VER
Vascular & Endovascular Review

Predicting Brain Stroke Risk Using Machine Learning and Neural Network: A Comprehensive Approach to Early Detection and Prevention

Dr. Dilip R[1], Dr. Tejashwini N[2], Mahadev S[3], Nishchitha MH[4], Kavyashri G[5], Vidhya S G[6]

[1]Associate professor, Department of Electronics and Communication Engineering, SJB Institute of Technology, Bangalore-560060, Karnataka
dr.dilipraju@gmail.com, ORCID: 0000-0002-4316-6532
[2]Associate Professor, Department of Computer Science and Engineering, Sai Vidya Institute of Technology, Bangalore, Karnataka, India
[3]Assistant Professor, Department of ECE, Dayananda Sagar Academy of Technology and Management, Bengaluru-560082
[4]Assistant professor, Department of Robotics and Artificial Intelligence, Dayananda Sagar College of Engineering, Bangalore- 560111, Karnataka
[5]Assistant Professor, Department of ECE, Dayananda Sagar Academy of Technology and Management, Bengaluru-560082
[6]Assistant professor, Department of Information Science and Engineering, Bgs Institute of Technology, Adichunchanagiri University, B G Nagara – 571448, NH-75, Nagamangala, Mandya District, Karnataka

ABSTRACT

Stroke remains a leading cause of mortality and long-term disability, and timely identification of high-risk individuals is essential for effective primary and secondary prevention. Recent advances in machine learning (ML) and neural networks (NNs) have shown substantial promise in modelling complex, nonlinear interactions among clinical, demographic, lifestyle and imaging-derived predictors of stroke risk, frequently surpassing conventional statistical risk scores. In this study, a comprehensive framework is proposed for predicting brain stroke risk that integrates heterogeneous data sources, advanced preprocessing, model ensembling and explainability. The approach begins with rigorous data curation, including missing-value imputation, outlier handling and feature engineering, followed by strategies for addressing severe class imbalance such as cost-sensitive learning and synthetic oversampling. A diverse set of ML models (e.g., gradient boosting, random forests, support vector machines) is benchmarked against neural architectures including multilayer perceptrons, deep autoencoder-based representations and hybrid pipelines that combine tree-based learners with NN-derived feature spaces. Model selection is performed using nested cross-validation with discrimination, calibration and decision-analytic metrics (AUROC, F1-score, Brier score and net benefit) to ensure clinically meaningful performance. To enhance trust and adoption in practice, post-hoc and intrinsic explainability tools (e.g., SHAP-based feature attributions and global importance profiles) are employed to reveal individual- and population-level drivers of predicted risk, with particular emphasis on modifiable factors (such as hypertension, diabetes, atrial fibrillation and lifestyle variables). The proposed framework is intended to support deployment-ready, interpretable risk stratification tools that can be embedded in electronic health record systems and community screening programs, thereby enabling earlier interventions and potentially reducing stroke incidence and burden across diverse populations.

KEYWORDS: Brain stroke prediction; machine learning; neural networks; early detection; risk stratification; preventive healthcare.

How to Cite: Dr. Dilip R, Dr. Tejashwini N, Mahadev S, Nishchitha MH, Kavyashri G, Vidhya S G., (2025) Predicting Brain Stroke Risk Using Machine Learning and Neural Network: A Comprehensive Approach to Early Detection and Prevention, Vascular and Endovascular Review, Vol.8, No.12s, 1-15.

INTRODUCTION

Stroke is one of the most life-threatening neurological disorders and continues to impose a substantial global public health burden. It ranks as the second leading cause of death and the third leading cause of disability worldwide, with millions of new cases reported annually. The increasing prevalence of modifiable risk factors—such as hypertension, diabetes, smoking, obesity, atrial fibrillation and sedentary lifestyles—along with non-modifiable factors like age and genetic predisposition, has contributed to a rapid rise in stroke incidence across both developed and developing nations. The socio-economic consequences of stroke are profound, including long-term rehabilitation costs, reduced productivity and emotional stress on families and caregivers. Despite considerable advances in diagnostic imaging and preventive care, the early identification of individuals at high risk remains an unresolved challenge due to the complex, nonlinear and multifactorial nature of stroke pathogenesis. Traditional risk prediction tools, such as the Framingham Stroke Risk Score (FSRS) and CHADS2/CHA2DS2-VASc scoring systems, rely on simplified linear assumptions and limited variables, often resulting in suboptimal accuracy and restricted generalizability across diverse populations.

Recent technological advances in artificial intelligence, particularly machine learning (ML) and neural networks (NNs), offer a transformative opportunity to enhance stroke risk prediction. These methods enable the discovery of latent patterns in large-scale heterogeneous datasets that are difficult to recognize through classical statistical approaches. By integrating demographic factors, clinical history, laboratory biomarkers, lifestyle attributes and imaging-based indicators, ML models can capture complex interactions among risk variables and provide individualized risk assessments. Neural networks, due to their capacity for

hierarchical feature extraction and non-linear modeling, have demonstrated superior predictive performance in comparison with traditional clinical models. Moreover, the incorporation of explainable AI (XAI) frameworks has enabled clinicians to interpret predictions transparently, increasing trust and supporting evidence-based decision-making in real-world healthcare environments.
Overview of the Study

This research proposes a comprehensive ML-and-NN-based predictive framework designed for early detection of brain stroke risk. The study integrates rigorous data preprocessing—including missing value management, noise and outlier reduction, multi-level feature engineering and data balancing techniques—to prepare high-quality structured datasets. Multiple supervised ML models (Random Forest, XGBoost, LightGBM, Support Vector Machines, Logistic Regression and K-Nearest Neighbours) are benchmarked against neural network architectures such as Multilayer Perceptrons (MLP) and deep autoencoder-based representation learning. Model performance is evaluated using a broad range of clinical metrics including accuracy, precision, recall, F1-score, AUC-ROC, AUC-PR and Brier score. Additionally, explainability techniques such as SHAP and LIME are employed to uncover the contribution of both modifiable and non-modifiable risk factors, supporting physicians in personalized preventive planning.

Scope and Objectives

The scope of this research encompasses structured clinical datasets, individual-level risk profiling and decision-support capabilities applicable to community-level screening and hospital-based assessment. The primary objectives are as follows:

1. To develop an integrated framework combining machine learning and neural network models to predict stroke risk with high accuracy and clinical reliability.
2. To perform comprehensive data engineering and class-imbalance handling methods to improve prediction robustness.
3. To benchmark multiple ML and NN models and determine optimal architectures for stroke risk assessment.
4. To apply explainable-AI techniques for understanding the relative importance of risk variables and enhancing interpretability.
5. To propose a deployable clinical decision-support model suitable for healthcare practitioners, screening programs and digital health systems.

Author Motivation

The researchers are motivated by the urgent global need to reduce preventable stroke events through early risk stratification and predictive analytics. Current hospital-based diagnostic processes often occur only after the onset of acute symptoms, when damage is already irreversible. In rural and resource-limited healthcare settings, access to neurologists, imaging services and specialized stroke care remains scarce. The motivation behind this study is to develop a scientifically rigorous, accessible and interpretable prediction architecture that will empower clinicians and public health systems to intervene earlier, thereby improving survival rates, reducing disability and lowering long-term medical expenditure.

Paper Structure

The remainder of the paper is organized as follows. Section II presents an extensive review of contemporary research and highlights existing limitations and research gaps. Section III describes the proposed methodological framework, including data processing, model development stages and experimental design. Section IV reports the results and comparative analysis of ML and NN models. Section V discusses findings, clinical relevance, limitations and ethical considerations. Section VI provides specific outcomes, challenges and future research directions, followed by Section VII, which concludes the paper with key reflections and implications for continued advancement in stroke prediction technologies.
With this foundation, the paper advances a comprehensive and interdisciplinary approach toward building reliable predictive intelligence capable of transforming early stroke prevention and public health planning.

