Heart Stroke Prediction Using Machine Learning Techniques

Kamrul Hasan [1], Ahmed Anan [2], MD. Arafat Islam [3], Munim Shahriar [5]
Mr. Faisal Muhammad Shah[5] and Mr. Md. Zahid Hossain [6]
[1,2,3,4,5,6]Department of Computer Science and Engineering,
Ahsanullah University of Science and Technology, Dhaka, Bangladesh

***Abstract*—Heart stroke poses significant global health challenges, necessitating timely intervention for risk mitigation. Leveraging machine learning (ML) techniques, this thesis explores predictive models for heart stroke using clinical data. A literature review underscores ML's role in enhancing accuracy and clinical decision-making. Algorithms like logistic regression, support vector machines, random forests, gradient boosting, and deep learning are scrutinized. Data collection and preprocessing involve demographics, medical history, lifestyle, and physiological parameters. Feature selection identifies key predictors for risk assessment. Comparative analysis, using metrics like accuracy and AUC-ROC, ensures model robustness. Emphasis on interpretability aids clinical integration for personalized intervention. This research contributes to advancing predictive analytics in cardiovascular healthcare, promising improved patient outcomes and reduced healthcare burden. Future directions and implementation challenges are discussed.**

***Index Terms*- Machine learning techniques, Logistic regression, Support vector machines, Random forests, gradient boosting, AUC-ROC**

I. INTRODUCTION

Introduction:

Heart stroke, also known as stroke or cerebrovascular accident (CVA), is a life-threatening medical emergency characterized by the sudden disruption of blood flow to the brain, leading to neurological deficits and potentially irreversible tissue damage. It is a leading cause of mortality and long-term disability worldwide, posing significant healthcare challenges and economic burdens on healthcare systems globally. According to the World Health Organization (WHO), stroke accounts for approximately 6.2 million deaths annually, with over 80% occurring in low- and middle-income countries.

Timely intervention and accurate prediction of heart strokes are paramount to mitigating their devastating consequences. The ability to identify individuals at high risk of experiencing a stroke enables healthcare professionals to implement preventive measures and intervene promptly, thereby reducing the likelihood of stroke occurrence and its associated morbidity and mortality.

In recent years, advancements in machine learning (ML) techniques have offered promising avenues for enhancing the prediction accuracy of various medical conditions, including cardiovascular diseases. ML algorithms, such as logistic regression, support vector machines, random forests, gradient boosting, and deep learning, have demonstrated efficacy in analyzing complex clinical data and identifying patterns that may be imperceptible to traditional statistical methods. By leveraging large-scale clinical datasets containing a wealth of patient information, ML models can extract valuable insights and facilitate personalized risk assessment for stroke prediction.

This thesis aims to investigate the application of ML techniques in predicting heart strokes using a clinical data approach. Specifically, it will focus on the utilization of relevant clinical data, encompassing patient demographics, medical history, lifestyle factors, and physiological parameters, to develop predictive models for identifying individuals at high risk of experiencing a stroke. The research will entail a comprehensive review of existing literature on stroke prediction, highlighting the role of ML techniques in augmenting prediction accuracy and informing clinical decision-making.

Furthermore, the thesis will delve into the methodology of collecting and preprocessing clinical data, emphasizing feature selection and engineering techniques to identify the most informative predictors for stroke risk assessment. Subsequently, various ML algorithms will be evaluated and compared in terms of their performance metrics, including accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC-ROC), to ascertain their effectiveness in stroke prediction.

Moreover, the interpretability and explainability of the developed ML models will be addressed to provide insights into the underlying factors contributing to stroke prediction. This will facilitate the seamless integration of predictive models into clinical practice, enabling healthcare professionals to tailor interventions and treatment strategies based on individual patient risk profiles.

Overall, this thesis seeks to contribute to the advancement of predictive analytics in cardiovascular healthcare, with the ultimate goal of improving patient outcomes and reducing the healthcare burden associated with heart strokes. By elucidating the potential of ML techniques in stroke prediction, this research endeavors to provide valuable insights and recommendations for future research directions and clinical implementation strategies.

II. LITERATURE REVIEW

Brain tumors represent an especially malignant form of cancer that can lead to substantial complications within the body. Consequently, identifying a brain tumor early and with precision can substantially enhance the likelihood of survival. However, distinguishing between various types of tumors poses a challenging task. Hence, creating an effective tumor representation through an optimization algorithm is essential for successful identification of brain tumors. This research explores ten existing methods for detecting

brain tumors, drawing inspiration from the limitations of each approach to develop a novel strategy for brain tumor detection.

