# PredictStr: a balanced benchmark dataset for improve stroke prediction

Taissir Fekih Romdhane, Mohamed Ibn Khedher, Mounim A El-Yacoubi

# PredictStr: A Balanced Benchmark Dataset for Improve Stroke Prediction

Taissir Fekih Romdhane*, Mohamed Ibn Khedher† and Mounim A. El-Yacoubi‡
*LIPSIC-FST, Faculty of Sciences, University of Tunis El Manar, Tunisia
taissir.fekih@fst.utm.tn
†IRT - SystemX, 2 Bd Thomas Gobert, 91120 Palaiseau, France
mohamed.ibn-khedher@irt-systemx.fr
‡Samovar, Telecom SudParis, Institut Polytechnique de Paris, 19 place Marguerite Perey, 91120 Palaiseau, France
mounim.el_yacoubi@telecom-sudparis.eu

***Abstract*—Predicting strokes is essential for improving healthcare outcomes and saving lives. This paper introduces a benchmarking dataset, PredictStr, specifically developed to enhance stroke prediction. This dataset improves upon a previously unique dataset identified in the literature. Our methodology comprises two main steps: firstly, we outline a series of preprocessing and cleaning measures to enhance data quality. Secondly, we present a novel algorithm, the Dynamic Hybrid Balancing Algorithm, which builds upon the ADSYSN algorithm by integrating consistency constraints to address class imbalances. Our contribution extends to the application of sophisticated analysis techniques, including histogram and boxplot analyses, feature distribution assessments, statistical explorations, correlation evaluations, feature importance rankings, and Individual Conditional Expectation (ICE) plots. These methodologies are designed to provide valuable insights into feature significance, thereby assisting researchers in identifying the most critical attributes for effective stroke detection.**

***Index Terms*—Stroke prediction, Balancing Algorithm, Data Analysis, Feature importance**

## I. Introduction

The evolution of digital technology has revolutionized the healthcare field. This significant advancement has become a necessity in our modern society for several reasons. First, the increasing aging population and the prevalence of chronic diseases demand more efficient and accessible healthcare systems. Second, the rise of the internet and connected devices has created a demand for remote health solutions, allowing for constant monitoring and unprecedented convenience for patients, particularly in rural or underserved areas.

E-health applications play a crucial role in detecting and managing a variety of serious diseases. Among the targeted conditions are diabetes [1], heart disease [2], mental health disorders [3], chronic respiratory diseases [4], stroke [5], [6], and neurodegenerative disorders [7], [8], [9], [10], [11].

Stroke remains a significant global health challenge, imposing a heavy burden in terms of morbidity and mortality [12]. Accurate prediction of stroke is crucial for the effectiveness of healthcare interventions and preventive measures. However, the development of machine learning models for stroke prediction faces significant challenges, including class imbalances, noise, and inherent biases in medical datasets [13]. To overcome these challenges, our paper introduces a preprocessing pipeline for the creation of the 'PredictStr' dataset derived from the well-established [14].

The key of our pipeline is the introduction of a novel variant of the ADASYN algorithm (Adaptive Synthetic Sampling) [15], called the Dynamic Hybrid Balancing Algorithm. This advanced mechanism is designed to maintain balance within the dataset, preserve its original distribution, and improve predictive accuracy.

Our contribution extends beyond simple dataset balancing; it includes a comprehensive dataset preparation protocol. The protocol encompasses a detailed description, data cleaning, preprocessing, quality assurance, and balancing using the Dynamic Hybrid Balancing Algorithm. Recognizing the scarcity of suitable datasets for stroke prediction, our study introduces the 'PredictStr' dataset as a valuable resource to address this deficiency.
Following its preparation, the dataset undergoes a thorough analysis with advanced techniques such as histogram and boxplot analyses [16], assessments of feature distribution, statistical explorations, correlation evaluations [17], importance feature rankings [18], and Individual Conditional Expectation (ICE) plots [19]. These advanced methodologies help to decode the complex dynamics of the dataset, clarify the relevance of specific features, and provide essential insights for enhancing stroke prediction models. With the enriched 'PredictStr' dataset and robust analytical tools, healthcare professionals are better equipped to make informed decisions, thereby improving stroke prevention strategies and tailoring patient care.

