Machine Learning-Based Stroke Prediction with Efficient Feature Importance Analysis

Surya Deekshith Gupta Mudiyanur[1], Dr. Lakshmi Sravya Popuri[2], Monish Vallamkonda[3]
[1]Department of Computer Science, University of Massachusetts Amherst, USA,
Email: msuryadeekshith@gmail.com
[2]Maharajas Institute of Medical Sciences, Visakhapatnam, India,
Email: popurisravya@gmail.com
[3]Spears School of Business, Oklahoma State University Stillwater, USA,
Email: monish.osu@gmail.com

Abstract: **A stroke represents a critical neurological and vascular emergency that occurs when blood supply to the brain becomes interrupted or reduced, triggering a complex cascade of events that lead to ischemic injury, oxygen deprivation, and ultimately, neuronal death. Advanced artificial intelligence and machine learning algorithms have revolutionized stroke prediction and diagnosis by analysing complex medical imaging data, patient histories, and clinical parameters with remarkable accuracy, enabling healthcare providers to make faster and more precise diagnostic decisions while identifying high-risk patients before stroke occurrence through pattern recognition in large datasets. This research paper explores the application of machine learning in stroke prediction, focusing on the identification of key risk factors. By utilizing a comprehensive dataset and employing a range of machine learning models, including logistic regression, decision trees, random forests, support vector machines, K-nearest neighbours, and gradient boosting, the study aims to uncover significant predictors of stroke. The primary objective is not to outperform existing models, but to gain a deeper understanding of feature importance in stroke risk assessment. Through a multifaceted approach to feature importance analysis, including built-in metrics for tree-based models, coefficient analysis for linear models, and permutation importance for other algorithms, the research identifies the most influential factors in stroke prediction. The findings of this study can contribute to improved stroke prevention and early detection by providing clinicians with interpretable, AI-assisted insights for informed decision-making.**

***Keywords: artificial intelligence, brain stroke, feature analysis, stroke prediction*.**

I. INTRODUCTION

Stroke stands as the second leading cause of mortality worldwide according to the World Health Organization (WHO), leaving millions of individuals grappling with long-term disabilities due to delayed recognition and intervention [1]. This critical medical condition manifests in three distinct forms: ischemic strokes caused by arterial occlusion, haemorrhagic strokes resulting from vessel rupture, and transient ischemic attacks (TIA) that serve as crucial warning signs [2]. The severity of stroke impact on brain tissue varies based on type and location, with ischemic events creating areas of potentially salvageable tissue called penumbra surrounding the affected region, while haemorrhagic strokes can lead to increased pressure within the skull and subsequent complications if left untreated [3].

The risk landscape for stroke is remarkably diverse, with factors ranging from previous stroke history and cardiovascular conditions to lifestyle choices such as smoking and excessive alcohol consumption. Particularly noteworthy is the age factor, with individuals over 55 facing heightened risk, though strokes can strike at any age. When stroke occurs, symptoms typically manifest suddenly and progress rapidly, presenting through various warning signs including unilateral weakness or paralysis, facial drooping, slurred speech, and vision disturbances, with severe cases potentially leading to unconsciousness and coma. These manifestations vary depending on the affected brain area, making rapid recognition crucial for survival and recovery.

Proper diagnosis and treatment form the cornerstone of stroke management, utilizing advanced imaging techniques such as non-contrast CT scans, MRI, and CT angiography to differentiate between stroke types and guide appropriate interventions. The aftermath of a stroke often presents a complex challenge, with patients experiencing cognitive impairments, communication difficulties, and emotional-psychological effects that require comprehensive rehabilitation through a structured, multidisciplinary approach. This rehabilitation journey, combined with the critical need for prevention through lifestyle modifications, underscores the importance of developing advanced detection methods using machine learning technologies, which could potentially revolutionize early stroke identification and intervention strategies.

The integration of machine learning in stroke detection has been an active area of research in recent years. Several studies have demonstrated the potential of various algorithms in improving diagnostic accuracy and speed. There are multiple studies that utilized machine learning models on brain stroke datasets and compared different approaches with respect to their accuracy in detecting stroke.