The study conducted by G. Sasikala et al. [1] presents a comprehensive analysis of stroke prediction using machine learning techniques applied to electromyographic (EMG) data. The research focuses on two datasets: Congenital Heart Disease (CHD) and International Stroke Trial (IST). Various machine learning algorithms including linear regression, logistic regression, support vector machine (SVM), and reinforcement learning were employed for classification. The performance evaluation highlighted that the Reinforcement Learning model exhibited superior accuracy and computational efficiency compared to other classifiers, achieving 80.5% accuracy on the CHD dataset and 80.9% on the IST dataset. Notably, data standardization, particularly deviation standardization, was identified as a crucial factor contributing to improved performance metrics such as accuracy. Despite the promising results, the study acknowledges limitations such as the small size of the datasets and potential overfitting, especially in the case of the IST dataset due to its limited data availability. Overall, the findings provide valuable insights for clinical data-based diagnosis and underscore the potential for enhancing medical efficiency.

Rakshit et al. [2] presents a study on heart stroke prediction utilizing machine learning algorithms. The dataset utilized in the study, obtained from Kaggle, comprises 5110 instances with 12 attributes including age, gender, average glucose level, smoking status, BMI, etc., with the target variable being stroke. The authors conducted data preprocessing, visualization, and splitting into training and testing sets. Various classification algorithms such as Random Forest, Logistic Regression, KNN, Decision Tree, Naïve Bayes, and SVM were employed and their performances were evaluated using metrics like accuracy, precision, recall, and F-measure. The Decision Tree algorithm was identified as the most effective with an accuracy of 100%. However, limitations such as the small dataset size and potential biases or missing features could affect the generalization of the model. Furthermore, the study could benefit from a more extensive dataset and validation on diverse populations to enhance the robustness of the predictive model.

The research paper presented by Mohapatra et al. [3] presents an adaptive model utilizing machine learning algorithms for predicting heart diseases. The study addresses the critical need for accurate prediction and prevention of heart diseases, which remain a leading cause of mortality globally. By leveraging data analytics, machine learning, and artificial intelligence, the paper aims to provide optimal solutions for heart disease management. The proposed stacking model incorporates seven different machine learning algorithms, including Random Forest, Naïve Bayes, Linear Regression, Decision Tree, Adaboost, K Nearest Neighbor, and Gradient Boosting. Through experiments with an 80:20 training-testing ratio, the study evaluates the efficiency of these algorithms based on measures such as Precision, Recall, F Score, and Accuracy. The results demonstrate that Gradient Boosting outperforms other approaches, achieving an impressive accuracy of 94.67%. Additionally, the paper discusses the system architecture, experimental results, and the significance of adopting a streamlined approach to machine learning development. The integration of MLOps framework enhances the efficiency and reliability of the heart disease prediction model. Furthermore, the study compares its findings with previous research, highlighting the advancements and varying performance of different ML algorithms in heart disease prediction. The conclusion underscores the effectiveness of Gradient Boosting in predicting heart disease and emphasizes the importance of selecting appropriate classifiers for optimal results. Future research directions include establishing the robustness and generalizability of the model's outcomes, as well as enhancing interpretability for guiding decision-making in clinical settings. The paper contributes to the growing body of literature on heart disease prediction using machine learning techniques and offers valuable insights for researchers and practitioners in the field.

The study conducted by Madduri et al. [4] focuses on predicting heart strokes using machine learning algorithms, with an emphasis on the dataset, methodology, results, and limitations. The dataset, sourced from Kaggle, comprises 304 patient records containing various health indicators, including hypertension, cardiovascular disease, marital status, type of work, residence, average blood glucose levels, and body mass index (BMI). The methodology involves data preprocessing, splitting, training, and testing, followed by the implementation of machine learning algorithms, including Naïve Bayes, Decision Tree, Random Forest, and K-nearest neighbors (KNN). The study achieves high accuracy rates, with Random Forest outperforming other algorithms, achieving an accuracy of 97.69%. However, it is essential to note the limitations of the study, such as potential biases in the dataset, reliance on retrospective data, and the need for validation in diverse populations. Overall, the ensemble solution based on machine learning algorithms demonstrates promising results in predicting cardiovascular diseases, particularly heart strokes, but further research is necessary to address limitations and enhance generalizability.

