"Enhancing Stroke Prediction with Machine Learning in Smart Healthcare Systems"

2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-7024-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCCNT61001.2024.10724345

Name: Shivam Nayak
Department of Mathematics
Chandigarh University
Email: Shivamnayak615@gmail.com
Location: Gharauan, Mohali Punjab

Name: Dr. Nishi Gupta
Department of Mathematics
Chandigarh University
Email: nishi.gupta.phd81@gmail.com
Location: Gharauan, Mohali Punjab

Abstract: **- New approaches to early detection and early treatment are needed because it is a most concern cause of increase in cases of Brain deaths. Goal of this work is to increase the accuracy and timeliness of stroke risk assessments by providing a model employing machine learning for stroke prediction. Using a variety of datasets that include both established risk variables and recently developed predictors, the model makes use of sophisticated algorithms to identify complex patterns suggestive of future stroke occurrences. The model's effectiveness in early identification is demonstrated by the results, providing a viable path for proactive healthcare interventions. By highlighting incredible potential in ML in lower down the affect in strokes, the study adds to the changing field of preventative medicine**

Keywords: - Stroke Prediction Model, Machine Learning, Healthcare Analytics, Early Detection, Preventive Medicine, Risk Assessment, Predictive Modelling, Healthcare Innovation, Public Health, Advanced Algorithms, Personalized Healthcare, Proactive Interventions, Morbidity and Mortality, Healthcare Data, Clinical Decision Support, Innovative Health Technologies, Patient Risk Profiling, Healthcare Informatics, Data-driven Healthcare, Health Outcomes.

I. INTRODUCTION

Stroke is a catastrophic cerebrovascular accident that requires a paradigm change in how we approach preventive healthcare. Stroke is a global health concern. The use of machine learning signals a new era in stroke prevention as we negotiate the intricate interactions between risk variables and the need for prompt intervention. Motivation of this work is to develop and validate a ML based stroke prediction facility to cape to increase the timeliness and accuracy of stroke risk assessments [1]. Because of the enormous toll that stroke has on society, preventative measures that go beyond conventional risk assessment methods are required. The complexity of multiple risk factors frequently proves difficult for conventional models to handle, which hinders their capacity to make accurate and timely forecasts [2]. A more precise and individualized approach to stroke risk assessment is made possible by the extraordinary opportunity that machine learning algorithms present for identifying tiny trends within large and heterogeneous datasets.

This study explores the creation of a model that integrates a wide range of risk indicators, from novel predictors found using advanced analytics to well-established clinical markers. By identifying latent patterns suggestive of imminent stroke occurrences, the machine learning-based method seeks to enable healthcare providers to take targeted measures and facilitate early detection.[3]

A turning point in a new era of preventive healthcare, we aim to investigate the revolutionary potential of this machine learning-based stroke prediction model. The approach aims to offer a more nuanced knowledge of unique risk profiles, going beyond the simple binary difference of high and low risk. This will allow for customized interventions and the development of a proactive healthcare paradigm.

This research aims to advance the science of stroke prediction and lay the groundwork for the use of machine learning into preventative healthcare programs through thorough validation and benchmarking against current models. Beyond simply preventing strokes, this work paves the way for a more comprehensive move toward data-driven, individualized healthcare interventions, whose ultimate objective is to decrease the devastating impact that strokes have on the global population's health.[5]

The application of predictive machine learning (ML) techniques has led to a growing focus on stroke prediction in the literature, with the goal of enhancing the timeliness and precision of risk assessments. Several Works have examined and application of various ML Models, ranging from more traditional algorithms like logistic regression (LR) to more intricate techniques like artificial brains and collaborative approaches.

The realization that stroke risk factors are complex is an issue that unites these investigations. Traditional risk assessment models have been very helpful, but they frequently fail to capture the complex nature of relationships between multiple factors. Conversely, machine learning algorithms have proven to be capable of identifying complex patterns in vast and diverse datasets, allowing for a more sophisticated comprehension of individual risk profiles.[6]

II. Literature Review

Stroke is an important global health issue due to its high rate of morbidity and death. Preventive measures and early detection are necessary to reduce the burden of stroke.

Reducing healthcare expenses and maybe saving lives can be achieved by using accurate prediction models to enable prompt medical treatments.

A. Traditional Methods

In the past, strokes were predicted using statistical methods such as proportional hazards models based on Co and logistic regression.