The structure of the paper is as follows: Section II discusses the background and related literature. Section III describes the dataset preparation including data preprocessing and balancing. Section IV presents a detailed analysis of the obtained dataset in terms of statistics and most important features. Finally, section V concludes the paper and suggests directions for future research.

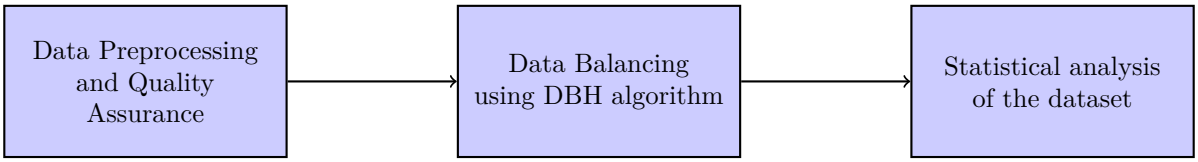


Fig. 1: Flowchart of our proposed approach

## II. Background and Literature Review

Stroke represents a major global health concern, leading to severe disabilities and death. Its substantial socioeconomic impacts underscore the urgent need for precise prediction and effective prevention strategies. Accurate stroke prediction not only enhances patient outcomes but also optimizes resource allocation, making innovative approaches essential for advancing stroke management.

In the realm of medical databases, notable repositories such as the Biobank [1], MIMIC-III Database [20], NHANES [2], and the Framingham Heart Study [21] have made significant contributions to medical research. Although these databases provide a broad spectrum of health-related information, their focus on stroke prediction is often limited. In contrast, the Stroke Prediction Dataset [14] is specifically designed for this purpose. This dataset encompasses a range of clinical attributes including age, hypertension status, presence of heart disease, and Body Mass Index (BMI), providing a solid foundation for developing robust predictive models.

The Stroke Prediction Dataset contains 5110 observations, each with 12 attributes; these include 11 clinical indicators and one target variable. The dataset offers valuable insights into stroke risk factors and predictive indicators. However, it also presents challenges such as significant class imbalance with stroke instances greatly outnumbered by non-stroke instances and missing values, especially in critical attributes like BMI. These issues may skew model performance and accuracy.

Addressing these challenges is crucial. Common techniques to manage class imbalance in predictive modeling include oversampling methods such as the Synthetic Minority Over-sampling Technique (SMOTE) [22] and Adaptive Synthetic Sampling (ADASYN) [15], and undersampling methods like Random Undersampling [23]. These methods aim to balance class distribution and improve model performance by better representing minority class patterns. However, care must be taken as imbalanced datasets can still bias models toward the majority class, reducing their practical utility. Developing strategies to effectively handle these imbalances is vital for enhancing the accuracy of stroke prediction.

[1]https://www.ukbiobank.ac.uk/

[2], https://www.cdc.gov/nchs/nhanes/index.htm

## III. Dataset Preparation

The proposed 'PredictStr' dataset, an improved version of the identified dataset from the literature, is a meticulously balanced dataset designed specifically for stroke prediction. It comprises 9,668 observations, each corresponding to a unique patient. This dataset includes 11 attributes, covering both predictive features and the target attribute (Table I). Among these, 10 features capture essential health-related variables, while one attribute indicates the likelihood of stroke occurrence. In this section, we present the different steps of data preparation including data preprocessing and balancing (Figure 1).

### *A. Preprocessing and Quality Assurance*

In the data preparation process, our initial step involved addressing data inconsistencies by checking for mismatched and missing values. We conducted data cleaning procedures, starting with the removal of the 'ID' column, which held no relevance for classification. Furthermore, we adjusted column names, renaming 'Ever_married' to 'Marital_Status' and 'avg_glucose_level' to 'Average_Glucose_Level'. The dataset contained 'N/A' values for the BMI attribute, 'unknown' values in the 'smoking_status' attribute, and 'other' values in the 'Gender' attribute. To clean the dataset, we implemented the following data cleaning actions:

1) Elimination of rows with the gender specified as 'Other': this affected one row, specifically row number 3117.
2) Replacement of 'N/A' values in the BMI column: we substituted these with the median BMI value of 28.1, impacting 201 rows.
3) Encoding of categorical values: we converted categorical variables such as gender, marital status, work type, residence type, and smoking status into numerical representations to facilitate analysis, simplifying subsequent analysis and modeling.
4) Substitution of 'unknown' values in the 'smoking_status' column: we replaced these with the median value of 0, affecting 1544 rows.

