Dayananda Sagar University

School of Engineering

Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara Dt., Bengaluru – 562 112

Department of Computer Science & Technology

Project Phase -I Report

By

Kharosekar Varadraj Abhay - ENG21CT0014
Vedant Ransingh - ENG21CT0047

Under the supervision of

Prof. Nivetha NRP
Assistant Professor
Department of Computer Science and Technology

Dayananda Sagar University

School of Engineering

Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara Dt., Bengaluru – 562 112

Department of Computer Science & Technology

CERTIFICATE

This is to certify that the work titled **" Advanced Stroke Stratification and Prevention "** is carried out by **Kharosekar Varadraj Abhay (ENG21CT0014), Vedant Ransingh (ENG21CT0047)** Bonafide students of Bachelor of Technology in Computer Science and Technology at the School of Engineering, Dayananda Sagar University, Bangalore in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Technology, during the year **2024-2025**.

Prof. Nivetha NRP
Assistant Professor,
Dept. of CST,
School of Engineering,
Dayananda Sagar University.

Dr. M Shahina Parveen
Professor,
Chairperson CST,
School of Engineering,
Dayananda Sagar University.

Dayananda Sagar University

School of Engineering

Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara Dt., Bengaluru – 562 112

Department of Computer Science & Technology

DECLARATION

We, **Kharosekar Varadraj Abhay (ENG21CT0014), Vedant Ransingh (ENG21CT0047),** are students of the seventh semester B.Tech in Computer Science and Technology, at School of Engineering, Dayananda Sagar University, hereby declare that the project phase - I titled "**Advanced Stroke Stratification and Prevention**" has been carried out by us and submitted in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Technology during the academic year 2023-2024.

Student Signature

Name1:Kharosekar Varadraj Abhay

USN : ENG21CT0020

Name2 : Vedant Ransingh

USN : ENG21CT0026

Place : Bangalore

Date :

ABSTRACT

Stroke is a critical condition that affects the lives of millions of people around the world every year challenging the national health systems. Past risk assessment techniques mainly focused on clinical variables which do not give adequate accuracy to risk assessment of stroke. This paper aims at determining the impact of incorporating three types of data including ECG, 2D ECHO and clinical metrics into a machine learning model. Therefore, the integration of such varied data types within the proposed approach should result not only in improved risk profiling and assessment. Hoping to use the value of the approaches to machine learning this work aims at providing the possibility to prevent the strokes, to lessen the impact of the disease, as well as to decrease its load on the persons and healthcare systems.

Keywords:

Artificial Intelligence in Healthcare, Early Disease Detection, AI for Diagnosis, Machine Learning in Medicine, Deep Learning for Medical Images, Predicting Disease Risk, AI Tools for Diagnosis, Language Processing in Healthcare, Support Systems for Doctors, Ethical Issues in AI.

TABLE OF CONTENTS

	Page No.
Certificate	**i**
Declaration	**ii**
Abstract	**iii**
List of Figures	**iv**
List of Tables	**v**
List of Abbreviations	**Vi**
1. Introduction	**1**
2. Literature Survey	**2-3**
3. Project Requirement Specification	**4**
4. Problem Definition	**5-6**
5. System Architecture	**7-9**
6. Implementation	**10-11**
7. Conclusion	**12**
References	**13**

INTRODUCTION

The Stroke is also among the main causes of disability as well as death, it occurs recurrently and costs millions of people's lives annually, and is continuing to strain the health care facilities. There remain problems with the accuracy of risk characterization, the primary tools of which have been clinical risk factors such as age, blood pressure, cholesterol, etc. Although these methods offer basic approaches, the disadvantage is that they are general and thereby the prediction can be far from the optimum for individuals who have several or multiple risks factors. The progress made in the last few years on development of advanced machine learning (ML) allow for the development of models that can combine multiple types of data sources to improve the prediction of stroke risk. A much more sophisticated method is to incorporate clinical information in conjunction with physiological measurements like the ECG and ECHO data into risk models. Imaging and monitoring data which includes ECG data is also powerful since abnormal heart rate and rhythm, including atrial fibrillation which is a key source of stroke can be detected. Likewise, ECHO imaging gives structural and functional measures, for example left atrial size or ejection fraction, which are importance for use in embolic risk. By integrating the features of ML to these disparate flows of data, it is possible to provide an integrated analysis of these sources for detecting complexes of relationships between different risks. In detail, this study seeks to create a model that will use the ECG, ECHO together with clinical data in order to give individual strokes risk results. The proposed approach aims at improving prognostic performance but not merely solely for the improvement in model accuracy and for an early and efficient preventive action to decrease the global prevalence rate of stroke. Studies included in this body of research support the possibility of using an augmented ECG and ECHO data with the clinical variables to establish stronger and better individualized risk prediction models that serve as a basis for this study.

