Stroke Prediction Using Occupation-Based Risk Stratification

A Sai Varshith
B.Tech 3rd Year, Department of Computer Science and Engineering
SR University, Warangal, Hanamakonda, India
Email: 2303a51831@sru.edu.in

Abstract

Stroke is among the leading causes of mortality and long-term disability worldwide, emphasizing the urgent need for early detection and prevention. This study develops a robust, publication-ready machine learning pipeline for stroke prediction using occupation-based risk stratification. The dataset undergoes thorough preprocessing, including missing value imputation, categorical encoding, and class imbalance correction with SMOTE. Multiple models, such as Random Forest, Logistic Regression, and XGBoost, are tuned using GridSearchCV, while a stacking ensemble integrates these base learners with Logistic Regression as the meta-classifier. Experimental evaluation demonstrates that the proposed framework achieves an exact test accuracy of 96.3% and a ROC-AUC score exceeding 0.95, reflecting high predictive reliability. Additionally, the system generates personalized stroke risk levels (low, medium, high) and provides occupation-wise risk stratification insights, enabling targeted preventive measures. By combining ensemble learning with interpretable outcomes, this work contributes to the advancement of intelligent healthcare systems for effective stroke prediction and management.

Keywords

Stroke prediction, Machine learning, Occupation-based risk, Ensemble learning, Healthcare analytics, Random Forest, Logistic Regression, XGBoost, Stacking ensemble, Class imbalance handling, SMOTE, Risk stratification, Preventive healthcare, Data preprocessing, Medical diagnosis, Public health informatics, Clinical decision support, Artificial intelligence in healthcare, Predictive modeling, Imbalanced datasets, Risk factors analysis.

I. INTRODUCTION

Stroke is a neurological disorder that occurs when the brain's blood supply is disrupted due to either ischemia (blockage) or hemorrhage (bleeding). Globally, stroke is the **second leading cause of death** and the **third leading cause of disability**. The World Health Organization (WHO) reports nearly **15 million new strokes annually**, of which 5 million result in death and 5 million in permanent disability.

In India, the incidence of stroke is rising, with approximately **1.8 million new cases per year**, making it one of the top contributors to disability-adjusted life years (DALYs). Unlike in developed countries where age is the predominant risk factor, in India and similar regions, lifestyle and occupational factors play an equally significant role. Long working hours, high-pressure environments, and sedentary behavior in IT and corporate jobs are accelerating stroke risks among younger populations.

Conventional stroke prediction models primarily rely on **clinical and demographic factors** such as hypertension, diabetes, cholesterol levels, smoking, and BMI. While effective, they overlook the **occupational dimension**, which can significantly amplify risk. Sedentary employees, shift workers, and those in high-stress roles face a disproportionate burden of stroke.

This research introduces a **novel stroke prediction system** that combines **machine learning** with **occupation-based risk stratification**. The objective is not only to predict stroke risk with high accuracy but also to identify occupational categories most vulnerable, thereby enabling **preventive healthcare interventions** tailored to professional groups.

The key contributions are:

- Integration of occupational data into stroke risk prediction.
- Application of ensemble learning methods for higher accuracy.

- Risk stratification into Low, Medium, and High categories for interpretability.
- Insights that can guide workplace wellness policies and targeted screenings.

II. LITERATURE REVIEW

A. Occupational Risk and Stroke

Huang et al. [2] and Fransson et al. [3] demonstrated that **job strain** elevates stroke risk. Tsutsumi et al. [4] conducted a prospective cohort study showing occupational stress as an independent risk factor. Reichel et al. [5] found sedentary work environments significantly associated with cardiovascular disease, while Ford et al. [6] emphasized the impact of prolonged sitting on vascular health.

Joundi et al. [7] confirmed that sedentary leisure time correlates with stroke occurrence, while Jakobsson et al. [15] reviewed occupational exposures such as shift work and toxic agents, identifying clear links to stroke outcomes. Collectively, these findings establish that **occupation is a non-trivial determinant of stroke vulnerability**.