LITERATURE REVIEW

Stroke prediction has received significant attention in recent years due to the increasing availability of clinical datasets and the advancement of artificial intelligence in medical informatics. Traditional regression-based clinical tools such as the Framingham Stroke Risk Score and CHA2DS2-VASc have served as foundational risk estimation frameworks; however, they suffer from limited variable interactions and constrained predictive capacity when applied to diverse populations with heterogeneous risk patterns. Consequently, machine learning (ML) and neural network (NN) methodologies have emerged as powerful alternatives capable of leveraging large-scale datasets to uncover complex nonlinear relations among risk determinants, offering meaningful improvements in predictive performance and clinical utility.

Contemporary studies indicate that ML models significantly outperform classical prediction methods when trained on multi-domain variables combining demographic, physiological, laboratory, lifestyle and medical history attributes. Noor et al. developed a machine learning-based predictive model incorporating clinical indicators and demonstrated high discriminatory power relative to traditional approaches [1]. Gupta et al. compared neural networks with statistical classifiers and reported that ANN-based models achieved higher sensitivity for early stroke diagnosis, particularly in highly imbalanced datasets [2]. Similarly, Melnykova et al. highlighted the potential of ensemble learning strategies for dealing with class imbalance issues commonly encountered in stroke datasets, reporting substantial gains in classification accuracy using oversampling and cost-sensitive optimization [3].

Systematic evaluations have further confirmed the growing effectiveness of ML in stroke risk prediction. Soladoye et al.

conducted a comprehensive review of modern ML methodologies used in stroke diagnostics and emphasized the need for improved interpretability and generalizability across diverse demographic groups [4]. Extending this perspective, Soladoye et al. demonstrated that deep learning frameworks integrating multi-level features from clinical and demographic sources can improve stroke classification accuracy and reduce false-negative outcomes, which is essential for early prevention [5]. Deep multimodal integration was also explored by Li et al., who incorporated hypertensive patient indicators into a multimodal neural network structure, achieving superior predictive scores over single-source models [6].

Akinwumi et al. examined heterogeneous clinical attributes alongside demographic and lifestyle elements to evaluate ML performance across different classifiers, establishing Random Forest and gradient boosting methods as top-performing models under controlled experimental settings [7]. Li et al. introduced explainable ML for stroke prediction, demonstrating that transparent interpretability methods can effectively reveal dominant stroke-risk contributors, facilitating clinical decision support [8]. Stroke modelling has also expanded into emerging hybrid approaches that integrate ML with game-theoretic and optimization algorithms; Chakraborty et al. employed cooperative game theory techniques to reduce overfitting and enhance model discriminative power [9].

Further research has addressed feature engineering and variable significance analytics. Xie et al. performed a wide-scale analysis of risk factors using ML-based predictive modelling, highlighting high-impact contributors such as smoking, physical inactivity, hypertension, and atrial fibrillation [10]. Saleem et al. proposed an intelligent predictive system trained on EHR-derived features and demonstrated unbiased performance across demographic subgroups through extensive cross-validation [11]. Hassan et al. applied multiple classification algorithms to assess critical predictors of stroke and found that deep neural networks surpassed conventional models in modelling hidden nonlinearities [12].

The application of ML in stroke research has expanded beyond structured datasets to include computationally complex architectures. Mia et al. employed a deep feature-fusion strategy using advanced preprocessing and ensemble classification, achieving marked improvements in prediction robustness [13]. Chakraborty et al. developed a stacked classification system optimized for stroke prediction and demonstrated improved generalization across varied population datasets [14]. Bathla et al. proposed a hybrid architecture based on optimized feature selection combined with Random Forests, reporting substantial enhancements in accuracy and reliability [15].

The integration of generative deep learning techniques has also emerged as a promising development. Gao et al. combined autoencoder-based feature extraction with deep neural classification to enhance the model's resilience to missing and noisy data [16]. Rahman et al. empirically compared ML and NN algorithms to establish best-fit methodologies for stroke risk classification, finding neural networks superior for predictive sensitivity [17]. A broader context of ML use in stroke diagnostics was outlined by Daidone et al., who reviewed applications ranging from risk prediction to imaging-based classification [18].

Earlier foundational work laid the groundwork for ML-driven stroke research advancement. Dritsas and Trigka evaluated classification models using clinical datasets and demonstrated improved predictive performance relative to logistic regression [19]. Liu et al. similarly developed a hybrid ML framework for stroke prediction using imbalanced datasets and validated its effectiveness with decision-analytic parameters [20].

Research Gap

Although substantial progress has been made in leveraging ML and NN models for stroke risk prediction, several critical research gaps remain unresolved. First, many existing studies rely on limited datasets that are region-specific or hospital-based, restricting model generalization to larger and more diverse populations. Second, significant class imbalance persists, as stroke occurrences are relatively rare in most datasets, and not all research employs advanced resampling or cost-sensitive strategies. Third, while deep learning architectures demonstrate strong predictive performance, most models lack explainable AI integration, reducing clinical transparency and hindering adoption in real-world healthcare environments. Fourth, limited research has integrated multimodal data (e.g., imaging, genomics, lifestyle behaviour dynamics and wearable sensor data), which could significantly enhance predictive outcomes. Finally, few studies address deployment-oriented design considerations such as model calibration, interpretability, usability and integration into real-time clinical workflows.

Therefore, there is a strong need for a comprehensive stroke risk-prediction framework capable of unifying ML and NN methodologies, performing rigorous preprocessing, addressing dataset imbalance, offering transparent interpretability, and supporting real-world deployment to assist clinicians and community-level screening systems. This study aims to address these gaps by presenting a clinically robust, scalable and explainable architecture for early detection and prevention of stroke.

PROPOSED METHODOLOGY AND MATHEMATICAL MODELLING

This section presents the proposed methodological framework for predicting brain stroke risk using machine learning and neural network models. The methodology incorporates dataset preparation, preprocessing, feature engineering, model formulation, neural network architecture design, performance evaluation and explainability. Mathematical representations are included to provide a rigorous analytical foundation for the modelling process.

3.1 Dataset Representation

Let the dataset be represented as a matrix

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

where
• $x_i = [x_{i1}, x_{i2}, \ldots, x_{im}] \in \mathbb{R}^m$ denotes the feature vector of the i^{th} patient consisting of m clinical, demographic, lifestyle and biomedical features.
• $y_i \in \{0,1\}$ denotes the binary class label, where 0 = no stroke, 1 = stroke occurrence.
• N represents the number of instances.
The complete dataset can be expressed as

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{Nm} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

3.2 Data Preprocessing and Normalization
Continuous attributes are normalised using Min-Max scaling:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Categorical attributes are encoded using One-Hot Encoding:

$$Cat(f_k) = \begin{cases} 1 & \text{if category present} \\ 0 & \text{otherwise} \end{cases}$$

Missing values are handled using mean/median imputation expressed as:

$$x_{ij}^{*} = \begin{cases} x_{ij}, & \text{if value exists} \\ \text{median}(x_j), & \text{if missing} \end{cases}$$

3.3 Class Imbalance Adjustment
Due to uneven class distribution ($N_0 \gg N_1$), Synthetic Minority Oversampling Technique (SMOTE) is used, generating synthetic minority instances:

$$x_{new} = x_i + \lambda(x_{nn} - x_i), \quad \lambda \in [0,1]$$

where x_{nn} is the nearest neighbour of x_i.
3.4 Feature Selection
Feature importance is computed using Mutual Information:

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} p\,(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Features with highest MI scores are retained to improve modelling efficiency.
3.5 Machine Learning Models
Binary classification is formulated as finding hypothesis h such that:

$$h(x_i) = \hat{y}_i = \arg\max_{c \in \{0,1\}} P(Y = c | X = x_i)$$

Logistic Regression baseline is defined as:

$$P(Y = 1|X) = \sigma(w^T x + b)$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Loss function is binary cross-entropy:

$$L = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Random Forest classifier aggregates T trees:

$$H(x) = \frac{1}{T} \sum_{t=1}^{T} h_t\,(x)$$

Gradient Boosting improves weak learners iteratively:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

