AHSANULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF CSE

TITLE OF THE PROJECT

"Heart Stroke Prediction Using Machine Learning Techniques"

TEAM DETAILS

MD. Kamrul Hasan
20200104009

Ahmed Anan
20200104017

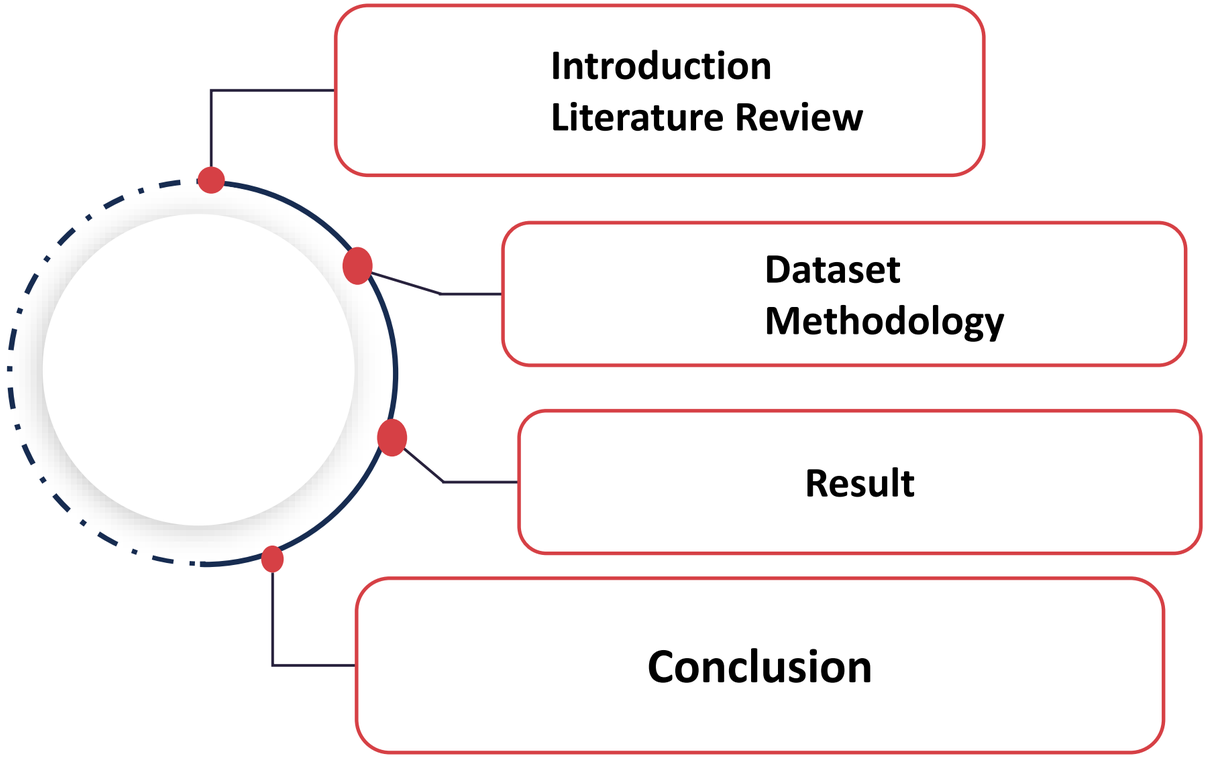
Md. Arafat Islam
20200104010

Munim Shahriar
20200104020

Abstract

Abstract—Heart stroke poses significant global health challenges, necessitating timely intervention for risk mitigation. Leveraging machine learning (ML) techniques, this work explores predictive models for heart stroke using clinical data.

Overview


Introduction
Literature Review
Dataset
Methodology
Result
Conclusion

PROJECT

INTRODUCTION

- Heart strokes, a significant health threat globally, require timely intervention and accurate prediction to mitigate their devastating effects.
- This thesis investigates ML application in predicting heart strokes using clinical data, focusing on patient demographics and medical history.
- Aims to advance predictive analytics in cardiovascular healthcare, improving patient outcomes and reducing healthcare burdens associated with heart strokes.
- Machine learning (ML) techniques offer promising avenues for enhancing stroke prediction accuracy by analyzing complex clinical data.
- It emphasizes feature selection and ML algorithm evaluation to develop effective stroke prediction models.

