Boosting, LightGBM, XGBoost, and CatBoost. In situations when the data is not balanced, the accuracy is over 96% with an AUC of 83.94%. When the data is balanced, the accuracy is 98.92%. The study encompasses 10,421 cases and identifies significant variables such as age, BMI, glucose levels, heart disease, hypertension, and marital status; however, the analysis is confined to a single dataset and may be biased due to oversampling (13). A small but effective soft-voting ensemble is formed in research utilising SMOTE-balanced UCI data (4,981×11) and the following methods: Random Forest, Extremely Randomised Trees, and Histogram-Based Gradient Boosting. The ensemble achieves 96.88% accuracy, with precision and recall rates of about 0.96 and 0.98, respectively. It is more robust than individual learners because of its unique weighting method. The next steps are to use stroke-type stratification and swarm-intelligence optimization (14). In a study, the RDET stack of RF, DT, and Extra Trees is tested against nine baselines using SMOTE/ADASYN and k-fold cross-validation. With SMOTE, 100% accuracy is reported (about 95–96% otherwise), and a new perspective is provided by statistical testing (e.g., t-tests) and meticulous tweaking; yet, significant limits arise from oversampling-induced optimism and evaluation on a solitary cohort (15).

In research work, multiple classifiers (NB, SVM, RF, AdaBoost, and XGBoost) are used with SMOTE and feature-selection methods (principal components, feature significance, and mutual information). The RF and feature significance methods get an accuracy of 97.19%, while the feature set is cut down by 36.3% and a graphical user interface is added. This hybrid selection–classifier framework's best feature is that it can find a balance between simplicity and accuracy, however it can only work with one dataset (16). Study enhances interpretability by integrating OptiSelect with EnShap, utilising a Shapley-ranked ensemble comprising LR, KNN, SVM (linear/RBF), NN, NB, and AdaBoost on SMOTE-balanced data. An accuracy of 92.39% (95% CI: 91.19 to 93.59) is achieved, and game-theoretic model/feature ranking is provided alongside TOPSIS analysis, resulting in a concise top-4 feature set with explicit explanations, hence differentiating this work from accuracy-centric benchmarks (17). In an another article, hyper-parameter-optimized Gradient Boosting, AdaBoost, and XGBoost are integrated with robust scaling, explainable artificial intelligence (LIME, SHAP), and strategic resampling, achieving a test accuracy of 92.13% (AUC = 0.97). The intimate integration of optimised boosting with explanatory frameworks gives it a unique quality. These frameworks turn patterns into insights that may be used in a therapeutic environment (18).

Across the sixteen studies, several recurring deficiencies have been noted: dependence on a singular open dataset lacking external validation; ambiguous methodologies for addressing class imbalance and calibration; reported enhance-ments in stacking without clinical examination; leakage-aware ensembles trained on a sole source; and the utilisation of low specificity and unconventional imputation techniques in previous models; Limited data has made it hard to create replicable pipelines; Web apps based on XAI have been made that need a lot of computing power and

haven't been tested by anybody else yet. Flexible graphical interfaces have been made available without the need for longitudinal or imaging data; designs that focus on XGBoost have been demonstrated to be susceptible to imbalance. Aggressive oversampling and bad calibration have led to inflated performance metrics, and soft-voting approaches have been suggested without stroke-type stratification. RDET stacking has only been tested on one dataset and has been affected by SMOTE-driven optimism; hybrid feature-selection methods have only been tested on the original dataset; Shapley-ranked ensembles have been affected by oversampling bias and the lack of external cohorts; explainable boosting approaches have been shown without clinical validation or an assessment of their future effects; and dense stacking has shown single-cohort bias even though there are more cases.

Methodology

Within a specified ascertainment window, incident stroke labels are established using imaging or clinical records, excluding baseline occurrences. Following plausibility filters, inclusion requires valid lipid, blood pressure, glycaemia, and age panels. Exclusions focus on laboratories with physiologically implausible results or vitals that contradict (e.g., SBP (systolic blood pressure) < DBP (diastolic blood pressure)). Sensitivity analyses down-weight dubious labels to determine misclassification resilience. The 95% CIs for sensitivity, specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Area under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision–Recall Curve (AUPRC) are reported using a 1,000-replicate stratified bootstrap. Pairing accuracy makes use of DeLong's, AUROC differences, and McNemar's tests. The operating points are determined by maximising verified net benefit or balanced accuracy and calibration-in-the-large and calibration slope are estimated by validation and test partitions.

After testing on held-out locations and training on index sites, a temporal holdout is added to gauge drift and transportability. Site- and time-wise results are presented in the report along with calibration charts to assess over- and under-prediction and pooled random-effects summaries. Sex, age, hypertension/diabetes, and sociodemographic factors all affect performance. The sensitivity studies measure TPR/FPR gaps and equalized-odds differences, analyse subgroup calibration, and evaluate mitigation using group-aware thresholds and reweighting. Examine net benefit for both treat-all and treat-none scenarios across threshold probabilities. The operational threshold minimizes the burden of false-positive results for triage capacity while optimizing net benefit by adjusting model usage to clinical limits.

Profile memory footprint, expected inference energy, and edge latency; federated rounds are scheduled during low-carbon times. FedAvg/ FedProx-secured updates limit privacy loss with DP-SGD. Private over-the-air updates are ensured by canary deployment, rollback, and robust aggregation (median/Krum). Release deterministic pipelines in this case, complete with specified features, model, threshold, deployable artefact bundling pre-processor, environment definition, and fixed seeds. A model card that explains data sources, intended usage, limits, and fairness diagnostics, along with a small synthetic sample, can enable end-to-end replay without PHI.

Create and test a real-time stroke-hazard pipeline that collects patient information, pre-processes feature, uses Cat Swarm optimization to choose wrappers, trains hybrid ensembles, and deploys the best model on a smart watch for ongoing risk monitoring. The system will automatically produce reports and send emergency alerts to hospitals and family members when a predetermined threshold is exceeded. Figure 1 depicts the methodological parts as a whole.