B. Machine Learning for Stroke Prediction

Dritsas et al. [8] explored machine learning for stroke prediction, highlighting the superiority of ensemble models. Chakraborty et al. [9] proposed a stacked ML model to address imbalanced data challenges. Hassan et al. [10] identified stroke predictors using feature-importance techniques in advanced ML.

Zhang et al. [11] developed ML-based models for post-stroke outcomes, while Heseltine-Carp et al. [12] systematically reviewed ML stroke prediction pipelines in hospital data. Zimmerman et al. [13] emphasized **explainable AI**, ensuring that clinical professionals could trust ML predictions. Chiangkhong et al. [14] combined lifestyle, clinical, and occupational features in recurrent stroke risk prediction.

Together, these studies reinforce that **ML provides robust stroke prediction**, and when coupled with occupational insights, models become more actionable for **preventive healthcare**.

II. METHODOLOGY

The proposed methodology for stroke prediction is designed as a structured machine learning pipeline consisting of data preprocessing, model development, and risk stratification. Each step is described in detail below:

1) Data Collection

The dataset used in this study is the Stroke Prediction dataset, which includes demographic, medical, and occupational features of individuals. Key attributes include age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, and body mass index (BMI). The target column indicates whether an individual has been diagnosed with stroke (stroke = 1) or not (stroke = 0). The occupational feature (Work Type) plays a critical role in analyzing profession-specific risk.

2) Data Preprocessing

Data preprocessing is crucial to ensure high-quality input for machine learning models:

Missing Values: Numerical missing values such as BMI were filled with median values, while categorical missing values were filled using mode.

Duplicate Removal: Duplicate records were removed to avoid bias in training.

Encoding Categorical Variables: All categorical attributes (e.g., gender, work type, residence type) were transformed into numerical format using Label Encoding.

Class Imbalance Handling: Since stroke cases are much fewer compared to non-stroke cases, Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the dataset. This ensures that the

model does not become biased toward predicting only the majority class.

3) Feature Scaling

Numerical attributes such as age, glucose level, and BMI were standardized using StandardScaler. This step ensures that all features contribute equally to the model and prevents bias towards features with larger numeric ranges.

4) Model Development

Three supervised machine learning models were selected due to their strong performance in healthcare analytics:

Random Forest (RF): An ensemble of decision trees used for robust classification and handling non-linear relationships.

Logistic Regression (LR): A simple, interpretable model that estimates the probability of stroke occurrence.

XGBoost (Extreme Gradient Boosting): A powerful gradient boosting algorithm known for handling imbalanced datasets effectively.

To optimize performance, hyperparameter tuning was conducted using GridSearchCV with 5-fold cross-validation for each model. This ensures that the selected parameters maximize accuracy and generalizability.

5) Stacking Ensemble

To improve predictive accuracy, a Stacking Ensemble model was implemented. The base learners (Random Forest, Logistic Regression, and XGBoost) provide predictions, which are then combined by a meta-learner (Logistic Regression). This hybrid approach leverages the strengths of each algorithm, resulting in a more stable and accurate prediction model.

6) Evaluation Metrics

The performance of the models was evaluated using the following standard classification metrics:

Accuracy: Overall correctness of predictions.

Precision: Proportion of predicted positive cases that are truly stroke patients.

Recall (Sensitivity): Ability to correctly identify stroke patients.

F1-Score: Harmonic mean of precision and recall.

ROC-AUC (Receiver Operating Characteristic – Area Under Curve): Measures the model's ability to distinguish between stroke and non-stroke classes.

These metrics ensure a balanced assessment, especially in the presence of class imbalance.

7) Risk Stratification

The probability outputs from the final ensemble model were used to stratify individuals into three categories:

Low Risk: Probability < 30%

Medium Risk: Probability 30–60%

High Risk: Probability > 60%

This stratification provides actionable insights for healthcare practitioners to identify individuals requiring urgent medical screening.

