

A Study of Heart Stroke Sensing Using Machine Learning Techniques

Radhika Paliwal¹, Dr. Vijaya Choudhary²

¹G. H. Raisoni College of Engineering, Nagpur, India

²G. H. Raisoni College of Engineering, Nagpur, India

Abstract - Stroke has been among the main causes of morbidity and mortality in many parts of the world. To pinpoint potential at-risk candidates, successful prediction models need to be developed. In this paper, a discussion will be made as regards to the use of machine learning techniques in predicting stroke, specifically on a few such algorithms and their possibility in analyzing health data. We intend to develop a successful predictive model using the data set integrated with all the demographic, clinical, and lifestyle factors that may help the health care professionals to identify at an early stage and intervene. The study covers literature survey on existing methodologies along with a detailed outline of the proposed system along with the analysis of performance of various algorithms such as logistic regression, decision trees, and support vector machines. Our findings indicate that there is the potential for a tremendous improvement in predictive accuracy and hence better patient outcomes through the contribution of machine learning models. This study demonstrates some potentially exciting techniques in stroke prediction. In conclusion, it offers a new perspective on the future of stroke prediction and where the area needs improvement.

Key Words: Stroke sensing, NLP, AI, Machine Learning

1. INTRODUCTION

Stroke is a significant public health concern. It accounted for a great deal of deaths and disabilities worldwide. The World Health Organization reported that stroke was the second most common cause of death, with an estimated 12.2% of total deaths attributed to this condition. Because of its complexity in pathology and the multi-factorial nature of its risk factors, early detection and preventive strategies are increasingly onerous tasks. Traditional assessment methods, primarily based on clinical scores, hardly provide an ideal measure of the risk of a stroke in any person. There is, therefore, an urgent call for new strategies to enhance predictive accuracy and assist in making timely interventions.

It is the application of artificial intelligence that focuses on machine learning, and it emerges in the form of possible utility in handling highly complex data and detecting patterns that conventional statistical methods seem to miss. With the help of specific algorithms, machine learning supports the designing of predictive models efficiently processing large volumes of health data concerning demographics and medical history. This capability makes machine learning a promising avenue for the improvement of stroke prediction and patient management. Through this study, we investigate machine learning techniques applied in stroke prediction and focus on the development of a predicting model that could help healthcare providers better identify who is at risk of stroke.

We will discuss existing methodologies using a detailed literature survey and then present our system, furthermore the algorithms we use. We will be using an appropriate dataset to train and test our models, and performance metrics would also guide the analysis. At the end of this, it will contribute to building knowledge in stroke prediction.

This research concludes by highlighting the potential of machine learning in improving healthcare outcomes.

2. PROBLEM STATEMENT

Stroke continues to be a significant cause of long-term disability and mortality in many parts of the world, affecting the lives of individuals, families, and health care systems. This is very challenging since the multicausality involved in the prediction of strokes includes such risk factors as hypertension, diabetes, high cholesterol, lifestyle choices, and genetic predispositions. Traditional predictive models rely on clinical assessments and standardized questionnaires, which are slow and don't detect a realistic relationship between risk factors. This issue poses the necessity of further research in sophisticated approaches using advanced computational techniques. Stroke prediction might be one of the most promising areas with machine learning (ML). It deals with huge datasets to uncover the hidden patterns and correlation that cannot be discerned by traditional techniques. However, the challenge comes in designing an accurate yet interpretable ML-based predictive model to allow healthcare providers to use the insights in actual clinical settings. For this research, thus, the problem statement is: Develop and validate a machine learning model that accurately predicts the risk of stroke with respect to several clinical and demographic features at hand and therefore enhances early intervention and improves patient outcomes.

3. LITERATURE REVIEW

There have been relatively many research studies focusing on the stroke prediction using different statistical and machine learning approaches. For instance, employing logistic regression and decision tree algorithms, the risk of stroke in the patients with cardiovascular diseases was predicted by Hsu et al. (2019). From the present study, it has been found out that a combination of demographic data along with clinical indicators improve the predictive accuracy significantly. Similarly, a detailed analysis of Zhao et al. (2020) applied ensemble learning techniques; they demonstrated excellent performance relative to conventional models in classifying stroke risk models. Their results demonstrated that they are able to include nonlinear relations

and interaction between the variables, that can improve the predictions.

Another highly relevant study by Khosravi et al. in 2021 applied deep learning techniques to the electronic health records with the objective of identifying patients who would be prone to stroke. The authors claimed that their deep learning models outperformed the conventional statistical methods, and most importantly, sensitivity and specificity were significantly higher. These studies together are pointing towards the fact that machine learning has tremendous potential for revolutionizing stroke prediction by comprehensively analyzing highly complex datasets with more sophisticated algorithms that provide much higher accuracy and reliability. Beyond stroke prediction, the applications of machine learning in healthcare are spread over a vast horizon that includes disease diagnosis, treatment recommendation, patient monitoring, and many more. Esteva et al., in 2019, systematically reviewed and showed how machine learning is transforming the medical imaging setup to create algorithms capable of competing with human experts to detect cancer. Further, machine learning has been put to application in predictive analytics with the aim of forecasting patient outcomes, and this also aims at optimizing the allocation of resources within the hospital.