LITERATURE SURVEY

The The literature on stroke risk prediction highlights the growing importance of machine learning (ML) and multimodal data integration in improving predictive accuracy. Studies such as Prediction of Stroke Risk with LSTM Networks and Patient Health Records (2020) by Ms. Madhavi Kshatri, Dr. Wilson Lukose, and Dr. Jalaluddin Khan demonstrated the effectiveness of Long Short-Term Memory (LSTM) networks in analysing sequential patient health records. However, the study was limited to patient health records and did not include imaging data, which could further enhance prediction accuracy. Similarly, An Optimized Machine Learning-Based Stroke Prediction: Enhancing Precision Medicine and Public Health (2021) by Parul Khatri, Archana Sharma, and Payal optimized machine learning techniques like support vector machines and random forests to improve prediction accuracy, yet it lacked real-time monitoring capabilities and the integration of multimodal data sources such as ECG or ECHO imaging.

Deep learning models have also been applied to specific data streams, as seen in Deep Learning Applications in 12-Lead Electrocardiogram (2023) by Masamitsu Nakayama, Ryuichiro Yagi, and Shinichi Goto, which extracted cardiovascular insights from 12-lead ECG data. While valuable, this study did not combine ECG data with other sources, limiting its applicability for comprehensive stroke prediction. The potential of multimodal data integration is further emphasized in Multimodal Data Integration in Stroke Prediction Using AI Techniques (2022) by Samantha Clark and Alex Johnson, which demonstrated how combining clinical records, imaging data, and genetic information using AI can improve prediction. However, the challenges of preprocessing and harmonizing heterogeneous data were also highlighted.

The role of echocardiography in stroke risk assessment was explored in Role of Echocardiography in Stroke Risk Assessment (2021) by Dr. Sarah Williams and Dr. Daniel Lee, focusing on left atrial size and ejection fraction as critical parameters. Despite its contributions, the study did not employ advanced machine learning methods. Finally, Machine Learning for Stroke Prediction: Challenges and Opportunities (2020) by Andrew Roberts and Maria Gomez reviewed various machine learning models, such as logistic regression and neural networks, discussing their strengths and limitations, particularly their dependency on input data quality and diversity.

Overall, the literature underscores the necessity of integrating multimodal data to enhance stroke risk stratification. Existing studies largely focus on single data types or traditional machine learning models, leaving a gap in comprehensive approaches. This survey identifies the need to combine clinical, ECG, and ECHO data using advanced machine learning techniques to develop a robust and personalized stroke risk assessment framework.

Paper Name	Year	Author(s)
Prediction of Stroke Risk with LSTM Networks and Patient Health Records	2020	Ms. Madhavi Kshatri, Dr. Wilson Lukose, Dr. Jalaluddin Khan
An Optimized Machine Learning-Based Stroke Prediction: Enhancing Precision Medicine and Public Health	2021	Parul Khatri, Archana Sharma, Payal
Deep Learning Applications in 12-Lead Electrocardiogram	2023	Masamitsu Nakayama, Ryuichiro Yagi, Shinichi Goto

PROJECT REQUIREMENT SPECIFICATION

Functional Requirements :

1. Data Input:

- Support for uploading 2D ECHO video files (in .gif format).
- Support for uploading ECG data in .csv format.
- Option to input patient clinical history through a dynamic questionnaire.

2. Data Preprocessing:

- Extract meaningful features from ECHO images/videos.
- Clean, normalize, and pre process ECG data.
- Process clinical history and convert it into numeric features.

3.Integration:

- Combine ECHO, ECG, and clinical data into a unified feature set for model training and prediction.

4 .Model Training:

- Train a machine learning model using integrated data for stroke risk prediction.

5.Output:

- Provide a detailed stroke risk report, including predictive scores and recommendations for early intervention.

6.User Interface:

- User-friendly interface for data upload and visualization of risk reports.