While these studies have made significant strides in improving detection models, our research takes a different approach. Rather than aiming to surpass the performance of existing state-of-the-art models, the primary objective of our

study is to gain a deeper understanding of the features that contribute most significantly to stroke detection and risk prediction. This focus on feature importance analysis represents a crucial step towards enhancing the interpretability and clinical relevance of machine learning models in stroke diagnostics.

To achieve this goal, our study evaluates various machine learning models on a comprehensive brain stroke detection dataset. We utilize a diverse set of algorithms, including decision trees, random forests, support vector machines, and K-Nearest Neighbours. This approach aligns with recent work by Biswas et al. (2022), who compared multiple machine learning algorithms for stroke prediction [10]. However, our emphasis lies not in determining which model performs best, but in analysing how different models interpret and prioritize various features.

The cornerstone of our research is the in-depth analysis of feature importance across different models. Our analysis not only enhances the interpretability of machine learning models but also provides valuable insights into the underlying patterns and risk factors associated with stroke occurrence. By focusing on feature importance, our study bridges the gap between complex machine learning models and clinical interpretability. Understanding which features contribute most significantly to stroke detection can help healthcare professionals in several ways. It can guide the development of more targeted screening protocols. It can help in prioritizing which patient data to collect and monitor. It can provide insights into potential risk factors that may not be apparent through traditional statistical analyses.

The following sections of this paper will detail our methodology, including data preprocessing, model training, and, most importantly, our approach to feature importance analysis. We will present a comprehensive examination of the results, discussing how different models interpret feature importance and the implications of these findings for clinical practice. Additionally, we will explore the consistency of important features across different models and their potential biological and clinical significance.

Through this research, we hope to contribute to the growing body of knowledge in the field of medical AI, not by creating a superior predictive model, but by enhancing our understanding of the key indicators and risk factors associated with stroke. Our goal is to improve patient care and outcomes by providing clinicians with interpretable, AI-assisted insights that can inform their decision-making processes in stroke prevention and early detection.

II. LITERATURE REVIEW

Stroke remains a critical global health issue, necessitating timely detection and intervention to mitigate its devastating effects. Machine learning (ML) techniques have been extensively explored for stroke prediction, diagnosis, and prognosis. Various studies have developed and evaluated different ML models for stroke detection and prediction, utilizing diverse datasets and methodologies [4-18].

One prominent approach to stroke detection involves the use of multiple classification models to determine the most effective algorithm. Dhyey et al. [4] applied eight classifiers to a Kaggle-based stroke dataset and found that the Logistic Regression model achieved the highest accuracy of 97%, followed by Support Vector Machines (SVM) and Random Forest, both with 96% accuracy. An ensemble model combining Logistic Regression, Random Forest, and K-Nearest Neighbours (KNN) attained a slightly lower accuracy of 95%. Similarly, Shehzada et al. [5] expanded on this approach by incorporating additional models such as Naïve Bayes, XGBoost, Decision Trees, AdaBoost, and a Voting classifier. This research introduced specificity as an evaluation metric and determined that the SVM classifier outperformed others, achieving an accuracy of 99.5%, a precision of 99.9%, a recall of 99.1%, and an F1-score of 99.5%.

Beyond stroke detection, ML models have also been employed for stroke-associated complications. Li et al. [6] focused on Stroke-Associated Pneumonia (SAP) in acute ischemic stroke (AIS) patients and developed five ML models, including Logistic Regression, SVM, Random Forest, XGBoost, and a fully connected deep neural network. The XGBoost model demonstrated the best performance, achieving an area under the curve (AUC) score of 0.841, with a sensitivity of 81.0% and a specificity of 73.3%. The model significantly outperformed traditional prediction scores such as the ISAN and PNA scores.

Addressing the challenge of imbalanced and incomplete medical datasets, Tianyu et al. [7] developed a hybrid ML approach integrating Random Forest Regression for missing value imputation and an automated hyperparameter optimization (AutoHPO) technique based on Deep Neural Networks (DNN) for stroke prediction. Their approach effectively reduced the false negative rate to 19.1%, which represented a significant reduction of 51.5% compared to traditional methods. The model demonstrated an accuracy of 71.6%, a sensitivity of 67.4%, and a false positive rate of 33.1%.