The decision to replace 'unknown' values with the median was due to the fact that removing (201 + 1544) instances would result in significant data loss. With 201 rows having missing BMI values out of a total of 5110, opting for median imputation rather than removal was motivated by the fact that the median minimizes the influence of outliers on imputation..

TABLE I: Summary of *PredictStr* Dataset Attributes.

| **Attribute** | **Values/Specification** | **Description** |
|---|---|---|
| Gender | 0: Male, 1: Female | Gender |
| Age | Numeric Value (year) | Age |
| Hypertension | 0: No, 1: Yes | Hypertension status |
| Heart_Disease | 0: No, 1: Yes | History of heart disease |
| Marital_Status | 0: Not Married, 1: Married | Marital status |
| Residence_Type | 0: Urban, 1: Rural | Type of residence |
| Work_Type | 0: Private, 1: Self-Employed, 2: Govt Job, 3: Children | Type of employment |
| Average_Glucose_Level | Numeric Value (mg/dl) | Average glucose level in the blood after meal |
| BMI | Numeric Value (kg/m$^2$) | Body Mass Index |
| Smoking_Status | 0: Never Smoked, 1: Formerly Smoked, 2: Smokes | Smoking status |
| Stroke_Prediction | 0: No Stroke, 1: Stroke | Prediction of stroke occurrence |

TABLE II: Statistical Overview of Top Five Important Features in the PredictStr Dataset

| Statistic | Average_Glucose | BMI | Age | Work_Type | Smoking_Status |
|---|---|---|---|---|---|
| Mean | 106.14 | 28.86 | 43.22 | 0.83 | 0.48 |
| Std Dev | 45.28 | 7.69 | 22.61 | 1.11 | 0.74 |
| Min | 55.12 | 10.30 | 0.08 | 0 | 0 |
| 25% | 77.24 | 23.80 | 25 | 0 | 0 |
| 50% | 91.88 | 28.10 | 45 | 0 | 0 |
| 75% | 114.09 | 32.80 | 61 | 2 | 1 |
| Max | 271.74 | 97.60 | 82 | 4 | 2 |

### *B. Balancing Using Dynamic Hybrid Balancing Algorithm*

In this step, we addressed the class imbalance challenge within the preprocessed dataset. The original dataset exhibited a notable disparity, containing 249 instances of the stroke class and 4,861 instances of the no-stroke class. To rectify this imbalance, we applied our novel Dynamic Hybrid Balancing Algorithm (DBH) with meticulous care to ensure that the generation of data adhered to the specifications and constraints of our medical attributes.

Our DBH algorithm involved formulating a set of constraints for both encoded categorical attributes (e.g., gender, hypertension, heart disease, marital status) and quantitative attributes (BMI and Average Glucose Level). This careful consideration ensured that the generated synthetic values maintained clinical validity, promoting accuracy and relevance. Specifically, constraints were applied to keep the Average Glucose Level within the original data range of 55 to 272 and BMI within the range of 14 to 98. Additionally, the 'age' attribute was constrained within the clinically relevant range of 1 to 100. This holistic constraint application guaranteed that the synthetic data not only preserved clinical accuracy but also remained meaningful in a medical context.

Our DBH algorithm incorporates ADASYN and the Louvain Modularity algorithm [24], dynamically generating synthetic samples within these constraints. Unlike static methods, it adjusts sampling ratios in real-time to maintain class balance, adapting to evolving dataset needs. ADASYN's adaptiveness enhances efficacy, ensuring balanced representation. Utilizing a graph representation via K-nearest neighbors [25], the Louvain Modularity algorithm identifies intricate dataset structures, enriching dataset composition. Integration of constraints for categorical and quantitative attributes ensures clinical relevance.

