Brain Stroke Prediction:
Using K-Nearest Neighbors

LaTeX template adapted from:
European Conference on Artificial Intelligence

Hamza Ali Khan[1]

Other group members:
Mohammad Ayaan Khan[2], **Hidayatullah Gul**[3], **Shaik Shoeb Hussain**[4]

Abstract. Stroke is a major global health concern. This report uses the KNN algorithm for stroke prediction with healthcare data. Preprocessing involved KNN imputation, SMOTE, and standardization. The initial model (94% accuracy) was biased, but optimization achieved balanced accuracy (91%), showcasing KNN's suitability for healthcare applications.

1 Introduction

Stroke is a leading global cause of death and long-term disability, affecting millions each year. Early prediction of stroke risk can improve outcomes through timely interventions. Traditional methods, relying on clinical expertise and statistical models, often fail to capture complex relationships between health and lifestyle factors [3].

Machine learning has shown promise in analyzing healthcare data to predict outcomes. This study uses the KNN algorithm, a non-parametric model that classifies data points based on their k-nearest neighbors using distance metrics such as Euclidean or Manhattan. Its simplicity and interpretability make it ideal for healthcare applications.

This report details the development of the KNN model, including preprocessing, model training, hyperparameter optimization, and evaluation, emphasizing the importance of balancing and tuning.

2 Background

This section discusses prior research on machine learning techniques used for stroke prediction, highlighting their advantages, limitations, and relevance to this study.

Machine learning has become a pivotal tool in healthcare, particularly for disease prediction and risk assessment. Stroke prediction, a critical healthcare challenge, has been extensively studied using various algorithms.

Support Vector Machines (SVM) are widely employed due to their robustness and ability to handle non-linear relationships. For instance, Liaqat et al. [5] applied SVM to a stroke dataset, achieving high accuracy. However, computational cost posed challenges for larger datasets.

Random Forest has also been explored for stroke prediction. Ahmed et al. [1] demonstrated its effectiveness in handling imbalanced healthcare datasets, achieving higher precision and recall compared to traditional methods. Neural Networks, such as Convolutional Neural Networks (CNNs), have shown superior performance in early stroke detection, as noted by Wang et al. [8]. However, their black-box nature raises interpretability concerns in clinical applications.

Traditional methods like logistic regression and decision trees are interpretable but often fail to capture complex relationships, as noted by Chen and Zhang [2].

In contrast, KNN offers simplicity and interpretability while performing competitively in healthcare applications. Studies like Sharma et al. [7] highlight KNN's adaptability when combined with techniques like SMOTE, addressing class imbalance issues. This study builds on these findings by applying KNN to a real-world stroke dataset, addressing challenges such as missing values and class imbalance.

3 Experiments and results

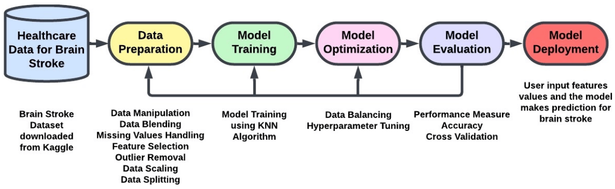


Figure 1. Model Architecture.

3.1 Data Import and Preprocessing

The dataset used for this study was sourced from Kaggle [4], which contains 12 features and 5110 samples. To prepare the data for model training, several preprocessing steps were performed. Missing values in the BMI column were filled using KNN imputation, which estimates missing values by leveraging the similarity between

[1] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: hk2108k@gre.ac.uk

[2] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: mk1693i@gre.ac.uk

[3] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: hg0915l@gre.ac.uk

[4] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: ss2221h@gre.ac.uk

data points. The patient_id column was removed as part of feature selection, as it doesn't contribute to the prediction task. Exploratory Data Analysis was conducted to identify outliers in the average_glucose_level, BMI, and age columns. These outliers were retained because they represent natural variations relevant to stroke risk and are important for making accurate predictions.

Categorical features, including gender and work type, were encoded into numerical values using Label Encoding to ensure compatibility with the machine learning model. To address the significant class imbalance in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied, generating synthetic samples for the minority class (stroke cases). Finally, the data was splittted into training(80%) and testing(20%) subsets, and features were standardized using StandardScaler to normalize the scales of all variables, ensuring that the distance-based KNN algorithm could perform effectively.

