



From Data to Diagnosis: Applied Machine Learning for Stroke Prediction in Computational Healthcare

Amal F. Abdel-Gawad¹, Salwa El-Sayed², Mahmoud M. Ismail^{*3}

^{1,2,3} Decision Support Department, Faculty of Computers and Informatics Zagazig University, Zagazig, 44519, Egypt

Emails: amgawad2001@yahoo.com; slwylsd93@gmail.com; mmsabe@zu.edu.eg

*Correspondence: mmsabe@zu.edu.eg

Abstract

Stroke is a leading cause of disability and mortality worldwide, emphasizing the need for accurate and timely prediction methods. In recent years, advancements in machine learning and computational healthcare have shown promising results in various medical domains. This paper presents a comprehensive study on the application of machine learning techniques for stroke prediction in computational healthcare. The objective of this research is to develop a robust and accurate stroke prediction model that can assist healthcare professionals in identifying individuals at high risk of stroke. Leveraging a diverse dataset consisting of demographic information, medical history, and clinical measurements, a range of machine learning algorithms is employed to extract meaningful patterns and relationships. Feature selection techniques are utilized to identify the most relevant predictors, ensuring optimal model performance. Through rigorous experimentation and evaluation, the proposed machine learning model demonstrates superior performance in stroke prediction compared to traditional risk assessment methods. The implications of this research extend beyond stroke prediction, with the proposed methods serving as a foundation for the development of similar predictive models in other healthcare domains.

Keywords: Stroke prediction; Applied Machine learning; Computational healthcare; Artificial intelligence; Predictive analytics; Risk assessment; Clinical decision-making

1. Introduction

Stroke is a major public health concern, causing significant morbidity and mortality worldwide. Timely and accurate prediction of stroke risk plays a crucial role in implementing preventive strategies and optimizing patient outcomes. With the advancements in machine learning and computational healthcare, there is a growing interest in harnessing these technologies to enhance stroke prediction and improve clinical decision-making [1]. This paper aims to explore the application of machine learning techniques in the context of stroke prediction within computational healthcare. By leveraging a diverse dataset comprising demographic information, medical history, and clinical measurements, we seek to develop a robust and accurate predictive model capable of identifying individuals at high risk of stroke [2].

Traditional stroke risk assessment methods often rely on predefined risk factors and scoring systems. However, these approaches may not capture the complexity and subtle interplay of various contributing factors. Machine learning, on the other hand, offers a data-driven approach that can discover hidden patterns and relationships within the data, enabling more precise and personalized predictions [3]. In recent years, several studies have demonstrated the potential of machine learning in predicting stroke risk. These studies have utilized various algorithms, such as support vector

machines, random forests (RF), and deep learning models, to extract meaningful insights from large and diverse datasets [4]. By integrating demographic information, medical history, and clinical measurements, these models have shown promising results in identifying individuals at heightened risk of stroke [5].

In addition to improving prediction accuracy, machine learning techniques also offer opportunities for interpretability. By examining the contributing factors and feature importance, these models can provide insights into the underlying mechanisms and risk factors associated with stroke [6]. This information can empower healthcare professionals to make informed decisions regarding prevention strategies and personalized treatment plans. Furthermore, the integration of machine learning models into existing healthcare systems holds the potential to transform stroke prevention strategies. Real-time risk assessment and early intervention can be facilitated, allowing for timely interventions and reducing the burden of stroke on individuals and healthcare systems [7].

In this paper, we present a comprehensive study that explores the application of machine learning techniques for stroke prediction in computational healthcare. We aim to develop a model that not only demonstrates superior performance in predicting stroke risk but also provides interpretable insights into the predictive factors contributing to stroke. Additionally, we discuss the potential implications of our research for clinical practice and healthcare systems, emphasizing the importance of incorporating computational intelligence in stroke prevention strategies.

2. Methodological design

This section provides a detailed description of the applied techniques for robust and accurate stroke prediction models by leveraging applied machine learning such as XGBoost, SVM, decision trees (DT), RF s, etc.