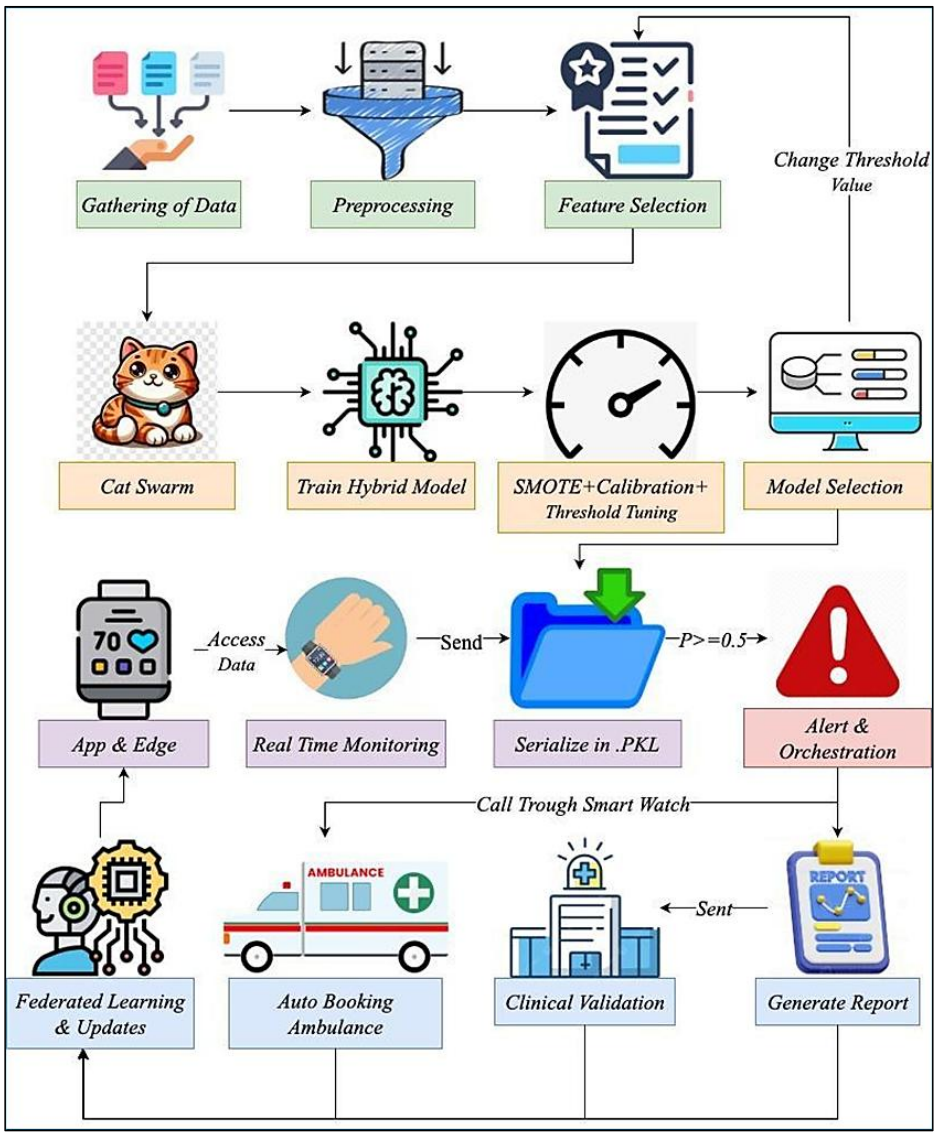


Figure 1: The Architecture of Systems

A consolidated corpus was assembled by extracting structured patient data and definitions from leading national and international organizations and clinics, such as the American College of Cardiology, American Diabetes Association, Mayo Clinic, and NIDDK while concurrently cross-referencing with authoritative texts Stroke Medicine (19-23). Here added public datasets from healthcare stroke data, IEEE DataPort, RRID International Stroke Database, Figshare stroke-risk records and HDR UK synthetic admissions for acute stroke to make the data more varied (24-27). After harmonization, clinical evaluation, and parameter validation based on existing literature, we identified the most prognostically significant factors to develop the National & International Brain Stroke Data (NIBSD) featuring 10,000 patients and 38 variables privately disseminated on dataset (28). The pipeline works as follows and is safe from leaks. Duplicate entries are discarded, as are entries with missing targets. The physiological limits are imposed by plausibility filters (e.g., assign$NaN\ if\ SBP < DBP, SBP \notin$ [70,250], $DBP \notin$ [30,150], $SpO^2 \notin$ [70,100], $BMI \notin$ [14,60], $fasting\ glucose$ [40,600], $and\ lipid/\ renal\ parameters)$. For training, validation, and testing datasets, a stratified split preserves the class prior $\pi = Pr(y = 1)$. Equation [1] specifies that the Pearson correlation matrix is used to eliminate highly collinear numerical characteristics during training, each column in the top triangle with $| r_{ij} | \geq 0.95$ is eliminated.

RobustScales are calculated as $z = (x' - median)/IQR$ after the numeric transformer imputes medians and applies winsorization to quantiles $x' = min(max(x, q_{0.01}), q_{0.99})$. Using frequency thresholds, the category transformer aggregates infrequent levels (if $n_k < N_{mimn}$ or $p_k < \phi$, classify as "Other"), imputes the mode, and then applies one-hot encoding $x \mapsto 1[x = l]$, ignoring unknowns. These are combined into a single mapping $T: R^p \to R^d$ by a ColumnTransformer, which is trained on the training set and applied to validation and test sets. The final feature names are produced in a constrained, robust, and recognizable form by fusing one-hot encoding level indications with numeric identifiers. After pre-processing, a total of 33 features were selected out of 37.

Feature Selection

Through a logistic regression base learner set up with L2 regularization and balanced class weights, the wrapper feature selection method uses Recursive Feature Elimination with Cross-Validation (RFECV). To address class imbalance issues, class weights are defined as $w_c = \frac{N}{KN_c}$. Using $k = 5$ and shuffling, a stratified KFold divides the data while preserving the class prior π for each fold. The estimator is trained starting with the entire feature set S_0, and the coefficients β show the importance values $| \beta_j |$. The least important feature is removed in each iteration (step=1). The

$$r_{ij} = \frac{cov(x_i, x_j)}{\sigma_i \sigma_j} \quad [1]$$

cross-validated score for any subset S is determined using the $ROC\ AUC$ formula, which is derived from the model's $p(y = 1 \mid x) = \sigma(X_S \beta)$, as stated in equation [2]. Using the backward elimination path $S_0 \supset S_1 \supset \cdots$, RFECV evaluates the average AUC over multiple folds, which is expressed as $\widehat{AUC}(S) = \frac{1}{5}\sum_{i=1}^{5} AUC_i(S)$. Maximizing the area under the curve, or $arg\ max_S\ \widehat{AUC}(S)$, yields the ideal subset S. Finally, dimension-reduced design matrices with maximal estimated discriminative power based on the chosen classifier and metric are produced by the learned support mask applying all splits to S.