As a result, the balanced dataset comprises 4,834 instances per class, providing a solid foundation for subsequent analyses. This algorithmic fusion not only balances the data but also refines the dataset structure, fostering more accurate stroke prediction. The advantages lie in its dynamic adaptability, structural refinement, and clinical relevance, enhancing predictive performance.

## IV. Experimental Results and Discussion

This section presents a detailed analysis of the experimental results and discussions based on the 'PredictStr' dataset. It explores class distribution, correlations between attributes, and the significance of features, providing insights into the characteristics of the dataset and comparisons with previous studies. The discussions emphasize the

dataset's potential and its relevance to stroke prediction research.

### *A. Dataset Distribution and Class Balance*

After preprocessing and applying our Dynamic Hybrid Balancing algorithm, the 'PredictStr' dataset achieves a balanced distribution with 4,834 instances per class. This dataset, formatted as a CSV file, is now readily available for researchers aiming to develop stroke prediction models. Researchers wishing to use this database should refer to the original database on which we have based this new and improved version. To facilitate this, we have divided the dataset into two subsets: 80% for training and 20% for testing. This split ensures adequate representation of both classes in each subset through stratification. These openly accessible datasets provide a robust foundation for predictive modeling efforts. This splitting of the dataset allows researchers to benchmark their AI algorithms and compare the performance of their algorithms on the same data split.

### *B. Evaluation of the 'PredictStr' dataset's quality and characteristics*

In this section, we present the results of our comprehensive evaluation, highlighting the critical aspects of our balanced and enhanced dataset. We begin our exploration by examining the significance of key attributes, as identified through importance ranking analysis. Aiming to provide a comprehensive understanding while effectively managing the extensive list of attributes, we concentrate on analyzing the most influential features. Statistical overview of the top five important features is presented in Table II. This focused approach enables us to reveal nuanced insights that enhance the accuracy of stroke prediction, aligning with our overarching goal of improving healthcare outcomes.

Table III showcases the key attributes for stroke prediction in the 'PredictStr' dataset. Particularly, 'Average_Glucose_Level' and 'BMI' are significant, each holding importance scores of approximately 26.30%, while 'Age' follows closely with a score of 22.40%. Additionally, 'Work_Type' and 'Smoking_Status' each contribute significantly with scores of 7.00%. These insights offer valuable guidance for understanding the dynamics of stroke prediction.

TABLE III: Top Five Most Important Features and Their Importance Scores

| **Feature** | **Importance Score (%)** |
|---|---|
| Average Glucose Level | 26.29 |
| BMI | 26.28 |
| Age | 22.40 |
| Work Type | 07.00 |
| Smoking_Status | 07.00 |

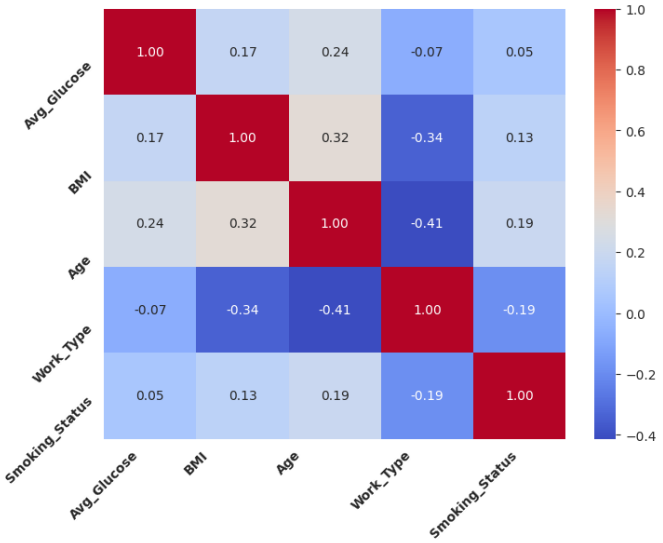


Fig. 2: Correlation Matrix of Top Five Features in the 'PredictStr' Dataset.