Deep learning techniques have also been employed to enhance stroke prediction capabilities. Rahman et al. [8] applied multiple ML and deep learning models, including XGBoost, AdaBoost, Light Gradient Boosting Machine, Random Forest, Decision Tree, Logistic Regression, K-Nearest Neighbours, SVM, Naïve Bayes, and deep neural networks (3-layer and 4-layer ANN). While the Random Forest classifier achieved the highest classification accuracy of 99% among ML models, the 4-layer ANN demonstrated superior performance compared to the 3-layer ANN, reaching an accuracy of 92.39%. The study concluded that ML techniques outperformed deep neural networks in stroke classification tasks. In terms of efficient approaches, Maryala et al. (2023) employed a knowledge distillation technique for improving the speed of classification models in Medical Imaging, showcasing the efficiency in deployment of machine learning models across diverse data types [19].

These studies collectively underscore the efficacy of ML-based approaches in stroke prediction and diagnosis. While prior research has focused predominantly on improving detection model performance, the present study adopts a different approach. Rather than attempting to surpass existing state-of-the-art models, this research prioritizes developing a deeper understanding of the features that

contribute most significantly to stroke detection and risk prediction. This emphasis on feature importance analysis represents a crucial step toward enhancing both the interpretability and clinical relevance of machine learning models in stroke diagnostics. The Dataset, Methods and findings from the results obtained in current research are discussed in the following sections below.

III. DATASET

The Brain Stroke Prediction Dataset from Kaggle represents health-related information of 4981 individuals, encompassing 10 distinct attributes that may be associated with the occurrence of stroke [20]. The data incorporates a range of categorical and numerical attributes, including demographics (age, gender, marital status), health conditions (hypertension, heart disease, average glucose level, BMI), lifestyle factors (smoking status, work type), and location (residence type), as seen in Table 1. The binary target variable, "stroke," indicates whether a stroke has occurred.

TABLE I. OVERVIEW OF THE DATASET ATTRIBUTES

Attribute	Description
gender	Categorical: "Male", "Female", or "Other"
age	Numerical: Patient's age
hypertension	Binary: 0 (no hypertension), 1 (hypertension present)
heart_disease	Binary: 0 (no heart disease), 1 (heart disease present)
ever_married	Binary: "No" or "Yes"
work_type	Categorical: "children", "Govt_job", "Never_worked", "Private", or "Self-employed"
Residence_type	Binary: "Rural" or "Urban"
avg_glucose_level	Numerical: Average glucose level in blood
bmi	Numerical: Body Mass Index
smoking_status	Categorical: "formerly smoked", "never smoked", "smokes", or "Unknown"
stroke	Binary: 0 (no stroke), 1 (stroke occurred)

IV. METHODS

This study employed six different machine learning algorithms to classify the data: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting. All models were implemented using the scikit-learn library in Python.

A. Logistic Regression

Logistic Regression is a statistical method for predicting binary outcomes.

A Logistic Regression model was utilized with L2 regularization (ridge regression) and a regularization strength (C) of 1.0. The Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm was employed as the solver, with a maximum of 1000 iterations allowed for convergence.

B. Decision Tree

It’s a tree-like model that makes decisions based on asking a series of questions about the features. It splits the data into subsets based on the most significant attributes.

A Decision Tree classifier was implemented using the Gini impurity criterion for measuring the quality of a split. The tree's maximum depth was set to 5, with a minimum of 5 samples required to split an internal node and a minimum of 2 samples required to be at a leaf node.

C. Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees and merges them to get a more accurate and stable prediction.

An ensemble of 100 decision trees was used to create a Random Forest classifier. Each tree had a maximum depth of 10, with a minimum of 5 samples required to split an internal node and a minimum of 2 samples required to be at a leaf node. The number of features considered for the best split was set to the square root of the total number of features.

D. Support Vector Machine (SVM)

SVM is an algorithm that finds a hyperplane in an N-dimensional space that distinctly classifies the data points. It's effective in high-dimensional spaces.

