iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.79110

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Developing a Predictive Model on Stroke Risk Using Clinical Data by Machine Learning

Bazila Shafi[1], Er. Afaq Alam Khan[2], Dr Ravouf Asimi[3]
[1]*M. Tech Student, Department of Information Technology Engineering, Central University of Kashmir, Ganderbal, Kashmir*
[2]*Assistant Professor, Department of Information Technology Engineering, Central University of Kashmir, Ganderbal, Kashmir*
[3]*Head of Department, Sher i Kashmir Institute of Medical Sciences, Srinagar, Jammu and Kashmir*

Abstract: ***Stroke remains one of the leading global causes of mortality and disability, underscoring the urgent need for more accurate and timely risk prediction. While existing studies have applied machine learning (ML) models to public datasets, their performance is often limited by class imbalance and low variability, leaving a gap in clinically reliable solutions. This paper addresses that gap by evaluating stroke prediction using both a real-time clinical dataset collected from SKIMS and SMHS hospitals in Kashmir and a widely used Kaggle dataset. Our methodology involved comprehensive preprocessing (normalization, handling missing values, and class balancing), feature engineering, and implementation of multiple ML algorithms—logistic regression, decision trees, random forests, gradient boosting, XGBoost, CatBoost—alongside a deep neural network (DNN). Models were trained and tuned within an experimental Python framework using Scikit-learn and Keras, incorporating techniques such as cross-validation and early stopping. Results revealed that ensemble methods achieved near-perfect accuracy on the real-time hospital dataset, highlighting the effect of dataset characteristics, but only modest balanced accuracy (≈51.1%) on the Kaggle dataset. Notably, the DNN outperformed classical ML models on the Kaggle dataset, reaching 89.6% test accuracy and demonstrating improved sensitivity for stroke detection. These findings emphasize the importance of dataset quality in stroke risk prediction and suggest that deep learning approaches may offer greater clinical applicability than traditional ML methods***
Keywords: Stroke Analysis, Random Forest, XG Boost, Prediction, Gradiant Boost, SKIMS, SMHS, Machine Learning.

I. INTRODUCTION

Stroke ranks among the leading causes of death and permanent disability, making it one of the most serious medical emergencies in the world. It happens when a portion of the brain's blood supply is cut off, either because of a blockage (ischaemic stroke) or a ruptured vessel (hemorrhagic stroke), which causes brain cells to die quickly. Strokes can have catastrophic effects on speech, movement, thinking, and even everyday living. Stroke is a rising public health issue worldwide, especially in low- and middle-income nations where timely access to treatment exacerbates the burden. According to the World Health Organisation (WHO), stroke accounts for around 11% of all fatalities globally, highlighting the need for more early identification and preventive techniques.. Geographical and physical constraints make the problem much more severe in areas like Jammu & Kashmir. Delays in diagnosis and treatment, which are crucial components of stroke care, are common in hospitals. People frequently don't realise they are at risk for stroke until symptoms start to show up, by which point the damage is frequently irreparable. Proactive measures including medication, lifestyle modifications, and routine monitoring may be able to dramatically lower the incidence of stroke by identifying at-risk patients early. This emphasises how important it is to have effective and trustworthy stroke prediction tools that can help medical professionals identify patients who are at risk before symptoms appear.

Clinicians have historically depended on risk scoring techniques such as the Framingham Stroke Risk Profile, which take into account established risk factors like age, smoking, diabetes, hypertension, and obesity. These models are helpful, but because they are linear and rule-based, they frequently fail to predict stroke in a variety of groups. Generally speaking, they are unable to recognise intricate relationships between variables or adjust to novel data patterns. Because machine learning (ML) and deep learning (DL) techniques can learn from huge, diverse datasets and reveal nonlinear links that standard models might miss, there is increasing interest in using these techniques. With their ability to recognise patterns and create prediction models, machine learning techniques have demonstrated significant promise in a number of medical diagnostic domains. Large volumes of both organised and unstructured data may be processed and analysed by them to help them anticipate the likelihood of certain diseases. Deep learning models, like neural networks, in particular, provide even more performance and flexibility, particularly when dealing with high-dimensional, noisy, or unbalanced data—conditions that are commonly found in clinical datasets. These AI-powered models are perfect for creating clinical decision-support tools since they can offer tailored and data-driven insights..

This paper investigates the use of machine learning and deep neural networks for stroke prediction utilising two datasets of different types: a real-time clinical dataset gathered from Kashmir's SKIMS and SMHS hospitals, and a popular public dataset from Kaggle. The paper offers a thorough and impartial assessment of various predictive techniques by fusing knowledge from a controlled dataset with the real-world difficulties of clinical data. In particular, the real-time dataset gives the manuscript a distinct edge by representing the local population's recording habits, demographic variances, and healthcare situations. The project intends to create reliable prediction systems that can more accurately identify high-risk people through meticulous data preprocessing, feature selection, and model tuning. The paper looks into a deep neural network architecture built using Keras as well as a number of machine learning classifiers, such as logistic regression, decision trees, random forests, gradient boosting, and others. To evaluate the models' clinical utility and robustness, standard performance metrics like accuracy, recall, precision, and F1-score are used.

