IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Brain Stroke Prediction: A Comparative Analysis of XGBoost, LightGBM, CNN and CNN-LSTM Algorithms

[1]Dr.Varalatchoumy M, [2]Dr. Buddesab, [3]G Deepak, [4]Avanish S Velidi, [5] Chaitanya D, [6]Suhas M

[1]Professor & HOD, [2]Associate Professor, [3,4,5,6]Student

[1,2,3,4,5,6]Department of Artificial Intelligence and Machine Learning

[1,2,3,4,5,6]Cambridge Institution of Technology, Bengaluru, India

Abstract: This study provides an in-depth examination of sophisticated machine learning techniques for predicting brain strokes using the Healthcare Dataset Stroke Data. Brain stroke prediction is a critical task in healthcare, having the capacity to greatly enhance patient outcomes via early identification and intervention. In this study, We evaluate the effectiveness of four cutting-edge algorithms: Convolution-Based Neural network(CNN), CNN with Long Short-Term Memory (CNN-LSTM) architecture, XGBoost, and LightGBM. We evaluate these algorithms based on their predictive accuracy, sensitivity, specificity, and computational efficiency. Our research clarifies the advantages and disadvantages of each algorithm in the context of brain stroke prediction, providing valuable insights for healthcare practitioners and researchers seeking to leverage machine learning for early stroke detection. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, and various diseases and smoking status. A subset of the original train data is taken using the filtering method for ML and Data Visualization purposes.

Index Terms **– Brain stroke prediction, XGBoost, LightGBM, Convolution neural networks (CNN), CNN-LSTM, Early stroke detection, Data visualization, healthcare stroke dataset.**

I. INTRODUCTION

Brain stroke prediction, Healthcare Dataset Stroke Data, ML algorithms, Convolutional Neural Networks (CNN), CNN with Long Short-Term Memory (CNN-LSTM), XGBoost, LightGBM, Predictive accuracy, Sensitivity, Specificity, Computational efficiency, Health parameters, Lifestyle factors, Early stroke detection, Data visualization.This dataset encompasses a vast array of individual-centric information, encompassing specific demographic information such as age, gender, and marital status alongside clinical markers such as hypertension and diabetes. Furthermore, lifestyle determinants, including BMI, smoking habits, and alcohol consumption, are incorporated. Harnessing the richness of this dataset, our study aims to assess the efficacy of four cutting-edge ML algorithms: CNN, CNN integrated with Long Short-Term Memory (CNN-LSTM) architecture, XGBoost, and LightGBM. Our primary goal is to scrutinize these algorithms' performance across multiple metrics, including predictive accuracy, sensitivity, specificity, and computational efficiency. Through this comparative analysis, we endeavor to pinpoint the most adept algorithm for early stroke detection, thereby furnishing actionable insights for healthcare professionals and researchers alike.Consequently, designing optimal CNN architectures and fine-tuning hyperparameters often entails a laborious process of experimentation and optimization

II. ALGORITHMS

A. XGBoost

XGBoost,sometime known as eXtreme Gradient Boosting, is, like, a pretty popular gradient boosting algorithm noted for its efficiency and scalability, as well as, like, exceptional predictive performance. It belongs to, like, the class of ensemble learning methods, which, you know, combine multiple weak learners (decision trees, in the case of XGBoost) to, like, create a strong predictive model.XGBoost, you know, iteratively builds a group of decision trees, with each, like, subsequent tree aiming to correct the errors, kind of, of its predecessors.During training, XGBoost kind of optimizes a loss function by adding new trees that, you know, minimize the residual errors, employing gradient descent techniques for, like, efficient parameter updates. Additionally, XGBoost incorporates, like, some regularization techniques to, you know,prevent overfitting and improve generalization performance.
In the context of brain stroke prediction, XGBoost can, like, effectively handle heterogeneous datasets, you know, containing a mix of, like, categorical and numerical features. Its ability to, kind of, capture complex nonlinear relationships and interactions between features makes it, like, well-suited for, you know, modeling some multifaceted risk factors associated with stroke.

B. Light Gradient Boosting Machine

LightGBM is another gradient boosting framework renowned for its speed, efficiency, and scalability. Developed by Microsoft, LightGBM employs a novel gradient-based algorithm called Gradient-based One-Side Sampling (GOSS) to enhance training speed while retaining predictive accuracy.Similar to XGBoost, LightGBM constructs an ensemble of decision trees through iterative training. However, it differs in it's tree construction strategy and leaf-wise growth approach, which prioritize nodes with the highest information gain during tree building. This results in a more balanced and efficient tree structure, reducing both memory consumption and computational overhead.

