2025 Global Conference in Emerging Technology (GINOTECH)
Pune, India. May 09 - 11, 2025

Data Driven Deep Learning Model for Effective Stroke Prediction

Monisha A
Faculty of Electronics and Communication Engineering
Adhi College of Engineering and Technology,
Kanchipuram, India
monisha.a.ece@adhi.edu.in

Mariaamutha R
Department of Electronics and Communication Engineering
Bannari Amman Institute of Technology,
Sathyamangalam, Erode
mariaamuthar@bitsathy.ac.in

Kannaki
Department of Electronics and Communication Engineering
Dhanalakshmi Srinivasan College of Engineering,
Coimbatore, 641105, India
Kannaki@dsce.ac.in

P Vinayagam
Department of Computer Science and Engineering
Saveetha Engineering College,
Chennai, India
vinayagamap@gmail.com

M. Ramaiah
Department of Computer Science and Engineering
Saveetha Engineering College,
Chennai, India
ramaiahmp24@gmail.com

M. Saranya
Department of Computer Science and Engineering
Sri Kaliswari College,
Sivakasi 626130, India
msaranya.skc@gmail.com

2025 Global Conference in Emerging Technology (GINOTECH) | 979-8-3315-0775-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/GINOTECH63460.2025.11076988

***Abstract*— The substantial impact of stroke on society has driven continuous efforts to enhance its diagnosis and management. The growing integration of technology with medical diagnostics empowers healthcare providers to optimize patient care by systematically mining and archiving medical records. This study proposes a Data-Driven Deep Learning (DDDL) model for effective stroke prediction, employing an Attention Residual Network to boost predictive accuracy and support timely clinical interventions. The input data is subjected to a preprocessing stage, where irrelevant and redundant information is cleaned and removed. The Exploratory Data Analysis (EDA) visualization phase incorporates data training and testing, offering valuable insights for further analysis. Classification is executed using the Attention Residual Network, ensuring accurate and dependable predictions. The predicted output aids in the early detection of stroke, enabling prompt clinical decision-making. Python software is utilized for simulation with a accuracy of 90%, validating the model's performance and demonstrating its potential to enhance stroke prediction and improve patient outcomes.**

Keywords*— *Attention Residual Network, Stroke, EDA Visualization, Data-driven model.

I. INTRODUCTION

With rapid technological advancements, human life expectancy is steadily increasing. Stroke diagnosis involves a combination of clinical assessment, imaging techniques, and laboratory tests to determine the type and severity of the condition. Physicians evaluate symptoms such as sudden numbness, confusion, and difficulty speaking. Imaging techniques like CT scans and MRIs help detect brain damage and locate blockages or bleeding. Additional tests, including blood tests and carotid ultrasounds, assess risk factors like clotting disorders and artery health. Stroke, a critical health concern, occurs primarily due to the disruption of blood supply to brain nerves caused by blood clotting. Depending on the severity, strokes categorized as major or minor. Minor strokes disrupt blood flow to specific brain regions, while major strokes can be fatal. As an emergency health condition, stroke requires immediate attention. Common symptoms include difficulty in movement, confusion, impaired verbal communication, and trouble understanding. Strokes often result in long-term neurological damage or death.

Strokes are generally classified into two categories: ischemic embolic and hemorrhagic. Ischemic embolic strokes occur when a blood clot forms in the heart, subsequently narrowing brain arteries. Haemorrhagic strokes also involve blood leakage from damaged arteries in the brain. For elderly individuals, strokes are particularly dangerous, often leading to life-threatening complications. Similar to rate the heart attacks damage the heart, strokes inflict significant damage on the brain. Once diagnosed, continuous health monitoring becomes essential for stroke patients.

A stroke often begins with a Transient Ischemic Attack (TIA), commonly known as a ministroke. TIAs signal that a full stroke could occur within days, necessitating prompt medical intervention. According to the World Health Organization (WHO), strokes contribute significantly to global mortality rates. However, early detection and diagnosis of strokes prevent death and severe brain damage in up to 85% of cases. Senior citizens require extra care, as strokes are more lethal for the aging population. Continuous observation and monitoring are critical in managing this disease effectively. The rising incidence of stroke cases is attributed to factors such as stress, physical inactivity, substance abuse, and poor dietary habits.