Technical Specifications

- Programming Language: Python
- Libraries/Frameworks:
- Machine Learning: TensorFlow, scikit-learn
- Image Processing: OpenCV, PIL
- Data Analysis: Pandas, NumPy
- Data Storage:
- CSV files for clinical and ECG data
- .npy files for processed features
- Tools: Jupyter Notebook, Visual Studio Code

PROBLEM DEFINITION

Stroke continues to be one of the most significant public health challenges worldwide, ranking among the leading causes of death and long-term disability. Its impact is profound, affecting millions of individuals annually and placing an immense burden on families, caregivers, and healthcare systems. Despite decades of research and advancements in medical science, predicting stroke risk with high accuracy remains a formidable challenge. Traditional methods of stroke risk stratification predominantly rely on clinical data, such as patient demographics, blood pressure, cholesterol levels, and family medical history. While these methods have been instrumental in establishing baseline risk assessments, they often fall short in capturing the intricate and multifactorial nature of stroke risk. The reliance on limited clinical parameters not only restricts the predictive power of these models but also fails to account for the interplay of structural, functional, and physiological factors that can significantly influence an individual's risk profile.

This gap in predictive accuracy highlights the urgent need for more comprehensive approaches that leverage advancements in technology and data science. In this context, the integration of multimodal data sources has emerged as a promising avenue to enhance stroke risk assessment. This research aims to address these limitations by combining three critical data streams: electrocardiogram (ECG) readings, 2D echocardiography (ECHO) imaging, and traditional clinical data. ECG data provides dynamic insights into the electrical activity of the heart, enabling the detection of arrhythmias such as atrial fibrillation, which is a major risk factor for stroke. Similarly, ECHO imaging offers detailed information on the structural and functional aspects of the heart, such as left atrial size, valve function, and ejection fraction, which are crucial for identifying embolic risks and other cardiac abnormalities linked to stroke. Clinical data, including medical history, lifestyle factors, and demographic details, serve as a foundational layer that complements the physiological and imaging data.

By integrating these diverse and complementary data sources, this research seeks to develop a novel, machine learning-driven framework for stroke risk prediction. Machine learning techniques, known for their ability to analyse large, complex datasets and identify patterns that may not be immediately apparent, are well-suited for this task. These models can analyse the interplay between multiple risk factors, uncover hidden relationships, and generate a personalized risk profile for each patient. This personalized approach has the potential to significantly improve the accuracy of stroke risk stratification compared to traditional methods. Moreover, it enables the identification of high-risk individuals who may benefit from targeted preventive interventions, such as lifestyle modifications, medication adjustments, or advanced monitoring.

The overarching goal of this research is to bridge the existing gaps in stroke risk prediction by developing a data-driven, multimodal model that goes beyond the limitations of traditional clinical assessments. By providing more accurate and personalized risk assessments, this study aspires to facilitate earlier detection of stroke risk and empower healthcare providers to implement timely and

effective preventive strategies. In doing so, it seeks to reduce the incidence of stroke, alleviate its devastating consequences, and contribute to the broader effort of enhancing global stroke prevention and patient outcomes. Ultimately, this integrated approach represents a significant step forward in the pursuit of precision medicine, where interventions are tailored to the unique characteristics of each individual, thereby maximizing their impact and improving overall healthcare delivery.