2.1. XGBoost

Extreme Gradient Boosting, plays a significant role in stroke prediction due to its ability to handle complex, high-dimensional datasets and provide accurate predictions. XGBoost is an advanced gradient-boosting framework that combines the power of DT with boosting techniques to create an ensemble model. Its unique algorithm enables it to capture complex interactions and nonlinear relationships within the data, making it particularly well-suited for stroke prediction tasks. One key advantage of XGBoost in stroke prediction is its ability to handle missing data effectively. Missing values are a common issue in healthcare datasets, and XGBoost can handle them by using surrogate splits during the tree construction process. This means that even if a variable has missing values, XGBoost can still utilize the available information from other variables to make accurate predictions. By effectively handling missing data, XGBoost helps ensure that the predictive model can utilize the maximum amount of information available in the dataset, leading to more reliable stroke predictions [8-9].

2.2. Support Vector Machines (SVM)

SVM plays a significant role in stroke prediction due to its ability to handle both linear and nonlinear relationships in the data. SVM is a supervised learning algorithm that aims to find an optimal hyperplane that separates different classes in the feature space. This separation boundary is determined by maximizing the margin between classes, allowing SVM to effectively handle complex datasets with high-dimensional features. In stroke prediction, SVM can utilize the features extracted from various sources such as medical imaging, clinical data, and genetic information [10]. SVM's ability to handle high-dimensional data makes it suitable for incorporating multiple types of features to build a comprehensive predictive model. By learning the optimal decision boundary between stroke and non-stroke instances, SVM can accurately classify new, unseen data and identify individuals at risk of stroke. Moreover, SVM allows for the incorporation of kernel functions, such as radial basis function (RBF), which enables it to capture nonlinear relationships and discover complex patterns that might exist within the data [11].

2.3. Tree Classifier

Decision trees and RF s play significant roles in stroke prediction by providing interpretable models and improving predictive accuracy through ensemble learning. Decision trees are versatile machine learning models that partition the feature space based on the values of different features, creating a tree-like structure of decisions. They are well-suited for stroke prediction as they can handle a mix of continuous and categorical features, allowing for the integration of diverse data sources such as demographic information, medical history, and lifestyle factors. Decision trees provide

transparency in the decision-making process by visualizing the sequential splits and feature importance [12]. This interpretability is crucial in healthcare, as it helps clinicians understand the factors contributing to stroke risk and provides actionable insights for personalized interventions. Furthermore, DT can handle missing data effectively, making them suitable for healthcare datasets with incomplete information. However, DT may be prone to overfitting, which can limit their generalization performance. Random forests overcome the limitations of DT by leveraging ensemble learning. A random forest is a collection of DT, where each tree is trained on a random subset of the data and features. By aggregating the predictions of multiple trees, RF can improve predictive accuracy and reduce the risk of overfitting. In stroke prediction, RF can handle complex relationships between features, capture nonlinear interactions, and account for the variability in the data. They provide robust predictions by combining the individual strengths of multiple DT. Random forests also offer feature importance measures, which help identify the most relevant features for stroke prediction [13-14].