Subsequently, the classification process plays a pivotal role in stroke prediction. Stroke prediction classification has historically relied on various traditional methods. One widely used technique is the Support Vector Machine (SVM), which establishes a relationship between input features and binary outcomes. While SVM is straightforward to interpret and implement, it often struggles with the complex, non-linear patterns characteristic of stroke data, thereby limiting its overall predictive effectiveness. Another commonly employed approach is the Decision Tree algorithm, which classifies data by creating splits based on feature values. Decision trees are intuitive, easy to visualize, and capable of handling both categorical and numerical data efficiently. However, they are prone to overfitting, particularly when dealing with noisy datasets, which diminishes their accuracy in real-world applications. A third notable method is the Artificial Neural Network (ANN), which aims to find the optimal hyperplane to separate different classes. ANNs excel at managing high-dimensional data and generally exhibit robustness against overfitting. However, ANNs demand

careful parameter tuning and are computationally intensive, especially when processing large datasets K-Nearest Neighbors (KNN) algorithm is simplicity and effectiveness in handling non-linear data in distribution. However, it is high computational cost, especially with large datasets, as it requires storing and searching for each prediction. To address these issues, the Attention Residual Network presents a significant advancement in stroke prediction classification. Attention Residual Network leverages hierarchical classification, allowing it to capture intricate patterns in data that traditional methods can't takes place. The main contributions are:

- To ensure high-quality data input, data pre-processing is used for accurate stroke prediction through efficient cleaning and removal of irrelevant or redundant information.
- To incorporate both data training and testing processes for robust analysis and to provide insightful visual representations of data patterns through EDA visualization.
- To improve the accuracy of skin cancer classification by employing an Attention Residual Network and to leverage hierarchical classification for capturing complex patterns in data.

II. Description of Proposed System

This work presents a novel approach for effective stroke prediction that integrates DDDL techniques with an Attention Residual Network for accurate classification.

Fig. 1 illustrates a DDDL model for effective stroke prediction. The process begins with input data, which undergoes data preprocessing to ensure high-quality data by cleaning and removing irrelevant or redundant information. This step enhances the accuracy of the predictive model. Following preprocessing, the data is passed to the EDA visualization stage, where the data is split into training and testing sets. This stage provides insightful visual representations of data patterns, enabling robust analysis and a deeper understanding of underlying trends. The training and testing datasets are then used for classification using an Attention Residual Network. This network captures complex hierarchical patterns in the data, significantly improving prediction accuracy. Finally, the model generates the predicted output, supporting early detection and timely clinical interventions for stroke.

III. System Modelling

A. Data Pre-processing

The process of cleaning and altering data, whether it be semi-structured, unstructured, or structured, is known as data preprocessing. This information is obtained from a variety of sources, such as application data, streaming data, and historical documents. Preprocessing involves applying various transformations to the data before feeding it into an algorithm. Typically, data collected from diverse sources is in a raw format, making it unsuitable for immediate analysis. Therefore, preprocessing ensures the data is cleaned, organized, and properly formatted, enhancing the performance and accuracy of analytical models.

B. EDA Visualization

In stroke prediction, EDA techniques play a crucial role in enabling data scientists to extract key insights from the data. EDA is essential for assessing and improving data quality, which directly impacts the accuracy of stroke prediction models.

Data scientists' expertise and domain knowledge are critical in performing effective EDA, as they help in identifying relevant patterns and relationships within the dataset. EDA not only assists in detecting data quality issues which also provides a comprehensive understanding of the dataset. This understanding is vital for making informed decisions regarding algorithmic and architectural choices for deep learning models used in stroke prediction. Additionally, EDA guides the data preprocessing steps by highlighting