The paper by Kamutam et al. [5] presents a comprehensive study on heart stroke prediction using machine learning techniques. Their research focuses on leveraging various data mining approaches, including KNN, Decision Tree, and Random Forest, to forecast the likelihood of heart stroke and categorize patient risk levels. They utilize the cardiac stroke dataset available on Kaggle, which comprises 12 attributes such as age, gender, hypertension, work type, residence type, heart disease, average glucose level, BMI, marital status, smoking status, and stroke history. The dataset is crucial in training and evaluating machine learning classifiers for heart stroke prediction. In their methodology, they divide the dataset into training and testing sets, with 80% used for training and 20% for testing. Performance metrics such as accuracy, precision, recall, and F-measure are employed to evaluate the effectiveness of the algorithms. The results indicate that the Random Forest algorithm outperforms KNN and Decision Tree with an accuracy of 99.17%. This high accuracy demonstrates the efficacy of the Random Forest approach in heart stroke prediction. However, it's essential to acknowledge limitations in the study, such as the size of the dataset and potential biases. Future research could involve utilizing larger datasets and

addressing these limitations to further enhance the accuracy and generalizability of the heart stroke prediction models.

The study conducted by Wiryaseputra (2017) et al. [6] investigates stroke prediction utilizing machine learning algorithms, focusing on classification techniques including Decision Tree, Random Forest, XGBoost, and Logistic Regression. The research employs a dataset comprising clinical features of 5110 observations with 11 attributes obtained from Kaggle. Data preprocessing involves steps such as filling null values, transforming string values into integers, checking and addressing class imbalance, and removing outliers using Z-Score. The performance of the models is evaluated using metrics such as accuracy, precision, recall, and F-1 score, with Random Forest exhibiting the best performance with an accuracy of 99.27%, as well as high precision, recall, and F-1 score. The study concludes that Random Forest is the most suitable algorithm for stroke prediction due to its robust performance. However, limitations such as the reliance on a single dataset and potential biases in data collection methods should be considered. Additionally, further research could explore the implementation of other machine learning algorithms and validate the findings on diverse datasets to enhance the generalizability of the results.

In their study, Maryam Poornajaf et al. [7] (2023) aimed to assess various machine learning algorithms for predicting heart disease, emphasizing the significance of feature selection and k-fold cross-validation in improving model performance. They compared classifiers such as KNN, Decision Trees (DT), and Random Forests (RF) on a heart disease dataset sourced from Kaggle, focusing on metrics like F1 score and area under the ROC curve. Notably, Random Forests exhibited the highest F1 score (92%) and AUC ROC (95%), indicating superior performance among the algorithms tested. However, they acknowledged variations in results compared to other studies, attributing this to differences in datasets. Other research by Sultana et al., Ali et al., Yadav et al., Shah et al., and others highlighted the effectiveness of different algorithms such as KStar, J48, SMO, Naïve Bayes, and SVM, with varying degrees of accuracy. While some achieved high accuracy rates, limitations in dataset size or selection were acknowledged. Additionally, methodologies varied, encompassing decision tree algorithms, neural networks, and ensemble methods. Despite advancements in techniques like deep learning and cloud-based systems for heart disease prediction, the study underscores the importance of robust dataset selection, methodological clarity, and continual improvement in machine learning approaches to enhance diagnostic accuracy and decision-making in heart disease management.

Saeed et al. [8] (2023) present a study focused on predicting cardiac diseases using artificial intelligence (AI) algorithms, particularly employing SelectKBest feature selection method. The dataset utilized in this research is crucial for understanding the effectiveness of the proposed approach; however, detailed information regarding its size, characteristics, and sources is lacking. The authors apply AI algorithms along with SelectKBest feature selection to predict cardiac diseases, although specific algorithms and techniques employed are not explicitly mentioned. The results section likely presents the outcomes of applying the proposed AI algorithms with SelectKBest on the cardiac disease dataset; however, detailed insights into achieved results, such as accuracy, sensitivity, specificity, or area under the curve (AUC) values, are missing. Despite its contributions, the study has several limitations, including the lack of transparency regarding the dataset and methodology, as well as the absence of a detailed discussion on the limitations of the proposed approach. Future research could address these limitations by providing more comprehensive descriptions of the dataset, methodology, and results, along with a critical analysis of the study's limitations and implications.