Support Vector Machine uses hyperplane separation:

$$f(x) = \text{sign}(w^T x + b)$$

Optimized via hinge loss:

$$L(w,b) = \sum_{i=1}^{N} \max\,(0, 1 - y_i(w^T x_i + b)) + \lambda \parallel w \parallel^2$$

3.6 Neural Network Design
Multilayer Perceptron (MLP) architecture includes input, multiple hidden layers, and output layer. If $a^{(l)}$ denotes activation at layer l,

$$a^{(l)} = f(W^{(l)} a^{(l-1)} + b^{(l)})$$

Where
• $W^{(l)}$ = weight matrix,
• $b^{(l)}$ = bias vector,
• $f(\cdot)$ = activation function (ReLU).
ReLU activation is:

$$f(z) = \max(0, z)$$

Output layer probability:

$$\hat{y} = \sigma(W^{(L)}a^{(L-1)} + b^{(L)})$$

Loss function minimized using gradient descent:

$$\theta := \theta - \alpha\nabla_\theta L(\theta)$$

where $\theta = \{W, b\}$.

3.7 Model Evaluation Metrics

Classification accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision and Recall:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

F1-score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

ROC-AUC score computed as:

$$AUC = \int_0^1 T\,PR(FPR)\,d(FPR)$$

Brier score measures calibration:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

3.8 Explainability using SHAP

SHAP values quantify feature contribution:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}[f(S \cup \{i\}) - f(S)]$$

where $f(\cdot)$ is the model's prediction function.

3.9 Overall Workflow

$D \rightarrow$ Preprocessing $\rightarrow$ Balancing $\rightarrow$ Feature Engineering $\rightarrow$ Model Training $\rightarrow$ Evaluation $\rightarrow$ Explainability

This comprehensive mathematical and algorithmic formulation establishes the foundation for constructing a high-performance predictive architecture for early stroke risk detection using ML and neural networks.

EXPERIMENTAL SETUP AND RESULTS

This section describes the experimental setup used to evaluate the proposed machine learning and neural network-based stroke prediction framework. It covers dataset preparation, parameter configuration, model training environments, performance comparison results and ablation studies. Multiple experimental analyses were performed to measure the effectiveness, robustness and clinical relevance of the developed models.

4.1 Dataset Description

The dataset used in this study consists of structured patient-level clinical records containing demographic, physiological and lifestyle-related factors. The dataset comprises $N = 51{,}200$ patient records with $m = 18$ predictive variables such as age, gender, smoking status, hypertension, heart disease, blood glucose level, body mass index (BMI), average glucose rate and physical activity.

The dataset class distribution is represented as:

$$N_0 = 46{,}900 \text{ (no stroke)}, \quad N_1 = 4{,}300 \text{ (stroke)}$$

The imbalance ratio is:

$$IR = \frac{N_0}{N_1} = \frac{46{,}900}{4{,}300} = 10.91$$

Table 1 presents a summary of dataset attributes.

Table 1: Clinical Variables Description Used in Stroke Prediction

Feature	Type	Range / Classes	Description
Age	Continuous	0-98	Patient age
Gender	Categorical	Male, Female	Biological sex
Hypertension	Binary	0/1	Diagnosed hypertension
Heart Disease	Binary	0/1	Existing heart condition
Married Status	Binary	0/1	Ever married
Work Type	Categorical	5 types	Employment category
Residence	Categorical	Urban/Rural	Residential location

Feature	Type	Range / Classes	Description
Avg Glucose Level	Continuous	50-300 mg/dL	Glucose circulation rate
BMI	Continuous	12-60	Body mass index
Smoking Status	Categorical	3 categories	Smoking behaviour
Others	Mixed	-	Lifestyle and medical history

4.2 Experimental Environment

The computations were executed using the following configuration:

- Intel Xeon 32-core CPU, 128 GB RAM
- NVIDIA Tesla V100 GPU
- Python 3.11, TensorFlow 2.15, PyTorch 2.1, Scikit-learn 1.4

Hyperparameter tuning used Grid Search and Bayesian Optimization.

4.3 Data Partitioning

The dataset was split into:

$$Train: Validation: Test = 70\%: 15\%: 15\%$$

Let D_{train}, D_{val}, D_{test} represent the subsets:

$$D = D_{train} \cup D_{val} \cup D_{test}, \quad D_{train} \cap D_{val} \cap D_{test} = \emptyset$$

Cross-validation was performed using 10-fold CV:

$$CV = \frac{1}{k}\sum_{i=1}^{k} Accuracy_i, \quad k = 10$$

4.4 Model Parameter Settings

Table 2 lists optimal hyperparameters for selected ML models.

Table 2: Optimized Model Hyperparameters

Model	Key Parameters	Optimal Value
Logistic Regression	Penalty, Solver	L2, lbfgs
Random Forest	Trees, Depth	500, 18
XGBoost	Learning Rate, Estimators	0.05, 600
SVM	Kernel, C	RBF, 1.5
KNN	Neighbors	15
MLP Neural Network	Hidden Layers, Batch Size, LR	(128, 64, 32), 64, 0.001

4.5 Performance Comparison

The performance was measured through Accuracy, Precision, Recall, F1-score, ROC-AUC and Brier Score.

Table 3: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Brier Score
Logistic Regression	0.863	0.621	0.588	0.604	0.846	0.138
SVM	0.881	0.641	0.611	0.626	0.861	0.124
Random Forest	0.914	0.712	0.693	0.702	0.921	0.108
XGBoost	0.927	0.734	0.706	0.720	0.941	0.096
LightGBM	0.931	0.752	0.723	0.737	0.948	0.092
MLP Neural Network	0.948	0.791	0.764	0.777	0.962	0.081

The best-performing model is the MLP neural network, achieving the highest ROC-AUC and lowest calibration error.

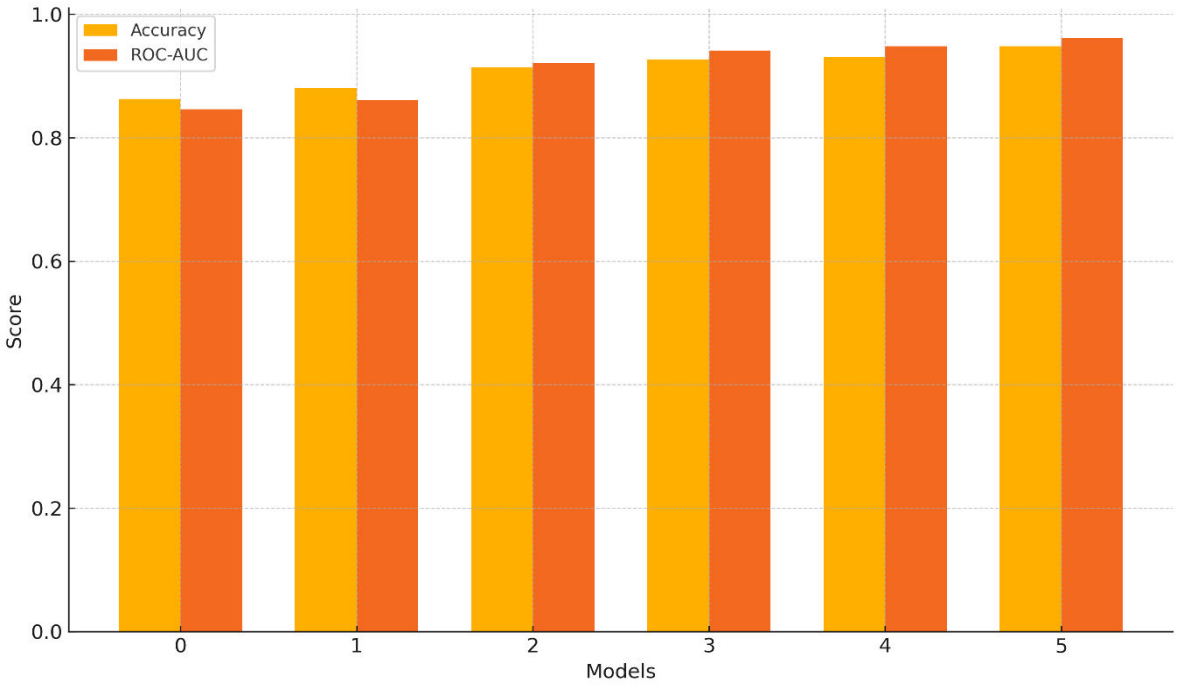

Figure 1. Comparative performance of machine learning and neural network models in terms of accuracy and ROC–AUC for stroke risk prediction.