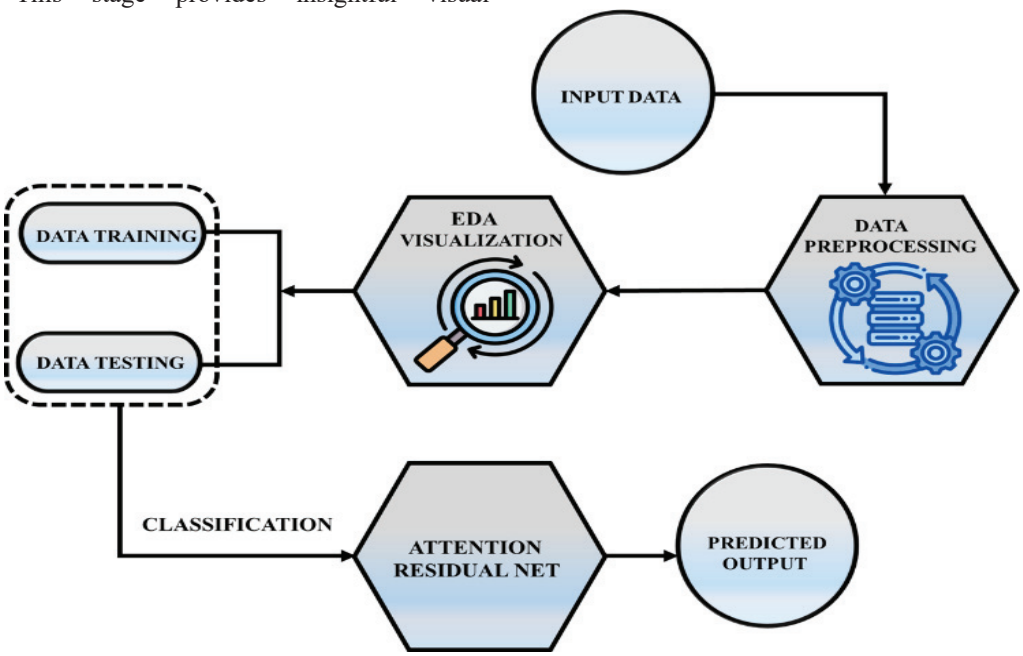

Fig. 1. Block Diagram of Proposed System

necessary transformations and cleaning processes, ensuring that the data is optimized for analysis.

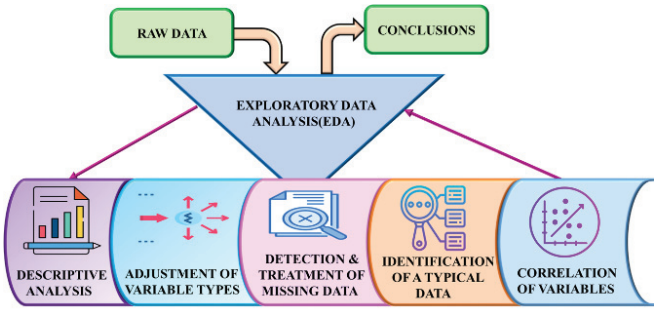

Fig. 2. Process of EDA

Fig. 2 illustrates the EDA process, which transforms raw data into meaningful conclusions. EDA involves several key steps: descriptive analysis for summarizing data characteristics, adjustment of variable types to ensure consistency, detection and treatment of missing data to enhance data quality, identification of atypical data for spotting anomalies and correlation of variables to uncover relationships.

C. Attention Residual Network

The data classification model is designed using an Attention Residual Network. Attention Residual Network, consisting of multiple stacked residual blocks, allows for easier enhancement of the network model structure. This research specifically focuses on improving the Attention Residual Network architecture. The original network structure of Attention Residual Network is illustrated in Fig. 3.

The Attention Residual Network lies in its residual structure, which significantly boosts network performance. This is achieved by enriching feature data through the combination of shallow and deep network feature information. By integrating these features, the residual structure overcomes issues like vanishing gradients and degradation in deep networks, ultimately enhancing accuracy and efficiency. The improvements introduced in this study leverage attention mechanisms alongside Attention Residual Network architecture to capture complex data patterns more effectively, resulting in a robust and precise classification model.

Attention Residual Network, aims to find the optimal weight values for feature extraction, represented by equation (1).

$$\theta^* = arg \max_{\theta} L(f(x,\theta), y) \tag{1}$$

Here, L is the loss function, and θ includes all parameters within the CNN's parameter space. The optimal parameter set $\theta *$ corresponds to the better weight values.

$$\theta_{iter} = \theta_{iter-1} - \eta g_{iter} \tag{2}$$

The function $f(x,\theta)$ represents the model, where x is the input data and y is the actual label. After calculating the weight gradient, the weight values are updated using equation (2).

$$\upsilon_{iter} = \beta_1 \upsilon_{iter-1} + (1-\beta_1) g_{iter} \tag{3}$$

$$\omega_{iter} = \beta_2 \omega_{iter-1} + (1-\beta_2) g_{iter} \cdot g_{iter} \tag{4}$$

$$\upsilon^{\wedge}_{iter} = \frac{\upsilon_{iter}}{1-\beta_1^{iter}} \tag{5}$$

The variable η indicates the learning rate, while $iter$ represents the number of iterations. The weight update process per iteration is described in equations (3) to (5). To address issues like slow convergence and extended training times, Attention Residual Network methods to enhance training efficiency, adjusts adaptive learning rates, and retains gradient information. Thus, Attention Residual Network classifier enhances feature extraction by focusing on critical patterns while minimizing irrelevant data and also improves prediction accuracy and reduces computation time with reliable classification performance.