Table II outlines key features from our stroke dataset. The top five most important features provide crucial insights into the dataset's demographic and health-related characteristics. Notably, the table highlights the variability across features, as evidenced by their standard deviation values. For instance, the mean age is 43.22 years, with a standard deviation of 22.61 years, indicating a wide range of ages. Similarly, BMI values vary from 10.30 to 97.60, underscoring the diversity in body mass indices. Overall, this statistical table provides a comprehensive snapshot of the dataset's composition and the variability of its key features.

The correlation matrix in Figure 2 reveals significant correlations among the top five features. BMI and Age exhibit a moderate positive correlation (0.32), indicating a tendency for higher BMI values with older age. However, Smoking Status shows a strong negative correlation with both BMI and Age (-0.34 and -0.41, respectively), suggesting non-smokers are more prevalent among individuals with higher BMI and older age. Additionally, Smoking Status and Work Type demonstrate a moderately negative correlation (-0.19), implying non-smokers are more common in certain work types. These correlations offer valuable insights into feature relationships, enhancing dataset interpretation.

### *C. Importance Ranking and ICE Plots using Random Forest classifier model*

In this section, we use the Random Forest classifier [26] to conduct importance ranking and Individual Conditional Expectation (ICE) plot analyses (Figure 3). These methods provide invaluable insights into the significance of each feature in predicting stroke outcomes. The Random Forest classifier serves as a tool for evaluating the relative impor-