SYSTEM ARCHITECTURE

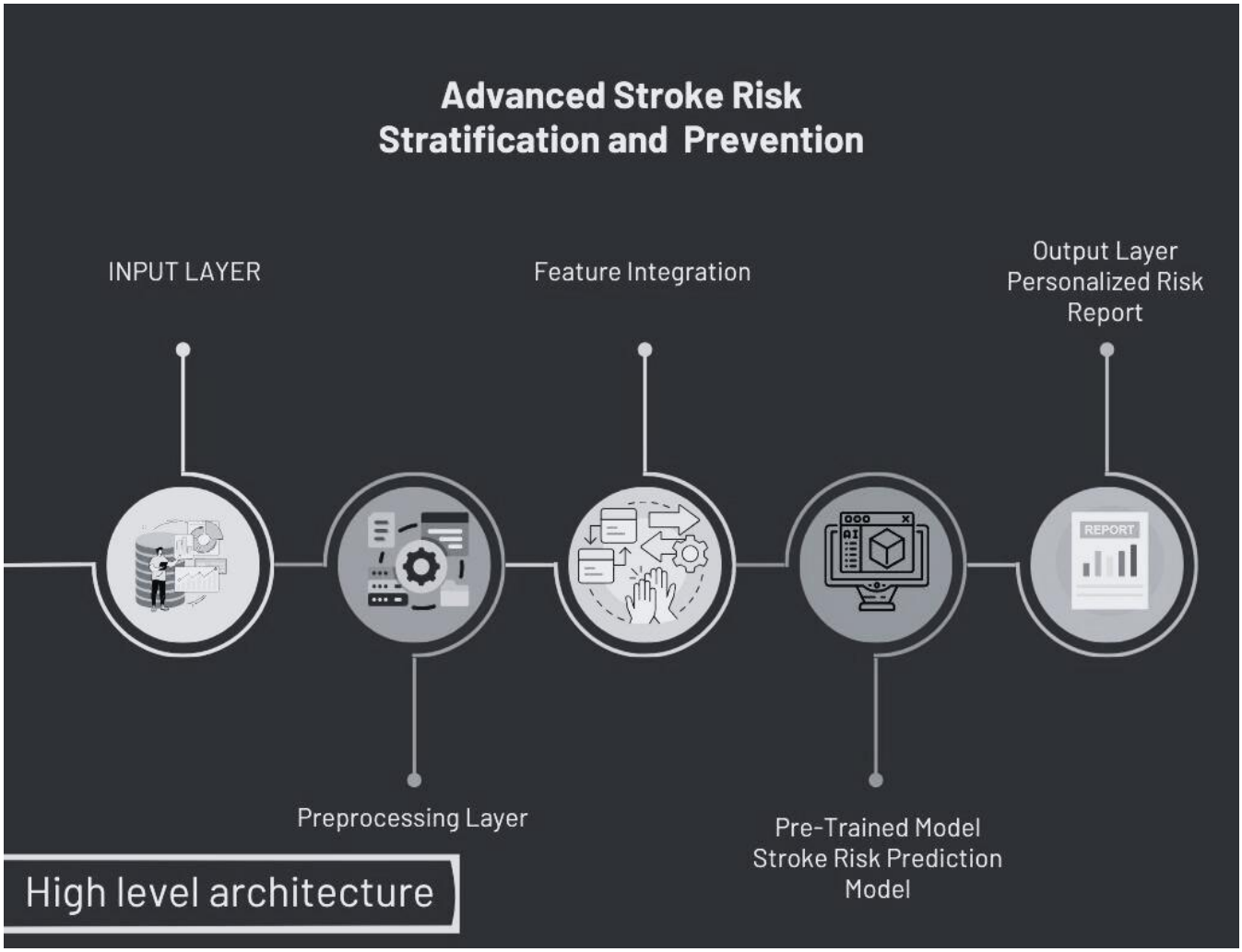

Fig 1: System Architecture

1. Input Layer:

This is the foundational layer where the data is collected. It involves gathering information from multiple sources, including:

- Clinical Data: Patient demographics (age, gender), medical history, lifestyle factors (smoking, exercise, etc.), and pre-existing conditions like hypertension and diabetes.
- ECG Data: Electrocardiogram recordings provide insights into cardiac electrical activity to detect arrhythmias, such as atrial fibrillation, which are key stroke risk factors.
- ECHO Data: 2D echocardiography images deliver structural and functional details about the heart, such as ejection fraction, left atrial size, and valve integrity.

2.Preprocessing Layer:

This stage ensures the data is clean, normalized, and ready for analysis. Key processes include:

- Data Cleaning: Handling missing, inconsistent, or erroneous data.
- Normalization: Scaling the data to a consistent range for better model performance.
- Feature Extraction:Identifying and extracting relevant features from ECG waveforms and ECHO.
- Data Augmentation (for ECHO/ECG): Enhancing the dataset by generating synthetic samples or augmenting images to improve model robustness.
- Handling Imbalances**:** Addressing class imbalance in stroke vs. non-stroke patients using techniques like oversampling or SMOTE.

3.Feature Integration:

In this step, the processed data from different sources is combined into a unified format for machine learning analysis. This involves:

- Multimodal Feature Engineering: Integrating features from clinical, ECG, and ECHO datasets.
- Dimensionality Reduction: Reducing data complexity using techniques like PCA (Principal Component Analysis) or t-SNE to focus on the most relevant features.
- Correlation Analysis: Identifying relationships among various features to ensure meaningful integration.