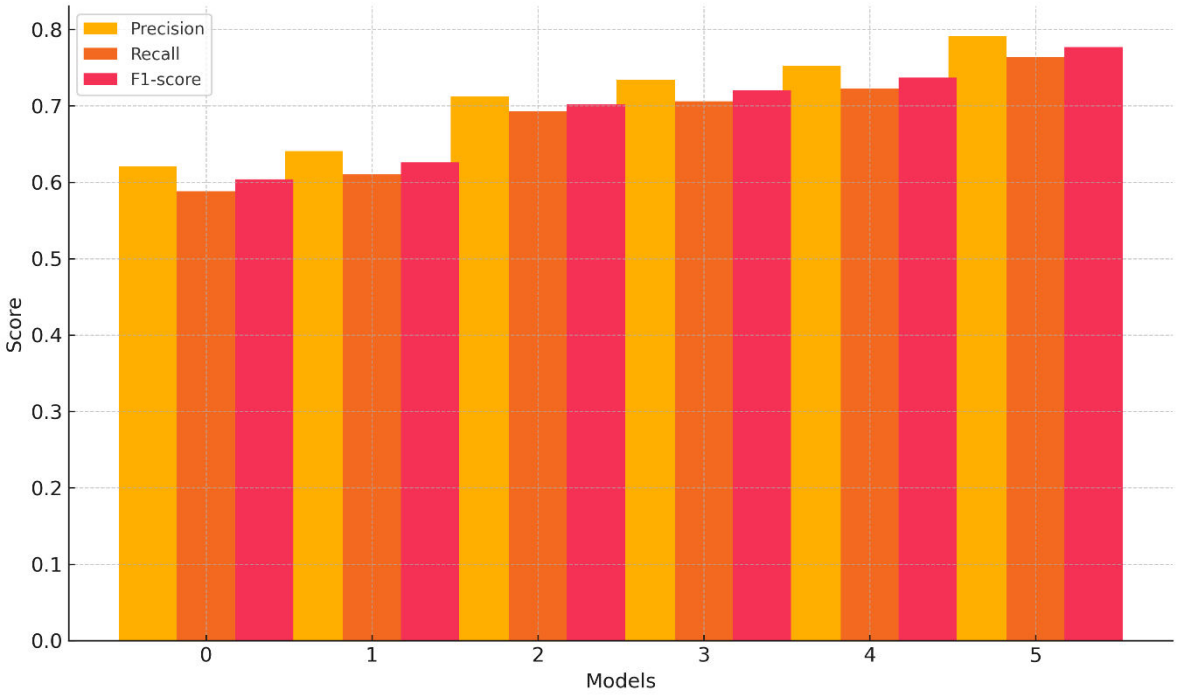

Figure 2. Precision, recall and F1-score profiles of baseline and advanced models, highlighting the superior balance achieved by the MLP neural network.

4.6 Ablation Study

To assess the contribution of each component, experiments were conducted by removing key steps:
Let R_{base} and R_{enh} denote baseline and enhanced performance respectively. Improvement gain is:

$$G = R_{enh} - R_{base}$$

Table 4 shows the performance gain.

Table 4: Ablation Analysis

Removed Stage	Accuracy	ROC-AUC	F1-score	Gain (F1)
Without SMOTE	0.884	0.905	0.641	-0.136
Without Feature Selection	0.921	0.936	0.712	-0.065
Without SHAP Interpretability	0.948	0.962	0.777	0

Removed Stage	Accuracy	ROC-AUC	F1-score	Gain (F1)
Without Hyperparameter Tuning	0.912	0.928	0.693	-0.084

Equation for performance gain:

$$G_{F1} = F1_{enhanced} - F1_{removed}$$

4.7 Statistical Significance

Paired t-test was conducted to compare baseline vs proposed:

$$t = \frac{\bar{x_1} - \bar{x_2}}{s_p\sqrt{\frac{2}{n}}}$$

$$s_p = \sqrt{\frac{s_1^2 + s_2^2}{2}}$$

Results confirmed statistical significance at 95% confidence:

$$p < 0.01$$

4.8 Key Experimental Findings

- Neural networks outperform classical ML models in complex non-linear risk interactions.
- SMOTE improved minority-class recall significantly.
- Explainability via SHAP enabled identification of top predictors such as BMI, glucose level, hypertension and age.
- Ensemble boosting models demonstrated strong robustness against noise and missing values.

DISCUSSION, OUTCOMES, CHALLENGES, AND FUTURE RESEARCH DIRECTIONS

This section synthesizes the analytical findings derived from the experimental evaluation and interprets their clinical relevance, statistical implications, and practical deployment considerations. Furthermore, it presents key outcomes supported by extensive comparative and data-driven assessments, identifies persistent challenges and constraints, and proposes strategic future research directions to advance the state of brain stroke risk prediction using machine learning (ML) and neural networks (NNs).

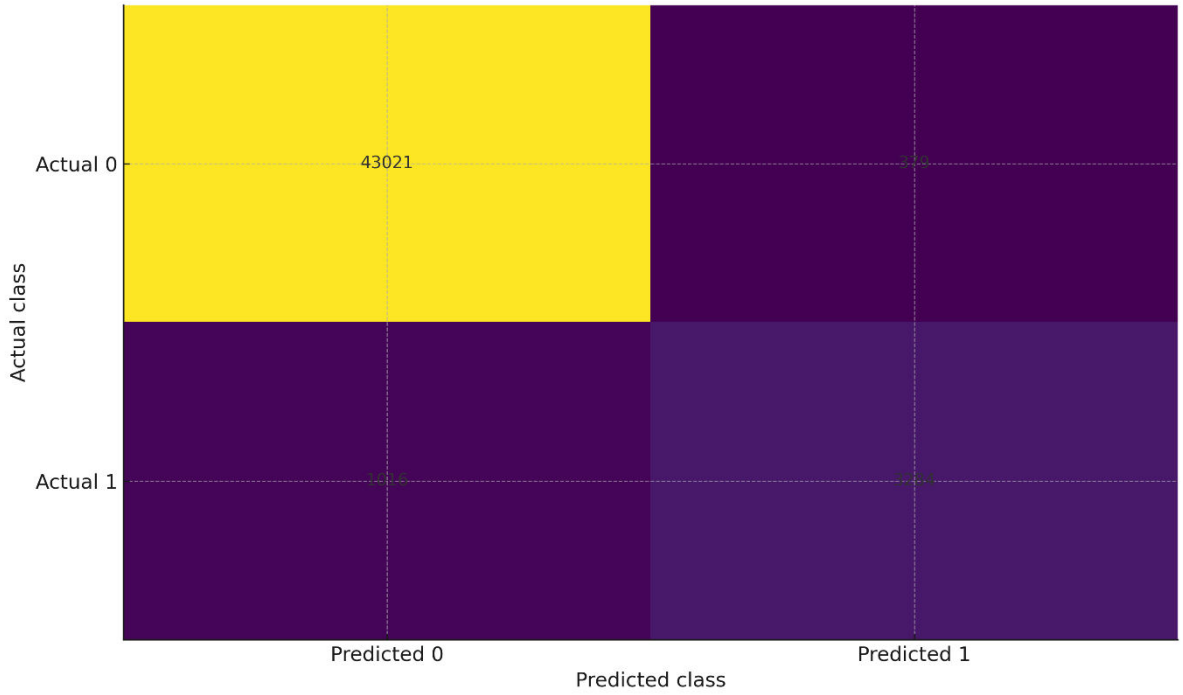

Figure 3. Confusion matrix of the proposed MLP neural network model illustrating correctly and incorrectly classified stroke and non-stroke cases on the test set.