3.2 Initial Model Training

The KNN model was initially trained using hyperparameters determined through trial and error:

- n_neighbors = 3
- weights = 'uniform'
- metric = 'euclidean'
- algorithm = 'kd_tree'

The Euclidean distance metric is defined as:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (1)$$

Figure 2 illustrates how Euclidean distance is calculated in a 2D feature space.

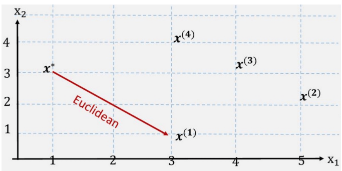


Figure 2. Illustration of Euclidean distance calculation. Adapted from lecture slides by Oroojeni [6].

The model achieved an accuracy of 94% on the test data. However, the classification report revealed significant bias, with the majority class (no stroke) achieving high precision and recall, while the minority class (stroke) showed poor performance:

[h]

Metric	Class 0 (No Stroke)	Class 1 (Stroke)
Precision	0.95	0.09
Recall	0.99	0.02
F1-score	0.97	0.03

Table 1. Initial Model Performance Metrics

This highlighted the need for addressing class imbalance and optimizing hyperparameters.

3.3 Model Optimization and Final Evaluation

To address bias, the dataset was balanced using SMOTE, and Grid Search was employed to optimize hyperparameters which tries every combination of the provided hyper-parameter values in order to find the best model. The best parameters identified were:

- n_neighbors = 3
- weights = 'distance'
- metric = 'manhattan'
- algorithm = 'auto'

The Manhattan distance metric is defined as:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \qquad (2)$$

Figure 3 illustrates how Manhattan distance is calculated in a 2D feature space.

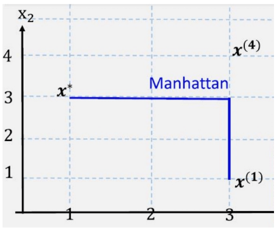


Figure 3. Illustration of Manhattan distance calculation. Adapted from lecture slides by Oroojeni [6].

The final model achieved an accuracy of 91%, with balanced performance across both classes:

[h]

Metric	Class 0 (No Stroke)	Class 1 (Stroke)
Precision	0.97	0.86
Recall	0.85	0.97
F1-score	0.90	0.91

Table 2. Final Model Performance Metrics

3.4 Cross-Validation and Loss

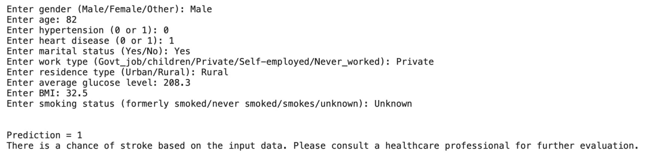
The model's robustness was validated using 5-fold cross-validation, resulting in an average accuracy of 83.55%. Additionally, binary cross-entropy loss was computed, confirming the model's reliability with a loss value of 2.0030.

3.5 Deployment

The final model was deployed to predict the likelihood of stroke based on user-provided input features. Figure 4 shows an example of the deployment, where a user enters relevant data (e.g., age, BMI, glucose level), and the model outputs the prediction and advises appropriate action. And this practical application demonstrates the feasibility of integrating KNN into healthcare decision support systems.

3.6 Team-member's Models Comparison

As this was a project so the group tested KNN(used by myself), SVM, Neural Network, and Random Forest for stroke prediction. Despite Random Forest's highest accuracy (95%), it was biased. SVM and Neural Network (92% accuracy) required longer training times. KNN achieved 91% balanced accuracy, with robust metrics, efficiency, simplicity, get trained quickly making it the optimal choice.

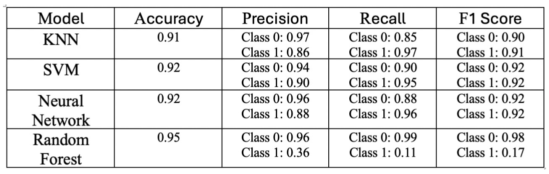


```
Enter gender (Male/Female/Other): Male
Enter age: 82
Enter hypertension (0 or 1): 0
Enter heart disease (0 or 1): 1
Enter marital status (Yes/No): Yes
Enter work type (Govt_job/children/Private/Self-employed/Never_worked): Private
Enter residence type (Urban/Rural): Rural
Enter average glucose level: 208.3
Enter BMI: 32.5
Enter smoking status (formerly smoked/never smoked/smokes/unknown): Unknown


Prediction = 1
There is a chance of stroke based on the input data. Please consult a healthcare professional for further evaluation.
```

Figure 4. Model Deployment: User input for stroke prediction and the corresponding output. The model predicts the likelihood of stroke and advises consulting a healthcare professional for further evaluation.