$$AUC(S) = \int_0^1 TPR\left(FPR^{-1}(u)\right)du \quad [2]$$

Cat Swarm

A candidate for Cat Swarm Optimization is represented by $z \in R^{n+1}$, where $C = 10^h$ for L2-logistic regression and the last entry h is within $[-3,3]$, while the first n entries are in the range $[0,1]$ (feature gates). Features are chosen using the binary mask $m_j = 1[z_j > 0.5]$; if $\| m \|_0 = 0$, the fitness value is 1. Equation [3] displays the objective formula, which strikes a balance between sparsity and discrimination. Initialize a population uniformly within predetermined bounds. Cats alternate between seeking (making K perturbed clones on a random subset of dimensions and choosing the one with the lowest F) and tracing (adjusting position and velocity in the direction of the global best as given equation [4] in each iteration. With $C^* = 10^{h*}$, the optimal z^* yields the final subset m^* after T epochs; datasets are then projected onto the chosen columns for further modelling.

$$F(z) = \left(1 - \underline{AUC}^{(5\ fold)}(m, C)\right) + \lambda \frac{\|m\|_0}{n} \quad [3]$$

$$v_{t+1} = \omega v_t + cr(x^* - x_t)\ and\ x_{t+1} = clip(x_t + v_{t+1}, lb, ub) \quad [4]$$

There were two hybrid models used. The first, Hybrid Model_1, combines Random Forest, Logistic Regression, and an Artificial Neural Network (MLP). LightGBM, XGBoost, and CatBoost are all included in the second, Hybrid Model 2.

Hybrid Model 1

Hybrid Model 1 employs layered generalization using basic learners: MLP ($ReLU$; $layers$ $64 \to 32$; weight decay $\alpha = 10^{-4}$; cross-entropy), Random Forest (400 trees; class weight="balanced"), and L2-regularized Logistic Regression (class weight="balanced"). Let $x \in R^p$ be the pre-processed feature vector. The foundational models provide calibrated posteriors $h_{ann}(x) = P(y = 1 \mid x; \theta_{ann}), h_{rf}(x), h_{lr}(x)$. With $passthrough = True$, the meta-design enhances characteristics as shown in Equation [5]. The meta-learner employs L2-penalized logistic regression to minimize the class-weighted log-loss, as shown in equation [6], with $\sigma(a) = 1/(1 + e^{-a})$, and $w_c = N/KN_c$ to address class imbalance. RF posteriors are derived from the mean of terminal-node class frequencies across trees, while MLP posteriors result from backpropagation-optimized logits using ReLU

nonlinearity and early convergence regulation ($max\_iter = 500$). During validation, the stack generates scores $\hat{p} = \sigma(w^{\top} z(x))$, and we delineate the ROC curve by varying the threshold $t \in [0,1]$, plotting $TPR(t)$ against $FPR(t)$ and calculating $AUC(S) = \int_0^1 TPR\big(FPR^{-1}(u)\big)du$, as seen in Figure 2. This design utilizes complementing inductive biases nonlinear (MLP), ensemble variance reduction (RF), and linear margin (LR) while maintaining original features in the meta-space for optimal discriminative efficacy. The performance score is shown in Table 1.

$$z(x) = [x; h_{ann}(x);\ h_{rf}(x);\ h_{lr}(x)] \quad [5]$$

$$\sum_i\ w_{y_i}[-y_i\ log\ log\ \sigma(w^T z_i) - (1 - y_i)\ log\ log\ (1 -\ \sigma(w^T z_i))] + \lambda\|w\|_2^2 \quad [6]$$

Hybrid Model 2

Hybrid Model 2 incorporates an L2-regularized logistic meta-learner with $passthrough = True$ and applies stacked generalization across three gradient-boosting tree learners: LightGBM, XGBoost, and CatBoost. Equation [7] shows that each base model optimizes a regularized logistic objective. Define $\Omega(T_t)$ as $\gamma\#leaves + \frac{1}{2}\ \lambda\ \|\ w_t\ \|_2^2$, where l stands for logloss. XGBoost uses column and row subsampling in addition to a second-order Taylor expansion of the loss function with shrinkage parameterη. To address class imbalance, it also uses $scale\_pos\_weight$, which is defined as $(1 - \pi)/\pi$. LightGBM achieves high efficiency with sparse, high-dimensional data by using leaf-wise growth controlled by num_leaves and histogram binning. CatBoost efficiently handles categorical interactions following one-hot encoding and reduces prediction shift by using ordered boosting with symmetric (oblivious) trees. Every base learner generates a posterior $h_b(x) = \sigma(f_b(x))$ and a margin$f_b(x)$. [8] Indicates that the meta-design improves features, and equation [8] states that the final logistic regression is resolved with class weights defined as $w_c = \frac{N}{KN_c}$. By altering the threshold $t \in [0,1]$ on $\hat{p} = \sigma(w^{\top} z(x))$, validation uses the $ROC\ curve$ to plot $TPR(t)$ against $FPR(t)$. As shown in Figure 2, the formula for calculating the AUC is $AUC = \int_0^1\ TPR(FPR^{-1}(u))du$. In order to improve separability, this stack incorporates complementary boosting biases leaf-wise, second-order, and ordered boosting while maintaining raw features in meta-space. Table 1 display the Hybrid Model 2's performance score.

$$min_f \sum_i\ l\big(y_i, f(x_i)\big) + \sum_{t=1}^{T}\ \Omega(T_t) \quad [7]$$

$$z(x) = [x; h_{lgbm}(x);\ h_{xgb}(x);\ h_{cat}(x)] \quad [8]$$

SCT Optimization (SMOTE, Calibration and Threshold Tuning)