Chen et al. [9] (2022) conducted a comprehensive study aiming to predict the 90-day prognosis of patients with transient ischemic attack (TIA) and minor stroke using machine learning methods. The study utilized data from the Third China National Stroke Registry (CNSR-III), which included demographic characteristics, physiological data, medical history, and laboratory results, among other variables. Machine learning algorithms, including CatBoost, XGBoost, Gradient Boosting Decision Tree (GBDT), Random Forest (RF), and AdaBoost, were employed and compared with traditional Logistic regression models. The results indicated that machine learning models, particularly CatBoost and XGBoost, outperformed the Logistic model in predicting poor prognosis at 90 days, with CatBoost exhibiting the highest predictive performance. The study highlighted the importance of considering TIA and minor stroke patients' prognosis, as they constitute a significant proportion of stroke cases and often receive less attention compared to non-minor stroke cases in clinical practice. However, the study is not without limitations. While the machine learning models demonstrated superior predictive performance, further validation and external validation are warranted to assess their generalizability and applicability in different clinical settings. Additionally, the study's reliance on data from a single registry may limit the generalizability of the findings. Nevertheless, the study contributes valuable insights into leveraging machine learning techniques for prognostic prediction in TIA and minor stroke patients, potentially facilitating more accurate risk assessment and tailored interventions in clinical practice.

The study by Zhang et al. [10] aimed to establish a prediction model and assess risk factors for postoperative stroke in elderly patients, utilizing machine learning (ML) prediction models on data from the MIMIC-III and MIMIC-VI databases. Various ML techniques were applied, addressing category imbalance and missing values. Among seven modeling approaches, the XGB model demonstrated the highest AUC of 0.78, outperforming others, especially with data balancing. Hypertension, cancer, congestive heart failure, chronic pulmonary disease, and peripheral vascular disease were identified as top predictors, with hypertension being notably significant. ML, a mature method, offered advantages in predictive accuracy. Utilizing comprehensive clinical data, this model could effectively predict postoperative stroke risk in elderly patients, emphasizing the importance of hypertension history over laboratory results for prevention, regardless of gender. The study enhances stroke prediction, contributing to improved prognosis and prevention strategies, crucial for elderly surgical patients.

III. METHODOLOGY

"Stroke prediction dataset" is a binary classification problem with multiple numerical and categorical features. We first preprocess the data by filling in the null values and balancing the dataset as it was highly imbalanced. Only the feature BMI has a total number of missing values of 201. From the table of descriptive statistics of the dataset, we observe that the mean and median values of BMI are very close to each other. Hence, we fill the missing values with the mean values.

We drop the id column as it is just a unique identifier. Here, features are defined as categorical if the attribute has less than 6 unique elements else it is a discrete feature. Then we label encode the data into categorical text data features. The dataset was highly unbalanced in favor of no stroke with a ratio of 19:1 of No Stroke : Stroke. As the bias is so high towards no stroke, predictions cannot be trusted. So, we need to carry out a feature engineering process for balancing the dataset. we will use SMOTE(Synthetic Minority Over-sampling Technique) analysis and feed the balanced dataset to the machine learning models.

We can do this balancing in two ways oversampling and undersampling but for the best performances combination of undersampling and oversampling may give us a good result. First, we will do undersample for the data of majority samples and it is followed by oversampling for the data of minority samples.

Sampling Strategy : (Samples of Minority Class) / (Samples of Majority Class) Majority class: no stroke: 4861 samples and the data of Minority class: stroke: 249 samples.

Undersampling:

Sampling Strategy = 0.1
0.1 = (249) / Majority Class Samples
After undersampling,
Majority Class: No Stroke: 2490 samples
Minority Class : Stroke: 249 samples

Oversampling:

Sampling Strategy = 1
1 = (Minority Class Samples) / 2490
After oversampling,
Majority Class Data: No Stroke: 2490 samples
Minority Class Data: Stroke: 2490 samples

Final Class Samples:

Majority Class: No Stroke: 2490 samples
Minority Class: Stroke: 2490 samples We made the data by reducing the data of majority samples then increasing the data of minority group to the majority group.

After that, we used the multiple-feature selection technique, Mutual Information Test, Chi-Squared Test, and finally ANOVA Test.

Mutual Information score of stroke with categorical features displays very low scores, According to the scores, none of the features should be selected for modeling.

For the Chi-Squared Test, We will reject features with scores less than 20. Hence, we will not use: smoking status, heart disease hypertension.