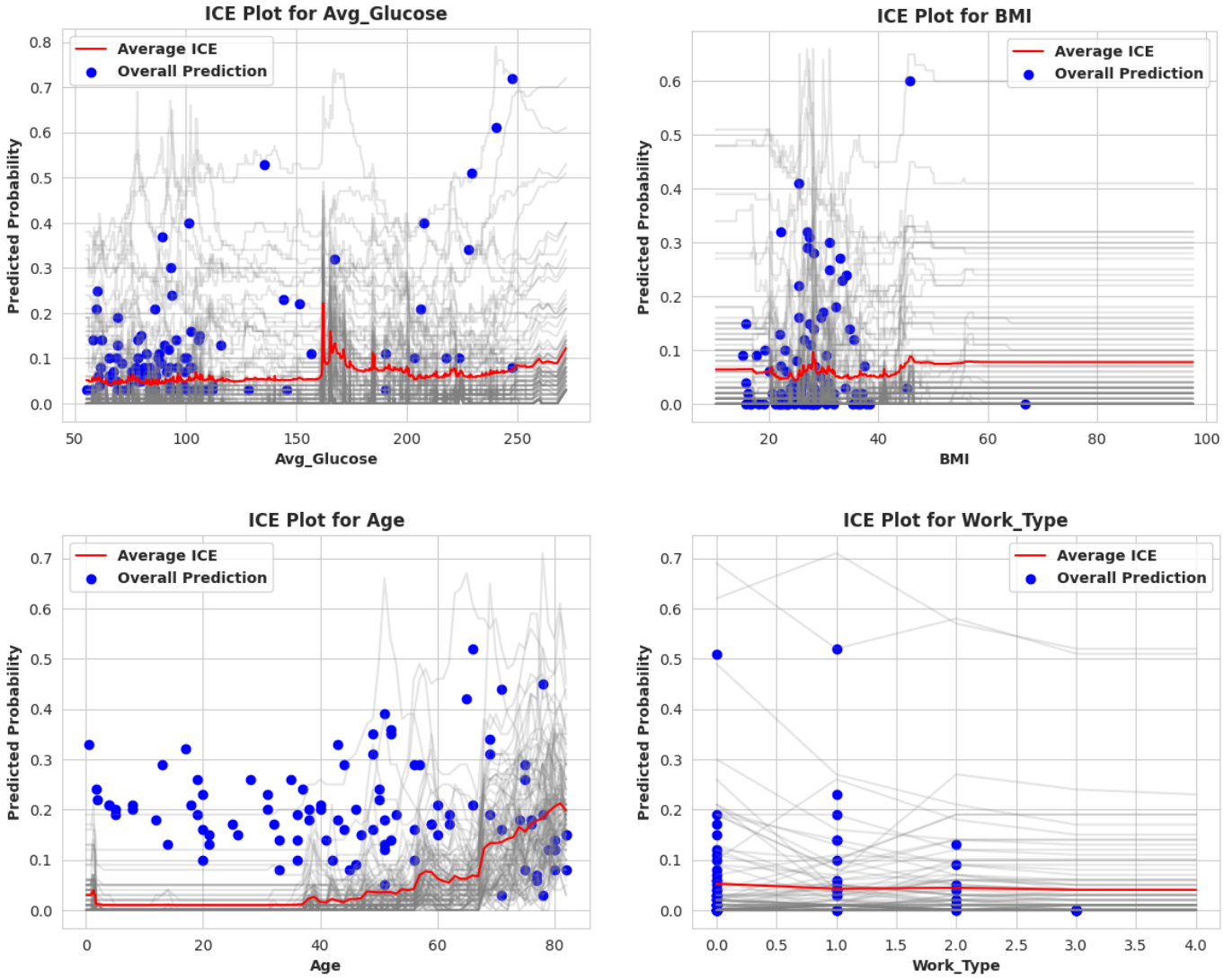


Fig. 3: ICE plots of the fourth most important features

tance of different attributes in our dataset. Additionally, our use of ICE plots introduces a unique visualization technique that reveals how individual features influence predictions, taking into account the values of other attributes. This dual approach significantly enhances our understanding of how attributes contribute to predictions and their nuanced effects on stroke prediction.

The importance ranking illuminates the overall contribution of each feature to the predictive accuracy of our model. At the same time, ICE plots offer a detailed perspective, enabling us to examine how specific feature values affect predictions in various scenarios. These comprehensive insights derived from the Random Forest classifier's analyses serve as a baseline for guiding the development of future stroke prediction models.

The Table IV presents the highest importance scores for the top five features identified in our dataset. These scores, measured using the ICE approach, quantify the relative importance of each feature in predicting the target variable. These values highlight the significance of each feature in influencing the target outcome.

The ICE values offer a numerical measure of feature importance, which aids in prioritizing variables for predictive modeling and feature selection processes. Understanding the relative importance of these features can guide further analysis and decision-making, ultimately contributing to the development of more accurate and reliable predictive models for the target variable.

## V. Conclusions

The paper introduces the 'PredictStr' dataset, a carefully designed and balanced dataset specifically for stroke prediction research, introducing a novel Dynamic Hybrid Balancing Algorithm. Through detailed analysis tech-

TABLE IV: Highest Average ICE values for the 5 most important attributes.

| Attribute | Average ICE Value |
|---|---|
| Average Glucose Level | 0.5233 |
| BMI | 0.4912 |
| Age | 0.5156 |
| Work Type | 0.5261 |
| Smoking_Status | 0.4814 |

niques like importance ranking, box plot analysis, correlation examination, and ICE plots, the study reveals intricate patterns in stroke prediction, underscoring the unique and high-quality nature of the dataset focused on this specific medical issue. This focus, combined with the innovative balancing algorithm, lays a solid foundation for future research to develop more precise and effective stroke prediction models. The importance of timely stroke detection and prevention highlights the potential of the 'PredictStr' dataset to significantly advance medical research and improve patient outcomes.

Future research directions include expanding the dataset's applicability, integrating diverse data types, and exploring cutting-edge machine learning models. Collaboration with healthcare professionals and the adoption of real-time monitoring could translate these findings into practical applications, enhancing proactive health interventions and patient care.

## References


[1] Serena Zanelli, Mounim A El Yacoubi, Magid Hallab, and Mehdi Ammi. Type 2 diabetes detection with light cnn from single raw ppg wave. IEEE Access, 2023.

[2] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, et al. Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience, 2021, 2021.

[3] Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. Deep learning in mental health outcome research: a scoping review. Translational Psychiatry, 10(1):116, 2020.

