Enhancing Predictive Accuracy For A Minority Class In Imbalanced Data: An Integrated Approach With ROSE and Tomek Link

Jeong-Wook Lee[a] • Young Eun Jeon[b] • Jung-In Seo[b]

[a] Institute of Data Science and Informatics, University of Missouri, Columbia, USA

[b] Department of Data Science, Andong National University, Andong, Korea

1. Abstract

Backgrounds

- **Severely imbalanced data** are commonly encountered in scenarios where **the event of interest occurs far less frequently than other events.**
- However, **machine learning models trained on such severely imbalanced data often exhibit bias toward the majority class, resulting in high overall accuracy but poor detection of minority class data points.**
- This issue is particularly problematic in applications such as fraud detection, disease diagnosis, and spam detection, where **correctly identifying the minority class is crucial.**

Purpose

- This study focuses on **enhancing the predictive performance for imbalanced data by proposing an integrated sampling technique with random over-sampling examples (ROSE) and Tomek link.**
- **The proposed sampling technique increases data diversity** by generating synthetic data points based on a probability distribution while eliminating noisy data points, leading to a more high-quality dataset.

2. Methodology

▪ Schematic Diagram

- An overall prediction framework with the integrated sampling technique of ROSE and Tomek link is presented in Fig 1 as a schematic diagram.

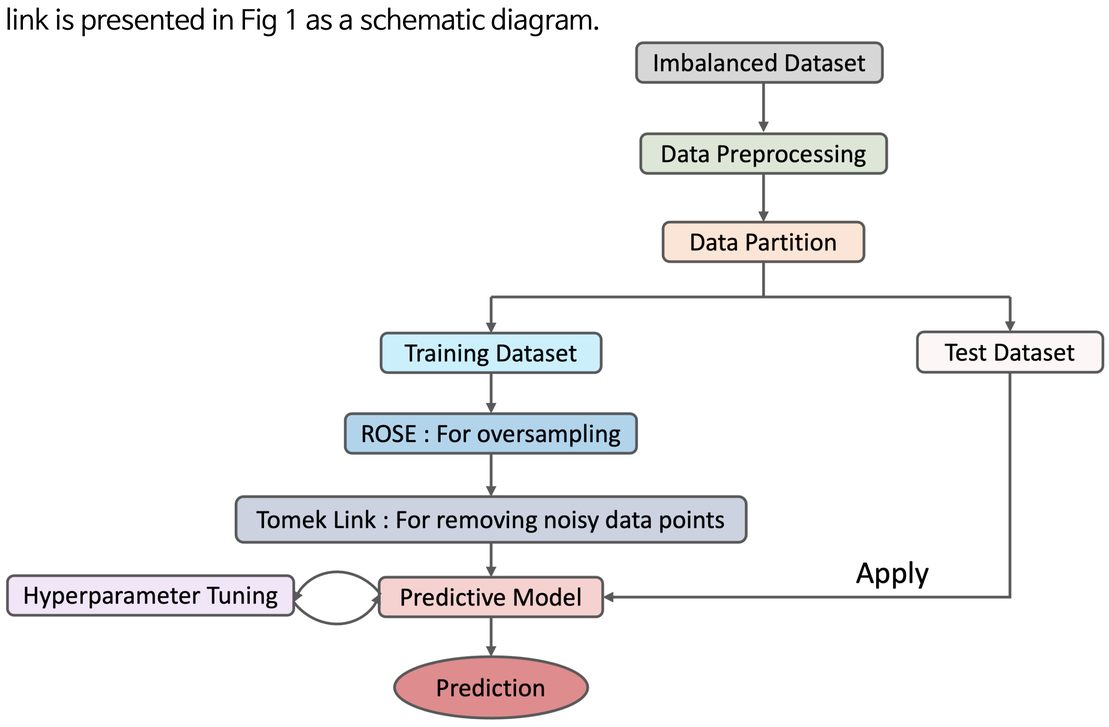


Fig 1. Prediction framework for analyzing an imbalanced dataset

- A concise description of the proposed sampling technique is provided in the subsequent section.

▪ Proposed Sampling Technique

• ROSE

✓ ROSE is an oversampling technique designed to address the class imbalance for a target variable by utilizing a smoothed bootstrap approach based on kernel density estimation.