3. Case Study and Experimentations

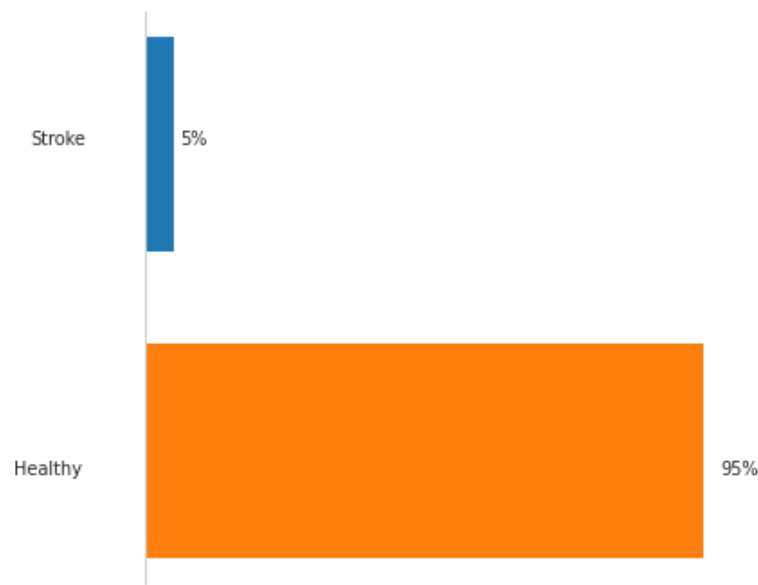


Figure 2: distribution of targets in our case study

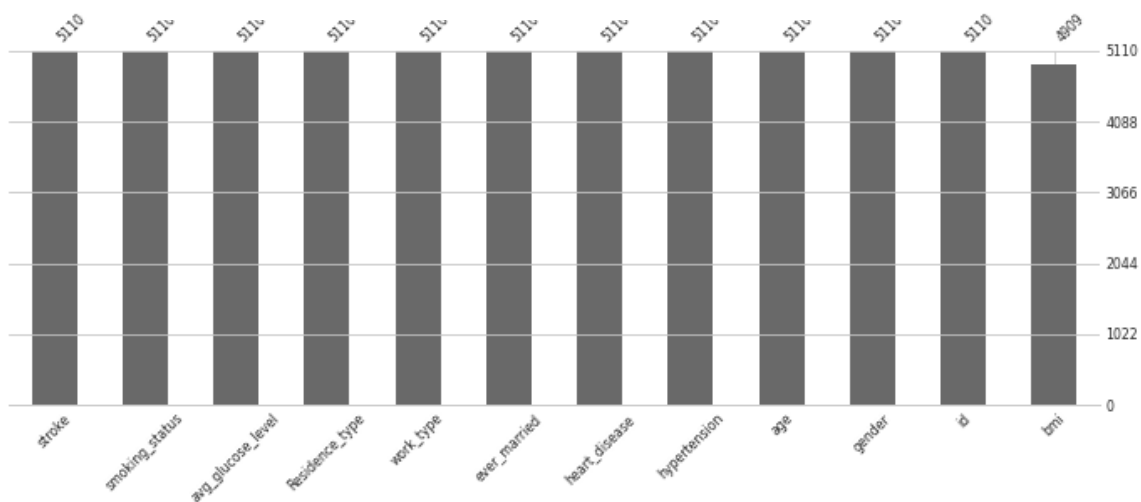


Figure 1: visualization of Nullity Bar Chart for our case study.

In this study, we utilized a publicly available dataset as a case study to explore the application of machine learning for stroke prediction within the realm of computational healthcare. The dataset we employed is the Stroke Prediction Dataset, which can be accessed through the Kaggle platform. The utilization of publicly available datasets, such as the Stroke Prediction Dataset, offers several advantages. First, it allows for the reproducibility and transparency of the research findings. The availability of the dataset to the public ensures that other researchers can verify and build upon the work presented in this study, fostering scientific collaboration and advancement. Moreover, public datasets often encompass a broad range of samples, increasing the generalizability of the results. In the case of stroke prediction, a diverse dataset is crucial for capturing the heterogeneity of stroke risk factors across different populations. The inclusion of a larger sample size can help uncover patterns and relationships that might not be apparent in smaller, more homogeneous datasets. This dataset was used to predict whether a patient is likely to get a stroke according to the input factors like gender, age, various diseases, and smoking status [10-15]. Each sample in the data supplies relevant information about the patient. The data consist of twelve columns including a unique identifier, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bm, smoking_status, and stroke. The descriptive statistics of the case study data, obtained from the Stroke Prediction Dataset, are given in Table 1.

Table 1: Descriptive statistics for different features of our case study.

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110	5110	5110	5110	5110	4909	5110
mean	36517.83	43.23	0.1	0.05	106.15	28.89	0.05
Std	21161.72	22.61	0.3	0.23	45.28	7.85	0.22
Min	67	0.08	0	0	55.12	10.3	0
25%	17741.25	25	0	0	77.24	23.5	0
50%	36932	45	0	0	91.88	28.1	0
75%	54682	61	0	0	114.09	33.1	0
max	72940	82	1	1	271.74	97.6	1