[4] Chengdi Wang, Jiechao Ma, Shu Zhang, Jun Shao, Yanyan Wang, Hong-Yu Zhou, Lujia Song, Jie Zheng, Yizhou Yu, and Weimin Li. Development and validation of an abnormality-derived deep-learning diagnostic system for major respiratory diseases. NPJ Digital Medicine, 5(1):124, 2022.

[5] Jialin Luo, Peishan Dai, Zhuang He, Zhongchao Huang, Shenghui Liao, and Kun Liu. Deep learning models for ischemic stroke lesion segmentation in medical images: A survey. Computers in Biology and Medicine, page 108509, 2024.

[6] Siqi Zhang, Miao Zhang, Shuai Ma, Qingyong Wang, Youyang Qu, Zhongren Sun, Tiansong Yang, et al. Research progress of deep learning in the diagnosis and prevention of stroke. BioMed Research International, 2021, 2021.

[7] Karim Haddada, Mohamed Ibn Khedher, and Olfa Jemai. Comparative study of deep learning architectures for early alzheimer detection. In 2023 International Conference on Cyberworlds (CW), pages 185–192. IEEE, 2023.

[8] Christian Kahindo, Mounim El Yacoubi, Sonia Garcia-Salicetti, Anne-Sophie Rigaud, and Victoria Cristancho-Lacroix. Characterizing early-stage alzheimer through spatiotemporal dynamics of handwriting. IEEE Signal Processing Letters, PP:1–1, 01 2018.

[9] Mounim A. El-Yacoubi, Sonia Garcia-Salicetti, Christian Kahindo, Anne-Sophie Rigaud, and Victoria Cristancho-Lacroix. From aging to early-stage alzheimer's: Uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning. Pattern Recognition, 86, 2019.

[10] S. Mirzaei, M. El Yacoubi, S. Garcia-Salicetti, J. Boudy, C. Kahindo, V. Cristancho-Lacroix, H. KerhervÃ©, and A.-S. Rigaud. Two-stage feature selection of voice parameters for early alzheimerâs disease prediction. IRBM, 39(6):430â435, December 2018.

[11] Holger Frohlich, Noemi Bontridder, Dijana Petrovska-Delacreta, Enrico Glaab, Felix Kluge, Mounim El Yacoubi, Mayca Marin Valero, Jean-Christophe Corvol, Bjoern Eskofier, Jean-Marc Van Gyseghem, Stephane Lehericy, Jurgen Winkler, and Jochen Klucken. Leveraging the potential of digital technology for better individualized treatment of parkinsonâs disease. Frontiers in Neurology, 13, February 2022.

[12] Diji Kuriakose and Zhicheng Xiao. Pathophysiology and treatment of stroke: present status and future perspectives. International journal of molecular sciences, 21(20):7609, 2020.

[13] Haibo He and Edwardo A Garcia. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9):1263–1284, 2009.

[14] Stroke prediction dataset. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset, 2021.

[15] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), pages 1322–1328. Ieee, 2008.

[16] David F Williamson, Robert A Parker, and Juliette S Kendrick. The box plot: a simple visual method to interpret data. Annals of internal medicine, 110(11):916–921, 1989.

[17] Shixia Liu, Weiwei Cui, Yingcai Wu, and Mengchen Liu. A survey on information visualization: recent advances and challenges. The Visual Computer, 30:1373–1393, 2014.

[18] Maksymilian Wojtas and Ke Chen. Feature importance ranking for deep learning. Advances in neural information processing systems, 33:5105–5114, 2020.

[19] Vanessa Buhrmester, David Münch, and Michael Arens. Analysis of explainers of black box deep neural networks for computer vision: A survey. Machine Learning and Knowledge Extraction, 3(4):966–989, 2021.

[20] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. Scientific data, 3(1):1–9, 2016.

[21] Syed S Mahmood, Daniel Levy, Ramachandran S Vasan, and Thomas J Wang. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. The lancet, 383(9921):999–1008, 2014.

[22] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

[23] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In Icml, volume 97, page 179. Citeseer, 1997.

[24] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10):P10008, 2008.

[25] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1):21–27, 1967.

[26] Tin Kam Ho. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition, volume 1, pages 278–282. IEEE, 1995.