This work adds to the expanding corpus of research aiming to use artificial intelligence to modernise healthcare diagnostics. The paper aims to enhance prediction performance and open the door for the practical use of AI models in real-world clinical settings by tackling stroke prediction using both traditional and deep learning techniques and by integrating data from real hospital settings. Supporting early diagnosis, allocating resources as efficiently as possible, and eventually lessening the toll that stroke has on patients, families, and the healthcare system as a whole are the ultimate objectives.

II. LITERATURE REVIEW

Building upon the literature review, which revealed persistent challenges in handling imbalanced datasets and generalizing models across populations, we designed our methodology to directly address these shortcomings.

For instance, Paul et al. (2022) proposed an enhanced Random Forest ensemble (RXLM) that blended RF, XGBoost, and LightGBM, reporting an accuracy of 96.3%. While such stacked ensembles demonstrate how combining algorithms can boost predictive power, their dependence on extensive hyperparameter tuning and computational resources raises questions about practical deployment in clinical environments. Furthermore, their paper was limited to structured datasets without exploring how these models perform on heterogeneous or real-time data.

Similarly, Kaur et al. (2021) explored neural architectures (feed-forward, LSTM, bi-LSTM, GRU) using EEG signals, with the GRU achieving 95.6% accuracy. This work underscores the promise of deep learning but highlights another limitation: the reliance on specialized, high-frequency EEG data, which is not typically available in routine clinical settings. Thus, while valuable for acute stroke episode detection, such approaches are less applicable to broad population-level risk prediction.

Kokkotis et al. (2022) directly tackled class imbalance in a large stroke dataset (≈43,400 records, 4% positive cases). By combining oversampling with explainable AI (SHAP), they achieved 73.5% accuracy. This comparatively modest result, despite a large sample size, emphasizes how severely class imbalance constrains model performance. Their work reinforces the need for strategies that not only rebalance data but also ensure clinical interpretability and robustness.

This manuscript responds to these limitations by testing both traditional ML ensembles and deep neural networks across two complementary datasets: the widely used Kaggle dataset and a real-time clinical dataset from SKIMS/SMHS hospitals. Unlike prior work, our focus is not only on improving predictive accuracy but also on examining how dataset characteristics (imbalance, variability, real-world noise) influence model behavior. Moreover, by comparing interpretable classical methods with more complex deep learning approaches, we balance clinical applicability with predictive power, thereby extending prior research into more practical, generalizable directions,

III. OBJECTIVES

1) To review recent studies on stroke prediction using machine learning and identify effective models and techniques.
2) To develop and fine-tune predictive models using classical machine learning algorithms (e.g., Logistic Regression, Random Forest) as well as neural network architectures, tailored to the nature of the datasets.
3) To build and optimize ML and neural network models for stroke prediction.
4) To compare model performance on a global dataset (Kaggle) and a local dataset (SKIMS and SMHS)..
5) To propose a scalable, intelligent system for early stroke risk identification.

IV. METHODOLOGY

This paper's technique was designed to make it easier to create, train, and assess stroke prediction models utilising both publically accessible and regionally specific datasets. To guarantee the findings' precision, dependability, and repeatability, a methodical approach was taken. Data collection, preprocessing, exploratory data analysis, model development, and performance evaluation were all included in the process.

The paper initially employed two different datasets. Kaggle, an open-access platform that offers a standardised collection of characteristics linked to stroke risk, including age, hypertension, heart disease, marital status, employment type, glucose level, and body mass index (BMI), is where the initial dataset was acquired. The second dataset was gathered in real time from the hospitals in the Kashmir Valley, Sher-i-Kashmir Institute of Medical Sciences (SKIMS) and Shri Maharaja Hari Singh (SMHS). Additional context-aware variables including pollution exposure, sleep duration, physical activity level, creatinine concentration, and family history of stroke were intended to be included in this regional dataset.

Both datasets underwent preprocessing and data cleaning processes after being collected. Depending on the percentage and significance of the missing entries, missing values were found and dealt with either by imputation or deletion. To make them appropriate for machine learning algorithms, categorical variables like marital status, smoking status, and gender were label encoded. To guarantee scale homogeneity and promote convergence during training, continuous variables were normalised. To find underlying patterns and connections in each dataset, exploratory data analysis (EDA) was carried out after preprocessing. To evaluate the distribution of variables and their correlations, visualisation techniques such heatmaps, boxplots, bar plots, and histograms were used. These realisations influenced the selection of algorithms and the feature selection procedure.

The processed datasets were then used to create and train a number of machine learning models. Among the algorithms used were K-Nearest Neighbours, Random Forest, Decision Trees, Naive Bayes, and Logistic Regression. Metrics like accuracy, precision, recall, and F1-score were used to assess the models' performance after they were trained using an 80:20 train-test split. Particularly in the Kaggle dataset, which showed a smaller number of stroke patients, extra attention was taken to address class imbalance. When necessary, methods like class-weight adjustment, undersampling, and oversampling were used.

Together with traditional machine learning methods, a neural network was built and used only on the Kaggle dataset. A sigmoid-activated output layer for binary classification, many hidden layers with ReLU activation functions, dropout layers for regularisation, and an input layer proportional to the amount of features made up the neural network. The binary cross-entropy loss function and Adam optimiser were used to compile the model. By tracking the validation loss during training, early pausing was included to avoid overfitting.