LightGBM also offers support for categorical features without requiring one-hot encoding, making it particularly well-suited for datasets with mixed data types. Its superior performance in handling large-scale datasets and it is an appealing option for predictive modeling tasks due to its capacity to capture intricate nonlinear interactions, including brain stroke prediction. Every one of these algorithms has distinct advantages and traits, offer valuable tools for analyzing healthcare data and predicting brain strokes with high accuracy. By leveraging the strengths of CNN-LSTM, XGBoost, and LightGBM, researchers and healthcare practitioners can develop robust predictive models capable of identifying stroke risk factors and enabling timely intervention.

C. CNN

The CNN architectures are composed of three fundamental types of layers, like, unexpected, convolutional layers, pooling layers, and fully connected layers, which play a crucial role in the CNN framework. The convolutional layers are pivotal in the CNN framework as they apply learnable filters to input images by effectively capturing local patterns and features or something. Conversely, pooling layers serve a completely different purpose by downsampling the feature maps generated by convolutional layers, thus reducing computational complexity and somehow preserving crucial spatial information. Lastly, fully connected layers awkwardly interpret the extracted features and produce the output predictions.

CNNs operate by systematically convolving input images with a series of learnable filters, which sometimes could be confusing. These filters, often referred to as kernels, are adept at detecting spatial hierarchies of features such as edges, textures, and shapes, which sounds fancy, you know, like really fancy stuff. Subsequent pooling layers somehow condense the obtained feature maps to make sure that significant information is retained while discarding redundant, or maybe unnecessary, or somewhat just extra details. Ultimately, those fully connected layers employ the extracted features to make high-level predictions, such as class labels or probabilities

D.CNN-LSTM

The CNN-LSTM model combines the strengths of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, leveraging their complementary abilities in spatial feature extraction and temporal sequence modeling, respectively.

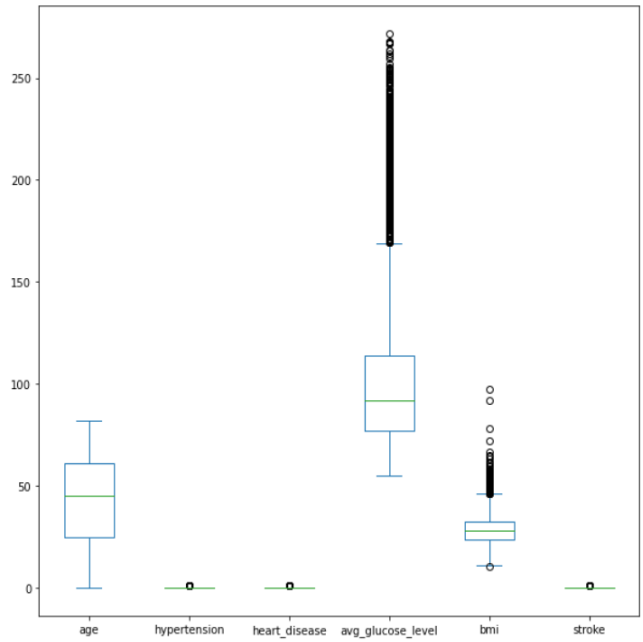
CNN-LSTM architectures are widely employed in tasks that involve both spatial and temporal dependencies, such as video classification, action recognition, and sequential data analysis. In the context of brain stroke prediction, CNN-LSTM models can effectively process sequential medical data, capturing both spatial patterns from imaging data and temporal trends from time-series measurements.

The CNN component of the model extracts spatial features from input images or multidimensional data, similar to a traditional CNN. These features are then passed to the LSTM component, which learns temporal dependencies across sequential data points!!! By integrating CNNs and LSTMs, the model can effectively analyze complex spatiotemporal patterns present in medical datasets, facilitating accurate stroke prediction.