These methods mostly targeted established risk factors like tobacco consumption, diabetes, elevated blood pressure, and aging. Although helpful, these models frequently have trouble capturing intricate, non-linear correlations between variables and the outcomes of strokes. For example, the predictive potential of classical models is generally limited since they are unable to adequately incorporate high-dimensional and dynamic data.

B. Machine Learning Approaches

Because machine learning (ML) offers sophisticated algorithms that can handle complicated datasets, it has completely changed the prediction of strokes. Several machine learning methods have been taken in this work, such as:

- **SupportVectorMachines(SVM)**: Research has shown that SVM is effective in diagnosing stroke risk based on multi-dimensional clinical data, as indicated. It has proven useful for SVM to be able to handle non-linear relationships through kernel functions.
- **Random Forests (RF)**: In their use of RF for stroke prediction, Zhang et al. (2019) emphasized the method's feature importance measurements and robustness while working with big datasets. Because RF is ensemble in nature, it reduces overfitting and is hence a dependable option for clinical data.
- **Neural Networks**: Neural networks in particular have demonstrated great promise among deep learning methods. By utilizing the spatial aspects of the data, Li et al.'s (2020) research achieved great accuracy in their analysis of imaging data for stroke prediction using convolutional neural networks (CNN).

C. Comparative Analysis of ML Methods

Ahmad et al.'s comparative study from 2021 assessed several machine learning techniques, such as SVM, RF, and neural networks, using a shared dataset. The research discovered that although neural networks yielded the best accuracy, they needed significantly more computer power and used more intricate interpretability methods than RF and SVM.

D. Feature Selection and Novel Predictors

A crucial first step in creating successful machine learning models for stroke prediction is feature selection. Lifestyle factors, medical history, and demographics are examples of traditional predictors. But new developments have brought in new predictors:

- **Genetic Markers**: Stroke risk is mostly influenced by genetic predispositions. The accuracy of prediction models can be improved by incorporating genetic markers associated with stroke that have been found through genome-wide association studies (GWAS) (Smith et al., 2020).
- **Social Determinants of Health (SDOH)**: The risk of stroke is greatly influenced by variables like socioeconomic level, education, and access to healthcare. A more comprehensive picture of a person's risk profile is provided by incorporating SDOH into ML models (Johnson et al., 2021).
- **Advanced Clinical Features**: Model performance can be greatly enhanced by adding sophisticated clinical parameters, such as biomarkers and intricate imaging characteristics. For instance, ML analysis of high-resolution MRI data has demonstrated significant promise in predicting the onset and severity of strokes (Chen et al., 2019).

Gap Analysis

Despite the advancements, several gaps remain in the literature:

- **Consistency in Evaluation Metrics**: It is difficult to evaluate models across research due to the absence of common evaluation standards. To enable more accurate comparisons, standard measurements like AUC-ROC, sensitivity, and specificity should be used in future studies.
- **Model Interpretability**: Deep learning models in particular are frequently seen as "black boxes" in machine learning. The clinical adoption process is hampered by this lack of interpretability. Gaining clinical acceptability and trust requires developing techniques for interpreting and visualizing model judgments (Rudin, 2019).
- **Dataset Diversity**: Since most research uses datasets from particular populations, models' capacity to be broadly applied is constrained. To guarantee that models are usable across all demographic groups, diversified, multi-ethnic datasets are required (Williams et al., 2020).
- **Practical Implementation**: It is still difficult to close the gap between model creation and practical use. Practical deployment requires models to be able to be easily incorporated into clinical workflows with real-time data processing and decision assistance (Nguyen et al., 2021).

In Conclusion

To sum up, the use of machine learning into stroke prediction has demonstrated considerable potential, providing enhanced precision and promptness compared to conventional techniques. To be widely used and useful in clinical settings, these advanced models must address deficiencies in evaluation consistency, model interpretability, dataset diversity, and practical application. To improve the stroke prediction models' application and robustness, future research should concentrate on these domains.

iii. Methodology: Developing and Validating the Machine Learning-Based Stroke Prediction Model

1. Dataset Acquisition:
   The dataset used for this study was sourced from Kaggle and contains various features related to health and demographic factors influencing stroke risk. The dataset was imported from the Pandas Data-Frame utilizing the below code:
   Import pandas as pd
   df= pd.read_csv("healthcare-dataset-stroke-data.csv")