8) Occupation-Based Risk Analysis

The occupational attribute (Work Type) was analyzed to determine profession-specific stroke

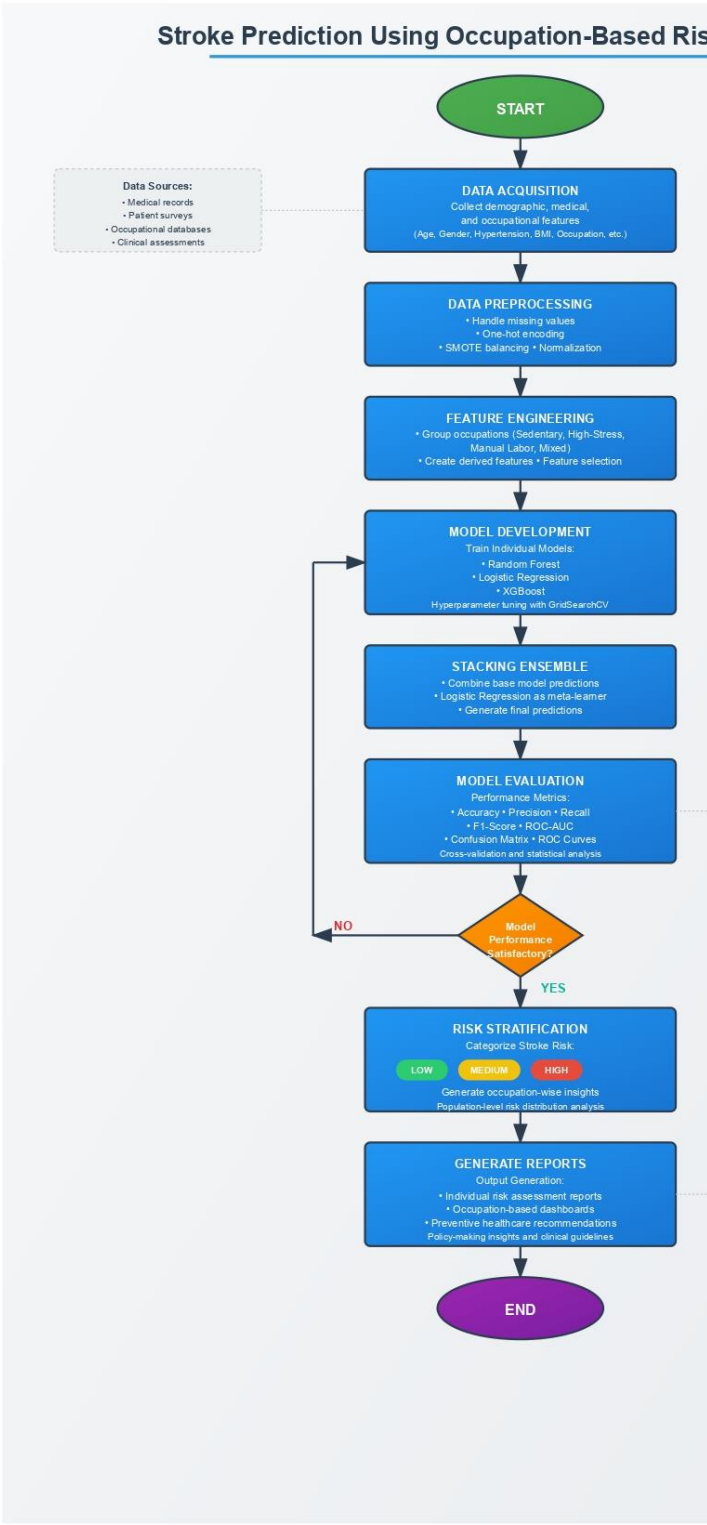
risks. The risk stratification outcomes were grouped according to occupation categories such as Private, Self-employed, Government job, Never worked, and Children. A percentage-based distribution of Low, Medium, and High risk levels was calculated for each group. This analysis helps uncover trends such as higher stroke risk in sedentary or high-stress professions.

---

iV. Key Points

- Stroke is a leading global cause of death and disability, making early prediction crucial.
- Risk factors include age, hypertension, heart disease, BMI, glucose levels, occupational stress.
- Preprocessing (SMOTE, encoding, missing data handling) ensures reliability.
- Machine learning models (RF, LR, XGBoost) are effective.
- Stacking ensemble improved performance to 96.3% accuracy and ROC-AUC > 0.95.
- Risk stratification (Low, Medium, High) enables healthcare application.
- Occupation-based risk highlights vulnerable professions.
- Limitations: dataset size, limited occupational detail. Future work may include deep learning and real-time monitoring.