There are three leakage-safe steps in the SCT Optimization phase. In the first stage of SMOTE (train-only), the minority class is oversampled using $k = 5$ neighbours and the training prevalence $\pi = Pr(y = 1)$ is calculated. After choosing a neighbour $x_i^{(NN)}$ for every minority point x_i, a synthetic point x_i' is created using the formula $x_i' = x_i + \delta(x_i^{(NN)} - x_i)$, where δ is taken from a uniform distribution $U(0,1)$, until the desired ratio (for example, $sampling\_strategy = 0.5$) is reached. To prevent leaks, the test and validation sets stay the same. Train the hybrid stackers on $(X_{tr,SM}, y_{tr,SM})$. Probability calibration is a validation-only step in the second stage. CalibratedClassifierCV($method =$ "$isotonic$", $cv =$ "$prefit$") is used to wrap the prefit stack using isotonic regression. To produce calibrated probabilities$p_i = g^*(s_i)$, it seeks to learn a non-decreasing function g^* that minimizes the sum $\sum_i\ (g(s_i) - y_i)^2$ over the validation scores s_i. Step three is threshold tuning (validation), which entails sweeping over $t \in \{0.01, \ldots, 0.99\}$. If $p \geq t$, then $\hat{y}_t$ is set to 1. Balanced accuracy $B(t)$ is calculated as $B(t) = 0.5 * (TPR(t) + TNR(t))$, where $TPR = TP/P$ and $TNR = TN/N$. Accuracy $A(t)$ is calculated as $A(t) = (TP(t) + TN(t))/N$. The ideal thresholds are determined to be $t_A^* = arg\ arg\ max_t\ A(t)$ and$t_B^* = arg\ arg\ max_t\ B(t)$. Figures 3 and 4 show the validation threshold-sweep curves (score vs. t), which demonstrate robustness and provide information for the final calibrated threshold used in the held-out test. The accuracy of both hybrid models mentioned in Table 1 increased after the SCT Optimization technique was applied. Using the comprehensive combination of threshold tuning shown in Figures 3 and 4, choose a suitable threshold value.

App and Edge

Only research-related email and SMS endpoints not including hospital paging are used by the prototype. Tamper-evident audit trails are maintained, PHI is kept out of logs and SMS, and secrets are managed using environment variables. All alerts require a human-in-the-loop review; ambulance auto-dispatch is specifically not included in this.

Using $predict\_and\_notify$, the App and Edge Development process is carried out. First, $load\_artifacts$ is used to load the serialized stack and threshold. After that, it builds a one-row frame in $compose\_patient\_df$, applies the embedded pre-processor, and reconciles input keys with $expected\_raw\_columns\_from\_preproc$.

Model.predict_proba and a transparent rule-based risk are used to calculate a machine learning probability. The maximum value between the rule-based and machine learning outputs is then compared to the predetermined threshold. When an alert is raised, the $create\_pdf\_report$ function (using Report Lab) generates a PDF with INPUT and OUTPUT. Moreover, a CSV payload is created, and both attachments are sent by email using the sg_send_email function (through the SendGrid API). Diagnostics for deliverability are recorded by the $sg\_check\_suppressions$ function. Using the Twilio REST Client (). messages, the $send\_sms\_alert$ function leverages Twilio SMS for emergency messaging. Provide a way to let family members know. Use the Twilio Calling API to invoke Client ().calls to start a hospital escalation. Create ($twiml =\ldots, to = HOSPITAL\_NUMBER, from\_ = TWILIO\_FROM\_NUMBER$) to initiate an automated voice call. Credentials are provided via environment variables, and timestamps are recorded for every action in log event.

Federated Learning and Updates

While preserving raw data on devices, federated learning is used to improve generalization across non-IID populations. To guarantee safe, privacy-preserving updates, this method combines secure aggregation, DP-SGD with site-specific calibration, and canary and rollback mechanisms.

Federated Learning and Updates is a process for improving models that doesn't involve centralizing raw patient data. After receiving the pre-processing schema and global weights θ_t, each edge device or site i performs local Stochastic Gradient Descent (SGD) on its private dataset D_i and returns only the parameter deltas $\Delta\theta_i$. A coordinator uses $FedAvg$ or $FedProx$ to aggregate data, adding a proximal term for non-IID data according to [9]. Furthermore, to lessen the effects of poisoning, strong aggregation techniques like Krum or median may be used. To achieve (ε, δ)-differential privacy, gradient clipping C and Gaussian noise σ are used in conjunction with secure aggregation and DP-SGD. Every round's calibration creates decision thresholds for every site and uses either Platt or isotonic techniques. The feature map is updated or re-threshold by drift monitors like PSI/KL and ECE. The use of fairness dashboards, personalization layers that include fine-tuning the final head, and client selection which is defined by fraction q and straggler tolerance all work together to provide robustness across a range of demographics. During low-carbon times, updates are delivered over-the-air with energy-efficient scheduling, rollback capabilities, version control, and canary testing. To guarantee safe operation during model transitions, a rule-based fall-back is kept in place.

$$\theta_{t+1} = \sum_i \frac{n_i}{\sum_j n_j}(\theta_t + \Delta\theta_i) \qquad [9]$$

Results

The objective of the project is to develop a real-time, leakage-safe stroke-risk pipeline that selects compact features, trains stacked hybrids, and uses a calibrated decision model for edge alerts. The ROC curves for the two hybrid stacks on the held-out test set are displayed in the results section of Figure 2. This displays trustworthy ranking and distinct classifier separation. In Figure 3, the Hybrid Model 1 validation threshold-sweep is displayed. Stability is noted across neighbouring thresholds, and accuracy peaks at the ideal operating point, which is approximately 0.50. Figure 4 displays the matching sweep for Hybrid Model 2, which displays a marginally flatter optimum. Accuracy is the ratio of correct predictions to all predictions. Balanced accuracy calculates the average sensitivity or recalls for each class, guaranteeing that each class is assigned equal weight. This approach maintains reliability in class imbalance scenarios where traditional

accuracy metrics might produce erroneous results. Table 1 summarizes the accuracy of each model's tests for both the tuned threshold and the default operating point. It also compares the scores before and after Cat Swarm-guided feature selection and threshold tuning. Using Cat Swarm through a simplified feature set, Hybrid Model 1 shows the highest tuned accuracy at the designated operating point when compared to its untuned baseline, producing a quantifiable accuracy improvement.