Let $\boldsymbol{T}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a training dataset with n observations, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ represents a set of p features for the ith observation and $y_i \in \{\mathcal{Y}_0, \mathcal{Y}_1\}$ is the corresponding class in a target variable. Additionally, let the number of observations in class $\mathcal{Y}_j$ $(j = 0, 1)$ be denoted by $n_j (< n)$. Then, ROSE is conducted through the following steps:

Step 1. Select class $y^* = \mathcal{Y}_j$ with equal probability (i.e., $P(\mathcal{Y}_j) = 0.5$).

Step 2. Select $(\boldsymbol{x}_i, y_i) \in \boldsymbol{T}_n$, such that $y_i = y^*$, with a probability $1/n_j$.

Step 3. Sample $\boldsymbol{x}^*$ from $K_{H_j}(\cdot, \boldsymbol{x}_i)$, where H_j denotes a smoothing matrix and K_{H_j} is a probability distribution centered at $\boldsymbol{x}_i$ and H_j. To be specific, $\boldsymbol{x}^*$ can be generated as $\boldsymbol{x}^* = \boldsymbol{x}_i + H_j \epsilon$, where ϵ is a realization from a standard normal distribution.

ROSE Algorithm

• Tomek Link

✓ Tomek link is used to identify and remove data points located at the boundaries between cl asses in a dataset, clarifying class boundaries.

3. Application

▪ Stroke Dataset

- To evaluate the validity and practical applicability of the proposed sampling technique, **a stroke dataset with a serious imbalance ratio of 98:2** is employed.
- For analysis, 43,400 patients are partitioned into training and test datasets at a 7:3 ratio, and Table 1 describes the variables used.

Table 1. Description of variables for the stroke dataset

Role	Variable	Discription	Value
Feature	*gender*	Gender	`Male, Female, Other`
	age	Age of the individual	
	hypertension	Hypertension status	`0: No 1: Yes`
	heart_disease	Heart disease status	`0: No 1: Yes`
	ever_married	Marital status	`No, Yes`
	work_type	Type of work	`Children, Private, Govt_job, Never_worked, Self-employed`
	Residence_type	Type of residence	`Rural, Urban`
	avg_glucose_level	Average glucose level	
	bmi	Body mass index	
Target	*stroke*	Stroke status	`0: No 1: Yes`