IV. Result and Discussion

This research presents an Attention Residual Network in the classification stage to enhance the accuracy and efficiency of stroke prediction. The simulation results are implemented using Python software.

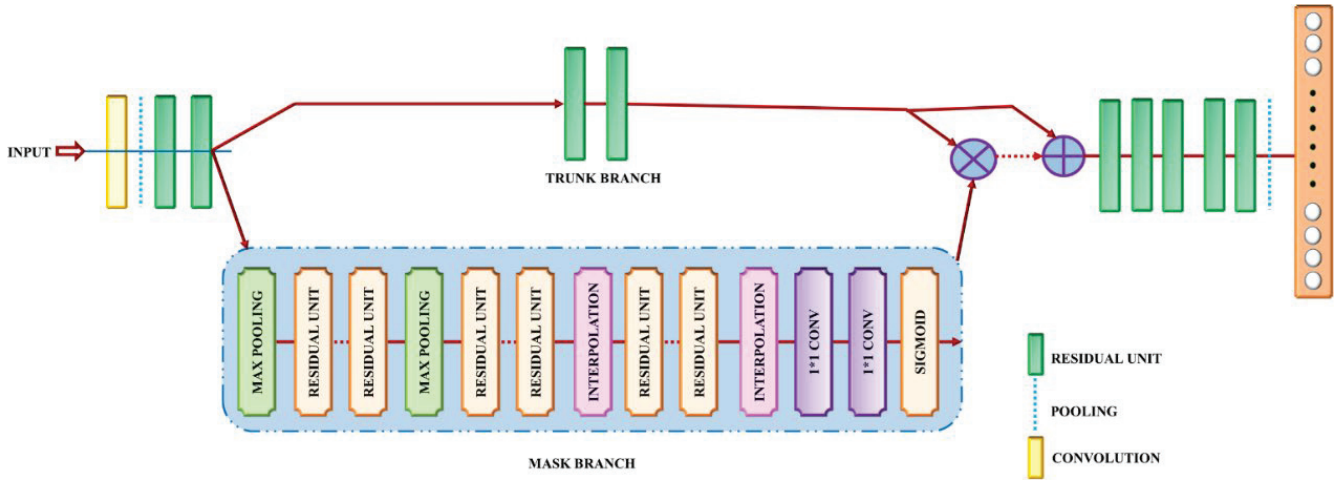

Fig. 3. Architecture of Attention Residual Network

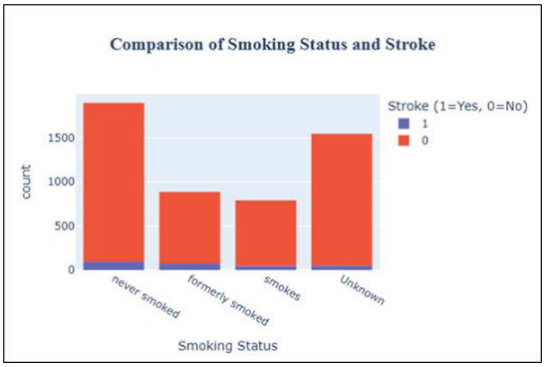

Fig. 4. Comparison of smoking status and stroke

Fig. 4 illustrates the relationship between smoking status and stroke occurrences. Most individuals who never smoked or had unknown smoking status did not experience a stroke, whereas stroke cases were comparatively fewer among current and former smokers.