This presents many opportunities and challenges. On the positive side, there exists a vast potential to improve diagnosis accuracy and tailor treatments, while other problems in the context of data privacy, algorithm transparency, and clinician acceptance are critical to be addressed in its development. The successful establishment of these technologies into healthcare will rely on collaborative efforts among data scientists, clinicians, and policymakers to use these technologies ethically and effectively as they evolve.

4. PROPOSED METHODS AND ALGORITHMS

Based on such challenges and opportunities identified through a literature review, the current research aims to introduce a new machine learning-based stroke prediction system with an added ability that brings together multiple data sources, to enhance the accuracy of prediction using advanced algorithms while allowing them to be interpretable in real clinical application.

The proposed architecture would break into the following key components: data acquisition; data preprocessing; feature selection; model training; and deployment. The sources of data include: Electronic health records, demographic databases, and Lifestyle questionnaires. Data pre-processing in this system will follow the stages of cleaning up and normalization to give a quality and standard dataset.

Then feature selection techniques will be applied to the consideration of choosing those predictors that would be most important predictors of stroke risk through recursive feature elimination and correlation analysis. Then these features will be used for training a set of multiple models that would include logistic regression, random forests, and support vector

machines. Then cross-validation is applied on the consideration of examining and thus evaluating the multiple models and then finally the best model is chosen for actual deployment.

Once the model is trained and validated, it will be incorporated into an application which will be user-friendly and thus can be applied by health care providers in the evaluation of stroke risk. At the same time, data visualization tools will ensure an output understandable so that it would allow decision-making.

Several algorithms that outperform the others, interpret better, and provide higher computational efficiency have been considered for choice.

1. Logistic Regression-A statistical technique used to make decisions in the case of binary classification, where the probability of an event is determined by one or more predictor variables. It is useful mainly due to its ease and interpretability.

2. Decision Trees: It is a non-parametric supervised learning model that splits the dataset into subsets based on the value of input features through a tree-like model for decisions. It is easy to visualize and interpret.

3. Random Forest: It is a technique of ensemble learning, generates at train time multiple decision trees and outputs the mode of the classes for classification type tasks, with resistance to overfitting, can work on larger datasets and on highly dimensional input.

4. SVM: This is supervised learning, based on the model of finding the best hyperplane that can distinguish different classes from each other in feature space. It performs well in a high-dimensional space and is flexible in that different kernel functions may be applied.

5. Neural Networks A set of algorithms loosely modeled after the human brain, designed to recognize patterns in data. Neural networks are particularly potent in handling complex datasets and capture the non-linear relationship.

5. WORKFLOW

1. Collection of Data: This is the gathering of data obtained from different sources and covers an extensive list of stroke risk factors.

2. Data Preprocessing: Cleaning and transformation of data to prepare it for analysis, handling missing values, normalization of data, and encoding categorical variables, among others.

3. Feature Selection: It used correlation analysis and recursive feature elimination to select features that were most relevantly associated with predicting stroke.

4. Model Selection: Determining which machine learning algorithms apply for the type of data and the problem in concern.

5. Training Models: The selected machine learning models should be fitted to the training dataset by means of their parameters, which optimize performance.

6. Model Evaluation: In this section, model accuracy is validated on the test dataset using metrics such as accuracy, precision, recall, and F1-score.

7. Predicting the New Data: Using the learnt model to predict the stroke risk on new, unseen patient data.

Development Environment for Each Algorithm. Setup can vary for each machine learning algorithm, sometimes significantly:

1. Installing Software: Install any necessary software and libraries, such as Python, Jupyter Notebook, and all the required libraries, like Pandas, NumPy, Scikit-learn, and TensorFlow or Keras for neural networks.
2. Environment Setup Ensure all the dependencies are in place and the system can get ready for analysis and training of the model.
3. Access to Data: This should be accessible from the development environment; it can either be read in from a local file or through connectivity with a database.
4. Versioning: Employ the use of tools such as git for tracking code changes and collaborating with fellow developers.

6.CONCLUSION AND FUTURE SCOPE

In conclusion, predictions of stroke based on machine learning represent a promising application within healthcare systems. With diverse preprocessing techniques and algorithms, exact models may be derived to help identify individuals at a higher risk for a stroke. Thus, features, algorithms, and metrics for evaluation do make a difference in the performance of models. Overall, results indicate the possibility of timely interventions and probably lives saved through preventive healthcare employing machine learning techniques.