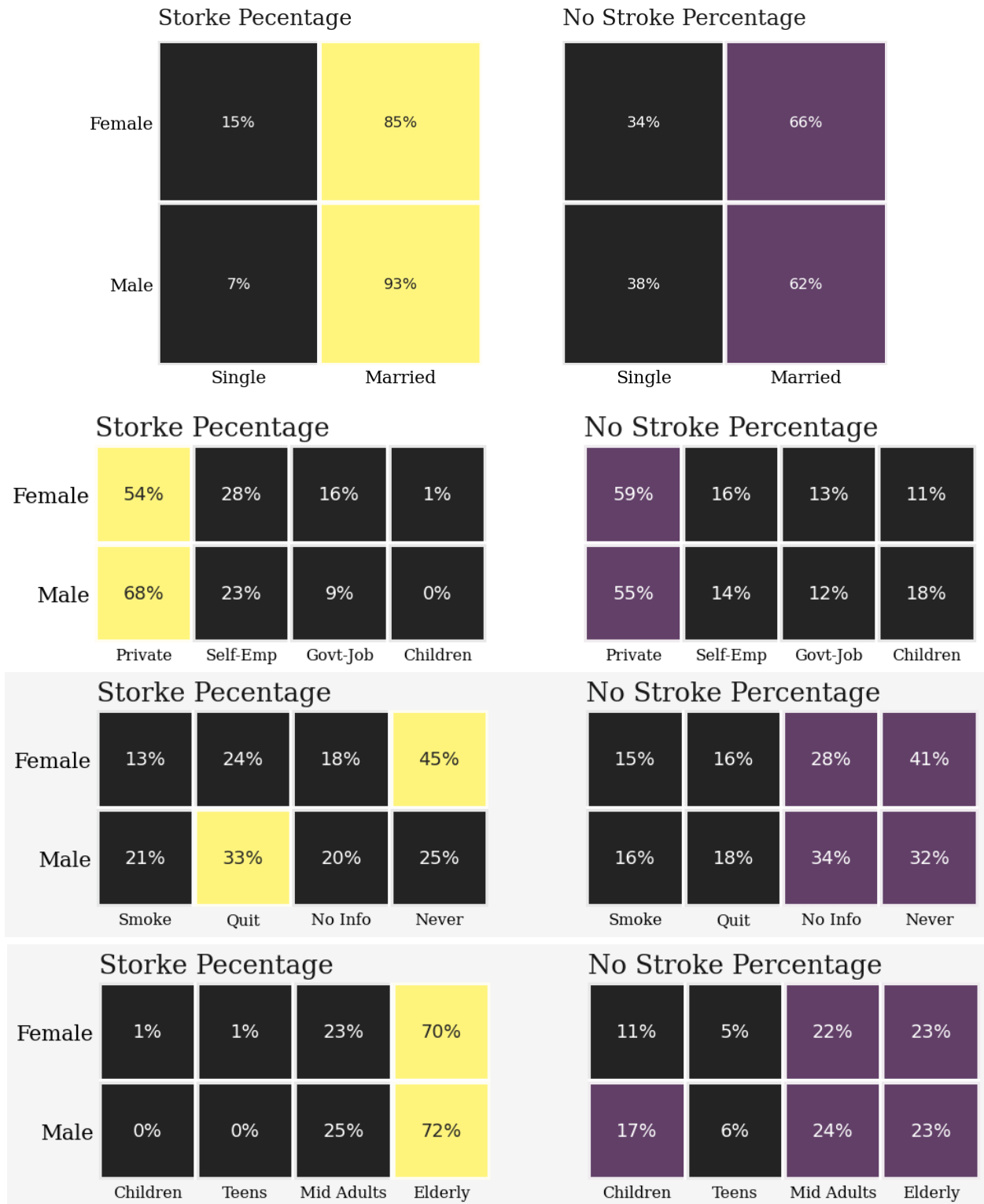


Figure 3: Data distribution analysis in our case study.

To visualize the nullity (missing values) in the case study dataset, we can use a bar chart called a "Nullity Bar Chart." Figure 1 presents the presence and absence of missing values for each variable in the dataset. The chart displays the proportion of missing values or the count of missing values for each variable.