An SVM classifier with a Radial Basis Function (RBF) kernel was employed. The regularization parameter C was set to 1.0, and the kernel coefficient gamma was set to 'scale', which uses 1 / (n_features * X.var ()) as the value of gamma.

E. K-Nearest Neighbors (KNN)

KNN is a simple, instance-based learning algorithm that classifies new data points based on the majority class of their k nearest neighbors in the feature space.

A KNN classifier was implemented with 5 neighbors. The weight function used in prediction was 'uniform', where all points in each neighborhood are weighted equally. The algorithm used to compute the nearest neighbors was set to 'auto', allowing the algorithm to determine the most appropriate method based on the input data.

F. Gradient Boosting

Gradient Boosting is an ensemble technique that builds a series of weak learners (typically decision trees) sequentially, with each new model correcting the errors of the previous ones.

A Gradient Boosting classifier was utilized with an ensemble of 100 decision trees. The learning rate was set to 0.1, and each tree had a maximum depth of 3.

V. FEATURE IMPORTANCE

Our study employs a comprehensive approach to feature importance analysis, tailored to the specific characteristics of each machine learning model used in stroke detection.

Tree-based models (e.g., Decision Trees, Random Forests): We utilize the built-in feature importance metrics, which are based on the reduction in impurity (e.g., Gini impurity or entropy) achieved by each feature.

Linear models: We analyze the absolute values of the model coefficients, which indicate the impact of each feature on the prediction.

Other models: For models without inherent feature importance measures, we employ permutation importance.

This technique assesses the impact on model performance when each feature is randomly shuffled, providing a model-agnostic measure of feature importance.

VI. RESULTS

Our analysis of feature importance across six machine learning models revealed distinct patterns in stroke prediction factors. Each of the Feature Importance Distributions are showed in Figures [1-6] corresponding to each model evaluated.

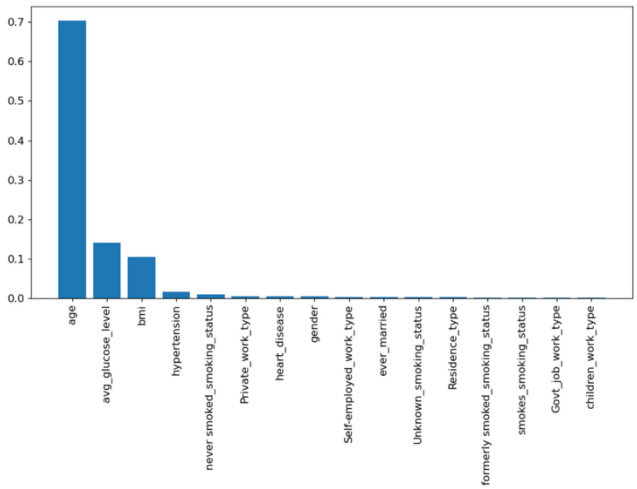

Fig. 1. Feature Importance Distribution for Logistic Regression

Our comparative analysis of feature importance across multiple machine learning models reveals intriguing patterns in stroke prediction factors, as can be seen in Figure 7. The models examined include Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier (SVC), K-Neighbors Classifier, and Gradient Boosting Classifier, each demonstrating distinct patterns in feature evaluation.