From the ANOVA Scores, we ignored the features with values less than 20. Hence, we rejected BMI for modeling. We ready the datasets for data scaling by dropping these features. Then we standardized and normalized the data using standardScalar and MinMaxScalar respectively.

We split the data into 85 - 15 train-test groups.

We used four classifier models xgboost, Random Forest, KNN, and Decision Tree with the ROC Curve to classify.

A Feature extraction technique PCA was applied to reduce the dimensionality to 2 dimensions. But the reduced dimension result was not used to train model.

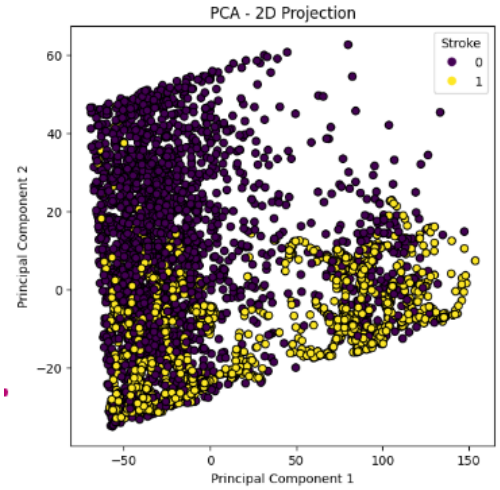

Fig. 1. model

IV. DATASET

Features	Details
Gender	Female: 59% Male: 41% Other: 0%
Hypertension	No: 4612 Yes: 498
Heart Disease	No: 4834 Yes: 276
Ever Married	True (3353) : 66% False (1757) : 34%
Work Type	Private: 57% Self-employed: 16% Other(1366): 27%
Residence Type	Urban: 51% Rural: 49%
BMI	N/A :4% 28.7: 1% Other (4868): 95%

Fig. 2. Dataset

The dataset for predicting heart strokes comprises various demographic and health-related features. Gender distribution shows a slight majority of females at 59%, followed by males at 41%, with no data for other genders. Hypertension is prevalent among a subset, with 498 individuals (10.2%) reporting a positive diagnosis

compared to 4612 (89.8%) without hypertension. Similarly, 276 individuals (5.4%) have a history of heart disease, while the majority, 4834 (94.6%), have not been diagnosed. Ever-married individuals constitute 66% of the dataset,

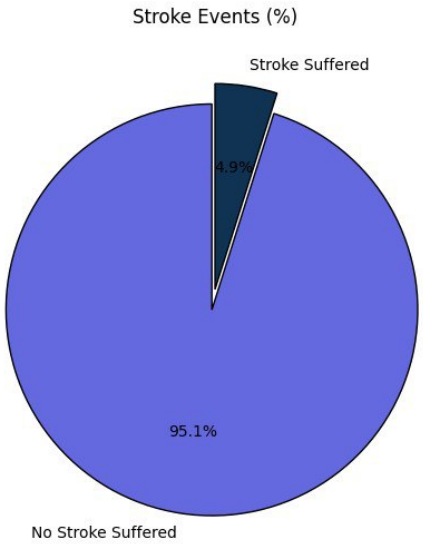

Fig. 3. Target Class Ratio

while 34% report being unmarried. Work types vary, with the majority being in private employment (57%), followed by self-employed individuals (16%), and others (27%). Urban and rural residence types are almost evenly split at 51% and 49%, respectively.

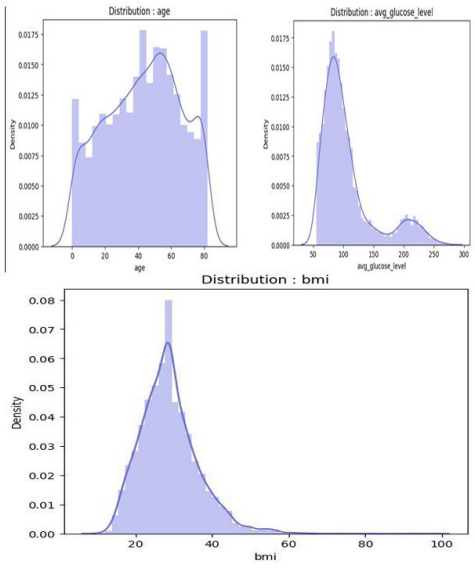

Fig. 4. Discrete Feature Distribution

BMI data is largely available (95%), with a mean value of 28.7, while 4% of the data is missing.