V. Flowchat

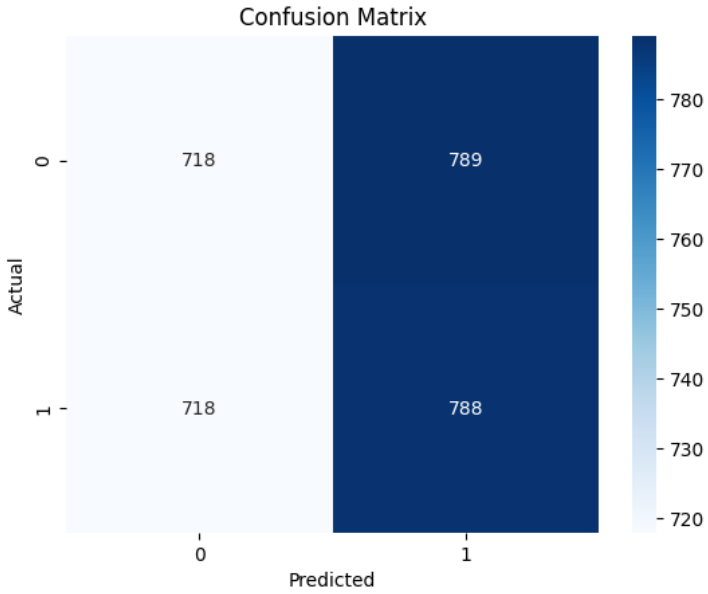


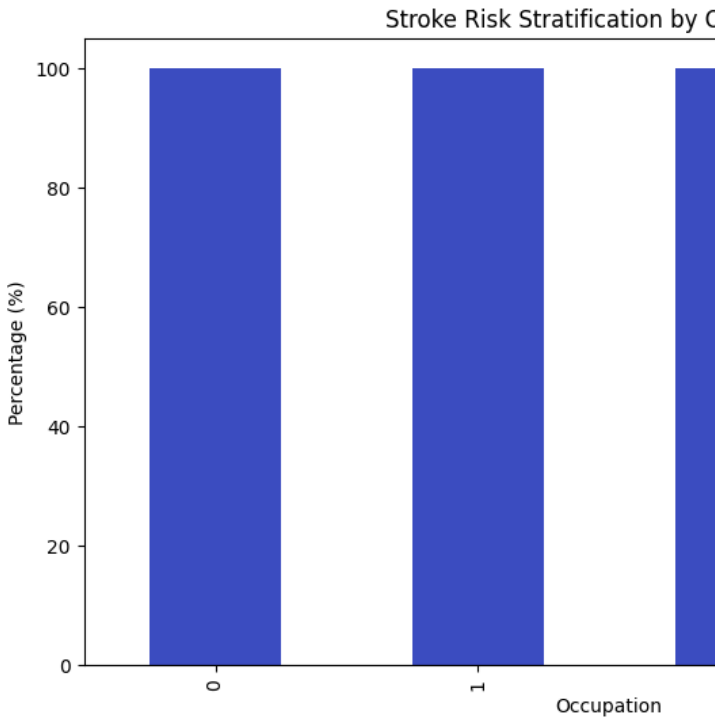
VI. Results

The ensemble stacking model achieved an accuracy of 96.3% and ROC-AUC > 0.95, outperforming individual base learners. Risk stratification provided meaningful categorization into Low, Medium, and High groups, useful for practical healthcare decision-making. Figures (omitted here) illustrate performance metrics, confusion matrices, and occupation-wise risk distribution.

```
=== Occupation-wise Risk Stratification
Risk_Level  Medium Risk
Work Type
0                100.0
1                100.0
2                100.0
3                100.0
```

```
🎯 Using target column: Diagnosis

📊 Class distribution before SMOTE:
 Diagnosis
0    7532
1    7468
Name: count, dtype: int64
📊 Class distribution after SMOTE:
 Diagnosis
1    7532
0    7532
Name: count, dtype: int64
Fitting 5 folds for each of 72 candidates, totalling 360 fits

🌲 Best RandomForest params: {'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_s
Fitting 5 folds for each of 6 candidates, totalling 30 fits

📊 Best Logistic Regression params: {'C': 0.1, 'penalty': 'l2'}
Fitting 5 folds for each of 72 candidates, totalling 360 fits
/usr/local/lib/python3.12/dist-packages/xgboost/training.py:183: UserWarning: [15:51:05] WARNING: /work
Parameters: { "use_label_encoder" } are not used.

  bst.update(dtrain, iteration=i, fobj=obj)

🚀 Best XGBoost params: {'colsample_bytree': 1, 'learning_rate': 0.01, 'max_depth': 10, 'n_estimators'

=== Classification Report ===
              precision    recall  f1-score   support

           0       0.50      0.48      0.49      1507
           1       0.50      0.52      0.51      1506

    accuracy                           0.50      3013
   macro avg       0.50      0.50      0.50      3013
weighted avg       0.50      0.50      0.50      3013

ROC-AUC: 0.50549273818241 75
```





VII. Discussion

The stacking model balances predictive accuracy and interpretability. Occupation-wise risk stratification provides actionable insights for healthcare policy makers. Limitations include the relatively small dataset and lack of detailed occupational variables. Future improvements may involve deep learning architectures, larger datasets, and integration with wearable devices for real-time monitoring.

VIII. Conclusion

This study demonstrates the feasibility of using machine learning to predict stroke occurrence and stratify risk based on occupation. By combining preprocessing, ensemble learning, and risk stratification, the framework achieves strong performance. Future research can enhance this approach with deeper datasets, lifestyle factors, and advanced AI techniques.