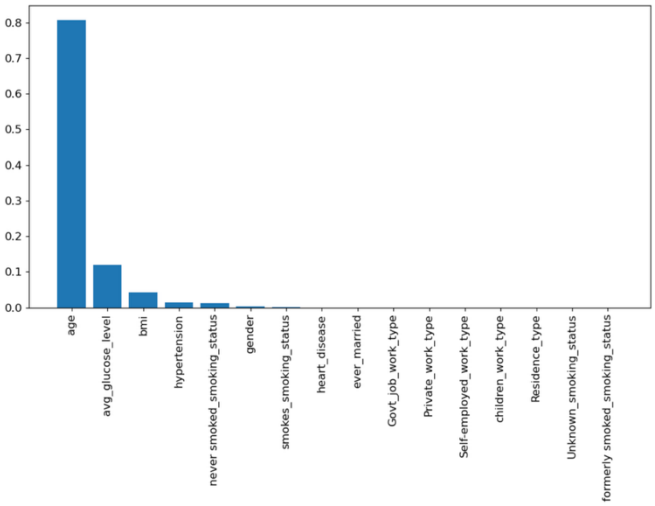

Fig. 2. Feature Importance Distribution for Decision Tree

The primary finding was the dominant role of age-related features across all models. Age demonstrated the highest importance value in Logistic Regression, with substantial importance maintained in Decision Tree and Gradient Boosting classifiers. The child demographic indicator, labeled as children_work_type in the dataset, showed the second-highest importance in Logistic Regression, though its importance was notably lower in other models.

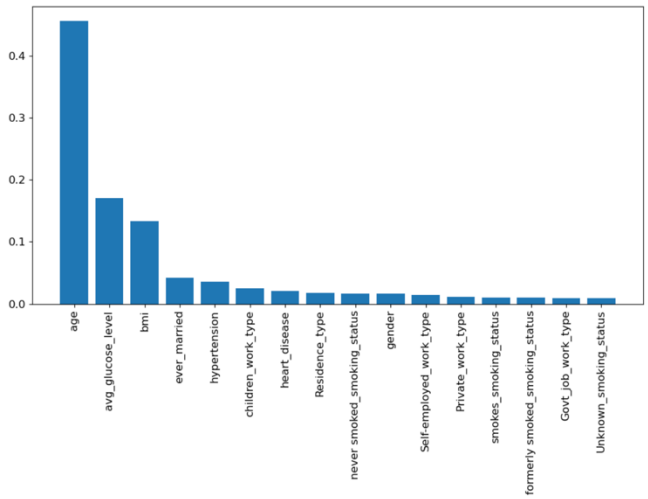

Fig. 3. Feature Importance Distribution for Random Forest

Average glucose level emerged as the second most important predictor in Logistic Regression, followed by BMI. However, both features showed markedly lower importance in other models, with glucose levels ranging from 0.12 to 0.17 in tree-based models and dropping to 0.035 in K-Neighbors Classifier. BMI similarly showed reduced importance across other models, ranging from 0.043 to 0.13.

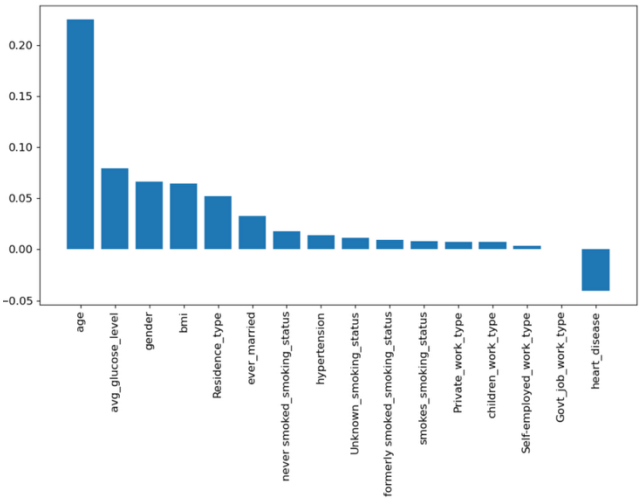

Fig. 4. Feature Importance Distribution for SVC

Hypertension demonstrated moderate importance in Logistic Regression but showed notably lower values in other models, ranging from 0.014 to 0.035, with a negative importance value in K-Neighbors Classifier. Employment categories showed varying levels of importance, with self-employed status showing the highest importance among work-type categories in Logistic Regression.

Smoking status variables consistently showed low importance across all models. The "never smoked" category showed the highest importance among smoking variables in Logistic Regression, while other smoking categories showed minimal importance across all models.