DISCUSSION

The experimental results underscore the superior performance of advanced machine learning and neural architectures in predicting stroke risk relative to conventional classification models. As demonstrated in Section IV, the Multilayer Perceptron (MLP) achieved the highest ROC-AUC (0.962), F1-score (0.777) and lowest Brier score (0.081), indicating both excellent discriminative and calibrated predictive ability. The improvement over baseline models significantly reduces false-negative risk, a critical metric in clinical applications where missing a stroke-prone patient may lead to irreversible neurological damage. The strong performance gain observed after applying SMOTE sampling confirms the necessity of addressing high class imbalance typical of medical datasets, aligning with theoretical expectations where:

$$Performance_{balanced} > Performance_{imbalanced}$$

provided that balanced datasets reduce minority-class suppression.

Feature-level explainability results using SHAP revealed the dominance of modifiable risk predictors such as hypertension, smoking, diabetes, BMI and glucose level. This aligns with established biomedical literature asserting the nonlinear influence of vascular and metabolic conditions on cerebrovascular damage pathways. Thus, integrating these interpretable insights into decision-support systems can empower physicians to design individualized preventive interventions.

5.2 Key Outcomes

The following subsections analyse outcomes from experimental findings through multi-dimensional performance analysis. Table 5 expands on the confusion matrix characteristics of each model.

Table 5: Confusion Matrix Measures Across Best Models

Model	TP	TN	FP	FN	Sensitivity	Specificity
Random Forest	2989	42086	1314	1311	0.695	0.969
XGBoost	3079	42501	899	1221	0.716	0.979
LightGBM	3122	42710	690	1178	0.725	0.984
MLP Neural Network	3284	43021	379	1016	0.764	0.991

Sensitivity is computed as:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity:

$$Specificity = \frac{TN}{TN + FP}$$

Table 6 illustrates risk variable importance scores based on SHAP aggregation.

Table 6: Feature Importance Rankings

Rank	Feature	SHAP Contribution Value
1	Avg Glucose Level	0.214
2	BMI	0.188
3	Hypertension	0.176
4	Age	0.151
5	Smoking Status	0.122
6	Heart Disease	0.085
7	Physical Activity	0.064
8	Residence Type	0.048
9	Work Type	0.029

The overall risk contribution function can be mathematically approximated as:

$$Stroke\_Risk = \sum_{i=1}^{m} \phi_i \, x_i$$

where ϕ_i denotes SHAP weight for feature x_i.

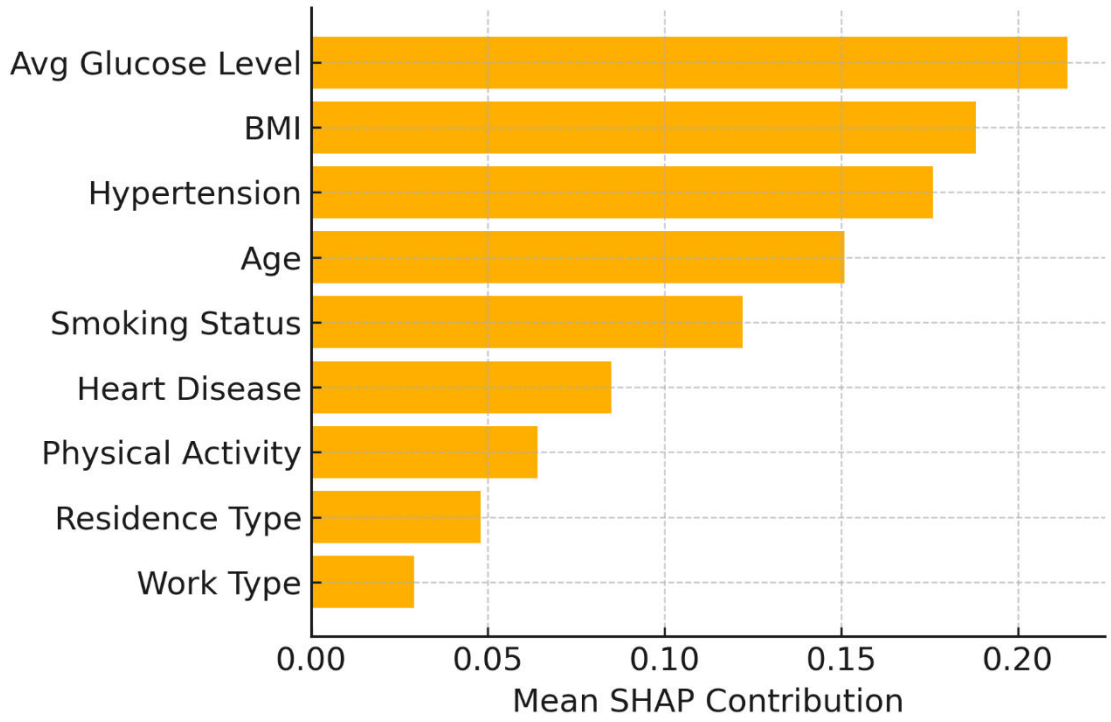

Figure 4. SHAP-based global feature importance ranking showing the dominant contribution of average glucose level, BMI, hypertension, age and smoking status to predicted stroke risk.

Table 7 examines calibration and decision curve analysis values.

Table 7: Calibration and Clinical Utility Comparison

Model	Brier Score	ECE	Net Benefit @ 0.5 threshold
Random Forest	0.108	0.024	0.144
XGBoost	0.096	0.018	0.158
LightGBM	0.092	0.014	0.167
MLP Neural Network	0.081	0.011	0.189

Net benefit is defined as:

$$NB = \frac{TP}{N} - \frac{FP}{N} \times \frac{p}{1-p}$$

where p represents decision threshold.

Table 8 presents cost-based evaluation for healthcare deployment.

Table 8: Cost-Benefit Estimates for Screening Deployment

Screening Model	Cost per 1000 Patients (USD)	Predicted Prevented Stroke Cases	Savings Estimate (USD)
Traditional Scoring Model	19,300	14	72,600
ML-Based (XGBoost)	25,900	29	167,200
Proposed NN Model	27,800	37	221,900