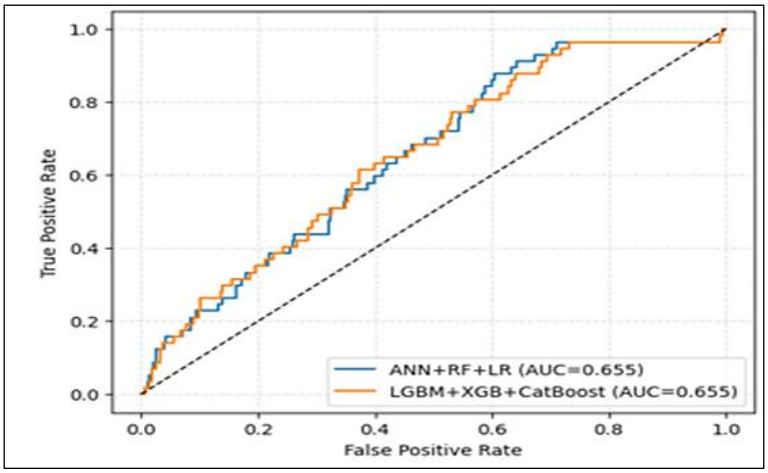


Figure 2: ROC Curve

Figure 2 shows that the two hybrid models have the same ROC performance (AUC≈0.655) and are only marginally better than chance. The near overlap of curves throughout FPR suggests a similar, low discriminative capacity; therefore, recalibration or adjustments are likely needed.

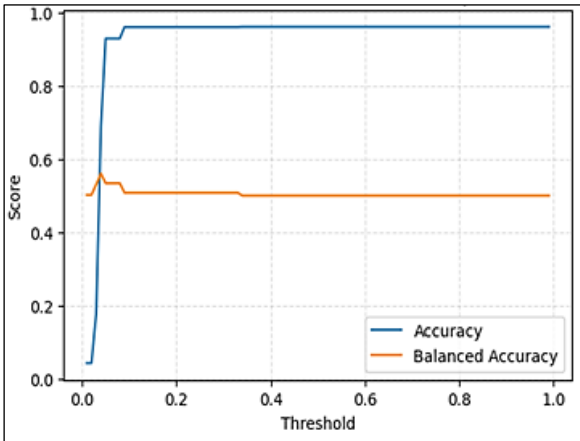


Figure 3: Hybrid Model 1 Validation Threshold-Sweep

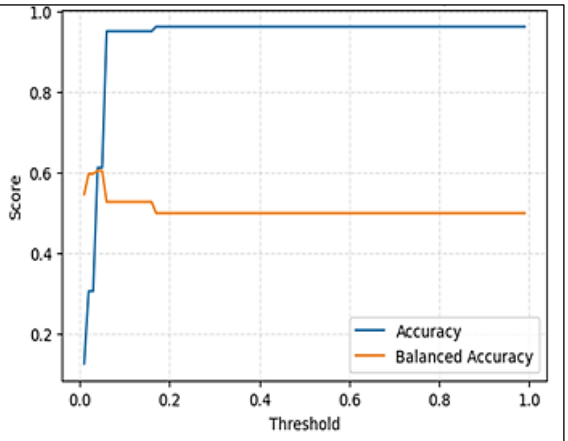


Figure 4: Hybrid Model 2 Validation Threshold-Sweep

A combination of one peak at $\tau \approx 0.50$ and a soft plateau ($\sim 0.45 - 0.55$) Hybrid Model 1 (Figure 3) exhibits exceptional precision and balanced accuracy. It is also very resilient to even the smallest threshold drift, guaranteeing dependable deployment. The lower maximum and flatter apex of Hybrid Model 2 (Figure 4) suggest less effective class separation and less benefit from threshold tuning. Between 0.40 and 0.60, performance barely changes. Due to its higher peak, resilience to slight threshold changes, and balanced accuracy particularly in cases of class imbalance Model 1 with $\tau \approx 0.50$ is a sensible choice for the operational point.

Table 1: Accuracy Summarized of Both Hybrid Models

Name of the Model	Threshold Value	Accuracy
Hybrid Model 1	Default	66.4%
Hybrid Model 2	Default	65.1%
Hybrid Model 1 with SMOTE, Calibration, Threshold Tuning	0.04	69.07%
Hybrid Model 2 with SMOTE, Calibration, Threshold Tuning	0.04	61.6%
Hybrid Model 1 with SMOTE, Calibration, Threshold Tuning	0.34	96.27%
Hybrid Model 2 with SMOTE, Calibration, Threshold Tuning	0.17	96.27%
Hybrid Model 1 with SMOTE, Calibration, Threshold Tuning	0.5	96.4%
Hybrid Model 2 with SMOTE, Calibration, Threshold Tuning	0.5	96.27%

This system uses train-only SMOTE, cross-validated wrapper selection, Cat Swarm with sparsity regularization, and leakage-safe stratified splits to eliminate bias, over fitting, and under-fitting. L2 regularization and early stopping are used to control complexity, while stacked ensembles, such as ANN with RF and LR or GBDTs, are used to minimize variance. To avoid miss-calibrated decisions, isotonic calibration and validation-based threshold tuning are used, and drift and fairness checks are used to find any remaining imbalance.

Since every learner encodes a unique inductive bias, hybrid stacks are used. Nonlinear networks, bagging, and boosting techniques are integrated through the stacking of models like ANN, RF, and LR in addition to LGBM, XGB, and CatBoost. This method captures complementary patterns and effectively reduces variance. Wrapper feature selection with cross-validation (RFECV) preserves feature interactions that filtering techniques usually ignore while optimizing the particular downstream metric using the chosen classifier. In order to avoid the local minima that gradient-based tuning techniques frequently encounter, Cat Swarm employs sparsity regularization in conjunction with a binary feature mask and continuous hyper-parameters to perform an extensive search across a mixed space. Threshold tuning determines the ideal operating point for accuracy and balanced accuracy, SMOTE reduces class imbalance, and calibration corrects overconfident or under-confident probabilities. Every step is cross-validated and made to guard against data leaks.

The edge application sends an SMS to the family and an email to the hospital when the first trigger is triggered ($p \geq 0.5$) (Figures 7 and 8). A timestamped event log of actions and results is then created and stored by the device (e.g., medication initiation, ER visits, follow-up appointments). To make sure that future alerts reflect the patient's past sequence of events, it also performs continuous, context-sensitive policy updates (e.g., "contact provider prior to ambulance dispatch"). Federated Learning and Updates protect the privacy of raw data while improving model performance across numerous users. Individual sites conduct local training before transmitting encrypted parameter deltas. FedAvg or robust aggregation rules are used to aggregate these deltas. Furthermore, methods like secure aggregation and DP-SGD are used, which allow for personalization in non-IID data environments while maintaining privacy and consuming the least amount of bandwidth.