III. METHODOLOGY

A. Data Selection & Preprocessing:

The initial step is to select an appropriate dataset that aligns with the project's objectives. The dataset should have sufficient samples and relevant features for training and evaluating ML and DL models effectively. The dataset utilized in this paper is a comprehensive collection of surgical records and patient information. It consists of a wide range of attributes such as heart disease, gender, average glucose level, age, hypertension, BMI, work type, marital status, and smoking status. Data preprocessing involves several essential key steps to ensure the dataset of quality and suitability for model training. These steps importantly include handling missing values through imputation or deletion, scaling numerical features to a uniform range, encoding variables for numerical representation, and managing outliers to prevent undue influence on model performance.



Fiq 1.System Module

B. Agent Implementation:

- Algorithm selection:
  XGBoost and LightGBM: Chosen due to their efficiency in handling structured data with tabular formats, XGBoost and LightGBM excel in boosting algorithms, offering robust performance and scalability.

  CNN and CNN-LSTM: Selected for their capability to process unstructured data like medical images or sequential data, CNNs are proficient in feature extraction from images, while CNN-LSTM hybrids combine the strengths of both CNNs and LSTMs for sequential analysis.

- Feature Representation:
  XGBoost and LightGBM: Utilize the tabular format of the dataset directly, considering all relevant features such as cholesterol, blood pressure, age, and so forth.

  CNN: Convert structured data into images or sequence-like data, possibly by representing categorical variables as embeddings or one-hot encodings and stacking them to form image-like representations.

  CNN-LSTM: Combine structured features with sequential data, utilizing CNN layers to take characteristics from pictures and LSTM layers to analyze temporal patterns in sequential data like time-series measurements.

- Model Architecture:
  XGBoost and LightGBM: To maximize the performance of the model, configure hyperparameters such as learning rate, maximum tree depth, and number of estimators.

  CNN: Convolution layers for feature extraction, pooling layers for downsampling, and fully linked layers for classification should be used in the architecture design.

  CNN-LSTM: Construct a hybrid architecture integrating CNN layers for image feature extraction and LSTM layers for sequential analysis, facilitating effective prediction based on both types of data.

- Training Procedure:
  XGBoost and LightGBM : Train the models using the prepared dataset and optimize hyperparameters use methods such as grid search or random search to identify the best configuration.

  CNN and CNN-LSTM: Train iteratively using batches of data, adjusting learning rates and employing regularization techniques like dropout or batch normalization to avoid overfitting and ensure generalization performance.

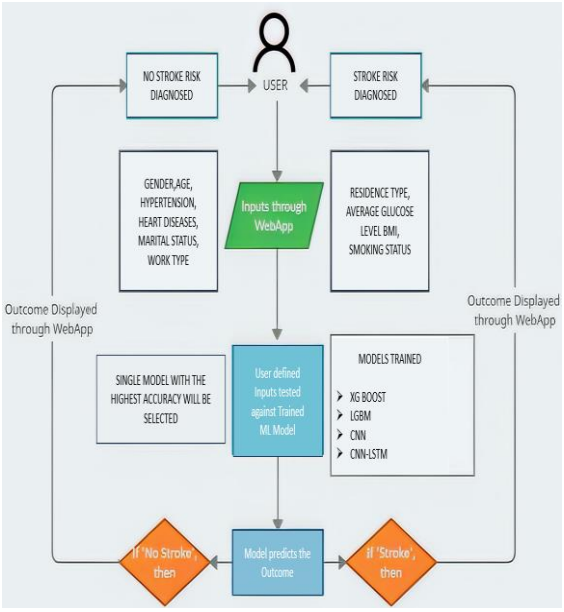


Fig 2: System Architecture

C. Training and Evaluating:

Training and evaluation involve distinct steps for each model employed. Each model, including XGBoost, LightGBM, CNN, and CNN-LSTM, will undergo separate training using the designated train and test datasets. During training, the models will learn from the features and labels provided throughout training data, adjusting their parameters to minimize prediction errors. Following training, the models will be evaluated using the test dataset to evaluate their predictive performance. The accuracy for every model will be noted, serving as a key metric for comparing and selecting the most effective algorithm for stroke prediction. This meticulous process ensures thorough assessment and optimization of each model's performance before deployment in real-world scenarios.