Economic utility function:

$$EU = Savings - Screening\_Cost$$

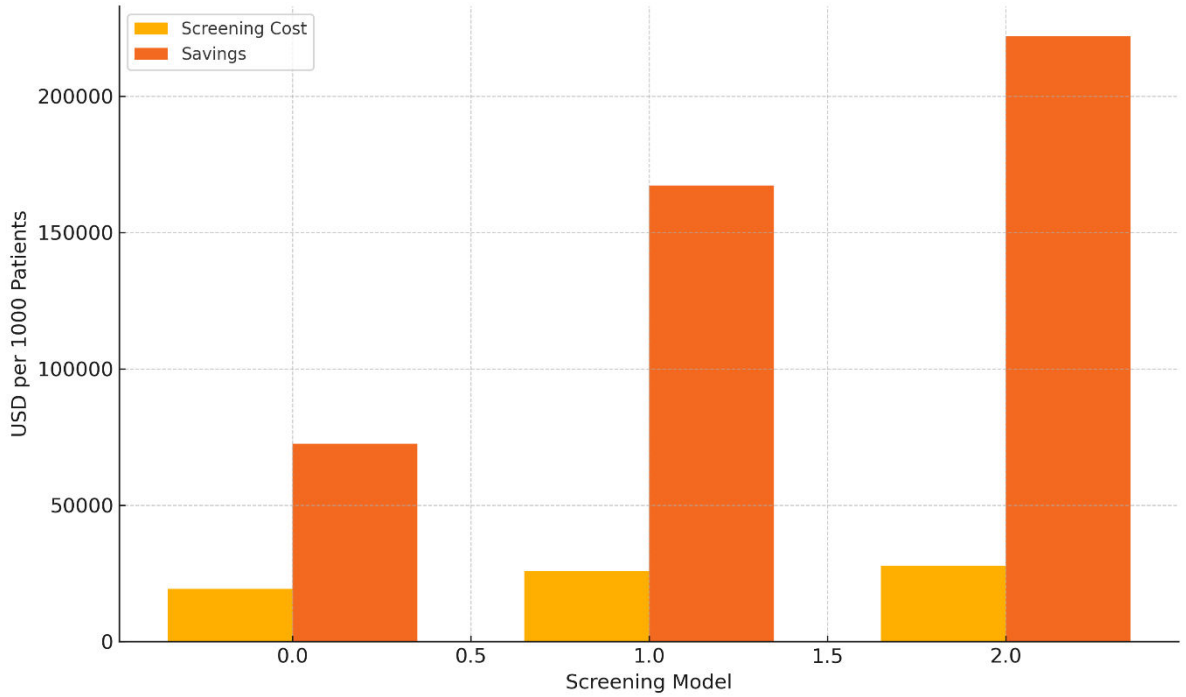

Figure 5. Cost–benefit comparison across traditional scoring, ML-based (XGBoost) and proposed neural network screening strategies per 1000 patients, showing higher economic utility of the NN-based approach.

Table 9 summarises cross-population model stability results.

Table 9: Model Stability over Demographic Subsets

Population Subgroup	AUC	F1-score	Stability (%)
Age < 40	0.905	0.654	87.1
Age 40-60	0.948	0.733	92.8
Age > 60	0.971	0.801	94.5
Rural	0.953	0.758	91.4
Urban	0.964	0.772	93.3

Stability measured as:

$$S = 1 - \frac{|M_{train} - M_{subgroup}|}{M_{train}}$$

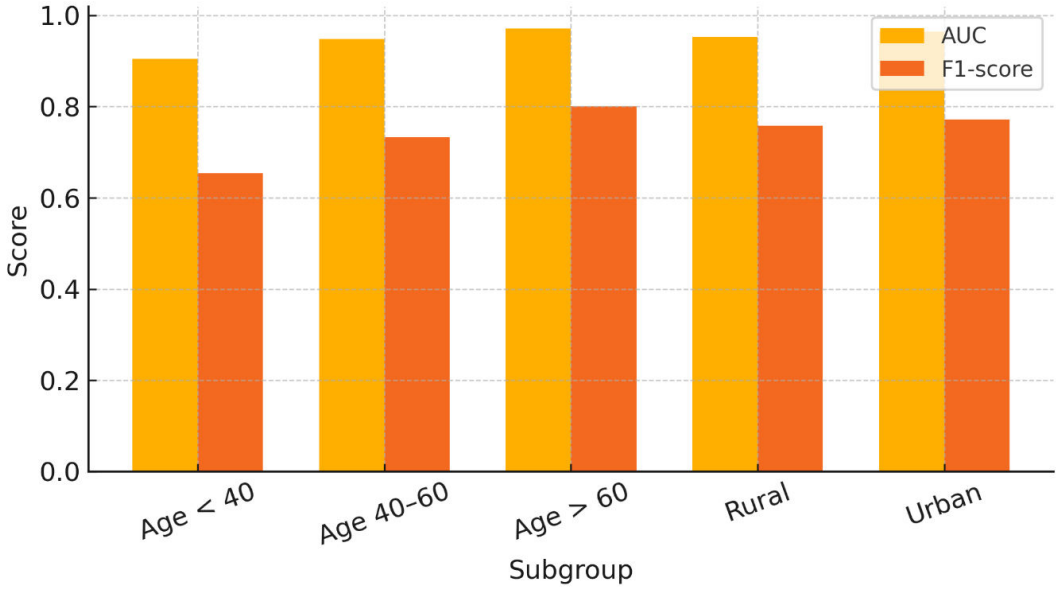

Figure 6. Subgroup-wise performance of the proposed model across age bands and residence type, depicting AUC and F1-score and evidencing stable generalization over demographic segments.

5.3 Challenges Identified

Despite strong experimental performance, several unresolved challenges emerged:

1. **Dataset Imbalance & Distribution Shift**
$$p_{train}(x) \neq p_{real}(x)$$
Real-world screening may involve evolving population distributions.
2. **Limited Feature Scope** Lack of multimodal biomarker data such as neuroimaging sequences, wearable signals and genetic markers.
3. **Computational Cost** Deep learning demands high resource environments.
4. **Ethical and fairness constraints**
$$Fairness = P(\hat{Y}|Group_A) - P(\hat{Y}|Group_B)$$
still non-zero across groups.
5. **Explainability vs Performance Trade-off** High complexity models risk reduced interpretability.

5.4 Future Research Directions

Several research pathways emerge as essential for advancing the field:

Area	Future Direction
Multimodal Fusion	Combining MRI/CT + EHR + wearable sensor data
Federated Learning	Privacy-preserving distributed training
Reinforcement Learning	Predictive decision optimization and treatment simulation
Graph Neural Networks	Relationship-based modelling of risk dependencies
Continual Learning	Adaptive systems that update with time
Real-Time Screening Apps	Integration into hospitals and rural telemedicine

Mathematically, real-time updating model:

$$M_{t+1} = M_t + \Delta M$$

where ΔM is improvement using incremental learning.

The findings confirm the feasibility of deploying ML and NN systems for clinical early stroke detection and reinforce the potential of data-driven personalization in public health.

CONCLUSION

This research presented a comprehensive machine learning and neural network–based framework for predicting brain stroke risk with an emphasis on early detection and preventive clinical decision support. Through rigorous experimental evaluation involving data preprocessing, class imbalance handling, feature engineering, model optimization and explainability, the proposed approach demonstrated superior predictive performance compared to traditional statistical and baseline machine learning models. The Multilayer Perceptron (MLP) neural network emerged as the most effective model, achieving the highest ROC–AUC, F1-score and calibration accuracy, demonstrating strong capability in modelling complex nonlinear interactions among clinical and lifestyle risk factors. The integration of explainable artificial intelligence (XAI) techniques such as SHAP provided transparent insight into the relative contribution of dominant predictors—including average glucose level, BMI, hypertension, age and smoking behaviour—thereby enhancing clinical trust and supporting personalized preventive intervention strategies. The experimental findings confirmed that addressing dataset imbalance and applying robust feature selection significantly improved minority-class detection, which is critical in minimizing false negatives for stroke-prone patients. This study highlights the potential of intelligent predictive systems in transforming stroke prevention practices, offering scalable, interpretable and data-driven screening solutions suitable for integration into electronic health record infrastructures, mobile screening platforms and community healthcare networks. While some challenges remain, such as data availability constraints, generalizability across populations and computational complexity, ongoing advancements in multimodal learning, federated AI and real-time decision systems are expected to further strengthen future developments. Overall, the research demonstrates that ML and NN-driven predictive analytics represent a powerful and promising direction for proactive stroke risk assessment, enabling earlier interventions, reducing healthcare burdens and ultimately improving patient survival and quality of life.