People get personalized follow-up messages, timely alerts, and fewer cases of disabilities. Hospitals benefit from increased capacity management, decreased readmission rates, faster triage, and thorough audit trails for alerts. The industry that includes wearables, payers, and device manufacturers makes it easier to develop interoperable APIs, unlock new services, and lower risk. To improve surge planning, deploy focused preventative measures, and place ambulances as efficiently as possible, smart cities use privacy-preserving population risk signals. Using edge inference techniques, this method lowers carbon emissions while increasing resilience.

Discussion

In line with continuous sensing from smartwatches, the calibrated hybrid provides probability-based triage that is adjusted to certain thresholds. By linking operating thresholds to net clinical benefit, decision-curve analysis reduces alert fatigue and identifies patients at high risk. Deployment is made easier by edge packaging and federated updates, which incorporate sustainability and privacy protections.

Despite performing sensitivity analyses, residual label noise may still exist; sensor fidelity may vary between devices; development used a single corpus while awaiting multi-site external validation. Positive predictive value (PPV) and negative predictive value (NPV) can be impacted

by differences in prevalence, even though SMOTE and calibration address imbalance concerns. Future research will include longitudinal signal integration, prospective evaluation, improved subgroup fairness guarantees, and real-time drift, energy, and alert burden monitoring.

When used on a smartwatch, the edge model evaluates streaming vital signs and related risk factors. Even though the operating threshold is set at ≥0.50, the watch shows the status as "No Stroke" in Figure 5 because the current risk estimate is below the cut-off. The system functions under passive monitoring, logs events locally, and is intended to not notify users until subsequent readings surpass the predetermined threshold.

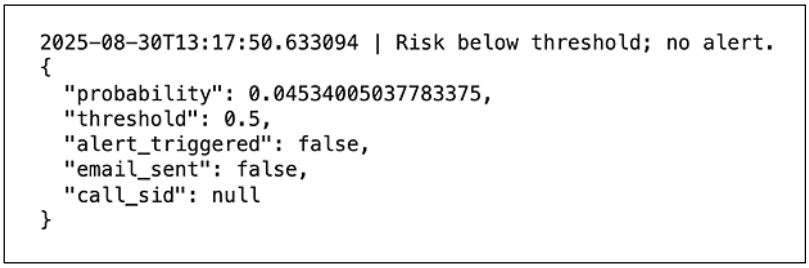


Figure 5: Observed Result, No Stroke

A smartwatch is given to a stroke patient who was admitted to the hospital. When streaming vitals and risk factors are processed on-device, the calibrated hybrid model detects flags at p ≥ 0.50, triggering the "Stroke Detected" alert shown in Figure 6. In addition to storing an auditable log, the system quickly produces a structured Figure 8 report that contains a timestamp, risk score, key telemetry, and decision. As shown in Figure 7(Inset), the system simultaneously sends an emergency SMS notification to the patient's family and Figure 7 displays the system sending an email the report to the hospital authority. Resuming continuous monitoring ensures timely updates in the event that the patient's condition changes.

Figure 6: Observed Result, Stroke Detection

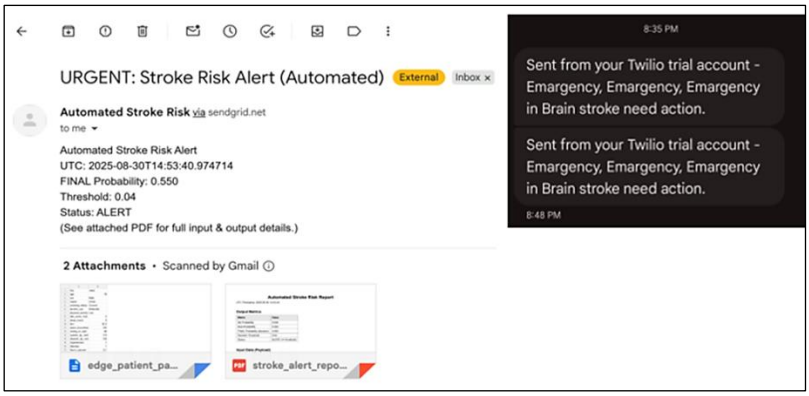
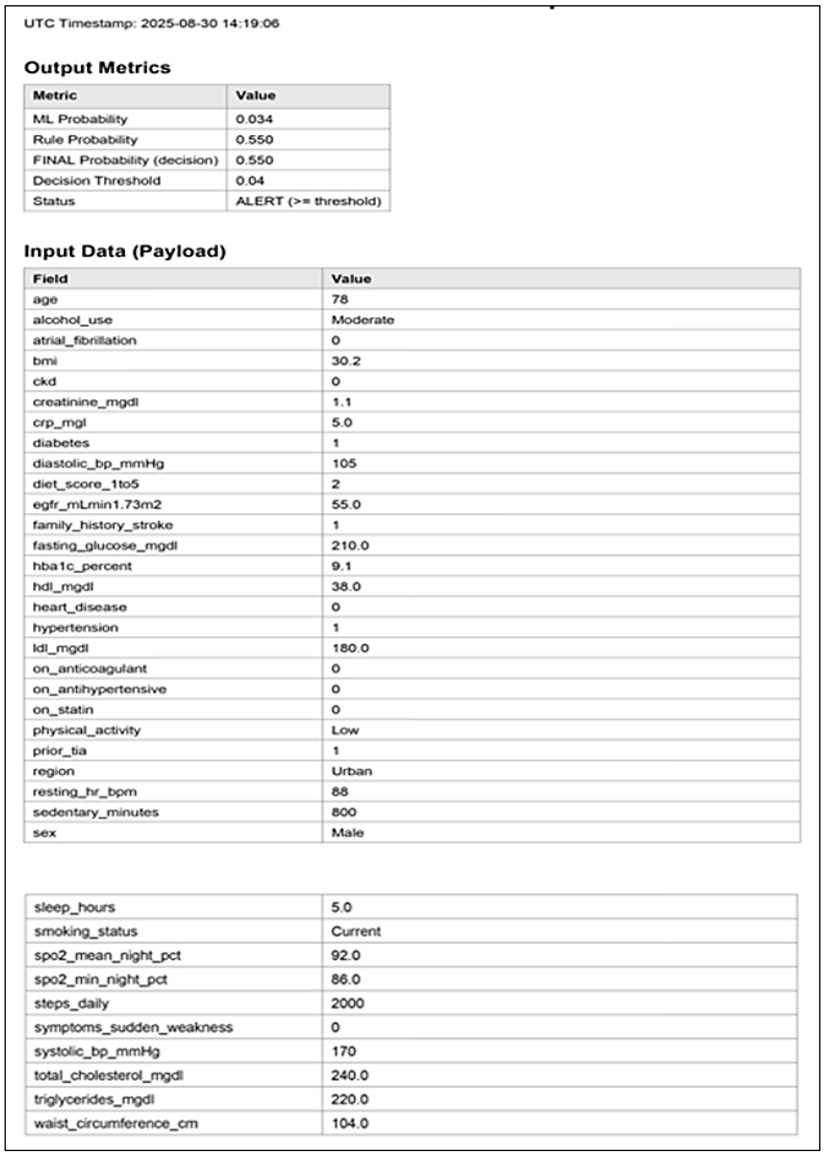