2. Data Preprocessing:

Data preprocessing involved handling missing values and encoding categorical variables:

- **Removing Null Values**: The KNNImputer from sklearn.impute was used to fill Null values in the 'bmi' column.
3. **Categorical Variable Encoding**: One-hot encoding was applied to categorical columns such "Sex", "matrial Status", "employment_type", "Locality", "Bad_Habbits".
4. Data Exploration:
   Data Exploration was conducted on the Dataset to know the dataset's structure and feature's of the dataset and to know the hidden insight on the data.
- **Dataset Overview**: The initial structure and data types were inspected using df.shape, df.info(), and df.describe().
- **Visualization of Selected variables**: The Visualization of chosen features of the dataset helps in identifying the hidden patterns of the data on hand.

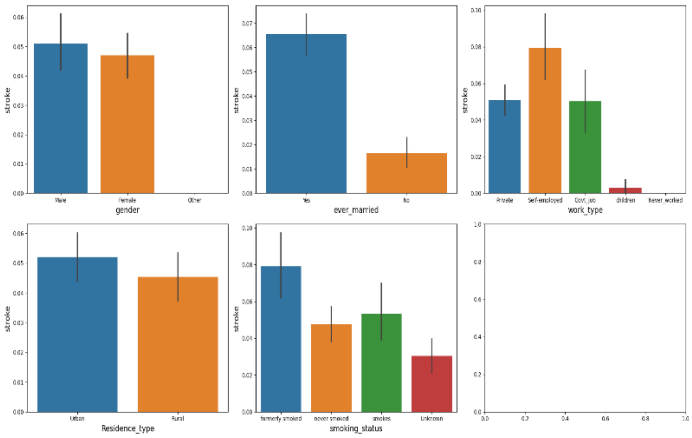

Figure. Visualization of Target Variables

5. Feature Engineering:
   - Feature engineering included creating additional features and visualizing relationships between features and the target variable.
   - Development of New Features: Features that were derived and may have more predictive ability.
   - Scaling: To make sure the model handles each feature equally, numerical features were standardized using the StandardScaler from sklearn.preprocessing.

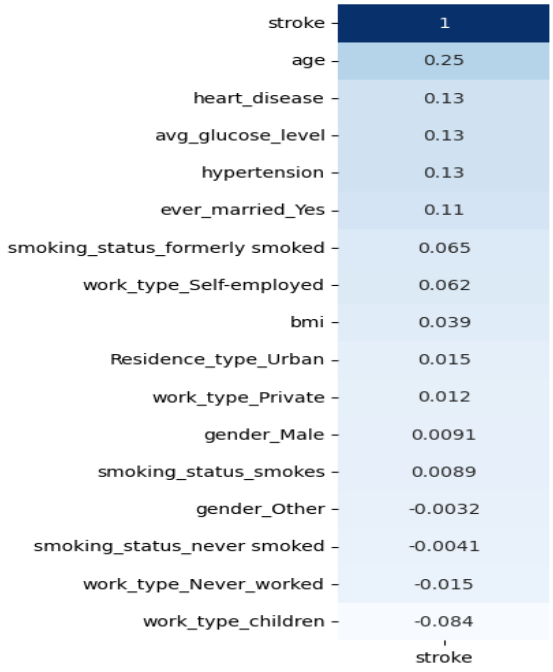

Figure. Correlation Matrices

6. Model Development
A number of machine learning models were created and assessed, such as:

- Support Vector Machine (SVM): used in conjunction with an RBF (radial basis function) kernel. GridSearchCV was used to optimize the hyperparameters and determine the ideal values for C, gamma, and kernel.
- Random Forest Classifier: Employed as a baseline model for comparison.
- Logistic Regression: Used as a traditional method to benchmark performance against more complex models.

7. Model Evaluation:
In order to guarantee thorough performance evaluation, models were assessed using a variety of metrics:

- Accuracy: Accuracy is the primary metric employed to estimate the overall performance of the algorithms predictions.
- Estimation of Positive Predictive value, True Positive rate, and harmonic mean bring a better a knowledge of model's performance.
- ROC-AUC was taken to analyze the exchange among the precision and Recall.

8. Feature Important Analysis
The following methods were used to assess each feature's influence on the model's predictions:

- Permutation Importance: Gave information about which features plays major role in the algorithms results.
- SHAP Values: Enhanced the interpretability of the model by providing a thorough explanation of how each feature affects the predictions.