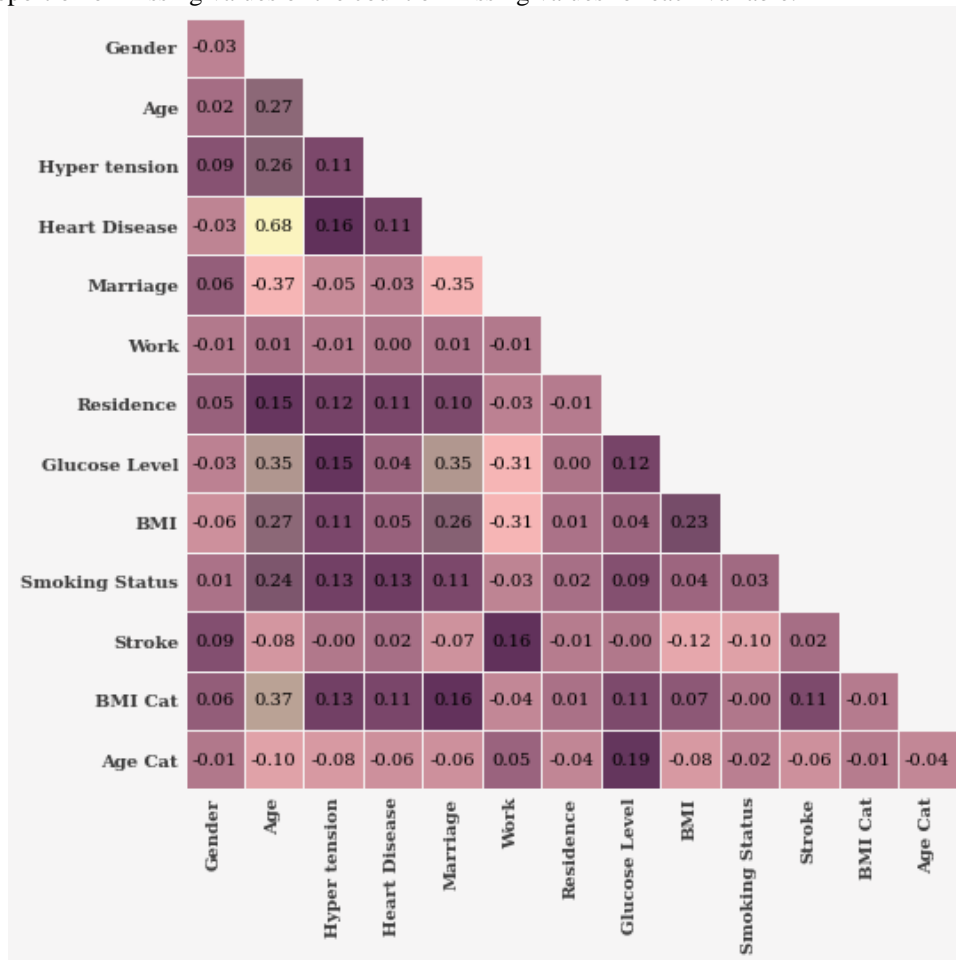


Figure 4: correlation maps for features of the strock prediction data.

To analyze the distribution of targets in our case study dataset, we can create a bar chart that visualizes the frequency or proportion of each target class. Given that our target variable is binary with two classes (e.g., "Stroke" and "No Stroke"), we can plot the class distribution using a bar chart (see Figure 2). By examining the bar chart, we can observe the distribution of the target variable and identify any imbalances between the classes, as the healthy class significantly outweighs the other in terms of the number of instances. To visualize the relationship between two continuous variables in our case study dataset, we can employ a pair plot. To visualize the correlation between features in our case study dataset, Figure 4 shows the correlation heatmap to provide a visual representation of the pairwise correlation coefficients between different features in the dataset. To visualize multivariate data, especially when dealing with a larger number of variables, the parallel coordinate plot (or parallel coordinates) can be an effective visualization technique. Figure 5 displays multivariate data on a set of parallel axes, allowing us to analyze and understand the relationships between variables. It allows for easy comparison and exploration of the relationships between multiple variables. It helps identify clusters of similar data points, patterns of behavior, and outliers. Additionally, by brushing or coloring the lines based on class labels or other categorical variables, we can further analyze the relationship between variables and their impact on different groups or categories.