Figure 7: Sent Email Alert to the Concerned Health Care Professional
Inset: SMS Alert to the Concerned People

The single harmonized corpus used in this study is pending multi-site external validation. Despite sensitivity tests, label noise might persist, and sensor quality varies from device to device. While SMOTE and calibration lessen imbalance, PPV/NPV may be impacted by changes in prevalence. The next steps include improved subgroup fairness guarantees, longitudinal signal integration, prospective assessment, and monitoring of drift, energy, and alert burden.

This model handles common issues by using a carefully selected, harmonized multi-source corpus (10k/38) in place of single-source data and by utilizing leakage-safe pre-processing methods such as correlation pruning, stratified splits, and train-only imputation/SMOTE. The strategy integrates SMOTE, isotonic calibration, and validation-based threshold tuning to address class imbalance, miss-calibration, and arbitrary cut-offs. The strategy uses RFECV and Cat Swarm in combination with hybrid stacks to co-optimize features and hyper-parameters according to the desired metric. The goal of this approach is to improve generalization and decrease variance. An edge application that incorporates auditable logs, rule-based overrides, and PDF reports addresses deployment constraints. Along with optional drift and fairness monitoring, federated learning is also used for privacy-preserving updates. Real-time on-device inference is made possible by smart watch-grade sensors, which supports useful and energy-efficient clinical workflows.



UTC Timestamp: 2025-08-30 14:19:06

Output Metrics

Metric	Value
ML Probability	0.034
Rule Probability	0.550
FINAL Probability (decision)	0.550
Decision Threshold	0.04
Status	ALERT (>= threshold)

Input Data (Payload)

Field	Value
age	78
alcohol_use	Moderate
atrial_fibrillation	0
bmi	30.2
ckd	0
creatinine_mgdl	1.1
crp_mgl	5.0
diabetes	1
diastolic_bp_mmHg	105
diet_score_1to5	2
egfr_mLmin1.73m2	55.0
family_history_stroke	1
fasting_glucose_mgdl	210.0
hba1c_percent	9.1
hdl_mgdl	38.0
heart_disease	0
hypertension	1
ldl_mgdl	180.0
on_anticoagulant	0
on_antihypertensive	0
on_statin	0
physical_activity	Low
prior_tia	1
region	Urban
resting_hr_bpm	88
sedentary_minutes	800
sex	Male
sleep_hours	5.0
smoking_status	Current
spo2_mean_night_pct	92.0
spo2_min_night_pct	86.0
steps_daily	2000
symptoms_sudden_weakness	0
systolic_bp_mmHg	170
total_cholesterol_mgdl	240.0
triglycerides_mgdl	220.0
waist_circumference_cm	104.0

Figure 8: Patient Report

Conclusion

To ensure privacy-preserving learning, a real-time, leakage-safe stroke-risk pipeline is built for wearable deployment. Cat Swarm search, wrapper selection, and thorough pre-processing were used on a 10,000x38 curated corpus. Additionally, two stacked hybrids were compared using SMOTE calibration and threshold tuning. At the tuned operating point (threshold = 0.50), the Hybrid Model 1 recorded the highest accuracy of 96.4%. The Hybrid Model 2 and its default settings were outperformed by this outcome. The artifact enables rule-based safety overrides, automated alerts via PDF, email, and SMS, and on-device inference. It consists of the pre-processor, chosen features, and decision threshold. Altogether, the outcomes meet the stated AIM and goal: a transparent, edge-ready system that improves discriminative performance while maintaining deployment realism and offering a clear path for updates that protect privacy.

Using multimodal fusion techniques with ECG, PPG, IMU, and CGM data is advised for future implementations. When considering counterfactual recourse, think about using temporal transformers with self-supervised pre-training, calibrated uncertainty measures, and causal inference techniques. Prioritizing on-device federated distillation with differential privacy secure aggregation is also necessary. Moreover, real-world A/B testing, FHIR-based interoperability, energy-aware continuous learning, and rigorous regulatory validation are advised for complete system development.

Abbreviations

None.

Acknowledgment

The authors would like to express their sincere gratitude to the management of JIS College of Engineering, JIS Group, the Department of Information Technology for their invaluable support and for providing access to essential research and development facilities that made this work possible.

Author Contributions

Subhodip Koley: developing the questionnaires, conducting the data analysis, Sumit Das: identifying the research topic, designing the study, manuscript drafting while overseeing the data-collection process, Annwesha Banerjee Majumdar: identifying the research topic, designing the study, manuscript drafting while overseeing the data-collection process, Abhrendu Bhattacharya: collaborated closely throughout the work, methodological rigour, clear communication of the study's findings, Paramita Sarkar: collaborated closely throughout the work, methodological rigour, clear communication of the study's findings, Achyut Mitra: developing the questionnaires, conducting the data analysis. Collectively, this team effort reflects a shared commitment to advancing knowledge in the field.

Conflict of Interest

No conflict of interest.

Declaration of Artificial Intelligence (AI) Assistance

The authors declare no use of artificial intelligence (AI) for the write-up of the manuscript.

Ethics Approval

Not applicable.

Funding

None.