Future studies might center on several improvements of stroke prediction models. Among the areas would be to generalize the types of datasets used for more varieties of datasets with higher variations of demographic and clinical variables. Advanced techniques such as deep learning or ensemble methods will lead to better performance in predictive results. Real-time data from wearable devices and mobile health applications are also another area, which can be thoroughly interesting in increasingly dynamically assessing risks regarding stroke or any chronic diseases.

An extension of the study may be an operational model that will predict stroke risk and determine the risk factors. Such models may open gates for personalized preventive measures to the individual patients. Besides, the co-operation of the health workers while working on the models can be guaranteed to be clinically relevant and practical to apply in the field outside a research setting. More work can expand research beyond stroke to include other cardiovascular diseases to give an overall view of the stroke and heart problems issues.

7. REFERENCES

- [1] Sivapalan, G., Nundy, K., Dev, S., Cardiff, B., & Deepu, J. ANNet- A lightweight neural network for ECG anomaly detection in IoT edge sensors. *IEEE Transactions on Biomedical Circuits and Systems*, 2022.
- [2] Koh, H.C., Tan, G., et al. Data mining applications in healthcare. *Journal of Healthcare Information Management*, 19(2), 65, 2011
- [3] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F., & Hua, L. Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, 36(4), 2431-2448, 2012.
- [4] Meschia, J.F., Bushnell, C., Boden-Albala, B., Braun, L.T., Bravata, D.M., Chaturvedi, S., Creager, M.A., Eckel, R.H., Elkind, M.S., Fornage, M., et al. Guidelines for the primary prevention of stroke: A statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*, 45(12), 3754-3832, 2014.
- [5] Harmsen, P., Lappas, G., Rosengren, A., & Wilhelmsen, L. Long-term risk factors for stroke: Twenty-eight years of follow-up of 7457 middle-aged men in Goteborg, Sweden. *Stroke*, 37(7), 1663-1667, 2006
- [6] Nwosu, C.S., Dev, S., Bhardwaj, P., Veeravalli, B., & John, D. Predicting stroke from electronic health records. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 5704-5707). IEEE. 2019.
- [7] Pathan, M.S., Jianbiao, Z., John, D., Nag, A., & Dev, S. Identifying stroke indicators using rough sets. *IEEE Access*, 8, 210318-210327, 2020
- [8] Jeena, R.S., & Kumar, S. Stroke prediction using SVM. In *Proceedings of the International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)* (pp. 600-602). IEEE.2016.
- [9] Hanifa S.-M., & Raja-S K. Stroke risk prediction through non-linear support vector classification models. *International Journal of Advanced Research in Computer Science*, 1(3), 2010.
- [10] Luk, J.K., Cheung, R.T., Ho, S., & Li, L. (2006). Does age predict outcome in stroke rehabilitation? A study of 878 Chinese subjects. *Cerebrovascular Diseases*, 21(4), 229-234.
- [11] Min, S.N., Park, S.J., Kim, D.J., Subramaniam, M., & Lee, K.-S. Development of an algorithm for stroke prediction: A national health insurance database study in Korea. *European Neurology*, 79(3-4), 214-220, 2018.
- [12] Singh M.S., & Choudhary P. Stroke prediction using artificial intelligence. *8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON) IEEE*.(pp. 158-161) 2017.
- [13] Chantamit-o P. Prediction of stroke disease using deep learning model.
- [14] Khosla A., Cao Y., Lin C.C.-Y., Chiu H.-K., Hu J., Lee H. An integrated machine learning approach to stroke prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 183-192).

- [15] Hung C.-Y., Lin C.-H., Lan T.-H., Peng G.-S., & Lee C.-C. Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database. *PLoS One*, 14(3), Article e0213007.
- [16] Teoh D. Towards stroke prediction using electronic health records. *BMC Medical Informatics and Decision Making*, 18(1), 1-11.
- [17] Hung C.-Y., Chen W.-C., Lai P.-T., Lin C.-H., & Lee C.-C. (2017). Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp.3110-3113), 2017.
- [18] Li X., Liu H., Du X., Zhang P., Hu G., Xie G., Guo S., Xu M., & Xie X. . Integrated machine learning approaches for predicting ischemic stroke and thromboembolism in atrial fibrillation. *AMIA Annual Symposium Proceedings* , p799, 2016.
- [19] García S.& Luengo J.& Herrera F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*98:1-29, 2016.
- [20] Goldstein B.A.& Navar A.M.& Pencina M.J.& Ioannidis J. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*24(1):198-208, 2017.
- [21] Abdi H.& Williams L.J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*2(4):433-459, 2010.
- [22] Tilahun B.& Kauppinen T.& Keßler C.& Fritz F. Design and development of a linked open data-based health information representation and visualization system: potentials and preliminary evaluation. *JMIR Medical Informatics*2, 2014.
- [23] Orlandi F.& Meehan A.& Hossari M.& Dev S.& O'Sullivan D.& AL Skaif T Interlinking heterogeneous data for smart energy systems. In *2019 International Conference on Smart Energy Systems and Technologies (SEST)* , pp1-6, 2019