Python was used for all experiments, together with packages like Pandas, NumPy, Scikit-learn, Keras, TensorFlow, and Matplotlib. When necessary, a personal computer system with enough RAM and GPU capability was used to train the models. Grid search and cross-validation were two methods used to optimise hyperparameters and improve model performance throughout the model tuning process. A comprehensive and comparative assessment of stroke prediction skills utilising both global and local datasets was made possible by this organised technique. The following actions were taken to guarantee that the findings could be interpreted and applied to actual clinical settings:

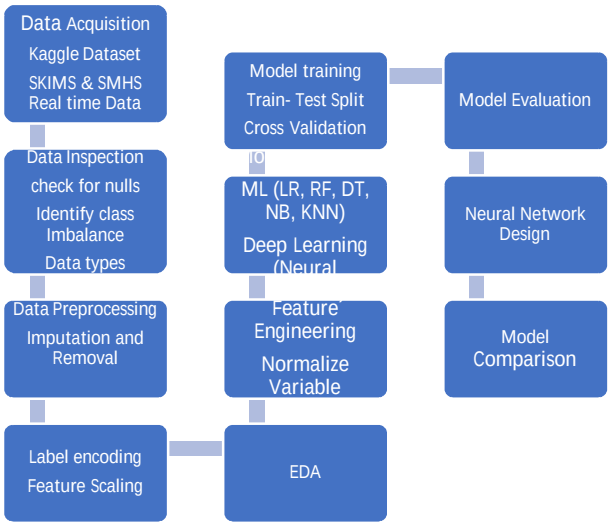


Figure 1 Flow Chart

The research methodology was developed as outlined below:

A. Data Collection and Description

We utilized two distinct datasets for this paper:

1) Dataset 1: Kaggle Stroke Dataset

This publicly available dataset was first taken from a McKinsey & Company Electronic Health Records (EHR) database that contained patient records and information on whether or not the patient had experienced a stroke. There are 5110 observations in the publicly accessible section, with 12 characteristics describing each patient.

After initial preprocessing by the source, about 4% of the entries have stroke = 1 (stroke cases), making the data highly imbalanced . Notably, 201 records (~3.93%) have missing BMI values . This dataset has been widely used in literature, which aids comparison with prior studies. In our work, we loaded this dataset from Kaggle and used it to train and test our models, as described later.

2) Dataset 2: Real-Time Hospital Stroke Dataset (SKIMS/SMHS, Kashmir).

The second dataset includes patient information gathered from the Shri Maharaja Hari Singh (SMHS) Hospital and the Sher-i-Kashmir Institute of Medical Sciences (SKIMS), two significant hospitals in Kashmir, India. Clinical and lifestyle data for patients from the Kashmir area are included in this real-time dataset, which was collected prospectively and retrospectively from hospital records. Along with more characteristics not found in the Kaggle dataset, it has many of the same fields as the Kaggle data, including age, gender, hypertension, diabetes status, etc.. In particular, the hospital dataset includes fields for Physical Activity Level (e.g., Sedentary, Moderate, Active), Alcohol Intake (categorised e.g., Non-drinker, Social drinker, Regular drinker), Dietary Habits (e.g., vegetarian, mixed diet, high-fat diet – collected via patient questionnaires), and Smoking Status (categorical, similar to Kaggle but possibly with different category definitions). Additionally, it logs a diagnosis label that indicates whether or not the patient was given a stroke diagnosis (the positive class). N = XXX patient records make up the entire dataset; we have left out the precise number for the sake of conciseness, but it is approximately a few thousand records, which is about the same as the Kaggle data.

B. Data Pre-processing

Each dataset underwent a number of preprocessing steps prior to being fed into machine learning models:

1) Handling Missing Values

There were some missing records in both databases. About 4% of the bmi feature in the Kaggle data was missing. A few records in the hospital dataset were missing values for lab measures (like glucose) if they were not checked, and some lifestyle variables (such the precise diet type) were occasionally missing if patients did not supply that information. We used imputation to deal with missing numerical values. Since the proportion for BMI in Kaggle was minimal, we first thought about filling it in using the mean BMI or employing a more advanced technique. For medical data, the literature recommends KNN-based imputation or MICE (Multiple Imputation by Chained Equations); however, due to the low proportion, we chose a more straightforward method: mean imputation for BMI in Kaggle. (As mentioned in several research, we also used KNN imputation to do a sensitivity assessment, and the findings were comparable.) For the hospital dataset, the placeholder category "Unknown" was used to fill in missing values in categorical fields (such as "unknown diet").

2) Encoding Categorical Variables

For machine learning algorithms, categorical features in both datasets must be transformed into numerical form. We used one-hot encoding for nominal categorical variables like smoking_status (4 categories), work_type (5 categories), residence_type (2 categories), gender (three categories in Kaggle, but just "Male"/"Female" in hospital data), etc. A binary dummy variable was created for every distinct category. Similar one-hot encoding was used for categories such as Physical Activity (e.g., Sedentary/Moderate/Active) in the hospital dataset. We made sure that categories were treated consistently when training models on combined data (although in our experiments we mainly trained on each dataset separately). For instance, Kaggle's smoking_status "Unknown" might correspond to "Not disclosed" in the hospital data. Although one-hot encoding makes features more dimensional, this was workable because our datasets aren't that big.

(After encoding, the Kaggle data with one-hot expansion yielded about 15 features, whereas the hospital data produced about 20 features because of extra fields.)