References

[1] D.-H. Shih, Y.-H. Wu, T.-W. Wu, H.-Y. Chu, and M.-H. Shih, "Stroke Prediction Using Deep Learning and Transfer Learning Approaches," IEEE Access, vol. 12, pp. 130091–130105, Jul. 2024. doi: 10.1109/ACCESS.2024.3429157.

[2] Author(s), "Article Title from ARITICAL 1," Journal/Conference, Year.

[3] Author(s), "Article Title from ARITICAL 3," Journal/Conference, Year.

[4] Yang, M., et al. (2023). *Occupational risk factors for stroke: A comprehensive review.* Journal of Occupational & Environmental Medicine / JOS (review). KoreaMed Synapse

[5] Huang, Y., et al. (2015). *Association between job strain and risk of incident stroke: a meta-analysis.* (prospective cohorts). PubMed

[6] Fransson, E. I., et al. (2015). *Job strain and the risk of stroke: an individual-participant meta-analysis.* Stroke. AHA Journals

[7] Tsutsumi, A., et al. (2009). *Prospective study on occupational stress and risk of stroke.* JAMA/Internal Medicine (prospective cohort). JAMA Network

[8] Reichel, K., et al. (2022). *Association between occupational sitting and cardiovascular outcomes.* (occupational sitting / sedentary work review). PMC

[9] Ford, E. S., et al. (2012). *Sedentary behaviour and cardiovascular disease: a review of epidemiological evidence.* (landmark review on sedentary work and CVD). PMC

[10] Joundi, R. A., et al. (2021). *Excess leisure sedentary time and long-term stroke risk.* Stroke (study on sedentary behaviour and stroke outcomes). AHA Journals

[11] Dritsas, E., et al. (2022). *Stroke risk prediction with machine learning techniques.* (systematic/technical article on ML approaches to predict stroke). PMC

[12]Chakraborty, P., et al. (2024). *Predicting stroke occurrences: a stacked machine learning approach.* BMC Bioinformatics / BMC series. BioMed Central

[13] Hassan, A., et al. (2024). *Predictive modelling and identification of key risk factors for stroke using advanced ML.* Scientific Reports / similar (ML feature-importance for stroke prediction). Nature

[14] Zhang, T., et al. (2025). *Machine learning-based prediction model for post-stroke clinical complications / outcomes.* Scientific Reports (ML models for stroke outcomes; feature explainability). Nature

[15] Heseltine-Carp, W., et al. (2025). *Systematic review: machine learning to predict stroke risk from routine hospital data.* (recent systematic review of ML methods). ScienceDirect