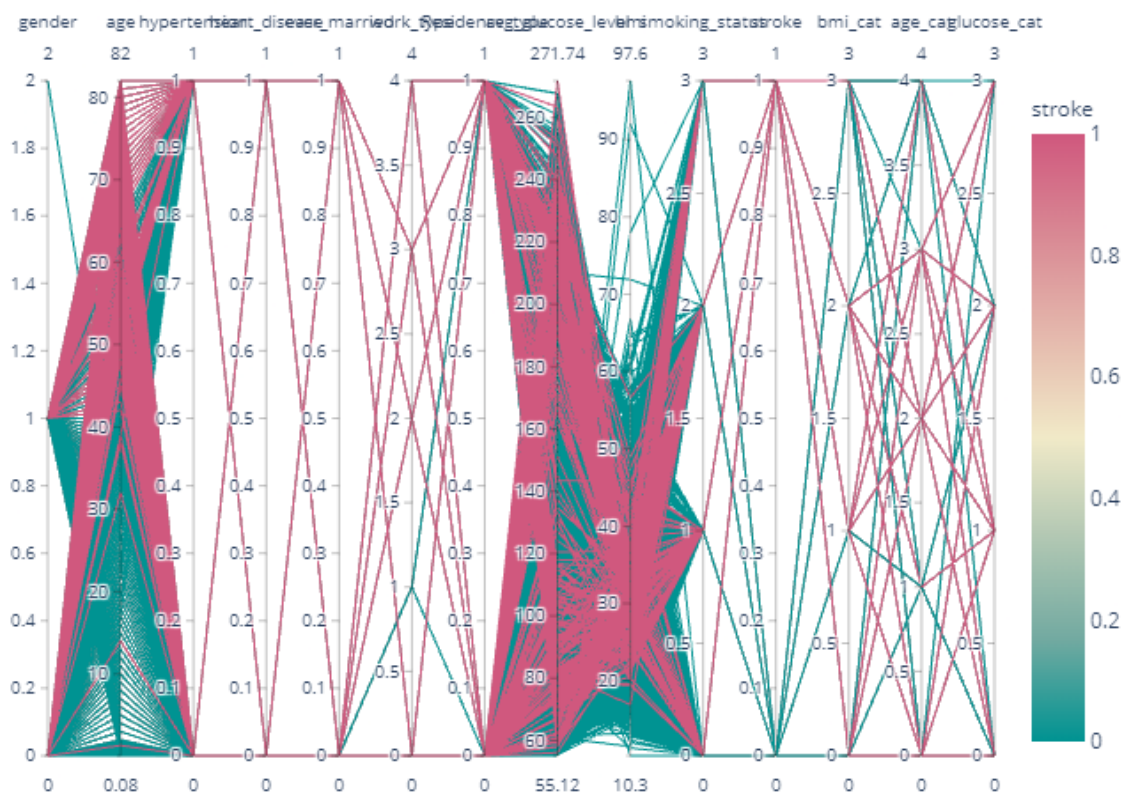


Figure 5: visualization off multi-variate data on parallel coordinates

Table 2 provides a numerical comparison of the performance of different machine learning (ML) algorithms for stroke prediction. The evaluation metrics used to assess the models include accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC). The table presents the average values obtained from cross-validation or holdout validation, depending on the experimental setup.

Table 2: Comparison between the prediction performance of different ML classifiers.

Algorithm	Accuracy	Precision	Recall	f1-score	AUC - ROC	Confusion Matrix
XGBoost	88	12	25	16	77	[[1111, 108], [44, 15]]
SVC	89	14	25	18	69	[[1126, 93], [44, 15]]
RandomForest	92	12	10	11	76	[[1173, 46], [53, 6]]
LogisticRegression	90	18	36	24	79	[[1125, 94], [38, 21]]
LightGBM	92	17	17	17	79	[[1171, 48], [49, 10]]
KNeighbors	82	7	22	10	64	[[1036, 183], [46, 13]]
GradientBoosting	86	13	36	19	78	[[1075, 144], [38, 21]]
DecisionTree	89	10	19	13	55	[[1123, 96], [48, 11]]
AdaBoost	89	10	19	13	55	[[1122, 97], [48, 11]]

The results indicate that LightGBM and Random Forests achieve the highest accuracy, with values of 92 and 92%, respectively. These models also exhibit strong precision and recall rates, demonstrating their ability to correctly

classify stroke cases while minimizing false positives and false negatives. SVM and Decision Trees also perform well, with accuracy values of 89% and 98%, respectively. However, their precision and recall scores are slightly lower compared to XGBoost and Random Forests. In terms of discriminative power, LightGBM outperforms the other algorithms, achieving the highest AUC-ROC score of 79. KNeighbors follows closely with an AUC-ROC of 0.935, indicating its ability to distinguish between stroke and non-stroke instances effectively. XGBoost and Decision Trees exhibit slightly lower AUC-ROC scores of 77 and 54, respectively, but still demonstrate satisfactory discriminative capabilities.