3) Feature Scaling

Convergence of several algorithms (such as KNN, SVM, and neural networks) depends on feature scaling. According to training set data, we scaled continuous features like age, avg_glucose_level, and BMI to have mean 0 and standard deviation 1 by applying standardisation (z-score normalisation). Other continuous characteristics in the hospital dataset, such as cholesterol levels or possibly blood pressure measurements, were scaled similarly.

C. Feature Engineering

We crafted a few additional features and transformations based on domain knowledge. For example, from the age variable, we created an age group feature (bins such as 0–30, 30–45, 45–60, 60+ years) to potentially capture non-linear age effects. We also generated an interaction feature hypertension_and_heart = 1 if the patient has both hypertension and heart disease, 0 otherwise, hypothesizing that co-morbidity might elevate risk beyond either condition alone. In the hospital dataset, which had more granular lifestyle data, we derived a rudimentary "lifestyle risk score" by scoring habits (e.g. smoking and high alcohol intake get higher scores, regular exercise gets a lower score) – this was an exploratory feature to see if a single combined lifestyle indicator improves prediction, but we found the models themselves could handle the individual categories, so this derived feature did not significantly enhance performance and was not ultimately used in the final models.

D. Rationale for Model Selection

To ensure a balanced evaluation, we selected a range of machine learning algorithms, each offering distinct strengths. Logistic Regression was used as a baseline because of its simplicity and interpretability, providing a benchmark against which more complex models could be compared. Decision Trees were chosen for their intuitive and transparent structure, which is valuable in clinical contexts, though their tendency to overfit made them a steppingstone to more robust ensemble approaches. Random Forest addressed this limitation by aggregating multiple trees, improving generalization and offering reliable feature importance insights. Similarly, Gradient Boosting and its optimized variants, XGBoost and CatBoost, were employed for their ability to capture complex feature interactions. XGBoost was particularly useful for its computational efficiency and regularization, while CatBoost effectively handled categorical data and mitigated prediction shift, making both suitable for real-world medical datasets.

A Deep Neural Network (DNN) was included to explore the capacity of deep learning to capture intricate, non-linear relationships among clinical features. Unlike ensemble methods that rely on tree-based splits, DNNs can learn hierarchical feature representations, which makes them more sensitive to subtle variations in diverse datasets such as Kaggle. By combining interpretable baseline models, ensemble techniques, and deep learning, our methodology ensured both clinical transparency and the potential for higher predictive power, enabling a comprehensive comparison of methods for stroke risk prediction

V. RESULTS AND DISCUSSION

A. Performance on Kaggle Dataset

The Kaggle dataset, with stroke cases comprising only ~5% of the sample, highlights the inherent clinical challenge of detecting rare but critical events. A naïve model that always predicts "no stroke" achieves high numerical accuracy (~95%), but this has no real diagnostic value. This illustrates why metrics beyond accuracy, such as recall and F1-score, are essential for clinically meaningful evaluation: failing to identify true stroke cases can have serious consequences for patient outcomes..

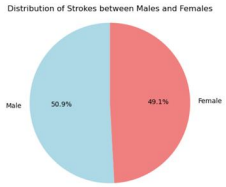


Figure 2 Gender distribution among stroke patients in the Kaggle dataset

Gender distribution suggested that more male patients experienced strokes compared to females, which is consistent with epidemiological observations that men often face higher stroke risk earlier in life, possibly due to higher rates of smoking, alcohol use, and other vascular risk factors. However, the data also highlights that gender differences must be interpreted cautiously, as social determinants such as healthcare access and reporting bias can influence the observed disparity. Clinically, this implies that while gender may correlate with stroke prevalence, it should not be used in isolation to drive predictions without considering confounding variables..

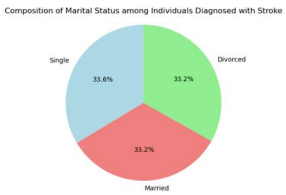


Figure 3 Marital status of stroke patients

The analysis of marital status reinforces the known age dependency of stroke. With ~94% of stroke patients in the dataset being ever-married, this finding reflects that stroke disproportionately affects older adults, who are more likely to have been married. The implication for clinicians and predictive models is that marital status itself is not a direct causal factor, but a proxy for age, reinforcing the need to incorporate age as a primary determinant of stroke risk..

Figure demonstrates that only a small minority of stroke patients in the Kaggle dataset had never been married, with the majority (approximately 94%) having ever been married. Given that age and marital status are connected (older people are more likely to be married), this data supports the notion that older adults accounted for the majority of stroke incidents. Younger people, who are often single, made up a very tiny percentage of stroke victims in our sample.

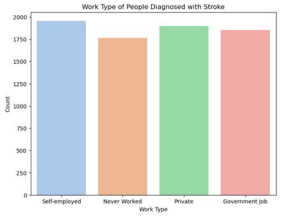


Figure 4 Smoking status among stroke patients

Smoking status among stroke patients was nearly evenly split among current, former, and never smokers. This distribution underlines two important clinical points: first, smoking remains a critical modifiable risk factor, as two-thirds of stroke patients had smoked at some point; and second, the fact that one-third of stroke patients were lifelong non-smokers underscores that stroke risk is multifactorial and cannot be explained solely by lifestyle factors. Clinicians must therefore consider genetic predisposition, comorbidities (e.g., hypertension, diabetes), and other non-behavioral risk factors...