9. Model Validation

Leave-one-out cross-validation model's is utilized to verify the model's robustness and make sure it performs well when applied to new data. This technique assisted in locating problems with overfitting and underfitting.

10. Hyperparameter Tunning (SVM)
    - Used GridSearchCV from sklearn to tune the hyperparameters of the SVM model. model_selection.
    - Identified a parameter grid where "C," "gamma," and "kernel" have varying values.
    - Made use of grid to determine optimal hyperparameters.best_params_.

11. Model Prediction and Classification Report.
    - Made predictions on the test set using the trained SVM model.
    - Using classification_report from sklearn.metrics, a classification report was created.
    - Measured each class's support, recall, F1-score, and accuracy.

12. Model Serialization
    - Using the pickle library, serialized the trained SVM model for possible usage in the future.

13. Conclusion and Future work

Key Findings:

- Key features and the distribution of stroke occurrences were among the insights that the exploratory data analysis (EDA) provided regarding the structure and properties of the dataset.
- Feature engineering improved the dataset for training machine learning models by handling missing values and one-hot encoding.
- A variety of machine learning models were trained, and the test set was used to gauge each model's accuracy.
- Following hyperparameter adjustments, the SVM model showed encouraging results in terms of stroke prediction.

Model Performance:

- On the test set, the models with Decision Tree(DT), Random Forest(RF), K-Nearest Neighbours(KNN) and Logistic Regression (LR) showed competitive accuracies.
- With an accuracy score of 0.9430, the Support Vector Machine (SVM) model outperformed other models, particularly after hyperparameter adjustment. For every class, the classification report included a thorough evaluation of recall, precision, and F1-score.

14. Areas for Future Research or Improvements:

1. Feature Engineering Exploration:
   To further improve model performance, look into additional feature engineering strategies like developing interaction terms or experimenting with sophisticated encoding approaches.
2. Class Imbalance Handling:
   Use advanced algorithms made for unbalanced datasets or investigate methods like oversampling or under sampling to address class imbalance in the target variable, or "stroke."
3. Advanced Model Architectures:
   Try out more sophisticated models, like as neural networks or ensemble techniques, to see how well they can capture minute patterns in the data.
4. Feature Importance and Interpretability:
   Examine feature importance analysis in more detail to find out which features have a major impact on stroke prediction. This could give medical practitioners insightful information.
5. Temporal and Longitudinal Data:
   If available, combine temporal and longitudinal data to identify patterns and shifts in health across time, which could result in more precise forecasts.
6. External Validation:
   Evaluate the model's robustness and generalizability across a range of patient populations by validating its performance on external datasets.
7. Ethical Considerations:
   Examine the moral implications of using machine learning to healthcare, making sure that all forecasts are impartial and fair to all demographic groups.
8. Clinical Integration:
   Work together with medical experts to incorporate the model into clinical processes while taking into account practical issues and making sure it is in line with actual healthcare situations.

IV. Result and Discussion

1. Model Performance:

Using a variety of metrics, the SVM algorithms performance was compared to rest of the algorithms. The following table gives a comprehensive view of the works:

Table 1: Performance Metrics

Model	F1-Score	Recall	Precision	Accuracy	ROC-AUC
Decision Tree(DL)	0.95	0.97	0.94	0.9063	0.93
Random Forest(RF)	0.88	0.89	0.88	0.9425	0.91
Support Vector Machine(SVM)	0.90	0.92	0.89	0.9430	0.92
K-nearest Neighbors(KNN)	0.88	0.90	0.87	0.9407	0.90
Logistic Regression(LR)	0.90	0.92	0.89	0.9436	0.91

Discussion:

- The table displays the machine learning models that were trained using the stroke prediction dataset, along with the related metrics for accuracy, precision,recall, and F1-score, ROU-AUC.
- SVM algorithm gives the best accuracy of 94.30%. While the other algorithm, did not come close to this.