BMI data is largely available within the dataset, encompassing 95% of the total data entries, indicating a comprehensive coverage of this important anthropometric measure. The mean BMI value of 28.7 suggests a noteworthy representation of individuals across a spectrum of body mass indices, ranging from underweight to obese

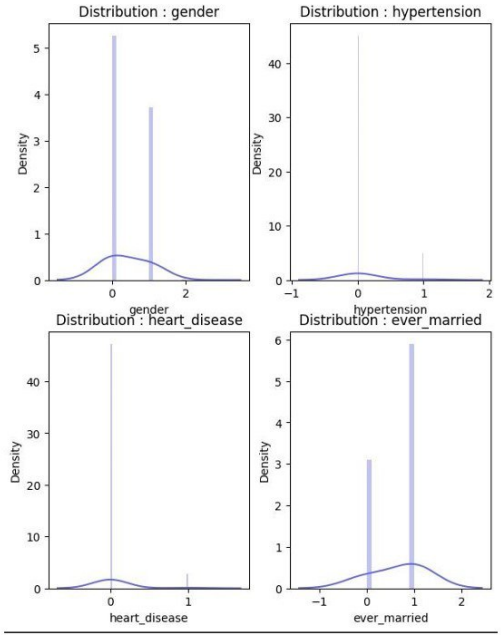

Fig. 5. Catagorical Feature Distribution

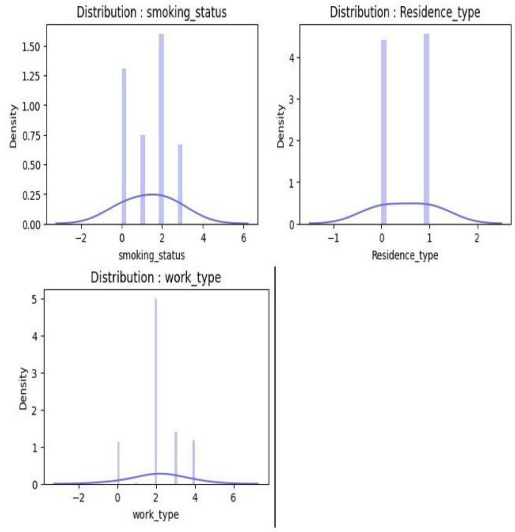

Fig. 6. Catagorical Feature Distribution

categories. This diverse distribution of BMI values reflects the heterogeneous nature of the population under study, capturing a wide range of body compositions and associated health risks. However, it is noteworthy that approximately 4% of the data is missing for BMI measurements. While this missing data represents a relatively small proportion of the overall dataset, its absence could potentially introduce biases or limitations in the analysis and modeling process. Addressing the missing data through imputation techniques or sensitivity analyses is crucial to ensure the robustness. Overall, the inclusion of BMI data, despite some missing values, enriches the dataset with a crucial anthropometric parameter that is strongly associated with cardiovascular health outcomes. This comprehensive array of features, including BMI, provides a solid foundation for the develop-

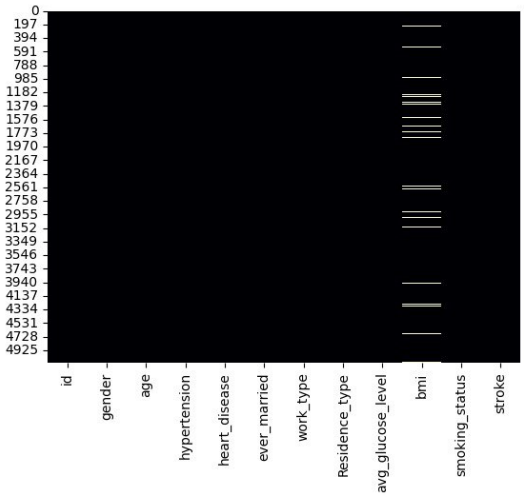

Fig. 7. Missing Data

ment of predictive models aimed at identifying individuals at elevated risk of heart stroke and facilitating targeted preventive interventions and healthcare management strategies.