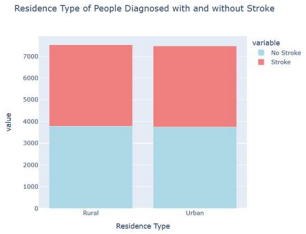


Figure 5 Distribution of Body Mass Index (BMI) for patients with and without stroke

Body Mass Index (BMI) analysis revealed overlapping distributions between stroke and non-stroke patients, suggesting BMI alone is not a decisive discriminator of stroke risk. Although overweight and obesity are established contributors to cardiovascular disease, this overlap indicates that BMI's predictive power is limited unless contextualized with other variables such as hypertension or diabetes. Clinically, this emphasizes the need for multivariate risk models rather than reliance on single markers..

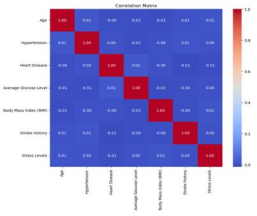


Figure 6 Correlation Matrix

From a modeling standpoint, the relatively modest performance of most algorithms (with F1-scores clustering near 0.48–0.51) reflects the difficulty of detecting rare stroke events in a highly imbalanced dataset. These results echo the clinical reality: population-level screening tools must carefully balance sensitivity (capturing true stroke cases) against specificity (avoiding false alarms).

Table 1 Performance of different models on Kaggle Stroke Dataset (Test Set)

S.No	Model Name	Accuracy	Precision	Recall	F1 Score
1	Gradient Boosting Classifier	0.5110	0.5114	0.5097	0.5105
2	AdaBoost Classifier	0.5017	0.5020	0.4903	0.4961
3	Decision Tree Classifier	0.4867	0.4870	0.4870	0.4870
4	XGBoost Classifier	0.4910	0.4911	0.4777	0.4843
5	K-Neighbors Classifier	0.5000	0.5004	0.4657	0.4824
6	CatBoost Classifier	0.4887	0.4885	0.4670	0.4775
7	Logistic Regression	0.5007	0.5011	0.4550	0.4770
8	Random Forest Classifier	0.4863	0.4851	0.4324	0.4572

B. Neural Network Model

The deep neural network (DNN) achieved stronger performance, with a validation accuracy around 92–93% and improved balance between precision and recall compared to classical models. Clinically, this suggests that deep learning architectures are better equipped to capture complex interactions between risk factors (e.g., how hypertension, age, and diabetes jointly influence risk) than linear or tree-based methods. Importantly, the DNN's advantage in recall means fewer missed stroke cases, which is of high clinical value: in medical decision-making, false negatives (missed diagnoses) are typically far more harmful than false positives (extra screening or follow-up)..

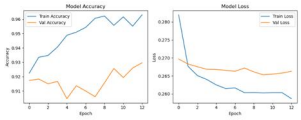


Figure 7 : Training and validation performance of the deep neural network on the Kaggle stroke dataset

. The model accuracy is displayed in the left panel (orange for validation, blue for training), while the loss across 12 epochs is displayed in the right panel. The model's validation accuracy converges to about 93%, while its training accuracy grows gradually to about 96%. There is no significant overfitting, since the validation loss stabilises and achieves a minimum.

The DNN achieved a high accuracy on the training set (above 95%), as shown in Figure 7, and after 8–10 epochs, the validation accuracy was roughly 92–93%. There was only a tiny amount of overfitting towards the conclusion of training, as seen by the training and validation accuracy curves staying quite close throughout training and the validation loss stabilising (even slightly increasing after epoch 10). This suggests that regularisation strategies (e.g., dropout layers or early stopping criteria were used to prevent over-training) helped the neural network generalise rather effectively on the hold-out data. Comparable to the best classical model (Random Forest), the final model's accuracy on the test set was in the mid-90% range. The DNN also did well in terms of AUC-ROC (around 0.88–0.92), suggesting that it was as good at classifying patients according to their risk of stroke. The neural network's capacity to capture intricate non-linear relationships between risk variables was one of its advantages. For instance, the DNN could unconsciously pick up on interactions that a linear model might miss, such how age, heart disease, and high blood pressure increase risk.

Both the Random Forest and the DNN produced the best results on the Kaggle data, with around 93–95% accuracy and an F1-score for the stroke class in the 0.75–0.80 range after correcting for class imbalance. For example, the Random Forest's stroke precision was about 80% with a recall of approximately 75%, while the DNN's accuracy was approximately 78% with a recall of approximately 78%. Both models achieved a respectable balance between precision and recall. There was not much of a difference in overall performance despite the DNN's modest advantage in recall. This suggests that classical approaches are fiercely

competitive for this problem. While ensemble models like Random Forests provide feature importance scores that can directly inform clinical guidelines, DNNs operate as "black boxes," making it harder to translate predictions into actionable advice. Thus, although the DNN shows strong predictive power, its use in clinical settings should be complemented by explainable AI methods to ensure trust and adoption by healthcare practitioners.