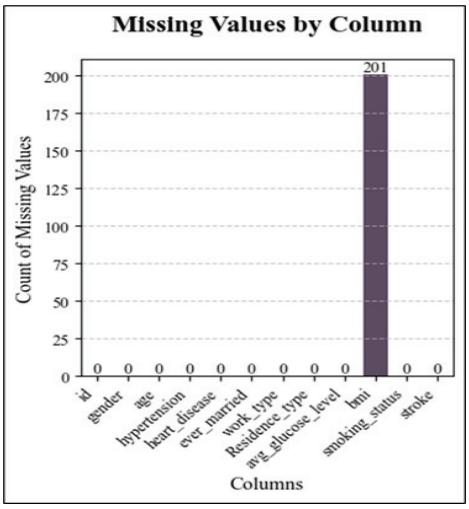

Fig. 5. Missing Values by Column

Fig. 5 shows missing values in the dataset. Only the smoking status column contains missing data, with 201 entries incomplete, while all other columns are fully populated and have no missing values.

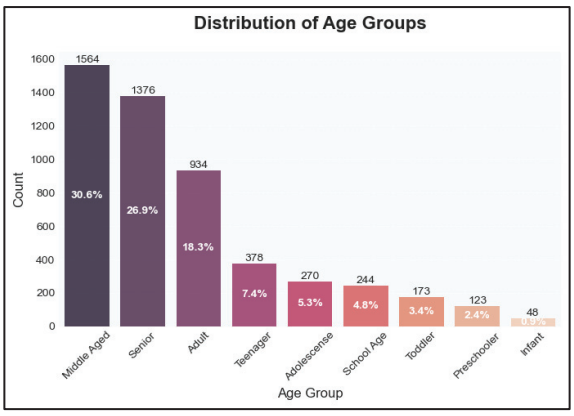

Fig. 6. Distribution of Age Groups

Fig. 6 presents the age group distribution, highlighting that middle-aged individuals (30.6%) and seniors (26.9%) make up the largest portions of the population. In contrast, infants (0.9%) and preschoolers (2.4%) represent the smallest demographic segments within the dataset.

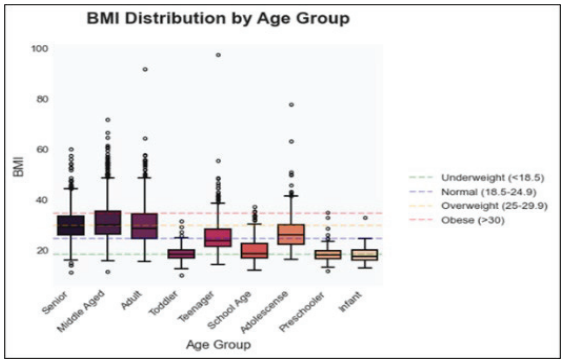

Fig. 7. BMI Distribution by Age groups

Fig. 7 illustrates BMI distribution across various age groups. Seniors, middle-aged, and adults exhibit higher BMI ranges with notable outliers, while younger groups like infants and preschoolers show lower BMI variability.

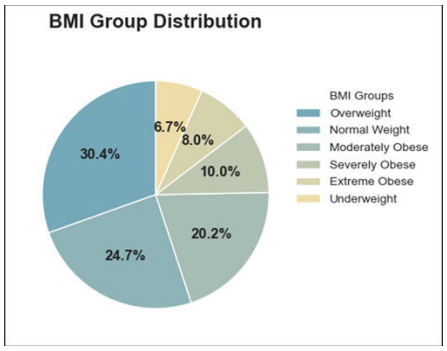

Fig. 8. BMI Group Distribution

Fig. 8 shows BMI group distribution, with 30.4% overweight, 24.7% of normal weight and 20.2% moderately obese. Severely obese, extreme obese, and underweight categories account for 10%, 8% and 6.7%, respectively.

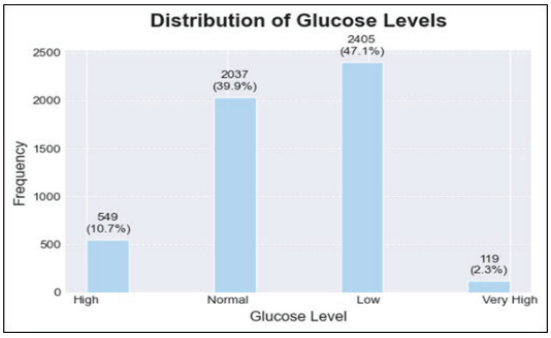

Fig. 9. Distribution of Glucose levels

Fig. 9 displays glucose level distribution: 47.1% have low levels, 39.9% normal, 10.7% high, and 2.3% very high. Low glucose levels are the most frequent among the groups analysed.