V. EXPERIMENTAL RESULT

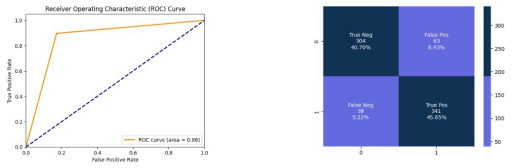

Fig. 8. Xgboost classifier

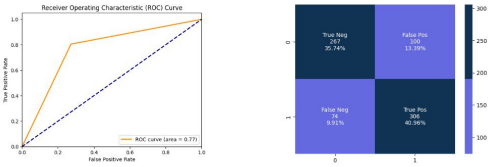

Fig. 9. Random Forest classifier

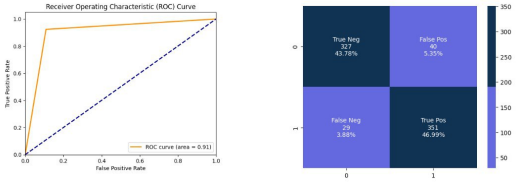

Fig. 10. KNN classifier

Xgboost classifier provided us with an accuracy of 0.84. Precision, recall, and f1-score of 0.82, 0.88, 0.85 respectively. And an acceptable ROC AUC Score of 83.72%.

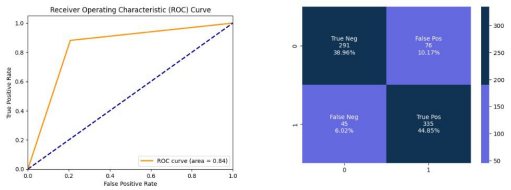

Fig. 11. Decision Tree classifier

Random Forest classifier provided us with an accuracy of 0.91. Precision, recall, and f1-score of 0.90, 0.92, and 0.91 respectively. And a very good ROC AUC Score of 90.73%. KNN classifier provided us with a not-so-good accuracy of 0.77. Precision, recall, and f1-score of 0.75, 0.81 and 0.78 respectively. And an ROC AUC Score of 76.64%.

TABLE I
XGBOOST CLASSIFIER RESULT

Precision	Recall	F1 score	Accuracy
0.82	**0.88**	0.85	**0.82**

TABLE II
RANDOM FOREST CLASSIFIER RESULT

Precision	Recall	F1 score	Accuracy
0.90	**0.92**	0.91	**0.91**

TABLE III
KNN CLASSIFIER RESULT

Precision	Recall	F1 score	Accuracy
0.75	**0.81**	0.78	**0.77**

Decision Tree classifier provided us with a not-so-good accuracy of 0.86. Precision, recall, and f1-score of 0.84, 0.90 and 0.87 respectively. And an ROC AUC Score of 86.29%.

VI. RESULT ANALYSIS

XGBoost Classifier: The XGBoost classifier demonstrates a good overall performance with an accuracy of 0.84. It achieves a balanced trade-off between precision and recall, indicating its ability to correctly classify positive instances (individuals at risk of heart stroke) while minimizing false positives. The F1-score of 0.85 suggests robustness in terms of both precision and recall. The ROC AUC score of 83.72% indicates acceptable discrimination ability, with a higher true positive rate and a lower false positive rate compared to random guessing.

Random Forest Classifier: The Random Forest classifier exhibits excellent performance with the highest accuracy of 0.91 among all classifiers. It demonstrates high precision and recall values of 0.90 and 0.92, respectively, indicating its ability to effectively identify individuals at risk of heart stroke while minimizing false positives and false negatives. The F1-score of 0.91 signifies a balanced performance in terms of precision and recall. The ROC AUC score of

90.73% reflects very good discrimination ability, with a high true positive rate and a low false positive rate.
KNN Classifier: The KNN classifier demonstrates moderate performance with an accuracy of 0.77. It achieves relatively lower precision and higher recall values compared to the Random Forest classifier, indicating a higher false positive rate but a lower false negative rate. The F1-score of 0.78 suggests a moderate balance between precision and recall. The ROC AUC score of 76.64% indicates acceptable discrimination ability, although it is lower compared to the Random Forest classifier.
Decision Tree Classifier: The Decision Tree classifier demonstrates good overall performance with an accuracy of 0.86. It achieves a balanced trade-off between precision and recall, similar to the XGBoost classifier. The F1-score of 0.87 indicates robustness in terms of both precision and recall. The ROC AUC score of 86.29% reflects good discrimination ability, comparable to the XGBoost classifier.
In summary, the Random Forest classifier exhibits the highest overall performance with the highest accuracy, precision, recall, F1-score, and ROC AUC score among all classifiers, followed by the Decision Tree classifier. The XGBoost classifier and Decision Tree classifier also demonstrate good performance, while the KNN classifier exhibits moderate performance compared to the other classifiers.