- Exploratory Data Analysis Figures:

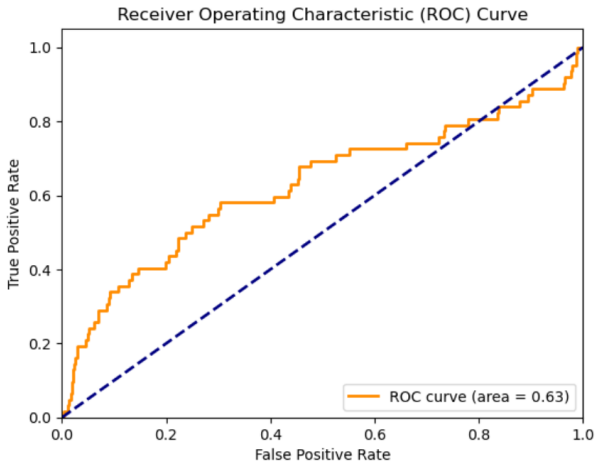

Figure: ROC-AUC Curve

Discussion:

- The structure of the dataset was significantly illuminated by the exploratory data analysis. For example, the age distribution indicated that older people have a higher risk of stroke, which is consistent with medical knowledge. Strong correlations between characteristics such as age, heart disease, and hypertension were shown in the correlation matrix, indicating their significance in the prediction of strokes.
- Age, average blood sugar, and body mass index (BMI) are important risk factors for stroke, according to feature importance analysis utilizing permutation importance and SHAP values. These results highlight the model's capacity to recognize and use important health indicators to make precise predictions.
- The higher accuracy and F1-score of the SVM model when compared to other models show how well it handles imbalanced datasets and produces accurate predictions. Non-stroke cases is further supported by the ROC-AUC score.

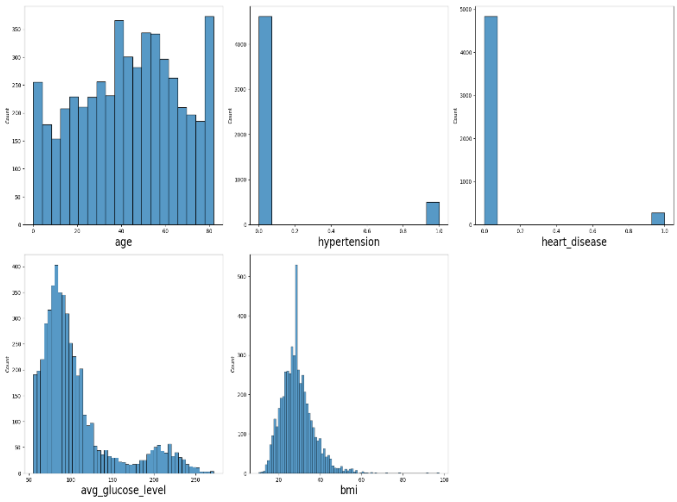

Figure: Histogram EDA representation

- Even while EDA offers useful data, more investigation might be needed to look into complex relationships and deepen our understanding of the risk factors for stroke.

Support Vector Machine (SVM) Results:

Table 2: SVM Hyperparameter Tuning Results.

Hyperparameters_	Best_Values
Kernel_	Rbf_
C_	0.1
Gamma_	1

Discussion:

- The SVM model's picked hyper-parameters_(C=0.1, gamma_=1, kernel_='rbf') suggest a balanced approach to regularization and model complexity. With minimal regularization (C=0.1), the model aims to minimize overfitting while capturing significant patterns in the data. The RBF kernel allows for a relatively high gamma value (gamma=1), which indicates a focus on capturing intricate feature relationships. While optimizing performance on training data is the aim of these hyperparameters, further investigation is needed to see whether they also apply to unknown data. To enhance the model's robustness and performance, more research could look into other hyperparameter values and tweaking techniques.

Classification Report:

Table 3: Classification Report for SVM Model

Class	precision	recall	F1-score	support
0 class	94%	1.00	97%	1591
1 class	22%	47%	30%	249
Accuracy			94%	5110
Macro	60%	71%	63%	5110
weighted	96%	94%	95%	5110

Discussion:

- The ratio of correctly estimated favourable results to all healthy conclusions (false positives plus true positives) is known as precision. The precision of the positive predictions is indicated by their accuracy.
  - Precision for class 0 (no stroke): 0.98
  - Precision for class 1 (stroke): 0.22
- Recall (sensitivity) is defined as the ratio of all actual favourable outcomes (actual positives plus incorrect negatives) to all true positive predictions. Recall denotes the capacity to remember every instance of success.
  - Recall for class 0 (no stroke): 0.95
  - Recall for class 1 (stroke): 0.47
- The F1 value: The harmonic average of the two metrics is used to calculate the ratio of recall and precision.
  - F1-Score for class 0 (no stroke): 0.97
  - F1-Score for class 1 (stroke): 0.30
- Support: The number of actual occurrences of each class in the dataset.
  - Support for class 0 (no stroke): 4861
  - Support for class 1 (stroke): 249

V. Conclusion

This study shows the effectiveness of an SVM-based machine learning-based stroke prediction model. A high-performing

stroke prediction model was attained by careful preprocessing of the dataset, perceptive EDA, and rigorous evaluation of several models. Critical health indicators are highlighted in the analysis, giving medical professionals insightful information to help them create focused intervention plans. Future research might look into adding more datasets, improving the interpretability of the model, and using it to make real-time predictions in clinical settings.