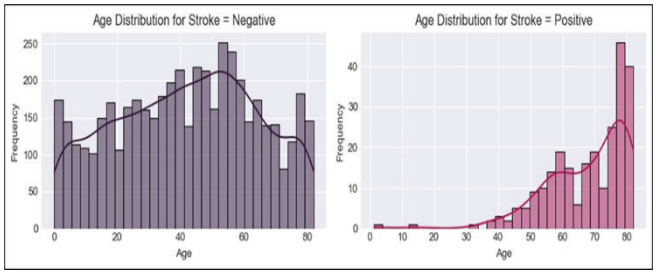

Fig. 10. Age Distribution Comparison

Fig. 10 illustrates the age distribution for stroke outcomes. Stroke-negative cases are relatively consistent across all age groups, whereas stroke-positive cases increase significantly after age 60, with the highest incidence observed between ages 70 and 80.

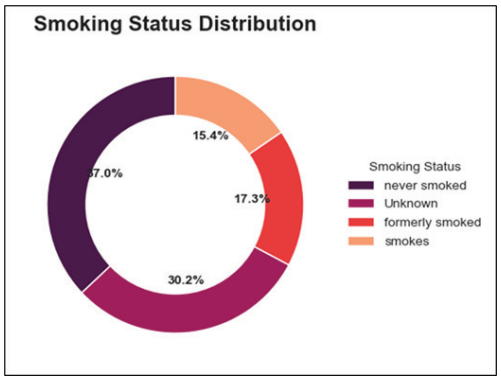

Fig. 11. Smoking status distribution

Fig. 11 displays the distribution of smoking status: 57% of individuals never smoked, 30.2% have unknown status, 17.3% are former smokers, and 15.4% currently smoke. This highlights the dominance of non-smokers in the dataset and a significant portion of missing or unreported smoking data.

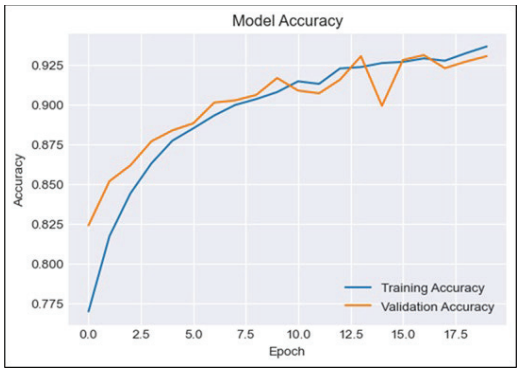

Fig. 12. Model Accuracy

Fig. 12 illustrates model accuracy over 20 epochs. Both training and validation accuracy progressively improve, reaching a peak of approximately 0.93, indicating strong model performance with minimal signs of overfitting.

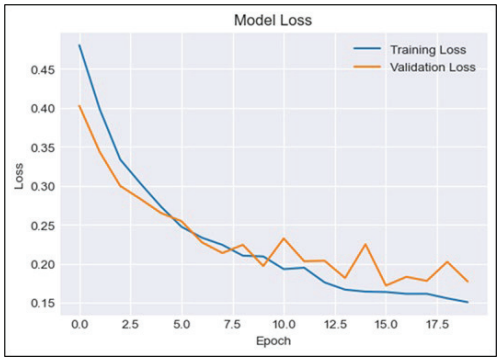

Fig. 13. Model loss

Fig. 13 shows model loss over 20 epochs. Both training and validation loss steadily decrease, stabilizing near 0.15. This indicates effective learning with minimal overfitting, although slight fluctuations in validation loss suggest occasional variance in model performance across epochs.

TABLE I. CLASSIFICATION REPORT

Classification Report				
	Precision	**Recall**	**F1-score**	**support**
No stroke	0.92	0.94	0.93	1457
Stroke	0.93	0.92	0.93	1460
Accuracy			0.93	2917
Macro Avg	0.93	0.93	0.93	2917
Weighted Avg	0.93	0.93	0.93	2917

The classification report in Table I demonstrates an overall accuracy of 93%, with balanced precision, recall, and F1-scores for both stroke and no-stroke predictions across 2,917 samples.