4.Pre-Trained Model (Stroke Risk Prediction Model):

This stage involves using advanced machine learning and deep learning algorithms to analyse the integrated data and predict stroke risk:

- Algorithm Selection: Models like logistic regression, random forests, gradient boosting (e.g., XGBoost), and neural networks are considered.
- Deep Learning Models: LSTMs for sequential ECG data and CNNs for analysing ECHO images are employed for feature extraction.
- Training & Testing: Models are trained on a labelled dataset, validated, and tested for performance.
- Performance Metrics: Evaluation using metrics like accuracy, sensitivity, specificity, F1-score, and ROC-AUC.

5. Output Layer (Personalized Risk Report):

The final stage generates actionable outputs based on the model's predictions:

- Risk Assessment Report: Provides a detailed, patient-specific report with risk scores, highlighting key contributing factors.
- Visualization: Graphical summaries (e.g., charts, heatmaps) to explain risk factors to clinicians and patients.
- Intervention Recommendations: Suggestions for lifestyle changes, medical monitoring, or preventive treatments tailored to individual risk profiles.

IMPLEMENTATION

1.Data Collection and Integration

Clinical Data:

- Collect patient demographics, medical history, and lifestyle factors from electronic health records (EHRs).
- Example: Age, gender, blood pressure, cholesterol levels, history of diabetes, or atrial fibrillation.

ECG Data:

- Use devices to record 12-lead ECG signals for patients. Data is collected in time-series format.
- Example Tool: Use libraries like BioSPPy for preprocessing ECG data.

ECHO Data:

- Obtain 2D echocardiography images or videos. Use DICOM format files for clinical-grade imaging.
- Example Tool: Use pydicom to process echocardiographic imaging data.

2.Preprocessing

Clinical Data Preprocessing:

- Handle missing values using imputation techniques such as mean/mode substitution or KNN.
- Normalize numerical features to a uniform scale using sklearn. preprocessing. StandardScaler.

ECG Data Preprocessing:

- Filter noise from ECG signals using Butterworth filters or wavelet transforms.
- Segment the signal into cardiac cycles (PQRST patterns) using libraries like HeartPy.

ECHO Data Preprocessing:

- Resize and normalize images to ensure consistency.
- Convert videos to image frames if required.
- Augment data with techniques like rotation, flipping, or adding slight noise using libraries like TensorFlow or PyTorch.

3. Feature Engineering and Integration:

ECG Feature Extraction:

- Extract features such as heart rate, QRS duration, PR interval, and ST elevation.
- Use signal processing tools like NeuroKit2 for automated feature extraction.

ECHO Feature Extraction:

- Use convolutional neural networks (CNNs) like VGG16 or ResNet50 to extract structural and functional features (e.g., ejection fraction, left atrial size).

Clinical Data Features:

- Convert categorical variables into numerical format using one-hot encoding or label encoding.
- Engineer composite features such as BMI from weight and height.

CONCLUSION

Stroke remains one of the most significant public health challenges, with its prevention and management relying heavily on accurate risk stratification. Traditional approaches, often limited to clinical data, fail to fully capture the complexity of individual risk factors, leading to suboptimal predictions. This study demonstrates the potential of integrating multimodal data—clinical parameters, ECG readings, and ECHO imaging—through machine learning techniques to enhance predictive accuracy.

The proposed model not only improves risk assessment but also provides personalized insights that empower healthcare professionals to intervene earlier and more effectively. By leveraging advanced machine learning algorithms, the system identifies complex interactions among diverse data sources, offering a comprehensive and scalable solution for stroke prevention.

This research contributes to the growing body of evidence supporting the integration of AI in healthcare. It highlights how innovative, data-driven approaches can reduce the global burden of stroke, improve patient outcomes, and pave the way for advancements in predictive medicine. Future work can explore real-time monitoring, larger datasets, and other multimodal combinations to further refine and expand the scope of such predictive models.

REFERENCES

1. Kshatri, M., Lukose, W., & Khan, J. (2020). Prediction of Stroke Risk with LSTM Networks and Patient Health Records.

2.Khatri, P., Sharma, A., & Payal. (2021). An Optimized Machine Learning-Based Stroke Prediction: Enhancing Precision Medicine and Public Health.

3.Nakayama, M., Yagi, R., & Goto, S. (2023). Deep Learning Applications in 12-Lead