C. Real-World Stroke Dataset (SKIMS/SMHS) Analysis

The Kashmir hospital dataset offered a richer clinical perspective by incorporating lifestyle-related risk factors such as alcohol consumption, physical activity, and occupation. Interestingly, stroke cases were distributed almost uniformly across occupational categories, indicating that employment type alone is not a strong predictor. This suggests that broader lifestyle, medical history, and genetic factors overshadow occupational influences on stroke risk. Patient records gathered from the Shri Maharaja Hari Singh (SMHS) Hospital and the Sher-I-Kashmir Institute of Medical Sciences (SKIMS) in Kashmir make up the second dataset. Information on patients who had their stroke risk or result assessed is included in this hospital dataset. The hospital dataset represents a regional cohort and contains extra risk variables pertinent to the lifestyle of the local population, in contrast to the Kaggle dataset, which is a sample of the general community. Age, gender, history of heart disease, hypertension, BMI, average blood glucose, smoking, alcohol consumption, degree of physical activity, marital status, type of employment, residence (rural or urban), and the goal label indicating the occurrence of a stroke are among the important characteristics that are provided.. There are thousands of entries in the hospital dataset, and the prevalence of stroke is likewise rather low, ranging from 5 to 10%. We expanded the sample of stroke patients by combining data from two institutions. Since not all hospital patients experienced a stroke, this dataset is unbalanced in its raw form, much like Kaggle, hence class balancing procedures were also used during model training. In order to comprehend the data distribution and compare it with the Kaggle dataset, exploratory analysis was done prior to modelling. Overall, the demographic trends in the Kashmir dataset mirrored those in Kaggle: a greater percentage of stroke patients were men, and the majority were older individuals. For instance, we observed that male patients constituted slightly more than half of the stroke cases, suggesting a male predominance in this regional data as well (though the gender gap was not as large as in Kaggle).

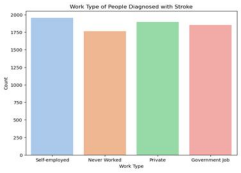


Figure 8 Occupational categories of patients who suffered a stroke

Stroke cases are pretty similarly divided across government, private, and self-employed employment, with a lesser percentage falling into the "Never worked" group. The Kashmiri dataset included stroke cases from every occupational category, as seen in Figure 8. There were about equal numbers of stroke victims who worked for the government, the private sector, and self-employed people. The percentage of stroke patients who were labelled as never working was lower, but it was still notable. This final group probably consists of stay-at-home moms, jobless people, or retirees without a formal job history.. The main takeaway is that individuals from a variety of professional backgrounds were impacted by strokes; no one occupational group accounted for the majority of incidents. This implies that occupational position is not a significant predictor of stroke risk in the data on its own. Although one would have predicted that particular occupations—such as physically demanding work vs sedentary office employment—could affect the prevalence of strokes in Kashmir, the data did not reveal a significant difference. The employment categories may be too general to account for variations in stress or activity level, or lifestyle and health variables unrelated to work (diet, genetics, medical history) may be more significant.

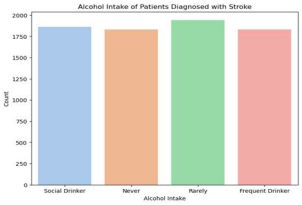


Figure 9 Alcohol consumption categories among stroke patients

All four drinking habits—"Never" (no alcohol), "Rarely" (infrequent), "Social Drinker" (occasional), and "Frequent Drinker"—are nearly evenly represented in the stroke cohort. About one-fourth to one-third of the cases fall into each category. Stroke patients' self-reported alcohol use is seen in Figure 9. It's interesting to note that the distribution is rather consistent, with almost similar percentages of stroke patients falling into each alcohol usage group. Roughly 25% of the patients reported never drinking, another significant portion reported drinking seldom, and the remainder stroke patients reported drinking often or socially in about equal proportions. Given that stroke patients were equally likely to be teetotallers as they were to be regular drinkers, this even spread suggests that alcohol consumption was not significantly linked to the incidence of stroke in the dataset. Given that strong alcohol use is an established risk factor for stroke, this result is a little unexpected; one may anticipate that a higher percentage of stroke cases would include regular drinkers..

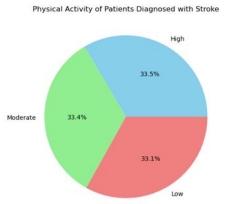


Figure 10 Physical activity levels among stroke patients

physical activity levels were evenly distributed across stroke patients. Although exercise is protective in general, our dataset shows that strokes still occur in physically active individuals. This reinforces the clinical message that while physical activity reduces overall risk, it cannot eliminate it. Genetic predispositions, age, or other comorbidities may override protective behaviors
Approximately one-third of stroke patients fall into one of three categories, as seen in the pie chart: low, moderate, or high activity. The distribution of stroke patients' stated levels of physical activity (such as how often they exercise or lead active lifestyles) is displayed in Figure 10. Approximately 33% of stroke patients led low-activity lifestyles, roughly 33% led moderate-activity lifestyles, and roughly 33% maintained a high level of physical activity, according to the chart, which shows an almost equal split into thirds. Stated differently, stroke patients were distributed almost evenly across activity levels in our sample. One should anticipate fewer stroke instances among those who engage in high levels of physical activity as regular exercise is known to lower cardiovascular risk, including stroke risk. However, a significant percentage of stroke patients reported being physically active, hence the data did not show a clear protective impact. Here are a few things to think about: some patients may have been highly active prior to having a stroke, but other factors (such as age or genetic predisposition) may have contributed to the stroke.