LITERATURE REVIEW

"G. Sasikala et al. demonstrates the effectiveness of reinforcement learning in predicting strokes using EMG data, emphasizing the importance of data standardization ." [1]

"Rakshit et al. conducted heart stroke prediction using machine learning, finding Decision Tree as the most effective algorithm, but noted limitations due to dataset size and potential biases" [2]

"Mohapatra et al. present an adaptive model utilizing Gradient Boosting and six other machine learning algorithms for highly accurate heart disease prediction" [3]

"Madduri et al. predict heart strokes using machine learning, achieving high accuracy with Random Forest (97.69%), but note limitations in dataset biases and retrospective data reliance." [4]

LITERATURE REVIEW

"Kamutam et al. employ machine learning techniques on a cardiac stroke dataset to predict heart stroke risk, finding Random Forest outperforms KNN and Decision Tree with 99.17% accuracy." [5]

" Wiryaseputra et al. find Random Forest outperforms other algorithms for stroke prediction, citing high accuracy and recommend further validation and exploration of alternative methods" [6]

"Maryam Poornajaf et al. compared machine learning algorithms for heart disease prediction, finding Random Forests with the highest F1 score and AUC ROC, emphasizing the significance of robust dataset selection and methodological clarity in improving diagnostic accuracy " [7]

[8]

"Adifa Vuyia et al. utilizies Extreme Gradient Boosting achieving the highest accuracy at 91.9%, underscoring the importance of accurate prediction."

METHODOLOGY

- **Preprocessing**

  →Only the feature **BMI** has a total number of missing values of 201.

  →We drop the **id** column as it is just a unique identifier.

  →features are defined as categorical if the attribute has less than 6 unique elements else it is a discrete feature

  →The dataset was highly unbalanced in favor of **no stroke** with a ratio of 19:1 of **No Stroke : Stroke.**

  →feature engineering process **SMOTE**(Synthetic Minority Over-sampling Technique) was applied to balance the dataset

  →**Sampling Strategy** : **(Samples of Minority Class) / (Samples of Majority Class)**

METHODOLOGY

UnderSampling: Trim down the majority class samples

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

OverSampling: Increase the minority class samples

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

Final Class Samples:

Majority Class: No Stroke: 2490 samples
Minority Class: Stroke: 2490 samples

METHODOLOGY

Feature selection

Used feature selection technique:

→**Mutual Information Test, Chi-Squared Test**, **ANOVA**

→**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.

→ According to **Chi-Squared** Test**,** features with scores less than 20 were rejected. Features: **smoking_status**, **heart_disease** & **hypertension**.

→According to **ANOVA** Test**,** features with scores less than 20 were rejected. Features: **BMI**.

Feature extraction:

→ Feature extraction technique **PCA** was applied to reduce the dimensionality to 2 dimensions.

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%

Features	Details
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%

RESULT

PERFORMANCE

Model: Xgboost classifier

Precision	Recall	F1 Score	Accuracy
0.82	0.88	0.85	0.84

TP: 44.85%	FP: 10.17%
TN: 38.96%	FN: 6.02%

Model: Random Forest classifier

Precision	Recall	F1 Score	Accuracy
0.90	0.92	0.91	0.91

TP: 46.99%	FP: 5.35%
TN: 43.78%	FN: 3.88%

RESULT

PERFORMANCE

Model:KNN classifier

Precision	Recall	F1 Score	Accuracy
0.75	0.81	0.78	0.77

TP: 40.96%	FP: 13.39%
TN: 35.74%	FN: 9.91%

Model: Decision Tree classifier

Precision	Recall	F1 Score	Accuracy
0.84	0.90	0.87	0.86

TP: 46.65%	FP: 8.43%
TN: 40.70%	FN: 5.22%

Limitation & Future Work

- The effectiveness of machine learning models for heart stroke prediction relies on the availability and quality of clinical data, with limitations such as missing or incomplete records potentially impacting model performance.

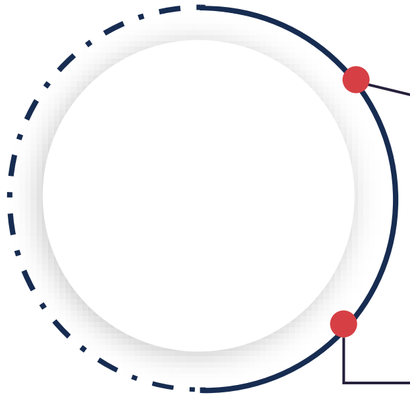
- Generalizability of developed machine learning models to diverse populations and healthcare settings may be limited, necessitating external validation on independent datasets from various demographics and geographic regions.

- Future research could explore integrating advanced machine learning techniques like ensemble methods, deep learning, and transfer learning to enhance predictive accuracy and robustness of heart stroke prediction models.

CONCLUTION

- The survey highlights machine learning's growing impact on healthcare, showcasing diverse applications from predicting heart strokes to enhancing diagnostic accuracy across medical domains.

- Despite promising advancements, challenges such as dataset limitations, algorithmic biases, and the need for interpretability persist, necessitating future research to address these issues for responsible integration into clinical settings.



THANKS EVERYONE

Any Questions?

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.

REFERENCES

[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] Islam, Taminul & Vuyia, Adifa & Hasan, Mahadi & Rana, Md.Masum. (2023). Cardiovascular Disease Prediction Using Machine Learning Approaches. 10.1109/CISES58720.2023.10183490.