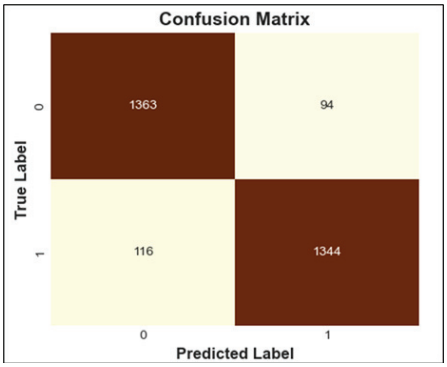

Fig. 14. Confusion Matrix

The confusion matrix in Fig. 14 shows 1,363 true negatives, 1,344 true positives, 94 false positives, and 116 false negatives. This reflects high classification accuracy with minimal misclassifications, highlighting the model's strong performance in correctly identifying both stroke and no-stroke cases.

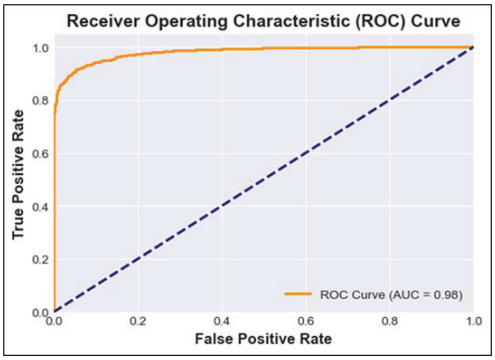

Fig. 15. ROC Curve

The ROC curve in Fig. 15 shows excellent model performance with an AUC of 0.98, indicating high true positive rates and minimal false positives, reflecting strong classification capability and predictive accuracy.

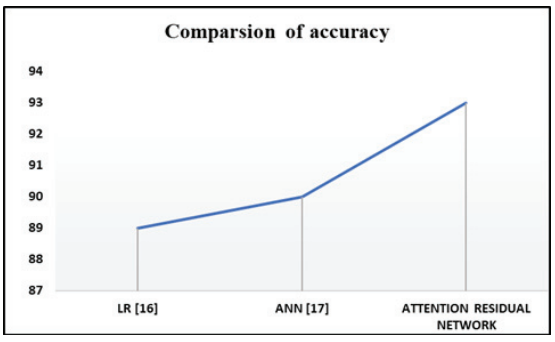

Fig. 16. Comparison of Accuracy

Fig. 16 represents the accuracy of the different method used for the classification stage such as LR [16], ANN [17] and Attention Residual Network with the accuracy of 89%, 90% and 93%.

TABLE II. COMPARISON OF PRECISION VALUE

SL.NO	Methods	Specificity
1.	RF [18]	0.856
2.	SVM [19]	0.871
3.	Deep Vision Net Classifier	0.895

Table II shows the precision values of different methods such as RF, SVM and Deep vision net classifier with values of 0.856, 0.871 and 0.895 respectively.

V. CONCLUSION

This study presents a DDDL based approach for stroke prediction, leveraging an Attention Residual Network to improve classification accuracy and support early diagnosis. The proposed model successfully streamlines stroke prediction through a systematic approach, from data preprocessing and exploratory analysis to precise classification. This comprehensive pipeline with clinical decision-making and patient management, ensuring reliable and effective healthcare outcomes is enhanced. The simulation results, achieved using Python, validate the model's effectiveness with an accuracy of 90%, demonstrating its potential in real-world medical applications. By integrating DDDL techniques, this model facilitates early detection, enabling timely interventions and reducing the risk of severe complications. The future work will focus on enhancing the model's interpretability and clinical applicability. Integrating explainable AI techniques will provide clinicians with transparent decision-making insights healthcare settings. Additionally, expanding the model's capabilities to incorporate multimodal data.

REFERENCES

[1] S. R. Sakhare, M. Kaur, F. Akter and K. Wanjale, "[Retracted] Early Stroke Prediction Methods for Prevention of Strokes," Behavioural Neurology, vol. 2022, no. 1, pp. 7725597, 2022.

[2] B. Noble, I. N. Sneddon and G. Eason, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[3] M. Trigka and E. Dritsas "Stroke risk prediction with machine learning techniques," Sensors, vol. 22, no. 13, pp. 4670, 2022.

[4] H. Wang, S. Dev, B. Veeravalli, N. Jain and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," Healthcare Analytics, vol. 2, pp. 100032, 2022.