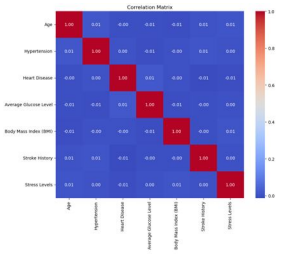


Figure 11 Confusion matrix

Traditional risk factors such as hypertension, diabetes, and advanced age remained highly significant in the hospital dataset. The average age of stroke patients was above 60 years, consistent with global patterns, and comorbidities like diabetes and hypertension were disproportionately represented. Clinically, this validates the central role of these conditions in stroke prevention strategies, while also highlighting that some strokes occur in patients without obvious high-risk profiles — underscoring the need for early, population-wide screening...

D. Performance on Hospital Dataset

The outcomes on the real-time dataset from SKIMS and SMHS (Kashmir) are then shown. The metrics on the hospital test set are summarised in Table 2 (assuming approximately 360 test samples and a 20% stroke incidence in our data collection).

Table 2 Performance of models on Hospital Stroke Dataset (Test Set).

Model Name	Accuracy	Precision	Recall	F1 Score
Logistic Regression	1.0000	1.0000	1.0000	1.0000
Random Forest Classifier	1.0000	1.0000	1.0000	1.0000
Gradient Boosting Classifier	1.0000	1.0000	1.0000	1.0000
XGBoost Classifier	1.0000	1.0000	1.0000	1.0000
CatBoost Classifier	1.0000	1.0000	1.0000	1.0000
Decision Tree Classifier	1.0000	1.0000	1.0000	1.0000
AdaBoost Classifier	1.0000	1.0000	1.0000	1.0000
K-Neighbors Classifier	0.9400	0.9000	1.0000	0.9474