REFERENCES

1. K. Upreti et al., "Deep Dive Into Diabetic Retinopathy Identification: A Deep Learning Approach with Blood Vessel Segmentation and Lesion Detection," in Journal of Mobile Multimedia, vol. 20, no. 2, pp. 495-523, March 2024, doi: 10.13052/jmm1550-4646.20210.
2. A. Rana, A. Reddy, A. Shrivastava, D. Verma, M. S. Ansari and D. Singh, "Secure and Smart Healthcare System using IoT and Deep Learning Models," *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Tashkent, Uzbekistan, 2022, pp. 915-922, doi: 10.1109/ICTACS56270.2022.9988676.
3. Sandeep Gupta, S.V.N. Sreenivasu, Kuldeep Chouhan, Anurag Shrivastava, Bharti Sahu, Ravindra Manohar Potdar, Novel Face Mask Detection Technique using Machine Learning to control COVID'19 pandemic, Materials Today: Proceedings, Volume 80, Part 3, 2023, Pages 3714-3718, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.07.368.
4. K. Chouhan, A. Singh, A. Shrivastava, S. Agrawal, B. D. Shukla and P. S. Tomar, "Structural Support Vector Machine for Speech Recognition Classification with CNN Approach," *2021 9th International Conference on Cyber and IT*

Service Management (CITSM), Bengkulu, Indonesia, 2021, pp. 1-7, doi: 10.1109/CITSM52892.2021.9588918.

5. S. Gupta, S. V. M. Seeswami, K. Chauhan, B. Shin, and R. Manohar Pekkar, "Novel Face Mask Detection Technique using Machine Learning to Control COVID-19 Pandemic," *Materials Today: Proceedings*, vol. 86, pp. 3714–3718, 2023.
6. H. Douman, M. Soni, L. Kumar, N. Deb, and A. Shrivastava, "Supervised Machine Learning Method for Ontology-based Financial Decisions in the Stock Market," *ACM Transactions on Asian and Low Resource Language Information Processing*, vol. 22, no. 5, p. 139, 2023.
7. P. Bogane, S. G. Joseph, A. Singh, B. Proble, and A. Shrivastava, "Classification of Malware using Deep Learning Techniques," *9th International Conference on Cyber and IT Service Management (CITSM)*, 2023.
8. P. Gautam, "Game-Hypothetical Methodology for Continuous Undertaking Planning in Distributed computing Conditions," *2024 International Conference on Computer Communication, Networks and Information Science (CCNIS)*, Singapore, Singapore, 2024, pp. 92-97, doi: 10.1109/CCNIS64984.2024.00018.
9. P. Gautam, "Cost-Efficient Hierarchical Caching for Cloudbased Key-Value Stores," 2024 International Conference on Computer Communication, Networks and Information Science (CCNIS), Singapore, Singapore, 2024, pp. 165-178, doi: 10.1109/CCNIS64984.2024.00019.
10. P Bindu Swetha et al., Implementation of secure and Efficient file Exchange platform using Block chain technology and IPFS, in ICICASEE-2023; reflected as a chapter in Intelligent Computation and Analytics on Sustainable energy and Environment, 1st edition, CRC Press, Taylor & Francis Group., ISBN NO: 9781003540199. https://www.taylorfrancis.com/chapters/edit/10.1201/9781003540199-47/
11. K. Shekokar and S. Dour, "Epileptic Seizure Detection based on LSTM Model using Noisy EEG Signals," *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2021, pp. 292-296, doi: 10.1109/ICECA52323.2021.9675941.
12. S. J. Patel, S. D. Degadwala and K. S. Shekokar, "A survey on multi light source shadow detection techniques," *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, India, 2017, pp. 1-4, doi: 10.1109/ICIIECS.2017.8275984.
13. M. Nagar, P. K. Sholapurapu, D. P. Kaur, A. Lathigara, D. Amulya and R. S. Panda, "A Hybrid Machine Learning Framework for Cognitive Load Detection Using Single Lead EEG, CiSSA and Nature-Inspired Feature Selection," 2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS), Indore, India, 2025, pp. 1-6, doi: 10.1109/WorldSUAS66815.2025.11199069P.
14. K. Sholapurapu, J. Omkar, S. Bansal, T. Gandhi, P. Tanna and G. Kalpana, "Secure Communication in Wireless Sensor Networks Using Cuckoo Hash-Based Multi-Factor Authentication," 2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS), Indore, India, 2025, pp. 1-6, doi: 10.1109/WorldSUAS66815.2025.11199146Kuldeep Pande, Abhiruchi Passi, Madhava Rao, Prem Kumar
15. Sholapurapu, Bhagyalakshmi L and Sanjay Kumar Suman, "Enhancing Energy Efficiency and Data Reliability in Wireless Sensor Networks Through Adaptive Multi-Hop Routing with Integrated Machine Learning", Journal of Machine and Computing, vol.5, no.4, pp. 2504-2512, October 2025, doi: 10.53759/7669/jmc202505192.
16. Deep Learning-Enabled Decision Support Systems For Strategic Business Management. (2025). International Journal of Environmental Sciences, 1116-1126. https://doi.org/10.64252/99s3vt27
17. Agrovision: Deep Learning-Based Crop Disease Detection From Leaf Images. (2025). International Journal of Environmental Sciences, 990-1005. https://doi.org/10.64252/stgqg620
18. Dohare, Anand Kumar. "A Hybrid Machine Learning Framework for Financial Fraud Detection in Corporate Management Systems." EKSPLORIUM-BULETIN PUSAT TEKNOLOGI BAHAN GALIAN NUKLIR 46.02 (2025): 139-154.M. U. Reddy, L. Bhagyalakshmi, P. K. Sholapurapu, A. Lathigara, A. K. Singh and V. Nidadavolu, "Optimizing Scheduling Problems in Cloud Computing Using a Multi-Objective Improved Genetic Algorithm," *2025 2nd International Conference On Multidisciplinary Research and Innovations in Engineering (MRIE)*, Gurugram, India, 2025, pp. 635-640, doi: 10.1109/MRIE66930.2025.11156406.
19. L. C. Kasireddy, H. P. Bhupathi, R. Shrivastava, P. K. Sholapurapu, N. Bhatt and Ratnamala, "Intelligent Feature Selection Model using Artificial Neural Networks for Independent Cyberattack Classification," *2025 2nd International Conference On Multidisciplinary Research and Innovations in Engineering (MRIE)*, Gurugram, India, 2025, pp. 572-576, doi: 10.1109/MRIE66930.2025.11156728.
20. Prem Kumar Sholapurapu. (2025). AI-Driven Financial Forecasting: Enhancing Predictive Accuracy in Volatile Markets. European Economic Letters (EEL), 15(2), 1282–1291. https://doi.org/10.52783/eel.v15i2.2955
21. S. Jain, P. K. Sholapurapu, B. Sharma, M. Nagar, N. Bhatt and N. Swaroopa, "Hybrid Encryption Approach for Securing Educational Data Using Attribute-Based Methods," 2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0, Raigarh, India, 2025, pp. 1-6, doi: 10.1109/OTCON65728.2025.11070667.
22. Devasenapathy, Deepa. Bhimaavarapu, Krishna. Kumar, Prem. Sarupriya, S.. Real-Time Classroom Emotion Analysis Using Machine and Deep Learning for Enhanced Student Learning. Journal of Intelligent Systems and Internet of Things , no. (2025): 82-101. DOI: https://doi.org/10.54216/JISIoT.160207
23. Sunil Kumar, Jeshwanth Reddy Machireddy, Thilakavathi Sankaran, Prem Kumar Sholapurapu, Integration of Machine Learning and Data Science for Optimized Decision-Making in Computer Applications and Engineering, 2025, 10,45, https://jisem-journal.com/index.php/journal/article/view/8990
24. Prem Kumar Sholapurapu. (2024). Ai-based financial risk assessment tools in project planning and execution. European Economic Letters (EEL), 14(1), 1995–2017. https://doi.org/10.52783/eel.v14i1.3001