VII. LIMITATIONS AND FUTURE WORKS

As a nobel work, there are a lot process that can be taken up by using this data. So we have some future thoughts to add in this work.

Limitations:

The effectiveness of machine learning models for heart stroke prediction heavily relies on the availability and quality of clinical data. Limitations in data collection, such as missing or incomplete records, may impact the performance of the predictive models.
The generalizability of machine learning models developed for heart stroke prediction to diverse populations and healthcare settings may be limited. External validation on independent datasets from different demographics and geographic regions is essential to assess the robustness and applicability of the models.

Future Works:

Future research can explore the integration of advanced machine learning techniques, such as ensemble methods, deep learning, and transfer learning, to enhance the predictive accuracy and robustness of heart stroke prediction models.
Incorporating multimodal data, including genetic information, wearable device data, and imaging studies, into predictive models can provide a more comprehensive understanding of heart stroke risk factors and improve prediction accuracy.
Future research should focus on the practical implementation and validation of machine learning models for heart stroke prediction in real clinical settings. Collaborations with healthcare institutions and stakeholders are essential for integrating predictive models into routine clinical practice and evaluating their impact on patient outcomes and healthcare delivery.

VIII. CONCLUSION

This survey underscores the burgeoning impact of machine learning in revolutionizing healthcare practices. It also reveals a remarkable spectrum of applications, from predicting heart stroke to enhancing diagnostic accuracy in various medical domains. Despite the promising strides, challenges such as dataset limitations, algorithmic biases, and the imperative need for interpretability persists. As the field continues to evolve, future research should prioritize addressing these challenges to ensure the responsible and effective integration of machine learning into clinical settings. The addition of advanced analytics, artificial intelligence, healthcare necessitates ongoing interdisciplinary collaboration and ethical considerations. By embracing these principles, the healthcare community can face the full potential of machine learning by fostering improved patient outcomes, personalized treatment strategies, and a more resilient healthcare ecosystem.

REFERENCES

[1] Gnanasekaran, Sasikala Roja, G Radhika, D. (2021). Prediction Of Heart Stroke Diseases Using Machine Learning Technique Based Electromyographic Data. 12. 4424-4431.
[2] Rakshit, Tanisha Shrestha, Aayush. (2021). Comparative Analysis and Implementation of Heart Stroke Prediction using Various Machine Learning Techniques. International Journal of Engineering and Technical Research. 10. 886-890.
[3] Mohapatra, Subasish Mishra, Indrani Mohanty, Subhadarshini. (2023). Stacking Model for Heart Stroke Prediction using Machine Learning Techniques. EAI Endorsed Transactions on Pervasive Health and Technology. 9. 10.4108/eetpht.9.4057.
[4] Madduri, Tarun Kumari J, Vimala Ayinapuru, Jaswitha Kodali, Nivas Prattipati, Vamsi. (2023). Heart Stroke Prediction Using Different Machine Learning Algorithms. 10.1007/978-981-99-5881-8-23.
[5] Vinay, Kamutam Yashwant, Marneni Mulla, Prashanth Dharam, Akhil. (2023). Heart Stroke Prediction using Machine Learning.
[6] Wiryaseputra, Michael. (2022). Stroke Prediction Using Machine Learning Classification Algorithm.
[7] Poornajaf, Maryam. (2023). Analysis of Accuracy Metric of Machine Learning Algorithms in Predicting Heart Disease. Frontiers in Health Informatics. 12. 135. 10.30699/fhi.v12i0.402.
[8] Hama Saeed, Mariwan. (2023). Cardiac disease prediction using AI algorithms with SelectKBest. Medical Biological Engineering Computing. 10.1007/s11517-023-02918-8.
[9] Chen, Si-Ding You, Jia Yang, Xiao-Meng Gu, Hong-Qiu Huang, Huan Feng, Jian-Feng Jiang, Yong Wang, Yong-Jun. (2022). Machine learning is an effective method to predict the 90-day prognosis of patients with transient ischemic attack and stroke. BMC Medical Research Methodology. 22. 10.1186.
[10] Zhang, Xiao Fei, Ningbo Zhang, Xinxin Wang, Qun Fang, Zongping. (2022). Machine Learning Prediction Models for Postoperative Stroke: Analyses of the MIMIC Database. Frontiers in Aging Neuroscience. 14. 897611. 10.3389.