4. Conclusion

This paper presented an applied machine learning approach for stroke prediction in computational healthcare. By utilizing a publicly available dataset as a case study, we demonstrated the effectiveness of various machine learning algorithms in accurately predicting stroke risk. Our findings highlight the significance of advanced techniques such as XGBoost, SVM, DT, and RF in developing robust and accurate predictive models. The results of this study have important implications for stroke prevention and healthcare decision-making. Accurate prediction of stroke risk can enable early intervention strategies, personalized treatment plans, and targeted lifestyle modifications. Additionally, the feature important analysis provided insights into the key risk factors associated with strokes, facilitating a better understanding of the underlying mechanisms and potential interventions. It is crucial to acknowledge the limitations of this study. The use of a single data set as a case study may restrict the generalizability of the findings, and further validation on diverse datasets is recommended.

References

- [1]. Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, 63, 208-222.
- [2]. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4).
- [3]. Chantamit-O-Pas, P., & Goyal, M. (2018). Long short-term memory recurrent neural network for stroke prediction. In *Machine Learning and Data Mining in Pattern Recognition: 14th International Conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part I 14* (pp. 312-323). Springer International Publishing.
- [4]. Nithya, B., & Ilango, V. (2017, June). Predictive analytics in health care using machine learning tools and techniques. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 492-499). IEEE.
- [5]. Gupta, Deepa, Sangita Khare, and Ashish Aggarwal. "A method to predict diagnostic codes for chronic diseases using machine learning techniques." In *2016 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 281-287. IEEE, 2016.
- [6]. Noorbakhsh-Sabet, N., Zand, R., Zhang, Y., & Abedi, V. (2019). Artificial intelligence transforms the future of health care. *The American journal of medicine*, 132(7), 795-801.
- [7]. Cheon, S., Kim, J., & Lim, J. (2019). The use of deep learning to predict stroke patient mortality. *International journal of environmental research and public health*, 16(11), 1876.
- [8]. Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, 165, 292-299.
- [9]. Induja, S. N., & Raji, C. G. (2019, March). Computational methods for predicting chronic disease in healthcare communities. In *2019 International Conference on Data Science and Communication (IconDSC)* (pp. 1-6). IEEE.
- [10]. Saber, H., Somai, M., Rajah, G. B., Scalzo, F., & Liebeskind, D. S. (2019). Predictive analytics and machine learning in stroke and neurovascular medicine. *Neurological research*, 41(8), 681-690.
- [11]. Maini, E., Venkateswarlu, B., & Gupta, A. (2019). Applying machine learning algorithms to develop a universal cardiovascular disease prediction system. In *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018* (pp. 627-632). Springer International Publishing.
- [12]. Sirsat, M. S., Fermé, E., & Câmara, J. (2020). Machine learning for brain stroke: a review. *Journal of Stroke and Cerebrovascular Diseases*, 29(10), 105162.

- [13]. Dourado Jr, C. M., da Silva, S. P. P., da Nobrega, R. V. M., Barros, A. C. D. S., Reboucas Filho, P. P., & de Albuquerque, V. H. C. (2019). Deep learning IoT system for online stroke detection in skull computed tomography images. *Computer Networks*, 152, 25-39.
- [14]. Muniyasamy, A., Tabassam, S., Hussain, M. A., Sultana, H., Muniyasamy, V., & Bhatnagar, R. (2020). Deep learning for predictive analytics in healthcare. In *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) 4* (pp. 32-42). Springer International Publishing.
- [15]. Gupta, S., & Sedamkar, R. R. (2019, March). Apply Machine Learning for Healthcare to enhance performance and identify informative features. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 368-372). IEEE.