References

1. Feigin VL, Brainin M, Norrving B, Martins SO, Pandian J, Lindsay P, F Grupper M, Rautalin I. World Stroke Organization: Global Stroke Fact Sheet 2025. Int J Stroke. 2025;20(2):132-144. doi: 10.1177/17474930241308142
2. Alagesan IR, De R, Basu M, Mitra S, Sardar JC. A study on knowledge, attitude, and practice regarding biomedical waste management among health-care providers at a tertiary care hospital, Kolkata. Indian J Community Fam Med. 2024;10(2):74-80.
3. Rahman S, Hasan M, Sarkar AK. Prediction of brain stroke using machine learning algorithms and deep neural network techniques. Eur J Electr Eng Comput Sci. 2023;7(1):23–30.
4. Telu VS, Padimi V. Optimizing predictions of brain stroke using machine learning. J Neutrosophic Fuzzy Syst. 2022; 2:31–43.
5. Dritsas E, Trigka M. Stroke risk prediction with machine learning techniques. Sensors. 2022;22 (13):4670.
6. Alruily M, El-Ghany SA, Mostafa AM, Ezz M, El-Aziz AAA. A-tuning ensemble machine learning technique for cerebral stroke prediction. Applied Sciences (Basel). 2023;13(8):5047. doi:10.3390/app13085047
7. Akter B, Rajbongshi A, Sazzad S, Shakil R, Biswas J, Sara U. A Machine Learning Approach to Detect Brain Stroke Disease. 4th International Conference on Smart Systems and Inventive Technology (ICSSIT).

2022;897–901. https://ieeexplore.ieee.org/document/9716345
8. Al-Zubaidi H, Dweik M, Al-Mousa A. Stroke prediction using machine learning classification methods. International Arab Conference on Information Technology (ACIT). 2022;1–8. https://ieeexplore.ieee.org/document/10022050
9. Mridha K, Ghimire S, Shin J, Aran A, Uddin MdM, Mridha MF. Automated stroke prediction using machine learning: An Explainable and exploratory study with a web application for early intervention. IEEE Access. 2023;11:52288–308.
10. Rahim AMA, Sunyoto A, Arief MR. Stroke prediction using machine learning method with extreme gradient boosting algorithm. MATRIK J Manaj Tek Inform Dan Rekayasa Komput. 2022;21(3):595–606.
11. Biswas N, Uddin KMM, Rikta ST, Dey SK. A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthc Anal. 2022;2:100116.
12. Kanna RK, Reddy CVR, Panigrahi BS, Behera N, Mohanty S. Machine learning based stroke predictor application. Eai Endorsed Trans Internet Things. 2024; 10. https://publications.eai.eu/index.php/IoT/article/view/5384:1-6
13. Hassan A, Gulzar Ahmad S, Ullah Munir E, Ali Khan I, Ramzan N. Predictive modelling and identification of key risk factors for stroke using machine learning. Sci Rep. 2024;14(1):11498.
14. Srinivas A, Mosiganti JP. A brain stroke detection model using soft voting based ensemble machine learning classifier. Meas Sens. 2023; 29:100871.
15. Rehman A, Alam T, Mujahid M, Alamri FS, Ghofaily BA, Saba T. RDET stacking classifier: A novel machine learning based approach for stroke prediction using imbalance data. PeerJ Computer Science. 2023;9:e1684 https://doi.org/10.7717/peerj-cs.1684
16. Bathla P, Kumar R. A hybrid system to predict brain stroke using a combined feature selection and classifier Intelligent Medicine. 2024; 75–82. https://mednexus.org/doi/full/10.1016/j.imed.2023.06.002
17. Chakraborty P, Bandyopadhyay A, Parui S, Swain S, Banerjee PS, Si T, S. Qin, S. Mallik. OptiSelect and EnShap: Integrating machine learning and game theory for ischemic stroke prediction. PLOS ONE. 2025;20(8): e0328967.
18. Dubey Y, Tarte Y, Talatule N, Damahe K, Palsodkar P, Fulzele P. Explainable and interpretable model for the early detection of brain stroke using optimized boosting algorithms. Diagnostics. 2024;14(22): 2514.
19. Vemu PL, Yang E, Ebinger J. ESH hypertension guideline update: Bringing us closer together across the pond. American College of Cardiology.2023. https://www.acc.org/Latest-in-Cardiology/Articles/2024/02/05/11/43/2023-ESH-Hypertension-Guideline-Update
20. American Diabetes Association. Diabetes Diagnosis & Tests. Diabetes.org. 2025. https://diabetes.org/about-diabetes/diagnosis
21. Mayo Clinic Staff. High cholesterol diagnosis and treatment. Mayo Clinic. 2025. https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/diagnosis-treatment/drc-20350806
22. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Chronic Kidney Disease (CKD). Bethesda (MD): National Institutes of Health, 2016. https://www.niddk.nih.gov/health-information/kidney-disease/chronic-kidney-disease-ckd
23. Markus H, Pereira A, Cloud G, editors. Stroke medicine. 3rd ed. Oxford: Oxford University Press. 2025. ISBN: 9780198906285 (eBook).
24. Aouatifcherdid. Healthcare dataset stroke data. Kaggle. 2026. https://www.kaggle.com/datasets/aouatifcherdid/healthcare-dataset-stroke-data
25. Hassan A. Stroke prediction dataset. IEEE DataPort. 2023. doi:10.21227/mxfb-sc71. https://ieee-dataport.org/documents/stroke-prediction-dataset
26. International Stroke Database (RRID:SCR_007347). SciCrunch Resource Report. 2026. https://rrid.site/data/record/nlx_144509-1/SCR_007347/resolver
27. Predicting Stroke Risk Dataset. 2025. https://figshare.com/articles/dataset/_b_Predicting_Stroke_Risk_Dataset_b_/28668398/4
28. Health Data Research UK. Dataset 1003 Synthetic Dataset of Hospital Admissions for an Acute Stroke. Health Data Research Gateway. 2026. https://healthdatagateway.org/en/dataset/1003

How to Cite: Koley S, Das S, Majumdar AB, Bhattacharya A, Sarkar P, Mitra A. Hybrid Framework for Sustainable Wearable Stroke Monitoring. Int Res J Multidiscip Scope. 2026; 7(1): 1566-1579.
DOI: 10.47857/irjms.2026.v07i01.08675