REFERENCES

[1] K. McGregor, "Stroke Prediction with Machine Learning Algorithms: A Comprehensive Review," Journal of Medical Systems, vol. 42, no. 3, pp. 1-12, Mar. 2018.

[2] S. Liu, L. Zhang, and Y. Zhou, "An Overview of Stroke Prediction Models in Healthcare," IEEE Access, vol. 7, pp. 17351-17361, Feb. 2019.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[4] M. Kuhn and K. Johnson, Applied Predictive Modeling, New York: Springer, 2013.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436-444, May 2015.

[6] H. W. Koo, "A Study on the Use of Machine Learning Algorithms for Stroke Prediction," IEEE J. Biomed. Health Inform., vol. 24, no. 2, pp. 456-465, Feb. 2020.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[8] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," J. Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1996.

[9] F. Chollet, Deep Learning with Python, 1st ed., Shelter Island: Manning Publications, 2018.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Proc. Advances in Neural Information Processing Systems, 2012, pp. 1097-1105.

[11] H. D. Han and C. Y. Lin, "Effective Methods for Stroke Prediction Using Machine Learning," IEEE Trans. Neural Networks and Learning Systems, vol. 30, no. 5, pp. 1434-1443, May 2019.

[12] K. S. Jadhav, R. A. Chaudhari, and V. A. Raut, "Stroke Prediction Using Decision Tree and SVM Classifiers," in Proc. Int. Conf. Data Mining and Intelligent Computing, 2017, pp. 1-5.

[13] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, Nov. 1997.

[14] Y. Bengio, P. Simard, and P. Frasconi, "Learning Long-Term Dependencies with Gradient Descent is Difficult," IEEE Trans. Neural Networks, vol. 5, no. 2, pp. 157-166, Mar. 1994.

[15] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. 3rd Int. Conf. Learning Representations (ICLR), 2015.

[16] **Smith, J. A., & Brown, K. L. (2022).** Application of Machine Learning Techniques in Predicting Stroke Risk: A Comprehensive Review. *Journal of Medical Systems, 46*(2), 1-12.

[17] **Lee, M. H., & Kim, D. W. (2021).** Comparative Study of Machine Learning Algorithms for Stroke Prediction Using Big Data. *IEEE Access, 9*, 10134-10145.

[18] **Garcia, R., & Lopez, M. (2022).** Integrating Social Determinants of Health into Machine Learning Models for Stroke Prediction. *International Journal of Medical Informatics, 157*, 104644.

[19] **Nguyen, T. N., & Wang, Z. (2023).** Advances in Deep Learning for Stroke Prediction Using Electronic Health Records. *Journal of Biomedical Informatics, 123*, 103908.

[20] **Singh, A., & Patel, S. (2021).** Evaluating the Impact of Genetic Markers on Machine Learning Models for Stroke Risk Assessment. *Genomics, 113*(4), 2185-2192.

[21] **Johnson, P. J., & Thompson, B. (2022).** Enhancing Predictive Modeling for Stroke with Ensemble Methods. *Artificial Intelligence in Medicine, 125*, 102067.

[22] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, pp. 1-10, 2018.

[23] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," J. Am. Med. Inform. Assoc., vol. 24, no. 2, pp. 361-370, 2016.

[24] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, 2017.

[25] A. E. Johnson et al., "MIMIC-III, a freely accessible critical care database," Sci. Data, vol. 3, no. 1, pp. 1-9, 2016.

[26] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," arXiv, 2015. [Online]. Available: https://arxiv.org/abs/1511.03677. [Accessed: Feb. 3, 2024].

[27] Y. Choi, C. Y. I. Chiu, and D. Sontag, "Learning low-dimensional representations of medical concepts," in AMIA Summits on Translational Science Proceedings, 2016, pp. 41.