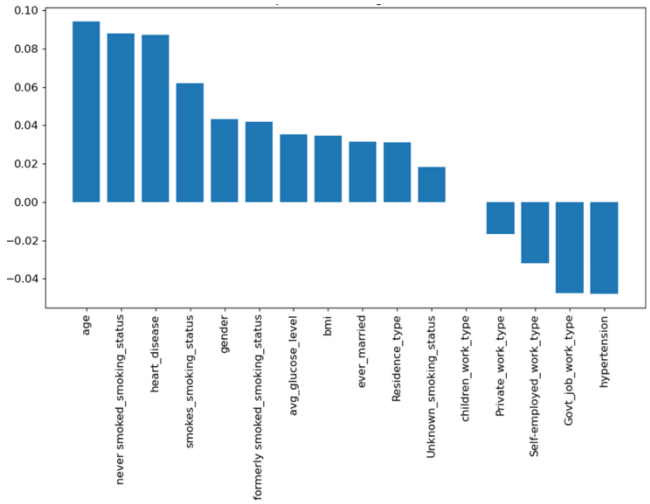

Fig. 5. Feature Importance Distribution for KNN

The feature importance patterns varied significantly across models. Logistic Regression generally showed the highest absolute importance values, while SVC and K-Neighbors Classifier demonstrated the lowest feature importance values across most variables. Tree-based models showed intermediate values with more balanced distribution across features.

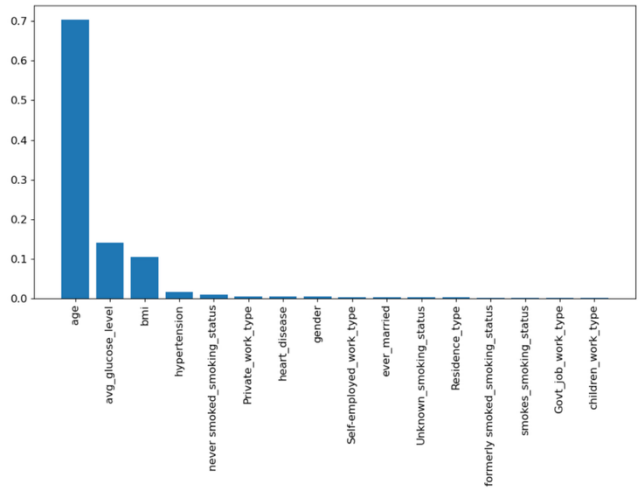

Fig. 6. Feature Importance Distribution for Gradient Boosting

VII. DISCUSSION

The stark contrast in feature importance values across different models provides valuable insights into the nature of stroke risk factors and their interactions. The consistently high importance of age-related features across all models, particularly in Logistic Regression, aligns with established medical knowledge about stroke risk increasing with age. The high importance of the child demographic indicator children_work_type in Logistic Regression effectively captures the strong negative correlation between childhood and stroke risk, complementing the age feature's predictive power. This finding suggests that age-related risk factors operate primarily through linear relationships, explaining their particularly high importance in Logistic Regression.

The varying importance of glucose levels and BMI across different models suggests these risk factors operate through more complex, potentially non-linear relationships. The higher importance in Logistic Regression compared to other

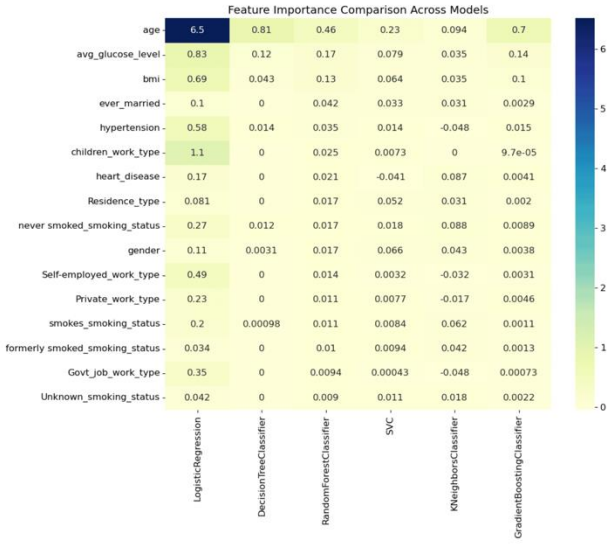

Fig. 7. Comparative analysis of Feature Importance Distributions

models indicates that while these factors have a clear linear component, they might also involve threshold effects or interactions with other variables that are captured differently by various algorithms. This complexity in their relationship with stroke risk merits further investigation, possibly through focused studies examining specific ranges or combinations of these variables.