▪ Prediction Framework

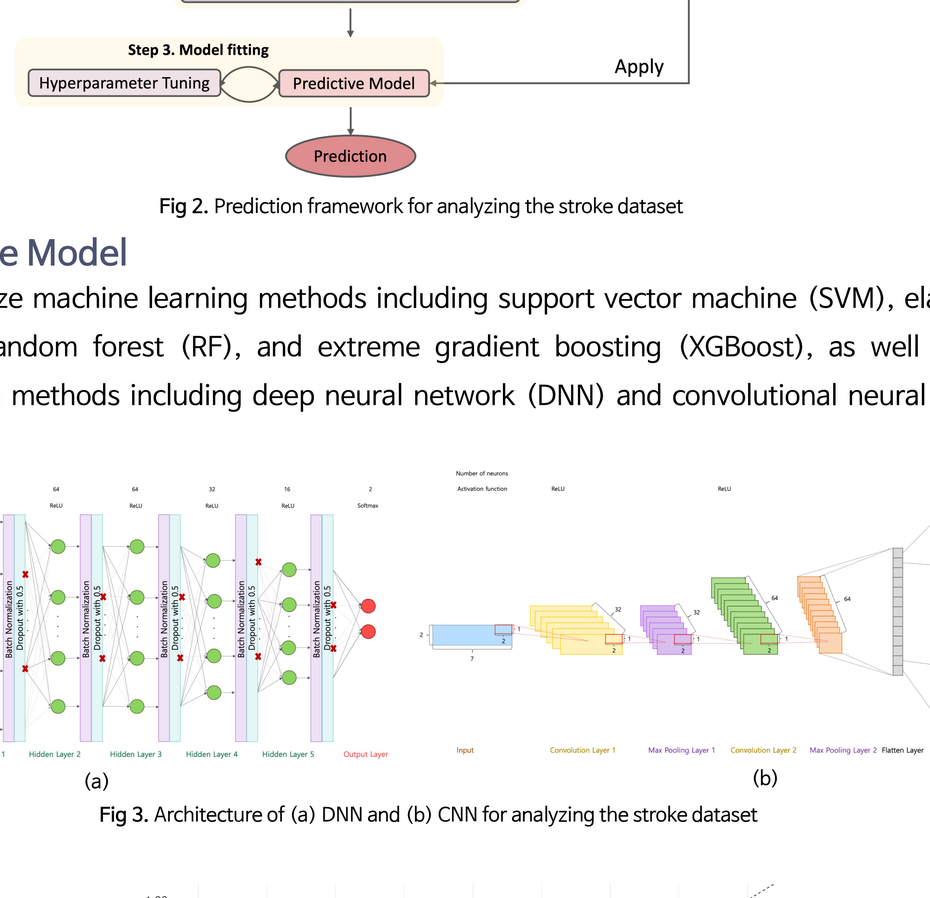


Fig 2. Prediction framework for analyzing the stroke dataset

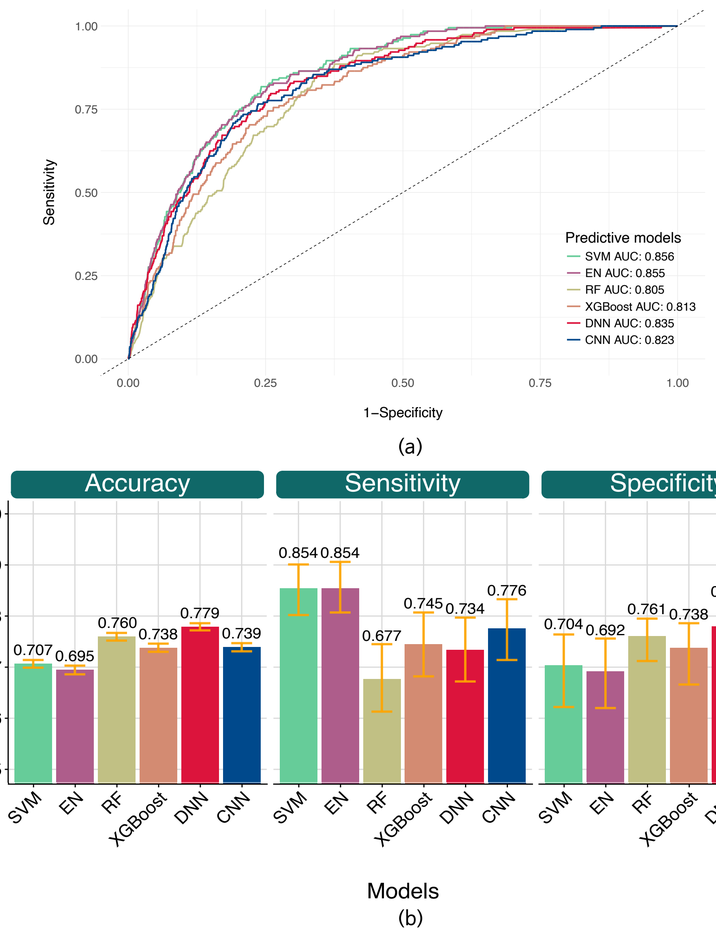
▪ Predictive Model

- We utilize machine learning methods including support vector machine (SVM), elastic-net (EN), random forest (RF), and extreme gradient boosting (XGBoost), as well as deep learning methods including deep neural network (DNN) and convolutional neural network (CNN).

(a) (b)

Fig 3. Architecture of (a) DNN and (b) CNN for analyzing the stroke dataset

▪ Results



(a) (b)

Fig 4. (a) ROC curves with AUC values and (b) bar plots for evaluation metrics with 95% CIs of all predictive models

4. Discussions and Conclusions

- **DNN and CNN demonstrate stable performance across all metrics**, which reveals that the two models are highly competitive in predicting whether a stroke occurs or not in patients when applying the proposed sampling method.
- Our findings reveal that **the proposed method can significantly improve the prediction of minority class data points, leading to more reliable and efficient predictive modeling in critical healthcare applications.**