[5] I. Hussain, M. S. Islam, M. Rahman and M .A. Hossain, "Explainable artificial intelligence model for stroke prediction using EEG signal," Sensors, vol. 22, no. 24, pp. 9859, 2022.

[6] J. Faura, A. Bustamante and J. Montaner, "Stroke-induced immunosuppression: implications for the prevention and prediction of post-stroke infections," Journal of neuroinflammation, vol. 18, pp. 1–14, 2021.

[7] S. Ghimire, K. Mridha, J. Shin, M. M. Uddin, and M. F. Mridha, "Automated stroke prediction using machine learning: an explainable and exploratory study with a web application for early intervention," IEEE Access, vol. 11, pp. 52288–52308, 2023.

[8] A. V. Hacimahmud and A. K. S. Alsajri, "Review of deep learning: Convolutional Neural Network Algorithm," Babylonian Journal of Machine Learning, vol. 2023, pp. 19–25, Apr. 2023.

[9] S. J. Forkel, L. Talozzi, V. Pacella, D. Pérennou, D. Tranel, M. Corbetta, C. Piscicelli, A. Boes and P. Nachev, "Latent disconnectome prediction of long-term cognitive-behavioural symptoms in stroke," Brain, vol. 146, no. 5, pp. 1963–1978, 2023.

[10] D. Kallmes, O. Joly, M. S. Jabal, T. Huynh, G. Harston, and W. Brinjikji, "Interpretable machine learning modeling for ischemic stroke outcome prediction," Frontiers in neurology, vol. 13, pp. 884693, 2022.

[11] V. Avula, C. J. Griessenauer, V. Abedi, S. Shahjouei, D. Chaudhary, J. Li, and R. Zand, "Prediction of long-term stroke recurrence using machine learning models," Journal of clinical medicine, vol. 10, no. 6, pp. 1286, 2021.

[12] S. Park, J. Yu, H. Lee, K. H. Cho and S. H. Kwon, "AI-based stroke disease prediction system using ECG and PPG bio-signals," Ieee Access, vol. 10, pp. 4363–4368, 2022.

[13] M. DÖLARSLAN, "CRISPR-Cas9 Mediated Gene Correction of CFTR Mutations in Cystic Fibrosis: Evaluating Efficacy, Safety, and Long-Term Outcomes in Patient-Derived Lung Organoids," SHIFAA, vol. 2023, pp. 41–47, May 2023.

[14] M. Hasan, A. K. Sarkar and S. Rahman "Prediction of brain stroke using machine learning algorithms and deep neural network techniques," European Journal of Electrical Engineering and Computer Science, vol. 7, no. 1, pp. 23–30, 2023.

[15] M. A. jaloud, "Advantages and disadvantages of X-rays," SHIFAA, vol. 2024, pp. 34–42, Feb. 2024.

[16] N. N. Dola, S. Bourouis T. Tazin, and M. Monirujjaman Khan, "Stroke disease detection and prediction using robust learning approaches," Journal of healthcare engineering, vol. 2021, no. 1, pp. 7633381, 2021.

[17] M. W. Merid, Y. M. Chekol, T. K. Tesfie, T. M. Tebeje, N. B. Gelaw, G. A. Tesema, N. B. Gebi, and W. S. Seretew, "Development and validation of a risk prediction model to estimate the risk of stroke among hypertensive patients in University of Gondar comprehensive specialized hospital, Gondar, 2012 to 2022," Degenerative neurological and neuromuscular disease, pp. 89–110, 2023.

[18] T. Hasan, A. Islam, A. Al Mehadi, S. M. Nasrullah, and M. R. Islam, "Exploring the performances of stacking classifier in predicting patients having stroke," In 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), pp. 242–247. IEEE, 2021.

[19] V. Mato-Abad, C. Fernandez-Lozano, S. Suárez-Garaboa, M. Rodríguez-Yáñez, I. López-Dequidt, T. Sobrino, F. Campos, A. Estany-Gestal, J. Castillo, and S. Rodríguez-Yáñez, "Random forest-based prediction of stroke outcome," Scientific reports, vol. 11, no. 1, pp. 10071, 2021.

[20] M. Hasan, S. Rahman and A. K. Sarkar, "Prediction of brain stroke using machine learning algorithms and deep neural network techniques," European Journal of Electrical Engineering and Computer Science, vol. 7, no. 1, pp. 23–30, 2023.