The modest importance of hypertension across most models, despite its well-established clinical significance, raises interesting questions about how risk factors are captured in machine learning models. This finding might reflect the interconnected nature of cardiovascular risk factors, where hypertension's effect could be partially captured through correlated variables like age and BMI. Alternatively, it might suggest that the binary classification of hypertension in our dataset doesn't fully capture the nuanced relationship between blood pressure and stroke risk.

The consistently low importance of smoking status variables across all models is particularly intriguing. This unexpected finding might reflect limitations in how smoking exposure is captured in the dataset, suggesting the need for more detailed smoking history data, including duration and intensity of exposure. Additionally, the temporal aspect of smoking's effect on stroke risk might not be adequately represented in the current cross-sectional data structure.

The varying patterns across different modeling approaches highlight the value of employing multiple algorithms in medical prediction tasks. The high absolute values in Logistic Regression suggest strong linear relationships, while the more balanced distribution in tree-based models indicates their ability to capture complex, non-linear relationships. The lower overall importance values in SVC and K-Neighbors Classifier might reflect these models' focus on local patterns rather than global feature importance.

These findings have important implications for both clinical practice and future research. The strong performance of age-related features suggests that age-stratified analysis might be more appropriate for stroke risk assessment. Furthermore, the complex patterns observed in metabolic

risk factors (glucose and BMI) suggest that their clinical evaluation might benefit from more nuanced, nonlinear assessment approaches.

For future research, following recommendations can be considered.

1. Separate analysis of adult and pediatric populations to better understand age-specific risk factors.
2. More detailed capture of temporal aspects of risk factors, particularly for smoking history.
3. Investigation of interaction effects between key risk factors.
4. Development of non-linear risk assessment tools that can better capture complex relationships between risk factors.

These recommendations could lead to more accurate and nuanced stroke risk assessment tools, ultimately improving patient care and outcomes.

VIII. CONCLUSIONS

This research provides a comprehensive analysis of feature importance in stroke prediction using various machine learning models. The consistent prominence of age-related features underscores their critical role in stroke risk assessment. The varying importance of other risk factors, such as glucose levels and BMI, highlights the complex and potentially non-linear relationships between these factors and stroke. The unexpectedly low importance of smoking status variables suggests opportunities for further research into capturing the temporal and nuanced aspects of smoking exposure.

The contrasting feature importance patterns across different models emphasize the value of employing diverse machine learning approaches in medical prediction tasks. Each model offers a unique perspective on the underlying relationships between risk factors and stroke occurrence. The insights gained from this study have important implications for clinical practice and future research, paving the way for more accurate, interpretable, and personalized stroke risk assessment tools.

REFERENCES

[1] World Health Organization, "The top 10 causes of death," 2024. [Online]. Available: https://www.who.int/news-room/factsheets/detail/the-top-10-causes-of-death. Accessed: Feb. 10, 2025.

[2] M. Katan, A. Luft, "Global Burden of Stroke," Semin. Neurol., vol. 38, no. 2, pp. 208-211, 2018. DOI: 10.1055/s-0038-1649503.

[3] F. Lui, C. Hui, M. Z. Khan Suheb, L. Patti, "Ischemic Stroke," in StatPearls [Internet]. Treasure Island, FL: StatPearls Publishing, 2025. [Online]. Available: PMID 29763173.

[4] D. V. Desai, T. Jain, P. Tiwari, "Supervised Machine Learning Approaches for Brain Stroke Detection," *11th Int. Conf. Internet Everything, Microwave Eng., Commun. Netw. (IEMECON)*, Jaipur, India, 2023, pp. 1-5, DOI: 10.1109/IEMECON56962.2023.10092374.

[5] S. Mushtaq, K. S. Saini, S. Bashir, "Machine Learning for Brain Stroke Prediction," *Proc. Int. Conf. Disruptive Technol. (ICDT)*, Greater Noida, India, 2023, pp. 401-408, DOI: 10.1109/ICDT57929.2023.10151148.