For real-world applications, where resources like time and computational power are limited, KNN provides a practical and effective solution for the chosen dataset. The evaluation metrics for each model are summarized in Figure 5.



Model	Accuracy	Precision	Recall	F1 Score
KNN	0.91	Class 0: 0.97 Class 1: 0.86	Class 0: 0.85 Class 1: 0.97	Class 0: 0.90 Class 1: 0.91
SVM	0.92	Class 0: 0.94 Class 1: 0.90	Class 0: 0.90 Class 1: 0.95	Class 0: 0.92 Class 1: 0.92
Neural Network	0.92	Class 0: 0.96 Class 1: 0.88	Class 0: 0.88 Class 1: 0.96	Class 0: 0.92 Class 1: 0.92
Random Forest	0.95	Class 0: 0.96 Class 1: 0.36	Class 0: 0.99 Class 1: 0.11	Class 0: 0.98 Class 1: 0.17

Figure 5. Comparison of model performance metrics for stroke prediction.

4 Discussion

The KNN algorithm demonstrated its effectiveness and interpretability for stroke prediction. Preprocessing steps such as KNN imputation for missing values and SMOTE for balancing classes addressed key challenges often encountered in healthcare datasets. Initially, the unbalanced model achieved a high accuracy of 94% but was biased towards the majority class, emphasizing the importance of addressing class imbalance to reduce false negatives, which can have severe clinical consequences.

After balancing the dataset and tuning hyperparameters using Grid Search, the final model achieved a balanced accuracy of 91%. Precision, recall, and F1-scores indicated improved fairness across both classes, justifying the slight reduction in overall accuracy. This highlights that building an effective model goes beyond accuracy alone, requiring attention to fairness and robustness. KNN's simplicity and interpretability further enhance its suitability for healthcare applications, particularly for clinical decision-making.

Limitations include KNN's sensitivity to hyperparameters and increased computational complexity with larger datasets. Additionally, testing on larger, more diverse datasets is necessary to validate the model's generalizability.

5 Conclusion and future work

This study successfully applied KNN to predict stroke risk using a healthcare dataset, demonstrating its potential as an interpretable and effective classification algorithm. Preprocessing steps, such as SMOTE and Grid Search, ensured balanced and robust performance with a final balanced accuracy of 91% and robust performance for both classes. These findings affirm KNN's relevance for healthcare applications, especially in settings where transparency and simplicity are essential.

Future Works

Future improvements include testing on larger datasets to enhance generalizability, exploring hybrid methods for scalability, and integrating real-time applications for continuous stroke risk monitoring. Additionally, incorporating explainability techniques like SHAP and LIME tailored for KNN could increase its clinical acceptability, bridging the gap between research and practical healthcare use.

ACKNOWLEDGEMENTS

I would like to thank my professor and tutors for their guidance and support throughout this project

REFERENCES

[1] Sara Ahmed and Naveed Younas, 'Random forest classifier for imbalanced healthcare datasets: An analysis', *Healthcare Informatics*, **18**, 45–54, (2020).

[2] Xiaoyu Chen and Hao Zhang, 'A comparative study of logistic regression and decision trees in healthcare applications', *Journal of Medical Informatics*, **15**, 123–134, (2018).

[3] Andre Esteva, Alexis Robicquet, and Bharath Ramsundar, 'A guide to deep learning in healthcare', *Nature Medicine*, **25**, 24–29, (2019).

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

[5] Muhammad Liaqat and Naveed Ahmad, 'Support vector machines for stroke prediction: A review', *Journal of Stroke Research*, **12**, 341–352, (2021).

[6] Hooman Oroojeni. Lecture slides: Introduction to neural networks. School of Computing and Mathematical Sciences, University of Greenwich, 2024. Accessed during lecture [Introduction to AI (COMP1827)].

[7] Anjali Sharma and Rajesh Patel, 'K-nearest neighbors for stroke prediction: A case study', *Health Informatics Journal*, **28**, 67–82, (2022).

[8] Li Wang and Hao Xu, 'Neural networks in healthcare: A case study on stroke prediction', *Computational Medicine*, **22**, 123–134, (2019).