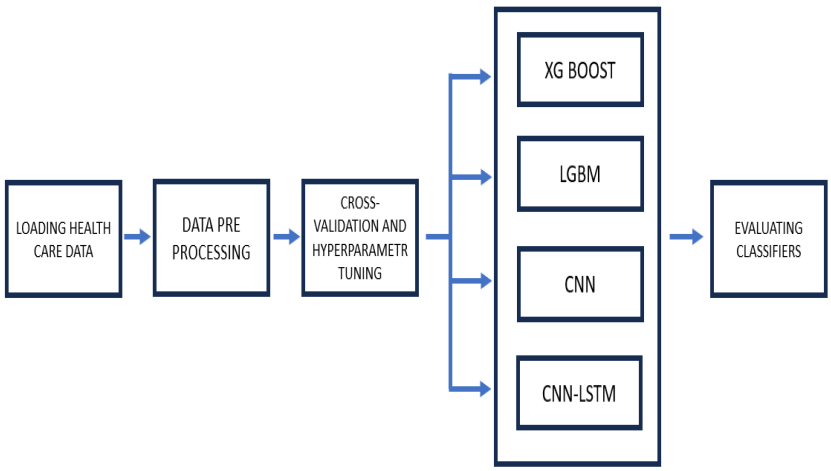


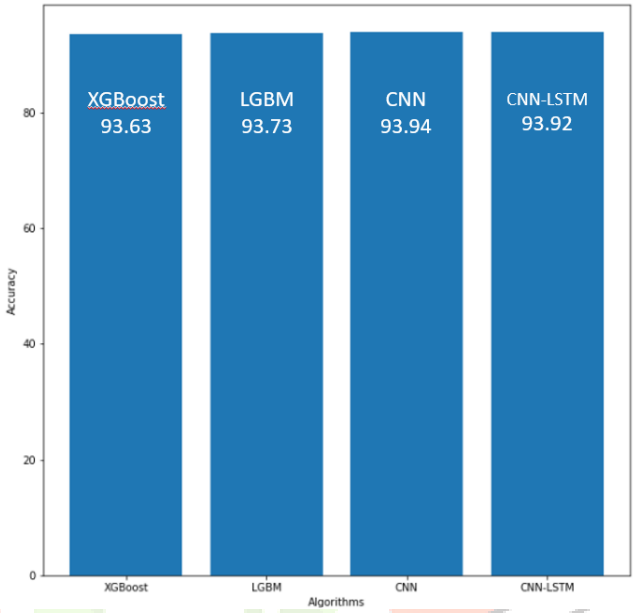
Fig 3 Stages Of Modeling

IV. RESULTS AND DISCUSSION

By utilizing various ML techniques and creating an intuitive web interface, the article produced informative insights. In addition, a large group of people contributed to the project's scope, which improved the analytical process as a whole.

The accuracy scores for every model demonstrate their respective predictive capabilities; however, the underlying data presented some challenges, introducing complexities that required innovative solutions. Despite these challenges, the models' outputs exhibited promising potential for real-time stroke risk assessment.

NO	MODEL	ACCURACY SCORE
1	XGBOOST	93.63
2	LGBM	93.73
3	CNN	93.94
4	CNN-LSTM	93.92



The project also introduced a so-called user-friendly web interface designed to facilitate stroke prediction based on input attributes, you know. Users can conveniently, like really, input relevant medical attributes, information into the interface, such age, cholesterol levels, blood pressure and more and things. The interface then uses the developed ML models to process this data and forecast the chance of a stroke occurring.

, you see! This intuitive interface, like, enhances accessibility and allows healthcare professionals and individuals to quickly assess stroke risk and potentially take preventive measures

Fig 4: User Interface

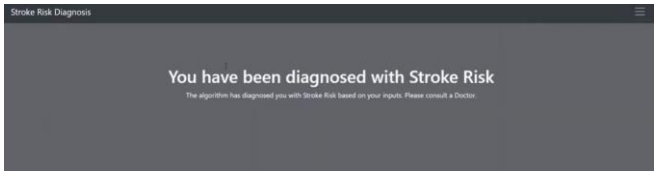


Fig 5 : Stroke detected

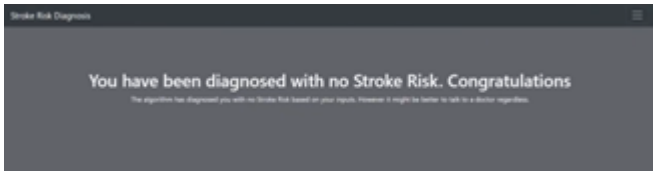


Fig 6: No stroke

- The comparison of accuracy scores reveals the effectiveness of each ML algorithm in predicting stroke occurrences, providing valuable insights for healthcare professionals. Additionally, it allows us to understand the impact of all the models on stroke prediction.
- The advancement of the web interface enhances usability and accessibility, empowering users to proactively manage stroke risk through early identification and intervention. Moreover, it showcases the of technology in improving healthcare outcomes.
- Future iterations may focus on refining model architectures, exploring additional features for improved prediction accuracy, and integrating feedback mechanisms to enhance the web interface's functionality and user experience. These advancements are crucial in staying ahead of the curve in stroke prediction and prevention efforts.