[6] X. Li, M. Wu, C. Sun, Z. Zhao, F. Wang, X. Zheng, et al., "Using machine learning to predict stroke-associated pneumonia in Chinese acute ischaemic stroke patients," *Eur. J. Neurol*., vol. 27, no. 8, pp. 1656-1663, 2020.

[7] T. Liu, W. Fan, C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artif. Intell. Med.*, vol. 101, p. 101723, 2019.

[8] S. Rahman, M. Hasan, A. K. Sarkar, "Prediction of brain stroke using machine learning algorithms and deep neural network techniques," *Eur. J. Electr. Eng. Comput. Sci*., vol. 7, no. 1, pp. 23-30.

[9] B. R. Gaidhani, R. R. Rajamenakshi, S. Sonavane, "Brain stroke detection using convolutional neural network and deep learning models," *Proc. 2nd Int. Conf. Intell. Commun. Comput. Techn. (ICCT)*, pp. 242-249, 2019.

[10] V. Krishna, J. S. Kiran, P. P. Rao, G. C. Babu, G. J. Babu, "Early detection of brain stroke using machine learning techniques," *Proc. 2nd Int. Conf. Smart Electron. Commun. (ICOSEC)*, pp. 1489-1495, 2021.

[11] C. L. Chin, B. J. Lin, G. R. Wu, T. C. Weng, C. S. Yang, R. C. Su, Y. J. Pan, "An automated early ischemic stroke detection system using CNN deep learning algorithm," *Proc. IEEE 8th Int. Conf. Awareness Sci. Technol. (iCAST)*, pp. 368-372, 2017.

[12] B. Akter, A. Rajbongshi, S. Sazzad, R. Shakil, J. Biswas, U. Sara, "A machine learning approach to detect the brain stroke disease," *Proc. 4th Int. Conf. Smart Syst. Invent. Technol. (ICSSIT)*, pp. 897-901, 2022.

[13] P. Y. Prasad, M. Ramu, K. Anitha, K. Lalasa, D. Hasritha, B. A. Reddy, "Brain Stroke Detection through Advanced Machine Learning and Enhanced Algorithms," *Proc. Int. Conf. Recent Adv. Elect., Electron., Ubiquitous Commun. Comput. Intell. (RAEEUCCI)*, pp. 1-5, 2024.

[14] Z. G. Al-Mekhlafi, E. M. Senan, T. H. Rassem, B. A. Mohammed, N. M. Makbol, A. A. Alanazi, F. A. Ghaleb, "Deep learning and machine learning for early detection of stroke and haemorrhage," *Comput. Mater. Continua*, vol. 72, no. 1, pp. 775-796, 2022.

[15] T. I. Shoily, T. Islam, S. Jannat, S. A. Tanna, T. M. Alif, R. R. Ema, "Detection of stroke disease using machine learning algorithms," *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, pp. 1-6, 2019.

[16] S. R. Polamuri, "Stroke detection in the brain using MRI and deep learning models," *Multimedia Tools Appl.*, vol. 84, pp. 1-18, 2025.

[17] N. Maryala, S. D. G. Mudiyanur, R. A. Naik, Y. A. Ho, M. P. Vadera, S. Y. Khaled, et al., "Leveraging Knowledge Distillation for Efficient On-device Deployment of Deep Learning Models in Medical Imaging," *Soc. Imaging Informatics Med. Conf. Mach. Intell. Med. Imaging*, pp. 1-3, 2020.

[18] P. Dhakan, A. Mandaliya, A. Limbachiya, H. N. Bhor, "Brain stroke detection using machine learning," *AIP Conf. Proc.*, vol. 3227, no. 1, p. 050002, 2023.

[19] N. Biswas, K. M. M. Uddin, S. T. Rikta, S. K. Dey, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Healthcare Analytics*, vol. 2, p. 100116, 2022.

[20] Kaggle, "Stroke Prediction Dataset," 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. Accessed: Feb. 10, 2025.