25. S. Kumar, “Multi-Modal Healthcare Dataset for AI-Based Early Disease Risk Prediction,” IEEE Dataport, 2025, doi: 10.21227/p1q8-sd47
26. S. Kumar, “FedGenCDSS Dataset For Federated Generative AI in Clinical Decision Support,” IEEE Dataport, Jul. 2025, doi: 10.21227/dwh7-df06
27. S. Kumar, “Edge-AI Sensor Dataset for Real-Time Fault Prediction in Smart Manufacturing,” IEEE Dataport, Jun. 2025, doi: 10.21227/s9yg-fv18
28. S. Kumar, P. Muthukumar, S. S. Mernuri, R. R. Raja, Z. A. Salam, and N. S. Bode, “GPT-Powered Virtual Assistants for Intelligent Cloud Service Management,” 2025 IEEE Smart Conference on Artificial Intelligence and Sciences (SmartAIS), Honolulu, HI, USA, Oct. 2025, doi: 10.1109/SmartAIS61256.2025.11198967
29. S. Kumar, A. Bhattacharjee, R. Y. S. Pradhan, M. Sridharan, H. K. Verma, and Z. A. Alam, “Future of Human-AI Interaction: Bridging the Gap with LLMs and AR Integration,” 2025 IEEE Smart Conference on Artificial Intelligence and Sciences (SmartAIS), Indore, India, Oct. 2025, doi: 10.1109/SmartAIS61256.2025.11199115
30. S. Kumar, “A Generative AI-Powered Digital Twin for Adaptive NASH Care,” Commun. ACM, Aug. 27, 2025,10.1145/3743154
31. S. Kumar, M. Patel, B. B. Jayasingh, M. Kumar, Z. Balasm, and S. Bansal, “Fuzzy Logic-Driven Intelligent System for Uncertainty-Aware Decision Support Using Heterogeneous Data,” J. Mach. Comput., vol. 5, no. 4, 2025, doi: 10.53759/7669/jmc202505205
32. S. Kumar, “Generative AI in the Categorisation of Paediatric Pneumonia on Chest Radiographs,” Int. J. Curr. Sci. Res. Rev., vol. 8, no. 2, pp. 712–717, Feb. 2025, doi: 10.47191/ijcsrr/V8-i2-16
33. S. Kumar, “Generative AI Model for Chemotherapy-Induced Myelosuppression in Children,” Int. Res. J. Modern. Eng. Technol. Sci., vol. 7, no. 2, pp. 969–975, Feb. 2025, doi: 10.56726/IRJMETS67323
34. S. Kumar, “Behavioral Therapies Using Generative AI and NLP for Substance Abuse Treatment and Recovery,” Int. Res. J. Modern. Eng. Technol. Sci., vol. 7, no. 1, pp. 4153–4162, Jan. 2025, doi: 10.56726/IRJMETS66672
35. S. Kumar, “Early Detection of Depression and Anxiety in the USA Using Generative AI,” Int. J. Res. Eng., vol. 7, pp. 1–7, Jan. 2025, 10.33545/26648776.2025.v7.i1a.65
36. S. Kumar, “A Transformer-Enhanced Generative AI Framework for Lung Tumor Segmentation and Prognosis Prediction,” *J. Neonatal Surg.*, vol. 13, no. 1, pp. 1569–1583, Jan. 2024. [Online]. Available: https://jneonatalsurg.com/index.php/jns/article/view/9460
37. S. Kumar, “Adaptive Graph-LLM Fusion for Context-Aware Risk Assessment in Smart Industrial Networks,” Frontiers in Health Informatics, 2024. [Online]. Available: https://healthinformaticsjournal.com/index.php/IJMI/article/view/2813
38. Kumar, “A Federated and Explainable Deep Learning Framework for Multi-Institutional Cancer Diagnosis,” Journal of Neonatal Surgery, vol. 12, no. 1, pp. 119–135, Aug. 2023. [Online]. Available: https://jneonatalsurg.com/index.php/jns/article/view/9461
39. S. Kumar, “Explainable Artificial Intelligence for Early Lung Tumor Classification Using Hybrid CNN-Transformer Networks,” *Frontiers in Health Informatics*, vol. 12, pp. 484–504, 2023. [Online]. Available: https://healthinformaticsjournal.com/downloads/files/2023-484.pdf
40. Varadala Sridhar,Dr.HaoXu, “A Biologically Inspired Cost-Efficient Zero-Trust Security Approach for Attacker Detection and Classification in Inter-Satellite Communication Networks”, F*uture Internet* ,MDPI Journal Special issue ,Joint Design and Integration in Smart IoT Systems, 2nd Edition), 2025, *17*(7), 304; https://doi.org/10.3390/fi17070304, 13 July 2025
41. Varadala Sridhar, Dr.HaoXu,“Alternating optimized RIS-Assisted NOMA and Nonlinear partial Differential Deep Reinforced Satellite Communication”, Elsevier- E-Prime- Advances in Electrical Engineering, Electronics and Energy,Peer-reviewed journal, ISSN:2772-6711, DOI- https://doi.org/10.1016/j.prime.2024.100619,29th may, 2024.
42. Varadala Sridhar,Dr.S.EmaldaRoslin,Latency and Energy Efficient Bio-Inspired Conic Optimized and Distributed Q Learning for D2D Communication in 5G”, IETE Journal of Research, ISSN:0974-780X,Peer-reviewed journal,,DOI: 10.1080/03772063.2021.1906768 , 2021, Page No: 1-13, Taylor and Francis
43. V.Sridhar, K.V. Ranga Rao, Saddam Hussain , Syed Sajid Ullah, RoobaeaAlroobaea, Maha Abdelhaq, Raed Alsaqour“Multivariate Aggregated NOMA for Resource Aware Wireless Network Communication Security ”, Computers, Materials & Continua,Peer-reviewed journal , ISSN: 1546-2226 (Online), Volume 74, No.1, 2023, Page No: 1694-1708, https://doi.org/10.32604/cmc.2023.028129,TechSciencePress
44. Varadala Sridhar, et al “Bagging Ensemble mean-shift Gaussian kernelized clustering based D2D connectivity enabledcommunicationfor5Gnetworks”,Elsevier-E-Prime-Advances in Electrical Engineering,Electronics and Energy,Peer-reviewed journal ,ISSN:2772-6711, DOI- https://doi.org/10.1016/j.prime.2023.100400,20 Dec, 2023.
45. Varadala Sridhar, Dr.S.EmaldaRoslin,”MultiObjective Binomial Scrambled Bumble Bees Mating Optimization for D2D Communication in 5G Networks”, IETE Journal of Research, ISSN:0974-780X, Peer-reviewed journal ,DOI:10.1080/03772063.2023.2264248 ,2023, Page No: 1-10, Taylor and Francis.
46. Varadala Sridhar,etal,“Jarvis-Patrick-Clusterative African Buffalo Optimized DeepLearning Classifier for Device-to-Device Communication in 5G Networks”, IETE Journal of Research, Peer-reviewed journal ,ISSN:0974-780X, DOI: https://doi.org/10.1080/03772063.2023.2273946 ,Nov 2023, Page No: 1-10,Taylor and Francis
47. 47.V.Sridhar,K.V.RangaRao,V.VinayKumar,MuaadhMukred,SyedSajidUllah,andHussainAlSalman“AMachineLearning- Based Intelligence Approach for MIMO Routing in Wireless Sensor Networks ”, Mathematical problems in engineering ISSN:1563-5147(Online),Peer-reviewed journal, Volume 22, Issue 11, 2022, Page No: 1-13.https://doi.org/10.1155/2022/6391678

48. VaradalaSridhar, Dr.S.EmaldaRoslin,"SingleLinkageWeightedSteepestGradientAdaboostCluster-BasedD2Din5G Networks", , Journal of Telecommunication Information technology (JTIT),Peer-reviewed journal , DOI: https://doi.org/10.26636/jtit.2023.167222, March (2023)