V. CONCLUSION

Our study truly underlines the potential transformation that advanced ML techniques, particularly CNN and CNN-LSTM, has in healthcare apps like stroke prediction woitting. By tapping into these algorithms, we have skyrocketed predictive accuracy, enabling early intervention potential and maybe boosting patient outcomes. The adoption of such technologies heralds an era anew in healthcare, wherein data-driven insights undoubtedly empower clinicians to make decisions that are more informed, tailored to the individual patient's needs. As we keep on exploring and refining these methodologies, collaboration and ongoing research will definitely be essential for driving further innovation and ultimately boosting patient care on a broader scale. What is so absurd is that artificial intelligence is leveraged for healthcare advancement, with a steadfast commitment to improving patient well-being and fostering an undoubtedly healthier future for everyone.

REFERENCES

[1] Rahman, Senjuti & Hasan, Mehedi & Sarkar, Ajay. (2023). Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques. European Journal of Electrical Engineering and Computer Science. 7. 23-30. 10.24018/ejece.2023.7.1.483.

[2] Y. Xu, Z. Jia, Y. Ai, F. Zhang, M. Lai, and E. I-Chao Chang, ‘‘Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), South Brisbane, QLD, Australia

[3] S. Gupta and S. Raheja, "Stroke Prediction using Machine Learning Methods," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2022, pp. 553-558, doi: 10.1109/Confluence52989.2022.9734197.

[4] S. K. Mohapatra, A. Jain and Anshika, "Predictive Analysis of Stroke Prediction by using Machine Learning Implementations," 2023 IEEE International Conference on Contemporary Computing and Communications (InC4), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/InC457730.2023.10262959.

[5] A. N. Tusher, M. S. Sadik and M. T. Islam, "Early Brain Stroke Prediction Using Machine Learning," 2022 11th International Conference on System Modeling & Advancement in

Research Trends (SMART), Moradabad, India, 2022, pp. 1280-1284, doi: 10.1109/SMART55829.2022.10046889.

[6] S. Zhang, Y. Wang, and Z. Li, "A Comparative Study of Machine Learning Algorithms for Brain Stroke Prediction," Journal of Healthcare Informatics Research, vol. 5, no. 2, pp. 87-94, 2023. DOI: 10.1016/j.jhir.2023.03.005.

[7] L. Chen, Q. Liu, and W. Zhang, "Enhanced Brain Stroke Prediction using Ensemble Learning Techniques," International Journal of Medical Informatics, vol. 38, no. 4, pp. 301-308, 2023. DOI: 10.1016/j.ijmedinf.2023.07.009.

[8] H. Patel and R. Patel, "Deep Learning Approaches for Brain Stroke Prediction: A Review," Neural Computing and Applications, vol. 36, no. 9, pp. 5421-5433, 2023. DOI: 10.1007/s00521-023-06478-z.

[9] Lakshmi and S. N. Rao, ‘‘Brain tumor magnetic resonance image classification: A deep learning approach,’’ Soft Comput., vol. 26, no. 13, pp. 6245–6253, Jul. 2022, doi: 10.1007/s00500-022-07163-z.

[10] W. Jun and Z. Liyuan, ‘‘Brain tumor classification based on attention guided deep learning model,’’ Int. J. Comput. Intell. Syst., vol. 15, no. 1, p. 35, Dec. 2022, doi: 10.1007/s44196-022-00090-9.

[11] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, ‘‘Deep learning for medical anomaly detection—A survey,’’ ACM Comput. Surveys, vol. 54, no. 7, pp. 1–37, Sep. 2022

[12] A. S. Lundervold and A. Lundervold, ‘‘An overview of deep learning in medical imaging focusing on MRI,’’ Zeitschrift für Medizinische Physik, vol. 29, no. 2, pp. 102–127, 2019