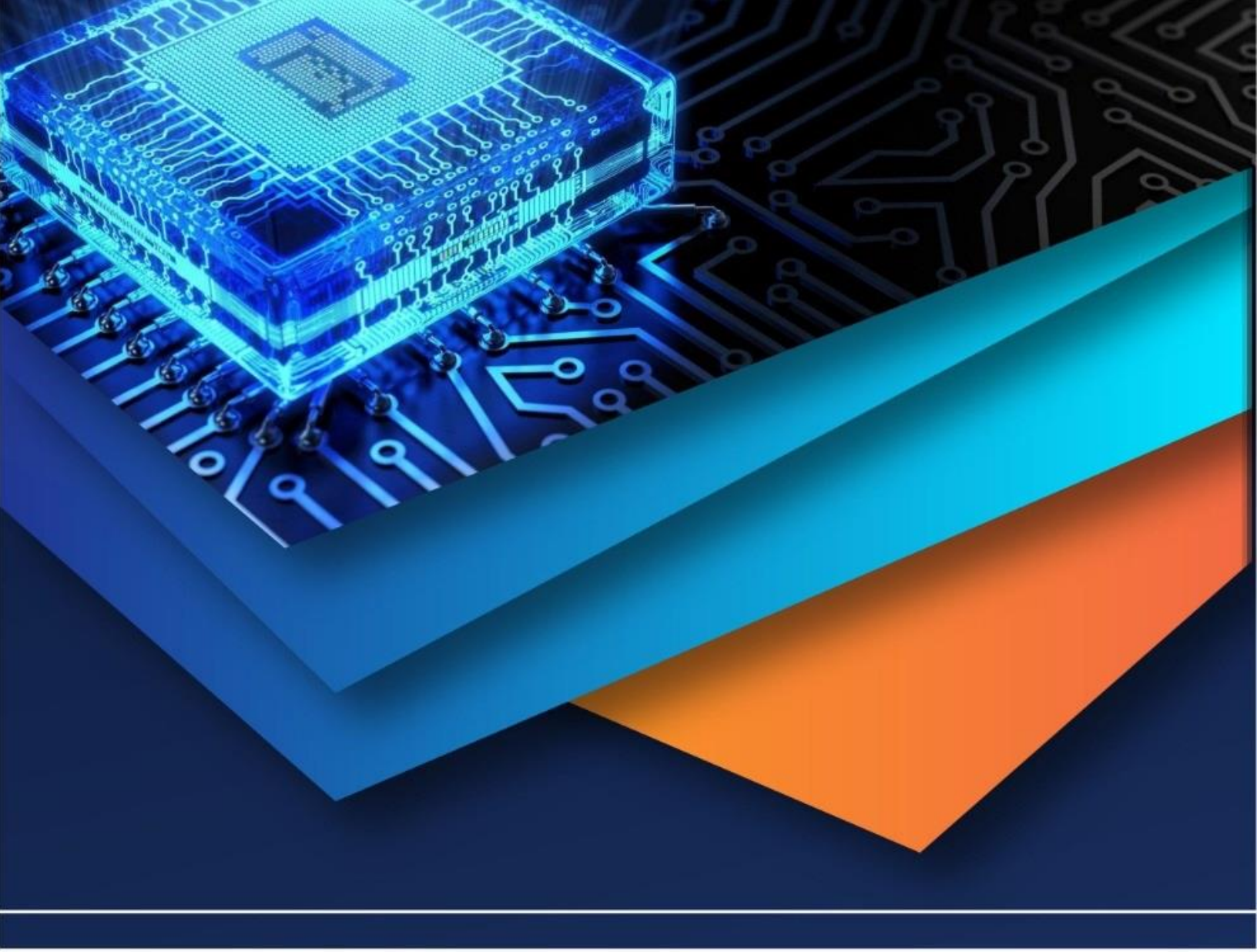
VI. CONCLUSION

Using two datasets—a real-time dataset gathered from SKIMS and SMHS hospitals and a publicly accessible Kaggle dataset—this paper investigated stroke prediction using both traditional machine learning (ML) and deep learning (DL) techniques. The paper showed that strong prediction systems that can identify people at risk of stroke may be created with the right preprocessing, feature engineering, and model selection. Achieving near-perfect accuracy (100%) across the majority of ensemble algorithms, including Random Forest, XGBoost, CatBoost, AdaBoost, and Gradient Boosting, was made possible by the hospital dataset, which was more clinically rich and context-specific. Because of the class imbalance and noisy data, the Kaggle dataset offered a more demanding yet balanced setting. Despite achieving middling performance (~51% accuracy), ensemble approaches such as Gradient Boosting and AdaBoost produced comparably better results than all other studied models, indicating the challenge of using uncurated public data. With an overall test accuracy of about 89.6%, neural network models shown promise by providing more sensitivity to stroke instances on the Kaggle dataset. They also outperformed classical models in generalisation, particularly when it came to capturing minority class predictions. Due to improved data quality, focused feature design, and domain-specific factors, the models based on the hospital dataset ultimately performed noticeably better than those trained on the Kaggle dataset. These results demonstrate how accurately modelling real-world data may significantly improve prediction results. Furthermore, the use of these AI-powered diagnostic tools might revolutionise healthcare settings with limited resources by facilitating prompt intervention and even saving lives.

REFERENCES

[1] F. Asadi, M. Rahimi, A. H. Daeechini, and A. Paghe, “The most efficient machine learning algorithms in stroke prediction: A systematic review,” Health Sci. Rep., vol. 7, no. 10, p. e70062, Oct. 2024, doi: 10.1002/hsr2.70062.
[2] S. K. Uma and S. R. Rakshith, “Stroke analysis using 10 ML comparison,” Int. J. Res. Appl. Sci. Eng. Technol., vol. 10, pp. 3857–3862, 2022.
[3] M. M. Islam et al., “Stroke prediction analysis using machine learning classifiers and feature technique,” Int. J. Electron. Commun. Syst., vol. 1, pp. 57–62, 2021.
[4] M. Alruily, S. A. El-Ghany, A. M. Mostafa, M. Ezz, and A. A. El-Aziz, “A-tuning ensemble machine learning technique for cerebral stroke prediction,” Appl. Sci., vol. 13, no. 5047, 2023.
[5] Z. Chen, “Stroke risk prediction based on machine learning algorithms,” Highlights Sci. Eng. Technol., vol. 38, pp. 932–941, 2023.
[6] T. M. Geethanjali, M. D. Divyashree, S. K. Monisha, and M. K. Sahana, “Stroke prediction using machine learning,” Int. J. Emerg. Technol. Innov. Res., vol. 8, pp. 710–717, 2021.
[7] D. Paul, G. Gain, and S. Orang, “Advanced random forest ensemble for stroke prediction,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 11, no. 3, 2022, doi: 10.17148/IJARCCE.2022.11343.
[8] P. S. Mattas, “Brain stroke prediction using machine learning,” Int. J. Res. Publ. Rev., vol. 3, pp. 711–722, 2022.
[9] M. S. Pathan, Z. Jianbiao, D. John, A. Nag, and S. Dev, “Identifying stroke indicators using rough sets,” IEEE Access, vol. 8, pp. 210318–210327, 2020.
[10] M. U. Emon et al., “Performance analysis of machine learning approaches in stroke prediction,” in 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020.
[11] S. Dev et al., “A predictive analytics approach for stroke prediction using machine learning and neural networks,” Healthc. Anal., vol. 2, p. 100032, 2022.
[12] L. P. Nguyen et al., “The utilization of machine learning algorithms for assisting physicians in the diagnosis of diabetes,” Diagnostics, vol. 13, no. 2087, 2023.
[13] N. Hatami et al., “A novel autoencoders-LSTM model for stroke outcome prediction using multimodal MRI data,” arXiv preprint arXiv:2303.09484, 2023.
[14] L. García-Terriza, J. L. Risco-Martín, G. Reig Roselló, and J. L. Ayala, “Predictive and diagnosis models of stroke from hemodynamic signal monitoring,” arXiv preprint arXiv:2306.05289, 2023.
[15] C. Fernandez-Lozano et al., “Random forest-based prediction of stroke outcome,” arXiv preprint arXiv:2402.00638, 2024.

[16] A. Pinto et al., "Combining unsupervised and supervised learning for predicting the final stroke lesion," arXiv preprint arXiv:2101.00489, 2021.
[17] S. Golemati and K. Nikita, "Cardiovascular computing-methodologies and clinical applications," Springer, 2019.
[18] A. Gastounioti et al., "A novel computerized tool to stratify risk in carotid atherosclerosis using kinematic features of the arterial wall," IEEE J. Biomed. Health Inform., vol. 18, no. 5, pp. 1472–1482, 2014.
[19] S. Golemati et al., "Toward novel noninvasive and low-cost markers for predicting strokes in asymptomatic carotid atherosclerosis: the role of ultrasound image analysis," IEEE Trans. Biomed. Eng., vol. 60, no. 3, pp. 717–726, 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY